# Analysis of Economics Data: An Introduction to Econometrics

© A. Colin Cameron, March 29, 2015.
Department of Economics, University of California - Davis.
Printed for ECN 102: Analysis of Economic Data

# Contents


**1 © A. Colin Cameron: Analysis of Economics Data** **1**
1.1 Statistical Methods . . . 1
1.2 Types of Data . . . 2
1.3 Regression Analysis . . . 4
1.4 Key Concepts . . . 5
1.5 Exercises . . . 6

**2 © A. Colin Cameron: Univariate Data Summary** **7**
2.1 Summary Statistics for Numerical Data . . . 7
2.2 Charts for Numerical Data . . . 15
2.3 Categorical Data . . . 20
2.4 Charts for Numerical Data by Category . . . 20
2.5 Key Concepts . . . 22
2.6 Exercises . . . 23

**3 © A. Colin Cameron: Economics Data** **27**
3.1 Output Measures . . . 28
3.2 Price Indexes and Price Inflation . . . 30
3.3 Labor Force Statistics . . . 32
3.4 Financial Data . . . 35
3.5 International Data . . . 38
3.6 Individual-Level Data . . . 39
3.7 Key Concepts . . . 39

**4 © A. Colin Cameron: Univariate Data Transformation** **41**
4.1 Natural Logarithm . . . 41
4.2 Other Transformations . . . 49
4.3 Key Concepts . . . 54
4.4 Exercises . . . 55
4.5 Appendix 4.A: Derivations for the Natural Logarithm . . . 58

**5 © A. Colin Cameron: The Sample Mean 61**
5.1 Sampling from a Finite Population . . . 62
5.2 Population and Sample . . . 64
5.3 Sample Generated by an Experiment . . . 67
5.4 Properties of the Sample Mean . . . 69
5.5 t Statistic . . . 72
5.6 Point Estimation . . . 74
5.7 Nonrandom Samples . . . 76
5.8 Key Concepts . . . 78
5.9 Exercises . . . 79
5.10 Appendix 5.A: Some Essentials of Probability . . . 83
5.11 Appendix 5.B: Results for the Sample Mean . . . 88
5.12 Appendix 5.C: Computer Generation of a Random Sample . . . 90

**6 © A. Colin Cameron: Statistical Inference for the Mean 91**
6.1 Example: Mean Annual Earnings . . . 91
6.2 t Statistic and t Distribution . . . 92
6.3 Confidence Intervals . . . 96
6.4 Two-Sided Hypothesis Tests . . . 100
6.5 Test Power . . . 105
6.6 Hypothesis Testing Examples . . . 107
6.7 Key Concepts . . . 110
6.8 Exercises . . . 112

**7 © A. Colin Cameron: Statistical Inference Extensions 117**
7.1 One-Sided Hypothesis Tests . . . 117
7.2 Generalization of Confidence Intervals and Hypothesis Tests . . . 121
7.3 Difference in Two Means . . . 123
7.4 Proportions Data . . . 126
7.5 Key Concepts . . . 129
7.6 Exercises . . . 130
7.7 Appendix 7.A: Standard Error for Difference in Means . . . 130
7.8 Appendix 7.B: Standard Error for Proportions Data . . . 130

**8 © A. Colin Cameron: Bivariate Data Summary 133**
8.1 Example: House Price and Size . . . 133
8.2 Two-way Tabulation . . . 135
8.3 Two-way Scatter Plot . . . 136
8.4 Correlation . . . 137
8.5 Regression Line . . . 141

8.6 Further Details on Regression . . . . . 145
8.7 Measures of Model Fit . . . . . 147
8.8 Computer Output following OLS Regression . . . . . 150
8.9 Regression and Correlation . . . . . 152
8.10 Causation . . . . . 153
8.11 Nonparametric Regression . . . . . 155
8.12 Key Concepts . . . . . 156
8.13 Exercises . . . . . 158
8.14 Appendix 8.A: Regression Computation Example . . . . . 159

**9 © A. Colin Cameron: The Least Squares Estimates 161**
9.1 Population and Sample . . . . . 161
9.2 Examples of Sampling from a Population . . . . . 165
9.3 Properties of the Least Squares Estimator . . . . . 169
9.4 t-Statistic . . . . . 175
9.5 Point Estimation . . . . . 176
9.6 Key Concepts . . . . . 178
9.7 Exercises . . . . . 179
9.8 Appendix 9.A: Conditional Mean and Variance . . . . . 179
9.9 Appendix 9.B: Population Covariance and Correlation . . . . . 182
9.10 Appendix 9.C Derivation of Properties of OLS . . . . . 183

**10 © A. Colin Cameron: Statistical Inference for Bivariate Regression 187**
10.1 Example: House Price and Size . . . . . 187
10.2 t Statistic . . . . . 188
10.3 Confidence Intervals . . . . . 189
10.4 Tests of Statistical Significance . . . . . 192
10.5 Two-Sided Hypothesis Tests . . . . . 195
10.6 One-Sided Hypothesis Tests . . . . . 200
10.7 Robust Standard Errors . . . . . 202
10.8 Presentation of Regression Results . . . . . 205
10.9 Prediction . . . . . 207
10.10 Examples . . . . . 211
10.11 Key Concepts . . . . . 211

**11 © A.Colin Cameron: Bivariate Case Studies 215**
11.1 Capital Asset Pricing Model . . . . . 215
11.2 Output and Unemployment . . . . . 219
11.3 Health Outcomes . . . . . 222
11.4 Health Expenditures . . . . . 226

11.5 Exercises . . . 229

**12 © A. Colin Cameron: Bivariate Data Transformation 231**
12.1 Example: Earnings and Age . . . 231
12.2 Indicator Variable Model . . . 232
12.3 Nonlinear Relationships . . . 234
12.4 Transformation of the Regressor . . . 238
12.5 Transformation of the Dependent Variable . . . 239
12.6 Elasticities and Semi-elasticities . . . 243
12.7 Comparison of Models . . . 246
12.8 Key Concepts . . . 247
12.9 Exercises . . . 248
12.10 Appendix 12.A: Calculus and Derivatives . . . 249

**13 © A. Colin Cameron: Data Summary with Multiple Regression 251**
13.1 Example: House Price and Characteristics . . . 251
13.2 Two-way Scatterplots . . . 254
13.3 Correlation . . . 255
13.4 Regression Line . . . 256
13.5 Estimated Partial Effects . . . 258
13.6 Model Fit . . . 261
13.7 Computer Output Following Multiple Regression . . . 264
13.8 Inestimable Models . . . 266
13.9 Key Concepts . . . 268
13.10 Exercises . . . 270

**14 © A. Colin Cameron: Inference for Multiple Regression 271**
14.1 Properties of the Least Squares Estimates . . . 271
14.2 Point Estimation . . . 276
14.3 Confidence Intervals . . . 277
14.4 Hypothesis Tests on a Single Parameter . . . 278
14.5 Joint Hypothesis Tests . . . 281
14.6 Inference with Robust Standard Errors . . . 288
14.7 More Precise Estimation . . . 293
14.8 Nonrandom Samples . . . 294
14.9 Prediction . . . 296
14.10 Key Concepts . . . 299
14.11 Exercises . . . 301
14.12 Appendix 14.A: OLS with Matrix Algebra . . . 301

**15 © A. Colin Cameron: Multivariate Case Studies 305**
15.1 School Academic Performance . . . 305
15.2 Cobb-Douglas Production Function . . . 310
15.3 Phillips Curve . . . 314
15.4 From Raw Data to Final Data . . . 319

**16 © A. Colin Cameron: Multivariate Data Transformation 325**
16.1 Example: Earnings, Age, Education and Type of Worker . . . 325
16.2 Single Indicator Variable . . . 326
16.3 Sets of Indicator Variables . . . 330
16.4 Quadratic Model and Polynomial Models . . . 335
16.5 Interacted Regressors . . . 339
16.6 Natural Logarithm Transformations . . . 341
16.7 Key Concepts . . . 343

**17 © A. Colin Cameron: Data and Model Misspecification 345**
17.1 Data Example: Democracy and Growth . . . 345
17.2 Model Diagnostics . . . 347
17.3 Model Specification Tests . . . 350
17.4 Data Problems: Multicollinearity . . . 351
17.5 Failure of Model Assumptions . . . 352
17.6 Incorrect Population Model . . . 353
17.7 Regressors Correlated with Errors . . . 355
17.8 Nonconstant Error Variance . . . 356
17.9 Correlated Errors . . . 357
17.10 Conclusion . . . 358
17.11 Summary . . . 358

**18 © A. Colin Cameron: DRAFT Causality 359**
18.1 Experiments . . . 360
18.2 Regression . . . 360
18.3 Differences in Differences . . . 360
18.4 Panel Data . . . 360
18.5 Matching . . . 360
18.6 Weighting . . . 360
18.7 Regression discontinuity design . . . 360
18.8 Instrumental Variables . . . 360

**19 © A. Colin Cameron: Cross-section Data** **361**
19.1 Data Example . . . 362
19.2 Heteroskedastic Errors . . . 363
19.3 Clustered Errors . . . 366
19.4 Models for Binary Outcome Data . . . 370
19.5 Survey Data . . . 376
19.6 Key Concepts . . . 377
19.7 Appendix 19.A: Weighted Least Squares . . . 378
19.8 Appendix 19.B: Nonlinear Models . . . 379
19.9 Appendix 19.C: Categorical Data . . . 384

**20 © A. Colin Cameron: Time Series Data** **385**
20.1 Example: Ten-year and One-year Interest Rates . . . 386
20.2 Single Time Series: Persistence . . . 387
20.3 Nonstationary Time Series . . . 392
20.4 Spurious Regression . . . 395
20.5 Static Regression . . . 400
20.6 Dynamic Regression . . . 404
20.7 Key Concepts . . . 408
20.8 Exercises . . . 409
20.9 Appendix 20.A: Autocorrelations . . . 409
20.10 Appendix 20.B: Unit Root and Cointegration Tests . . . 410
20.11 Appendix 20.C: Additional Time Series Models . . . 413

**21 © A. Colin Cameron: Panel Data** **417**
21.1 Data Example: NBA Team Revenue . . . 418
21.2 Pooled OLS Estimation . . . 421
21.3 Random Effects Estimation . . . 425
21.4 Fixed Effects Estimation . . . 428
21.5 Long Panels . . . 434
21.6 Key Concepts . . . 435
21.7 Additional Models and Estimators . . . 436

**A © A. Colin Cameron: Computing** **439**
A.1 General Issues . . . 439
A.2 Stata . . . 443
A.3 Eviews . . . 447
A.4 Gretl . . . 450
A.5 R . . . 453
A.6 Excel and Google Sheets . . . 457

A.7 Critical values and p-values . . . . . 461

# List of Figures


1.1 Linear regression example . . . . . 5

2.1 Normal distribution: Probability of being within one, two or three standard deviations of the mean. . . . . 12
2.2 Histograms for symmetric, right-skewed and left-skewed data. . . . . 13
2.3 Box Plot: Annual earnings of 30 year-old female full-time workers in 2010. . 16
2.4 Histogram with two different bin widths: Annual earnings of 30 year-old female full-time workers in 2010. . . . . 18
2.5 Smoothed histogram (kernel density estimate) with two different window widths: Annual earnings of 30 year-old female full-time workers in 2010. . . . 19
2.6 Line Chart: Real GDP per capita in U.S. (in 2005 dollars). . . . . 19
2.7 Pie Chart: Fishing site. . . . . 21
2.8 Column Chart: Health expenditures in U.S. in 2013 (in $billion). . . . . 23

3.1 Annual growth rate in U.S. real GDP from 1930. . . . . 28
3.2 Annual U.S. price inflation (CPI-U) from 1930. . . . . 31
3.3 U.S. Employment since 1939. . . . . 33
3.4 U.S. Unemployment rate since 1948. . . . . 35
3.5 U.S. Labor Force Participation Rate since 1948. . . . . 36
3.6 U.S. Treasury security interest rates since 1954. . . . . 37

4.1 Exponential function and natural logarithm function . . . . . 42
4.2 Standard and Poors 500 index: Levels and natural logarithm. . . . . 44
4.3 Annual earnings in levels and in natural logarithms: Histograms and kernel density estimates. . . . . 48
4.4 Monthly total revolving consumer credit outstanding: Smoothing by moving average and by seasonal adjustment. . . . . 50
4.5 GDP and per capita GDP in U.S.: nominal in current dollars and real in 2009 dollars. . . . . 52

5.1 1880 Census: Histogram of age for the entire population. . . . . 62

5.2 1880 Census: Histogram for one randomly-drawn sample (n = 25) and histogram with kernel density estimate for 100 sample means of age (n = 25). . 63
5.3 Coin tosses: Histogram for one sample (n = 30) and histogram with kernel density estimate for 400 sample means (n = 30). . . . . 68
5.4 $\Pr[X \leq 1.5]$ and $\Pr[0.5 \leq X \leq 1.5]$ for standard normal distribution. . . . . 87

6.1 Student's $t$ distribution: $T(4)$ and $T(30)$ compared to the standard normal. . 94
6.2 Student's t distribution: t-values $t_{v,\alpha}$ and $t_{v,\alpha/2}$ for $v = 170$ and $\alpha = 0.05$. . . 95
6.3 Two-sided hypothesis test when $t = 0.724$, $n = 171$ and $\alpha = 0.05$. . . . . 103

7.1 One-sided hypothesis test (upper one-tailed alternative) when $t = 0.724$, $n = 171$ and $\alpha = 0.05$. . . . . 120

8.1 House price and size: Scatterplot and the four quadrants used to calculate the correlation coefficient. . . . . 137
8.2 Correlation: Examples of different strengths and direction of correlation. . . 140
8.3 Residual (here negative): The vertical difference between the data point and the regression line. . . . . 142
8.4 Regression of house price on size. . . . . 144
8.5 R-squared components: Total sum of squares and explained sum of squares for five observations. . . . . 148
8.6 Nonparametric regression of house price on size: local linear and lowess. . . . 156

9.1 Error is difference between y and the population line while residual is difference between y and fitted regression line. . . . . 163
9.2 Generated data: Scatterplot with true population line and with fitted regression line. . . . . 166
9.3 Generated data: Scatterplots and fitted regression lines from three random samples (n = 30). . . . . 167
9.4 Generated data: Histogram with kernel density estimates for slopes and intercepts from 400 regressions (n=30). . . . . 168
9.5 1880 Census data: Histogram with kernel density estimates for slopes and intercepts from 400 regressions (n=200). NOTE: PANELS SAY 100 IN ERROR – IT WAS 400. . . . . 169

10.1 Hypothesis tests: $p$-value approach and critical region approach. . . . . 196
10.2 . . . . . 210

11.1 CAPM Model: Coca Cola and market excess returns . . . . . 218
11.2 Okun's Law: Percentage change in output and change in unemployment rate 220
11.3 Health Outcomes: Life expectancy and health spending per capita . . . . . 224

11.4 Health Expenditures: Relationship with GDP per capita . . . . . 228

12.1 Earnings OLS Regressions: Age as regressor and indicator variable as regressor. 233
12.2 Nonlinear model examples . . . . . 235
12.3 Marginal effect computation: Calculus method and finite difference method . 236

13.1 House price data: Two-way scatterplots . . . . . 254

14.1 F distribution: F(3,50) and F(10,50) . . . . . 282

15.1 Academic Performance Index: Histogram and Density . . . . . 307
15.2 Academic Perfomance Index: Regression on Parents' Education . . . . . 308
15.3 Cobb-Douglas Production Function: Predicted Output and Fitted Isoquants 313
15.4 Inflation and unemployment rates from 1949 to 1969 and the associated Phillips curve. . . . . 316
15.5 Inflation and unemployment rates from 1970 to 2014 and the associated Phillips curve. . . . . 317

16.1 Indicator variable (left panel) and Interacted Indicator Variable (right panel) as Regressor. . . . . 328
16.2 Various quadratic models: The coefficient of $x^2$ is positive for the top three panels and negative for the lower three panels. . . . . 336

17.1 Democracy and Growth: Scatter plot . . . . . 347
17.2 Democracy regression: Actual values plotted against fitted values and OLS residuals plotted against fitted values . . . . . 349

19.1 Heteroskedastic data: Scatter plots of earnings and earnings residual against education. . . . . 365
19.2 Binary outcome data: predictions from logit, probit and OLS models. . . . . 372

20.1 Annualized rate of return at constant maturity on 1-year and 10-year U.S. Treasury bills from January 1982 to January 2015. . . . . 387
20.2 Changes in annualized rate of return on 1-year and 10-year U.S. Treasury bills from February 1982 to January 2015. . . . . 399

21.1 Natural logarithm of team earnings plotted against number of wins in previous season for 29 teams over 10 seasons. . . . . 419
21.2 Panel data: (1) Natural logarithm of team earnings plotted against season for 8 teams. (2) Number of wins least season earnings plotted against season for 8 teams. . . . . 420

21.3 Within variation in natural logarithm of team earnings plotted against wintin variation in number of wins in previous season for 29 teams over 10 seasons. 431

A.1 Stata: Sample output. . . . . 446
A.2 Eviews: Sample output. . . . . 451
A.3 Gretl: Sample output. . . . . 454
A.4 R: Sample output. . . . . 458
A.5 Excel: Sample output. . . . . 460

# List of Tables


1 Book Chapters. . . . . iv

2.1 Summary statistics: Annual earnings of 30 year-old female full-time workers in 2010. . . . . 8
2.2 Frequencies in bins of width 15,000: Annual earnings of 30 year-old female full-tine workers in 2010. . . . . 17
2.3 Tabulation of frequencies for fishing sites. . . . . 20
2.4 U.S. Health Expenditures in 2013 in billions of dollars by category . . . . . 22

3.1 Contractions in U.S. Economy since 1930. . . . . 29
3.2 U.S. Consumer Price Index (CPI-U) since 1930. . . . . 31

4.1 Approximating ln(1+x) by x. . . . . 43
4.2 Approximating exp(x) by 1+x. . . . . 43
4.3 Approximating proportionate change in x by the change in ln(x). . . . . 46
4.4 Number of periods for investment at percentage rate r per period to double. 47
4.5 Nominal and real data: Calculations. . . . . 53
4.6 Exponential and natural logarithm functions: Properties. . . . . 59

6.1 Sample mean: Annual earnings of female full-time workers aged 30 in 2010. . 92
6.2 Student's t distribution: t-values for various degrees of freedom and confidence levels. . . . . 98
6.3 Hypothesis Tests: Type I and Type II Errors. . . . . 106
6.4 Descriptive Statistics: Gasolin price per gallon at 32 gas stations. . . . . 107
6.5 Descriptive Statistics: Annual earnings of male full-time workers aged 30 in 2010. . . . . 108
6.6 Descriptive Statistics: Annual inflation rate using quarterly data from 1948 to 2012. . . . . 109

7.1 Summary statistics: Annual earnings of female and male full-time workers aged 30 in 2010. . . . . 125

8.1 House price and size: Complete listing of data. . . . . 134
8.2 House price and size: Summary statistics. . . . . 134
8.3 House price and size: Cross tabulation with row percentages. . . . . 135
8.4 House price and size: Cross tabulation with expected frequencies. . . . . 136
8.5 House price: Computer output from regression on house size. . . . . 151
8.6 Regression: Computation example for a sample of size five. . . . . 160

9.1 Generated data: Model y = 1 + 2x + u where u is N(0, 4) distributed. . . . 165

10.1 House price and size: Regression estimates with default standard errors. . . . 187
10.2 Hypothesis Tests: Type I and Type II Errors. . . . . 198
10.3 Hypothesis tests: Summary for tests on the slope parameter. . . . . 202
10.4 Regression estimates: Various ways to report the results. . . . . 207

11.1 CAPM data: Variable definitions and summary statistics . . . . . 217
11.2 Output and unemployment data: Variable definitions and summary statistics 220
11.3 Health data: Variable definitions and summary statistics . . . . . 222
11.4 Health data: Variable definitions and summary statistics . . . . . 227

12.1 Annual Earnings of male full-time workers aged 25-65 in 2010: Variable definitions and summary statistics. . . . . 232
12.2 Summary for models involving natural logarithms . . . . . 245
12.3 Regression of Earnings (y) on Age (x) for various transformation models of the variables. . . . . 246
12.4 Derivatives: Leading Examples. . . . . 250

13.1 House price data: Variable definitions and summary statistics . . . . . 252
13.2 House Price data: Complete listing. . . . . 253
13.3 House price data: correlations for 29 houses sold in Central Davis . . . . . 256
13.4 House price: Computer output from multiple regression. . . . . 265

14.1 Summary for hypothesis tests on the slope parameter in multiple regression. 279
14.2 House price: Computer output from multiple regression. . . . . 281
14.3 F distribution with q and n-k degrees of freedom: test critical values. . . . . 284
14.4 Robust standard errors: Leading examples. . . . . 288
14.5 House price data: OLS regression with heteroskedastic-robust standard errors. 290

15.1 Academic Performance Index data: Variable definitions and summary statistics (n=807) . . . . . 306
15.2 Academic Performance Index data: pairwise correlations . . . . . 308
15.3 Academic Performance Index: Results from multiple regression. . . . . 309

15.4 Cobb-Douglas Data: Output (Q), Capital (K) nd Labor (L) for U.S. from 1899 to 1922 (n=24). . . . . 311
15.5 Phillips curve data: Variable definitions and summary statistics . . . . . 315
15.6 Survey data: Format of raw data file. . . . . 320
15.7 Summary statistics: Survey data before cleaning. . . . . 321
15.8 Summary statistics: Survey data after cleaning. . . . . 324

16.1 Annual Earnings of male full-time workers aged 25-65 in 2010: Variable definitions and summary statistics. . . . . 326
16.2 Sets of indicator variables: Earnings regression with default standard errors in parentheses and t-statistics in square brackets. . . . . 332

17.1 Democracy and Growth: Variable definitions and summary statistics. . . . . 346

19.1 Annual Earnings of male full-time workers aged 25-65 in 2010: Variable definitions and summary statistics . . . . . 362
19.2 Earnings: OLS estimates with default standard errors. . . . . 362
19.3 Earnings: OLS regression with heteroskedastic-robust standard errors. . . . . 364
19.4 Earnings: OLS regression with cluster-robust standard errors. . . . . 370
19.5 Earnings: OLS, random effects and fixed effects estimates with default and cluster-robust standard errors. . . . . 374
19.6 High earnings: Classification table folloiwng logit regression. . . . . 374
19.7 Nonlinear models: Leading examples for different data types. . . . . 381

20.1 Ten-year and one-year U.S.Treasury Notes from January 1982 to January 2015: Variable definitions and summary statistics . . . . . 386
20.2 Interest rates: Autocorrelations and residual autocorrelations. . . . . 391
20.3 Interest rates: HAC standard error of slope coefficient for various choices of lag length m . . . . . 402
20.4 Static and Dynamic Models: Change in ten-year interest rates. . . . . 406

21.1 NBA Team revenue: Variable definitions and summary statistics (n=29, T=10).418
21.2 NBA Team Revenue: Pooled OLS with various standard error estimation methods. . . . . 423
21.3 NBA Team Revenue: Pooled OLS with various models and cluster-robust t-statistics. . . . . 424
21.4 NBA Team Revenue: Pooled OLS with various models and cluster-robust standard errors. . . . . 427

A.1 Statistical packages: File-name extensions for script files and data files. . . . 441

A.2 Statistical packages: commands to calculate 95 percent confidence interval critical values. . . . . 461

A.3 Statistical packages: commands to calculate p-values for a two-sided test. . . 462

# Preface

### Motivation

Anyone with a personal computer, or even a tablet with an internet connection, has extraordinary ability to obtain and analyze data compared even to a university professor at the cutting edge a couple of decades ago. Such data analysis should be part of the skill set of anyone graduating with a major in an economics-related discipline, yet not all economic programs succeed in providing this training.

The goal of this book is to provide an introduction to these methods. This is a more modest and realistic goal than that in many first courses in econometrics that aim high and can be the hardest course in an economics major. Analysis of economics data using a computer is emphasized. The key statistical methods, notably confidence intervals and hypothesis tests, are explained but there is less emphasis on theory than in a typical econometrics course.

Ideally students using this book can be encouraged to take further course(s) in econometrics and statistics, just as students take several courses (introductory, intermediate and then field courses) rather than a single course in order to learn basic microeconomics. It is unrealistic to expect mastery of econometrics in just one course.

### Data Analysis

The book takes a learning-by-doing approach. The key requirement is use of an econometrics or statistical package. The website for the book provides all the datasets and computer code for repeating the analysis in the book using leading commercial econometrics packages – in alphabetical order **Eviews** and **Stata** – and using the free packages **Gretl** and **R**. The tables and figures in this book were based on output from Stata 13. The spreadsheet programs **Excel** and **Google Sheets** can also be used, but are more limited.

For the particular statistical package that is chosen to use with this book, the instructor and student can easily work through each chapter using the datasets and computer code that are all available at the book website. **This is by far the best way to learn the material.** The book itself is limited to presenting key summary tables, but not the specific commands and consequent computer output, as these vary with the package used in instruction.

Appendix A summarizes key commands for the various statistical packages. Datasets are referred to in the book using capital letters without any file extension, and are available in three formats that should be readable using even very old versions of software. For example, the dataset called HOUSE in the text is available as file HOUSE.dta, a Stata 11 dataset, as file HOUSE.csv, a comma-separated values text file, and as file HOUSE.xls, an Excel 97-2002 worksheet. The datasets generally have less than 1,000 observations, enabling analysis using the cheaper student versions of the commercial packages that restrict the sample size.

### How to Use this Book

The key to understanding the material is to analyze data using a statistical package. As already noted, the data and commands for the data analysis in each chapter are available at the book website.

The exercises at the end of each chapter are mainly learning-by-doing exercises using a very wide range of datasets that can be obtained from the book website.

**Book Outline**

These notes, prepared in February 2015, are an almost final draft of a book to be published by W.W. Norton.

This version is exclusively for University of Sydney students taking ECMT1020. At this stage the book website is not ready, but datasets will be made available to students. And at this stage many of the chapter exercises are not included. Finally, there will be typographical errors in this draft, for which I apologise.

The book is divided into four parts. The first three parts consider, in turn, univariate, bivariate, and multivariate data analysis. This enables the reader to be exposed three times to the same fundamental statistical concepts of confidence intervals and hypothesis tests that are used to extrapolate from the sample at hand to the population. The fourth part of the book presents material beyond basic regression.

The four parts of the book are as follows.

I. Chapters 2–7: Univariate analysis – data summary, data transformation, and statistical inference for the sample mean.

II. Chapters 8–12: Bivariate analysis – data summary, statistical inference, case studies and data transformation for two series.

III. Chapters 13–16: Multivariate analysis for several series – multiple regression, statistical inference, case studies and data transformation using multiple regression.

IV. Chapters 17–21: Further topics – model misspecification, causality and special issues that arise with cross-section data, time series data and panel data.

An appendix provides instructions on how to analyze data using a statistical package.

**For the Instructor**

The heart of the book is regression analysis: modelling the relationship between one variable and one or more related variables. The book provides students with sufficient background to run basic regressions and to read regression output tables in economics articles based on more advanced estimation methods. A goal of this book is to prepare students to read such articles, so that economic field courses can include readings with regression applications.

Table 1: Book Chapters.

| PART | Ch. | Title | Essentials |
|---|---|---|---|
| I: UNIVARIATE | 1 | Analysis of Economics Data | x |
| | 2 | Univariate Data Summary | x |
| | 3 | Economics Data | |
| | 4 | Univariate Data Transformation | |
| | 5 | The Sample Mean | |
| | 6 | Inference for the Sample Mean | x |
| | 7 | Statistical Inference Extensions | |
| BIVARIATE | 8 | Bivariate Data Summary | x |
| | 9 | The Least Squares Estimates | |
| | 10 | Inference for Bivariate Data | x |
| | 11 | Case Studies for Bivariate Regression | |
| | 12 | Bivariate Data Transformation | |
| MULTIVARIATE | 13 | Multiple Regression | x |
| | 14 | Inference for Multiple Regression | x |
| | 15 | Case Studies for Multiple Regression | |
| | 16 | Multivariate Data Transformation | |
| FURTHER TOPICS | 17 | Model Misspecification | x |
| | 18 | Causality | |
| | 19 | Cross-section Data | |
| | 20 | Time Series Data | |
| | 21 | Panel Data | |
| APPENDIX | A | Computer Packages | |

It is assumed that students have studied introductory economics. This is not really essential for the basic examples, but for some examples, such as the Cobb-Douglas production function or the Phillips curve, it is helpful to have taken an introductory course in microeconomics or macroeconomics. The data transformation chapters emphasize the natural logarithm and elasticities that are used much more in economics than in other fields.

No previous exposure to regression analysis is assumed.

Ideally an introductory course in probability and statistics is taken before a course based on this book. This means exposure to univariate statistical inference, essentially confidence intervals and hypothesis tests on the population mean $\mu$ based on the sample mean. My own experience is that even if students have taken such a course, many do not understand or do not recall statistical inference. Accordingly students benefit greatly from seeing the material a second time. Furthermore, some instructors may prefer to teach this course to students with no background in statistics.

I have therefore written the book to accommodate students with a weak or nonexistent statistical background. The essentials of probability are covered briefly in Chapter 5, with

more thorough coverage given in appendices to that chapter. A complete presentation of statistical inference for univariate data is given in Chapter 6. More generally, compared to an introductory probability and statistics course Chapters 5-7 place more emphasis on statistical inference and less on probability theory.

The goal of this book is to present methods with minimal use of mathematics. Some mathematics is nonetheless necessary, especially summation notation presented in Chapter 2.1. More mathematical material and some derivations are pushed into chapter appendices. Changes in one variable with respect to another are presented using delta notation, though at times connections to derivatives is made for the benefit of those with a calculus background. Less-prepared students may find it possible to gloss over much of the mathematics. The emphasis of the book is in the interpretation of statistical output rather than dexterity with statistical formulae. For those who want more, some derivations for the OLS estimator are given in an appendices to Chapter 9 and 14.

I have generally presented regression results based on default standard errors, as these can be obtained using any spreadsheet or package. But robust standard errors are introduced early, in Chapter 10, and particularly in the case studies chapter there is discussion of instead using appropriate robust standard errors and what difference this makes.

The chapters can be covered in one or two seventy-five minute lectures. To simplify the exposition of methods, my approach is to work with the one data example throughout a chapter, or even across several chapters in the case of modelling the sale price of a house. At the same time many additional datasets are introduced throughout the book – at the end of some of the chapters, in the case studies chapters, and in the many exercises at the end of this chapter. Some data examples come from empirical research articles published in leading economics journals that were deliberately chosen in the belief that the articles would be intelligible to undergraduate students.

**Course Outline**

In a ten-week quarter-long course with two lectures a week it is possible to cover the key material in Chapters 1-17, including perhaps one of the case studies in each of the case studies chapters. In the book chapter summary above I have indicated what are the most essential chapters. In a semester-long course one could cover most of the book. For those with students who have a solid enough statistical background to skip Part I it may be possible to cover all of the remainder of the book.

# Chapter 1

# © A. Colin Cameron: Analysis of Economics Data

Statisticians specialize in data analysis, and offer courses that cover many of the statistical techniques in this course.

What distinguishes this book is the use of economics data, and interpretation of this data as an economist. This leads to additional issues that are not routinely emphasized in statistics courses and, going the other way, emphasis on only a small subset of the many available statistical methods. The term **econometrics** is used to describe this discipline, just as the term biostatistics is used to describe the statistical analysis of biological science data such as those from medical studies.

## 1.1 Statistical Methods

The starting point of statistical analysis is a dataset, a collection of measurements that most often come from a survey or from an experiment.

Statistical analysis begins with a **summary** or **description** of what can be a bewilderingly large set of numbers. There are several standard statistics that are used to summarize features of the data such as the central tendency of the data and the spread of the data. For example, given data on annual income of a number of individuals we may compute the average income for these individuals. This is relatively straightforward.

Most data analyses seek to go further and use such summary measures to extrapolate to the world beyond the particular dataset at hand. For example, if the average annual income in a sample of forty Californians is $35,000, what can we say about the average income of all Californians? Or if forty tosses of a coin lead to 18 heads and 22 tails, can we conclude that the coin is not a fair coin?

This extrapolation entails the much more challenging methods of **statistical inference**

- inferring details of a population from the sample at hand. The two main statistical tools used are **confidence intervals** and **hypothesis tests**; much of the book is focused on learning how to use these tools in a variety of settings.

Additionally, steps should be taken to ensure that a sample is representative and obtained in such a way that the phenomena of interest can be measured sufficiently precisely. Like other books at this level these issues are only briefly considered; they are covered in detail in separate statistics courses on survey sampling and experimental design.

In some special cases the dataset may be large enough and precise enough that there is no need to control for randomness due to sampling. For example, this would be the case if we had a complete census of the population or if we could toss the coin a million times. While very large datasets are increasingly available, such as those from internet transactions, in typical economics applications one needs to control for uncertainty.

## 1.2 Types of Data

The discipline of statistics covers a wide range of data types and associated methods that are summarized in this section.

Within this wide range economists, and hence this book, focus on observational data on continuously measured variables analyzed using regression methods. An example is sale price data from a sample of individual house sales.

### 1.2.1 Economics Data

There are several types of data that may demand different statistical methods:

- Numerical data that are continuous.
- Numerical data that are discrete.
- Categorical data.

Economics data are usually **numerical data** that are naturally recorded and interpreted as numbers. Furthermore, they often potentially take so many different values that they are viewed to be **continuous numerical data**. Examples are individual annual income or national GDP. Less often the data are **discrete numerical data** that take only integer values. Examples are the number of jobs held at a point in time or the number of patents awarded to a firm in a year. **Categorical data** are an alternative to numerical data where the data are recorded as belonging to one or more categories, such as whether or not a person is employed. Such data may be coded as numbers, e.g. 1 if employed and 0 if not employed, but are not intrinsically numerical.

This book emphasizes the study of **economics data** that are continuous numerical data. Many examples will be provided, particularly leading relationships that are discussed in introductory microeconomics and macroeconomics courses.

Before analysis begins, data are often transformed to a more suitable form. For example, suppose interest lies in modelling improvements in living standards over time. A standard measure to use is the annual growth rate in real per capita gross domestic product (GDP). This entails transformation of the original GDP data to adjust for inflation and to calculate year-to-year changes. Such transformations are used frequently enough to be treated in a separate chapter in each of Parts I, II and III.

### 1.2.2 Observational Data

Economics data are most often **observational data**, meaning they are based on observations of actual behavior in an uncontrolled environment. By contrast many physical and biological sciences in particular use **experimental data** that are observations on the results of experiments which can be controlled by the investigator.

A particular challenge of using observational data is that while it is easy to detect a relationship between two data series, it can be very difficult to determine cause and effect. Determining **causality** is studied in detail in Chapter 17.

### 1.2.3 Three Types of Observations

Within the class of observational data distinction is made between three types of data:

- Cross-section data on different individuals at a point in time.
- Time-series data on the same quantity at different point of time.
- Panel (or longitudinal) data on different individuals at different points of time.

**Cross-section data** are data on different entities, such as individuals, households, firms or countries, collected at a common point in time. Examples are earnings of individuals and output of firms. Such data are most often used in **microeconomics**. Standard notation is to use the subscript $i$ to denote the typical observation. The sample of size $n$ is denoted $x_1, ..., x_n$ with $i^{th}$ observation $x_i$.

**Time-series data** are data on the same quantity collected at different points in time. Examples are gross domestic product and the interest rate on a 13-week Treasury bill. Such data are most often used in **macroeconomics** and **finance**. Standard notation is to use the subscript $t$ to denote the typical observation. The sample of size $T$ is denoted $x_1, ..., x_T$ with $t^{th}$ observation $x_t$. In this course we use subscript $i$ as much as possible, but revert

occasionally to subscript $t$ in some time series settings. In particular, the one-period change in a time series variable is

$$\Delta x_t = x_t - x_{t-1}.$$

**Panel data** or **longitudinal data** are data on different individuals, such as firms or people or countries, each observed at several points in time. Examples include analysis of individual income over several years, and analysis of GDP in several countries over time. Such data are used in both microeconomics and macroeconomics. The typical observation is $x_{it}$, data for the $i^{th}$ individual at time $t$.

The same basic statistical methods may be applied to all three types of data. However, each type of data also adds its own special considerations for statistical inference and for model specification. These are detailed in Chapters 18-20.

## 1.3 Regression Analysis

Introductory statistics courses focus on data on a single variable considered in isolation, such as the individual income and coin toss examples. Some economic statistics such as the unemployment rate or the growth rate in real GDP or median earnings are also of interest on their own.

We first analyze **univariate data**, studying a single data series such as house price, with individual observations denoted $x_i$ or denoted $x_t$. The treatment of univariate data is similar to that in an introductory statistics course.

Most economic data analysis, however, is focused on measuring the relationship between two or more variables. The statistical method used to measure such inter-relationships is called **regression analysis**. Most of this book studies regression analysis.

For example, in Chapter 7 we consider the relationship between house price (in dollars) and house size (in square feet) for a sample of 29 house sales. Figure 1.1 presents a scatter plot of the data which suggests that, as expected, larger houses are associated with a higher price. Superimposed on this scatter plot is a line, called a regression line, that is the best fitting line for these data using a criterion given in Chapter 7. The regression line has slope coefficient equal to 74, approximately, so an increase in house size of one square foot is associated with an increase in house price of \$74.

We next study **bivariate data** on two related data series, with individual observations denoted $y_i$ and $x_i$. Linear regression for bivariate data (**bivariate regression)** fits a line of form

$$y = b_1 + b_2 x$$

to a sample on $y$ and $x$.

Finally **multivariate data** consider three or more related series. Usually one of those variables, say $y$, is explained by several other variables, say $x_1$, $x_2$, .... using **multiple**

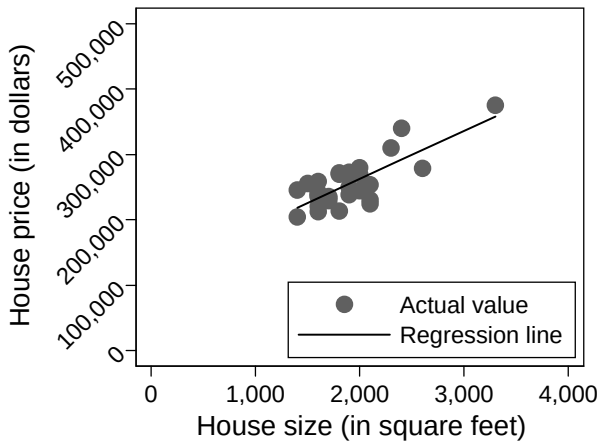


Figure 1.1: Linear regression example

**regression**. For example, we may consider the relationship between house price and several features of the house, such as size, number of bedrooms and lot size.

Regressions may be used to measure how an outcome variable ($y$) changes as one of the regressors ($x$) changes, or may be used to predict the outcome variable ($y$) for a given level of the regressors ($x$).

## 1.4 Key Concepts

1. There are two aspects to statistical analysis of data: description and inferential statistics. The latter attempts to extrapolate from the sample to the population, often using confidence intervals and/or hypothesis tests.

2. The analysis of economics data uses a subset of statistical methods, most notably regression analysis for continuous numerical data, and emphasizes economic interpretation of economics-related data.

3. Economics data are usually observational rather than experimental. This makes it difficult to establish causal effects. For pedagogical reasons this complication is deferred to Chapter 17, though much econometrics research seeks to estimate causal relationships, even with observational data.

4. Cross-section data (denoted $x_i$) are data on different individuals at a point in time; time-series data (denoted $x_t$) are data on the same quantity at different points of time; panel data or longitudinal data (denoted $x_{it}$) are data on different individuals at different points of time.

5. The book covers, in turn, univariate data (single series), bivariate data (two series), and multivariate data (several series).

6. Key Terms: Summary statistics; sample; population; statistical inference; continuous numerical data; discrete numerical data; categorical data; observational data; experimental data, cross-section data; time series data; panel data; longitudinal data; univariate data; bivariate data; multivariate data; regression analysis; bivariate regression; multiple regression.

## 1.5 Exercises

1. For each of the following examples state whether the data are numerical or categorical, and state whether the data are cross-section, time series or panel data.

   (a) Quarterly data on the level of U.S. new housing construction from 2000 to 2012.

   (b) Data on number of doctor visits in 2012 for a sample of 192 individuals.

   (c) Data on annual health expenditures for each U.S. state from 2000 to 2012.

   (d) Data on usual mode of transportation used to commute to work for a sample of 151 individuals.

2. For each of the following examples state whether the data are numerical or categorical, and state whether the data are cross-section, time series or panel data.

   (a) Data on annual health expenditures in 2012 for the U.S. by use of funds.

   (b) Data on whether the Dow Jones Index at the close of trading was at a higher or lower value than at the close of trading the preceding trading day.

   (c) Data on sales this quarter by each of 23 sales representatives.

   (d) Data on the price of 1 gigabyte of computer disk storage each year from 1980 to 2012.

3. For each of the following state whether the data are observational or experimental

   (a) Data on earnings for individuals some of whom chose to participate in a training program and some who were not.

   (b) Data on earnings for individuals some of whom were randomly assigned to a training program and some who were not.

   (c) Data on school outcomes for charter schools and for traditional schools.

# Chapter 2

# © A. Colin Cameron: Univariate Data Summary

Univariate data are a single series of data that are observations on one variable. A numerical data example is annual earnings for each person in a sample of women. A categorical data example is expenditures in each of a number of categories.

The chapter begins with presentation of summary statistics for numerical data. These are useful both in their own right and as a tool for checking that there are no obvious errors in data entry, such as negative values for a variable that should be nonnegative.

The chapter then presents charts that can provide a very quick way to grasp the essential features of univariate data. The graphical methods used vary with the type of data. While the key charts are given, there are many possible variations. The graphs presented here are quite basic. Presentation quality graphics entail much more preparation and are beyond the scope of this book. Useful resources for graph styles are leading publications such as *The Economist*, *The New York Times* and *The Wall Street Journal* that frequently present charts for economics data.

Statistical inference, using data from a sample to make inferences about the population from which the data is sampled, is deferred to later chapters.

## 2.1 Summary Statistics for Numerical Data

**Summary statistics** or **descriptive statistics** provide a summary of data on a numerical variable.

Consider data on the annual earnings of a sample of 171 women who are 30 years of age in 2010, all of whom worked full-time (35 or more hours per week and 48 or more weeks per year). The data are in dataset EARNINGS.

Table 2.1 presents various summary statistics, rounded to the nearest dollar, that are

Table 2.1: Summary statistics: Annual earnings of 30 year-old female full-time workers in 2010.

| **Statistic** | **Value** |
|---|---|
| Mean | 41,413 |
| Standard deviation | 25,527 |
| Minimum | 1,050 |
| Maximum | 172,000 |
| Number of Observations | 171 |
| Variance | 651,630,282 |
| Upper quartile (75th percentile) | 50,000 |
| Median (50th percentile) | 36,000 |
| Lower quartile (25th percentile) | 25,000 |
| Skewness | 1.71 |
| Kurtosis | 7.32 |

explained in this section. A summary statistics command in a statistical package usually automatically reports at least the first five of these.

As a quick check of the data we note that there are 171 observations that range from \$1,050 to \$172,000. The minimum value is surprisingly low as it implies earnings of less than \$1 per hour for this sample of full-time workers. From a more detailed check of the original survey data, this individual was self-employed, so such a low value is possible. The second lowest sample value of annual earnings is \$9,000.

The observations for a **sample** of size $n$ are denoted

$$x_1, x_2, ..., x_n.$$

Here $x_1$ is the first observation, $x_2$ is the second observation, .... and $x_n$ is the $n^{th}$ observation. For cross-section data the typical observation is the $i^{th}$ observation, denoted $x_i$, while for time series data it is more customary to use the subscript $t$, in which case $x_t$ is the $t^{th}$ observation.

### 2.1.1 Central Tendency

A measure of **central tendency** or **central location** describes the center of the distribution of the data.

The most common measure is the **sample mean**, which is the arithmetic average of the data. For example, if the data take values 8, 3, 7 and 6, then the sample mean is $(8 + 3 + 7 + 6)/4 = 6$. More generally, for a sample of size $n$, the sample mean $\bar{x}$ is defined as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

A shorthand notation to present this formula, and many other formulas for summary statistics, uses **summation notation**. In general $\sum_{i=1}^{n} x_i$ denotes the sum of all the $x_i$ from $i = 1$ to $n$, so that $\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$. In the current example, $x_1 = 8$, $x_2 = 3$, $x_3 = 7$ and $x_4 = 6$ so $\sum_{i=1}^{4} x_i = 8 + 3 + 7 + 6 = 24$. As a second example, if $x_i = 5 + 2i^2$ then $\sum_{i=1}^{n} x_i = \sum_{i=1}^{3}(5 + 2i^2) = (5 + 2 \times 1^2) + (5 + 2 \times 2^2) + (5 + 2 \times 3^2) = 7 + 13 + 23 = 43$.

Using summation notation, the sample mean can be written as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The other leading estimate of central tendency is the **median**. The data are first ordered from the lowest value to the highest value, and the median is that value that divides the ordered data into two halves. This is directly obtained as the midpoint of the ordered data if there is an odd number of observations. For an even number of observations one chooses the average of the two observations in the middle. For sample 8, 3, 7 and 6, the ordered sample is 3, 6, 7 and 8, and the median equals $(6 + 7)/2 = 6.5$.

The median has the advantage of being more resistant to outliers than the mean. For example, mean income will change a lot if Bill Gates is included in the sample, whereas the median is essentially unchanged. The mean is more often used, however, and this book focuses on statistical inference for the mean rather than the median.

A third measure, less commonly-used, is the **mid-range** which is the average of the smallest and largest values in the sample.

A fourth measure, the **mode**, is the most commonly occurring value. This is only useful when the data is discrete, or if the underlying data are intrinsically continuous but the observed data are greatly rounded, so that a given value can occur multiple times in the sample. Even then the mode is not necessarily a good measure of central tendency, especially if the distribution has more than one mode or if the distribution is asymmetric, defined below.

From Table 2.1, earnings are on average \$41,413. For these data with 171 observations the median is the $86^{th}$ of the ordered observations, and this equals \$36,000. So half the women in the sample earn less than \$36,000 and half earn more than this amount. Note that mean earnings in this example are substantially greater than median earnings. This is often the case for data on incomes, earnings and prices. The midrange is $(172000 + 1050)/2$ or \$86,525; this is much higher than the mean and is not particularly meaningful here. The mode, not given in Table 2.1, is \$25,000. In practice it is unlikely that any two women in this sample of size 171 would have exactly the same earnings. Here, due to rounding in reporting, ten women reported earnings of exactly \$25,000.

### 2.1.2 Quartiles, Deciles and Percentiles

The median is the point that equally divides an ordered sample. One can consider other divisions of the ordered sample.

The **lower quartile** is that point where one-quarter of the ordered sample lies below and three-quarters of the ordered sample lies above. The **upper quartile** is that point where three-quarters of the ordered sample lies below and one-quarter of the ordered sample lies above. Adjustment, similar to that for the median with an even number of observations, is needed when more than one data point could be the quartile. The median is the middle quartile.

Even more detailed divisions of the sample are possible. **Percentiles** split the ordered sample into hundredths. The $p^{th}$ percentile is the value for which $p$ percent of the observed values are equal to or less than the value. The upper quartile, median, and lower quartile are, respectively, the $75^{th}$, $50^{th}$, and $25^{th}$ percentiles. **Deciles** split the ordered sample into tenths and are often used, for example, to summarize the distribution of individual income. A **quantile** is a percentile reported as a fraction of one rather than as a percentage. For example the .81 quantile is the $81^{st}$ percentile.

From Table 2.1 the lower and upper quartiles of earnings are, respectively, \$25,000 and \$50,000, so the middle half of 30 year-old female full-time workers earned between \$25,000 and \$50,000 per year.

### 2.1.3 Data Dispersion or Spread

A measure of **dispersion** describes the **spread** or **variability** of the data. The most commonly-used measure is the standard deviation.

An obvious measure to use is the average of the deviations $(x_i - \overline{x})$ of the data $x_i$ from the sample mean $\overline{x}$. But this can be shown to always equal zero, because in sum the negative deviations exactly balance the positive deviations. Instead these deviations are squared, before averaging, to get the **sample variance** $s^2$ where

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

The division by $(n-1)$ rather than the more obvious $n$ is explained in Chapter 5.

The sample variance is measured in units that differ from those in the original data, due to the squaring. For example, if the data were in units of dollars then the variance is in units of dollars squared. To return to the original units we take the square root. This yields the **sample standard deviation** $s$, defined as

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

If one sample has a larger sample standard deviation than another then we view the sample as having greater variability.

As an example consider the sample 8, 3, 7 and 6 which has $n = 4$ and $\bar{x} = 6$. Then the sample variance

$$s^2 = \frac{(8-6)^2 + (3-6)^2 + (7-6)^2 + (6-6)^2}{4-1} = \frac{14}{3} \simeq 4.667,$$

and the sample standard deviation is $s = \sqrt{14/3} \simeq 2.16$.

In some cases it is useful to measure the variability of the data relative to the mean, using the **coefficient of variation**

$$\text{CV} = \frac{s}{\bar{x}}.$$

For example, in comparing variability of individual income in 2010 with that in 1980 the coefficient of variation will control for any change in mean income from 1980 to 2010.

Three other measures of variation in the data are the range, the interquartile range, and the average absolute deviation.

The **range** is the difference between the **maximum** and **minimum** values in the data set.

An **outlying observation**, or **outlier**, is an observation that is unusually large or small. The **interquartile range**, the difference between the upper quartile and the lower quartile, has the advantage of being more resistant to outliers than the standard deviation or the range.

The **average absolute deviation**, $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$ is also more resistant to outliers than the standard deviation or the range.

From Table 2.1, the sample standard deviation of earnings is \$25,527. The coefficient of variation is 25,527/41,413 = 0.62, so the standard deviation of earnings is 62% of mean earnings. The range is (172000−1050) or \$179,950. The interquartile range is (50000−25000) or \$25,000.

### 2.1.4 Interpretation of the Standard Deviation

The standard deviation is the commonly-used measure of variability, as is clear from subsequent chapters. It is not as easy to understand as the mean, which is simply the average.

A useful way to interpret the standard deviation is to use results for the normal distribution. For a random variable $X$ that is normally distributed with mean $\mu$ and standard deviation $\sigma$, the probability of being within one, two or three standard deviations of the mean is, respectively, 0.684, 0.955 and 0.997. This is illustrated in Figure 2.1.

It follows that approximately two-thirds of the sample is within one standard deviation of the mean, 95% is within two standard deviations and 99.7% is within three standard deviations of the mean. These results can also provide an approximate guide for data that are not normally distributed.

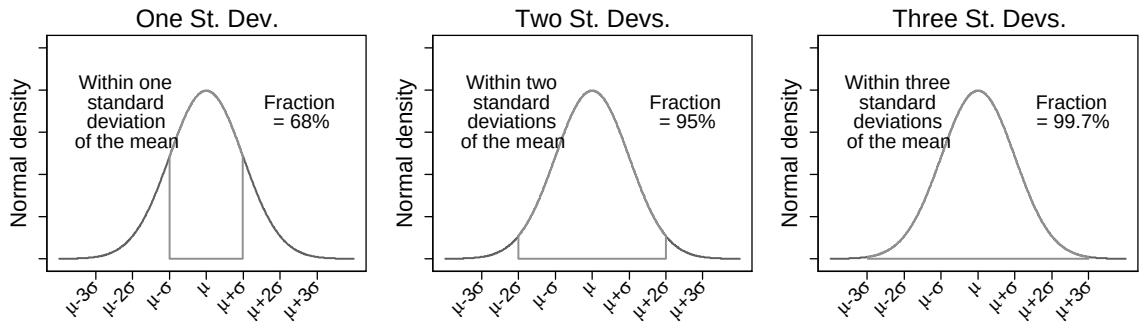


Figure 2.1: Normal distribution: Probability of being within one, two or three standard deviations of the mean.

Regardless of the actual distribution, a result called **Chebychev's inequality** implies that it always the case that at least three-quarters of a random sample is within two standard deviations of the mean, and at least eight-ninths is within three standard deviations of the mean.

As an example, consider the earnings data. These data have $\bar{x} = 41,413$ and $s = 25,527$, so the interval (14,886, 66,940) is within one standard deviation of the mean since, for example, $\bar{x} - s = 41,413 - 25,527 = 14,886$. For these data 77% of the observations are within this interval, compared to 68% predicted by the normal approximation. Similarly, for these data 96% of the observations are within two standard deviations of the mean, compared to 95% predicted by the normal approximation.

### 2.1.5 Symmetry

A **symmetric distribution** is one whose shape is the same when reflected around the median. The normal distribution is an example.

**Positive skewed** or **right-skewed** data have a much longer tail on the right. Most of the data are bunched on the left, but there is a continued presence of high values on the right. **Negative skewed** or **left-skewed** data have a much longer left tail.

Skewness can sometimes be visually detected using a histogram or kernel density estimate, introduced in the next section. Figure 2.2 presents histograms for symmetric, right-skewed and left-skewed data.

A formal measure of asymmetry is the **skewness statistic**, calculated as a scale-free measure by normalizing by the standard deviation. Different statistical packages can use slightly different formulae in computing the skewness statistic. The simplest measure, used

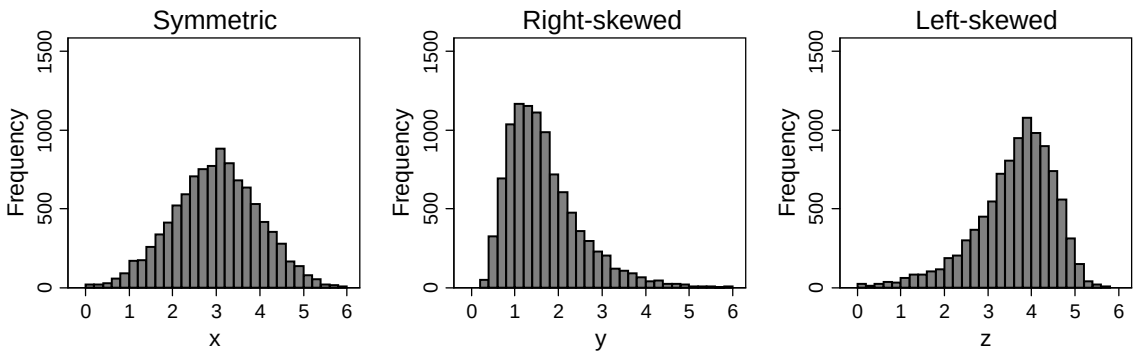


Figure 2.2: Histograms for symmetric, right-skewed and left-skewed data.

by most econometrics packages, is

$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}.$$

Some statistical packages, including Excel, multiply this measure by $\frac{\sqrt{n(n-1)}}{n-2}$. This adjustment is felt to lead to a better measure in small samples. In large samples the difference between the two measures disappears and both approximately equal $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$, where $s$ is the sample standard deviation.

A zero value indicates symmetry since there is then no skewness. A positive value indicates positive or right-skewness and a negative value indicates negative skewness. There is no clear-cut rule for when data are highly skewed; a skewness measure in excess of one in absolute value indicates at least mild skewness. Note also that in small samples the skewness statistic is a less precise estimate of data skewness. For the three examples in Figure 2.2 the skewness measure equals, respectively, $-0.04$, $1.92$, and $-2.31$.

Appreciable difference between the sample mean and sample median is also a sign of skewness. For right-skewed data the sample mean usually exceeds the sample median. For left-skewed data the sample mean usually is less than the sample median.

If economics data are skewed then they are usually right-skewed. For the earnings data, the histogram given below in Figure 2.4 clearly displays right skewness with a long right tail. For example, 94% of observations lie below the midpoint of $\$93,000$ and only 6% lie above the midpoint. And, from Table 2.1, the mean of \$41,413 exceeds the mean of \$36,000 and the skewness measure is 1.71.

Much economic analysis centers on modelling central tendencies. If skewness leads to an appreciable difference between the mean and the median, then both may be reported

or, depending on the purpose, only one of the mean or median may be reported. For example, household income is right-skewed and government statistical reports emphasize median household income rather than mean household income. This reports the income of the household in the middle of the household income distribution. At the same time, other government reports emphasize real per capita GDP which is a mean. This is in part because in that case the median cannot be computed, as data are not collected on GDP at the individual level. But it is also because in measuring the resources available to the economy interest lies in how much is available per person rather than how much is available to the median person.

### 2.1.6 Kurtosis

Different statistical packages can use slightly different formulae in computing the **kurtosis statistic**. The simplest measure, used by most econometrics packages, is

$$\text{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}.$$

The normal distribution, with Kurt $= 3$, is often used as a benchmark, especially if the distribution is reasonably symmetric. **Excess kurtosis** measures kurtosis relative to the normal distribution, yielding

$$\text{ExcessKurt} \simeq \text{Kurt} - 3.$$

Some statistical packages use an alternative measure of excess kurtosis that is felt to be better in small samples. One such measure, used by Excel, multiplies the kurtosis measure given above by $\frac{(n+1)(n-1)}{(n-2)(n-3)}$ and then computes excess kurtosis by subtracting $3\frac{(n-1)^2}{(n-2)(n-3)}$ rather than 3. In large samples the difference between different measures disappears and they approximately equal $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4$, where $s$ is the sample standard deviation.

There is some debate about how to interpret excess kurtosis. Originally a distribution with positive excess kurtosis was viewed as having a higher peak than the normal, but this need not be the case especially if the distribution is asymmetric or bimodal.

A more useful and intuitive interpretation is that it measures the relative importance of observations in the tail of the distribution, since $x_i - \bar{x}$ is raised to the fourth power. **Positive excess kurtosis** means that there is likely to be greater area in the tails than for the normal distribution with the same mean and variance. The kurtosis measure is most often used for financial data. Fat tails are a feature of data on investment returns, and the greatest interest may lie in the tails since unusual events can provide the greatest opportunity to make a profit (or a loss). Economists uses this interpretation of kurtosis – large positive excess kurtosis means that the tails of the distribution are much fatter than those for the standard normal distribution.

From Table 2.1 the earnings data has kurtosis statistic (Kurt) of 7.32, substantially greater than 3, suggesting that the sample distribution has fatter tails than the normal distribution. For the three examples in Figure 2.2 the kurtosis measure equals, respectively, 3.04, 11.68, and 16.57.

### 2.1.7 Box-and-Whisker Plot

A **box-and-whisker plot** or, more simply, a **box plot**, provides some of the key summary statistics for the data in a simple graphic.

All box-and-whisker plots give the lower quartile, median and upper quartile; these form the "box." Simple box-and-whisker plots additionally give the minimum and maximum; these form the "whiskers." More complicated box-and-whisker plots additionally plot outlying values. In that case the whiskers are data-determined lower and upper bounds, ones appropriate for data that are not too greatly dispersed, and outlying values are observations that exceed these bounds.

Figure 2.3 gives a box-and-whisker plot, of the more complicated form, for the earnings data. The solid shaded region ranges from the lower quartile of \$25,000 to the upper quartile of \$50,000. The solid white line within the shaded region is the median of \$36,000. The upper bar equals the upper quartile plus 1.5 times the inter-quartile range. Here this equals $50,000 + 1.5 \times 25,000$ or \$87,500. The six dots represent the six distinct values of earnings above \$87,500 in the sample. (In fact due to one duplicate there are seven observations in excess of \$87,500). The lower bar is the minimum sample value of \$1,050, as in this example the minimum exceeds the lower quartile minus 1.5 times the inter-quartile range.

The plot clearly shows the right-skewness of the data. The difference between the upper quartile and the median is much greater than the difference between the median and the lower quartile. And there are quite a few outlying sample points that take large values.

## 2.2 Charts for Numerical Data

Histograms are the leading method for graphical inspection of cross-section numerical data. Histograms can also be useful for time series numerical data, provided that the data have been transformed to have little overall trend. As an example, histograms may be useful for summarizing real GDP growth rates or price inflation rates over time, but are of very limited use for describing GDP and price levels which trend upward over time.

### 2.2.1 Histograms

Table 2.2 summarizes the earnings data grouped into intervals of width \$15,000. Each interval is called a **bin**; here there are 13 bins, each of equal **bin width** of \$15,000. The **frequency**

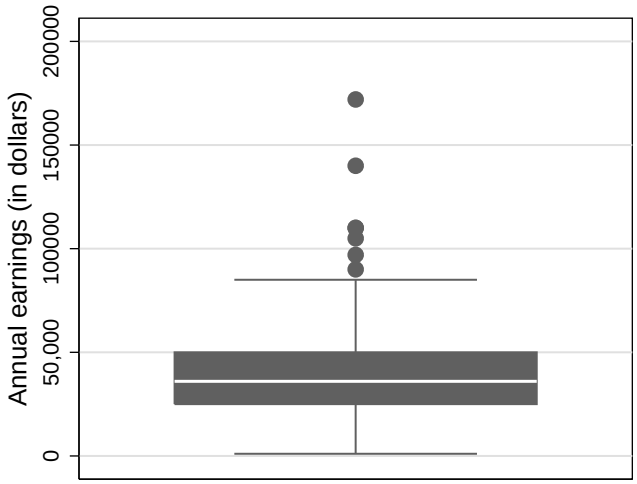


Figure 2.3: Box Plot: Annual earnings of 30 year-old female full-time workers in 2010.

is the number of observations that fall into a given bin, and the **relative frequency** is the proportion (or percentage) that fall into a given bin. For example, 53 observations or 31.0% of the sample have earnings between \$15,000 and \$29,999.

A **histogram** is a graph of the frequency distribution, with values on the horizontal axis and frequency or relative frequency or density (the relative frequency divided by the bin width) on the vertical axis. A histogram for the data summarized in Table 2.2 is presented in the first panel of Figure 2.4, while the second panel provides a more detailed histogram that groups the data over a narrower range.

The histogram varies with the number of bins, with a trade-off between few bins providing not enough detail and too many bins yielding a histogram that is very choppy. Given $n$ observations, a common default choice for the number of bins is $\sqrt{n}$. The class intervals are then of width approximately equal to the highest value minus the lowest value divided by the number of bins, with possible modification for unusually small or large observations. For $n = 171$ this yields 13 bins of equal width $(172000 - 1050)/13 = 13,150$. Table 2.2 and the first panel of Figure 2.4 instead round these defaults to 12 bins of equal width \$15,000 with a start value of \$0. The second panel of Figure 2.4 doubles the number of bins, by halving the bin width to \$7,500.

A variation on a histogram, one that gives more detail on the actual values taken by the data, is a **stem and leaf display**. This splits each data point into leading digits, called a stem, and remaining digits, called a leaf. For example for the earnings data the ten thousands may be the stem and the remaining digits the leaf. The data are then presented in tabular form where each row corresponds to a stem value and has first column the stem value and

Table 2.2: Frequencies in bins of width 15,000: Annual earnings of 30 year-old female full-tine workers in 2010.

| Range (or bin) | Frequency | Relative frequency (%) |
|---|---|---|
| 0-14,999 | 12 | 7.0 |
| 15,000-29,999 | 53 | 31.0 |
| 30,000-44,999 | 52 | 30.4 |
| 45,000-59,999 | 20 | 11.7 |
| 60,000-74,999 | 11 | 6.4 |
| 75,000-89,999 | 16 | 9.4 |
| 90,000-104,999 | 2 | 1.2 |
| 105,000-119,999 | 3 | 1.8 |
| 120,000-134,999 | 0 | 0.0 |
| 135,000-149,999 | 1 | 0.6 |
| 150,000-164,999 | 0 | 0.0 |
| 165,000-180,000 | 1 | 0.6 |

the second column the leaf values for that column.

Histograms can be used for numerical data that is either discrete or continuous. For **discrete data** that take low values, such as the number of different jobs held by a person during the year, each distinct value forms a bin so the bin width is one.

### 2.2.2 Smoothed Histograms

Data that take many different values, such as earnings data, have an underlying continuous probability density function rather than a discrete probability mass function. A classic example is the normal distribution which has a bell-shaped density. Probabilities are determined by areas under the curve and the total area under a density is one; see Appendix 5.A. It is then natural to directly estimate the density, using a smoothed histogram.

A **smoothed histogram** smooths the histogram in two ways. First, it uses rolling bins (or **windows**) that are overlapping rather than distinct. Second, in counting the fraction of the sample within each bin it gives more weight to observations that are closest to the center of the window and less to those near the ends of the window.

The smoothed histogram varies greatly with choice of **window width**, just as the histogram varies with the bin width. It varies less with the weights that are used. Different statistical packages may have different rules for choosing the default window width, and use different weights, leading to different smoothed histograms.

The most commonly-used smoothed histogram is a **kernel density estimate**. Two kernel density estimates for the earnings data are presented in Figure 2.5. The first panel uses a window width close to the statistical package's default width, while the second is

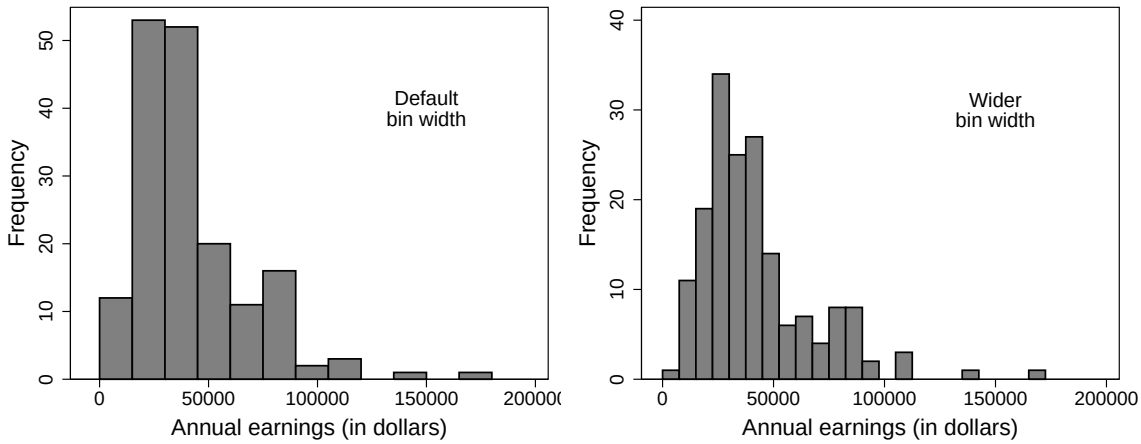


Figure 2.4: Histogram with two different bin widths: Annual earnings of 30 year-old female full-time workers in 2010.

smoother as it uses a window width that is twice as large. The kernel density estimate is not bell-shaped, implying that the data are not normally distributed. The density appears to be right-skewed, and suggests that a lognormal distribution may be appropriate.

### 2.2.3 Line Charts

A **line chart** plots the successive values $x_1, x_2, ...$ of the data against the successive index values 1,2,...

The leading application is to time series data that are ordered by time. This leads to graphs that plot the variable of interest against time.

Figure 2.6 presents a line chart of data from 1929 to 2011 for U.S. real gross domestic product (GDP) per capita in constant 2005 dollars. The data are in dataset REALGDPPC. The line chart clearly indicates great improvement in living standards, with per capita real GDP quintupling over the eighty-two years.

More generally, line charts can be useful whenever there is a natural ordering of the observations. For example, given data on test scores for 31 students it may be helpful to arrange the scores in descending order and produce a line chart of test score against student rank.

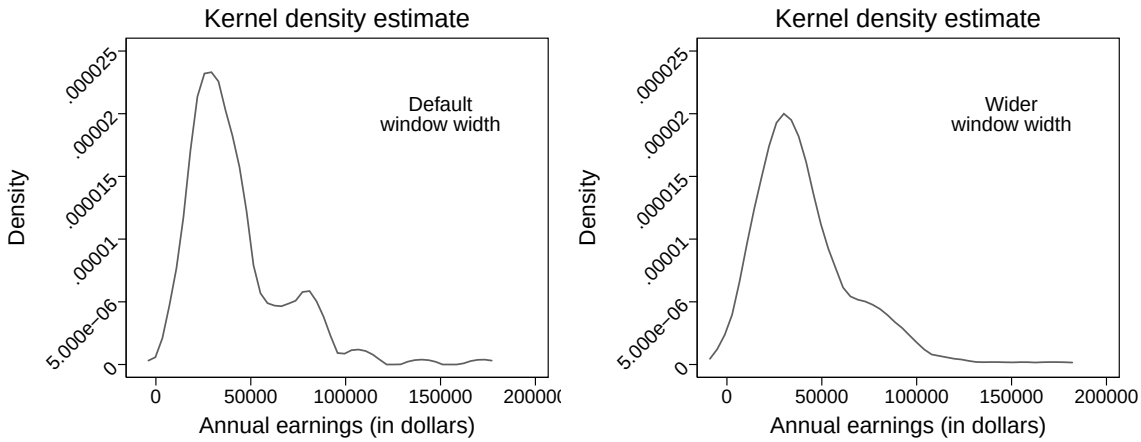


Figure 2.5: Smoothed histogram (kernel density estimate) with two different window widths: Annual earnings of 30 year-old female full-time workers in 2010.

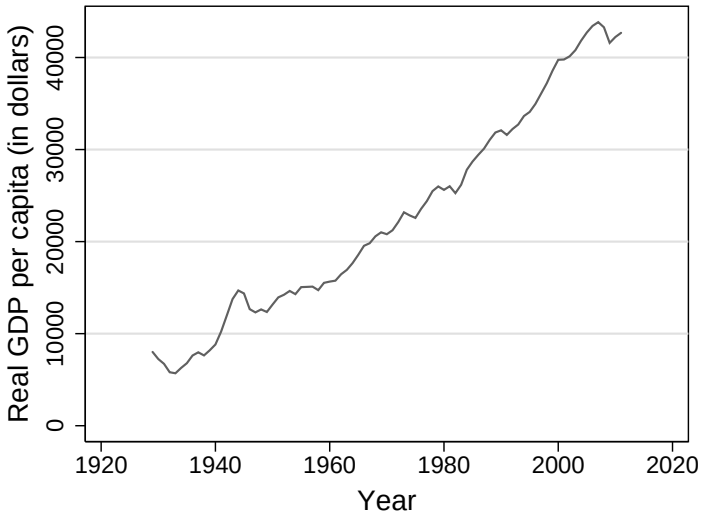


Figure 2.6: Line Chart: Real GDP per capita in U.S. (in 2005 dollars).

Table 2.3: Tabulation of frequencies for fishing sites.

| Category | Frequency | Relative frequency (%) |
|---|---|---|
| Beach | 134 | 11.34 |
| Pier | 178 | 15.06 |
| Private Boat | 418 | 35.36 |
| Charter Boat | 452 | 38.24 |

## 2.3 Categorical Data

As an example of categorical data consider choice of fishing site for a sample of 1,182 fishers given in dataset FISHING. There are four possibilities – fishing may be from a beach, pier, charter boat or private boat.

### 2.3.1 Data Summary

The fishing site data may be recorded as text, such as "beach", "pier", "charter" and "private". Or they may be recorded as numbers, such as 1, 2, 3 and 4. But even in the latter case the possibilities are intrinsically categorical. Furthermore there is no natural ordering of the categories.

For such data it is meaningless to compute summary statistics such as the sample mean. Instead the data are summarized using a tabulation of the frequencies for each category. For the fishing site data this is given in Table 2.3. It is clear that more people fished from a boat (private or charter) than from the shore (beach or pier).

### 2.3.2 Pie Charts

A **pie chart** splits a circle into slices, where the area of each slice corresponds to the relative frequency of observations in each category. Pie charts are most useful for visually representing each categories' share of the total, provided there are not too many categories.

Figure 2.7 presents a pie chart using the fishing site data. Again this makes clear that the largest categories are charter boat and private boat fishing.

## 2.4 Charts for Numerical Data by Category

As an example of categorical data consider U.S. health expenditures in 2013 of $2,920 billion (18% of GDP), broken into its main subcomponents. The data in dataset HEALTHCAT are completely listed in Table 2.4.

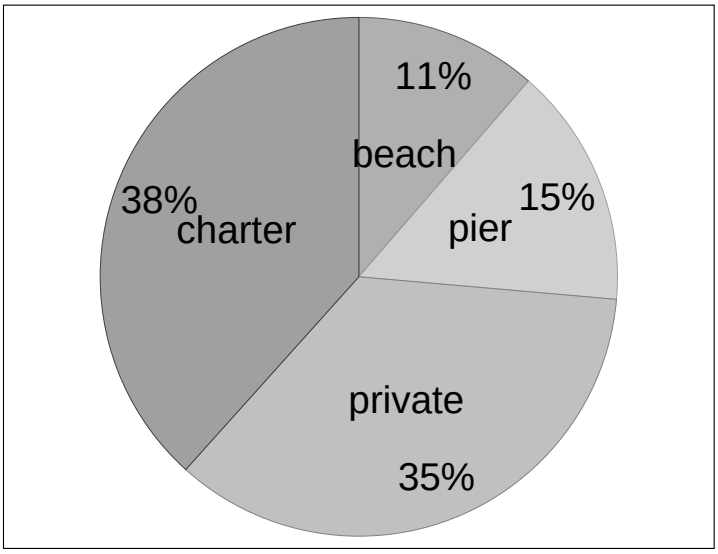


Figure 2.7: Pie Chart: Fishing site.

### 2.4.1 Bar Charts

A **bar chart** provides a bar for each category where the length of the bar is determined by the category value, here expenditures on the category of health.

A **column chart** or **vertical bar chart** puts the values on the vertical axis and the category on the horizontal axis. A **horizontal bar chart** instead puts the category on the vertical axis and the value on the horizontal axis. The choice of which to use is determined in part by whether one wants a short and wide chart, in which case a column chart is most often used, or a tall and narrow chart, in which case a horizontal bar chart is most often used.

Figure 2.8 presents a column chart for U.S. health expenditure data in 2010. This chart, ordered by size of category, makes it clear that hospital and physician expenditures are by far the largest components of total health expenditures.

### 2.4.2 Pie Charts

The same data could be presented using a pie chart, if one was interested in the shares of each category. But this would be difficult to read as there are too many categories. Instead it would be best to aggregate the smallest categories. For example, one might use hospital, physician, drugs and supplies, and all other categories combined.

Bar and pie charts can also be used if the categories are themselves numerical data, after first forming different categories according to what range of values the numerical data falls

Table 2.4: U.S. Health Expenditures in 2013 in billions of dollars by category

| Category | Amount ($ billions) |
|---|---|
| Hospital Care | 937 |
| Physician and Clinical Services | 587 |
| Dental | 111 |
| Other Professional | 80 |
| Other Health and Personal | 148 |
| Home Health Care | 80 |
| Nursing Care | 156 |
| Drugs and Supplies (Retail Sales) | 370 |
| Government Administration | 37 |
| Net Cost of Health Insurance | 174 |
| Government Public Health | 75 |
| Noncommercial Research | 47 |
| Structures and Equipment | 118 |

into. For example one might group years of completed schooling into 0-11 (less than high school), 12 (high school graduate), 13-15 (some college), 16 (4-year college graduate) and 17 and above (postgraduate). Then give a bar chart of average income by schooling category. A histogram is just a column chart of frequencies plotted against the class boundaries.

## 2.5 Key Concepts

1. Commonly-used statistics for numerical data include the mean and median (for central tendency), the standard deviation, inter-quartile range and range (for dispersion), quantiles and percentiles, and symmetry and kurtosis statistics.

2. An outlying observation, or outlier, is an observation that is unusually large or small.

3. A box plot provides a visual summary of key sample statistics. Some box plots also plot outlying observations.

4. Commonly-used charts that can provide a useful visual presentation of the data are histograms, kernel density graphs, column charts, line charts, bar charts and pie charts. Which is best to use depends on whether the data is numerical (continuous or discrete) or categorical, and whether the data are cross-section or time series data.

5. Key Terms: sample; summary statistics; central tendency; central location; sample mean; median; mid-range; mode; quartile; decile; percentile; quantile; dispersion; sample variance; standard deviation; coefficient of variation; range; outlying observation;

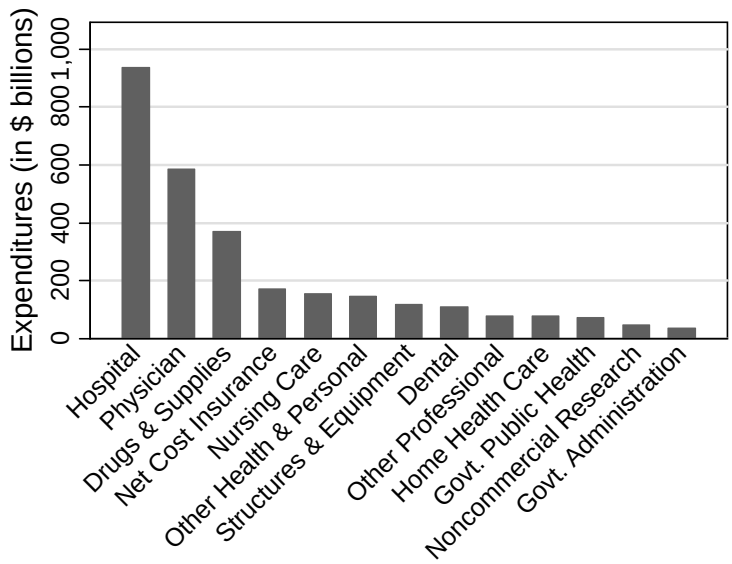


Figure 2.8: Column Chart: Health expenditures in U.S. in 2013 (in $billion).

outlier; inter-quartile range; symmetry; skewness; right-skewed; positive skewed; kurtosis; normal distribution; box plot; histogram; frequency; relative frequency; stem and leaf display; kernel density estimate; line chart; horizontal bar chart; vertical bar chart; column chart; pie chart.

## 2.6 Exercises

1. Obtain $\sum_{i=1}^{n} z_i$ for the following cases:

   (a) $z_i = 1$ and $n = 5$.

   (b) $z_i = i$ and $n = 5$.

   (c) $z_i = 2i^2$ and $n = 5$.

   (d) $z_i = 1/i$ and $n = 5$.

2. For the panel variable $x_{it}$ that takes values $x_{11} = 5$, $x_{12} = 3$, $x_{13} = 7$, $x_{21} = 8$, $x_{22} = 6$, and $x_{23} = 4$:

   (a) Calculate $\sum_{t=1}^{3} x_{it}$ for $i = 1$ and for $i = 2$.

   (b) Calculate $\sum_{i=1}^{2} x_{it}$ for $t = 1$, for $t = 2$ and for $t = 3$.

3. Compute from first principles (i.e. using the formula and a calculator) the mean, standard deviation, coefficient of variation, symmetry statistic and kurtosis statistic for the sample 4, 2, 0, 2. Show all calculations.

4. Repeat the previous question when the observations are ten times larger, so the sample is now 40, 20, 0, and 20. Which of the measures are scale-free measures?

5. Repeat the first question when the observations are translated by 2, so the sample is now 6, 4, 2, and 4. Which of the measures are unchanged by translation?

6. A sample of size 200 has mean of 20 and standard deviation of 5. If the data are normally distributed, what range of values do you expect 95% of the sample to lie in?

7. IQ scores have a mean of 100, standard deviation of 14 and are approximately normally distributed. What range of IQ scores do you expect 99.7% of the population lie in?

8. For a sample of size 1,000 the central two-thirds of the observations lie between 60 and 100. If these data are normally distributed, provide an estimate of the mean and standard deviation.

9. A sample of 30 people had the following years of completed schooling: 12, 12, 14, 12, 12, 12, 12, 12, 16, 12, 14, 12, 12, 13, 14, 12, 17, 12, 12, 16, 12, 12, 8, 14, 16, 12, 12, 17, 12, 16.

   (a) Read the data into your statistical package.
   (b) Obtain summary statistics.
   (c) Give the inter-quartile range.
   (d) List the first five observations.
   (e) Obtain a table of frequencies for these data.
   (f) Give a histogram, with a bin width of one for these discrete data. Do the data appear to be normally distributed?
   (g) Provide a pie chart - what is the most common educational level?

10. Repeat the previous question for the following samples:

    (a) 20 people age 30 with the following number of annual doctor visits: 0, 0, 3, 4, 2, 5, 5, 2, 31, 2, 2, 2, 3, 0, 8, 0, 8, 1, 2, 4.
    (b) 25 families with the following number of family members: 3, 3, 4, 7, 4, 3, 5, 2, 2, 4, 7, 3, 4, 3, 3, 5, 3, 4, 4, 1, 6, 5, 4, 5, 5.

11. The unemployment rate for college graduates (bachelor's degree or higher) aged 25 to 34 years in April in each of the years 2000 to 2013 was 1.3, 2.0, 2.7, 2.9, 2.6, 2.1, 2.2, 1.9, 2.2, 4.3, 4.7, 3.9, 3.6, 3.6.

    (a) Read the data into your statistical package.
    (b) Obtain key summary statistics.
    (c) Give the inter-quartile range.
    (d) List the first five observations.
    (e) Give a line chart.
    (f) Give a bar chart.

12. Repeat the previous question for the following cases:

    (a) High school graduates aged 25 to 34 years with April unemployment rates of 4.4, 5.2, 8.7, 7.9, 7.1, 6.4, 6.1, 5.5, 6.8, 13.6, 13.9, 13.7, 10.1, 9.9.
    (b) People with 1-3 years of high school aged 25 to 34 years with April unemployment rates of 7.7, 9.0, 14.2, 11.7, 11.8, 12.6, 13.9, 10.4, 12.8, 23.5, 19.8, 24.5, 20.1, 15.4.

13. Table 2.4 gives U.S. health expenditures by category for 2010. For 2009 the corresponding amounts were, respectively, 776, 503, 103, 66, 139, 335, 188, 164, 76, 46, 100.

    (a) Give a column chart.
    (b) Give a pie chart.

14. Repeat the previous question for the following years

    (a) 1980 when the corresponding amounts were, respectively, 101, 48, 13, 2, 15, 26, 12, 12, 6, 5, 15.
    (b) 2000 when the corresponding amounts were, respectively, 416, 291, 62, 32, 85, 178, 102, 83, 43, 26, 62.

15. For each of the following situations state whether the median price or the mean price of cars sold is a more useful measure of central tendency.

    (a) You want to know the typical price of a car.
    (b) You also know the number of cars sold and want to calculate sales tax receipts when car sales are subject to a 5% tax.

16. In each of the following situations state whether or not the data are likely to be positively skewed, or whether there is not enough information to know.

    (a) The mean is 50 and the median is 20.

    (b) The skewness statistic is 0.1.

    (c) The excess kurtosis statistic is 5.

    (d) The $10^{th}$ percentile is 20, the median is 50 and the $90^{th}$ percentile is 200.

17. The dataset PRICEEARNINGS has annual data on the price-earnings ratio (variable *pe*) for S&P500 firms (in real terms) from 1881 to 2012).

    (a) Obtain the summary statistics for *pe*. Do the data appear to be skewed? Do the data appear to have greater kurtosis than the normal distribution? Explain.

    (b) Plot the histogram. Do the data appear to be skewed?

    (c) Provide a time series plot of the data. Comment on any unusual features.

    (d) Do the data to be unusually high or low in 2012? Explain.

18. The dataset AUSREGWEALTH has data on average net wealth of households in 517 regions in Australia in 2003-04.

    (a) Obtain the summary statistics. Do the data appear to be skewed? Do the data appear to have greater kurtosis than the normal distribution? Explain.

    (b) Plot the histogram. Do the data appear to be skewed?

    (c) If your software does this, plot the kernel density estimate. Do the data appear to be skewed?

    (d) Now take the natural logarithm of average net worth and repeat parts a-c.

# Chapter 3

# © A. Colin Cameron: Economics Data

This chapter describes some of the leading variables used in economic analysis and often featured in media articles about the economy. It includes convenient internet sources for these data. The material provides a resource for empirical analysis, and this chapter might be skipped on a first reading.

A single source for many of the standard U.S. macroeconomic and financial variables, including the series presented in the first four sections of this chapter, is the FRED® economic database provided free by the Federal Reserve Bank of Saint Louis at website research.stlouisfed.org/fred2. The data can be downloaded as Excel or comma-separated files from the FRED website. Alternatively, and more simply, given an internet connection the data can be directly loaded from within Eviews, from within Stata using user-written command `freduse`, or from within R using one of several packages including `Quantmod` and `fimport`. Some data series are subject to one or more revisions after their initial release to reflect more complete information that becomes available over time. It is therefore important to note what version of the data is being used. Furthermore, many macroeconomic time series are seasonally adjusted and seasonal adjustment is imperfect. Finally many data series, such as employment statistics, are based on surveys so there is sampling error.

While this chapter provides considerable detail for macroeconomics data, most economics applications use microeconomics data. There are too many sources of microeconomics data to detail. This chapter merely concludes by presenting internet websites that provide convenient access to some of the leading individual-level data sets such as the U.S. Census, the U.S. Current Population Survey and the American Community Survey.

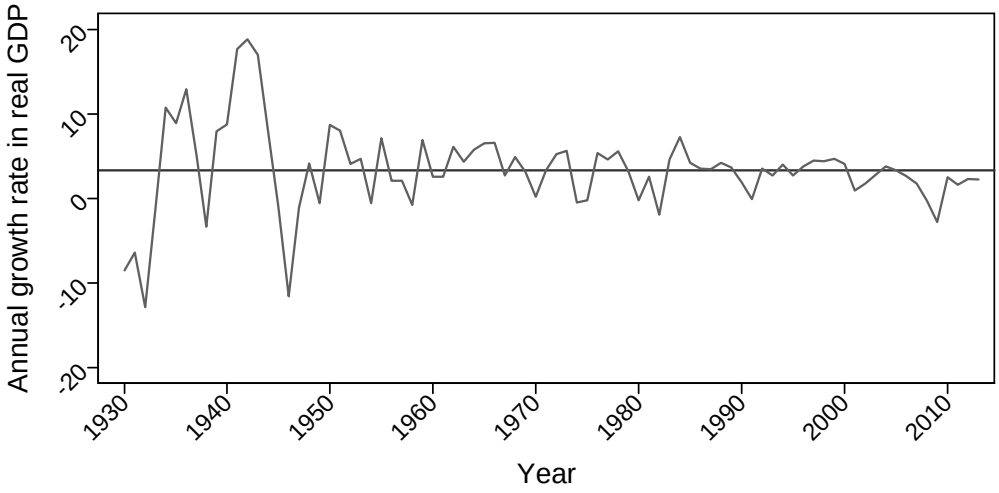


Figure 3.1: Annual growth rate in U.S. real GDP from 1930.

## 3.1 Output Measures

### 3.1.1 Gross Domestic Product

The main measure of the aggregate output of the economy is **gross domestic product** (**GDP**), the value of all goods and services produced within the country. The textbook definition of GDP, based on expenditures, is that it equals the sum of personal consumption, investment, government expenditure, and net exports (exports minus imports). Other approaches to equivalently compute GDP are based on net production (or value-added) and on income received. **Real GDP** is most often analyzed as it controls for price inflation; see Chapter 4.2 for further discussion.

Figure 4.1 plots the annual growth in U.S. real GDP in each year from 1930. The average growth rate over the period 1930–2013 was 3.3 percent, given by the solid line. The period since the mid-1970's has seen lower growth of an average of 2.7 percent per year. Real GDP was much more volatile in the earlier periods, and was relatively stable from the mid-1980's to 2007, a period known as the Great Moderation.

A **contraction** or **recession** is a substantial slowdown in the country's economic activity. Table 4.1 presents dates for the 14 contractions in the U.S. economy that have occurred since 1929, as determined by the National Bureau of Economic Research. These contractions roughly correspond to two or more quarters of negative growth in real GDP. The contractionary periods are relatively short, averaging 14 months in length and summing to just 16 of the past 83 years. Economic recovery is by no means complete at the end of the con-

Table 3.1: Contractions in U.S. Economy since 1930.

| Years | Start | Duration (months) | Notable Recessions |
|---|---|---|---|
| 1929-33 | 1929 August | 43 | Great Depression |
| 1937-38 | 1937 May | 13 | Great Depression continued |
| 1945 | 1945 February | 8 | |
| 1948-49 | 1948 November | 11 | |
| 1953-54 | 1953 July | 10 | |
| 1957-58 | 1957 August | 8 | |
| 1960-61 | 1960 April | 10 | |
| 1969-70 | 1969 December | 11 | |
| 1973-75 | 1973 November | 16 | First OPEC oil shock |
| 1980 | 1980 January | 6 | Second OPEC shock and monetary tightening |
| 1981-82 | 1981 July | 16 | Monetary tightening continued |
| 1990-1991 | 1990 July | 8 | |
| 2001 | 2001 March | 8 | Dot-com bust (Tech-wreck) |
| 2007-09 | 2007 December | 18 | Global Financial Crisis (Great Recession) |

traction. For example, in 2013 the economy had not completely recovered from the 2007-09 contraction.

Major recessions are the Great Depression of the 1930's, the mid-1970's recession that followed the first Organization of Petroleum Exporting Countries (OPEC) oil shock, the early 1980's recession that followed the second OPEC oil shock and subsequent tightening of monetary policy in response to high price inflation of over ten percent, the early 2000's recession that followed a major stock market correction (the "dot-com bust" or "tech-wreck") and the late 2000's Great Recession following the Global Financial Crisis. The seemingly large contraction in 1945 was short-lived and was due to great reduction in government wartime spending at the end of World War II and substantial price inflation with the relaxation of wartime price controls.

Media reports emphasize the estimated quarterly growth in seasonally adjusted real GDP presented as an annualized rate (by multiplication by four); Chapter 4.2 details these adjustments. The GDP data are part of the National Income and Product Accounts (NIPA) produced quarterly by the Bureau of Economic Analysis in the U.S. Department of Commerce.

### 3.1.2 Other Output Measures

A related measure, **gross national product** (**GNP**), was the main output measure used in the U.S. before the switch to GDP in 1991. GNP is the value of all goods and services produced by nationals of the country. This equals GDP plus income earned in other countries

minus income of non-nationals in the country. For example, suppose a French company makes wine in the U.S. Then U.S. GDP rises, but U.S. GNP does not rise as the French are non-nationals. Conversely, French GDP does not rise, as production was in the U.S., but French GNP does rise.

The GDP data are released only quarterly, so great attention is additionally paid to data released more frequently that provide more up-to-date information about the direction of the macroeconomy. These data include the industrial production index, new housing starts, new manufacturing orders, the trade balance and the University of Michigan index of consumer sentiment.

## 3.2 Price Indexes and Price Inflation

A remarkably large amount of popular discussion fails to control for changes in prices over time. For example, a news item about rising house prices might say that houses are six times as expensive today as in 1970, neglecting to observe that overall prices also increased six times over this period.

A **price index** measures how the overall price of goods, services or production inputs varies over time or across regions.

The simplest price index begins with a representative collection of goods and computes how the cost of this basket of goods changes over time. Then

$$\text{Price index} = 100 \times \frac{\text{Price of basket in current period}}{\text{Price of basket in base period}} .$$

Thus if the basket cost \$160 in the base period and \$240 in the current period, the index is 100 in the base period and 150 ($= 100 \times 240/160$) in the current period.

### 3.2.1 Consumer Price Index

The most commonly-used price index is the **Consumer Price Index (CPI)**, a monthly index produced by the U.S. Bureau of Labor Statistics (BLS). The BLS computes price indexes for 211 categories of goods and services in 38 geographical areas. Weighted averages of these indexes are then used to compute several different aggregate price indexes. The most commonly reported indexes are the CPI-U that is intended to reflect prices for urban consumers and the CPI-W that is intended to reflect prices for wage earners. These indexes are often used for **cost of living adjustments** (COLAS). For example, the CPI-W is used to adjust Social Security pensions. The sampling error for the CPI is very small.

Prices have changed greatly over time. Table 4.2 gives the CPI-U index from 1930 to 2013. Over this time prices have increased more than thirteen-fold. One 1930 dollar is

Table 3.2: U.S. Consumer Price Index (CPI-U) since 1930.

| | 1930 | 1950 | 1970 | 1990 | 2010 | 2013 |
|---|---|---|---|---|---|---|
| CPI-U (1982-84=100) | 17.1 | 23.5 | 37.8 | 127.4 | 216.7 | 230.3 |
| Amount one 1930$ equals in various years | 1.00 | 1.37 | 2.21 | 7.45 | 12.67 | 13.47 |
| Amount one 2013$ equals in various years | 0.07 | 0.10 | 0.16 | 0.55 | 0.94 | 1.00 |

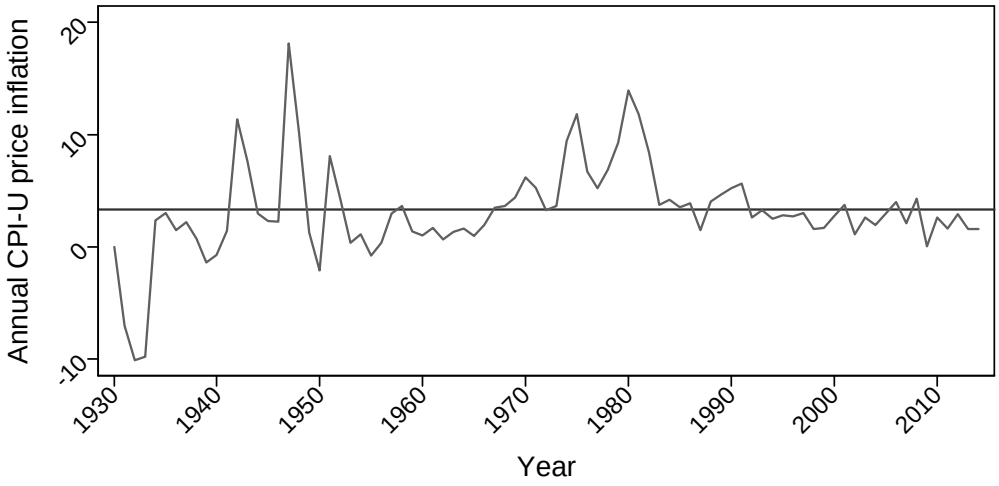


Figure 3.2: Annual U.S. price inflation (CPI-U) from 1930.

equivalent to $13.47 in 2013, since 230.3/17.1 = 13.47. Going the other way, a dollar in 2013 has the same purchasing power as $0.07 in 1930 dollars since 17.1/230.3 = 0.07.

Figure 4.2 plots the annual price inflation rate measured using the CPI-U in each year from 1930. The average inflation rate over the period 1930–2013 was 3.3 percent, given by the solid line. There was substantial price deflation during the 1930's Great Depression and substantial price inflation immediately following World War II and in the 1970's. The period since the mid-1980's has seen low and stable price inflation.

Major challenges in calculating a price index are that (1) the representative basket of goods and services changes over time due to substitution towards goods that become relatively cheaper and away from goods that become relatively more expensive, (2) the representative basket of goods and services changes over time with the introduction of new goods, and (3) the quality of goods in the basket changes over time.

An influential 1996 study calculated that the CPI indexes appreciably overstated inflation due to failure to adequately account for these three complications, by 1.1 percent per year

for the CPI-U. The BLS has since changed its methods so that this overestimation of price inflation is much reduced. Additionally the BLS introduced the C-CPI-U, the **chained** CPI-U, that better accounts for substitution across broad categories of goods and services and measures annual inflation to be 0.2-0.3 percent per year lower than the CPI-U.

### 3.2.2 Other Price Indexes

Real GDP is calculated from nominal GDP using the **GDP price deflator**, a quarterly index produced by the Bureau of Economic Analysis as part of NIPA. The deflator is a chained index that controls for substitution in response to changes in relative prices.

The CPI has the attraction of being released monthly and beginning in 1913. The other most closely watched monthly index is the **Personal Consumption Expenditures (excluding food and energy) Price Index**. This is a measure of **core inflation** as it excludes the two components that can fluctuate greatly for seasonal and other factors unrelated to overall price inflation. Additionally the monthly **Producer Price Index** (PPI), produced by the BLS, measures changes over time in the prices received by domestic producers for their output. The PPI is used to provide a measure of real output of producers, rather than nominal output. As manufacturing has become a smaller share of the economy the BLS has widened the range of goods and services covered by the index.

Another price index that is followed by the media is the monthly Standard and Poors **Case-Shiller 20-city home price index**. Prices for key commodities, such as crude oil prices (most often the price of west Texas intermediate at Cushing, Oklahoma) are also tracked.

## 3.3 Labor Force Statistics

Key labor force statistics in the U.S. come from two distinct surveys. One survey leads to the most reliable measure of employment, while the other survey additionally measures unemployment and the size of the labor force.

### 3.3.1 Employment

Employment statistics are based on the U.S. Bureau of Labor Statistics (BLS) **establishment survey**, a monthly survey of about 145,000 businesses and government agencies. The key statistic reported by the media is the monthly change in seasonally adjusted total non-farm payroll employment. Currently an increase of 200,000 jobs or more is viewed as good, while an increase of 100,000 jobs or less is insufficient to accommodate population growth.

The media take no account of the measurement error in this important statistic. The BLS computes the margin of error for a 90% confidence interval in this statistic to be 90,000

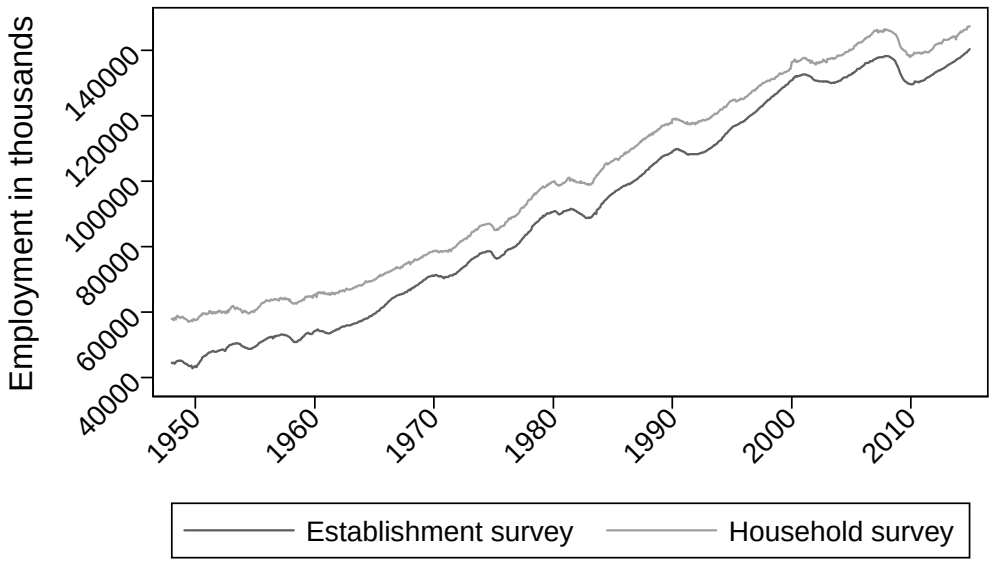


Figure 3.3: U.S. Employment since 1939.

jobs. So a monthly gain of 74,000 jobs, for example, has a wide 90% confidence interval of (-16,000, 164,000). And seasonal adjustment is challenging as employment varies greatly across the year. For example, from November 2013 to December 2013 seasonally adjusted employment was estimated to rise by 74,000 jobs while seasonally unadjusted employment actually fell by 244,000 jobs.

A second, and quite different, source for employment data is the monthly **Current Population Survey (CPS)**, a **household survey** of around 60,000 households administered by the U.S. Bureau of the Census for the BLS.

Figure 4.3 presents employment data from 1940. The establishment survey figures, plotted in the lower line, provide the more reliable measure of month to month changes as they come from a larger sample. The household survey measures higher employment as the survey covers all work, not just nonfarm payroll employment. Both series show declines in employment in the mid-1970's, the early 1990's, the early 2000's and the late 2000's. These declines correspond to the contractions listed in Table 4.1.

Great attention is paid to the monthly government employment statistics. Another much watched series is the monthly ADP employment figure, based on data collected by the payroll-processing firm Automatic Data Processing (ADP) and constructed in collaboration with Moody's Analytics. This series is released ahead of the BLS establishment survey and is intended to predict the establishment figures.

### 3.3.2 Unemployment

The CPS household survey provides considerably more detail on labor force statistics.

The civilian population (aged 16 years and over) is split into three categories as follows

$$\text{Population} = \text{Employed} + \text{Unemployed} + \text{Not in the Labor Force.}$$

Those with a job are **employed**. This includes any part-time or temporary work and temporary absence from a job for reasons such as illness. Those without a job but who looked for a job in the past four weeks and are available to work immediately are **unemployed**. A wide range of job seeking methods are counted, including contacting friends or relatives. The **labor force** is defined as the sum of the employed and the unemployed; this measures those who either have a job or do not have a job but would like a job. Those not employed or unemployed are determined to be **not in the labor force**.

The **unemployment rate** gives the percentage of those in the labor force who do not have a job, thus

$$\text{Unemployment Rate} = 100 \times \text{Unemployed} \;/\; (\text{Employed} + \text{Unemployed}).$$

Figure 4.4 presents three measures of the unemployment rate from 1948 for all those 16 years and older and in the civilian population.

The middle series in Figure 4.4 gives the base unemployment rate, measured as described above. It is highly countercyclical, with marked increases in each of the contractionary periods listed in Table 4.1.

The lower series in Figure 4.4 is restricted to just the long-term unemployed, those unemployed for at least half a year, and gives this number as a percentage of the labor force. Long-term unemployment is much, much higher in the most recent recession than in previous recessions, and some of the long-term unemployed may never become reemployed.

The highest series in Figure 4.4 provides a measure of underemployment as well as unemployment. This is the U-6 unemployment rate, a broader definition that adds to the unemployed all marginally attached workers (those who have looked for a job in the past 12 months though not in the past 4 weeks) and those employed part-time that are willing to work full-time. This rate reached 17 percent of the labor force in 2010.

### 3.3.3 Labor Force Participation Rate

The **labor force participation rate** gives the labor force as a percentage of the population. Thus

$$\text{Labor Force Participation Rate} = 100 \times (\text{Employed} + \text{Unemployed}) \;/\; \text{Population}.$$

Figure 4.5 presents this rate for all ages (16 years and older), the lower curve, and for those aged 25-54 years, the upper curve. The increase from 1950 to 2000 is due to the rising

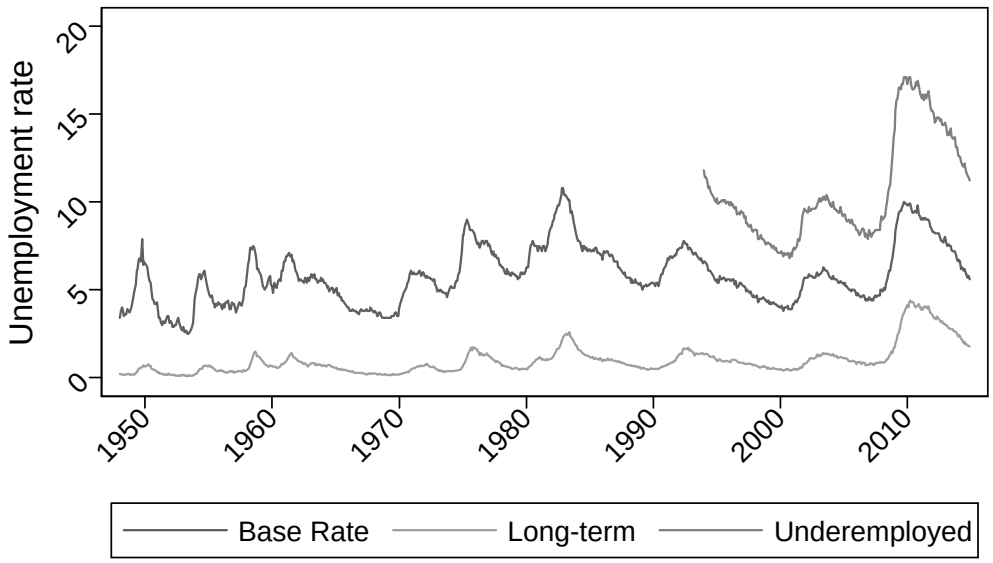


Figure 3.4: U.S. Unemployment rate since 1948.

labor-force participation of women. The rate is much less cyclical than the **employment-population ratio** or the unemployment rate.

The labor-force participation rate is higher for those aged 25-54 years, since most in this age group have completed their education and are not yet retired. The labor-force participation rate has dropped considerably since 2000, even for those aged 25-54 years. This decline is of considerable concern to policy-makers, with very slow recovery from the global financial crisis.

## 3.4 Financial Data

### 3.4.1 Interest Rates

Interest rates, usually expressed as the annual percentage return on a financial asset, vary with the safety of the asset and the length of maturity.

The safest assets are viewed to be U.S. government bonds. Representative rates are the ten-year **Treasury bond** at long maturity and the one-year **Treasury bill** at short maturity.

Figure 4.6 plots these rates from 1954 on. The longer rate is generally higher than the short rate, and varies less with the business cycle. Interest rates increased steadily from 1965

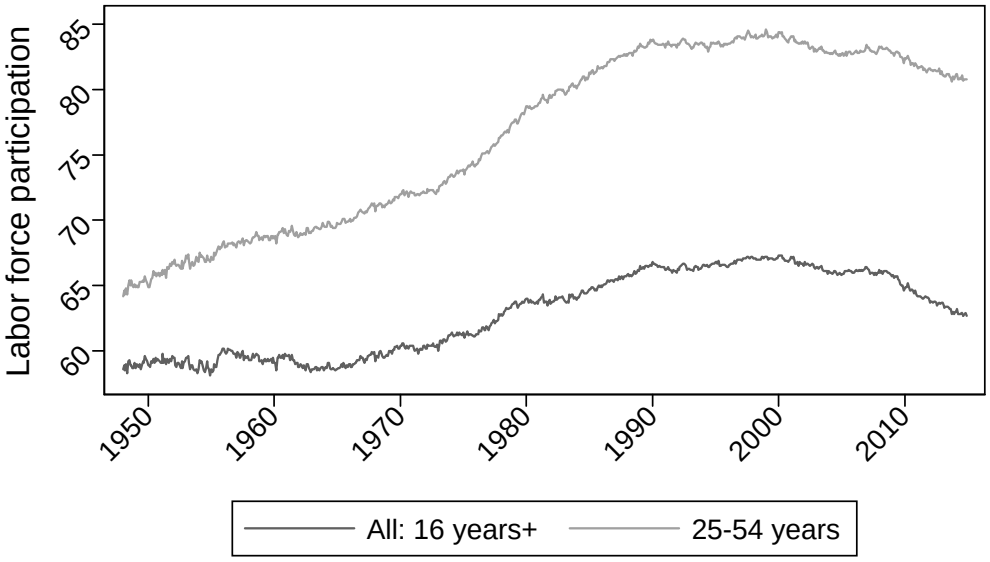


Figure 3.5: U.S. Labor Force Participation Rate since 1948.

to 1980 and then declined steadily, reflecting similar trends in price inflation rates given in Figure 4.2. The pronounced dips in 1-year interest rates reflect attempts by the Federal Reserve Bank to stimulate the economy following contraction. In the 2000's the difference between the short rate and long rate is much higher than in the past. The one-year rate is essentially at its floor of zero in the early 2010's. Due to this rigidity, policy-makers currently pay more attention to the two-year Treasury bond rate.

A key policy tool of the Federal Reserve is the **federal funds rate**, the interest rate at which banks trade reserve balances with each other, usually overnight. The Federal Reserve uses purchases and sales of government securities to keep the federal funds rate within a desired range.

A benchmark rate for short-term loans is the **bank prime loan rate**. For long-term corporate borrowing a standard rate is Moody's long-term **corporate bond rate** that includes bonds with remaining maturities as close as possible to 30 years. Moody's Aaa bond rate is for highly-rated corporations and provides an alternative measure of long-term interest rates to Treasury bond rates.

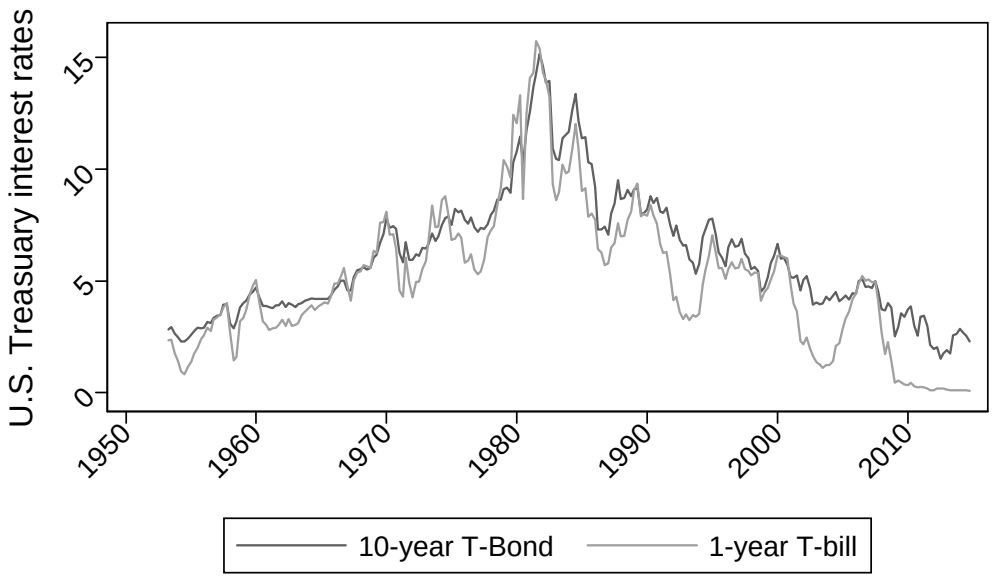


Figure 3.6: U.S. Treasury security interest rates since 1954.

### 3.4.2 Stock Price Indexes

A **stock price index** is a weighted average of the price of each stock, with

$$I_t = \sum\nolimits_{j=1}^{J} w_{jt} p_{jt}$$

where $p_{jt}$ is the price of the $j^{th}$ stock at time $t$ and $w_{jt}$ is the weight given to the $j^{th}$ stock.

A **market-value weighted index** or **capitalization weighted index** weights by the total market value of outstanding shares, i.e. the stock price times the number of shares outstanding. A **float weighted index** considers only those stock that are held by the general public (and hence traded on the market) rather than, say, the government or company insiders. A **price weighted index** chooses weights $w_{jt}$ that are proportional to the price of the stock, so a stock that sells for \$100 a share gets five times the weight as one that sells for \$20 a share.

For construction of any index, issues include how frequently to adjust the weights and change the companies in the index over time. Typically companies that become too small are dropped in favor of companies that grow sufficiently large to be included. This leads to a positive bias in the index since more successful companies are replacing less successful companies. The indexes only record changes in stock prices and do not add in stock dividends, nor do they adjust for price inflation. More complete analysis of these indexes uses the real returns after reinvestment of any dividends in the index.

There are many stock price indexes. The most commonly reported U.S. stock indexes are the following. The **Standard and Poors 500** (S&P 500) index is a market-value and float weighted index of 500 large U.S. companies. The **Dow Jones Industrial Average** is a price weighted index of 30 large U.S. companies. The Dow Jones gets a lot of attention for historical reasons, but the S&P 500 gives a much better picture of the overall stock market as it covers more companies and is market-weighted rather than price-weighted. The **Nasdaq Composite** index is a modified market-value weighted index of over 3,000 U.S. common stocks, and securities that are similar to stocks, that are traded on the NASDAQ stock market. The index especially tracks technology and growth companies. The **Russell 2000** index is a market-value and float weighted index of 2,000 small capitalization U.S. companies.

## 3.5 International Data

The following international datasets are available free, aside from some of the IMF data and more detailed OECD data.

The **Penn World Tables** (PWT), available at pwt.sas.upenn.edu, provide annual national accounts data for many countries. The data are made comparable across countries by determining the value of different countries' currencies using **purchasing power parity**, under which a representative basket of goods costs the same in each country, rather than using market exchange rates. These data are especially useful for cross-country comparisons of GDP per capita and productivity growth. PWT version 8 covers 167 countries over the years 1950-2011.

The **World Bank** website data.worldbank.org makes available annual time series data for over two hundred countries on a wide range of topics, including macroeconomic data and various socioeconomic, demographic and development measures. Data are presented in both the local currency and in U.S. dollars at market exchange rates. These data can also be accessed from within Stata using user-written command `wbopendata`.

The **Organisation for Economic Co-operation and Development** (OECD) has 34 member countries that include the most developed countries in Europe, North America, Asia and Australasia. The OECD website stats.oecd.org has very detailed data for these countries, including macroeconomic data and a wide range of socioeconomic, demographic and development measures. Additionally the OECD provides more detailed datasets, such as the Program for International Student Assessment (PISA).

The **International Monetary Fund** (IMF) has a number of global statistical databases. In particular, the **International Financial Statistics** (IFS) database covers international and domestic finance for over two hundred countries, beginning in 1948. IFS data is updated monthly and is available for free to subscribers from developing countries, as well as to students and researchers in the many academic institutions that subscribe to the database.

Other IMF databases include those on international trade flows, balance of payments and government finance.

**Eurostat**, a directorate-general of the European Commission, provides at the website epp.eurostat.ec.europa.eu harmonized statistics for European countries. The data cover a wide range of economic and related activities.

## 3.6 Individual-Level Data

Individual-level datasets can be difficult to manage due to their size, missing data, and the need for care in interpreting values. For example, a gender indicator variable may be coded as taking values 0 or 1 or as taking values 1 and 2, and one needs to also know which value indicates which gender. As another example, a zero may mean the variable takes value zero, or it may mean that the person was not asked the question because, for example, only women aged between 18 and 65 years were asked the question. When downloading data from these datasets one often needs to also download the codebook that defines the variable and its possible values, and possibly also download the survey questionnaire.

A major convenient source for individual-level data is the **Integrated Public Use Microdata Series (IPUMS)** at website www.ipums.org. IPUMS datasets include large subsamples of the decennial U.S. Censuses from 1850 onwards, the U.S. Current Population Surveys (CPS) since 1962, the U.S. American Community Surveys (ACS) from 2001, the U.S. National Health Interview Surveys from 1963 and the American Time Use Surveys from 2003. Additionally IPUMS has harmonized data from the cencuses of over seventy countries.

The IPUMS datasets are repeated cross-sections, asking similar questions in each survey but to a different sample of individuals. Panel datasets follow the same individual for several years. The main two U.S. panel datasets are the Panel Survey of Income Dynamics (PSID) at psidonline.isr.umich.edu and the National Longitudinal Surveys (NLS), most conveniently available at nlsinfo.org/investigator.

Many additional cross-section and panel datasets are available from the Inter-university Consortium for Political and Social Research (ICPSR) at the website icpsr.umich.edu.

## 3.7 Key Concepts

1. The main measure of aggregate output of the economy is GDP.

2. Price indexes measure how overall prices vary over time or regions,

3. The most commonly-used price indexes for the U.S. are the monthly CPI and the quarterly GDP price deflator.

4. Labor force statistics break the population into three categories – employed, unemployed and not in the labor force.

5. Key labor force statistics are the unemployment rate, the employment-population ratio, and the monthly change in employment.

6. Key interest rates for U.S. macroeconomic policy are the one-year treasury bill rate, the ten-year treasury bill rate, and the federal funds rate.

7. The three most reported U.S. stock price indexes are the S&P 500, the Dow Jones Industrial Average and the Nasdaq composite Index. These are computed in quite different ways.

8. A convenient source for the preceding macroeconomic data is FRED.

9. Good sources for international data include the Penn World Tables, World Bank, OECD, IMF and Eurostat.

10. Good sources for individual-level U.S. data are IPUMS and ICPSR.

11. The two main U.S. panel datasets on individuals are the PSID and the various NLS surveys.

# Chapter 4

# © A. Colin Cameron: Univariate Data Transformation

An important component of economic analysis is appropriately transforming the original data before analysis. For example, business cycle analysis focuses on real GDP and its growth rate rather than on nominal GDP.

The most commonly-used transformations of economics data are presented in this chapter. These transformations are relatively straightforward to understand, with the notable exception of the natural logarithm. The natural logarithm is extensively used in economics, much more so than in other areas of applied statistics. All applied economists should be familiar with the use of natural logarithms.

## 4.1 Natural Logarithm

We begin by defining the natural logarithm, and the related exponential, before presenting several ways in which the natural logarithm is used in economic data analysis.

### 4.1.1 Natural Logarithm and Exponential Functions Defined

A **logarithmic function** is the reverse operation to raising a number to a power. For example, $10^2 = 100$ implies that $\log_{10} 100 = 2$. In words, if 10 raised to the power 2 equals 100 then the logarithm to the base 10 of 100 is 2. More generally, for the logarithm to the base $a$ of $x$ equals $b$ we have that if $a^b = x$ then $\log_a x = b$.

There are many possible choices of the base for logarithms. The most obvious choice, given the decimal system, is to use the logarithm to base 10.

In economics analysis it is instead more common to use the **natural logarithm**, or

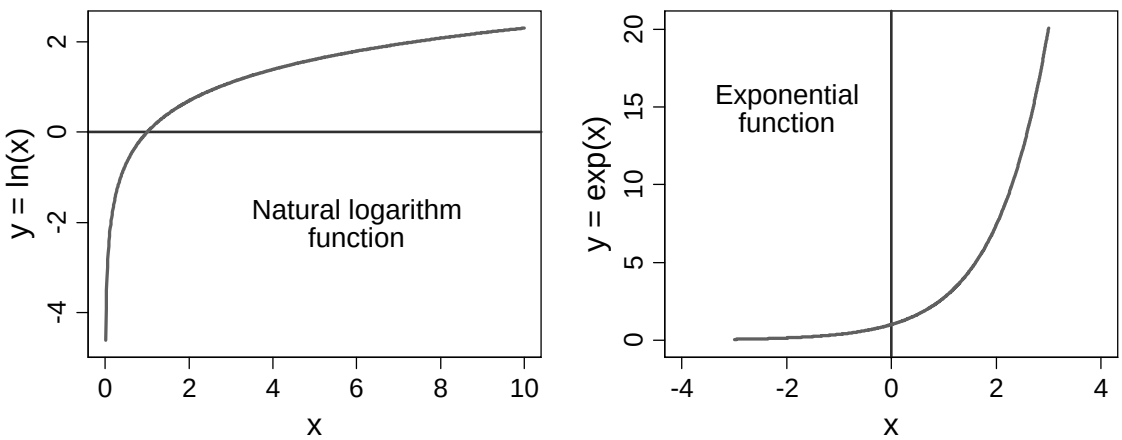


Figure 4.1: Exponential function and natural logarithm function

**logarithm to base** $e$. The natural logarithm of $x$ is usually denoted $\ln x$, so

$$\ln x = \log_e(x), \quad x > 0,$$

where $e \simeq 2.7182818...$ is an irrational number. More specifically, the number $e$ is a transcendental number, like $\pi \simeq 3.1415927....$ While $\pi$ can be easily defined as the area of a circle with radius 1, or as the perimeter of a circle with diameter 1, there is no such simple physical interpretation for $e$.

The natural logarithm function is plotted in the left panel of Figure 4.1 for $0 < x < 10$. The function $\ln x$ is defined only for positive $x$ and is always increasing in $x$, but at a decreasing rate as $x$ becomes larger.

The natural logarithm is the reverse operation to exponentiation. The **exponential function** is denoted

$$\exp(x) = e^x.$$

The relationship between the natural logarithm and exponential function is that

$$y = e^x \quad \Rightarrow \quad x = \ln y.$$

For example, $e^2 \simeq 7.38906$ so $\ln 7.38906 \simeq 2.0$.

The right panel of Figure 4.1 plots the exponential function for $-3 < x < 3$. The function $\exp(x)$ is always positive and always increasing in $x$, at an increasing rate as $x$ becomes larger. Note that $\exp(-x) = e^{-x} = 1/e^x = 1/\exp(x)$.

### 4.1.2 Approximating the Natural Logarithm and the Exponential

The use of the base $e$ for logarithms, rather than other bases, is attractive in part due to the following approximation which is the basis for many subsequent results.

For small $x$,

$$\ln(1+x) \simeq x, \quad \text{for, say, } x < 0.1.$$

See the Appendix 4.A for a derivation.

Table 4.1 compares the approximation $x$ to $\ln(1+x)$. The approximation works quite well for small $x$, though $x$ increasingly overestimates $\ln(1+x)$. For $x < 0.10$ the approximation is within five percent of $\ln(1+x)$, whereas the approximation $x = 0.2$, for example, is ten percent larger than $\ln 1.2 = 0.1823$. For many purposes the approximation is adequate for $x < 0.1$.

Table 4.1: Approximating ln(1+x) by x.

| | | **"Small" x** | | **"'Larger" x** | | |
|---|---|---|---|---|---|---|
| | | x=0.05 | x=0.10 | x=0.15 | x=0.20 | x=0.50 |
| True Value | ln(1+x) | 0.0488 | 0.0953 | 0.1398 | 0.1823 | 0.4055 |
| Approximation | x | 0.05 | 0.10 | 0.15 | 0.20 | 0.50 |

A related approximation for the exponential, also derived in Appendix 4.A, is that

$$e^x \simeq 1 + x, \quad \text{for, say, } x < 0.1.$$

While this approximation also works well for low $x$, it increasingly underestimates $e^x$ as $x$ increases.

Table 4.2 compares the approximation $1 + x$ to $\exp(x)$. The approximation works quite well for low $x$, though $1 + x$ increasingly underestimates $\exp(x)$ as $x$ increases.

Table 4.2: Approximating exp(x) by 1+x.

| | | **"Small" x** | | **"Larger" x** | | |
|---|---|---|---|---|---|---|
| | | x=0.05 | x=0.10 | x=0.15 | x=0.20 | x=0.50 |
| True Value | exp(x) | 1.0513 | 1.1052 | 1.1618 | 1.2214 | 1.6487 |
| Approximation | 1+x | 1.05 | 1.10 | 1.15 | 1.20 | 1.50 |

**Remark 1** *For small $x$ we can approximate $\ln(1+x)$ by $x$ and approximate $\exp(x)$ by $1+x$. For many purposes these approximations are adequate for $x < 0.10$.*

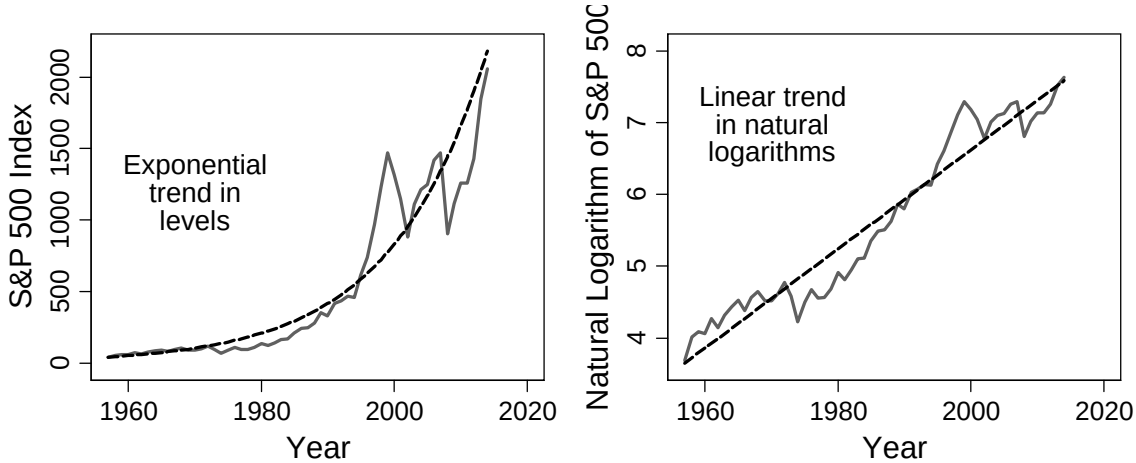


Figure 4.2: Standard and Poors 500 index: Levels and natural logarithm.

### 4.1.3 Linearizing Exponential Growth

Many economic quantities grow over time according to a power law, or exponentially. Examples include investments and price levels.

As an example consider the Standard and Poors 500 Stock Market Index. Year-end annual data in dataset SP500INDEX are plotted in the left panel of Figure 4.2 for the period 1957-2011. There is exponential growth until 1999, followed by seemingly erratic behavior with a second peak in 2007. The two dips in the 2000's are the "tech wreck" in 2003, when greatly inflated technology stocks lost much of their value, and the global financial crisis in 2008.

Let $x_t$ be a variable with **growth rate** that is constant over time. Then an **exponential time trend** with initial value $x_0$ at time 0 and growth rate $r$ leads to value $x_0(1+r)$ after one period, value $x_0(1+r) \times (1+r) = x_0(1+r)^2$ after two periods, and so on. In the $t^{th}$ period

$$x_t = x_0(1+r)^t.$$

When $r > 0$ there is **exponential growth**, while for $r < 0$ there is **exponential decay** such as in the case of depreciation of an asset.

For example, consider \$100 invested at annual interest rate of three percent for ten years. Converting from percentages, the annual growth rate is $r = 3/100 = 0.03$, and the investment is worth $100 \times (1.03)^{10}$ or \$134.39 after ten years.

The smoother curve in the left panel of Figure 4.2 is an exponential time trend for the S&P 500 data, a line of the form $x_0(1+r)^t$.

Taking the natural logarithm of both sides of the preceding equation yields

$$\begin{aligned} \ln x_t &= \ln(x_0(1+r)^t) \\ &= \ln x_0 + \ln(1+r)^t \qquad \text{using } \ln(ab) = \ln a + \ln b \\ &= \ln x_0 + \ln(1+r) \times t \quad \text{using } \ln(a^b) = (\ln a) \times b. \end{aligned}$$

It follows that the **natural logarithm** of $x_t$ is a **linear function** of $t$. For small $r$, $\ln(1+r) \simeq r$, so

$$\ln x_t \simeq \ln x_0 + r \times t.$$

The slope approximately equals $r$ for small $r$.

The right panel of Figure 4.2 plots the natural logarithm of the S&P 500 index against time. Superimposed is a **linear time trend**, a line of the form $a + rt$. This graph more clearly shows departures from trend than does the left panel of Figure 4.2. It shows another major dip in the early 1970's, a consequence of the first OPEC oil shock in 1973-74. And it indicates that stocks were rising much above trend rates in the late 1990's. The slope of the line is $r = .069$, so the index grew at a compound rate of 6.9 percent per year. This gain is a nominal gain that overstates real gains; complete analysis should go further and convert to a real return. At the same time the index understates nominal gains as it fails to include stock dividends.

In some examples in the physical sciences exponential growth is best represented as

$$x_t = \exp(a + rt).$$

Then taking the natural logarithm yields immediately that

$$\ln x_t = a + r \times t.$$

The natural logarithm is again linear in time, though in this case no approximation is needed for the slope to equal $r$.

**Remark 2** *For an exponential time trend with constant growth rate the natural logarithm is linear in time. The slope equals $r$ if $x_t = \exp(a + rt)$ and approximately equals $r$ if $x_t = x_0(1+r)^t$. Going the other way, if the natural logarithm is linear in time then growth in the original series is exponential.*

### 4.1.4 Approximating Proportionate Changes

Let $\Delta x = x_1 - x_0$ define the change in $x$ when $x$ changes from $x_0$ to $x_1$. Then the **proportionate change** in $x$ is

$$\frac{\Delta x}{x_0} = \frac{x_1 - x_0}{x_0}.$$

For example, consider a change from $x_0 = 40$ to $x_1 = 40.4$. Then $\Delta x = 40.4 - 40 = 0.4$, and the proportionate change in $x$ is $\Delta x/x_0 = 0.4/40 = 0.01$.

Interest often lies in proportionate changes in a series. For example, rather than being interested in the price level or changes in the price level, interest may lie in the price inflation rate, the proportionate change in prices. In that case it can be useful to transform data to the natural logarithm due to the following result.

For small proportionate changes we can use the following approximation

$$\Delta \ln x \simeq \frac{\Delta x}{x} \quad \text{for, say, } \frac{\Delta x}{x} < 0.1.$$

Multiplying by 100 yields percentage changes, so equivalently

$$100 \times \Delta \ln x \simeq \text{ Percentage change in } x.$$

For example, when $x$ changes from $x_0 = 40$ to $x_1 = 40.4$, then the approximation yields $\ln(40.4) - \ln(40) = 3.69883 - 3.68888 \simeq 0.00995$, close to 0.01000.

**Remark 3** *The natural logarithm function has the property that the change in the natural logarithm of $x$ approximately equals the proportionate change in $x$, for small proportionate changes in $x$. Multiplying by 100 gives percentage changes.*

Table 4.3 compares $\Delta \ln x$ to $\Delta x/x$. The approximation under-estimates the actual proportional change, and does less well as $\Delta x/x$ increases. By $\Delta x/x = 0.2$ the approximation instead yields 0.1823, and a 50% change is instead estimated as a 40.55% change. The approximation works best for series with relatively small changes, such as quarterly prices or annual prices or annual real interest rates.

As an example, return to the S&P 500 data plotted in Figure 4.2. From the second panel, the natural logarithm increased from 3.7 to 7.6 over the 56 years from 1957 to 2013. The annual increase if $(7.6 - 3.7)/56 = 0.070$, an increase of 7.0 percent per year.

Table 4.3: Approximating proportionate change in x by the change in ln(x).

| | | **"Small"** $\Delta x/x$ | | **"Larger"** $\Delta x/x$ | | |
|---|---|---|---|---|---|---|
| | | $\Delta x/x$ =0.05 | $\Delta x/x$=0.10 | $\Delta x/x$=0.15 | $\Delta x/x$=0.20 | $\Delta x/x$=0.50 |
| True Value | $\Delta x/x$ | 0.05 | 0.10 | 0.15 | 0.20 | 0.50 |
| Approximation | $\Delta \ln x$ | 0.0488 | 0.0953 | 0.1398 | 0.1823 | 0.4055 |
| Percentage difference | | -2.4% | -4.7% | -6.8% | -12.5% | -18.9% |

### 4.1.5 Compounding and the Rule of 72

Suppose money is invested at 3 percent per annum. How many years will it take for the investment to double in value? Similarly if inflation averages four percent a year, in how many years will price levels double?

More generally, suppose money is invested at **proportionate rate** $r$ per period. Then after $n$ periods the investment is $(1+r)^n$ times larger. It follows that the number of periods $n$ that it takes for money to double is the solution to the equation $(1+r)^n = 2$. Taking the natural logarithm of both sides and solving, see Appendix 4.A, yields

$$n = \ln 2/[\ln(1+r)].$$

We make the following approximations. First, $\ln(1+r) \simeq x$ for small $x$. Second, $\ln 2 = 0.6931 \simeq 0.72$. So $n \simeq 0.72/r$.

It is more convenient to work with growth rates expressed as a percentage. Then for money invested at **percentage rate** $r$ per period it takes approximately

$$n = 72/r$$

for the investment to double.

**Remark 4** *The **rule of 72** states that it takes approximately $n = 72/r$ periods for a series to double if it is growing at the percentage rate $r$.*

Table 4.4 gives the number of years that it takes for money to double when invested at $r$ percent per annum, using the exact result and using the rule of 72 approximation. The approximation for $r < 10\%$ is very good, only slightly overestimating the number of periods to double.

Table 4.4: Number of periods for investment at percentage rate r per period to double.

| | | **r=1%** | **r=2%** | **r=3%** | **r=4%** | **r=6%** | **r=8%** | **r=9%** |
|---|---|---|---|---|---|---|---|---|
| True Value | $\ln 2/\ln(1+r/100)$ | 69.66 | 35.00 | 23.45 | 17.67 | 11.90 | 9.01 | 8.04 |
| Rule of 72 approximation | $72/r$ | 72 | 36 | 24 | 18 | 12 | 9 | 8 |

While $\ln 2 \simeq 0.6931$, it is more convenient to use the cruder approximation $\ln 2 \simeq 0.72$ as the integers 2, 3, 4, 6, 8 and 9 all divide exactly into 72. Variations of this rule that are more precise for low $r$ but not for moderate $r$ are the **rule of 70**, using $70/r$, and even the **rule of 69**, using $69/r$.

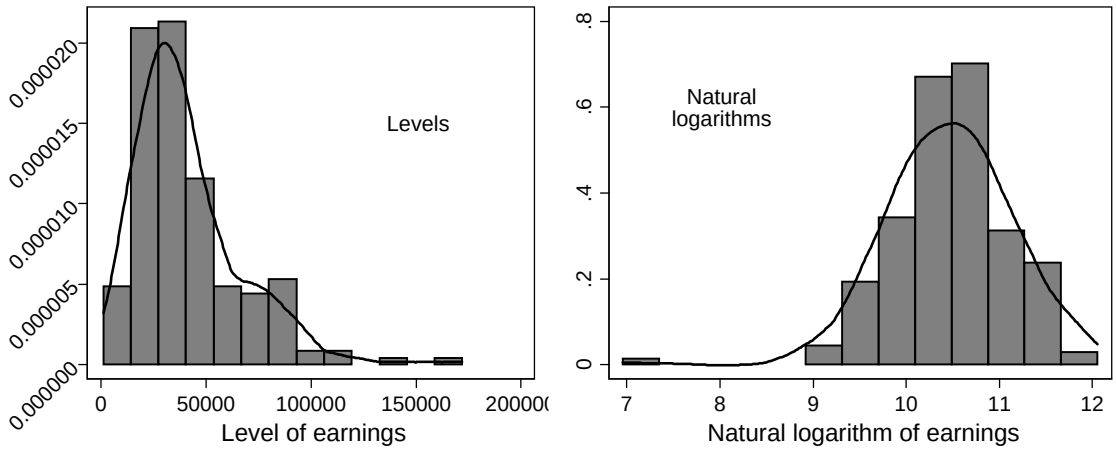


Figure 4.3: Annual earnings in levels and in natural logarithms: Histograms and kernel density estimates.

### 4.1.6 Eliminating Right Skewness

Many cross-section data sets can be right-skewed. For example, data on income or wages of a sample of individuals are often right-skewed. The natural logarithm transformation can lead to a transformed data series that is more symmetrically distributed. If a variable $x$ is such that $\ln x$ is normally distributed, then $x$ itself is said to follow the **lognormal distribution**.

The left panel of Figure 4.3 presents a histogram of earnings of female full-time workers aged 30 in 2010, along with a kernel density estimate, using data from dataset EARNINGS introduced in Chapter 2. The data are clearly right-skewed. The second panel of Figure 4.3 shows the histogram after transformation to natural logarithms. The histogram is close to symmetric, aside from one very small value (the sample includes an observation with unusually low annual earnings of \$1,050 and corresponding low natural logarithm of 6.96), and is approximately normally distributed.

### 4.1.7 Compound Interest Rates

Interest rates are often reported as annual interest rates, even though interest on the investment may be calculated more frequently than annually, such as monthly or even daily. In that case one needs to be careful in defining the annual interest rate.

Suppose that an interest rate is compounded monthly at a monthly interest rate of 1 percent. Then the nominal annual interest rate is defined to be the monthly interest rate times the number of months, or $12 \times 1 = 12$ percent. The effective annual interest

rate is more than 12 percent, due to compounding. Compounding over twelve months, $(1 + .01)^{12} = 1.12683$, leading to an effective annual interest rate of 12.683 percent.

In general if the **nominal annual interest rate** is $r$ and there are $n$ compounding periods per year then the **effective annual interest rate** is

$$r_{effective} = (1 + r/n)^n - 1.$$

The nominal annual interest rate is also called the **annual percentage rate** (APR) while the effective annual interest rate is also called the **annual percentage yield** (APY).

What if we compound daily rather than monthly? Then the annual effective interest rate becomes 12.750 percent, since $(1 + 12/365)^{365} = 1.12750$.

If one **continuously compounds** an interest rate at progressively smaller intervals, then

$$(1 + r/n)^n \to e^r \quad \text{as } n \to \infty.$$

In the current example $\exp(0.12) = 1.12750$, which equals the earlier result for daily compounding to four decimal places and the earlier result for monthly compounding to three decimal places.

## 4.2 Other Transformations

In this section we present a range of commonly-used transformations other than the natural logarithm.

### 4.2.1 Standardized Scores (z scores)

A **standardized score** is obtained by subtracting the mean and dividing by the sample standard deviation. Thus

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, ..., n,$$

where $x_i$ is the original value, $\bar{x}$ is the sample mean and $s$ is the sample standard deviation. The resulting score has mean zero and standard deviation one. A standardized score is often called a **z-score** as its distribution may be well approximated by a standard normal distribution, which also has mean 0 and variance 1. Note also that the symmetry and kurtosis statistics given in Chapter 2.1 approximately equal the sample averages of the standardized scores for each observation raised to, respectively, the third and fourth power.

A standardized score is immediately interpretable – a one unit increase in $z_i$ equals a one standard deviation increase in the original score $x_i$.

Standardized scores are useful for comparing data series that are scaled differently. For example, suppose we wish to compare student performance on two tests with different total

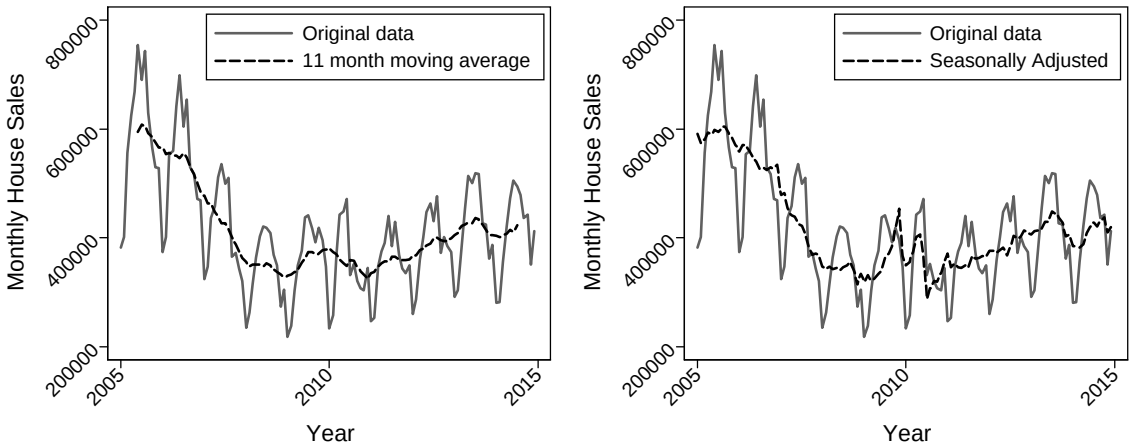


Figure 4.4: Monthly total revolving consumer credit outstanding: Smoothing by moving average and by seasonal adjustment.

points or of different difficulty, so that the sample means and standard deviations differ across the tests. Then we compare the sample values of the standardized scores $z_{1i} = (x_{1i} - \bar{x}_1)/s_1$ and $z_{2i} = (x_{2i} - \bar{x}_2)/s_2$, where the subscripts 1 and 2 denote the first and second tests.

### 4.2.2 Moving Averages

A **moving average** smooths data by taking the average of observations in several successive periods. This is especially useful for time series data. For data that bounce around from period to period, averaging can smooth the data. Visual analysis of long-term trends in the data are easier to see, since period-to-period variation is reduced.

A **simple moving average** averages the current and immediate past observations. For example, a five-period moving average takes the average of the data over the current and preceding four periods, or $(x_t + x_{t-1} + x_{t-2} + x_{t-3} + x_{t-4})/5$.

If instead the current observation appears in the middle, then the moving average is a **centered moving average.** For example, a centered five-period moving average takes the average of the data two periods ago, one period ago, this period, next period, and the period after that, or $(x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2})/5$. The centered moving average at time $t$ has the disadvantage that it is not immediately available at time $t$ as its computation also uses data from some future time periods.

As an example consider U.S. monthly data from 2005 to 2014 on sales of existing homes, compiled by the National Association of Realtors. The data are in dataset HOUSING. The first panel of Figure 4.4 plots the original data along with an eleven-month centered moving

average. The original data are relatively variable within a year, with a low point in January and February and a peak in summer to early fall. The data also indicate the large decrease accompanying the global financial crisis, with a thirty percent decline compared to 2005. The moving average smooths the data considerably. Note that centering the moving average comes at the expense of it not being computable for the most recent months as it requires data in future months.

### 4.2.3 Seasonal Adjustment

For data that fluctuate within the year due to seasonal influences it can be useful to smooth out the seasonal variation. For example, since retail sales are highest in December it would be misleading to compare January sales to December sales and interpret the decrease in sales in January as an economic slump.

**Seasonal adjustment** smooths data to control for seasonal variation in the data. For example, monthly data are decreased in months that have relatively high values every year and are increased in months that have unusually low values every year.

The monthly data on total revolving consumer credit outstanding given in Figure 4.4 show a clear spike in credit outstanding in December of each year, due to the Christmas shopping season. Seasonal adjustment will ideally eliminate the spike and just show the underlying trend in the data.

The second panel of Figure 4.4 presents published seasonally adjusted data for the same series, using the widely-used X-11-ARIMA seasonal adjustment program developed by the U.S. Census Bureau. The seasonal adjusted series is much smoother than the original, and essentially eliminates the seasonal variation.

Many macroeconomics series are released as **seasonally adjusted data**. Analysts interpreting these data should be aware that there is no indisputable best way to seasonally adjust.

### 4.2.4 Real and Nominal Data

Economics data are often measured in dollars. Any meaningful interpretation of these data over time requires conversion to the purchasing power of a dollar in a benchmark year. The original data are called **nominal** data, measured in **current dollars**. Thus 1990 data are measured in 1990 dollars, 1991 data are measured in 1991 dollars, and so on. The data after conversion are called **real data**, measured in **constant dollars**. Then data in various years are reported in dollars of a given year, say 2009 dollars for example. Similar conversion using exchange rates or purchasing power parity indexes is needed to compare nominal data across countries with different currencies.

The **index number problem** is that there is no perfect way to create a price index (or a quantity index) when both prices and quantities of the various goods and services that are

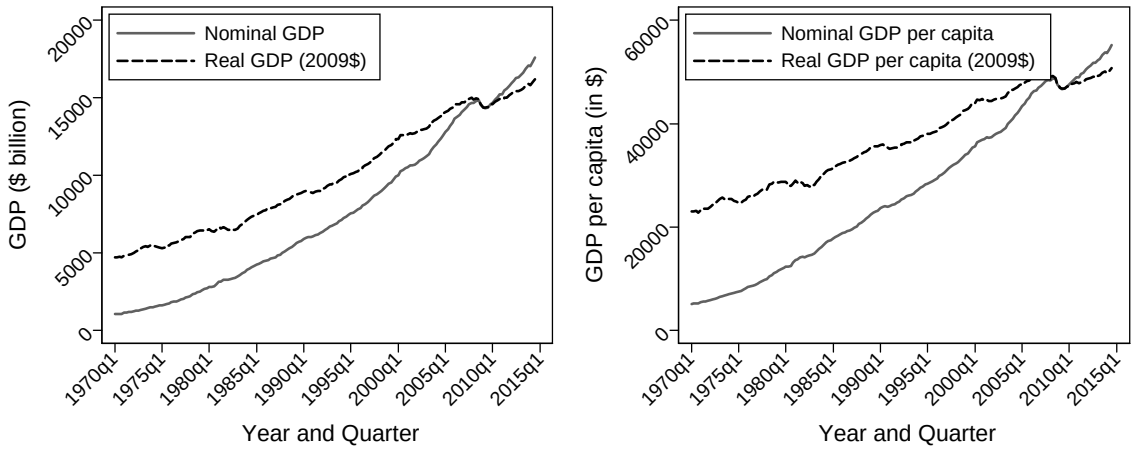


Figure 4.5: GDP and per capita GDP in U.S.: nominal in current dollars and real in 2009 dollars.

components of the index change over time. The leading published indexes use methods that control partially for this problem; see Chapter 3.2 for discussion of price indexes.

As an example of use of real data rather than nominal data, consider U.S. Gross Domestic Product (GDP), the standard measure of the economy's total output. The solid line in the first panel of Figure 4.5 gives nominal GDP from 1970 to 2014. The data in dataset REALGDPPC are seasonally adjusted quarterly data, annualized by multiplying by four. Nominal GDP has increased 16.7 times, from \$1,054 billion to \$17,600 billion. The fall in GDP in the recession of 2007-2009 is clearly visible.

Part of this large increase in nominal GDP reflects price inflation – a dollar in 1970 had much more purchasing power than a dollar in 2014. The conversion from nominal to real data is done by using a **price index,** which measures prices relative to a value of 100 in a base year. Here we use the GDP chain-type price index, normalized to equal 100 in 2009. The index in the first quarter of 1970 was 22.382, so a 1959 dollar was worth $100/22.382 = \$4.47$ in 2009 dollars, and 1959 first quarter nominal GDP of \$1,054 billion was worth \$4,710 billion ($1054 \times 100/22.382$) in 2009 dollars. Similarly in the third quarter of 2014 the index was 108.603 and nominal GDP of \$17,600 billion was worth \$16,210 billion in 2009 dollars. Table 4.5 summarizes these calculations.

The dashed line in the first panel of Figure 4.5 gives real GDP from 1970 to 2014, measured in 2009 dollars. Real GDP increased 3.4 times, from \$4,710 billion to \$16,210 billion. This is still a substantial increase, but it is much less than the 16.7 times increase in nominal GDP. The difference is due to a 4.9 times (108.603/22.382) increase in prices over this period, leading to real GDP rising $16.7/4.8 = 3.4$ times.

Table 4.5: Nominal and real data: Calculations.

| **Time Period** | **Nominal Value** Current \$ billions | **Index** 2009=100 | **Real Value** 2009 \$ billions |
|---|---|---|---|
| 1970 Q1 | 1,054 | 22.382 | 1054×100/22.382=4,710 |
| 2014 Q3 | 17,600 | 108.603 | 17,600×100/108.603=16,210 |

The recessions in 1973-74, 1980, 1982 and 1991 become more pronounced using real GDP data, with more pronounced dips due to eliminating increases in nominal GDP that occur due to price inflation.

What if we used a year different from 2009 as the base year? Then the price index will differ and real GDP will differ as it is no longer measured in 2009 dollars. But the resultant proportionate changes will be unchanged, with a 4.9 times rise in prices and real GDP becoming 3.4 times larger over the 44 years.

### 4.2.5 Per Capita Data

**Per capita data** are data formed from an original series by dividing by the size of the population.

In some cases interest lies in aggregate data and in some cases per capita data. For example, to compare the size of the economy over time use real aggregate GDP, but to compare living standards over time use real per capita GDP.

As already noted, real GDP increased 3.4 times from 1970 to 2014. But the U.S. population is 1.56 times larger, with increase from 204 million to 319 million. Thus real per capita GDP has grown about 2.2 times, as $3.4/1.56 = 2.2$. This is illustrated in the second panel of Figure 4.5, with increase in real GDP per capita from \$23,000 to \$50,800. This is still an appreciable improvement over time, and is about 1.8 percent per annum, since $1.018^{44} \simeq 2.2$. But it is nowhere near as large as the initial starting point of a 16.7 times increase in nominal GDP.

### 4.2.6 Growth Rates and Percentage Changes

If interest lies in changes over time it can be convenient to transform to **percentage changes** or **growth rates**. For example, to analyze changes in living standards we consider percentage changes in real per capita GDP over time.

The **one-period percentage change** is calculated as

$$\text{Percentage change } = 100 \times \frac{x_t - x_{t-1}}{x_{t-1}}.$$

In many cases this is converted to an **annualized rate**. For example, for quarterly data the quarterly change multiplied by four gives the annualized quarterly change. Alternatively, for quarterly data one can instead compute a four-period percentage change, $100 \times (x_t - x_{t-4})/x_{t-4}$, which also expresses the change as an annual rate. For data that are not already seasonally adjusted, this latter method can smooth out quarterly seasonal fluctuations. Similarly for monthly data we may use $100 \times (x_t - x_{t-12})/x_{t-12}$.

An alternative calculation method for computing approximate percentage changes is to use

$$\text{Percentage change } \simeq 100 \times (\ln x_t - \ln x_{t-1}).$$

This has already been presented in Chapter 4.1.

Leading examples of use of percentage changes are the real GDP growth rate and the price inflation rate, presented in Chapter 3.

Potential confusion can arise when statements are made about changes in growth rates or interest rates. For example, suppose the growth rate increases from 3 percent in one year to 5 percent the next year. It is misleading to call this a 2 percent increase in the growth rate, since this literally means that an increase in the growth rate from 3.0 percent to $3.0 \times 1.02 = 3.06$ percent. Instead the correct term to use is that the growth rate increased by two **percentage points**. Very small changes are described in **basis points**, where a basis point is one-hundredth of a percentage point. For example, an increase from 3.0 percent to 3.15 percent is an increase of fifteen basis points.

Often U.S. real GDP growth is compared to that in western Europe or Japan. U.S. growth in real GDP is higher, but so too is its population growth. In fact the growth rate for per capita real GDP in the U.S. is similar to that in western Europe and Japan.

## 4.3 Key Concepts

1. The natural logarithm $\ln x$ is the reverse of exponentiation.

2. For small $x$, say $x < 0.1$, $\ln(1 + x) \simeq x$ and $\exp(x) \simeq 1 + x$.

3. If $x$ grows at a constant exponential (or multiplicative) rate then $\ln x$ grows linearly.

4. For small $\Delta x/x$, say $x < 0.1$, $\Delta \ln x \simeq \Delta x/x$. In words, the change in $\ln x$ approximates the proportionate change in $x$.

5. For a series growing at percentage rate $r$ per period it takes approximately $72/r$ periods to double.

6. If the distribution of data $x$ is right-skewed then the distribution of $\ln x$ may be close to symmetric.

7. A standardized score (or z-score) $z = (x - \bar{x})/s$ has mean zero and standard deviation one.

8. A moving average smooths data by taking the average of several successive observations.

9. Seasonal adjustment smooths data by controlling for seasonal fluctuations.

10. Nominal data are data measured in current dollars.

11. Real data are measured in constant dollars of a baseline period, to control for changes in purchasing power.

12. Per capita data are obtained by dividing by the size of the population.

13. Growth rates are percentage changes in a series. Monthly and quarterly rates are often reported at an annualized rate.

14. A change from, for example, 3% to 5% is a 2 percentage points change.

15. Key words: Logarithm; natural logarithm; exponential function; linear trend; exponential trend; rule of 72; lognormal distribution; moving average; seasonal adjustment; nominal data; real data; current dollars; constant dollars; price index; per capita data; growth rate; percentage change; percentage point change; basis points.

## 4.4 Exercises

1. Answer the following.

   (a) Suppose $y = \exp(a + bx)$. What is $\ln y$?

   (b) Suppose $\ln y = c + dx)$. What is $e^y$?

   (c) Suppose $y = (a + bx)^c$. What is $\ln y$?

   (d) Suppose $y = x \times z$. What is $\ln y$?

   (e) What does $\ln 1.02$ approximately equal?

2. Suppose $x_t = x_0 \times 1.04^t$ and $x_0 = 100$.

   (a) Plot $x_t$ against $t$ for $t = 1, ..., 100$. Comment on the shape of the curve.

   (b) Plot $\ln x_t$ against $t$ for $t = 1, ..., 100$. Comment on the shape of the curve.

3. Suppose $x_t$ grows exponentially at 6% per year with initial value $x_0 = 80$.

(a) Give the formula for $x_t$.

(b) Exactly how many years (or part thereof) will it take before $x_t = 160$?

(c) Give the formula for $\ln x_t$.

(d) Using the rule of 72, approximately how many years will it take before $x_t = 160$?

(e) Compare your answers in parts b and d and comment.

4. For the function $y = x^3$:

   (a) Find $y$ when $x$ changes by $\Delta x$ to $x + \Delta x$.

   (b) Hence find $\Delta y$.

   (c) Hence find $\Delta y/\Delta x$.

   (d) Hence find $\Delta y/\Delta x$ as $\Delta x \to 0$; this is the derivative $dy/dx$.

5. For the exponential function $y = e^x$:

   (a) Find $y$ when $x$ changes by $\Delta x$ to $x + \Delta x$.

   (b) Hence show that $\Delta y = e^x(e^{\Delta x} - 1)$. Hint: $e^{a+b} = e^a e^b$.

   (c) Hence find $\Delta y/\Delta x$.

   (d) Simplify using the result that $e^{\Delta x} = (1 + \Delta x)$ for small $\Delta x$ to obtain $\Delta y/\Delta x$ as $\Delta x \to 0$; this is the derivative $dy/dx$.

6. For the time series variable $x_t$ that takes consecutive values 5, 3, 7, 8 for $t = 1, ..., 4$:

   (a) Calculate $\sum_{t=1}^{4} x_t$.

   (b) Calculate $\Delta x_t$ for $t = 2, 3, 4$ and hence $\sum_{t=2}^{4} \Delta x_t$.

7. Suppose $x$ increases from 500 to 520.

   (a) What is the proportionate change in $x$?

   (b) What is the absolute change in $\ln x$?

   (c) Compare your answers in parts a and b and comment.

   (d) Repeat parts a-c if $x$ increases from 500 to 1000.

8. Monthly data for various stock indexes are given in dataset STOCKINDEX.

   (a) Plot the Dow Jones index against time from January 1957 to November 2011. Does the growth appear to be (approximately) linear or exponential?

(b) Which period had the greatest absolute decrease in the index?

(c) Now take the natural logarithm of the index and plot against time. Do you think the original index grows (approximately) exponentially? Explain.

(d) State which decade(s) had relatively high proportionate growth in the index.

(e) Given the value of the natural logarithm at the start and end dates, what do you think is the approximate average monthly rate of growth in the index? Explain.

9. Repeat the analysis of the previous question using the following stock indexes in dataset STOCKINDEX.

   (a) Nasdaq index from February 1971 to November 2011.

   (b) Russel index from September 1987 to November 2011.

   (c) Harder: Dow Jones index from January 1901 to December 1956 using monthly data that you obtain from the web.

10. Use data in dataset STOCKINDEX from September 1987 to November 2011.

    (a) Calculate the z-score for each of the Dow Jones, Nasdaq and Russel indexes.

    (b) Do these z-scores have mean zero and standard deviation one?

    (c) Give histograms (or kernel density estimates) for each of the three z-scores. Do they appear to be normally distributed?

    (d) On the same graph plot each of the three z-scores against time. Do the three series appear to move together? Explain.

11. Use data in dataset STOCKINDEX from January 1965 to December 1972.

    (a) Calculated a simple five-period moving average for the Dow Jones index.

    (b) On the same graph plot the original series and the moving average. Comment.

    (c) Would a centered moving average be better?

12. Suppose the GDP chain-type price index (1996 = 100) is 59 in 1980 and is 108 in 2000.

    (a) How much higher are prices in 2000 than in 1980?

    (b) Given your answer in a, what is the average inflation rate over this period if inflation is compounded annually?

13. Use data in dataset GDPAUSTRALIA from January 1960 to September 2013. The data are quarterly data on nominal GDP (at an annual rate in millions of Australian dollars), a price index (=100 in 2011) and population (in millions).

    (a) Plot nominal GDP and real GDP (which you need to create) against time. Comment.

    (b) Compute nominal GDP per capita and real GDP per capita and plot these against time. Comment.

    (c) Take the natural logarithm of real per capita GDP and use the rule of 72 to calculate the approximate quarterly growth rate.

14. Use data in dataset GDPAUSTRALIA, described in the previous question.

    (a) Compute a four period moving average for nominal GDP. Has this reduced seasonal variation?

    (b) Compute annual growth rate in real GDP as four times the proportionate change from one quarter to the next.

    (c) Compute annual growth rate in real GDP as the proportionate change over the last four quarters.

    (d) Compare the two growth rate measures and comment.

## 4.5 Appendix 4.A: Derivations for the Natural Logarithm

Several properties of the exponential function and the natural logarithm function are given in Table 4.6. Here $n! = n \times (n-1) \times \cdots \times 1$ so, for example, $3! = 3 \times 2 \times 1 = 6$.

To obtain the result that $\ln(1+x) \simeq x$ for small $x$, note that from Table 4.6,

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots .$$

For small $x$ the first term in the sum is much larger than the others, yielding $\ln(1+x) \simeq x$. For example, $\ln(1.1) = \ln(1+0.1) = 0.10 - 0.1^2/2 + 0.1^3/3 - \cdots = 0.10 - 0.005 + 0.00033 - \cdots$ If we consider just the first term we obtain $\ln(1.1) \simeq 0.1$, compared to $\ln(1.1) \simeq 0.09533$. Of course if we sum the first three terms, say, we obtain a better approximation with $\ln(1.1) \simeq 0.09533$.

Similarly the result that $e^x \simeq 1 + x$ for small $x$ follows from

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots .$$

Table 4.6: Exponential and natural logarithm functions: Properties.

| Exponential Function | Example |
|---|---|
| $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$; | $e^{.05} = 1 + \frac{.05}{1} + \frac{.05^2}{2} + \frac{.05^3}{6} + \cdots$ |
| $e^x = \lim_{n\to\infty} (1 + \frac{x}{n})^n$ | $e^{.05} = \lim_{n\to\infty} (1 + \frac{.05}{n})^n$ |
| $e^x \times e^y = e^{x+y}$ | $e^2 \times e^3 = e^5$ |
| $e^x \simeq 1 + x$ for small $x$ | $e^{.05} \simeq 1 + .05 = 1.05$ |
| $de^x/dx = e^x$ | |
| **Logarithmic Function** | **Example** |
| $\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$ | $\ln 1.1 = \ln(1 + .1) = .1 - \frac{.1^2}{2} + \frac{.1^3}{3} + \cdots$ |
| $\ln x^a = a \ln x$ | $\ln x^2 = 2 \ln x$ |
| $\ln(x \times y) = \ln x + \ln y$ | $\ln(2 \times 3) = \ln 2 + \ln 3$ |
| $\ln(1 + x) \simeq x$ for small $x$ | $\ln(.95) = \ln(1 - .95) \simeq -.05$ |
| $d \ln x/dx = 1/x$ | |

For small $x$ the first term in the sum is much larger than the others, yielding $e^x \simeq 1 + x$.

To obtain the result that $\Delta \ln x \simeq \Delta x/x_0$, for small $\Delta x/x_0$, note that for $x$ changing from $x_0$ to $x_1$ we have

$$\begin{aligned}
\Delta \ln x &= \ln x_1 - \ln x_0 \\
&= \ln\left(\frac{x_1}{x_0}\right) && \text{from the property of logarithms} \\
&= \ln\left(1 + \frac{x_1 - x_0}{x_0}\right) && \text{manipulating} \\
&= \ln\left(1 + \frac{\Delta x}{x_0}\right) \\
&\simeq \frac{\Delta x}{x_0} && \text{for small } \frac{\Delta x}{x_0}.
\end{aligned}$$

The key step is using the approximation $\ln(1 + \Delta x/x_0) \simeq \Delta x/x_0$ for small $\Delta x/x_0$.

For those familiar with differential calculus, the property that $\Delta \ln x \simeq \Delta x/x$ follows almost immediately from the result that the derivative $d \ln x/dx = 1/x$, so $\Delta \ln x/\Delta x \simeq 1/x$ for small $x$ and rearranging yields $\Delta \ln x \simeq \Delta x/x$.

The number of periods $n$ that it takes an investment at proportionate rate $r$ to double is the solution to $(1 + r)^n = 2$. Then

$$\begin{aligned}
& (1+r)^n &&= 2 \\
\Rightarrow\ & \ln[(1+r)^n] &&= \ln 2 && \text{taking the natural log of both sides} \\
\Rightarrow\ & n \ln(1+r) &&= \ln 2 && \text{as } \ln x^b = b \ln x \\
\Rightarrow\ & n &&= \ln 2/\ln(1+r) && \text{solving for } n.
\end{aligned}$$

This is the exact result given in the main text.

# Chapter 5

# © A. Colin Cameron: The Sample Mean

Obtaining the sample mean $\bar{x}$, and other summary statistics, is straightforward. But different samples will yield different values of these sample statistics, due to the inherent randomness in the data. How can this randomness be controlled for if we want to make statements about the unchanging features of the distribution for the entire population? More simply, how can we extrapolate from the sample to the population?

For example, dataset EARNINGS introduced in Chapter 2 has data on individual annual earnings for a sample of 30-year-old female full-time workers. The sample mean from a sample of size 171 was \$41,413. What can be said about the likely range of values of mean earnings for all 30-year-old female full-time workers in the country? Are mean earnings in this population really as high or as low as \$41,413? Or is the observed sample mean of \$41,413 just an artifact of this particular sample? To make progress we need to define the concept of a mean in the population, and provide a measure of how precisely the sample mean estimates the population mean.

The chapter is relatively dense. While selecting only the essential material, it introduces a considerable amount of the probability theory covered in an introductory probability and statistics course. The focus is on the concepts of population mean and population variance of an individual observation, and the consequent distribution of the sample mean. The simplest material is presented in the text, with additional background material on probability and derivations for the sample mean presented in two chapter appendices.

For readers not wanting this depth of analysis, the fundamental insights are presented in Chapters 5.1-5.3. For those who skip the remainder of the chapter, the essential properties of the sample mean are restated in Chapter 6.2.

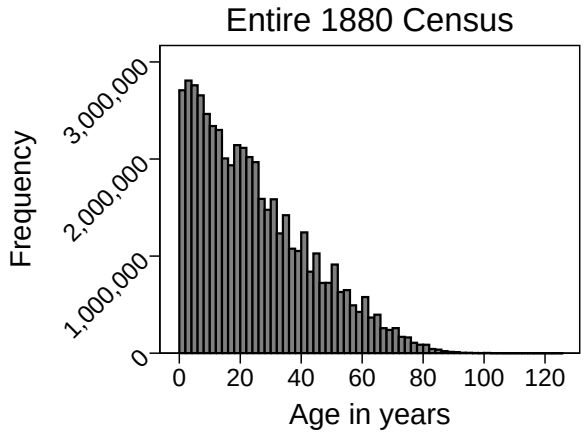


Figure 5.1: 1880 Census: Histogram of age for the entire population.

## 5.1 Sampling from a Finite Population

The definition of a **population** is simplest when the population is finite, such as the collection of all people in a country. Then the **population mean**, denoted $\mu$ and pronounced "**mew**", is the average for all values in the population. Similarly the **population standard deviation**, denoted $\sigma$ and pronounced "**sigma**", is the standard deviation average for all values in the population.

To illustrate the distinction between sample and population, we take a series of random subsamples from the population, for each subsample obtain the sample mean $\bar{x}$, and compare the various sample means $\bar{x}$ to the population mean $\mu$.

### 5.1.1 Example: 1880 U.S. Census

The 1880 Census provides a complete enumeration of the U.S. population in 1880. We consider one of the variables that was recorded, that on age in years.

Figure 5.1 provides a histogram of age for all 50,169,452 people recorded as living in the U.S. in 1880. The distribution is basically declining in age. The blips are due to individuals rounding their age to the nearest five years or ten years.

For a complete census such as this, the observed distribution is actually the **population** distribution. The population average age is 24.13 years, so $\mu = 24.13$, and the population standard deviation of age is 18.61, so $\sigma = 18.61$.

Now consider taking one randomly-drawn **sample** of size $n = 25$ drawn from this population of size $N = 50{,}169{,}452$. The left panel of Figure 5.2 presents the histogram for this single sample of size $n = 25$. For this sample, the average age was 27.84 years, so $\bar{x} = 27.84$, and

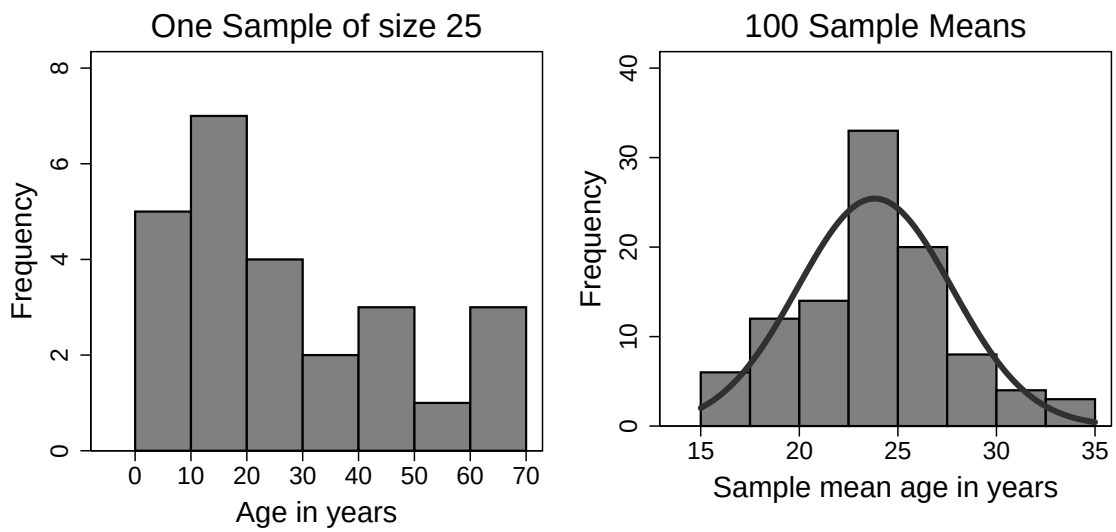


Figure 5.2: 1880 Census: Histogram for one randomly-drawn sample (n = 25) and histogram with kernel density estimate for 100 sample means of age (n = 25).

the standard deviation of age was 20.71, so $s = 20.71$. Due to the randomness of sampling, these are similar to, but not exactly equal to, $\mu$ and $\sigma$.

Now randomly draw 100 distinct samples of size 25, leading to 100 different sample means. The first three such samples turned out to have means $\bar{x}_1 = 27.84$, $\bar{x}_2 = 19.40$ and $\bar{x}_3 = 23.28$ years. The right panel of Figure 5.2 presents a histogram, along with the corresponding kernel density estimate (a smoothed histogram), for these 100 sample means that are stored in dataset AGEMEANS. Several things are apparent. First, the histogram is roughly centered on the population mean. In fact the average of the 100 means is 23.83, close to $\mu = 24.13$. Second, there is much less variability in these 100 means than in the original population. Here the standard deviation of the 100 means is 3.92, roughly one-fifth of the population standard deviation of $\sigma = 18.61$. Third, the estimated density is roughly that of a normally distributed random variable.

The preceding example illustrates the more general result that for random samples from a population

- the average of many sample means is close to the population mean
- the sample mean is much less variable than the individual underlying observations
- the sample mean is approximately normally distributed, provided the sample is sufficiently large.

## 5.2 Population and Sample

We now provide a more formal treatment that defines the population mean and population standard deviation in settings such as the fraction of coin tosses that result in heads more general settings than the previous finite population example. Additional details are provided in Appendix 5.A.

### 5.2.1 Population and Sample

The following formal terminology is used.

- The **population** is the set of all observations or measurements or experimental outcomes.
- The **sample** is a subset selected from the population.

To make clear the distinction between the sample and the population, the statistics convention is to use capitalized letters for **random variables**, e.g. $X$, and lower-case letters, e.g. $x$, to denote **sample realizations** (or outcomes) of the random variable. Thus

$$\begin{array}{ll} X & \text{denotes a random variable} \\ x & \text{denotes a realization of } X. \end{array}$$

For example, if the population consists of five possible values 1, 2, 3, 4 and 5, and the value 3 is randomly drawn, then the random variable $X$ has realized value $x = 3$.

A sample of size $n$ consists of $n$ draws from the population. Each draw is a realization of a random variable. The random variable $X_1$ may be the earnings of the first person chosen randomly from the population, $X_2$ is the second value, and so on. The observed value of $X_1$ is denoted $x_1$, the observed value of $X_2$ is denoted $x_2$, and so on.

**Remark 5** *A sample of size $n$ has observed values $x_1, x_2, ..., x_n$ that are realizations of the random variables $X_1, X_2, ..., X_n$.*

### 5.2.2 Population Mean

The **population mean** of $X$, denoted $\mu$ or $\mu_X$, is the probability-weighted average of all values of $X$ in the population. The notation $\mu$, pronounced "mew", is used to denote the mean as $\mu$ (or mu) is the Greek letter for m.

The population mean is also denoted $\mathrm{E}[X]$, the **expected value** of the random variable $X$. It is called the expected value of $X$ as it is the long-run average value that is expected if we draw a value of $X$ at random, draw a second value of $X$ at random, and so on, and then obtain the average of these values.

The population mean is easily defined for a **finite population of size** $N$. Then $\mu$ equals the average of the $N$ population values of $X$. Denote the population values for population members $1, 2, ...., N$ by $x_1^*, x_2^*, ..., x_N^*$ . Then **the population mean is the population average**

$$\mu \equiv \mathrm{E}[X] = \frac{x_1^* + x_2^* + \cdots + x_N^*}{N} = \frac{1}{N}\sum\nolimits_x x,$$

where $\sum_x x$ denotes the sum over the population members. This definition of $\mu$ is adequate for understanding the concept of the population mean. An example is the mean age of all people in the country.

More generally, consider an **experiment** that leads to possible values $x_1^*, x_2^*, ...$ with potentially different probabilities $\Pr[X = x_1^*]$, $\Pr[X = x_2^*], ...$ These probabilities necessarily sum to one. Then **the population mean is the probability-weighted average**

$$\mu \equiv \mathrm{E}[X] = \Pr[X = x_1^*] \times x_1^* + \Pr[X = x_2^*] \times x_2^* + \cdots = \sum\nolimits_x \Pr[X = x] \times x,$$

where $\sum_x$ denotes summation over all the possible distinct values that $X$ may take. Appendix 5.A provides additional discussion.

For example, let $X$ be the number of heads obtained in two coin tosses. $X = 0$ if neither toss lands heads, $X = 1$ if either heads then tails occurs or tails then heads occurs, and $X = 2$ if both tosses land heads. For a fair coin with probability of heads equal to 0.5 on each toss, it follows that $X = 0$ (tails and tails) with probability 0.25, $X = 1$ (tails and heads or heads and tails) with probability 0.5, and $X = 2$ (heads and heads) with probability 0.25. In summary we have

$$X = \begin{cases} 0 & \text{with probability } 0.25 \\ 1 & \text{with probability } 0.5 \\ 2 & \text{with probability } 0.25. \end{cases}$$

Then

$$\mu = \sum\nolimits_x \Pr[X = x] \times x = 0.25 \times 0 + 0.5 \times 1 + 0.25 \times 2 = 1.0.$$

The population mean number of heads in two coin tosses equals one.

### 5.2.3 Sample Mean

The **sample mean**, by contrast, is the average of the $n$ sample realizations $x_1, ..., x_n$, or

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

For example, for a finite population that consists of 5 equally likely values 1, 2, 3, 4 and 5, the population mean $\mu = (1 + 2 + 3 + 4 + 5)/5 = 3$. If a sample of size 3 is drawn from this population and consists of the values $x_1 = 2$, $x_2 = 5$ and $x_3 = 4$, then the sample mean

$\bar{x} = (2+5+4)/3 \simeq 3.67$. Note that in general the sample mean will not exactly equal the population mean.

As a second example, return to the two-coin toss example. We found that the population mean number of heads in two coin tosses equaled 1.0, assuming the coin was fair. If performing five consecutive two-coin tosses results in the number of heads equal to, respectively, 1, 0, 0, 2 and 1, then the sample mean $\bar{x} = (1+0+0+2+1)/5 = 0.8$. Again in general the sample mean will not exactly equal the population mean.

The sample $x_1, ..., x_n$ is a realization of the random variables $X_1, X_2, ..., X_n$. It follows that **the sample mean $\bar{x}$ is a realization of the random variable**

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

**Remark 6** *The population mean $\mu$ is the probability-weighted average of all values that the random variable $X$ may take in the population. The sample mean $\bar{x}$ is the average of the observed values in a sample from the population.*

### 5.2.4 Population Variance and Standard Deviation

The **population variance** of $X$, denoted $\text{Var}[X]$ or $\sigma^2$ or $\sigma_X^2$, is **the probability-weighted average of the squared deviations of $X$ from the mean** $\mu$ for all possible values of $X$ in the population. The population variance is also denoted $\text{E}[(X-\mu)^2]$, the **expected value** of the $(X-\mu)^2$ as it is the long-run average value that is expected if we draw a value of $X$ at random, draw a second value of $X$ at random, and so on, and then obtain the average of these values..

For sampling from a finite population of size $N$ where $x$ takes values, the population variance is simply the average of each value of $(x^* - \mu)^2$, so the population variance of $X$ is

$$\sigma^2 \equiv \text{E}[(X-\mu)^2] = \frac{(x_1^* - \mu)^2 + (x_2^* - \mu)^2 + \cdots + (x_N^* - \mu)^2}{N} = \frac{1}{N}\sum\nolimits_x (x-\mu)^2.$$

More generally for an experiment, the population variance is a probability-weighted average, so

$$\begin{aligned}
\sigma^2 &\equiv \text{E}[(X-\mu)^2] \\
&= \Pr[X = x_1^*] \times (x_1^* - \mu)^2 + \Pr[X = x_2^*] \times (x_2^* - \mu)^2 + \cdots \\
&= \sum\nolimits_x \Pr[X = x] \times (x-\mu)^2.
\end{aligned}$$

The **population standard deviation** $\sigma$ is the square root of the variance. The notation $\sigma$, pronounced "sigma", is used to denote the standard deviation as $\sigma$ is the Greek letter for s.

Continuing the earlier example of number of heads in two coin tosses, with $\Pr[X = 0] = 0.25$, $\Pr[X = 1] = 0.5$, $\Pr[X = 2] = 0.25$, and $\mu = 1$,

$$\sigma^2 = 0.25 \times (0-1)^2 + 0.5 \times (1-1)^2 + 0.25 \times (2-1)^2 = 0.5.$$

The population variance is 0.5 and the population standard deviation is $\sqrt{0.5} \simeq 0.707$.

### 5.2.5 Sample Variance and Standard Deviation

The **sample variance** is the average of the squared deviations of $x$ around $\bar{x}$, rather than around $\mu$, since $\mu$ is unknown. From Chapter 2

$$s^2 = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2.$$

The divisor $(n-1)$ is called the **degrees of freedom** because only $(n-1)$ terms in the sum are free to vary since they are linked by the relationship $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Taking the square root of $s^2$ yields the **sample standard deviation** $s$

Like the sample mean, the sample variance is the realization of a random variable, namely

$$S^2 = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (X_i - \bar{X})^2.$$

Similarly, the sample standard deviation $s$ is a realization of the random variable $S$.

**Remark 7** *The population variance $\sigma^2$ is the probability-weighted average of all values that $(X-\mu)^2$ may take in the population. The sample variance $s^2$ is the average (after a degrees of freedom adjustment) of $(x_i - \bar{x})^2$.*

## 5.3 Sample Generated by an Experiment

We consider an example of a sample generated by an experiment. In the particular experiment considered probability theory can be used to determine the population mean an d the population standard deviation.

We next take a series of samples, by running the experiment many times, and for each sample obtain the sample mean $\bar{x}$. We again find that (1) the average of many sample means is close to the population mean, (2) the sample mean is much less variable than the individual underlying observations, and (3) provided the sample is sufficiently large, the sample mean is approximately normally distributed.

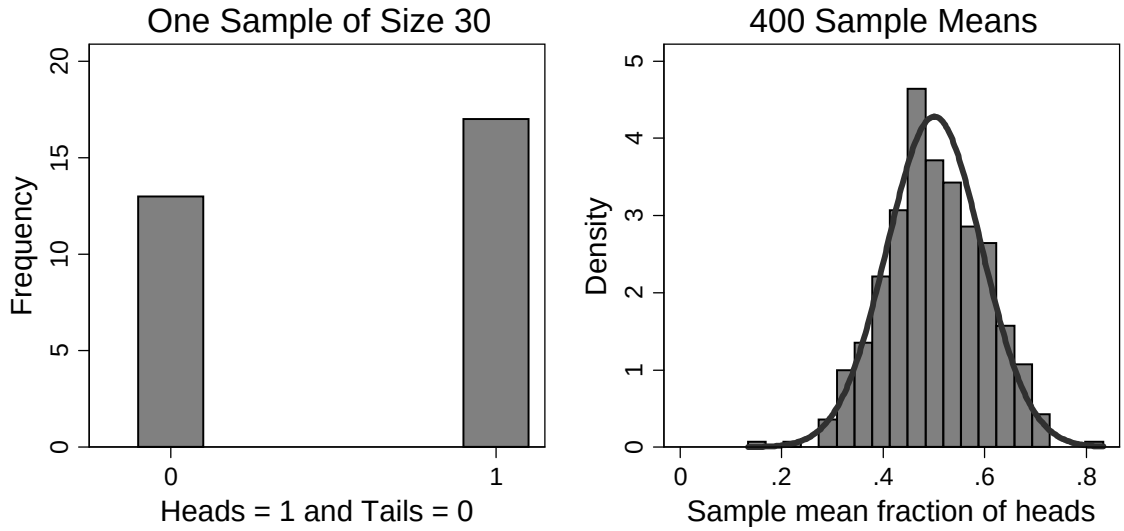


Figure 5.3: Coin tosses: Histogram for one sample (n = 30) and histogram with kernel density estimate for 400 sample means (n = 30).

### 5.3.1 Example: Coin Tosses

We consider the fraction of times that a fair coin lands heads in 30 tosses.

The random variable $X = 1$ if heads and $X = 0$ if tails. Since the coin is fair, the toss leads to heads half the time and tails half the time, so the population mean $\mu = \mathrm{E}[X] = 0.5$ and the population standard deviation $\sigma = 0.5$ since $\sigma^2 = 0.5 \times (0-0.5)^2 + 0.5 \times (1-0.5)^2 = 0.25$.

The left panel of Figure 5.3 shows a histogram for one sample of 30 tosses. In this sample there were 17 heads and 13 tails, so $\bar{x} = 17/30 = 0.567$, and $s = 0.504$.

Now randomly draw 400 samples, each of 30 coin tosses. In this example the first three such samples have means $\bar{x}_1 = .567$, $\bar{x}_2 = .467$ and $\bar{x}_3 = .700$. Dataset COINTOSSMEANS has all 400 sample means, with histogram given in the right panel of Figure 5.3. The histogram is roughly centered on the population mean, with the average of the 400 means equal to .501, close to $\mu = .5$. There is much less variability in these 400 means than in the original population data, with the standard deviation of the 400 means equal to .093, between one-fifth and one-sixth of the population standard deviation of $\sigma = .5$. The right panel of Figure 5.3 also includes the corresponding histogram and kernel density estimate of the 400 means that are roughly those of a normally distributed random variable.

In the preceding example we did not actually toss a coin 12,000 times to obtain the results for 400 samples, each with 30 coin tosses. Instead a computer was used to simulate the coin tosses. The method to do so is explained in Appendix 5.C.

## 5.4 Properties of the Sample Mean

The 1880 Census and coin tosses examples show that the sample mean $\bar{x}$ is on average close to the population mean $\mu$, that the variability of the sample mean $\bar{x}$ is much less than that of the underlying individual observations, and that the sample mean may be approximately normally distributed.

In this section these results are formalized in a general setting.

### 5.4.1 Population Assumptions

The statistical properties of the sample mean depend on the assumptions made about the process generating the random variables $X_1, X_2, ..., X_n$.

We assume

1. $X_i$ has common mean $\mu$
$$\mathrm{E}[X_i] = \mu \text{ for all } i.$$
2. $X_i$ has common variance $\sigma^2$
$$\mathrm{Var}[X_i] = \sigma^2 \text{ for all } i.$$
3. Different observations are statistically independent
$$X_i \text{ is statistically independent of } X_j, i \neq j.$$

Here statistical independence implies, for example, that the value taken by $X_2$ is not influenced by the value taken by $X_1$; see Appendix 5.A for a more formal definition.

Short-hand notation is that in the population

$$X \sim (\mu, \sigma^2),$$

where $\sim$ means "is distributed as", and the terms in parentheses are, respectively, the population mean and the population variance.

The assumptions are met when data are obtained from a **simple random sample**, often called more simply a **random sample**, where $X_1 \sim (\mu, \sigma^2), ...., X_n \sim (\mu, \sigma^2)$ and $X_1, ..., X_n$ are statistically independent of each other.

Assumptions 1 and 2 presume that the mean $\mu$ and standard deviation $\sigma^2$ exist. In theory some random variables may not have a mean and variance, but in practice their existence is a reasonable assumption for typical economics data. Assumption 2 (identical variances) and assumption 3 can d 2 The assumptions of a simple random sample, namely identical variances and statistical independence, can be relaxed. And the reason for using regression models in later chapters is that these relax assumption 1 of identical means.

In addition to these population assumptions it is also assumed that the sample is such that there is variation in the data. Then not all observations in the sample are identical, so the sample standard deviation is greater than zero.

### 5.4.2 Mean of the Sample Mean

The **(population) mean of the sample mean** $\bar{X}$ is

$$\mu_{\bar{X}} \equiv \mathrm{E}[\bar{X}] = \mu,$$

In words, the expected value of the sample mean equals the population mean, so the average of $\bar{X}$ from many samples is expected to equal $\mu$. This result holds under assumption 1; see Appendix 5.B for a derivation.

This result means that if we were able to obtain many random samples and for each sample obtain the sample mean, then on average the sample means equal the population mean.

### 5.4.3 Standard Deviation of the Sample Mean

The variability of $\bar{X}$ around its mean of $\mu$ is measured using the variance and standard deviation of $\bar{X}$.

The **(population) variance of the sample mean** is

$$\sigma^2_{\bar{X}} = \mathrm{Var}[\bar{X}] \equiv \mathrm{E}[(\bar{X} - \mu_{\bar{X}})^2] = \frac{\sigma^2}{n},$$

where $\sigma^2$ is the population variance of $X$. This result holds under assumptions 1-3; see Appendix 5.B for a derivation. The **(population) standard deviation of the sample mean** is then

$$\sigma_{\bar{X}} \equiv \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

The variance $\sigma^2_{\bar{X}}$ measures the variability in $\bar{X}$. The result that $\sigma^2_{\bar{X}} = \sigma^2/n$ implies that the sample mean is less variable than the underlying data, as demonstrated in Figure 5.2 for the 1880 Census example and in Figure 5.3 for the coin toss example.

Furthermore the variability of the sample mean as an estimate of the population mean decreases greatly as the sample size increases, at rate $n$ for the variance and at rate $\sqrt{n}$ for the standard deviation. For the 1880 Census example the standard deviation of the 100 means was 3.92, close to $\sigma/\sqrt{n} = 18.61/\sqrt{25} = 3.72$. And for the coin toss example with the standard deviation of the 400 means was .093, close to $\sigma/\sqrt{n} = .5/\sqrt{30} \simeq .091$.

As expected, **larger samples lead to greater precision** in estimating $\mu$. Furthermore, $\sigma^2_{\bar{X}} = \sigma^2/n \to 0$ as $n \to \infty$, so the sample mean will be very close to $\mu$ as the sample size $n \to \infty$.

**Remark 8** *Under simple random sampling the sample mean $\bar{x}$ is the realization of a random variable $\bar{X}$ that has mean equal to the population mean $\mu$ and standard deviation $\sigma/\sqrt{n}$ that gets smaller as the sample size increases.*

### 5.4.4 Standard Error of the Sample Mean

The variance and standard deviation of $\bar{X}$ depend on the population variance $\sigma^2$ which is unknown. Replacing $\sigma^2$ by its estimate $s^2$, leads to the following estimates.

The **estimated variance** of $\bar{X}$ is

$$s^2_{\bar{X}} = \frac{s^2}{n} = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}{n}.$$

Taking the square root, the **estimated standard deviation** of $\bar{X}$, called the **standard error of the sample mean**, is

$$se(\bar{X}) = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}}{\sqrt{n}},$$

where $s$ is the sample standard deviation, the sample estimate of the standard deviation of $X$.

Note that in general the term "**standard error**" means **estimated standard deviation**. The various estimators considered in this book each have a distinct standard error. In many situations computer output will include a reported "standard error", but this is not necessarily the standard error of the sample mean $\bar{X}$.

**Remark 9** *Under simple random sampling the standard error of the mean, the estimated standard deviation of the sample mean, equals $s/\sqrt{n}$ where $s$ is the sample standard deviation.*

### 5.4.5 Normal Distribution and the Central Limit Theorem

From the right panels of Figures 5.2 and 5.3, the sample means appear to be approximately normally distributed, even though each observation is clearly not from the normal distribution. Remarkably this is the case in quite general settings, provided the sample size is sufficiently large.

The preceding results imply that the random variable

$$\bar{X} \sim (\mu, \sigma^2/n).$$

In general subtracting the mean and dividing by the standard deviation leads to a random variable with mean 0 and variance 1; see Appendix 5.A. Here we denote this standardized random variable by $Z$, so

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim (0, 1).$$

Furthermore, if the sample is a simple random sample and the sample size $n \to \infty$, then a result from statistics called the **central limit theorem**, states that $Z$ has the **standard normal distribution**, so then

$$Z \sim N(0,1).$$

Often $n > 30$ is sufficient for this result to be a good approximation. It follows that $\bar{X}$ is then normally distributed (with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$).

The remarkable power and usefulness of the central limit theorem cannot be understated. Regardless of the distribution of the underlying random variable $X$, if assumptions 1-3 hold then averaging leads to a standardized random variable that is standard normally distributed. It can also be extended to cases where not all of assumptions 1-3 hold; see Appendix 5.B.

**Remark 10** *Under simple random sampling the standardized random variable* $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ *is standard normally distributed as the sample size goes to infinity.*

## 5.5 t Statistic

Using the sample mean $\bar{x}$ to perform statistical inference on the population mean $\mu$ requires knowledge of the distribution of the sample mean. Thanks to the central limit theorem we know that under assumptions 1-3 the sample mean is normally distributed.

However, to immediately apply the central limit theorem result requires knowledge of the population standard deviation $\sigma$. In practice this is not known, so we instead estimate it by the sample standard deviation $s$. Since this is an estimate it adds noise. As a result we obtain a statistic that has a distribution that has fatter tails than the standard normal. As explained next, this distribution is the $t$ distribution.

### 5.5.1 T Distribution

The sample standard deviation $s$ is used instead of the unknown population standard deviation $\sigma$. Then the **distribution of the sample mean** $\bar{X}$ is defined in terms of the transformed random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

where $s$ is a realization of the random variable $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

The distribution of the random variable $T$ is in general complicated. Approximately

$$T \sim T(n-1),$$

where $T(n-1)$ denotes the $t$ **distribution** with $(n-1)$ degrees of freedom .

Under Assumptions 1-3 the $t$ distribution approximation can be shown to be **exact** in two distinct situations. First, if the underlying random variable $X$ is itself normally distributed, regardless of the sample size. Second, if the sample size $n \to \infty$, then $s$ is a very good estimate of $\sigma$ and we expect that $T$ is standard normal distributed by the original central limit theorem result. This is indeed the case as the $T(\infty)$ distribution is the same as the standard normal.

The $t$ distribution is similar to the standard normal distribution and discussed in some detail in Chapter 6.2. An added complication is that there is a range of $t$ distributions, depending on the value taken by the degrees of freedom parameter. As the degrees of freedom parameter goes to infinity the $t$ distribution collapses to the standard normal.

### 5.5.2 The Sample t-Statistic

We observe a single sample with sample mean $\bar{x}$, sample standard deviation $s$, sample standard error $se(\bar{x}) = s/\sqrt{n}$, and corresponding sample value of the $t$-statistic. So the sample $t$ statistic is a single realization of a $T(n-1)$ distributed random variable. For simplicity we write

$$t = \frac{\bar{x} - \mu}{se(\bar{x})} \sim T(n-1).$$

More formally $t$ is a realization of the random variable $T = (\bar{X} - \mu)/(S/\sqrt{n})$ that is viewed as $T(n-1)$ distributed.

As already noted, this result is exact if the sample data are normally distributed and is a good approximation for nonnormal data as $n \to \infty$. A common rule-of-thumb is that the approximation will be a good one if $n > 30$.

A less common alternative uses the standard normal distribution for $t$. This has the advantage of simplicity. It coincides with using the $T(n-1)$ distribution for very large $n$. For small $n$, however, using the standard normal distribution leads to confidence intervals that may be too narrow and to tests that may over-reject; see Chapter 6.

**Remark 11** *Under simple random sampling the t-statistic $t = (\bar{x} - \mu)/(s/\sqrt{n})$ is the realization of a randomly variable that is $T(n-1)$ distributed, exactly if data are normally distributed and approximately for nonnormal data if $n$ is sufficiently large.*

### 5.5.3 Summary for the Sample Mean

The distinction between variability in $X_i$, the random variable leading to the $i^{th}$ sample value $x_i$, and the variability in $\bar{X}$, the random variable leading to the sample mean $\bar{x}$, can cause confusion. A summary for a simple random sample is the following:

1. Individual $X_i$ are assumed to have common mean $\mu$ and variance $\sigma^2$.

2. The average $\bar{X}$ of the $n$ draws of $X_i$ has mean $\mu$ and variance $\sigma^2/n$.

3. The standardized statistic $Z = (\bar{X} - \mu)/\sigma/\sqrt{n} \sim (0, 1)$ has mean 0 and variance 1.

4. As sample size $n \to \infty$, $Z$ is standard normal distributed by the central limit theorem.

5. Replacing the unknown $\sigma$ by the sample standard deviation leads to (approximately) a $t$ distribution with $(n-1)$ degrees of freedom, denoted $T(n-1)$.

6. The sample $t$-statistic $t = (\bar{x} - \mu)/(s/\sqrt{n})$ is a realization of a random variable that is approximately $T(n-1)$ distributed.

7. For normally distributed data this last result is exact. For nonnormal data the approximation improves as $n$ gets larger. Often $n > 30$ is sufficient.

## 5.6 Point Estimation

The first goal of inference on the population mean $\mu$ is to have a single estimate of $\mu$, called a **point estimate**.

What makes a good estimate of $\mu$? Ideally its distribution should be centered on $\mu$ and have as little variability around $\mu$ as possible.

In this more technical section we explain why, for data from a simple random sample, the sample mean is often a good estimate of $\mu$.

### 5.6.1 Unbiased and Consistent Estimators

The two commonly-used criteria for being "centered on $\mu$" are unbiasedness and consistency.

**Remark 12** *An **unbiased estimator** of a population parameter is a statistic whose expected value equals the population parameter.*

Since $\mathrm{E}[\bar{X}] = \mu$, the sample mean has the attraction of being an unbiased estimator for the population mean. It can also be shown that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ has the property that $\mathrm{E}[S^2] = \sigma^2$, so the sample variance is unbiased for the population variance. For this reason the formula for $s^2$ divides by $n-1$ rather than the more obvious $n$.

**Remark 13** *A **consistent estimator** of a population parameter is a statistic that is almost certainly arbitrarily close to the population parameter, as the sample size gets very large.*

More precisely, **consistency** of $\bar{X}$ for $\mu$ means that the probability that $\bar{X}$ is more than a small amount $\varepsilon$ away from $\mu$ goes to zero as the sample size gets large, no matter how small the value of the deviation $\varepsilon$. That is, the probability that the difference $|\bar{X} - \mu|$ exceeds $\varepsilon$ goes to zero as $n \to \infty$ for any choice of $\varepsilon > 0$.

The sample mean $\bar{X}$ is consistent for $\mu$ under simple random sampling (assumptions 1-3) as it is unbiased and has variance $\sigma^2/n$ which goes to zero as $n \to \infty$. This convergence of $\bar{X}$ to $\mu$ as the sample size gets large is an example of a **law of large numbers**.

Unbiasedness is easier to understand, but it is consistency that is most important if the sample size is large. The two concepts are related. An unbiased estimator that additionally has variance that goes to zero as the sample size gets large is consistent. A biased estimator can also be consistent, however, provided that both its bias and its variance go to zero as the sample gets large.

**Remark 14** *Under simple random sampling the sample mean is unbiased for $\mu$, meaning that in repeated samples it will on average equal $\mu$. And the sample mean is consistent for $\mu$, meaning loosely that as the sample size gets large the probability that $\bar{X}$ differs from $\mu$ approaches zero.*

### 5.6.2 Best Estimators

The sample mean is unbiased for $\mu$ and is consistent for $\mu$.

Other potential estimators of $\mu$ can also have these properties. For example, the sample median can be shown to also be unbiased and consistent for $\mu$ if $X$ is symmetrically distributed. In that case we discriminate between such estimators on the basis of the size of their variance.

**Remark 15** *A **best estimator** or **efficient estimator** in a class of estimators has **minimum variance** among the class.*

The class considered is usually either the class of consistent estimators or the class of unbiased estimators.

Smaller variance is desired as then there will be less variability in the estimator from sample to sample. As an example of a poor choice for an unbiased estimator, suppose we just used the first observation in each sample of size $n$ to estimate $\mu$. Then this estimator is unbiased from sample to sample as $\mathrm{E}[X_1] = \mu$. But it has variance $\sigma^2$ which is high relative to other possible unbiased estimators.

For simple random samples the sample mean $\bar{X}$ has variance $\sigma^2/n$. Whether alternative unbiased estimators for $\mu$ can better this depends on the distribution for $X$. For some common distributions of $X$, notably the normal, Bernoulli, binomial and Poisson, it can be shown that, given data from a simple random sample, no other unbiased estimator or consistent estimator for $\mu$ has smaller variance.

The sample mean is generally used, for simplicity and because even in situations where the sample mean is not the most efficient estimator, its variance is usually not much greater than the minimum possible variance, so the efficiency loss in using the sample mean is not great.

## 5.7 Nonrandom Samples

The standard assumption is that data are generated from a simple random sample. This requires that the observations are statistically independent. And it requires that they are realizations of a random variable with common mean and common variance, i.e. $X_i \sim (\mu, \sigma^2)$. In practice these assumptions need not hold. Then in some cases standard results can be modified to yield valid estimates and statistical inference, while in other cases estimates may be biased and inconsistent.

This issue is particularly relevant for samples based on a survey. It has become relatively inexpensive to conduct a survey by means such as telephone or the internet. Due to nonrepresentativeness of the grouped surveyed, or high nonresponse rates even if the group surveyed is representative, the sample may be a very skewed sample. If a sample reveals a surprising result, it may be an artifact of being nonrepresentative.

### 5.7.1 Representative Sample but Not Statistically Independent

Even if a sample is representative, the observations may not be statistically independent. In this case assumption 3 in Chapter 5.4 does not hold.

For example, for monthly data on price inflation it is very likely that inflation rates in successive months are related. Then $\bar{X}$ still has mean $\mu$, provided each $X_i$ has mean $\mu$. But the variance of $\bar{X}$ is no longer $\sigma^2/n$; it will be more than this. Similar complications arise with clustered sampling, where observations in the same cluster are correlated.

In such cases alternative methods have been developed to estimate the standard deviation of $\bar{X}$. Denote the alternative estimate by $se(\bar{x})$. Then we base inference on

$$t = \frac{\bar{x} - \mu}{se(\bar{x})},$$

where now $se(\bar{x}) \neq s/\sqrt{n}$. The correct $se(\bar{x})$ is most easily obtained by OLS regression on just an intercept and using appropriate robust standard errors; see Chapter 10.7.

### 5.7.2 Nonrepresentative Samples

A more serious complication arises if the sample is not representative of the population, so that assumption 1 in Chapter 5.4 does not hold. Then $\bar{X}$ may be biased and inconsistent for $\mu$.

For example, a survey of readers of Golf Digest will provide an inconsistent estimate of the golfing habits of all Americans, since it oversamples active golfers. (The survey might, however, provide a consistent estimate of the golfing habits of all readers of Golf Digest. This would be of use to the advertising department of Golf Digest. In this latter case the population is viewed as readers of Golf Digest rather than all Americans.)

A famous example of a nonrepresentative sample is the incorrect prediction of the winner of the 1948 U.S. presidential election. Opinion polls predicted that the Republican candidate John Dewey would defeat the incumbent Democrat, President Harry Truman. Yet Truman won convincingly. The opinion polls were not representative for two reasons. First, the last opinion polls were taken well before the election, so a late surge to Truman meant that they were not representative of opinions on election day itself. Second, the opinion polls were not based on random sampling - the interviewers were given too much discretion as to who they interviewed.

### 5.7.3 Weighted Mean

Samples obtained from government surveys and from political polling surveys are often not representative of the population. Yet the leading national surveys can nonetheless be adjusted to provide valid estimates of the population mean.

For example, the unemployment rate in the United States is obtained from the Current Population Survey (CPS), a monthly survey of 60,000 households. This survey is not a simple random sample. Households in smaller states are oversampled to provide more reliable state-level data. Similarly, minority and disadvantaged populations are oversampled. And the surveyed households are clustered geographically to reduce interview costs.

To overcome these complications, surveys such as the CPS, provide sampling weights that make possible unbiased estimation of the population mean using a **weighted mean**

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

where the **sample weights** $w_i$ are the reciprocal of the **probability** that the $i^{th}$ individual in the population is included in the sample.

Most standard statistical software enables computation of the weighted mean and the standard error of the weighted mean, provided the sample weights are known. Statistical software for **survey data** allows for additional complications of surveys, briefly discussed in Chapter 19.5.

## 5.8 Key Concepts

1. Random variables $X$ are denoted in upper case and realized values $x$ are denoted in lower case.

2. A sample of size $n$ has observed values $x_1, x_2, ..., x_n$ that are realizations of the random variables $X_1, X_2, ..., X_n$.

3. The population mean $\mu$ is the probability-weighted average of all values that the random variable $X$ may take in the population.

4. The population variance $\sigma^2$ is the probability-weighted average of all values that $(X - \mu)^2$ may take in the population.

5. The population standard deviation is $\sigma$.

6. Statistical inference seeks to infer properties of the population from the sample at hand.

7. The sample statistics, such as the sample mean, are random variables whose statistical properties are determined by those of the random variables whose realizations produced the sample.

8. In particular, the sample mean $\bar{x}$ is a realization of the random variable $\bar{X}$.

9. A simple random sample is one whose observations are independent draws from the same distribution with $X_i \sim (\mu, \sigma^2)$. Then assumptions 1-3 are satisfied.

10. Under simple random sampling the sample mean $\bar{x}$ is the realization of a random variable $\bar{X}$ that has mean equal to the population mean $\mu$ and standard deviation $\sigma/\sqrt{n}$ that gets smaller as the sample size increases.

11. The estimated standard deviation of $\bar{X}$, called the standard error of $\bar{X}$, is denoted $se(\bar{x})$.

12. Under simple random sampling, $se(\bar{x}) = s/\sqrt{n}$.

13. For statistical inference on $\mu$ we use the $t$-statistic $t = (\bar{x} - \mu)/se(\bar{x})$.

14. Under simple random sampling the t-statistic $t$ is the realization of a randomly variable that is $T(n-1)$ distributed, exactly if data are normally distributed and approximately for nonnormal data if $n$ is sufficiently large.

15. In most cases for $n > 30$ it is reasonable to use the $T(n-1)$ distribution.

16. An unbiased estimator of a population parameter is a statistic whose expected value equals the population parameter.

17. A consistent estimator of a population parameter is a statistic that is almost certainly arbitrarily close to the population parameter, as the sample size gets very large.

18. A best estimator or efficient estimator has minimum variance among the class of consistent estimators (or the class of unbiased estimators).

19. Given a simple random sample, the sample mean is unbiased and consistent. Furthermore it is the best estimator in the special cases that the distribution of $X$ is normal, Bernoulli, binomial or Poisson.

20. Adjustment to methods may be needed if the sample is not a simple random sample.

21. Key Terms: population; population mean; population variance; population standard deviation; sample; simple random sample; sample mean; sample standard deviation; standard error of the sample mean; central limit theorem; standard normal distribution; $t$ statistic; $t$ distribution; degrees of freedom; normal distribution; parameter; estimator; unbiased; consistent; best estimator; minimum variance; nonrandom samples; weighted mean; random number generator.

## 5.9 Exercises

1. Let $X$ denote annual health costs for an individual and suppose $X = 1000$ with probability 0.8 and $X = 5000$ with probability 0.2.

   (a) Obtain $\mu = \mathrm{E}[X]$ from first principles.

   (b) Obtain $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ from first principles.

   (c) Hence find the standard deviation of $X$.

2. Let $\bar{X}$ be the mean of a random sample of size 100 from a random variable that is distributed with mean 200, variance 400, and a distribution that is not the normal distribution.

   (a) Give the mean of $\bar{X}$.

   (b) Give the variance and standard deviation of $\bar{X}$.

   (c) Is $\bar{X}$ likely to be normally distributed? Explain.

3. Use a computer and a random number generator to obtain 1000 random numbers between 0 and 1, setting the seed to 10101. These are generated in such a way that they can be viewed as independent draws of the uniform random variable $X$ with mean $\mu = 0.5$ and variance $\sigma^2 = 1/12$.

   (a) Are the sample mean and sample variance approximately equal to $\mu$ and $\sigma^2$?

   (b) How many of the 1,000 random numbers do you expect to lie between 0.0 and 0.1, and between 0.1 and 0.2, etc? Hint: Any value between 0 and 1 is equally likely.

   (c) Plot a histogram of the random numbers drawn, with starting value 0, 10 bins, and frequency on the vertical axis. Do you (approximately) get what you expected from part b.

   (d) Give a scatter plot of the random numbers against the observation number. Do they appear to be randomly draws between 0 and 1?

   (e) Give a line plot of the random number against observation number for the first 50 observations. Do consecutive random numbers appear to be related to each other, or do they appear to be independent?

4. For random sampling from $X \sim (\mu, \sigma^2)$ state which of the following statements are true

   (a) $\bar{X} = \mu$.

   (b) $\bar{X}$ has population mean $\mu$.

   (c) $\bar{X}$ has population variance $\sigma^2$.

5. Consider random sampling from $X \sim (\mu, \sigma^2)$. State what happens to the size of $\mathrm{E}[\bar{X}]$, $\mathrm{Var}[\bar{X}]$ and the standard deviation of $\bar{X}$ when the sample size is made four times as large.

6. Let $X = 1$ with $\Pr[X = 1] = 1/6$ and $X = 0$ with $\Pr[X = 0] = 5/6$. (One way this would arise is if we tossed a six-sided die and set $X = 1$ if a five, say is obtained, and let $X = 0$ otherwise.)

   (a) Obtain $\mu = \mathrm{E}[X]$ from first principles.

   (b) Obtain $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ from first principles.

   (c) Now use a computer and a random number generator to obtain a sample of size 100 for this example. Hint: A random number is less than 1/6 with probability 1/6.

   (d) Compare the mean $\bar{x}$ and variance $s^2$ of this sample to your answers in parts a-b.

(e) Obtain the histogram. Does $X$ appear to be normally distributed?

7. The preceding computer experiment was run 400 times, yielding 400 samples of size 100. The sample mean $\bar{x}$ for each sample is given in dataset DIETOSS.

   (a) Obtain the descriptive statistics for the 400 values of $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

   (b) Obtain the histogram (or better still the kernel density estimate). Is this what you expect? Explain.

8. Suppose $X$ takes value 1 with probability 0.4, value 2 with probability 0.2, and value 3 with probability 0.4.

   (a) Obtain $\mu = [X]$ from first principles.

   (b) Obtain $\sigma^2 = \mathrm{E}[(X - \mu)^2]$ from first principles.

   (c) Now use a computer and a random number generator to obtain a sample of size 1,000 for this example. Hint: Let $u$ be the random number. Then $x = 1$ if $u < 0.4$, $x = 2$ if $0.4 \leq u < 0.6$, and $x = 3$ if $u \geq 0.8$.

   (d) Compare the mean and standard deviation of this sample to your answers in parts a-b.

9. The preceding computer experiment was run 400 times, obtaining 400 samples of size 100. The sample mean $\bar{x}$ for each sample is given in dataset ONETWOTHREE.

   (a) Obtain the descriptive statistics for $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

   (b) Obtain the histogram (or better still the kernel density estimate). Is this what you expect? Explain.

10. The dataset TDIST4 has the sample means $\bar{x}$ and corresponding standard standard deviations $s$ from 1000 random samples of size 4 where $X \sim N(100, 16^2)$.

    (a) Obtain the descriptive statistics for $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

    (b) Obtain the descriptive statistics for $se(\bar{x}) = s/\sqrt{n}$ for these data. Is the mean what you expect? Explain.

    (c) Compute $z = (\bar{x} - 100)/16$. Explain why $z$ is standard normal distributed. For these data does $z$ appear to be normally distributed?

(d) Compute $t = (\bar{x} - 100)/se(\bar{x})$. Explain why $t$ is $t_3$ distributed. For these data does $t$ appear to be $t_3$ distributed? [The $t_3$ distribution has mean 0 and variance 3 and for example, $\Pr[T_3 > 1.638] = 0.10$.]

11. The dataset TDIST25 has the sample means $\bar{x}$ and and corresponding standard standard deviations $s$ from 1000 random samples of size 25 where $X \sim N(200, 50^2)$.

    (a) Obtain the descriptive statistics for $\bar{x}$. Are the mean and standard deviation what you expect? Explain.

    (b) Obtain the descriptive statistics for $se(\bar{x})$ for these data. Is the mean what you expect? Explain.

    (c) Compute $z = (\bar{x} - 200)/50$. Explain why $z$ is standard normal distributed. For these data does $z$ appear to be normally distributed?

    (d) Compute $t = (\bar{x} - 200)/se(\bar{x})$. Explain why $t$ is $T(24)$ distributed. For these data does $t$ appear to be $T(24)$ distributed? [The $T(24)$ distribution has mean 0 and variance 24/22 and, for example, $\Pr[T_{24} > 1.318] = 0.10$.]

12. Each of the following questions requires computer software that enables creation of many samples and saving key statistics from each sample.

    (a) Create 1000 random samples of size 4 where $X \sim N(100, 16^2)$ and save the sample means $\bar{x}$ and corresponding standard errors $se(\bar{x})$ from each sample. Use these results, rather than dataset TDIST4, to answer question 10.

    (b) Create 1000 random samples of size 4 where $X \sim N(100, 16^2)$ and save the sample means $\bar{x}$ and corresponding standard errors $se(\bar{x})$ from each sample. Use these results, rather than dataset TDIST25, to answer question 11.

13. An insurance company offers insurance to 10,000 people with independent loss distributions that have mean \$5,000 and standard deviation \$20,000. Let $\bar{X} = \frac{1}{10000}\sum_{i=1}^{10000}$ denote the average loss per individual.

    (a) Find the mean and standard deviation of $\bar{X}$.

    (b) Suppose the insurance company sells insurance that provides complete coverage for \$5,400, and for simplicity suppose that the insurance company has no costs aside from paying out any insurance claims. Is the insurance company likely to make a loss? Explain your answer. Hint: By the central limit theorem $\bar{X}$ is normally distributed.

14. Suppose we take a simple random sample from $X \sim (\mu, \sigma^2)$.

(a) We estimate $\mu$ by $X_1$, the first value of $X$ in the sample. Is this estimate unbiased for $\mu$? Is this estimator consistent for $\mu$? Explain.

(b) We estimate $\mu$ by $\bar{X} + \frac{1}{n}$. Is this estimate unbiased for $\mu$? Is this estimator consistent for $\mu$? Explain.

(c) Now suppose $X \sim N(\mu, \sigma^2)$ and we estimate $\mu$ by an estimator $\widetilde{\mu}$ that has $\mathrm{E}[\widetilde{\mu}] = \mu$ and $\mathrm{Var}[\widetilde{\mu}] = 2\sigma^2/n$. Is this estimator a best unbiased estimator for $\mu$? Explain.

15. State whether the following samples are likely to be representative or nonrepresentative of the population.

(a) Every twentieth person is sampled. All people respond.

(b) Every twentieth person is sampled. But only ten percent of those sampled respond.

(c) Every person is sampled. Only ten percent of those sampled respond. We question every twentieth person who did respond.

16. The dataset AUSREGWEALTH has data on average net wealth of households ($x_i$) in 517 regions in Australia in 2003-04. Calculate the weighted mean where weight by household size as follows. This is an example of weighting by frequency weights.

(a) Let $n_i$ equal number of households in each region. Compute $\sum_{i=1}^{n} n_i$.

(b) Generate the variable $n_i x_i$ and hence the weighted mean $\bar{x}_w = \sum_{i=1}^{n} n_i x_i / \sum_{i=1}^{n} n_i$.

(c) Compare the weighted mean to the unweighted mean.

(d) Calculate the weighted standard deviation as $\sum_{i=1}^{n} n_i (x_i - \bar{x}_w)^2 / \sum_{i=1}^{n} n_i$.

(e) Compare the weighted standard deviation to the unweighted standard deviation.

(f) If your software computes weighted means and standard deviation, reproduce these results using your software.

## 5.10 Appendix 5.A: Some Essentials of Probability

This appendix presents further details on the material presented in the main text of this chapter.

A **random variable** is a variable whose value is determined by the outcome of an experiment, where an **experiment** is an operation whose outcome cannot be predicted with certainty. Standard notation is to denote the random variable in upper case, say $X$ (or $Y$ or $Z$), and to denote the realized value of the random variable in lower case, say $x$ (or $y$ or $z$).

The **probability distribution** of a random variable $X$ describes the random behavior of $X$.

### 5.10.1 Discrete Random Variables

A **discrete random variable** is a random variable that can only take a finite number of values (or a countably infinite number of variables such as 0, 1, 2, 3, ....). As an example, $X$ may measure whether or not a person is currently employed, so $X$ may take values 1 (employed) or 0 (not employed). As a second example, $X$ may be the number of consultations with a doctor over the past year; then $X$ may take the values 0, 1, 2, ..... .

In general a discrete random variable takes values $x_1$, $x_2$, ... The **probability mass function** gives the probabilities for each value taken by the random variable:

$$\Pr[X = x], \quad x = x_1, x_2, ...$$

Probabilities lie between 0 and 1 and sum to one over all possible values of $X$, so

$$\sum\nolimits_x \Pr[X = x] = 1,$$

where $\sum_x$ denotes summation over all possible values taken by $X$.

The **cumulative probability distribution function** gives the probability that the random variable $X$ is less than or equal to a particular value:

$$\Pr[X \leq x] \quad x = x_1, x_2, ...$$

The probability that $X$ lies in a given range can be calculated using either the probability mass function or the cumulative distribution function. We have

$$\begin{aligned} \Pr[a < X \leq b] &= \Pr[X = a] + \cdots + \Pr[X = b] \\ &= \Pr[X \leq b] - \Pr[X \leq a]. \end{aligned}$$

The **expected value** of a function $g(X)$ of the random variable $X$ is the long-run average value that we expect if we draw a value $x_1$ of $X$ at random and compute $g(x_1)$, draw a second value and so on, and then obtain the average of these values. Equivalently, for each value that $x$ might take, compute $g(x)$ and then calculate the probability-weighted average of these values by weighting this value by the probability of that value $x$ occurring. Then the **expected value** of $g(X)$

$$\begin{aligned} \mathrm{E}[g(X)] &= g(x_1) \times \Pr[X = x_1] + g(x_2) \times \Pr[X = x_2] + \cdots \\ &= \sum\nolimits_x g(x) \times \Pr[X = x]. \end{aligned}$$

The two most commonly-used expected values are the **population mean** $\mu \equiv \mathrm{E}[X]$ that sets $g(X) = X$ and the **population variance** $\sigma^2 \equiv \mathrm{E}[(X - \mu)^2]$ that sets $g(X) = (X - \mu)^2$. Additionally the **population standard deviation** $\sigma$ is the square root of the variance. These have been presented in some detail in Chapter 5.2.

The discrete probability distributions analyzed in introductory probability courses are the **Bernoulli**, **binomial** and **Poisson** distributions. Basic analysis of economics data uses only the first of these, which is presented next.

### 5.10.2 Bernoulli Distribution

The **Bernoulli distribution** is the term used to describe the distribution of a random variable that takes just one of two values: 0 or 1. This is the simplest example of a discrete random variable.

Denote the probability that $X = 1$ by $p$. For example, $p = 0.5$ in the case of a coin toss if the the coin is fair. Then it must be the case that $X = 0$ with probability $1 - p$, since the probabilities over all possible outcomes must sum to one, and $p + (1 - p) = 1$. So

$$\Pr[X = x] = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0. \end{cases}$$

The quantity $p$ that determines the probability distribution is called a **parameter**. It is unknown, but can be estimated given a sample.

A Bernoulli random has mean $\mu = p$ and variance $\sigma^2 = p(1 - p)$.

These properties can be obtained using the following algebra. For the mean, $\mathrm{E}[X] = 0 \times \Pr[X = 0] + 1 \times \Pr[X = 1] = 0 \times (1 - p) + 1 \times p = p$. And the variance $\sigma^2 = \mathrm{E}[(X - \mu)^2] = (0 - p)^2 \times \Pr[X = 0] + (1 - p)^2 \times \Pr[X = 1] = p^2 \times (1 - p) + (1 - p)^2 \times p = p(1 - p)\{p + (1 - p)\} = p(1 - p)$.

### 5.10.3 Linear Transformation of a Random Variable

If we **add a fixed amount to a random variable** then the mean is changed by the amount $a$ and the variance is unchanged.

If we **multiply a random variable by a fixed multiple** then the mean is multiplied by $b$and the variance is multiplied by $b^2$ (hence the standard deviation is multiplied by $b$).

It follows that if $X$ has mean $\mu$ and variance $\sigma^2$ then $X - \mu$ has mean $\mathrm{E}[X] - \mu = \mu - \mu = 0$ and variance $\sigma^2$. Subsequent division by $\sigma$ leads to mean $0/\sigma = 0$ and variance $\sigma^2/\sigma^2 = 1$. Thus the random variable $Y = (X - \mu)/\sigma$ is a **standardized random variable** with mean zero and variance 1. So if we subtract the mean and divide by the standard deviation we transform the random variable $X$ to the new random variable $Y$ that necessarily has mean 0 and variance 1.

More generally, if $X$ has mean $\mu$ and variance $\sigma^2$ then the random variable $Y = a + bX$, a **linear transformation** of $X$, has mean $\mathrm{E}[Y] = a + b\mathrm{E}[X]$ and variance $\mathrm{Var}[Y] = b^2\mathrm{Var}[X]$.

### 5.10.4 Statistical Independence

The two random variables $X$ and $Y$ are **statistically independent** or, more simply, **independent** if the value taken by $X$ is unrelated to the value taken by $Y$.

For example, let the two random variables $X$ and $Y$ represent the outcomes from two consecutive tosses of the coin. Then the two random variables are statistically independent

if the result of the first toss, say a head, has no bearing on the result of the second toss. If the probability of heads on the first toss is $p$ then, regardless of whether the first toss results in heads or tails, the probability of heads on the second toss is still $p$. (This is the case even if $p = 0.4$, say, so that the coin is not fair).

Random variables are not necessarily statistically independent. But under simple random sampling $X_1, ..., X_n$ are statistically independent.

### 5.10.5 Sums of Independent Random Variables

The mean of a **weighted sum of random variables** equals the weighted sum of their means. That is,

$$\mathrm{E}[aX + bY] = a \times \mathrm{E}[X] + b \times \mathrm{E}[Y].$$

The variance of a sum of random variables is more complicated, as it depends on the statistical relationship between $X$ and $Y$. Simplification occurs if the random variables are statistically independent. Then

$$\mathrm{Var}[aX + bY] = \mathrm{Var}[aX] + \ \mathrm{Var}[bY] = a^2 \times \mathrm{Var}[X] + b^2 \times \mathrm{Var}[Y].$$

This is a weighted sum of the individual variance, with weights that are the square of the original weights.

Applying this result, $X + Y$ has variance that is the sum of the variance of $X$ and the variance of $Y$. Similarly, the difference $X - Y$ has variance that is the sum of the variance of $X$ and the variance of $Y$.

These results are used in the subsequent Appendix 5.B to obtain the mean and variance of the sample mean.

### 5.10.6 Continuous Random Variables

Not all random variables take just discrete values.

A **continuous random variable** $X$ can take an uncountably infinite number of values, such as any real value, or any positive real value, or any real value between zero and one. As an example, $X$ may be annual income of an individual. Or $X$ may be the length of time that the individual has been employed at their current job.

Since $X$ can take any value the probability that it equals any particular value is infinitesimally small. So it is meaningless to consider the probability of $X$ taking a particular value. Instead we evaluate the probability that $X$ lies in a range of values.

A **continuous probability distribution** is defined by the **probability density function** $f(x)$. This function has the property that the probability that $X$ lies between two values, say $a$ and $b$, is given by the area under the function $f(x)$ between $a$ and $b$. The

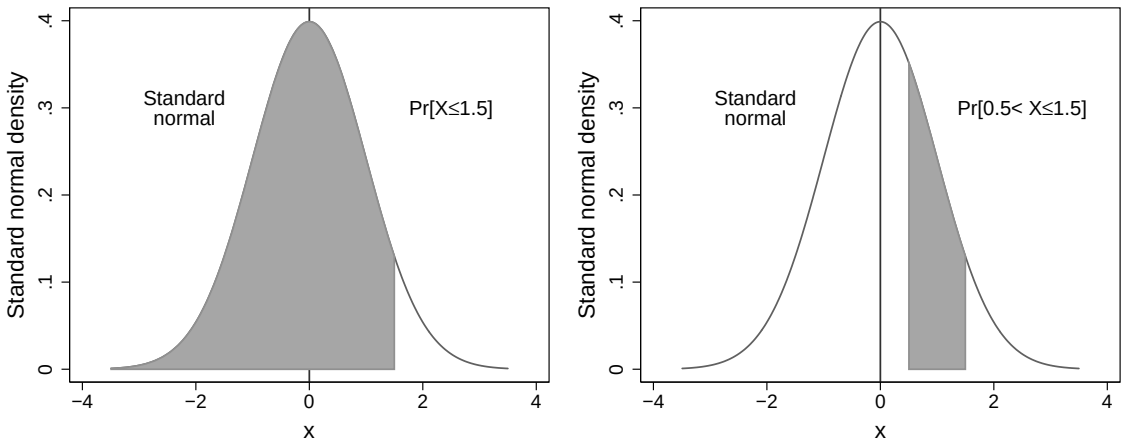


Figure 5.4: $\Pr[X \leq 1.5]$ and $\Pr[0.5 \leq X \leq 1.5]$ for standard normal distribution.

total area under the curve, from the minimum to maximum value of $X$, equals one since probabilities sum to one.

The associated **cumulative probability distribution function** $\Pr[X \leq x]$, is given by the area under the curve from the lowest value that $X$ can take to $x$.

### 5.10.7 Standard Normal Distribution

The leading example of a continuous random variable is a standard normal random variable. The **standard normal distribution** is defined by its probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right), \quad -\infty < x < \infty.$$

The standard normal can be shown to have mean $\mu = 0$ and standard deviation $\sigma = 1$. The notation $N(0, 1)$ is used to denote the standard normal distribution. The standard normal density is the curve given in Figure 5.4.

The shaded region in the left panel of Figure 5.4 gives $\Pr[X \leq 1.5]$, since this is the area under the curve from $-\infty$ to 1.5. The total area under the curve is necessarily 1, and it appears visually that $\Pr[X \leq 1.5] \simeq 0.9$ since the shaded region is about 90% of the total area under the curve. From standard normal tables or a computer in fact $\Pr[X \leq 1.5] = 0.9332$.

The right panel of Figure 5.4 gives $\Pr[0.5 < X \leq 1.5]$. This appears to be approximately equal to 0.2. From standard normal tables or a computer in fact $\Pr[0.5 < X \leq 1.5] = 0.2317$.

For those familiar with **integral calculus** the area under the curve is obtained by taking the integral. Thus, for example, $\Pr[X \leq 1.5] = \int_{-\infty}^{1.5} f(x)dx$ where $f(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$.

For the standard normal this integral has no exact solution. Instead one uses numerical approximations that are given in statistical tables or are calculated by the computer.

### 5.10.8 Other Continuous Distributions

The normal distribution is a generalization of the standard normal distribution that allows for a nonzero mean and a standard deviation other than one. In the more general case the **normal distribution,** with mean $\mu$ and standard deviation $\sigma$, has probability density function $f(x) = (1/\sqrt{2\pi\sigma^2})\exp(-(x-\mu)^2/2\sigma^2)$ and is denoted $N(\mu, \sigma^2)$.

A powerful property of the normal, shared by few other distributions, is that linear combinations of normally distributed random variables are also normally distributed. If $X \sim N(\mu, \sigma^2)$ then $Y = a + bX \sim N(a + b\mu, b^2\sigma^2)$. One consequence of this result is that if $X \sim (\mu, \sigma^2)$ then $Y = (X - \mu)/\sigma \sim N(0, 1)$. In words, for a normally distributed random variable, subtracting the mean and then dividing by the standard deviation leads to a random variable that is standard normal distributed.

The continuous probability distributions most often used in econometrics are the normal, $t$, $F$ and chi-squared distributions. The $t$ **distribution** is discussed in some detail in Chapter 6.2. The $F$ **distribution** and the **chi-squared distribution** are introduced in Chapter 14.5.

## 5.11 Appendix 5.B: Results for the Sample Mean

We know derive statistical properties of the sample mean that were simply stated in the main text of this chapter.

### 5.11.1 Mean and Variance

The sample mean is a realization of the random variable $\bar{X}$ that is the sum of $X_1$ to $X_n$ divided by $n$. This can be written as a weighted sum of random variables

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n.$$

We first show that $\mathrm{E}[\bar{X}] = \mu$ given assumption 1 in Section 5.4. The mean of $\bar{X}$ equals the weighted sum of the means of $X_1$ to $X_n$. If we assume that the $X_i$ have common mean $\mu$, then $\mathrm{E}[X_i] = \mu$ and hence

$$\mathrm{E}[\bar{X}] = \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu.$$

Next we want to show that $\mathrm{Var}[\bar{X}] = \sigma^2/n$ given assumptions 1-3 in Section 5.4. Then the $X_i$ are statistically independent and have common variance $\sigma^2$. The variance of $\bar{X}$ is a

weighted sum of each variance, but the weights $1/n$ are now squared, as stated in Appendix 5.A. We have

$$\begin{aligned}\text{Var}[\bar{X}] &= \text{Var}\left[\tfrac{1}{n}X_1 + \tfrac{1}{n}X_2 + ... + \tfrac{1}{n}X_n\right] \\ &= \text{Var}\left[\tfrac{1}{n}X_1\right] + \text{Var}\left[\tfrac{1}{n}X_2\right] + \cdots + \text{Var}\left[\tfrac{1}{n}X_n\right] \\ &= \left(\tfrac{1}{n}\right)^2 \text{Var}\left[X_1\right] + \cdots + \left(\tfrac{1}{n}\right)^2 \text{Var}\left[X_n\right].\end{aligned}$$

The second equality uses statistical independence of the $X_i$, and the third equality uses $\text{Var}[bX] = b^2\text{Var}[X]$. Given the common variance $\sigma^2$, $\text{Var}[\bar{X}] = \left(\frac{1}{n}\right)^2 \sigma^2 + \cdots + \left(\frac{1}{n}\right)^2 \sigma^2$, which equals $n$ times $\left(\frac{1}{n}\right)^2 \sigma^2$, which equals $\frac{1}{n}\sigma^2$.

### 5.11.2 Law of Large Numbers

Now consider the behavior of $\bar{X}$ as the sample size gets large. Then $\bar{X}$ has mean $\mu$ and variance that goes to zero, since the variance $\sigma^2/n \to 0$ as $n \to \infty$.

So the distribution of $\bar{X}$ is centered on $\mu$ with very little variation around $\mu$. The formal statistical term used is that $\bar{X}$ **converges in probability** to $\mu$ if the probability that $|\bar{X} - \mu| > \varepsilon$ goes to zero as $n \to \infty$, no matter how small $\varepsilon > 0$ is chosen to be.

A **law of large numbers** states that, under some assumptions, an average of random variables converges in probability to its expected value; here that $\bar{X}$ converges in probability $\mu$. The simplest law of large numbers assumes that $X_i$ are statistically independent and identically distributed and that the mean $\mu$ exists. This is the case for simple random sampling.

### 5.11.3 Central Limit Theorem

As explained in Chapter 5.4, the standardized variable $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ has mean zero and variance one.

The **central limit theorem** states that, under some assumptions, $Z$ is standard normally distributed as the sample size gets large. That is, as $n \to \infty$

$$Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Note that this result does not require that $X_i$ is normally distributed. It follows that in large samples $\bar{X} \sim N(\mu, \sigma^2/n)$.

The simplest central limit theorem is the Lindberg-Levy central limit theorem that assumes that $X_i$ are statistically independent and identically distributed with mean $\mu$ and variance $\sigma^2$. This is the case for simple random sampling.

More general central limit theorems do not require a common variance and/or statistical independence. IN the more general case $Z = (\bar{X} - \mu)/\sqrt{\text{Var}[\bar{X}]}$ where $\text{Var}[\bar{X}]$ may no longer simplify to $\sigma/\sqrt{n}$.

## 5.12 Appendix 5.C: Computer Generation of a Random Sample

In the coin toss example in Section 5.3 we did not actually toss a coin 12,000 times to obtain the results for 400 samples, each with 30 coin tosses. Instead a computer was used to simulate the coin tosses.

A **uniform random number generator** creates values between 0 and 1 such that any value between 0 and 1 is equally likely and successive values appear to be independent of each other.

Then to simulate 30 coin tosses we draw 30 uniform random numbers and let the result be heads if the uniform random number exceeds 0.5, and tails if the uniform random number is less than 0.5.

Similarly for the Census example, if the uniform random number is between 0 and $1/N$, where $N = 50{,}169{,}452$, we choose the first person. If the uniform random number is between $1/N$ and $2/N$ we choose the second person, and so on.

The random numbers are more properly called **pseudo random numbers**, as a deterministic rule is used to create the sequence of numbers $u_1, u_2, ....$ For example, one method given value $u_j$ specifies the next value to be $u_{j+1} = (69069u_j + 1234567) \bmod 2^{32}$, where $a \bmod b$ is the remainder when $a$ is divided by $b$. Remarkably this rule leads to $u_{j+1}$ appearing to be unrelated to $u_j$ and to the different values of $u_j$ being equally likely. The sequence depends on the starting value $u_0$, called the **seed**. For example, we might set the seed equal to 10101.

When using random numbers it is always good practice to **set the seed**, as then results can be replicated exactly in future simulations.

# Chapter 6

# © A. Colin Cameron: Statistical Inference for the Mean

As explained in Chapter 5, the sample mean $\bar{x}$ is a random outcome – different samples lead to a different value of the sample mean. The sample at hand is viewed as being one from a population, where the mean in the population is called the population mean and is denoted by $\mu$. The goal is to make inference on the population mean given the observed sample mean. For example, is the view that population mean earnings equal \$40,000, say, consistent with sample mean earnings equal to \$41,413.

This chapter analyzes inference based on the sample mean. It presents the fundamentals of statistical inference, notably confidence intervals and hypothesis tests.

While the focus is on statistical inference for the mean, the concepts discussed for the mean carry over to other univariate statistics, such as the median and the standard deviation, though this is much less common in practice. The concepts also carry over to regression, the subject of later chapters. A good understanding of statistical inference is essential as it lies at the heart of analysis of economics data.

The chapter continues directly from the previous chapter. For readers who bypassed the details in Chapter 5, the key result for statistical inference on the mean is reproduced in Chapter 6.2.

## 6.1 Example: Mean Annual Earnings

Dataset EARNINGS introduced in Chapter 2 has data on individual annual earnings for a sample of 30 year-old full-time workers in 2010. The population considered is all 30 year-old female full-time workers in 2010 in the United States, with unknown population mean earnings denoted $\mu$. We wish to make inference about $\mu$, using data from the sample which is a random sample from the population.

Table 6.1: Sample mean: Annual earnings of female full-time workers aged 30 in 2010.

| Variable | Mean | Standard Deviation | Standard Error of the Mean | 95% confidence interval | |
|---|---|---|---|---|---|
| Earnings | 41412.69 | 25527 | 1952.10 | 37559.21 | 45266.17 |

Table 6.1 presents several key sample statistics, generated by a statistical package, that are used to make statistical inference on the **population mean** $\mu$.

The **sample mean** is the commonly-used estimate of the population mean. From Table 6.1, $\bar{x} = 41412.69$, so the estimate of population mean earnings is \$41,413.

The **sample standard deviation** $s = 25527$ is an estimate of the population standard deviation $\sigma$.

The **standard error of the sample mean**, where standard error is the statistical term for estimated standard deviation, measures the precision of $\bar{x}$ as an estimate of $\mu$. A smaller standard error means greater precision of $\bar{x}$ as an estimate of $\mu$. From Table 6.1, the standard error of the sample mean equals \$1,952. This is much smaller than the standard deviation of \$25,227 for just one observation, because averaging reduces the variability. In fact, from Chapter 5.4, the standard error given simple random sampling equals the standard deviation of a single observation divided by the square root of the sample size. Here $s/\sqrt{n} = 25527/\sqrt{171} = 1952$.

A **confidence interval** gives a range of values that includes the true (unknown) population mean $\mu$, where "confidence interval" is defined precisely in Chapter 6.3. From Table 6.1, a 95 percent confidence interval for population mean earnings is (\$37,559, \$45,266).

An **hypothesis test** is a test of whether or not the data support a hypothesized value for the population mean $\mu$, such as $\mu = 40000$, or a range of values, such as $\mu \leq 40000$. Hypothesis tests, not included in Table 6.1, are presented in Chapter 6.4.

This example sketches out key aspects of statistical inference on the population mean based on the sample mean. The remainder of this chapter provides complete details.

## 6.2 t Statistic and t Distribution

Statistical inference on the population mean is based on the $t$-statistic, a transformation of the sample mean, rather than the sample mean itself. In this section we repeat the results given in Chapter 5.5, and present the $t$ distribution.

### 6.2.1 The $t$ Statistic

The sample $x_1, ..., x_n$ is assumed to be a simple random sample, meaning that each sample observation $x_i$ is the independent realization of a random variable $X_i$ with common (unknown) population mean $\mu$ and common (unknown) population variance $\sigma^2$. In that case assumptions 1-3 of Chapter 5.4 are satisfied.

A remarkable result, the central limit theorem, is that if the simple random sample is of large size then the sample mean $\bar{x}$ is normally distributed, regardless of the distribution of the individual observations $x_i$. This distribution, however, depends on the unknown population standard deviation variance $\sigma$, so we replace $\sigma$ by its estimate, the sample standard deviation $s$. From Chapter 5.5, this leads to the $t$-statistic that is $t$ distributed.

**Remark 16** *From a **simple random sample** $x_1, ..., x_n$ calculate the **sample mean** $\bar{x}$, the sample standard deviation $s$ and the **standard error of** $\bar{x}$, $se(\bar{x}) = s/\sqrt{n}$. The **t-statistic***

$$t = \frac{\bar{x} - \mu}{se(\bar{x})} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

*is a realization of a random variable that is approximately $T(n-1)$ distributed, where $T(n-1)$ denotes the **t distribution** with $n-1$ degrees of freedom.*

The $T(n-1)$ approximation is exact if additionally the data are normally distributed or if $n \to \infty$, and is often good if $n > 30$.

To form the $t$-statistic the only summary statistics of the sample needed are the sample mean $\bar{x}$ and the sample standard deviation $s$. Additionally the $t$-statistic depends on the population mean $\mu$, which is unknown. The knowledge that the $t-$statistic is $T(n-1)$ distributed used to make statistical inference on $\mu$, as detailed in subsequent sections of this chapter.

### 6.2.2 The $t$ Distribution

A $t$-distributed random variable is a continuous random variable. In that case probabilities are given by the area under the probability density function. The formula for the probability density function for the standard normal distribution was given in Appendix 5.A of the previous chapter. The formula for the $t$-distribution. is considerably more complex and is not given here. Instead the properties of the $t$ distribution are outlined.

The probability density function for the $t$ **distribution**, or **Student's** $t$ **distribution**, is a bell-shaped curve centered on zero, and symmetric about zero, that is a slightly squashed version of the standard normal. It has one parameter, denoted $v$ here, called the **degrees of freedom**. The $t$ distribution with $v$ degrees of freedom, denoted $T(v)$, has mean 0, provided $v > 1$, and variance $v/(v-2)$, provided $v > 2$. The standard normal has the same mean of 0, but smaller variance of 1.

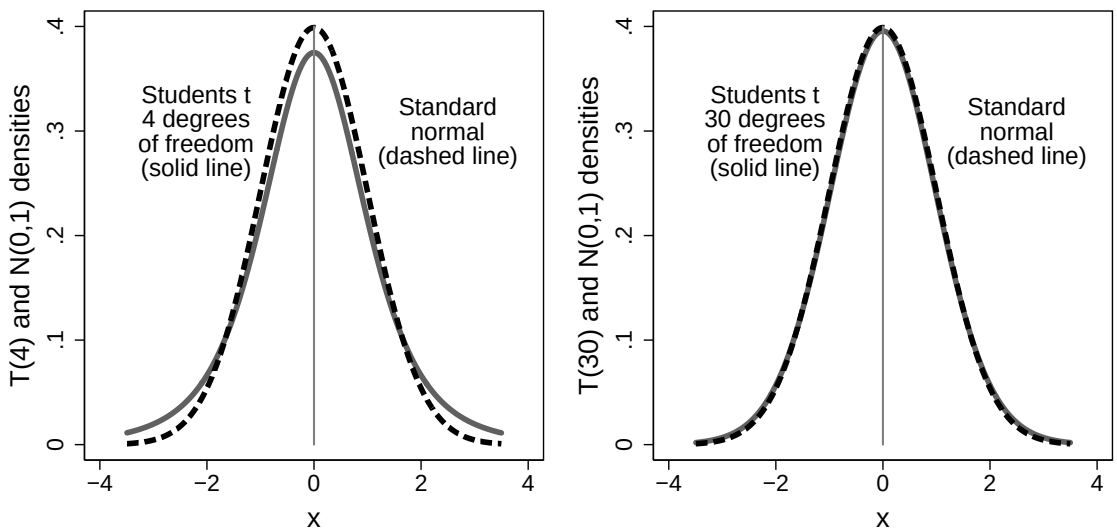


Figure 6.1: Student's $t$ distribution: $T(4)$ and $T(30)$ compared to the standard normal.

Figure 6.1 presents, respectively, the $T(4)$ and $T(30)$ distributions and compares them in each case to the standard normal. The $t$ distribution has greater area in the tails than the standard normal, a difference that gets smaller as $v$ gets larger and disappears as $v \to \infty$.

In most economics applications where the $t$ distribution is used, $v > 30$. From the second panel of Figure 6.1 there is seemingly little difference between the $T(30)$ and the standard normal distributions. But there is still an appreciable difference in the tails of these distributions. In fact $\Pr[|T_{30}| > 2] = 0.0546$, where $T_{30}$ denotes a random variable that is $T(30)$ distributed. This is approximately 20% larger than $\Pr[|Z| > 2] = 0.0455$ for $Z$ standard normal distributed.

Such differences can be large enough to matter for confidence intervals and hypothesis tests because they use tail probabilities. For this reason statistical packages and this book base inference on the $t$ distribution rather than the standard normal distribution.

As the degrees of freedom $v \to \infty$, the difference disappears since the $t$ distribution then collapses to the standard normal distribution. For example, for $T_{1000} \sim T(1000)$ we have $\Pr[|T_{1000}| > 2] = 0.0458$, very close to 0.0455 for the standard normal.

**Remark 17** *The $t$ distribution with $v$ degrees of freedom, denoted $T(v)$, is like a squashed version of the standard normal distribution with fatter tails. As $v \to \infty$ the $t$ distribution goes to the standard normal.*

### 6.2.3 Probabilities and Critical values for the $t$ Distribution

The computation of **probabilities** for the $t$ distribution requires advanced numerical methods. Until recently statisticians needed to refer to published tables. Now one can directly use

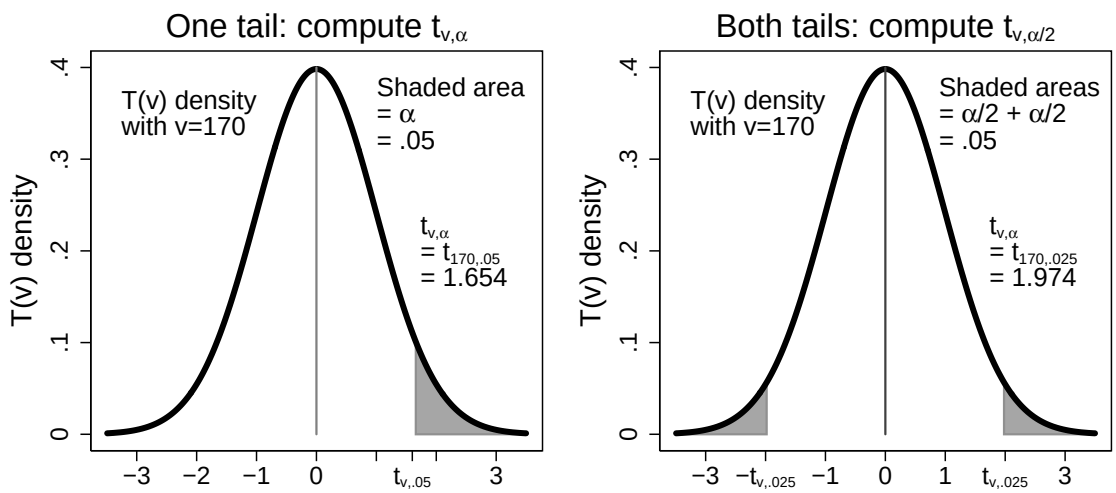


Figure 6.2: Student's t distribution: t-values $t_{v,\alpha}$ and $t_{v,\alpha/2}$ for $v = 170$ and $\alpha = 0.05$.

a computer. For example, to compute $\Pr[T_{170} > 1.5]$, the probability that a $T(170)$ random variable exceeds 1.5, one can use the Stata function ttail(170,1) which returns a probability of 0.06760.

In some situations this computation needs to be inverted. The probability is set and we wish to calculate the associated value of $t$ that gives this probability. For example we may wish to find the value $c$ such that the probability that a $T(170)$ distributed random variable exceeds $c$ is equal to 0.05. Then one can use the Stata function invttail(170,.05) which returns a value of 1.6539. We have that $c = 1.6539$ solves $\Pr[T_{170} > c] = 0.05$.

Appendix A of this book gives corresponding commands for other statistical packages.

More generally the following notation is used. The **critical value** $c = t_{v,\alpha}$ satisfies

$$\Pr[T_v > t_{v,\alpha}] = \alpha.$$

In words, the critical value $t_{v,\alpha}$ is that value such that a $T(v)$ distributed random variable exceeds $t_{v,\alpha}$ with probability $\alpha$. Even more simply, **the area under the curve to the right of $t_{v,\alpha}$ equals** $\alpha$. Appendix A of this book gives commands to obtain probabilities and critical values using several different statistical packages.

The left panel of Figure 6.2 presents the example $\Pr[T_{170} > 1.654] = 0.05$. Then $\alpha = 0.05$ is the shaded area in the right tail, and the critical value $t_{v,\alpha} = t_{170,.05} = 1.654$ is given on the horizontal axis.

**Remark 18** *The **critical value** $t_{v,\alpha}$ is that value for which a $T(v)$ distributed random variable exceeds $t_{v,\alpha}$ with probability $\alpha$, i.e.* $\Pr[T_v > t_{v,\alpha}] = \alpha$.

Sometimes we want the **combined area** in left and right tails to equal $\alpha$. Given the symmetry about 0 of the $t$ distribution we have

$$\Pr[|T_v| > t_{v,\alpha/2}] = \Pr[T_v < -t_{v,\alpha/2}] + \Pr[T_v > t_{v,\alpha/2}] = \alpha/2 + \alpha/2 = \alpha.$$

The **combined area under the curve to the left of $-t_{v,\alpha/2}$ and to the right of $t_{v,\alpha/2}$ equals** $\alpha$.

For example, $\Pr[T_{170} > 1.974] = 0.025$, so $t_{170,.025} = 1.974$. Combining both tails of the $t$ distribution it follows that $\Pr[|T_{170}| > 1.974] = 0.05$.

The left panel of Figure 6.2 presents the example $\Pr[|T_{170}| > 1.974] = 0.05$. Then the shaded area in each tail is 0.025, $\alpha = 0.05$ is the combined area in the two tails, and the critical value $t_{v,\alpha} = t_{170,.025} = 1.974$ is given on the horizontal axis.

**Remark 19** *A $t$ distributed random variable with $v$ degrees of freedom exceeds* ***in absolute value*** *the* ***critical value*** $t_{v,\alpha/2}$ *with probability $\alpha$, i.e.* $\Pr[|T_v| > t_{v,\alpha/2}] = \alpha$.

Note that some books define $t_{170,.05}$, for example, to be the .05 quantile or $5^{th}$ percentile, so the area in the **left tail** of the distribution is .05. Throughout this book, however, $t_{170,.05}$ is that value for which the area in the **right tail** of the distribution is .05. This makes no difference in practice due to the symmetry of the $t$ distribution about zero, for example, $t_{170,.05} = t_{170,.95}$.

## 6.3 Confidence Intervals

Different samples will lead to different estimates of the population mean. A confidence interval for an unknown parameter, such as the population mean, gives a range of values that the parameter lies in with a certain "confidence level", defined below.

### 6.3.1 Example

Before presenting the general result, a 95% confidence interval is derived from first principles using a simpler specific example.

Consider a sample with $\bar{x} = 30$ and $n = 61$, and standard error of the mean $s/\sqrt{n} = 5$. Then the $t$-statistic $t = (\bar{x} - \mu)/(s/\sqrt{n})) = (30 - \mu)/5$ is a draw from the $T(60)$ distribution.

Now from $T(60)$ tables

$$\Pr[-2.0003 < T_{60} < 2.0003] = 0.95,$$

which we round to $\Pr[-2 < T < 2] = 0.95$. Substituting $t = (30 - \mu)/5$ it follows that

$$\Pr\left[-2 < \frac{30 - \mu}{5} < 2\right] = 0.95.$$

This interval can be converted to an interval that is centered on $\mu$ as follows

$$\begin{aligned}
& \Pr\left[-2 < \tfrac{30-\mu}{5} < 2\right] &&= 0.95 \\
\Rightarrow\ & \Pr\left[-10 < 30-\mu < 10\right] &&= 0.95 \quad \text{multiplying all terms by the standard error 5} \\
\Rightarrow\ & \Pr\left[-40 < -\mu < -20\right] &&= 0.95 \quad \text{subtracting the mean 30 from all terms} \\
\Rightarrow\ & \Pr\left[40 > \mu > 20\right] &&= 0.95 \quad \text{multiplying by } -1 \text{ reverses the inequalities.}
\end{aligned}$$

Re-ordering the final inequality yields

$$\Pr\left[20 < \mu < 40\right] = 0.95.$$

The interval $(20, 40)$ is called a 95% confidence interval for $\mu$.

### 6.3.2 Confidence Interval

More generally, consider **confidence intervals** with unspecified values for the sample mean and its standard error, and general **confidence level** $100(1-\alpha)\%$.

Different values of $\alpha$ correspond to different levels of confidence. In particular, $\alpha = .05$ corresponds to a 95% confidence interval since $100(1-.05) = 100 \times .95 = 95$.

Let $T_{n-1}$ denote a random variable that is $T(n-1)$ distributed. Then

$$\Pr[-t_{n-1,\alpha/2} < T_{n-1} < t_{n-1,\alpha/2}] = 1-\alpha.$$

Substituting in for the sample value of the $t$-statistic yields

$$\Pr\left[-t_{n-1,\alpha/2} < \frac{\bar{x}-\mu}{se(\bar{x})} < t_{n-1,\alpha/2}\right] = 1-\alpha.$$

Several algebraic steps, similar to those in the preceding example, lead to

$$\Pr\left[\bar{x} - t_{n-1,\alpha/2} \times se(\bar{x}) < \mu < \bar{x} + t_{n-1,\alpha/2} \times se(\bar{x})\right] = 1-\alpha.$$

This leads to the following result.

**Remark 20** *A* ***100*$(1-\alpha)$** ***percent confidence interval for the population mean*** *is*

$$\bar{x} \pm t_{n-1,\alpha/2} \times se(\bar{x}),$$

*where $\bar{x}$ is the sample mean; $t_{n-1,\alpha/2}$ is that value such that a $T(n-1)$ distributed random variable exceeds it in absolute value with probability $\alpha$; and $se(\bar{x}) = s/\sqrt{n}$ is the standard error of the sample mean.*

The confidence interval is centered around the estimate of $\mu$, the sample mean $\bar{x}$, and is symmetric.

Intuitively the confidence interval is narrower the more precise is our estimate of $\mu$. This is indeed the case, as from the formula the confidence interval is narrower the smaller is the standard error of $\bar{x}$. In particular, we have the following result.

**Remark 21** *The confidence interval narrows as the sample size gets larger, since larger samples lead to a smaller standard error.*

Table 6.2: Student's t distribution: t-values for various degrees of freedom and confidence levels.

| **Confidence Level** | $\mathbf{100(1-\alpha)}$ | **90%** | **95%** | **99%** |
|---|---|---|---|---|
| Area in both tails | $\alpha$ | 0.10 | **0.05** | 0.01 |
| Area in single tail | $\alpha/2$ | 0.05 | **0.025** | 0.005 |
| t value for $v = 10$ | $t_{10,\alpha/2}$ | 1.812 | **2.228** | 3.169 |
| t value for $v = 30$ | $t_{30,\alpha/2}$ | 1.697 | **2.042** | 2.750 |
| t value for $v = 100$ | $t_{100,\alpha/2}$ | 1.660 | **1.980** | 2.626 |
| t value for $v = \infty$ | $t_{\infty,\alpha/2}$ | 1.645 | **1.960** | 2.576 |
| standard normal value | $z_{\alpha/2}$ | 1.645 | **1.960** | 2.576 |

### 6.3.3 What Level of Confidence?

Ideally there is both a high level of confidence and a narrow confidence interval. For example, having 95% confidence that $\mu$ lies between 20 and 40 is preferred to having only 90% confidence that $\mu$ lies between 20 and 40. And having 95% confidence that $\mu$ lies between 20 and 40 is preferred to having 95% confidence that $\mu$ lies in the broader range of 10 to 50.

Unfortunately there is a trade-off between these two considerations. The confidence interval widens as the confidence level is increased. Algebraically, at higher confidence levels $\alpha$ becomes smaller and $t_{n-1,\alpha/2}$ becomes larger.

So what value of confidence should be used? There is no best value in general, but **it is most common to use a 95% confidence interval**, the special case $\alpha = .05$ (so $\alpha/2 = .025$). The other common choices are to use 90% confidence intervals, with $\alpha = .10$, and 99% confidence intervals with $\alpha = .01$.

Table 6.2 presents $t_{v,\alpha/2}$ for various confidence levels, corresponding to different values of $\alpha$, and for selected different numbers of observations, corresponding to different values of $v = n - 1$. The value $t_{v,\alpha/2}$ decreases as the sample size increases. For 95% confidence intervals, presented in bold, the $t$ value is 2.042 for the $t_{30}$ distribution falling to 1.960 for the $t_\infty$ distribution which is equivalent to the standard normal distribution.

A 95% **percent confidence interval for the population mean** is

$$\bar{x} \pm t_{n-1,.025} \times se(\bar{x}).$$

In typical econometrics applications the sample size $n > 30$, in which case from Table 6.2 the critical value $t_{n-1,.025}$ approximately equals 2. This leads to the following.

**Remark 22** *It is most common, though arbitrary, to use a 95% confidence interval. An* ***approximate 95% confidence interval*** *for the population mean is a* ***two-standard error interval****: the sample mean plus or minus two times the standard error,*

This is a useful guide. And it makes clear that if we are willing to tolerate an error range of plus or minus two standard errors, then a good choice of the confidence level is 95%. In any published work or in assignments, however, use the more precise interval $\bar{x} \pm t_{n-1,.025} \times se(\bar{x})$.

### 6.3.4 Example: Mean Annual Earnings

For the female annual earnings data in dataset EARNINGS, introduced in Chapter 2, $n = 171$, $\bar{x} = 41413$, $s = 25527$, and $se(\bar{x}) = s/\sqrt{n} = 1952$. From $T_{170}$ tables, $t_{170,.025} = 1.974$.

It follows that a 95% confidence interval for population mean earnings of thirty year-old female full-time workers is

$$\bar{x} \pm t_{n-1,\alpha/2} \times se(\bar{x}) = 41413 \pm 1.974 \times 1952 = 41413 \pm 3853 = (37560,\ 45266).$$

This is the 95% confidence interval that was given in Table 6.1.

A 90% confidence interval uses $t_{170,.05} = 1.654$ leading to narrower interval (\$38,184, \$44,641), and a 99% confidence interval uses $t_{170,.005} = 2.605$ leading to wider interval (\$36,328, \$46,498). As mentioned earlier, higher levels of confidence lead to wider confidence intervals.

### 6.3.5 Interpretation of Confidence Intervals

Interpretation of confidence intervals is conceptually difficult. Consider the annual earnings example with 95 percent confidence interval (\$37,560, \$45,266). The correct interpretation of this confidence interval is that **the calculated 95 percent confidence interval** from this sample **includes the true population mean $\mu$ with probability 0.95.**

To understand this interpretation it is necessary to imagine that there are many separate samples of the population, each of size $n = 171$ in this example. From each sample we form a 95% confidence interval. Then we expect 95% of such confidence intervals to include the true (unknown) mean $\mu$.

Return to the two examples in Chapter 5 where $\mu$ was actually known. For the 1880 Census example in Chapter 5.2 we know $\mu = 24.13$. Further analysis of the 100 samples of size 25 yields a 95% confidence interval (17.99, 34.81) for the first sample, (13.12, 25.54) for the second sample, and so on. In total 98 of the 100 samples had 95% confidence intervals that included $\mu = 24.13$; the two that did not were the $20^{th}$ sample with 95% confidence interval (8.57, 23.90) and the $50^{th}$ sample with 95% confidence interval (11.49, 21.45). In theory we expect 95% of the 95% confidence intervals to include $\mu$. The reason 98% rather than 95% included $\mu$ reflects randomness with just 100 confidence intervals. If we had obtained one million 95% confidence intervals, say, and the $t$ statistic was exactly $T_{24}$ distributed, then very close to 95% of these intervals would include $\mu$. Similarly, for the coin toss example in Chapter 5.3, 386 of the 400 95% confidence intervals, or 96.5%, included the true parameter value $\mu = 0.5$.

As these examples demonstrate, a confidence interval will sometimes fail to include the population mean $\mu$, due to the randomness inherent in sampling. A 95 percent confidence interval has the property that if we were able to obtain many separate random samples, 95 percent of the resulting confidence intervals will include the population mean $\mu$, and 5 percent will not.

In fact we have only one sample, and we say that the calculated 95 percent confidence interval from this sample includes the true population mean $\mu$ with probability 0.95. This probabilistic statement refers to the confidence interval, which is random, and not to $\mu$ which is fixed. It is **wrong** to instead interpret this confidence interval as meaning that with probability 0.95 the population mean $\mu$ lies inside (\$37,560, \$45,266) and with probability 0.05 it lies outside this range.

**Remark 23** *A calculated 95 percent confidence interval for the population mean is an interval that includes the true population mean $\mu$ with probability 0.95.*

## 6.4 Two-Sided Hypothesis Tests

A confidence interval for $\mu$ provides a likely range of values for $\mu$. By contrast, a two-sided hypothesis test is a test of whether or not $\mu$ takes a specified value.

### 6.4.1 Null and Alternative Hypotheses

The particular hypothesis under test is called the **null hypothesis** and is denoted $H_0$. The alternative to the hypothesis test is called the **alternative hypothesis** and is denoted $H_a$.

Here we consider test of whether $\mu$ takes a particular value. Let $\mu^*$ denote this value. Then the null hypothesis is $H_0 : \mu = \mu^*$, and the alternative hypothesis is $H_a : \mu \neq \mu^*$. Because the alternative hypothesis includes both $\mu < \mu^*$ and $\mu > \mu^*$ the test is called a two-sided test.

For example, consider the claim that population mean earnings equal \$40,000. To test this claim we test $H_0 : \mu = 40000$ against $H_a : \mu \neq 40000$.

**Remark 24** *A **two-sided test** or **two-tailed test** for the population mean $\mu$ is a test of the null hypothesis*

$$H_0 : \mu = \mu^*,$$

*where $\mu^*$ is a specified value for $\mu$, against the alternative hypothesis*

$$H_a : \mu \neq \mu^*.$$

### 6.4.2 Significance Level of a Test

The result of a test is to either reject or not reject the null hypothesis.

This decision made may be in error. In particular, we may reject the null hypothesis when in fact it was true. This type of error is called a type I error.

For example, suppose the null hypothesis is that someone is innocent. Then a type I error is made if we reject the null and find the person guilty, when in fact the person was innocent.

In the earnings example a type I error occurs if we reject $H_0 : \mu = 40000$ when in fact $\mu = 40000$.

**Remark 25** *A **type I error** occurs if $H_0$ is rejected when $H_0$ is true.*

Ideally the probability of making a type I error is small. The following terminology is used.

**Remark 26** *The **significance level** of a test, denoted $\alpha$, is the pre-specified maximum probability of a type I error that will be tolerated.*

The level of statistical significance to use is discussed in some length in a later section. It is most common to tolerate up to a 5% chance of making a type I error, in which case $\alpha = 0.05$.

Whatever the choice of $\alpha$, we reject $H_0$ at significance level $\alpha$ if the probability of making a type I error is less than 0.05, and do not reject $H_0$ otherwise.

Note that if we do not reject the null hypothesis then we simply say that we "fail to reject the null hypothesis." We do not say that we "accept the null hypothesis." The reason for doing so is that there are other null hypotheses that we also fail to reject. For example, in the earnings example we show below that we do not reject $H_0 : \mu = 40000$. But for these data other null hypotheses, such as $H_0 : \mu = 41000$, will also be not rejected at level 0.05.

The ability of tests to discriminate between various values of $\mu$ is pursued in Chapter 6.5 where we introduce a second type of error, a type II error that occurs if we fail to reject $H_0$ when $H_0$ is false. Discrimination is easier the more precisely $\bar{x}$ is estimated. With more data $\bar{x}$ becomes more precisely estimated.

### 6.4.3 The t-test Statistic

Implementing an hypothesis test requires a decision rule for rejecting the null hypothesis, and calculation of the probability of making a type I error for this decision rule.

The obvious decision rule is to reject $H_0 : \mu = \mu^*$ if the sample mean $\bar{x}$ is far from the hypothesized population mean $\mu^*$. It is convenient to transform from this difference $(\bar{x} - \mu^*)$ to $(\bar{x} - \mu^*)/se(\bar{x})$, since the latter is known to have a $t$ distribution.

**Remark 27** *The $t$-test statistic for test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$ is $t = (\bar{x} - \mu^*)/se(\bar{x})$. Under $H_0$, and assuming simple random sampling, $t$ is the realization of a random variable that is approximately $T(n-1)$ distributed.*

For the data on earnings of 30 year-old female full-time workers in 2010 in the United States, $n = 171$, $\bar{x} = 41413$ and $se(\bar{x}) = 1952$.

For test of $H_0 : \mu = 40000$, the $t$-test statistic is therefore

$$t = (\bar{x} - 40000)/se(\bar{x}) = (41413 - 40000)/1952 = 0.724.$$

Under the null hypothesis, this is a draw from the $T_{170}$ distribution, since $n - 1 = 170$.

### 6.4.4 Rejection using $p$-values

Continuing immediately with the earnings example, we found that $t = 0.724$. How likely are we to obtain a draw from the $T_{170}$ distribution (here $n - 1 = 170$) that is equal to 0.724 or greater in absolute value? This probability is called the $p$-value, and in this example it is quite high.

**Remark 28** *The $p$-**value** is the probability of observing a $t$-test statistic at least as large in absolute value as that obtained in the current sample. For a two-sided test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$ the $p$-value is $p = \Pr[|T_{n-1}| \geq |t|]$. $H_0$ is rejected at significance level $\alpha$ if $p < \alpha$, and is not rejected otherwise.*

For the earnings example

$$p = \Pr[|T_{170}| \geq 0.724] = 0.470.$$

There is high probability ($p = 0.470$) of observing a $t$ value of 0.724 or larger in absolute value, even if the mean really is the hypothesized value of \$40,000. Since $p > 0.05$ we do not reject the null hypothesis that $\mu = 40000$ at significance level 0.05.

The left panel of Figure 6.3 displays the probability that $|T_{170}|$ exceeds the observed $t$-test statistic of 0.724.

We are more likely to reject the null hypothesis the larger is the absolute value of the $t$-statistic. So, other things equal, we are more likely to reject the null hypothesis the larger is $se(\bar{x})$, since computation of $t$ entails division by $se(\bar{x})$. So a larger sample is better, as then $se(\bar{x})$ is smaller.

**Remark 29** *More precise estimation of $\mu$, such as through a larger sample size, makes it more likely that the null hypothesis is rejected.*

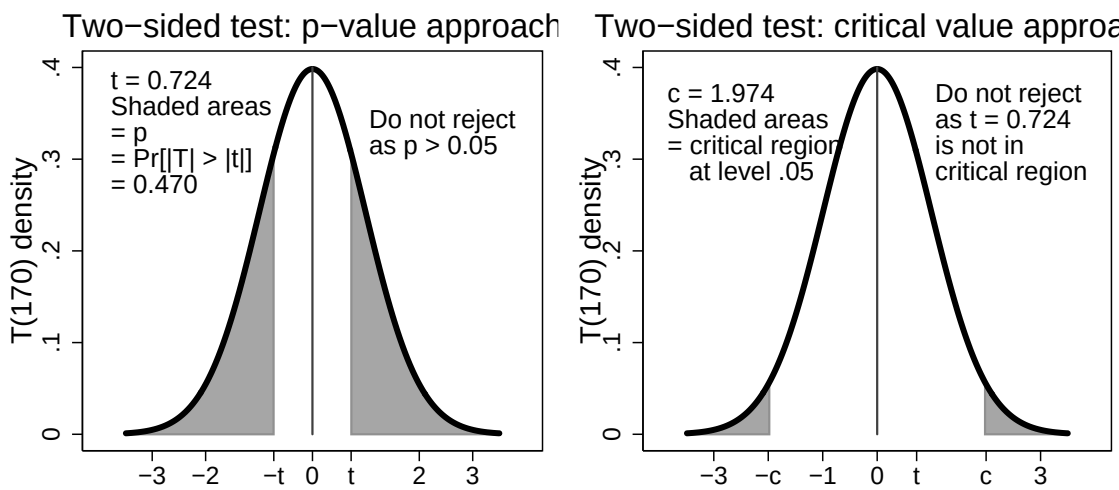


Figure 6.3: Two-sided hypothesis test when $t = 0.724$, $n = 171$ and $\alpha = 0.05$.

### 6.4.5 Rejection using Critical Regions

The $p$-value method requires access to a computer, in order to precisely compute the $p$-value. Before widespread access to computers, an alternative method was used that, for given significance level $\alpha$, leads to the same conclusion.

This **alternative equivalent method** defines a **critical region** or **rejection region**, which is the range of values of $t$ that would lead to rejection of $H_0$ at the specified significance level $\alpha$. Then reject $H_0$ if the computed value of $t$ falls in this range.

**Remark 30** *For a two-sided test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$, and for specified $\alpha$, the **critical value** $c$ is such that $c = t_{n-1,\alpha/2}$; equivalently $\Pr[|T_{n-1}| \geq c] = \alpha$. $H_0$ is rejected at significance level $\alpha$ if $|t| > c$, and is not rejected otherwise.*

Return to the female earnings example. For significance level 0.05 and $n - 1 = 170$, the critical value

$$c = t_{170,.025} = 1.974.$$

$H_0$ is not rejected at significance level 0.05, since $t = 0.724$ does not exceed 1.974 in absolute value. This conclusion is the same as that using the $p$-value approach.

The shaded region in the right panel of Figure 6.3 shows the rejection region, the range of values for which $|T_{170}|$ exceeds 1.974 since this occurs with probability 0.05. The sample $t$-test statistic equals 0.724 which does not fall in the shaded region. So we do not reject $H_0$.

The critical value approach has the advantage that it does not require computing the $p$-value. Instead one refers to a printed table for the $t$ distribution. Typically $t$ tables are

given for area in the right table equal to 0.25, 0.10, 0.05, 0.01, and 0.005, and for all degrees of freedom from 1 to 30, every fifth integer from 30 to 60, then 70, 80, 90, 100 and $\infty$.

The attraction of reporting $p$-values, rather than $t$ values, is that the reader can then easily test using his or her own preferred value of $\alpha$. For this reason the $p$-value method is preferred. The alternative critical value method was developed for an earlier time when reliance on published tables made it difficult to accurately calculate $p$-values.

### 6.4.6 Which Significance Level?

Increasing the significance level $\alpha$ makes it less likely that the null hypothesis is rejected. This should be clear from, for example, the second panel of Figure 6.3, where the rejection region will get smaller as the error $\alpha$ in the two tails gets smaller. What significance level should be used?

**It is most common to use $\alpha = 0.05$, called a test at the 5% significance level**. Then a type I error is made 1 in 20 times.

This is a convention and in many applications other values of $\alpha$ may be warranted. For example, in testing the null hypothesis that there will be no nuclear war the significance level may be chosen to be much higher than 0.05, since the consequence of failing to reject the null hypothesis when it should be rejected is so high.

The choice of significance level is discussed in more detail in Chapter 6.5.

### 6.4.7 Relationship to Confidence Interval

Two-sided tests can be implemented using confidence intervals. If the null hypothesis value $\mu^*$ falls inside the $100(1-\alpha)$ percent confidence interval then do not reject $H_0$ at significance level $\alpha$. Otherwise reject $H_0$ at significance level $\alpha$.

For the female earnings data, from Table 6.1 the 95 percent confidence interval for population mean female earnings is $(37559, 45266)$. Since this interval includes 40000 we do not reject $H_0 : \mu = 40000$ at significance level 0.05.

### 6.4.8 Summary

A summary of the preceding earnings hypothesis test example is the following.

| | |
|---|---|
| Hypotheses | $H_0 : \mu = 40000$, $H_a : \mu \neq 40000$ |
| Significance level | $\alpha = 0.05$ |
| Data | $\bar{x} = 41413$, $s = 25527$, $n = 171$ |
| Test statistic | $t = (41413 - 40000)/(25527/\sqrt{171}) = 0.724$ |
| (1) p-value approach | $p = \Pr[\lvert T_{170}\rvert \geq \lvert 0.724\rvert] = 0.470$ |
| (2) Critical value approach | $c = t_{170,.025} = 1.974$ |
| Conclusion | Do not reject $H_0$ at level .05 as (1) $p > .05$ or (2) $\lvert t\rvert < c$. |

The p-value and critical value approaches are alternative methods that always lead to the same conclusion.

## 6.5 Test Power

The hypothesis testing method considered in the previous section considers only a type I error, the error of rejecting the null hypothesis when it is in fact true. An alternative error that may be made is to fail to reject the null hypothesis when it is false, called a type II error.

This more technical section discusses the trade-off between the two. The bottom line is that for the standard testing procedure

- the probability of mistakenly rejecting the null hypothesis is set at significance level $\alpha$ that is typically low
- the probability of mistakenly failing to reject the null hypothesis can then be very high
- the probability of making this second type of mistake, for given level of $\alpha$, is lower the more precise is the estimate of $\mu$.

So it is best to use as precise an estimate of $\mu$ as is possible. This is a major reason for using the sample mean $\bar{x}$ to estimate $\mu$ as in most circumstances it is a relatively precise estimator of $\mu$. From Chapter 5.6, the sample mean has the smallest variance among unbiased and consistent estimators if the data are distributed as normal, Bernoulli, binomial or Poisson. For other types of data the sample mean often has variance that is not much greater than the minimum possible variance.

Another way to improve precision is to obtain more data as then there is a reduction in $se(\bar{x})$, equal to $s/\sqrt{n}$ in the case of simple random sampling.

### 6.5.1 Type I and II errors

Consider a medical test of whether or not someone has a disease. A positive test result indicates disease while a negative result indicates the absence of disease.

Medical tests are not always perfect. A **false negative** occurs when the test fails to find disease that is present. And a **false positive** occurs when the test finds disease when in fact there is none.

Ideally a medical test leads to both few false positives and few false negatives. But there is a trade-off between the two. A more sensitive test will detect disease more often leading to fewer false negatives, but this will be at the expense of more false positives. And if there are many positive false positives, leading to expensive and stressful false alarms, it may not even be worthwhile doing the test.

Table 6.3: Hypothesis Tests: Type I and Type II Errors.

| **Decision** | **Truth** | |
|---|---|---|
| | $H_0$ really true: No disease | $H_0$ really false: Disease |
| Do not reject $H_0$ : Find no disease | Correct decision | Type II error (false positive) |
| Reject $H_0$ : Find disease | Type I error (false negative) | Correct decision |

Similar issues arise in hypothesis testing. Let the null hypothesis ($H_0$) be that there is no disease and the alternative hypothesis ($H_1$) be that disease is present. Then a false negative corresponds to rejecting the null hypothesis when we should not reject, a type I error used extensively in the previous section. A false positive corresponds to failing to reject the null hypothesis when we should reject. The standard terminology is the following.

**Remark 31** *A **type I error** (or false negative) occurs if $H_0$ is rejected when $H_0$ is true. A **type II error** (or false positive) occurs if $H_0$ is not rejected when $H_0$ is false.*

The situation is summarized in Table 6.3.

### 6.5.2 Test Size and Test Power

Ideally both Type 1 and type II errors occur with low probability. But unfortunately decreasing one means increasing the other. For example, we can reduce the probability of a type I error by being very conservative and rarely rejecting, but then the type II error probability will rise as we then are unlikely to reject even if we should.

**Remark 32** ***Test size** is the **probability of a type I error**.*

Test size is set at $\alpha$, the significance level of the test.

By contrast, studies rarely report the probability of a type II error because this varies with the parameter value under the alternative hypothesis. For example, we are very likely to erroneously fail to reject $H_0 : \mu = 40000$ if in fact $\mu = 39999$, and much less likely to do so if in fact $\mu = 20000$. The further apart the $H_0$ and $H_a$ values, the easier it is to discriminate, and the less likely we are to make a type II error.

Type II errors are controlled for by using testing procedures that keep the probability of a type II error as small as possible for the specified probability of a type I error. The following term is introduced.

**Remark 33** ***Test power** is **one minus the probability of a type II error**.*

Then high test power corresponds to low probability of a type II error.

The best tests, those with the most power for a given test size, are called **most powerful tests**. Such tests are based on the unbiased estimator (or consistent estimator) with minimum variance. Intuitively, the more precise our estimate of $\mu$ the less likely we are to make a type II error (for a specified probability of a type I error).

For tests of hypotheses on the population mean based on a simple random sample, the $t$ test is the most powerful test if the data are normally distributed.

It is common to choose a small value of $\alpha$, usually $\alpha = 0.05$, $\alpha = 0.01$ or $\alpha = 0.10$. These choices of small $\alpha$ reflect a reluctance to make a type I error through unwarranted rejections of the null hypothesis. But this comes with the potential cost of a high probability of a type II error by failing to reject the null hypothesis when we should have rejected.

## 6.6 Hypothesis Testing Examples

We consider three examples. The first is a two-sided test and the other two are one-sided tests. Additionally we discuss complications that can arise – survey data should be from a representative sample and for time series data usually the standard error of the mean, $se(\bar{x})$, no longer equals $s/\sqrt{n}$.

### 6.6.1 Example: Gasoline Prices

Test the claim that the mean price of regular gasoline in Yolo County is neither higher nor lower than the norm for California.

The dataset GASPRICE comes from a website that provides daily data on gas prices. Data are available for 32 Yolo County gas stations on a day when the average price for all California gas stations was \$3.81. Descriptive statistics are given in Table 6.4. The standard error of the sample mean $se(\bar{x}) = s/\sqrt{n} = 0.1510/\sqrt{32} = .0267$.

Table 6.4: Descriptive Statistics: Gasolin price per gallon at 32 gas stations.

| Variable | Obs | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Earnings | 32 | 3.6697 | 0.1510 | 3.49 | 4.09 |

The null hypothesis is $\mu = 3.81$, tested against the alternative $\mu \neq 3.81$. The $t$-statistic is

$$t = \frac{3.6697 - 3.81}{.0267} = -5.256.$$

Large values of $t$ in absolute value favor the alternative, as then $\bar{x}$ is very different from 3.81. Using the $p$-value method we have

$$p = \Pr[|T_{31}| > |-5.256|] = 0.000.$$

Reject $H_0$ at level .05 since $p < .05$. Using the critical value method

$$c = t_{31,.025} = 2.040.$$

Reject $H_0$ at level .05 since $|t| = 5.256 > c = 2.040$. Therefore reject the claim that population mean Yolo County gas prices equal the California state-average price.

### 6.6.2 Example: Male Earnings

Test the claim that mean earnings of male full-time workers in 2010 exceed $50,000.

The dataset EARNINGSMALE is a small subsample from the very large American Community Survey (ACS). The subsample is selected in such a way that it is a random sample of the population of 30 year-old male full-time workers in 2010. Descriptive statistics are given in Table 6.5. The minimum value of 1,000 is possible as the person was self-employed. The next lowest value was 8,000. The standard error of the sample mean $se(\bar{x}) = s/\sqrt{n} = 65035/\sqrt{191} = 4706$.

Table 6.5: Descriptive Statistics: Annual earnings of male full-time workers aged 30 in 2010.

| **Variable** | **Obs** | **Mean** | **St. Dev.** | **Min** | **Max** |
|---|---|---|---|---|---|
| Earnings | 191 | 52354 | 65035 | 1000 | 498000 |

The first decision is in setting up the null and alternative hypotheses. The claim that mean earnings exceed 50000 is set up as the alternative hypothesis, so $H_0 : \mu \leq 50000$ is tested against $H_a : \mu > 50000$. The $t$-statistic is

$$t = \frac{52354 - 50000}{4706} = .500.$$

Large positive values of $t$ favor the alternative, as then $\bar{x}$ is much greater than 50000. Here

$$p = \Pr[T_{190} > .498] = 0.309.$$

Do not reject $H_0$ at level .05 since $p = .310$ is not less than .05. Alternatively, the critical value

$$c = t_{190,.05} = 1.653.$$

Do not reject $H_0$ at level .05 since $t = .498$ is not less than $c = 1.653$. The data do not support the claim that population mean earnings are more than \$50,000 at significance level .05.

With the same data, now test the claim that population mean annual earnings for 30 year-old male full-time workers in 2010 are less than \$60,000. The claim is set up as the alternative hypothesis, so $H_0 : \mu \geq 60000$ is tested against $H_a : \mu < 60000$. The $t$-statistic is $(52354 - 60000)/4706 = -1.625$. Then $p = \Pr[T_{190} < -1.625] = 0.053$. Do not reject $H_0$ at level .05 since $p = .053$ exceeds .05. Alternatively $c = -t_{190,.05} = -1.653$, so do not reject $H_0$ as $t > c$. The data do not support the claim that population mean earnings are less than \$50,000 at significance level .05.

Note that it is important that the sample be a representative sample. National government-sponsored surveys are usually not representative of the U.S. population, as they tend to over-sample low population segments of interest to policy-makers, such as racial minorities, people with low income, and people in low population states. This is likely to lead to over-sampling of low-earnings individuals.

For **nonrepresentative samples** with sampling weights we should base inference on the sample mean $\bar{x}_w$ and its standard error $se(\bar{x}_w)$ that are defined in Chapter 5.4. Then the $100(1-\alpha)\%$ confidence interval for $\mu$ is $\bar{x}_w \pm t_{n-1,\alpha/2} \times se(\bar{x}_w)$ and the $t$-statistic becomes $t = (\bar{x}_w - \mu^*)/se(\bar{x}_w)$.

This issue was avoided in this illustrative example, however, by using the sampling weights to select a subset of the original ACS dataset in a way that ensured that the sample considered here is a representative sample of 30 year-old males. Similarly the female earnings data analyzed in Chapter 2 were selected in such a way as to be a representative sample.

### 6.6.3 Example: Price Inflation

Test the claim that price inflation in the United States averaged 3.0% over the period 1948 to 2012. Do the test at significance level $\alpha = .05$.

There are several commonly-used measures of price inflation. Here we use the year-to-year percentage changes in the quarterly GDP price deflator. Descriptive statistics for the dataset INFLATION are given in Table 6.5. Assuming observations are independent, the standard error $se(\bar{x}) = 2.3781/\sqrt{259} = .1478$.

Table 6.6: Descriptive Statistics: Annual inflation rate using quarterly data from 1948 to 2012.

| Variable | Obs | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Inflation | 259 | 3.3593 | 2.3781 | -2.07 | 11.09 |

The null hypothesis is $\mu = 3.0$, tested against the alternative $\mu \neq 3.0$. The $t$-statistic is

$$t = \frac{3.3593 - 3.0}{.1478} = 2.432.$$

Large absolute values of $t$ favor the alternative hypothesis, as then $\bar{x}$ is very different from 3.0. Using the $p$-value method we have

$$p = \Pr[|T_{258}| > 2.432] = 0.016.$$

Reject $H_0$ at level .05 since $p < .05$. Using the critical value method

$$c = t_{258,.025} = 1.970.$$

Reject $H_0$ at level .05 since $t = 2.432 > c = 1.970$. Therefore reject the claim that population mean inflation was 3.0% as significance level 0.05.

An important caveat in this example is that the underlying theory assumes that observations in the sample are statistically independent or unrelated with each other. In fact for time series data there can be dependence as, for example, high inflation in one quarter is likely to recur again the next quarter. Failure to control for this dependence can lead to an overestimate of the precision of estimation, i.e. the reported standard error is too small.

Time series methods introduced in Chapter 10.7, provide statistical methods that are valid even with such dependence. In this particular example there is very high dependence from one quarter to the next, and appropriate methods lead to much larger standard error of $\bar{x}$. From Chapter 10.7, allowing for this complication yields $se(\bar{x}) = .5562$, so $t = (3.3593 - 3.0)/.5562 = 0.646$. With this adjustment $H_0$ is now not rejected at level .05.

## 6.7 Key Concepts

1. The key tools of statistical inference are confidence intervals and hypothesis tests.

2. Inference for the population mean $\mu$ these are based on the $t$-statistic $t = (\bar{x} - \mu)/se(\bar{x})$. This is the distance between $\bar{x}$ and $\mu$, normalized by the standard error of the mean.

3. Under simple random sampling $t$ is a realization of the $T(n-1)$ distribution.

4. The $T(v)$ distribution, the $t$ distribution with $v$ degrees of freedom, is like a squashed version of the standard normal distribution with fatter tails. As $v \to \infty$ the t distribution goes to the standard normal.

5. $T_v$ denotes a random variable that is $T(v)$ distributed.

6. The critical value $t_{v,\alpha}$ is that value such that a $T(v)$ distributed random variable exceeds $t_{v,\alpha}$ with probability $\alpha$, i..e., $\Pr[T_v > t_{v,\alpha}] = \alpha$.

7. The critical value $t_{v,\alpha/2}$ is that value such that a $T(v)$ distributed random variable exceeds in absolute value $t_{v,\alpha}$ with probability $\alpha$, i..e., $\Pr[|T_v| > t_{v,\alpha/2}] = \alpha$.

8. A $100(1-\alpha)\%$ confidence interval for $\mu$ is $\bar{x} \pm t_{n-1,\alpha/2} \times se(\bar{x})$. This interval will include $\mu$ with probability $\alpha$.

9. It is most common to use a 95% confidence interval, so $\alpha = .05$ and $\alpha/2 = .025$. A higher degree of confidence leads to a wider confidence interval.

10. An approximate 95% confidence interval for $\mu$ is the two standard error interval $\bar{x} \pm 2se(\bar{x})$ where $se(\bar{x}) = s/\sqrt{n}$.

11. A two-sided hypothesis test is a test of $H_0 : \mu = \mu^*$ against $H_a : \mu \neq \mu^*$.

12. A type I error occurs if $H_0$ is rejected when $H_0$ is true.

13. The significance level of a test, denoted $\alpha$, is the pre-specified maximum probability of a type I error that will be tolerated.

14. The $t$-test statistic $t = (\bar{x} - \mu^*)/se(\bar{x})$ is the realization of a random variable that is approximately $T(n-1)$ distributed under $H_0 : \mu = \mu^*$.

15. The $p$-value is the probability of observing a t-test statistic at least as large in absolute value as that obtained in the current sample.

16. For a two-sided test $p = \Pr[|T_{n-1}| \geq |t|]$. $H_0$ is rejected at significance level $\alpha$ if $p < \alpha$,and is not rejected otherwise.

17. For a two-sided test the critical value $c = t_{n-1,\alpha/2}$. $H_0$ is rejected at significance level $\alpha$ if $|t| > c$, and is not rejected otherwise.

18. It is most common to test at significance level $\alpha = 0.05$.

19. A type II error occurs if $H_0$ is not rejected when $H_0$ is false.

20. Decreasing $\alpha$ decreases the chance of making a type I error but increases the chance of making a type II error.

21. It is common to use methods that keep the probability of a type II error as small as possible for a given significance level $\alpha$.

22. Key terms: confidence interval; two standard error interval; margin of error; hypothesis test; null hypothesis; alternative hypothesis; two-sided test; significance level; $t$-statistic; $p$-value; critical region; rejection region; critical value; type I error; type II error; test size; test power.

## 6.8 Exercises

1. For a random variable $T$ that is $T(22)$ distributed use a statistical package or a table of the $T(22)$ distribution to find

   (a) $\Pr[T > 2.0]$.

   (b) $\Pr[T < -2.0 \text{ or } T > 2.0]$.

   (c) $t^*$ such that $\Pr[T > t^*] = .05$.

   (d) $t^*$ such that $\Pr[T < -t^* \text{ or } T > t^*] = .05$.

2. For a standard normal distributed random variable $Z$, give the following (approximately) without using a computer or referring to a table. Hint: A normally distributed random variable lies within two standard deviations of its mean with approximate probability of 0.95.

   (a) $\Pr[Z > 2.0]$.

   (b) $\Pr[Z < -2.0 \text{ or } Z > 2.0]$.

   (c) $z^*$ such that $\Pr[Z > z^*] = .05$.

   (d) $z^*$ such that $\Pr[Z < -z^* \text{ or } Z > z^*] = .05$.

3. For a random variable $T$ that is $T(30)$ distributed, $\Pr[-1.3 < T < 1.3] = 0.80$. Using this result, derive an 80% confidence interval for the population mean $\mu$ given a random sample with $n = 51$, $\bar{x} = 40$ and $s/\sqrt{n} = 10$.

4. Consider a random variable $T$ that is $T(31)$ distributed.

   (a) Find $\Pr[|T| > 1.2]$. Show your answer on a hand-drawn graph similar to Figure 6.2.

   (b) Find $t_{.025,31}$. Show your answer on a hand-drawn graph similar to Figure 6.2.

5. Suppose a 95% confidence interval for $\mu$ is (20, 30). Do you expect a wider, narrower, or similar confidence interval in the following situations?

(a) A 99% confidence interval is constructed.

(b) The sample size is much larger.

(c) The sample mean is much larger.

(d) The standard deviation is much larger.

6. The IQ score for a random sample of 88 people has sample mean 102 and sample standard deviation 14.

   (a) Give a 95% confidence interval for the population mean IQ. [Hint: Be sure to use the standard error of the sample mean and not the standard deviation].

   (b) Perform a test at significance level 0.05 of the null hypothesis that population mean IQ equals 100 against the alternative that it does not equal 100. Use the $p$-value approach.

   (c) Repeat part b. using the critical value approach.

   (d) The claim is made that population mean IQ exceeds 100. Perform an appropriate hypothesis test at significance level 0.05 State clearly your conclusion.

7. Suppose a random sample of 25 economists forecast economic growth for the next year. The range of forecasts is from growth of -2.0 percent to growth of 3.5 percent with average 1.2 percent and with standard deviation of 2.0 percent.

   (a) Give a 95 percent confidence interval for the population mean forecast.

   (b) Test at significance level 0.05 the claim that growth will be zero next year. State the null and alternative hypotheses and your conclusion.

   (c) Test at significance level 0.05 the claim that the next year will be a recessionary year, i.e. that growth will be negative. State the null and alternative hypotheses and your conclusion.

   (d) What distributional assumptions on the underlying forecasts are needed to justify the methods used in parts a and b?

   (e) You are told that the economists sampled are all top advisers to the current President of the country. How, if at all, would the analysis in this question be effected?

8. Suppose we fail to reject $H_0$ at significance level 0.05. Do you expect it to be more likely that we reject $H_0$ in the following situations?

   (a) The test is at significance level 0.01.

(b) The sample size is much larger.

9. The dataset TDIST4 has the sample means $\bar{x}$ and corresponding standard standard deviations $s$ from 1000 random samples of size 4 where $X \sim N(100, 16^2)$.

   (a) Calculate a 95% confidence interval for $\mu$ for each of the 1000 samples.

   (b) Calculate the percentage of the 1000 confidence intervals that include $\mu = 100$. Comment.

   (c) Calculate the $t$-statistic for two-sided test of $H_0 : \mu = 100$ for each of the 1000 samples.

   (d) Calculate the percentage of the 1000 $t$-tests at significance level 0.05 that lead to rejection of $H_0$. Comment.

10. The dataset TDIST25 has the sample means $\bar{x}$ and and corresponding standard standard deviations $s$ from 1000 random samples of size 25 where $X \sim N(200, 50^2)$.

    (a) Calculate a 95% confidence interval for $\mu$ for each of the 1000 samples.

    (b) Calculate the percentage of the 1000 confidence intervals that include $\mu = 200$. Comment.

    (c) Calculate the $t$-statistic for two-sided test of $H_0 : \mu = 200$ for each of the 1000 samples.

    (d) Calculate the percentage of the 1000 $t$-tests at significance level 0.05 that lead to rejection of $H_0$. Comment.

11. The summary statistics for usual hours worked per week (variable *hours*) for a simple random sample of women aged 30 years are the following

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| *hours* | 109 | 32.81 | 19.73 | 0 | 90 |

    (a) Give a 95% confidence interval for population mean usual hours worked.

    (b) The claim is made that the population mean usual hours worked is 35 hours. Test this claim at significance level 0.01. State the null and alternative hypotheses and your conclusion.

    (c) The claim is made that the population mean usual hours worked is less than 35 hours. Test this claim at significance level 0.10. State the null and alternative hypotheses and your conclusion.

12. The dataset NAEP has scores for 51 U.S. states (including the District of Columbia) on the National Assessment of Educational Progress (NAEP) for eighth-grade mathematics for the years 2003, 2005, 2007 and 2009. Consider the change in the score for each state from 2003 to 2005 (which you need to calculate).

    (a) Generate data for the change in the score from 2003 to 2005.

    (b) Calculate a 95% confidence interval for the mean score change.

    (c) Test at significance level 0.95 the claim that there has been no change in the mean score. State the null and alternative hypotheses and your conclusion.

    (d) Test at significance level 0.05 the claim that the mean score improved. State the null and alternative hypotheses and your conclusion.

13. The dataset KNEEREPLACE has 2011 data for a number of New York hospitals on the average posted charge and average cost of knee joint replacement for cases of moderate severity, where the average is over all such cases the hospital treated in 2011.

    (a) Calculate the ratio of average posted charge to mean cost for each hospital.

    (b) Calculate a 95% confidence interval for the mean of this ratio. Comment.

    (c) Test at significance level 0.05 the claim that the mean ratio is equal to 2.5. State the null and alternative hypotheses and your conclusion.

    (d) Test at significance level 0.05 the claim that the mean ratio is less than 2.5. State the null and alternative hypotheses and your conclusion.

14. The dataset CALELECTRICITY has data for the spot price and one-day ahead forward price in the California wholesale electricity market for the one hour period 5-6 p.m. for each day from April 1 1998 to February 12 2000. Restrict analysis to days in 1998.

    (a) Generate daily data on the difference between the spot price and the one-day ahead price.

    (b) Calculate a 95% confidence interval for the mean difference.

    (c) Test at significance level 0.95 the claim that there is no difference between the spot and one-day ahead price. State the null and alternative hypotheses and your conclusion.

    (d) Test at significance level 0.05 the claim that the one-day ahead price exceeds the spot price. State the null and alternative hypotheses and your conclusion.

# Chapter 7

# © A. Colin Cameron: Statistical Inference Extensions

In some cases we wish to test whether a parameter exceeds a certain value, or is less than a certain value. For example, we may wish to test the claim that the population mean earnings exceeds \$40,000. The previous chapter considered only test of whether or not population mean earnings equal \$40,000. This chapter begins with extension of two-sided hypothesis tests to one-sided hypothesis tests.

The results presented to date consider only statistical inference on the population mean. But many other parameters may also be of interest, especially when we move on to regression in later chapters. The methods for statistical inference for the population mean extend naturally to inference on other parameters. In this chapter a general treatment is given, followed by specialization to inference on the difference between two population means.

Univariate data that appear in the media are often proportions data that come from opinion polls. For example, a poll may report that 52% of likely voters plan to vote for a Democratic candidate. In principle sampling randomness can be controlled for by directly applying the methods of Chapter 6, but in practice variations are used for this particular type of data. These are presented in this chapter.

## 7.1 One-Sided Hypothesis Tests

A one-sided hypothesis test is used to test a one-sided claim such as that the population mean earnings exceeds \$40,000. A one-sided test uses the same test statistic as a two-sided test but the $p$-value and critical value are calculated differently, leading to a different rejection region. And computer output that gives results for two-sided tests may not necessarily do so for one-sided tests.

### 7.1.1 One-Sided Hypotheses

For a **one-sided hypothesis test** care needs to be used in specifying the null and alternative hypotheses as the conclusion can differ according to which hypothesis is set up as the null and which is the alternative. As justified below, the following rule is used.

**Remark 34** *For one-sided tests the statement being tested is specified to be the alternative hypothesis. And if a new theory is put forward to supplant an old, the new theory is specified to be the alternative hypothesis.*

For example, if we wish to test the claim that the population mean earnings exceed \$40,000, we should test $H_0 : \mu \leq 40000$ against $H_a : \mu > 40000$. By contrast, to test the claim that the population mean earnings are less than \$40,000, we should test $H_0 : \mu \geq 40000$ against $H_a : \mu < 40000$.

**Remark 35** *An **upper one-tailed alternative test** is a test of $H_0 : \mu \leq \mu^*$, where $\mu^*$ is a specified value for $\mu$, against $H_a : \mu > \mu^*$. A **lower one-tailed alternative test** is a test of $H_0 : \mu \geq \mu^*$ against $H_a : \mu < \mu^*$.*

Some textbooks use a different notation for the null hypothesis in one-sided tests, defining it to be $H_0 : \mu = \mu^*$ rather than $H_0 : \mu \leq \mu^*$ (or $H_0 : \mu \leq \mu^*$). This alternative notation makes no difference to the subsequent analysis.

### 7.1.2 P-values and Critical Regions

Inference for both types of one-sided test is based on the same calculated test statistic

$$t = (\bar{x} - \mu^*)/se(\bar{x}),$$

as used for two-sided hypothesis tests. As usual this statistic is viewed as being the realization of a $T(n-1)$ distributed random variable. What differs in the one-sided case is calculation of the $p$-values and critical values.

For an **upper** one-tailed alternative test large positive values of $t$ are grounds for rejection of $H_0$, since then $\bar{x}$ (the estimate of $\mu$) is much larger than $\mu^*$. Thus the $p$-value is the probability of being in the upper tail of the $T(n-1)$ distribution, so $p = \Pr[T_{n-1} \geq t]$. And the critical region for a test at significance level $\alpha$ is $t > c$ where $c$ is such that $\Pr[T_{n-1} > c] = \alpha$. From Chapter 6.2 this is denoted $c = t_{n-1,\alpha}$. $H_0$ is rejected at significance level $\alpha$ if $p < \alpha$ or, equivalently, if $t > c$.

For a **lower** one-tailed alternative test large negative values of $t$ lead to rejection of $H_0$, since then $\bar{x}$ is much smaller than $\mu^*$. Now the $p$-value is the probability of being in the lower tail of the $T(n-1)$ distribution, so $p = \Pr[T_{n-1} \leq t]$. And the critical region for a

test at significance level $\alpha$ is $t < c$ where $c$ is such that $\Pr[T_{n-1} < c] = \alpha$. From Chapter 6.2 this is denoted $c = -t_{n-1,\alpha}$. $H_0$ is rejected at significance level $\alpha$ if $p < \alpha$ or, equivalently, if $t < c$.

**Remark 36** *Let $t$ be the usual $t$-test statistic. For an upper one-tailed alternative test the $p$-value is $p = \Pr[T_{n-1} \geq t]$, the critical value at significance level $\alpha$ is $c = t_{n-1,\alpha}$, and we reject $H_0$ if $p < \alpha$ or, equivalently, if $t > c$. For a lower one-tailed alternative test the $p$-value is $p = \Pr[T_{n-1} \leq t]$, the critical value at significance level $\alpha$ is $c = -t_{n-1,\alpha}$, and we reject $H_0$ if $p < \alpha$ or, equivalently, if $t < c$.*

It is most common to test at significance level 0.05. Consider an upper one-tail alternative test. For large $n$ the critical $c = t_{\infty,.05} = 1.645$, while for $n = 31$, a small $n$, $t_{n-1,.05} = 1.697$. Suppose $t = 1.8$. Then for most sample sizes an upper one-tailed test will reject at significance level .05, since $t > c$. By contrast a two-sided test at significance level .05 will not reject, since then the critical value $c = t_{.025,n-1} \leq 1.960$ so $|t| < c$.

More generally, the one-sided test requires less evidence to reject the null hypothesis, provided the test statistic $t$ is in the correct tail of the distribution.

### 7.1.3 Example: Mean Annual Earnings

Suppose we wish to evaluate the claim that the population mean exceeds \$40,000, again using the female earnings data studied in Chapter 6. A test of this claim is implemented as a test of $H_0 : \mu \leq 40000$ against $H_a : \mu > 40000$. This is an example of an upper one-tailed alternative test.

The $t$-statistic has already been calculated in Chapter 6.4, with $t = 0.724$. Large positive values of $t$ support rejection of $H_0$ since then $\bar{x}$ is much greater than the hypothesized population mean value of \$40,000. Absolute values are not relevant for this one-tailed test since a negative value of the $t$-statistic automatically does not favor rejection of $H_0 : \mu \leq 40000$.

The $p$-value, the probability that a $t$ distributed random variable exceeds the observed $t$ value of 0.724, is

$$p = \Pr[T_{170} \geq .724] = 0.235.$$

Since $p$ is larger than 0.05, we do not reject $H_0$ at significance level 0.05.

The shaded region in the left panel of Figure 7.1 gives the probability that $T_{170} \geq .724$. Since this region has area 0.235 that exceeds 0.05, we do not reject $H_0$ at significance level 0.05.

Using the alternative equivalent critical value method instead, the critical value $c$ solves $\Pr[T_{170} \geq c] = 0.05$ if testing is at the significance level 0.05. Then

$$c = t_{170,.05} = 1.654.$$

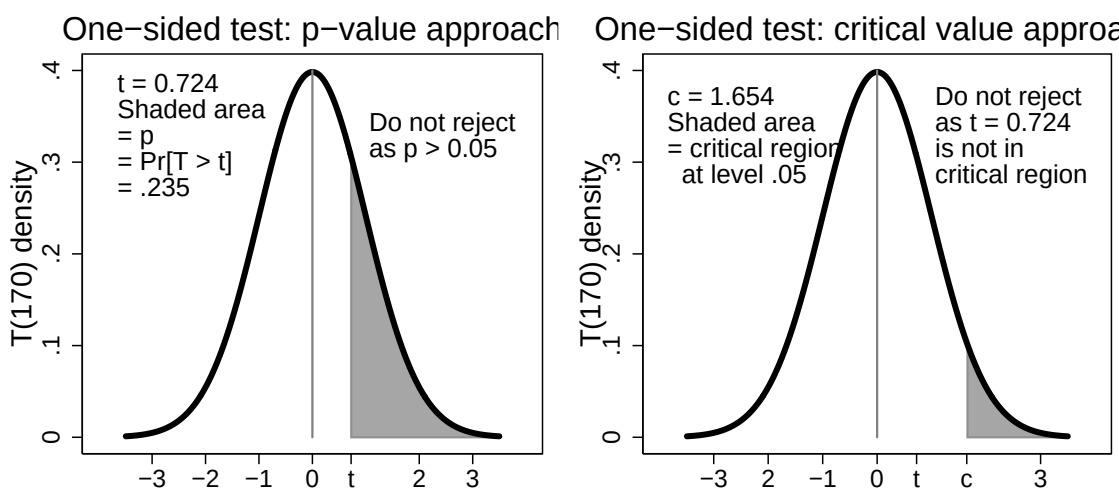


Figure 7.1: One-sided hypothesis test (upper one-tailed alternative) when $t = 0.724$, $n = 171$ and $\alpha = 0.05$.

We do not reject $H_0$ at significance level 0.05, since $t = 0.724 \leq 1.654$.

This shaded region in the right panel of Figure 7.1 is the critical region. This is $t$-statistics in excess of $c = 1.654$ since there is probability 0.05 of a $T(170)$ random variable exceeding 1.654. Since $t = 0.724$ does not fall in this critical region we do not reject $H_0$ at significance level 0.05.

Using either method we do not reject at significance level 0.05 the hypothesis that the population mean earnings is less than or equal to \$40,000. The claim that mean earnings exceed \$40,000 is not supported.

### 7.1.4 Specifying the Null Hypothesis for One-Sided Tests

Suppose the claim is made that population mean earnings are more than \$40,000. Should we perform an upper one-tailed alternative test or a lower one-tailed alternative test?

There are two potential ways to proceed, though only the first should be used as explained in what follows.

1. Test $H_0 : \mu \leq 40000$ against $H_a : \mu > 40000$.

2. Test $H_0 : \mu \geq 40000$ against $H_a : \mu < 40000$.

Suppose we take the first approach. The claim that $\mu > 40000$ is supported if $H_0$ is rejected. Rejection of $H_0$ requires a sample mean of considerably more than 40000, say

$\bar{x} > 43000$ (it can be shown that we need $\bar{x} > 43000$ if $t_{n-1,\alpha} \times s/\sqrt{n} = 3000$). So the claim that $\mu > 40000$ is supported if $\bar{x} > 43000$.

Suppose instead the second approach is taken. Then the claim is supported if $H_0$ is not rejected. Rejection of $H_0$ requires a sample mean of considerably less than 40000, say $\bar{x} < 37000$ (if again $t_{n-1,\alpha} \times s/\sqrt{n} = 3000$). Non-rejection of $H_0$ then occurs if $\bar{x} > 37000$. Thus the claim that $\mu > 40000$ is supported if $\bar{x} > 37000$.

To summarize, the claim is that mean earnings exceed \$40,000. The first specification of the null and alternative hypotheses leads to support of the claim if the sample average exceeds \$43,000, while the second specification leads to support of the claim if the sample average exceeds \$37,000. The philosophy of hypothesis testing is to require strong evidence to support a claim. The first specification is therefore used, with the claim made specified as the alternative hypothesis.

**Remark 37** *For one-sided tests the claim being tested is specified to be the alternative hypothesis, as stronger evidence is then needed to support the claim than if the claim was set up as the null hypothesis.*

There can be considerable debate as to which hypothesis should be the null. For example, suppose we wish to determine whether women at a workplace have been discriminated against, a not unusual issue to be determined in court. One approach is to specify the alternative hypothesis to be that women are paid less than men (the claim made) while another approach is to specify this as the null hypothesis. Lawyers for the employer may favor the first approach, lawyers for the employee may favor the second approach, and the statistical methodology presented here selects the first approach.

## 7.2 Generalization of Confidence Intervals and Hypothesis Tests

Confidence intervals and hypothesis tests can be applied to parameters other than the population mean $\mu$. Leading examples include the difference in two means $(\mu_2 - \mu_1)$, presented in Chapter 7.3, and the slope of a regression line, the subject of many later chapters.

The approach of Chapter 6 for inference on $\mu$ extends easily to such settings.

### 7.2.1 Generalizations of Confidence Intervals

Recall that inference on $\mu$ is based on $t = (\bar{x} - \mu)/se(\bar{x})$. In words, the $t$-statistic equals the estimate minus the parameter divided by the standard error, where the standard error measures how precisely the parameter has been estimated. More generally we have the following result.

**Remark 38** *For the models analyzed in this book, for sufficiently large sample size the statistic*

$$t = \frac{estimate - parameter}{standard\ error}$$

*is the sample realization of a $T(v)$ distributed random variable, where the degrees of freedom $v$ for the $t$ distribution varies with the application, and the standard error is the estimated standard deviation of the estimate.*

From Chapter 6, a $100(1-\alpha)\%$ confidence interval for $\mu$ is $\bar{x} \pm t_{v,\alpha/2} \times se(\bar{x})$. This generalizes as follows.

**Remark 39** *A $100(1-\alpha)\%$ confidence interval for the unknown parameter is*

$$estimate \pm t_{v,\alpha/2} \times standard\ error.$$

The most commonly-used confidence level is 95 percent and, as discussed, $t_{v,.025} \simeq 2$ for $v > 30$. This immediately leads to the following simple rule-of-thumb.

**Remark 40** *An approximate 95% confidence interval for the unknown parameter is the two-standard error interval*

$$estimate \pm 2 \times standard\ error.$$

The term **margin of error** is used to describe the half-width of a confidence interval, or $t_{v,\alpha/2} \times se(\cdot)$. The term is most often used in the context of 95% confidence intervals, since these are the most commonly-used confidence intervals. Then since $t_{v,.025} \simeq 2$,

$$\text{Margin of error} = 2 \times \text{Standard error}.$$

As an example of the above, suppose we estimate a parameter $\theta$, where $\theta$ is the Greek letter theta. Our sample estimate of $\theta$ is 11 with standard error of the estimate equal to 3, and the sample size is large. Then an approximate 95% confidence interval for $\theta$ is $11 \pm 2 \times 3$ or $(5, 17)$. Since the standard error is 3, the margin of error is said to be $2 \times 3 = 6$, or $\pm 6$.

### 7.2.2 Generalizations of Hypothesis Tests

For hypothesis testing we again use the more general form of the $t$-statistic, assumed to be approximately $T(v)$ distributed.

**Remark 41** *Consider a two-sided test at significance level $\alpha$ of the null hypothesis $(H_0)$ that a parameter equals a hypothesized value against the alternative hypothesis $(H_a)$ that it does not. Calculate the $t$-statistic*

$$t = \frac{estimate - hypothesized\ parameter\ value}{standard\ error}.$$

*Under the null hypothesis $t$ is the sample realization of a $T(v)$ distributed random variable. The p-value approach is to reject $H_0$ if $p < \alpha$ where $p = \Pr[|T_v| > t]$. The critical value approach is to reject $H_0$ if $|t| > c$ where $c = t_{v,\alpha/2}$ satisfies $\Pr[T_v > t_{v,\alpha/2}] = \alpha$. The two methods lead to the same conclusion.*

To repeat, in Chapter 6 the estimate is $\bar{x}$, the parameter is $\mu$, the standard error is $se(\bar{x}) = s/\sqrt{n}$, and $v = n - 1$.

For both confidence intervals and hypothesis tests, the standard normal distribution is sometimes used rather than the $T(v)$ distribution.

## 7.3 Difference in Two Means

Suppose we have two independent random samples on the same measure, and want to test whether the population mean is the same in the two samples. For example, we might observe that the average annual earnings in a sample of men exceeds the average annual earnings in a sample of men. A test of whether this implies that population mean earnings of men differs from that for women needs to control for sampling variability.

### 7.3.1 Statistical Inference for Difference in Two Means

Let subscript 1 denote the first sample and subscript 2 denote the second sample, where simple random sampling is assumed. The first sample of size $n_1$ is based on the random variable $X_1 \sim (\mu_1, \sigma_1^2)$ and yields sample mean $\bar{x}_1$ and sample variance $s_1^2$. The second sample of size $n_2$ is based on the random variable $X_2 \sim (\mu_2, \sigma_2^2)$ and yields sample mean $\bar{x}_2$ and sample variance $s_2^2$.

Interest lies in estimating $\mu_1 - \mu_2$, the **difference in population means**. The obvious estimator is the difference in sample means $\bar{x}_1 - \bar{x}_2$. In general we base inference on $t =$ (estimate−parameter)/(standard error). Here

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{se(\bar{x}_1 - \bar{x}_2)} \sim T_{n_1+n_2-2},$$

where from Appendix 7.A

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{se(\bar{x}_1)^2 + se(\bar{x}_2)^2}.$$

The $T_{n_1+n_2-2}$ distribution is used as two degrees of freedom are used in estimating both $s_1^2$ and $s_2^2$, so the remaining degrees of freedom are $n_1 + n_2 - 2$.

This result leads immediately to the following confidence interval.

**Remark 42** *A* $100(1-\alpha)\%$ *confidence interval for the difference in two means* $\mu_1 - \mu_2$ *is*

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2} \times se(\bar{x}_1 - \bar{x}_2).$$

The corresponding test of whether or not the two means are equal is the following.

**Remark 43** *A two-sided test of the difference in two means is a test of* $H_0 : \mu_1 - \mu_2 = 0$ *against* $H_a : \mu_1 - \mu_2 \neq 0$. *We reject* $H_0$ *at level* $\alpha$ *if the t-statistic*

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{se(\bar{x}_1 - \bar{x}_2)}$$

*if* $p < \alpha$ *where* $p = \Pr[|T_{n_1+n_2-2}| > |t|] < \alpha$, *or if* $|t| > c$ *where* $c = t_{n_1+n_2-2;\alpha/2}$.

Statistical packages often provide the option to implement two variations on this method. First, in the case assumed here with different population variances in the two samples, a more precise approximation, called Satterthwaite's formula, is used to obtain the degrees of freedom for the $t$ distribution. Second, if instead it is assumed that the variances are equal, so $\sigma_1^2 = \sigma_2^2$, then an alternative formula for $se(\bar{x}_1 - \bar{x}_2)$ may be used. In economics applications it is almost always assumed that $\sigma_1^2 \neq \sigma_2^2$.

Comparison of more than two means can be done using **analysis of variance** (ANOVA) techniques that are used extensively in branches of applied statistics that do not regularly use regression methods.

While statistical software includes specialized commands for test of the difference in means, in econometrics the test is implemented as a special case of regression; see Chapter 12.2.

### 7.3.2 Example: Difference in Means of Male and Female Earnings

We again consider data on annual earnings for 30 year-old full-time workers. Table 7.1 presents data for the sample of 171 women from dataset EARNINGS introduced in Chapter 2, and for similar data for a sample of 191 men from dataset EARNINGSMALE introduced in Chapter 6.6. The combined data are in dataset EARNINGSBOTH.

In terms of the preceding notation $n_1 = 171$, $\bar{x}_1 = 41413$, $s_1 = 25527$, $se(\bar{x}_1) = s_1/\sqrt{n_1} = 1952$, $n_2 = 191$, $\bar{x}_2 = 52345$, $s_2 = 65035$ and $se(\bar{x}_2) = s_2/\sqrt{n_2} = 4706$.

The estimated difference in means is $\bar{x}_1 - \bar{x}_2 = -10,941$, so on average female 30-year-old full-time workers annually earn \$10,941 less than males. From Table 7.1 the higher mean male earnings arises because the male earnings distribution is much more right skewed than that for women. The lower quartiles and medians for the two samples are similar, but the upper quartile is \$10,000 higher for men and the skewness coefficient is 5.34 for men compared to 1.71 for women. And more detailed inspection of the data reveals that five men earned more than \$172,000, the highest earnings of any woman in the sample.

Table 7.1: Summary statistics: Annual earnings of female and male full-time workers aged 30 in 2010.

| **Statistic** | **Female** | **Male** |
|---|---|---|
| Sample size ($n$) | 171 | 191 |
| Mean ($\bar{x}$) | 41,413 | 52,354 |
| Standard deviation ($s$) | 25,527 | 65,035 |
| Standard error of the mean ($s/\sqrt{n}$) | 1,952 | 4,706 |
| Minimum | 1,050 | 1 |
| Maximum | 172,000 | 498,000 |
| Upper quartile (75th percentile) | 50,000 | 60,000 |
| Median (50th percentile) | 36,000 | 38,000 |
| Lower quartile (25th percentile_ | 25,000 | 25,000 |
| Skewness | 1.71 | 5.35 |
| Kurtosis | 7.32 | 35.76 |

Interest lies in knowing whether the observed difference arises solely due to sampling randomness, or whether there really is a gender difference in the population mean earnings. Here $se(\bar{x}_1)^2 + se(\bar{x}_2)^2 = 1952^2 + 4706^2 = 25956740$, and taking the square root yields the standard error $se(\bar{x}_1 - \bar{x}_2) = 5095$. Using $t_{360,.025} = 1.967$, since $n_1 + n_2 - 2 = 360$, a 95% confidence interval for the mean difference between female and male earnings is

$$-10941 \pm 1.967 \times 5095 = -10941 \pm 10021 = (-20962, -920).$$

The 95% confidence interval is (-\$20,962, -\$920). This confidence interval is quite broad, a consequence of the large variation in earnings leading to noisy estimates of mean earnings, even given 362 observations.

Is the difference statistically significant? A big decision to make at this stage is whether to use a one-sided or two-sided test.

If we just want to test whether the difference in means is statistically significant at level $\alpha$ then a two-sided test of $\mu_1 - \mu^* = 0$ against $H_a : \mu_1 - \mu^* \neq 0$ is warranted. Here

$$t = (\bar{x}_1 - \bar{x}_2)/se((\bar{x}_1 - \bar{x}_2)) = -10941/5095 = -2.147.$$

Since $p = \Pr[|T_{360}| < |-2.147|] = 0.032 < 0.05$ we reject $H_0$ at level 0.05. Equivalently, since $|t| = 2.147 < c = t_{360,.025} = 1.967$ we reject $H_0$ at level 0.05. The conclusion is that population mean earnings for women differ from those for men.

The usual prior belief is that women earn less than men, and this is often a reason for performing statistical analysis of earnings differences by gender. In that case a one-sided test may instead be appropriate. Making the claim that women earn less than men on average the alternative, we perform a one-sided test of $H_0 : \mu_1 - \mu^* \geq 0$ against $H_a : \mu_1 - \mu^* < 0$.

Again $t = -2.147$. Now, however, $p = \Pr[T_{360} < -2.147] = 0.016$. Since $p < 0.05$ we reject $H_0$ at level 0.05. Equivalently, since $t = -2.147 < c = -t_{360,.05} = -1.649$ we reject $H_0$ at level 0.05. The conclusion is that population mean earnings for women are less than those for men.

Note that for the one-sided test the $p$-value halved, from 0.032 to 0.016, making it easier to reject the null hypothesis and find support for the claim.

## 7.4 Proportions Data

Proportions data are data on the fraction of times that a given event occurs. Examples are unemployment rates, employment rates and the fraction intending to vote for a given political candidate. Given survey data we wish to extrapolate from the sample proportion to the population proportion.

Statistical packages often have procedures specific to proportions data. It is important to note that different statistical packages can give slightly different confidence intervals and $p$-values for hypothesis tests depending on what formula is used to compute the standard error, whether the standard normal or $T(n-1)$ distribution is used, and whether "small sample" methods, not covered here, are used in place of the methods given below.

### 7.4.1 Statistical Inference for Proportions Data

For **proportions data** the underlying random variable $X$ for each surveyed individual is viewed as taking value 1 with probability $p$ and value 0 with probability $1-p$. For example, we may let $X = 1$ if the person intends to vote Democrat, so then $p$ is the unknown population probability of voting Democrat. This is the **Bernoulli distribution** with population mean $\mu = p$ and population variance $\sigma^2 = p(1-p)$; see Appendix B.2.

Suppose we have a sample of size $n$, where the $i^{th}$ observation $x_i$ equals 1 if the $i^{th}$ interviewed person says they intend to vote Democrat, and equals 0 if this is not the case. We assume the sample of size $n$ is a simple random sample.

The sample mean $\bar{x}$ is then the fraction or proportion of those interviewed who intend to vote Democrat. The sample mean $\bar{x}$ is an unbiased estimate of the population mean $p$.

Proportions data are a very special case as the underlying random variable takes only two values. This leads to several different ways to estimate the standard error of $\bar{x}$, including the general method presented in Chapter 5.4. As a result different statistical packages with specialized commands for proportions data may compute the standard error of the sample mean in different ways, leading to somewhat different confidence intervals and different results for hypothesis tests. Additionally confidence intervals and hypothesis tests for proportions data most often use the standard normal distribution rather than the $T(n-1)$ distribution. This

will make little difference as usually $n$ is quite large in economics applications with proportions data.

Here we present the most commonly-used method for computing confidence intervals and for performing hypothesis tests. Further details are given in Appendix 7.B.

For confidence intervals it is most common to use $se(\bar{x}) = \sqrt{\bar{x}(1-\bar{x})/n}$ and the standard normal distribution.

**Remark 44** *A $100 \times (1-\alpha)\%$ confidence interval for $p$ is most often*

$$\bar{x} \pm z_{\alpha/2} \times \sqrt{\bar{x}(1-\bar{x})/n}.$$

For hypothesis tests on $p$ it is most common to use $se(\bar{x}) = \sqrt{p^*(1-p^*)/n}$, which replaces the estimate $\bar{x}$ of $p$ by the hypothesized value $p^*$, and the standard normal distribution.

**Remark 45** *A two-sided hypothesis test of $H_0 : p = p^*$ against $H_a : p \neq p^*$ is based on the statistic*

$$z = \frac{\bar{x} - p^*}{\sqrt{p^*(1-p^*)/n}} \sim N(0,1).$$

*We reject $H_0$ at level $\alpha$ if the $p = \Pr[|Z| > |z|] < \alpha$, or if $|z|$ exceeds the critical value $c = z_{\alpha/2}$.*

For both approaches the approximation is felt to be good if both $np > 10$ and $n(1-p) > 10$. Thus for low $p$ or high $p$ considerably more than 30 observations are needed. In most economics applications with proportions data there are many observations, so the approximation will be good.

If this is not the case, one can instead use exact statistical inference on $p$ based on the binomial distribution. Some statistical packages provide commands to calculate exact (or "small sample") confidence intervals and hypothesis tests; these are not presented here.

### 7.4.2 Example: Voting Intentions

Suppose we have a random sample of 921 voters of whom 480 intend to vote Democrat and 441 who intend to vote Republican. For statistical inference we use the large sample approach adapted to the special case of proportions data.

For these data

$$\bar{x} = [480 \times 1 + 441 \times 0]/921 = 0.5212.$$

For confidence intervals we use as standard error of the mean

$$se(\bar{x}) = \sqrt{0.5212(1-0.5212)/921} = 0.01646.$$

For the standard normal, $z_{.025} = 1.960$. So a 95 percent confidence interval for the population mean proportion of Democrat voters is

$$0.521 \pm 1.964 \times 0.01646 = 0.5212 \pm 0.0323 = (0.4889, 0.5535).$$

Converting to percentage of votes, a 95 percent confidence interval for the population percentage of voters who intend to vote Democrat is (48.9%, 55.3%).

Now suppose we wish to test the belief that the Democrat candidate will win the election. This is a one-sided test with the belief specified to be the alternative hypothesis. Thus we test $H_0 : p \leq 0.5$ against $H_a : p > 0.5$. For hypothesis testing we use $se(\bar{x}) = \sqrt{0.5 \times (1 - 0.5)/921}$, so the test statistic

$$z = \frac{0.5212 - 0.5}{\sqrt{0.5 \times (1 - 0.5)/920}} = 1.287,$$

though as noted there are other ways to estimate $se(\bar{x})$. The $p$-value is $p = \Pr[|Z| > 1.287] = 0.198$. Since the $p$-value exceeds 0.05 we do not reject the null hypothesis at significance level 0.05. Alternatively, the critical value $c = z_{.025} = 1.964$, and we do not reject $H_0$ at level .05 since $|z| = 1.287 < c$. We cannot conclude that the Democrat candidate will win the election.

### 7.4.3 Margin of Error

The term **margin of error** when used in the media refers to the half width of a 95% confidence interval for the mean proportion. The margin of error is approximately two times the standard error. So if a poll finds that 45% favor a particular policy with a margin of error of 4% then a 95% confidence interval for the population mean percentage in favor of the policy is $(41\%, 49\%)$.

An important question is how many people need to be polled to achieve a desired margin of error. For proportions data $\bar{x}(1 - \bar{x}) \leq 0.25$ since $\bar{x}(1 - \bar{x})$ takes a maximum value of 0.25 when $\bar{x} = 0.5$. It follows that $se(\bar{x}) = \sqrt{\bar{x}(1 - \bar{x})/n} \leq \sqrt{0.25/n} = 0.5/\sqrt{n}$. Hence for a 95% confidence interval the margin of error, approximately equal to $2 \times se(\bar{x})$, is less than $\sqrt{1/n}$.

**Remark 46** *For a 95% confidence interval on the population mean percentage* $100p\%$*, the margin of error is at most* $(100/\sqrt{n})\%$.

Published opinion polls typically interview between 600 and 2,000 people. Then the corresponding maximum margin of error is, respectively, 4.1% and 2.2% since, for example, $100/\sqrt{600} \simeq 4.1$. For many purposes these margins of error are tolerable, but they will be too large for predicting the result of a close election.

## 7.5 Key Concepts

1. For one-sided hypothesis tests the claim being made is specified as the alternative hypothesis.

2. An upper one-tailed alternative test is a test of $H_0 : \mu \leq \mu^*$ against $H_a : \mu > \mu^*$. We reject $H_0$ at level $\alpha$ if $p = \Pr[T_{n-1} \geq t]$ is less than $\alpha$. Equivalently reject if $t > c = t_{n-1,\alpha}$.

3. A lower one-tailed alternative test is a test of $H_0 : \mu \geq \mu^*$ against $H_a : \mu < \mu^*$. We reject $H_0$ at level $\alpha$ if $p = \Pr[T_{n-1} \leq t]$ is less than $\alpha$. Equivalently reject if $t < c = -t_{n-1,\alpha}$.

4. For one-sided tests at significance level 0.05 a rough guide is to use as critical value 1.645 for an upper one-tail alternative and $-1.645$ for a lower one-tail alternative, as $t_{n-1,.025} = 1.645$ for large $n$.

5. In many settings the statistic $t$ = (estimate−parameter)/(standard error) can be viewed as the realization of a $T(v)$ distributed random variable where the degrees of freedom $v$ varies with the application.

6. An approximate 95% confidence interval for a parameter is then the estimate plus or minus two times the standard error.

7. The half-width of a confidence interval is called the margin of error.

8. The margin of error for a 95% confidence interval is approximately two times the standard error.

9. A test of the difference in two means is a test of whether or not $\mu_1 - \mu_2 = 0$ where $\mu_1$ and $\mu_2$ denote the population means in the two populations.

10. While statistical software includes specialized commands for test of the difference in means, in econometrics the test is implemented as a special case of regression.

11. Proportions data are data on the fraction of times that a given event occurs.

12. While these data can be analyzed by directly using the results of Chapter 6, statistical packages often have procedures specific to proportions data that lead to slightly different results.

13. Key terms: one-sided test; upper one-sided alternative; lower one-sided alternative; two standard error interval; margin of error; difference in means; proportions data; polling data; voting intentions.

## 7.6 Exercises

1. A statistical package gives the following output:

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      85    62.46812    10.07956    92.92892    42.42382    82.51242
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   1.7330
Ho: mean = 45                                    degrees of freedom =       84

    Ha: mean < 45               Ha: mean != 45                 Ha: mean > 45
 Pr(T < t) = 0.9566         Pr(|T| > |t|) = 0.0868          Pr(T > t) = 0.0434
```

   (a) Perform a test of whether or not $\mu = 45$ at significance level 0.05.

   (b) The claim is made that the population mean exceeds 45. Perform an appropriate test at significance level 0.05.

   (c) Calculate a 90% confidence interval for $\mu$.

   (d) Perform a two-sided test of whether or not $\mu = 40$.

## 7.7 Appendix 7.A: Standard Error for Difference in Means

Since $\bar{X}_1 \sim (\mu_1, \sigma_1^2/n_1)$ and $\bar{X}_2 \sim (\mu_2, \sigma_2^2/n_2)$, if the two samples are independent

$$\bar{X}_1 - \bar{X}_2 \sim ((\mu_1 - \mu_2),\ (\sigma_1^2/n_1) + (\sigma_2^2/n_2)).$$

This uses the property that the mean of the difference is the difference in means, while the variance of the difference of two independent random variables is the sum of the variances; see Appendix B. The variance $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$ can be estimated by $s_1^2/n_2 + s_1^2/n_2$. Taking the square root

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{se(\bar{x}_1)^2 + se(\bar{x}_2)^2}.$$

## 7.8 Appendix 7.B: Standard Error for Proportions Data

The most obvious way to proceed is to calculate the usual standard error of the mean given in Chapter 5.4. Simplification occurs because, for data on $x_i$ that take only values 0 or 1 it can be shown that $\sum_{i=1}^n (x_i - \bar{x})^2 = n\bar{x}(1 - \bar{x})$. It follows that the sample variance of $X$ is $s^2 = n\bar{x}(1 - \bar{x})/(n - 1)$, so the usual formula for the standard error of $\bar{x}$ yields

$$se(\bar{x}) = s/\sqrt{n} = \sqrt{\bar{x}(1 - \bar{x})/(n - 1)}.$$

However, it is more common to instead use the following variation that specializes to the Bernoulli distribution – random variables that can take only two values are necessarily Bernoulli distributed. For a Bernoulli random variable the population variance $\sigma^2 = p(1-p)$, so dividing by the sample size the variance of the sample mean is $p(1-p)/n$. Since $p$ can be estimated by $\bar{x}$, the variance of the sample mean can be simply estimated by $s^2 = \bar{x}(1-\bar{x})/n$. This leads to an alternative estimate of the standard error of $\bar{x}$

$$se_{Bernoulli}(\bar{x}) = \sqrt{\bar{x}(1-\bar{x})/n}.$$

This is very close to $se(\bar{x})$ since the only change is to divide by $n$ rather than $n-1$ and typically $n$ is very large when proportions data are used.

For hypothesis tests on $p$ it is standard to estimate the variance of $\bar{x}$ using the additional knowledge that $p = p^*$ under $H_0$. Then the variance of the sample mean $p(1-p)/n$ is estimated by $p^*(1-p^*)/n$ rather than $\bar{x}(1-\bar{x})/n$. Then

$$se_{H_0}(\bar{x}) = \sqrt{p^*(1-p^*)/n}.$$

This measure can differ quite a bit from the first two measures, depending on how much the sample mean $\bar{x}$ differs from the hypothesized $p^*$.

# Chapter 8

# © A. Colin Cameron: Bivariate Data Summary

We now summarize the relationship between two variables, such as that between earnings and education or between house price and house size, rather than analyzing just one variable in isolation.

Without knowledge of the particular variables being analyzed, the two variables should be treated equally. In practice one variable is often viewed as being explained by another variable. The standard notation used follows that of mathematics, where $y$ is a function of $x$. Thus the variable $y$ is viewed as being explained by the variable $x$. It is important to note, however, that without additional information the roles of the two variables may in fact be reversed, so that it is $x$ that is being explained by $y$. It is safest to say that we measure the association between variable $y$ and variable $x$.

This chapter presents data summary using cross tabulation, a two-way scatter plot, the correlation coefficient and least squares regression. Statistical inference for regression is deferred to Chapters 9 and 10.

## 8.1 Example: House Price and Size

A common real-estate refrain is that the most important determinant of house price is "location, location, location." The data analyzed here already control for location by restricting analysis to houses within a small homogeneous community. The houses are of similar vintage and in the same school district. Given this control for location, it is reasonable to believe that house size will be the key determinant of house price.

Dataset HOUSE includes data on the price (in dollars) and size (in square feet) of 29 houses sold in a small homogeneous community. These data are used extensively in the next few chapters.

Table 8.1: House price and size: Complete listing of data.

| House Price | House Size | House Price | House Size | House Price | House Size |
|---|---|---|---|---|---|
| 375,000 | 3,300 | 255,000 | 1,500 | 235,000 | 1,700 |
| 340,000 | 2,400 | 253,000 | 2,100 | 233,000 | 1,700 |
| 310,000 | 2,300 | 249,000 | 1,900 | 230,000 | 2,100 |
| 279,900 | 2,000 | 245,000 | 1,400 | 229,000 | 1,700 |
| 278,500 | 2,600 | 244,000 | 2,000 | 224,500 | 2,100 |
| 273,000 | 1,900 | 241,000 | 1,600 | 220,000 | 1,600 |
| 272,000 | 1,800 | 239,500 | 1,600 | 213,000 | 1,800 |
| 270,000 | 2,000 | 238,000 | 1,900 | 212,000 | 1,600 |
| 270,000 | 1,800 | 236,500 | 1,600 | 204,000 | 1,400 |
| 258,500 | 1,600 | 235,000 | 1,600 | | |

Table 8.2: House price and size: Summary statistics.

| Statistic | House Price (in dollars) | House Size (in square feet) |
|---|---|---|
| Mean | 253,910 | 1,883 |
| Standard deviation (of $x$) | 37,391 | 398 |
| Standard error (of $\bar{x}$) | 6,943 | 74 |
| Maximum | 375,000 | 3,300 |
| Median (50th percentile) | 244,000 | 1,800 |
| Minimum | 204,000 | 1,400 |
| Skewness | 1.56 | 1.73 |
| Kurtosis | 5.61 | 6.74 |

The complete listing of the data in Table 8.1 orders the observations by decreasing price, making interpretation easier. It does appear that larger houses are higher priced. For example, the five most expensive houses are all 2,000 square feet or more, while the four cheapest houses are all less than 2,000 square feet in size. The goal of this chapter is to better quantify this relationship.

Table 8.2 presents various summary statistics for these 29 observations. The house sale price ranges from \$204,000 to \$375,000 with mean \$253,910 and standard deviation \$37,391. (By comparison in December 2014 existing single-family homes in the U.S. sold on average for \$256,600). House size ranges from 1,400 to 3,300 square feet with mean 1,883 square feet and standard deviation 398 square feet.

Table 8.3: House price and size: Cross tabulation with row percentages.

| | Size range | | | |
|---|---|---|---|---|
| **Price range** | **Small** | **Medium** | **Large** | **Total** |
| **Low** | 11 | 6 | 0 | 17 |
| | 64.71 | 35.29 | 0.00 | 100.00 |
| **High** | 2 | 7 | 3 | 12 |
| | 16.67 | 58.33 | 25.00 | 100.00 |
| **Total** | 13 | 13 | 3 | 29 |
| | 44.83 | 44.83 | 10.34 | 100.00 |

## 8.2 Two-way Tabulation

A **two-way tabulation** or **cross tabulation** of the two variables $x$ and $y$ is a two-way frequency table that lists the number (or fraction) of observations equal to each of the distinct values taken by the pair $(x, y)$. This generalizes a frequency table for a single variable that tabulates the distinct values taken by a variable along with the number of observations or the fraction of the observations equal to that value; see Chapter 2.

A cross tabulation is easiest to read if the variables $x$ and $y$ take relatively few values. These values can be for data that are intrinsically categorical. For example, we might cross tabulate voting intentions (Democrat, Republican or Other) with race (White, Hispanic, ...). Or one or both variables may be numerical.

If either variable takes too many values then it is best to first aggregate the variable into a limited number of values. As an example, consider the house price and size data. Define the new variable *Pricerange* that takes a low value if *Price*<\$250,000 and a high value if *Price*≥\$250,000. And define the new variable *Sizerange* that takes the value small if *Size*<1,800, the value medium if 1,800≤*Size*<2,400, and the value large if *Size*≥2,400.

Table 8.3 presents the two-way tabulation for these aggregated variables. The main entry is the number of observations with a given combination of price and size. For example, there were 11 houses that were of low price and small size. The table also includes **row sums** and **column sums**. For example in the row for low price range there are a total of $11+6+0 = 17$ observations. So 17 of the 29 houses sold were in the low price range.

The table includes a second optional entry, a **row percentage**, that for each value of *Pricerange* gives the percentage of observations in each of the size ranges. For example, for low-priced houses 11 out of 17 are small, or $100 \times 11/17 = 64.71\%$. The row percentages indicate that most low-priced were small (64.71%), while most high-priced houses were medium or large (58.33% + 25.00% = 83.33%). The table could also include similarly constructed **column percentages**.

A two-way tabulation can also include **expected frequencies**, created under the as-

Table 8.4: House price and size: Cross tabulation with expected frequencies.

| | Size range | | | |
|---|---|---|---|---|
| **Price range** | **Small** | **Medium** | **Large** | **Total** |
| **Low** | 11 | 6 | 0 | 17 |
| | 7.62 | 7.62 | 1.76 | 17.00 |
| **High** | 2 | 7 | 3 | 12 |
| | 5.38 | 5.38 | 1.24 | 12.00 |
| **Total** | 13 | 13 | 3 | 29 |
| | 13.00 | 13.00 | 3.00 | 29.00 |

sumption that the two variables are statistically independent. If two variables are statistically independent then the joint probability of occurrence equals the product of each individual probability of occurrence. For example, the joint probability of low-price and small-size then equals $(17/29) \times (13/29)$. Multiplying by 29 gives the expected frequency of $17 \times 13/29 = 7.62$. This is less than the observed frequency of 11.

Table 8.4 presents both the observed frequencies and the expected frequencies. More of the low-price houses are small then would be expected if price and size were independent (11 versus 7.62) and fewer of the high-price houses are small then would be expected if price and size were independent (2 versus 5.38).

The difference between observed and expected frequency values is the basis for Pearson's chi-squared goodness-of-fit test of statistical independence of two categorical variables. This test is presented in Appendix 19.C.

**Remark 47** *A **two-way tabulation** or **cross tabulation** of the two variables $x$ and $y$ is a two-way frequency table that lists the number (or fraction) of observations equal to each of the distinct values taken by the pair $(x, y)$. It can include additional information including row percentages, column percentages and expected frequencies.*

## 8.3 Two-way Scatter Plot

The standard visual method for bivariate data is a **two-way scatter plot**. This is a plot of one variable on the vertical axis against the other variable on the horizontal axis. A major decision to make is which variable to place on the vertical axis and which to place on the horizontal axis. In many situations it is clear which variable it is that we want to explain, and this variable appears on the vertical axis. As expanded upon in Chapter 8.10, however, it need not be the case that the variable on the horizontal axis is causing the variable on the vertical axis.

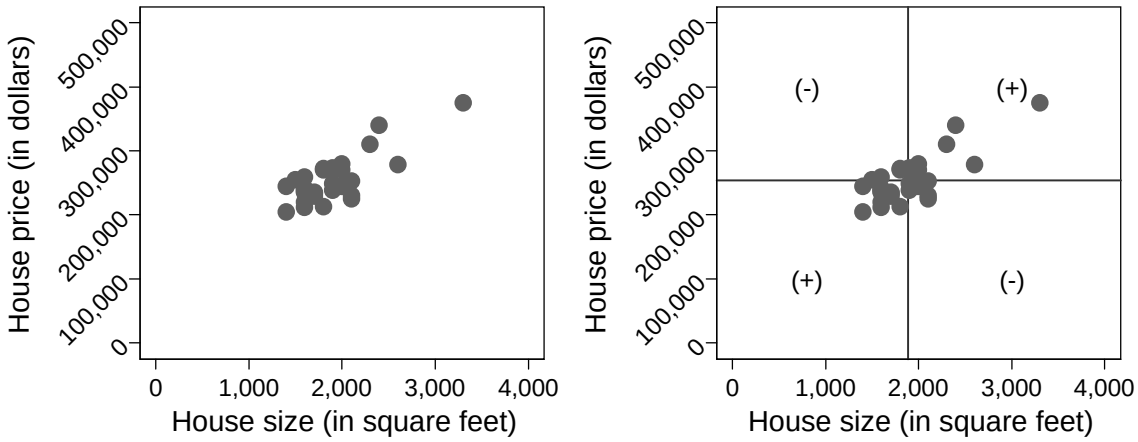


Figure 8.1: House price and size: Scatterplot and the four quadrants used to calculate the correlation coefficient.

The left panel of Figure 8.1 presents a scatter plot of the house price data. House price appears on the vertical axis and house size appears on the horizontal axis, since we wish to explain house price given house size. Each point represents a combination of sale price and size of house. For example, the upper right point is for a house that sold for \$375,000 and was 3,300 square feet in size. The scatter plot clearly illustrates that, on average, larger houses sell for more.

**Remark 48** *A two-way scatter plot is the standard tool for visualizing the relationship between two variables.*

## 8.4 Correlation

The sample correlation coefficient is the standard single measure of the association between two variables. It is based on the sample covariance between two variables, which is presented first.

### 8.4.1 Sample Covariance

Recall the formulas for the sample variances of $x$ and $y$, respectively $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. The **sample covariance** between $x$ and $y$ is the analogous

cross-product

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Note that the formula is symmetric in $x$ and $y$, so $s_{yx} = s_{xy}$.

Also, the sample covariance between $x$ and itself equals the variance of $x$. Algebraically, $s_{xx} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = s_x^2$.

The sign of the covariance is easily interpreted, with $s_{xy} > 0$ if there is positive association, $s_{xy} = 0$ if there is no association, and $s_{xy} < 0$ if there is negative association.

To see this, consider the case of positive covariance. Then $s_{xy} > 0$ if the cross-product $(x_i - \bar{x})(y_i - \bar{y})$ is mostly positive. This happens if most observations have both $(x_i - \bar{x}) > 0$ and $(y_i - \bar{y}) > 0$ or both $(x_i - \bar{x}) < 0$ and $(y_i - \bar{y}) < 0$. The association in this case is positive, since above-average values of $x$ tend to be associated with above-average values of $y$, and below-average values of $x$ tend to be associated with below-average values of $y$.

The second panel of Figure 8.1 indicates a positive covariance for the data on house price and size. The vertical line is $\bar{x} = 1883$, and the horizontal line is $\bar{y} = 253910$. The top-right quadrant, denoted (+), has positive values of $(x_i - \bar{x})(y_i - \bar{y})$ since in this quadrant $(x_i - \bar{x}) > 0$ and $(y_i - \bar{y}) > 0$. Similar considerations lead to the signs in the other three quadrants. The covariance $s_{xy} = 11701613$ is positive, as most of the observations lie in the two quadrants with positive value for $(x_i - \bar{x})(y_i - \bar{y})$. Furthermore, for the observations in the two quadrants with negative value for $(x_i - \bar{x})(y_i - \bar{y})$, this negative cross product is relatively small in absolute value.

**Remark 49** *The sample covariance $s_{xy}$ is positive if $x$ and $y$ tend to move together in the same direction, and negative if $x$ and $y$ tend to move together in the opposite direction. A covariance of zero means there is no linear association between $x$ and $y$.*

### 8.4.2 Sample Correlation

A weakness of the sample covariance is that its magnitude is not easily interpreted. For the house price and house size data $s_{xy} = 11701613$. This is a large number, but it does not necessarily imply that the association between $x$ and $y$ is large. Furthermore, $s_{xy}$ changes with the units of measurement. For example, if house price is measured in thousands of dollars rather than dollars, the covariance is divided by one thousand since $(y_i - \bar{y})$ is then reduced by a factor of 1,000.

The **sample correlation coefficient** is a standardized measure, or unit-free measure, of association between $x$ and $y$ that is defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\text{Covariance of } x \text{ and } y}{(\text{Standard deviation of } x) \times (\text{Standard deviation of } y)}.$$

Given the definitions of covariance and standard deviation,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

since the multiplicative factor $1/(n-1)$ in the numerator and denominator cancels out. It can be shown that $-1 \leq r_{xy} \leq 1$.

The sign of $r_{xy}$ is the same as the sign of $s_{xy}$. So a positive value for $r_{xy}$ means a positive association, a negative value for $r_{xy}$ means a negative association, and a zero value for $r_{xy}$ means no association. The advantage of the correlation coefficient, compared to the sample covariance, is that it is bounded between $-1$ and 1.

**Remark 50** *The sample correlation coefficient $r_{xy}$ lies between $-1$ and 1. It is positive if $x$ and $y$ tend to move together in the same direction and equals 1 if they move exactly together in the same direction. It is negative if $x$ and $y$ tend to move together in the opposite direction and equals $-1$ if they move exactly together in the opposite direction. It is zero if there is no linear association between $x$ and $y$.*

Some algebra shows that the sample correlation coefficient is $(n-1)$ times the sample covariance between the z-scores $(x_i - \bar{x})/s_x$ and $(y_i - \bar{y})/s_y$. It is also the sample correlation between these z-scores.

There are several ways to measure association between two variables – the correlation coefficient is more completely called the **Pearson product moment correlation coefficient**. It measures linear association between two variables but can miss nonlinear association. As a result, while two variables that are statistically independent have zero correlation, zero correlation does not imply that the variables are statistically independent. In practice for most economics data the correlation coefficient is an adequate measure of association.

Figure 8.2 illustrates, in turn, strong positive correlation ($r = .78$), moderate positive correlation ($r = .54$), almost no correlation ($r = .07$), and moderate negative correlation ($r = -.54$).

The formula for the correlation coefficient treats $x$ and $y$ symmetrically, so $r_{yx} = r_{xy}$. Thus while the correlation coefficient detects association, it is neutral on whether it is $x$ that is causing $y$ or $y$ that is causing $x$.

### 8.4.3 Autocorrelations for Time Series Data

Time series data are often correlated over time, with this correlation reducing as observations become further apart in time.

For a time series $y_t$, $t = 1, ..., T$, let $y_{t-j}$ denote the data lagged $j$ periods. Then the correlation between $y_t$ and $y_{t-j}$ is called the **autocorrelation at lag** $j$ and is denoted $\rho_j$.

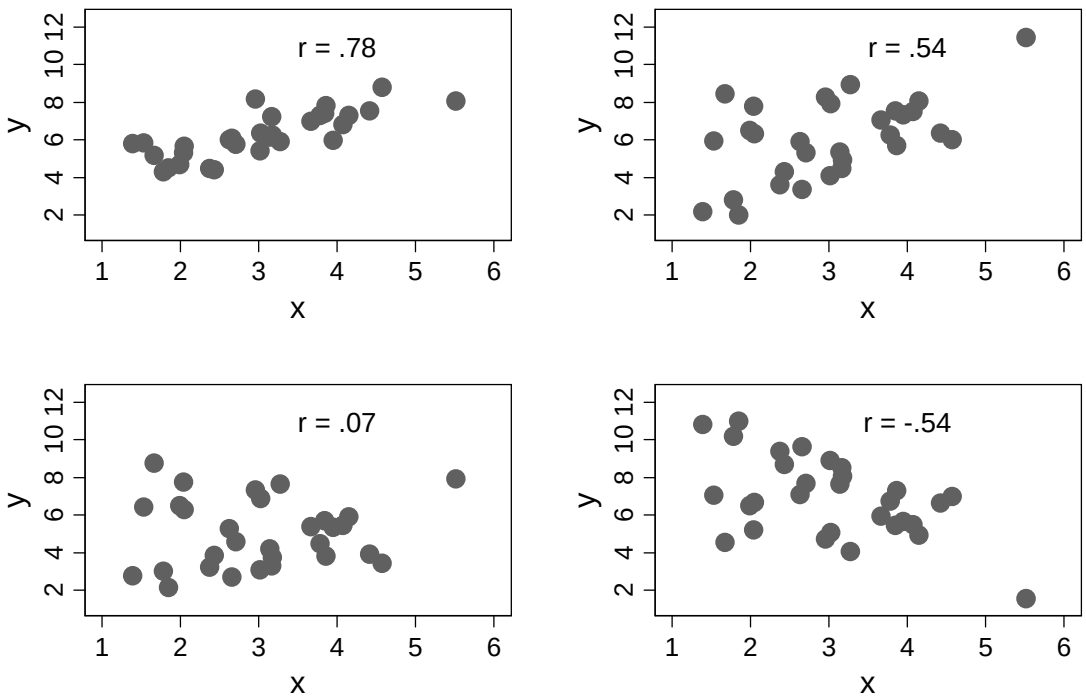


Figure 8.2: Correlation: Examples of different strengths and direction of correlation.

The simplest way to compute $\widehat{\rho}_j$ is to use the usual sample correlation coefficient between $y_t$ and $y_{t-j}$. Note that this is computed using only $T-j$ observations since $y_{t-j}$ is observed only for $t = j+1, ..., T$. More specialized econometrics software uses a more refined formula for $\widehat{\rho}_j$ that leads to somewhat different estimates in finite samples but to the same value as the sample size gets very large.

As an example, consider the annual data on annual U.S. price inflation from 1930 to 2014 in dataset INFLATION introduced in Chapter 6.6. The sample correlation coefficient yields autocorrelations at lags 1 to 3 of $\widehat{\rho}_1 = 0.606$, $\widehat{\rho}_2 = 0.243$, and $\widehat{\rho}_3 = 0.142$. A more specialized command to compute autocorrelations gives $\widehat{\rho}_1 = 0.604$, $\widehat{\rho}_2 = 0.233$, and $\widehat{\rho}_3 = 0.128$. Both sets of estimates indicate considerable correlation in inflation over time.

**Remark 51** *For time series data the autocorrelation at lag $j$ is the correlation between current data and the data lagged $j$ periods.*

### 8.4.4 Example: House Price Data

For the house data, $r_{xy} = 0.786$, so there is high positive association between house sale price and house size. A larger house is associated with a higher price.

Regression analysis, presented next, provides an estimate of the magnitude of the relationship – as house size increases by 100 square feet, say, what is the associated change in house price?

## 8.5 Regression Line

The regression line is the data analysis tool most used by economists. Here the method and interpretation of regression estimates is presented. The subsequent chapters consider statistical inference.

### 8.5.1 Regression Line

The **regression line** obtained from regression of $y$ on $x$ is denoted

$$\widehat{y} = b_1 + b_2 x.$$

The following terminology is used for the variables in this regression.

- $y$ is called the **dependent variable**
- $\widehat{y}$ is called the **predicted value** or **fitted value** of the dependent variable

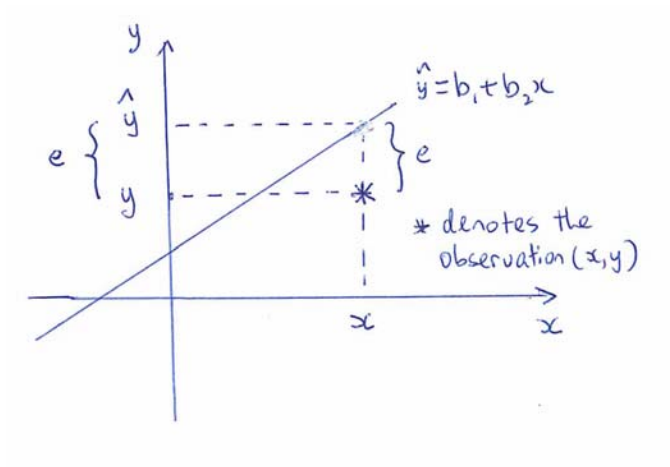


Figure 8.3: Residual (here negative): The vertical difference between the data point and the regression line.

- $x$ is called the **independent variable** or **explanatory variable** or **regressor variable** or **covariate.**

The data are used to estimate the coefficients $b_1$ and $b_2$, where $b_1$ is the estimated $y$-axis **intercept (or constant**) and $b_2$ is the estimated **slope coefficient**.

### 8.5.2 The Residual

The difference between the actual value of $y$ and its fitted value from the regression line is called the residual – the amount left over after fitting the line. Thus the **residual**

$$e = y - \widehat{y}.$$

Commonly-used alternative notation for the residual is $\widehat{u} = y - \widehat{y}$.

Figure 8.3 illustrates the residual for one observation. For the pictured observation the residual is negative, since $\widehat{y}$ exceeds $y$ so $y - \widehat{y} < 0$.

For the first observation, with subscript 1, the residual is $e_1 = y_1 - \widehat{y}_1$. For the second observation the residual is $e_2 = y_2 - \widehat{y}_2$, and so on. For a representative observation, say the $i^{th}$ observation, the residual is given by

$$e_i = y_i - \widehat{y}_i = y_i - b_1 - b_2 x_i,$$

where the second equality uses $\widehat{y}_i = b_1 + b_2 x_i$.

If the regression line fitted the data perfectly then the residuals would all equal zero. This ideal rarely occurs in practice. Instead we seek to make the residuals as small as possible. There are several ways to proceed, depending on what penalty function is used for nonzero residuals. The obvious method is to minimize the sum of the $n$ residuals, $\sum_{i=1}^{n} e_i$. But it turns out that there are many ways to do so. For example if we let all fitted observations equal the sample mean (so $\widehat{y}_i = \bar{y}$ for all $i$), it can be shown that the sum of residuals always equals zero.

Instead we minimize the sum of squared residuals, $\sum_{i=1}^{n} e_i^2$.

### 8.5.3 Least Squares Regression

Given data $(y_1, x_1), ..., (y_n, x_n)$ the **least squares regression method** chooses $b_1$ and $b_2$ to **minimize the sum of squares of the residuals**,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2.$$

The resulting formula for the **least squares slope coefficient** is

$$b_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

and the **least squares intercept** is

$$b_1 = \bar{y} - b_2 \bar{x}.$$

These formulas for $b_1$ and $b_2$ are obtained using calculus methods. Differentiating with respect to $b_1$ and $b_2$ and setting the derivatives equal to zero gives two equations in two unknowns. These equations can be solved for $b_1$ and $b_2$ after considerable algebra.

A detailed example of computation by hand of $b_1$ and $b_2$ from these formulas is given in Appendix A. These calculations are easily done by a computer. The OLS estimates can always be computed, provided there are at least two observations and that there is some variation in the regressors, so that $\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$.

The origin of the term **regression** is given in Chapter 8.10. A more complete term for least squares regression is **ordinary least squares (OLS) regression**. The adjective "ordinary" is added as there are other variations on least squares; see Chapter 14.7.

### 8.5.4 Interpretation of the Slope Coefficient

The slope coefficient $b_2$ can be directly interpreted as the change in the fitted value of $y$ when $x$ increases by one unit.

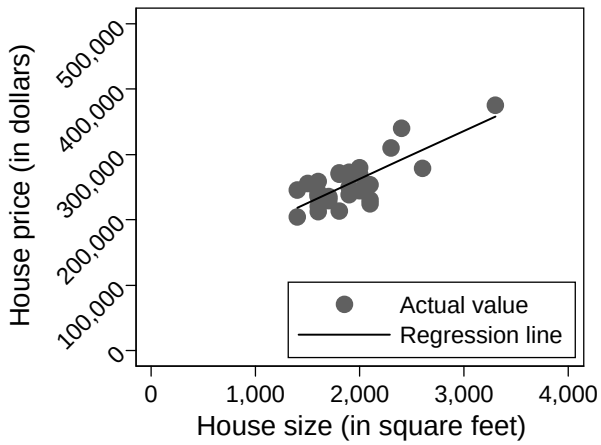


Figure 8.4: Regression of house price on size.

To see this, suppose the regressor changes by $\Delta x$ from $x$ to $(x + \Delta x)$. Then the fitted value $\widehat{y}$ changes from $b_1 + b_2 x$ to $b_1 + b_2(x + \Delta x) = b_1 + b_2 x + b_2 x \Delta x$. This is a change of $b_2 \Delta x$. It follows that $\Delta \widehat{y} = b_2 \Delta x$, which implies

$$\frac{\Delta \widehat{y}}{\Delta x} = b_2.$$

For a one-unit change in $x$, $\Delta x = 1$ so $\Delta \widehat{y} = b_2$.

The same result can be obtained using calculus methods, since $\widehat{y} = b_1 + b_2 x$ has derivative $d\widehat{y}/dx = b_2$.

### 8.5.5 Example: House Price

OLS regression of house sale price (variable *Price*) on house size (variable *Size*) using dataset HOUSE leads to fitted regression line

$$\widehat{Price} = 115,017 + 73.77 \times Size.$$

The slope coefficient of 73.77 implies that one more square foot in size is associated with a \$73.77 increase in the house price. Equivalently an additional small room of size ten feet by ten feet, or 100 square feet, is associated with a $100 \times \$73.77 = \$7{,}377$ increase in house price.

Figure 8.4 presents the regression line, superimposed on a scatter plot of the data.

### 8.5.6 Summary for the Regression Line

Combining the preceding results we have the following.

**Remark 52** *The regression line is $\widehat{y} = b_1 + b_2 x$. The OLS estimates $b_1$ and $b_2$ are obtained by minimizing the sum of squared residuals, where the residual $e = y - \widehat{y}$ is the difference between the actual and predicted value of $y$. The slope coefficient $b_2$ gives the change in the fitted value of $y$ when $x$ changes by one unit.*

## 8.6 Further Details on Regression

This section considers prediction following regression and outlier detection. Additionally we consider regression on just an intercept, and link this to the sample mean.

### 8.6.1 Prediction

The regression line can be used to predict values of $y$ for given values of $x$. For $x = x^*$ the **prediction** is

$$\widehat{y}^* = b_1 + b_2 x^*.$$

For example, for the house sales data a house of size 2000 square feet is predicted to sell for \$262,557, since $\widehat{y} = 115017 + 73.77 \times 2000 = 262557$.

Prediction can be **in-sample**, in which case $\widehat{y}_i = b_1 + b_2 x_i$ is called the **fitted value** of $y_i$, $i = 1, ..., n$. For in-sample prediction, a property of least squares regression is that on average the in-sample predictions equal the actual value of $y$. That is, $\frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$. Equivalently, the **least squares residuals** $e_i = y_i - \widehat{y}_i$ **always sum to zero**. These results hold provided an intercept is included in the regression.

Prediction can also be **out-of-sample**. In that case it can be tempting to extrapolate and make prediction for $y$ for values of $x$ beyond the range of the sample. Such predictions may not be very reliable.

Chapter 10.9 provides a more detailed discussion of prediction. Such predictions are most reliable for $x$ values close to the sample mean $\bar{x}$. And the better the fit of the model, that is the closer the sample data points are to the regression line, the better will be the forecast.

**Remark 53** *A property of OLS regression with an intercept is that the residuals sum to zero. Equivalently, on average the in-sample fitted values are equal to the sample values of $y$.*

### 8.6.2 Outlying Observations

An **outlier** is or **outlying observation** is one that is a relatively large distance from the bulk of the data. In the bivariate context the outlier could have wither an unusual value of $x$ or of $y$. A scatter plot is a useful visual tool.

If the scatter plot includes the fitted regression line, then points far from the line correspond to observations with a large residual (recall $e = y - \widehat{y}$). In such cases, the regression line is indicating that the actual value of $y$ is unusual given the value of $x$.

Outlying observations can have an unusually large effect on the OLS estimates $b_1$ and $b_2$. From the numerator formula for $b_2$, an observation with a large value for $(x_i - \bar{x})(y_i - \bar{y})$ can have a big influence on $b_2$. This is case for observations that are a long way from both $\bar{x}$ and $\bar{y}$.

### 8.6.3 Intercept-only Regression

OLS regression of $y$ on just an intercept yields fitted value $\widehat{y} = b_1$ where $b_1 = \bar{y}$. Then the fitted value $\widehat{y}_i = \bar{y}$ for all observations.

To obtain this result, note that in general the intercept $b_1 = \bar{y} - b_2\bar{x}$. This simplifies to $b_1 = \bar{y}$ when there is no regressor $x$, since in that case the slope coefficient $b_2 = 0$.

Also for intercept-only regression, the standard error of the regression reduces to the usual sample standard deviation of $y$. To see this, note that then $\widehat{y}_i = \bar{y}$ and the degrees of freedom are $n - 1$, so $s_e = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$.

**Remark 54** *Regression of $y$ on only an intercept yields the sample mean $\bar{y}$. Univariate statistics based on the sample mean is just a special case of OLS regression.*

This result provides a justification for using least squares regression – it can be viewed as an extension of the sample mean for univariate data.

The result can be useful when the data $y_1, ..., y_n$ are correlated. In that case $se(\bar{y}) \neq s_y/\sqrt{n}$, so the methods of Chapters 5-7 need modification. Standard software for the sample mean does not handle this complication. But some regression commands do; see Chapter 10.7 for illustration.

Intercept-only regression is an option available in most computer packages that have a command for OLS regression. Giving the OLS command with dependent variable $y$ and no regressor given yields estimated intercept equal to the sample mean. Alternatively, if the package has an option to run a regression with the intercept set to zero then OLS regress $y$ on $z$, say, with the zero-intercept option, where the created variable $z$ equals 1 for every observation.

## 8.7 Measures of Model Fit

Two standard measures of model fit, the closeness of the data points to the fitted regression line, are the standard error of the regression and $R$-squared, a measure that is standardized to lie between 0 and 1.

### 8.7.1 Standard Error of the Regression

An obvious measure of how close the fitted values $\widehat{y}$ are to the actual values $y$ is the standard deviation of the residual $\widehat{e} = y - \widehat{y}$. Instead a minor variation is used.

The **standard error of the regression** is

$$s_e = \sqrt{\frac{1}{n-2}\sum\nolimits_{i=1}^{n}(y_i - \widehat{y}_i)^2}.$$

The measure $s_e$ is essentially the standard deviation of the residual, except that division is by $n-2$ rather than $n-1$. The divisor $(n-2)$, called the **degrees of freedom**, is used because only $(n-2)$ terms in the sum are free to vary since computation of $\widehat{y} = b_1 + b_2 x$ is based on the two estimates $b_1$ and $b_2$ .

Lower values of $s_e$ mean that the fitted values are closer to the actual values of $y$. Another name for $s_e$ is the **root mean squared error (MSE) of the residual**. It is also sometimes called the **standard error of the residual**.

**Remark 55** *The standard error of the regression measures the standard deviation of the residual, and hence the variability of the dependent variable around the regression line.*

### 8.7.2 Definition of R-Squared

The standard error of the residual is not a scale free measure. For example, if we measured house price in thousands of dollars rather than dollars, then $s_e$ after regression of house price on size would be 0.001 times as large. A scale-free measure is provided by R-squared.

R-Squared, denoted $R^2$, measures the **fraction of the variation of $y$ (around the sample mean $\bar{y}$) that is explained by the regressor**.

Variability in $y$ around the sample mean $\bar{y}$ is measured using the **total sum of squares**, denoted $TSS$, where

$$TSS = \sum\nolimits_{i=1}^{n}(y_i - \bar{y})^2,$$

is the sum of squared deviations of the dependent variable $y$ around $\bar{y}$.

The left panel of Figure 8.5 presents the deviations $(y_i - \bar{y})$ for five data points with $(x, y)$ equal to, respectively $(1, 4.7)$, $(2, 4.7)$, $(3, 4.5)$, $(4, 7.4)$, and $(5, 8.7)$.

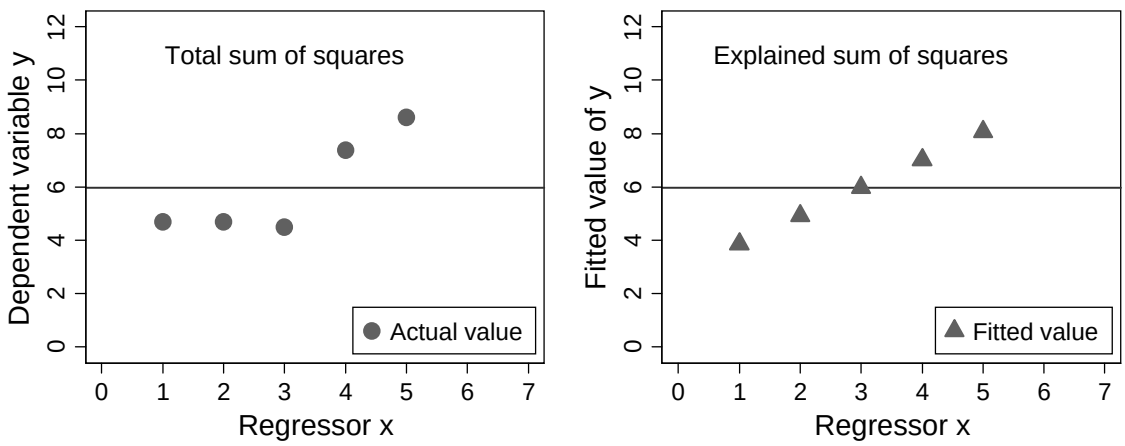


Figure 8.5: R-squared components: Total sum of squares and explained sum of squares for five observations.

Variability around $\bar{y}$ after regression is measured using the **explained sum of squares**, denoted *ExpSS*, where

$$ExpSS = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2,$$

is the sum of squared deviations of the fitted value $\widehat{y}$ around $\bar{y}$. The right panel of Figure 8.5 presents the deviations $(\widehat{y}_i - \bar{y})$ for the same data points, where the fitted regression was $\widehat{y} = 2.81 + 1.05x$, and the approximate fitted values are, respectively, 3.9, 4.9, 6.0, 7.0 and 8.1. The explained sum of squares is also referred to as the **regression sum of squares** or as the **model sum of squares**.

**R-squared** equals the explained sum of squares as a fraction of the total sum of squares

$$R^2 = \frac{ExpSS}{TSS} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

and is also called the **coefficient of determination**.

For the five-observation example illustrated in Figure 8.5 the $y$ values are, respectively, 4.7, 4.7, 4.5, 7.4 and 8.7. Since $\bar{y} = 6.0$, it follows that $TSS = (-1.3)^2 + (-1.3)^2 + (-1.5)^2 + 1.4^2 + 2.7^2 = 14.88$. The fitted $\widehat{y}$ values are, respectively, 3.9, 4.9, 6.0, 7.0 and 8.2. It follows that $ExpSS = (-2.1)^2 + (-1.1)^2 + (0.0)^2 + 1.0^2 + 2.2^2 = 11.46$. Thus $R^2 = 11.46/14.88 = 0.77$. So 77 percent of the variation in $y$ is explained by regression on $x$.

$R^2$ is necessarily positive, since the two terms in the ratio are positive. Provided the regression includes an intercept, the explained sum of squares are at most equal to the total sum of squares, so $R^2 \leq 1$. If the regression does not include an intercept then $R^2$ should not be used.

**Remark 56** *$R^2$ measures the fraction of the variation in the dependent variable explained by the regressor. Provided the regression includes an intercept, $R^2$ lies between 0 and 1. The minimum value occurs if there is no relationship between $y$ and $x$, so $\widehat{y}_i = \bar{y}$ for all observations. $R^2 = 1$ if the regression line perfectly fits the dependent variable, so $\widehat{y}_i = y_i$ for all observations.*

### 8.7.3 Alternative Computation of R-Squared

The **residual sum of squares**, denoted *RSS*, is defined by

$$RSS = \sum\nolimits_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

*RSS* is the sum of squared deviations of the fitted value $\widehat{y}$ around the actual value $y$. The residual sum of squares is sometimes referred to as the **error sum of squares**.

It can be shown that

$$TSS = ExpSS + \ RSS.$$

It follows that $R^2$ can be equivalently defined as

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

The least squares coefficients therefore maximize $R^2$, since they minimize the residual sum of squares.

For the five-observation example illustrated in Figure 8.5 the residuals $(y_i - \widehat{y}_i)$ are, respectively, 0.8, $-0.2$, $-1.5$, 0.3 and 0.5. It follows that $RSS = 0.8^2 + (-0.2)^2 + (-1.5)^2 + 0.3^2 + 0.5^2 = 3.4$. From previous calculation $TSS = 14.4$. So $R^2 = 1 - 3.4/14.4 = 0.76$. Again $R^2 = 0.76$.

### 8.7.4 Interpretation of R-Squared

It is clear that a value of $R^2 \simeq 0$ represents a poor fit and a value of $R^2 \simeq 1$ represents an excellent fit. But there is no rule for where $R^2$ becomes large enough that the fit moves from poor to good.

The value of $R^2$ increases as data become more aggregated. For example, $R^2$ will be quite low, around 0.1 to 0.3, for regression of earnings on education using individual-level data. If instead more aggregated data is used, such as analysis of state-average earnings across states or of national average earnings over time, $R^2$ will be much closer to 1.

For bivariate regression, $R^2$ can be used to compare models with the same dependent variable. Then models with higher $R^2$ are preferred. But it makes little sense to compare the $R^2$ across models with different dependent variable $y$, because then $R^2$ explains variation

around different sample means $\bar{y}$. For example, a model explaining the level of annual GDP will in practice have a much higher $R^2$ than a model for the annual change in GDP.

Low values of $R^2$ do not mean that regression analysis is without merit. For example, regression of earnings on education usually indicates a substantial effect of education, such as one more year of education being associated with a six percent increase in annual earnings. At the same time the $R^2$ in earnings-education regressions using individual-level data is very low. These seemingly contradictory results can be explained as follows. There is a large effect of schooling on earnings, **on average**. But at the **individual level**, education is only one of many determinants of earnings and there is considerable variability in earnings even for people with the same level of education. If data analysis shows that, on average, schooling substantially raises earnings, then it is a worthwhile investment for government to encourage high levels of schooling. At the same time, however, there is great uncertainty as to whether any one given individual will necessarily increase their earnings given this additional education.

It can be shown that $R^2$ **equals the squared sample correlation coefficient between** $y$ and $x$:

$$R^2 = r_{xy}^2.$$

Thus one by-product of regression analysis is to obtain the squared sample correlation coefficient. It follows that for the four datasets plotted in Figure 8.2, least squares regression leads to fitted models with $R^2$ of, respectively 0.61 ($= 0.78^2$), 0.29, 0.01, and 0.29. And since $r_{xy}^2 = r_{yx}^2$ this result also implies that $R^2$ from regression of $y$ on $x$ equals $R^2$ from regression of $x$ on $y$.

The correlation coefficient is unchanged when data are subject to a linear transformation. Since $\widehat{y} = b_1 + b_2 x$ it follows that $r_{\widehat{y}y}^2 = r_{xy}^2$. So equals the squared sample correlation coefficient between $y$ and $\widehat{y}$.

**Remark 57** *For bivariate regression $R^2$ equals the squared correlation between $y$ and $x$. And $R^2$ equals the squared correlation between $y$ and $\widehat{y}$.*

$R^2$, unlike $r^2$, can be easily extended to regression with additional regressors; see Chapter 13. That chapter also introduces the adjusted $R^2$, denoted $\bar{R}^2$ that adjusts $R^2$ for the number of regressors.

## 8.8 Computer Output following OLS Regression

Commands for OLS regression and examples of resultant output for several statistical packages are given in Appendix A. Different packages report results in somewhat different ways and provide different levels of detail. Here we present quite detailed regression results.

Table 8.5: House price: Computer output from regression on house size.

| **ANOVA Table** | | | | | |
|---|---|---|---|---|---|
| **Source** | **SS** | **df** | **MS** | **F** | **p-value** |
| Explained | $2.4171\times10^{10}$ | 1 | $2.4171\times10^{10}$ | 43.58 | 0.000 |
| Residual | $1.4975\times10^{10}$ | 27 | $5.546\times10^{8}$ | | |
| Total | $3.9146\times10^{10}$ | 28 | $1.3981\times10^{9}$ | | |

| **Dependent Variable** *Price* | | | | | | |
|---|---|---|---|---|---|---|
| **Regressor** | **Coefficient** | **Standard Error** | **t statistic** | **p value** | **95% conf. int.** | |
| *Size* | 73.77 | 11.17 | 6.60 | 0.000 | 50.84 | 96.70 |
| *Intercept* | 115017 | 21489 | 5.35 | 0.000 | 70925 | 159110 |

| **Summary Statistics** | |
|---|---|
| Observations | 29 |
| F(1,27) | 43.58 |
| p-value for F | 0.0000 |
| R-squared | 0.618 |
| Adjusted $R^2$ | 0.603 |
| St. error of regression | 23551 |

### 8.8.1 Computer Output for House Price Example

Table 8.5 presents fairly complete results from OLS regression of variable *Price* on variable *Size.*

The results can be broken into three parts – ANOVA table, regression coefficients, and diagnostic statistics – that are considered in turn.

### 8.8.2 ANOVA Table and R-squared

The analysis of variance (ANOVA) table is less important than the other results, and is for this reason not reported by all statistical packages.

In the column headed SS the total sum of squares is decomposed into the explained (or regression or model) sum of squares and the residual sum of squares. The column headed df gives the associated degrees of freedom, the column headed MS gives the mean square which equals the sum of squares divided by its degrees of freedom. The remainder of the ANOVA table gives the $F$-statistic and its $p$-value; see Chapter 14.5.

The ANOVA table provides data that is used to compute R-squared and the standard error of the regression.

The R-squared equals the explained sum of squares ($2.4171 \times 10^{10}$) divided by the total sum of squares ($3.9146 \times 10^{10}$) which equals 0.618, as given near the bottom of Table 8.5.

Equivalently $R$-squared equals one minus the the explained sum of squares ($1.4975 \times 10^{10}$) divided by the total sum of squares ($3.9146 \times 10^{10}$) which equals $1 - 0.382 = 0.618$.

The standard error of the regression (also called the standard error of the residual and the root mean square error) equals $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$. Here the residual sum of squares $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ equals $1.4975 \times 10^{10}$ or, more precisely, 14,975,101,655. Dividing by $n-2 = 27$ yields value $5.546 \times 10^8$, which is the entry for residual mean square (MS) in the ANOVA table. Taking the square root yields $s_e = 23551$, which is the entry at the bottom of Table 8.5.

### 8.8.3 Regression Coefficients

The second block in Table 8.5 gives the main results. The intercept (or constant) equals 115017 and the slope coefficient ($b_2$) equals 73.77. Some statistical packages report the intercept last, as in Table 8.5, while others report the intercept first.

The remaining columns give the standard error, $t$-statistic, $p$-value and 95% confidence interval for the intercept and slope estimates. These quantities are the subject of Chapter 9.

### 8.8.4 Regression Diagnostics

The third block in Table 8.5 gives key regression diagnostics. The $F$-statistic, its $p$-value and the adjusted $R^2$ are relevant for multiple regression, rather than bivariate regression, so are presented in later chapters. Computation of $R^2$ and the standard error of the regression, from the results in the ANOVA table, has already been discussed.

$R^2 = 0.618$ can be viewed as a good fit. The standard error of the regression equals 23551, substantially smaller than the standard deviation of variable *Price* which from Table 8.2 equals 37391.

### 8.8.5 Intercept-only Regression

Now consider intercept-only regression. For the house price data, regressing variable *Price* on just an intercept yields intercept estimate of 253910 and standard error of the regression of 37391. As expected, these equal the mean and standard deviation of variable *Price* given in Table 8.2.

## 8.9 Regression and Correlation

For bivariate data there is a very close connection between regression and correlation. The term $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ appears in the definitions of both $b_2$ and $r_{xy}$. The slope coefficient

can be re-expressed in terms of the sample correlation coefficient as

$$b_2 = r_{xy} \times \frac{s_y}{s_x}.$$

(To obtain this result, observe that $b_2 = s_{xy}/s_x^2$ upon cancellation of the common multiplicative factor $1/(n-1)$ in $s_{xy}$ and $s_x^2$. Then multiplying by $(s_y/s_y)$ and rearranging yields $b_2 = s_{xy}/s_x^2 = [s_{xy}/(s_x s_y)] \times (s_y/s_x)] = r_{xy} \times (s_y/s_x)$.)

**Remark 58** *The slope coefficient is a multiple $s_y/s_x$ of the sample correlation coefficient.*

It follows that if there is positive correlation, i.e. $r_{xy} > 0$, then the regression slope coefficient $b_2$ is positive, and similarly $r_{xy} < 0$ implies a negative regression slope coefficient. So the sample correlation coefficient and the slope coefficient always lead to the same conclusion regarding whether the association between the two variables is positive or negative.

Furthermore, it can be shown that if we **standardize** the data and regress the z-score $(y_i - \bar{y})/s_y$ on the z-score $(x_i - \bar{x})/s_x$, then the least squares slope coefficient equals $r_{xy}$. This implies the following.

**Remark 59** *The correlation coefficient $r_{xy}$ measures the number of standard deviations that $y$ changes by when $x$ changes by one standard deviation.*

In economic data analysis regression is the most commonly-used method. In some other disciplines correlation analysis is more extensively used. The two methods lead to exactly the same conclusions regarding association between $x$ and $y$.

## 8.10 Causation

There is an important distinction between correlation and regression analysis. The correlation coefficient always treats $x$ and $y$ neutrally, since $r_{xy} = r_{yx}$. Such symmetry is not the case for regression. As shown below, the slope coefficient from regression of $y$ on $x$ is not the reciprocal of that from reverse regression of $x$ on $y$. Furthermore, without further information we can only measure association, and not causation. Thus even if there is an association between $y$ and $x$, it could be that $y$ causes $x$, $x$ causes $y$, or that a third variable is causing both $x$ and $y$.

### 8.10.1 Reverse Regression

The regression $\widehat{y} = b_1 + b_2 x$ is called **regression of $y$ on $x$**. We could instead regress $x$ on $y$, called a **reverse regression**, estimating

$$\widehat{x} = c_1 + c_2 y.$$

One might expect that one slope is the reciprocal of the other, so $c_2 = 1/b_2$, but this is not the case.

Using the formula for the slope coefficient with $x$ and $y$ reversed, the slope coefficient from the reverse regression is

$$c_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

and the intercept $c_1 = \bar{x} - b_2\bar{y}$. By inspection $c_2 \neq 1/b_2$. In fact the relationship between the two slope coefficients is

$$c_2 = b_2 \times \frac{s_x^2}{s_y^2}.$$

(To obtain this result, recall that $b_2 = r_{xy}s_y/s_x$. Similarly, $c_2 = r_{xy}s_x/s_y = [s_{xy}/s_x s_y] \times (s_x/s_y)$ given the definition of $r_{xy}$. Some rearrangement yields $c_2 = [s_{xy}/s_x^2] \times (s_x^2/s_y^2) = b_2 \times (s_x^2/s_y^2)$.)

**Remark 60** *The slope coefficient from reverse regression of $x$ on $y$ does not equal the reciprocal of the slope coefficient from regression of $y$ on $x$.*

For the house price regression, $b_2 = 73.77$. The reverse regression of house size on house price yields slope $c_2 = 0.0084$. This differs from the reciprocal $1/b_2 = 1/73.77 = 0.0136$.

### 8.10.2 Causation

The data alone cannot tell us which direction, if any, is the appropriate direction to run a regression, as the only information in the data is the size of the correlation. Thus given a fitted model $\widehat{y} = b_1 + b_2 x$, without further information we can only say that a one unit increase in $x$ **is associated with** a $b_2$ increase in $y$. We **cannot** say that a one unit increase in $x$ causes a $b_2$ increase in $y$.

For example, a medical study might find that alcohol consumption is associated with depression. But is it alcohol consumption that causes depression, or is it depression that leads to alcohol consumption?

With additional information we may be able to say that one causes the other. For example, for the relationship between parents' height and height of their fully-grown child, the direction of causation is clear. Indeed it is from this example that Galton coined the term **regression** over one hundred years ago. Let $y$ be child's height and $x$ be midparent height (the average of mother's and father's height). Galton found that $y \simeq \bar{x} + b_2(x - \bar{x})$ where $b_2 = 0.33 < 1$. Thus if the parents have above average height, so that $(x - \bar{x}) > 0$, then the child will on average also be of above average height, but not as much above average as the parents. Galton termed this "regression towards mediocrity", where mediocrity is the average or mean.

Many examples exist where the direction of causation is questionable. Often it is due to a third variable that may be driving both $y$ and $x$. For example, higher education is positively associated with higher earnings. But this may be due solely to unobserved innate ability that leads to both higher earnings, due to higher productivity, and higher education, due to ability to study more advanced material.

**Remark 61** *Regression of $y$ on $x$ measures the association between $y$ and $x$ but does not on its own measure the direction of causation between $y$ and $x$.*

Most of this book focuses merely on association. Methods to determine causation, in the absence of strong prior information, are presented in Chapter 18.

## 8.11 Nonparametric Regression

The purpose of bivariate regression is to predict $y$ given $x$. In this chapter we restrict attention to the linear model, with $\widehat{y} = b_1 + b_2 x$. Later chapters, notably Chapter 12, considers specific nonlinear models such as $\widehat{y} = b_1 + b_2 \ln x$.

**Nonparametric regression** instead obtains fitted values $\widehat{y}$ without specifying (or parameterizing – hence the term nonparametric) a model for $y$ given $x$.

Nonparametric regression would be straightforward if we had many values of $y$ for each value of $x$, which can happen if the regressor takes a few discrete values. Then $\widehat{y}|x$ is simply the average of the values of $y$ for each distinct value of $x$, called a **local average**. For example, if there were nine observations for which $x = 2$ then $\widehat{y}|x = 2$ is simply the average of $y$ for the nine observations for which $x = 2$.

There are two limitations to this approach. First, there may not be multiple observations for each potential value of $x$. Second, even if they are the local averages $\widehat{y}$ can vary considerably across adjacent values of $x$. This is analogous to forming a histogram with a very narrow bin width.

The standard nonparametric regression estimators smooth the local averages in a manner similar to kernel density estimation that leads to a smoother histogram. First, the averages are computed using rolling bins or windows that are overlapping rather than distinct. Second, in forming the average of $y$ within each window more weight is given to observations that are closest to the center of the window and less to those near the ends of the window.

Figure 8.6 illustrates two commonly-used nonparametric regression estimation methods applied to the house price data. The solid curve is obtained by **local linear regression**. The dashed line is obtained by locally weighted scatter plot smoothing (**lowess**) estimation, a refinement of local linear regression. The two methods give similar results in this example, though in other cases there can be bigger differences, especially at the endpoints (low age and high age) where estimation is noisier due to sparser data. The estimates vary with the window width, with wider windows leading to a smoother estimate, and to a lesser extent

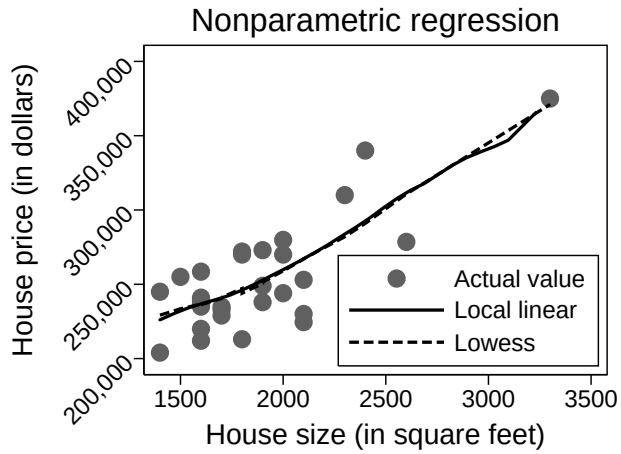


Figure 8.6: Nonparametric regression of house price on size: local linear and lowess.

with the weighting function. The graphs in Figure 8.6 use the window widths and weighting function that were automatically selected by the statistical package for lowess, but for local linear a much wider window width was used.

The two nonparametric regression estimates suggest that a linear model is a good model for these data. The linear model, assuming that it is a correctly specified model, has the advantages over the nonparametric regression of more precise estimation, and the possibility of prediction for values of $x$ outside the range of sample values of $x$. Model-based methods also have the advantage of being more easily extended to regression with several regressors.

Nonparametric regression is most often used in exploratory data analysis, providing a visual summary of the relationship between two variables that can be useful in its own right and can aid in model specification and testing. Most, but not all, statistical packages include nonparametric regression.

## 8.12 Key Concepts

1. A two-way tabulation or cross tabulation of two variables is a two-way frequency table that lists the number (or fraction) of observations equal to each of the distinct values taken by the pair of variables.

2. A two-way scatter plot is the standard tool for visualizing the relationship between two variables.

3. The sample correlation coefficient $r_{xy}$ is a measure of association between two variables

that lies between 1 (perfect positive association) and -1 (perfect negative association).

4. The autocorrelation at lag $j$ of a time series is the correlation between current data and the data lagged $j$ periods.

5. The least squares regression line is $\widehat{y} = b_1 + b_2 x$.

6. The residual $e = y - \widehat{y}$ is the difference between the actual and fitted values of the dependent variable.

7. The least squares coefficients $b_1$ and $b_2$ minimize the sum of squared residuals.

8. The least squares estimates can always be computed, provided there is some variation in the regressor value across observations.

9. The slope coefficient $b_2$ measures the change in the fitted value of $y$ when $x$ changes by one unit.

10. Regression on only an intercept yields the sample mean.

11. The standard error of the regression ($s_e$) measures the standard deviation of the residual, and hence the variability of $y_i$ around the regression line.

12. R-squared ($R^2$) measures the fraction of the variation of $y$ (around the sample mean $\bar{y}$) that is explained by the regressor.

13. For bivariate regression R-squared equals the squared correlation coefficient between $x$ and $y$. And it equals the squared correlation between $y$ and $\widehat{y}$.

14. The slope coefficient is a multiple $s_y/s_x$ of the sample correlation coefficient.

15. The correlation coefficient $r_{xy}$ measures the number of standard deviations that $y$ changes by when $x$ changes by one standard deviation.

16. The slope coefficient from reverse regression of $x$ on $y$ does not equal the reciprocal of the slope coefficient from regression of $y$ on $x$.

17. Least squares regression measures correlation, rather than causation. Without further information, we cannot say that a change in $x$ causes $y$. The causal path may be reversed. Or both $x$ and $y$ may be associated with a third variable that is driving the association.

18. Nonparametric regression is a useful exploratory data analysis for providing a visual summary of the relationship between two variables.

19. Two commonly-used nonparametric methods are local linear and lowess.

20. Key Terms: two-way tabulation; cross tabulation; two-way scatter plot; sample covariance; sample correlation coefficient; positive correlation; negative correlation; autocorrelation; regression line; dependent variable; independent variable; explanatory variable; regressor variable; covariate; least squares regression; residual; fitted value; slope coefficient; ordinary least squares; intercept coefficient; prediction; in-sample prediction; out-of-sample prediction; standard error of the regression; intercept-only regression; R-squared; total sum of squares; explained (or regression or model) sum of squares; residual (or unexplained) sum of squares; error sum of squares; reverse regression; causation; nonparametric regression; local linear; lowess.

## 8.13 Exercises

1. The dataset HOMEPRICEINDEX has monthly house price indices (=100 in January 2000) for twenty U.S. cities from January 1991 to September 2013.

   (a) Which city is least highly correlated with the composite10 index?

   (b) Which city is most highly correlated with the composite10 index?

   (c) Give a scatter plot for the price indices for the cities you found in parts a-b, with the part a city on the vertical axis.

   (d) Regress that part a city on the part b city. Explain in words the relationship between the two indices.

   (e) Find the squared correlation coefficient between the two cities. Compare this to $R^2$ from part d.

2. The dataset ANSCOMBE has 11 observations on variables x1–x4 and y1–y4.

   (a) Verify that x1–x4 have the same means and standard deviations.

   (b) Verify that y1–y4 have the same means and standard deviations.

   (c) Verify that the correlation between x1 and y1 equals that between x2 and y2, x3 and y3, and x4 and y4.

   (d) Verify that the intercept and slope coefficients from regression of y1 on x1 equals that from regression of y2 on x2, y3 on x3 and y2, x2 and y2, x3 and y3, and x4 and y4.

   (e) Now compare scatter plots of y1 on x1m, y2 on x2, y3 on x3 and y4 on x4. Are you surprised? What is the lesson?

3. The dataset ELECTRICITYPRICE has data for the California wholesale electricity market price on the price of electricity (in dollars per megawatt hour), and the volume of electricity (in megawatt hour) for the 3p.m.-4p.m. hourly period each day from January 1 2000 to November 30 2000.

   (a) Does the electricity price fluctuate very much? At what time of the year is it highest? Explain.

   (b) Does the volume of electricity fluctuate very much? At what time of the year is it highest? Explain.

   (c) Produce a scatter plot of price against volume, along with a fitted regression line.

   (d) Regress price against volume. Summarize your results.

   (e) Is this relationship estimating the supply curve or the demand curve? Hint: Electricity demand in California at this time was relatively price inelastic in the short run, but varied considerably with how hot it is. Short-run marginal costs increase with amount supplied as less efficient power plants come on line. Draw a price-quantity diagram with appropriate demand and supply curves.

4. The dataset KNEEREPLACE has 2011 data for New York hospitals on the mean (the average over cases) posted charge and mean cost of knee joint replacement for cases of moderate severity.

   (a) On average how much larger in relative terms is the mean charge than the mean cost.

   (b) Regress mean charge on mean cost. Are you surprised by the value of the slope coefficient? Explain.

   (c) Now give a scatter plot of the data. Note that two observations are outliers.

   (d) Drop these two outliers and regress mean charge on mean cost. Are you surprised by the value of the slope coefficient? Explain.

## 8.14 Appendix 8.A: Regression Computation Example

This section provides an example of manual computation of the OLS coefficients, the standard error of the regression, R-squared and the sample correlation coefficient.

Consider artificial data on number of vehicles per household ($y$) and household size ($x$) measured by number of people of all ages. There are five sample observations: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 2)$, $(x_3, y_3) = (3, 2)$, $(x_4, y_4) = (4, 2)$, and $(x_5, y_5) = (5, 3)$.

To calculate the intercept and slope of the regression line we compute the components for each observation and sum over observations. The calculations are presented in Table 8.6.

Table 8.6: Regression: Computation example for a sample of size five.

| $i$ | $x_i$ | $y_i$ | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | -2 | -1 | 2 | 4 |
| 2 | 2 | 2 | -1 | 0 | 0 | 1 |
| 3 | 3 | 2 | 0 | 0 | 0 | 0 |
| 4 | 4 | 2 | 1 | 0 | 0 | 1 |
| 5 | 5 | 3 | 2 | 1 | 2 | 4 |
| Sum | 15 | 10 | 0 | 0 | 4 | 10 |
| Mean | $\bar{x} = 3$ | $\bar{y} = 2$ | | | | |

Plugging these sums into the OLS formulas yields regression slope coefficient

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{4}{10} = 0.4,$$

and regression intercept

$$b_1 = \bar{y} - b_2\bar{x} = 2 - 0.4 \times 3 = 0.8.$$

Thus the computed regression line is $\widehat{y} = 0.8 + 0.4x$.

The fitted values of $\widehat{y} = 0.8 + 0.4x$ for the five observations are, respectively, 1.2, 1.6, 2, 2.4, and 2.8. Then the residual sum of squares

$$\sum_{i=1}^n (y_i - \widehat{y}_i)^2 = (1 - 1.2)^2 + (2 - 1.6)^2 + (2 - 2)^2 + (2 - 2.4)^2 + (3 - 2.8)^2 = 0.4.$$

The standard error of the regression is

$$s_e = \sqrt{\sum_{i=1}^n (y_i - \widehat{y}_i)^2/(n-2)} = \sqrt{0.4/3} = 0.365148.$$

To compute the associated $R^2$ we also need to calculate the total sum of squares.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (1 - 2)^2 + (2 - 2)^2 + (2 - 2)^2 + (2 - 2)^2 + (3 - 2)^2 = 2.0.$$

So

$$R^2 = 1 - \frac{0.4}{2.0} = 0.8.$$

Thus 80% of the variation in number of cars is explained by household size. Note that the explained sum of squares is $2.0 - 0.4 = 1.6$.

The sample correlation coefficient is computed as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \times \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{4}{\sqrt{10 \times 2}} = \frac{2}{\sqrt{5}} = 0.894.$$

This is close to one and indicates strong positive association between cars and household size. As expected, $r_{xy}^2 = (2/\sqrt{5})^2 = 4/5 = 0.8$ is equal to $R^2$.

# Chapter 9

# © A. Colin Cameron: The Least Squares Estimates

Regression curve fitting is relatively easy. Now we consider the more difficult topic of extrapolation from the sample to the population.

Recall that in univariate statistics the sample mean was used to make inference about the population mean. Similarly the sample fitted regression line $\widehat{y} = b_1 + b_2 x$ can be used to make inference about the population line.

Different samples will lead to different fitted regression lines, due to different random departures in the data from the population conditional mean. Statistical inferential methods control for this randomness. For example, if the regression slope coefficient is greater than zero in our single sample, suggesting that $y$ increases as $x$ increases, does this necessarily hold in the population? Or is it just this particular sample that has this relationship?

This chapter presents the statistical properties of the least squares estimates. The subsequent chapter uses these properties to construct confidence intervals and perform hypothesis tests.

## 9.1 Population and Sample

For bivariate data we relate the sample fitted regression line to the population line, defined next. Interest lies especially in using the estimated slope coefficient of the fitted regression to make inference about the (unknown) slope of the population line.

### 9.1.1 Population Line or Population Conditional Mean

We wish to estimate the population relationship between $y$ and $x$. That is, we would like to know what value of $y$ to expect on average for each possible value that the regressor $x$ might

take.

Formally this is the **population conditional mean** of $Y$ given $X = x$, denoted $\mathrm{E}[Y|X = x]$, the **expected value** of $Y$ given $X = x$. This is the probability-weighted average of all possible values of $Y$ in the population given $X = x$. The word "conditional" is added to indicate that the value of the mean of $Y$ varies with the value $x$ of the regressor.

Appendix 9.A provides more detail on conditional distributions and the conditional mean. It presents a specific example in which $X$ can take two values (0 or 1) and the conditional mean of $Y$ when $X = 1$ equals 40, whereas the conditional mean of $Y$ when $X = 0$ takes the different value of 15.

For linear regression it is assumed that **conditional mean** of $Y$ given $X = x$ is a **linear** function of $x$, so

$$\mathrm{E}[Y|X = x] = \beta_1 + \beta_2 x.$$

Here $\beta$ is "beta", the Greek letter $b$. Then $\beta_1$ is the **population intercept parameter** and $\beta_2$ is the **population slope parameter**. The OLS coefficients $b_1$ and $b_2$ are the sample estimates of $\beta_1$ and $\beta_2$.

The linear function is a simplification as the conditional mean can in general be a quite complicated function of $x$. Nonlinear relationships are studied in Chapter 12.

In the econometrics literature it is standard to write $\mathrm{E}[Y|X = x]$ more simply as $\mathrm{E}[y|x]$. Thus we write that the **conditional mean** or **conditional expectation function** or **population line** is

$$\mathrm{E}[y|x] = \beta_1 + \beta_2 x.$$

The goal is to estimate the population line, as then we can make predictions and estimate the effect of changing $x$ on the conditional mean $\mathrm{E}[y|x]$.

**Remark 62** *$E[y|x]$ denotes the conditional mean of $Y$ given $X = x$, the probability-weighted average of all possible values of $Y$ in the population given $X = x$. For linear regression $E[y|x] = \beta_1 + \beta_2 x$.*

### 9.1.2 Error Term

The dependent variable $y$ does not exactly equal $\beta_1 + \beta_2 x$, it only equals $\beta_1 + \beta_2 x$ on average (more precisely, in expectation).

The difference between $y$ and $\beta_1 + \beta_2 x$ is viewed as arising due to an **error term** or **disturbance term** that is denoted $u$. Thus we suppose that the data on $y$ are generated by the **population model**

$$y = \beta_1 + \beta_2 x + u.$$

The error $u$ is not observed. It can be positive, leading to a $y$ value above the population line, or negative, in which case the $y$ value lies below the population line.

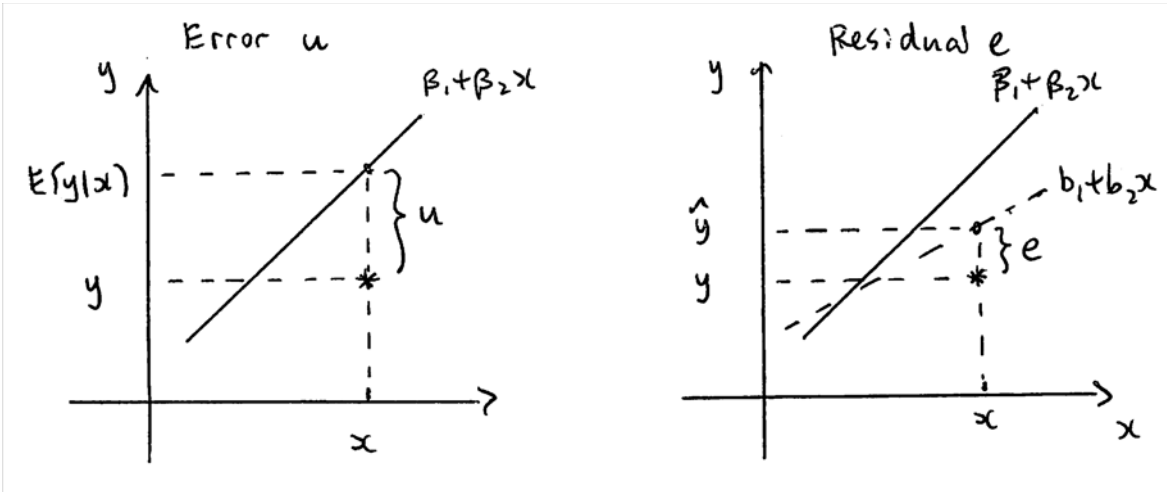


Figure 9.1: Error is difference between y and the population line while residual is difference between y and fitted regression line.

The first panel of Figure 9.1 shows a population line $\beta_1 + \beta_2 x$. The actual value of $y$ is the value predicted by the line plus the error $u$. For the displayed observation the error $u$ is negative, leading to value $y$ that is less than $\beta_1 + \beta_2 x$.

To ensure that the model $y = \beta_1 + \beta_2 x + u$ implies that $\mathrm{E}[y|x] = \beta_1 + \beta_2 x$ we need to assume that

$$\mathrm{E}[u|x] = 0,$$

so the **conditional mean of the error is zero**. This means that for each value taken by the regressor variable $x$, on average the error term equals zero.

### 9.1.3 Error Term versus Residual

The error, denoted $u$, should not be confused with the residual, denoted $e$, obtained from regression using the sample data. The error term is unobserved, whereas the residual can be computed given $y$ and the fitted value $\widehat{y}$.

The second panel of Figure 9.1 adds a fitted regression line $b_1 + b_2 x$. Due to random sampling this will differ from the population line. The residual $e$ is the difference between the fitted value $\widehat{y} = b_1 + b_2 x$ and the actual value $y$. For the displayed observation the residual is negative as the observed value $y$ is less than $\widehat{y}$.

In this example the residual $e$ is also negative and is smaller in absolute value than the error $u$. More generally the residual may be smaller or larger than the error and need not be of the same sign.

**Remark 63** *The error term $u$ is the difference between $y$ and the population line $\beta_1 + \beta_2 x$ where $\beta_1$ and $\beta_2$ are unknown population intercept and slope parameters. The residual $e$ is the difference between $y$ and the fitted regression line $b_1 + b_2 x$ where $b_1$ and $b_2$ are intercept and slope coefficients estimated by OLS regression.*

### 9.1.4 Population Conditional Variance

The error term $u$ provides the only variation in the distribution of $Y$ given $X = x$. It follows that the **population conditional variance** of $Y$ given $X = x$, denoted simply as $\text{Var}[y|x]$, equals the conditional variance of $u$ given $x$. Thus

$$\text{Var}[y|x] = \text{Var}[u|x].$$

The simplest models assume that this variance does not vary with $x$. This means that for each value taken by the regressor variable $x$ there is the same variation around the population line. This constant variance is denoted $\sigma_u^2$.

**Remark 64** *The conditional variance of $Y$ given $X = x$ is a measure of the variance of $Y$ around the population line (for given $X = x$). This conditional variance equals $\sigma_u^2 = Var[u|x]$.*

### 9.1.5 Summary

In the univariate case it was assumed that $y_1, ..., y_n$ is a simple random sample from

$$Y \sim (\mu, \sigma^2),$$

and inference on $\mu$ is based on the sample mean $\bar{y}$.

For bivariate regression it is assumed that $(x_1, y_1), ..., (x_n, y_n)$ is a sample with data independent across observations and that

$$Y|X = x \sim (\beta_1 + \beta_2 x_i, \sigma_u^2).$$

Inference on $\beta_1$ and $\beta_2$ is based on the least squares estimates $b_1$ and $b_2$.

**Remark 65** *Compared to univariate analysis, bivariate regression allows the population mean of $Y$ to vary with the value taken by $X$, according to a linear relationship.*

Table 9.1: Generated data: Model y = 1 + 2x + u where u is N(0, 4) distributed.

| Observation | x | E[y\|x]=1+2x | u | y=1+2x+u |
|---|---|---|---|---|
| 1 | 1 | 3 | 1.689889 | 4.689889 |
| 2 | 2 | 5 | -.3187171 | 4.681283 |
| 3 | 3 | 7 | -2.506667 | 4.493333 |
| 4 | 4 | 9 | -1.63328 | 7.366720 |
| 5 | 5 | 11 | -2.390764 | 8.609236 |

## 9.2 Examples of Sampling from a Population

This section begins by generating a single example of five observations to make clear the difference between the population line and the sample line. This is followed by two examples that illustrate how different samples from the same population lead to different fitted regression lines. One example uses artificially generated data, while the other randomly draws observations from a well-defined finite population.

These examples demonstrate the more general result that

- the regression coefficients $b_1$ and $b_2$ on average are close to the population slope parameters $\beta_1$ and $\beta_2$
- the regression coefficients are approximately normally distributed, provided the sample size is sufficiently large.

### 9.2.1 Single Sample Generated from an Experiment

Suppose that five observations on $y$ and $x$ are generated by the model

$$\begin{aligned} y_i &= 1 + 2x_i + u_i, \quad i = 1, ..., 5 \\ u_i &\sim N(0, 4) \text{ independent over } i \\ x_1 &= 1,\ x_2 = 2, ....,\ x_5 = 5. \end{aligned}$$

Here the population line is $\mathrm{E}[y|x] = 1 + 2x$ and the error term $u$ has mean 0 and variance 4 that does not vary with $x$. There are five observations corresponding to values of $x$ equal to 1, 2, ..., 5.

For the first observation, with $x_1 = 1$, suppose we draw error $u_1 = 1.689889$ from the $N(0, 4)$ distribution. Then $y_1 = 1 + 2 \times 1 + 1.689889 = 4.689889$. Other observations are similarly obtained, leading to the sample given in Table 9.1.

The first panel of Figure 9.2 presents the true regression line $\mathrm{E}[y|x] = 1 + 2x$ and the five generated observations for $y = 1 + 2x + u$. Note that the sample points deviate from

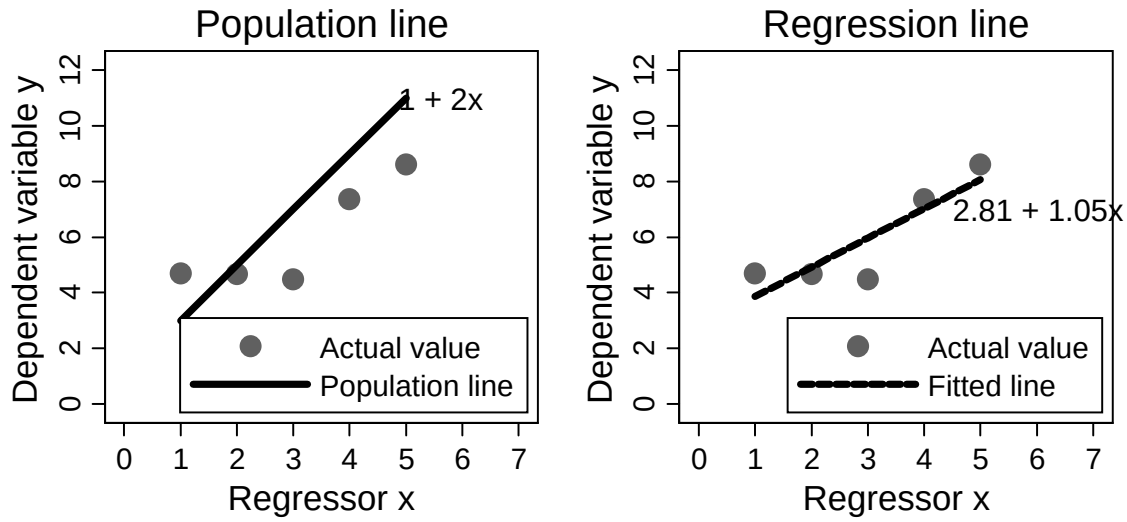


Figure 9.2: Generated data: Scatterplot with true population line and with fitted regression line.

the population regression line, due to the error term $u$. In this case the first observation $(x, y) = (1, 4.69)$ is above the line due to a positive error $u$, while the other four observations are below the population line due to negative draws of $u$.

Generated data: Scatterplot with true population line and with fitted regression line.

The second panel of Figure 9.2 presents the fitted regression line $b_1 + b_2 x \simeq 2.81 + 1.05x$, obtain by OLS regression for the data listed in Table 9.1, along with the five data points. Clearly the fitted regression line differs from the population regression line.

The difference between the fitted regression line and the population regression line is due to sampling variability. The challenge is to make inference about the true regression line, which in practice is unknown, controlling for this sampling variability.

### 9.2.2 Many Samples Generated from an Experiment

Now suppose that many samples, each of 30 observations, are generated by the model

$$\begin{aligned} y_i &= 1 + 2x_i + u_i \quad i = 1, ..., 30 \\ u_i &\sim N(0, 4) \text{ independent over } i \\ x_i &\sim N(3, 1). \end{aligned}$$

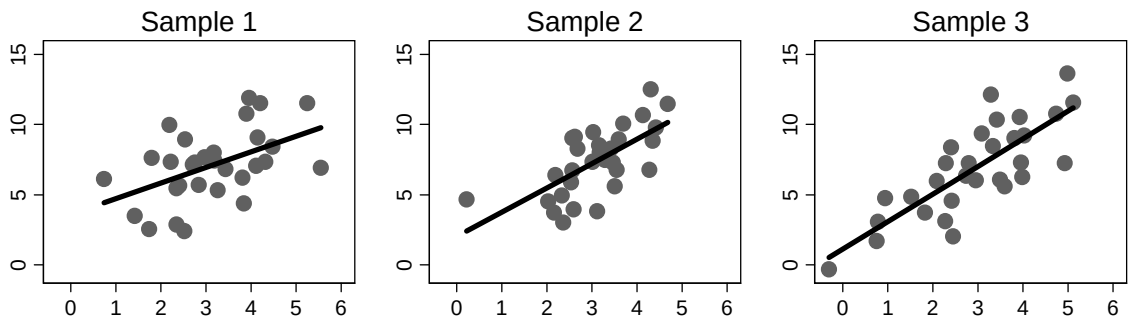


Figure 9.3: Generated data: Scatterplots and fitted regression lines from three random samples (n = 30).

Then $y = \beta_1 + \beta_2 x + u$ where $\beta_1 = 1$, $\beta_2 = 2$ and $\sigma_u^2 = 4$. This is similar to the preceding example, except now $x$ takes a wider range of values than simply $1, 2, ..., 5$.

Figure 9.3 presents scatter plots of the data, and the associated fitted regression lines, for three randomly generated samples of size $n = 30$ from this population model. Clearly the data, and the consequent estimated intercept and slope coefficients $b_1$ and $b_2$, vary across the samples.

This exercise is repeated 400 times, yielding 400 samples of size 30 and hence 400 estimated slope coefficients and 400 intercepts.

The left panel of Figure 9.4 presents a histogram for the 400 estimated slope coefficients $b_2$. The histogram is roughly centered on the population slope coefficient $\beta_2 = 2$; in fact the average of the 400 estimated slopes is equal to 1.979.

The right panel of Figure 9.4 is a similar histogram for the 400 intercepts $b_1$, with average 1.039 close to the population value $\beta_1 = 1$.

It appears that on average the least squares coefficients are close to their population values. Furthermore, the histograms and kernel density estimates suggest that the least squares coefficients are approximately normally distributed.

### 9.2.3 Many Samples from a Finite Population: 1880 U.S. Census

The 1880 Census provides a complete enumeration of the U.S. population in 1880. We consider the relationship between labor force participation and age for males aged sixty to seventy years. The variable *Lfp* takes value 1 if in the labor force and value 0 if not in the labor force. The variable *Age* is age in years.

The Census gives the population. There are 1,058,475 males aged 60-70 years, of whom

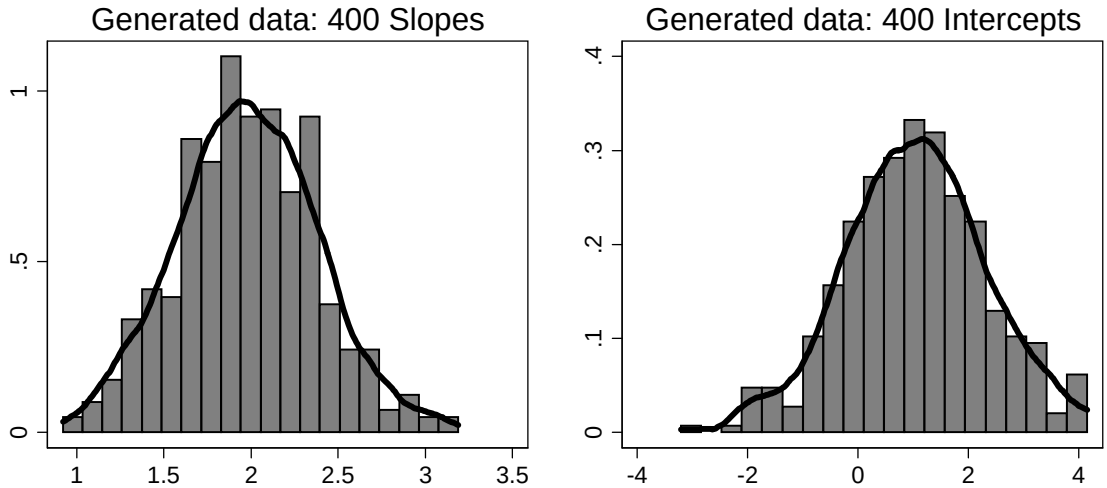


Figure 9.4: Generated data: Histogram with kernel density estimates for slopes and intercepts from 400 regressions (n=30).

89.45% were in the labor force. The population line is simply the regression line fitted using all 1,058,475 population values. This yields

$$\mathrm{E}[Lfp|Age] = 1.593 - 0.0109 \times Age,$$

so $\beta_1 = 1.593$ and $\beta_2 = -0.0109$. The negative coefficient on variable *Age* means that labor force participation declines with age, as expected. The population mean labor force participation rate is $1.593 - 0.0109 \times 60 = 0.939$ or 93.9% at age 60 and falls to 0.830 or 83.0% at age 70.

Now we take random samples of size 200 from this population and obtain estimates $b_1$ and $b_2$ for each sample. This corresponds to simple random sampling from the population. The larger sample size of 200 is used in this example because the fit for this regression is poor, with typical $R^2 \simeq 0.01$. The first three estimated samples yield regression estimates $1.708 - 0.0133 \times Age$, $1.535 - 0.0102 \times Age$, and $1.174 - 0.0040 \times Age$.

400 samples of size 200 were randomly drawn from the Census data and fitted. This yields 400 fitted slopes and 400 intercepts.

The left panel of Figure 9.5 presents a histogram for the 400 estimated slope coefficients $b_2$. The histogram is roughly centered on the population slope coefficient $\beta_2 = -0.0109$; in fact the average of the 400 slopes is equal to $-0.0104$.

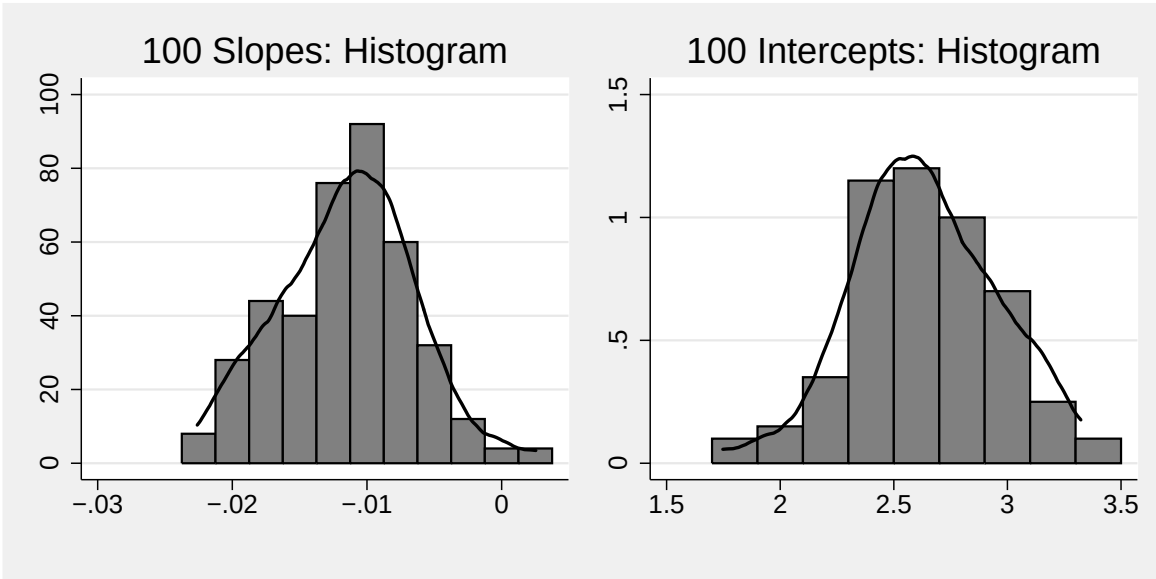


Figure 9.5: 1880 Census data: Histogram with kernel density estimates for slopes and intercepts from 400 regressions (n=200). NOTE: PANELS SAY 100 IN ERROR – IT WAS 400.

The right panel of Figure 9.5 is a similar histogram for the 400 intercepts $b_1$, with average 1.560 close to the population value $\beta_1 = 1.593$.

It again appears that on average the least squares coefficients are close to their population values. And the histograms and kernel density estimates suggest that the least squares coefficients are approximately normally distributed.

## 9.3 Properties of the Least Squares Estimator

From Chapter 8, the least squares slope and intercept estimates are

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ b_1 &= \bar{y} - b_2 \bar{x}. \end{aligned}$$

Alternative notation that is also commonly-used is to denote $b_1$ and $b_2$ as $\widehat{\beta}_1$ and $\widehat{\beta}_2$, since they are viewed as estimates of $\beta_1$ and $\beta_2$ and a carat (ˆ) denotes estimate.

In this section we detail the properties of the parameter $b_2$ as estimates of $\beta_1$ and $\beta_2$. Results are simply stated here. More detailed derivations are given in Appendix 9.C.

### 9.3.1 Data Assumptions

Throughout it is assumed that there is some variation in the regressors, so that $\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0$. Otherwise it is not possible to compute the OLS coefficients $b_1$ and $b_2$.

Additionally it is assumed that there are at least three observations. Otherwise it is not possible to compute the standard error of the regression.

The sampling process can be simple random sampling, where $(x_i, y_i)$ are jointly sampled from the population. Or the $x's$ may be fixed in advance, as is the case in a controlled experiment, with just the $y's$ randomly determined.

### 9.3.2 Population Assumptions

To obtain the statistical properties of the OLS coefficients we need to make assumptions about the population model and the sampling process that yielded the sample $(x_1, y_1), ...., (x_n, y_n)$.

Standard assumptions are that

1. The **population model** is

$$y_i = \beta_1 + \beta_2 x_i + u_i \text{ for all } i.$$

2. The **error for the $i^{th}$ observation has zero mean conditional on the regressor**:

$$\mathrm{E}[u_i|x_i] = 0 \text{ for all } i.$$

3. The **error for the $i^{th}$ observation has constant variance conditional on the regressor**:

$$\mathrm{Var}[u_i|x_i] = \sigma_u^2 \text{ for all } i.$$

4. The **errors for different observations are statistically independent**

$$u_i \text{ is independent of } u_j \text{ for all } j \neq i.$$

Assumptions 1-2 are the crucial assumptions that ensure that

$$\mathrm{E}[y_i|x_i] = \beta_1 + \beta_2 x_i,$$

so that the data are generated by a linear population relationship. In that case it makes sense to fit the data with a regression line.

Assumptions 3 and 4 are additional assumptions that are used in determining the precision and distribution of the estimates $b_1$ and $b_2$.

### 9.3.3 How Reasonable are the Assumptions?

Assumption 1 specifies the basic relationship to be **linear**. Least squares regression can be adapted to nonlinear relationships, however; see Chapter 12.

Assumption 2 is crucial, even in more general models than the single regressor linear model of assumption 1. It means that for any value of $x$, the error $u$ on average equals zero. This crucial assumption rules out, for example, the possibility of high values of $x$ being associated with high values of $u$.

To investigate the reasonableness of assumption 2, consider regression of earnings on years of schooling. In that case the error term includes all things other than schooling that may effect the level of earnings, such as innate ability. People with high ability might be expected to have, on average, higher levels of schooling ($x$). But even given their level of schooling, they might be expected to have an additional boost to their earnings due to innate activity, in which case the error ($u$) is positive. This suggests that, for larger values of $x$, $\mathrm{E}[u|x] > 0$. This is a violation of assumption 2.

It can be shown that assumption 2 implies that the error term $u$ is uncorrelated with the regressor $x$, so

$$\mathrm{Cor}[u_i, x_i] = 0.$$

So an alternative way to state assumption 2 is that the **error term is uncorrelated with the regressor**.

Assumption 2 can be accommodated somewhat by adding additional regressors in the model, in the hope that after doing so the remaining error is uncorrelated with the regressors. In the earnings-schooling example we could include as an additional regressor a variable measuring innate ability, if such data were available, and use multiple regression presented in later chapters. Then the remaining error may be uncorrelated with the regressors.

Throughout chapters 1-16 it is assumed that assumption 2 holds. Chapter 18 on causality discusses relaxation of assumption 2.

Assumption 3 implies that regardless of the value taken by $x$, the error term $u$ has conditional (on $x$) variance that does not vary with $x$. This rules out, for example, that in a regression of earnings on schooling, the variability in earnings is greater for those with high levels of schooling than it is for low levels of schooling.

Assumption 3 is called the assumption of conditionally **homoskedastic errors**, where the term "skedastic" is based on the Greek word for scattering and "homos" is the Greek word for same.

Note that there is an asymmetry. Assumptions 1-2 imply that the conditional mean varies with $x$ while assumption 3 is that the conditional variance does not vary with $x$. Assumption 3 can easily be relaxed, however; see Chapter 10.7.

Assumption 4 means that the value taken by the error $u$ for one observation is unrelated to the value of the error for other observations. This rules out, for example, a time series model for GDP where, say, a positive error in one period (so GDP is unusually high given

the regressor) is quite likely to be followed by a positive error in the subsequent period. Assumption 4 can be relaxed, however; see Chapter 10.7.

**Remark 66** *Population assumptions 1 and 2 are essential assumptions that imply that* $E[y|x] = \beta_1 + \beta_2 x$. *Population assumptions 3-4 are used to obtain an estimate of the precision of* $b_1$ *and* $b_2$, *and can be relaxed. Additionally it is assumed that there is some variation in the regressors and that* $n \geq 3$.

### 9.3.4 Mean and Variance of the Least Squares Slope Coefficient

Given assumptions 1-2 the **mean of the least squares slope coefficient** $b_2$ is

$$\mathrm{E}[b_2] = \beta_2.$$

So $b_2$ is unbiased for $\beta_2$.

This result means that if we were able to obtain many random samples and for each sample estimate the slope coefficient, then on average the slope coefficients equal the population slope coefficient.

If we additionally make assumptions 3-4 then the **variance of the least squares slope coefficient** $b_2$ is

$$\mathrm{Var}[b_2] \equiv \sigma^2_{b_2} = \frac{\sigma^2_u}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Assumptions 3-4 can be relaxed, in which case we get a different expression for $\mathrm{Var}[b_2]$.

**Remark 67** *Under assumptions 1-2 the estimated slope coefficient* $b_2$ *has mean equal to the population slope parameter* $\beta_2$. *Under assumptions 1-4 the estimated slope coefficient* $b_2$ *has variance* $\sigma^2_{b_2} = \sigma^2_u / \sum_{i=1}^n (x_i - \bar{x})^2$.

### 9.3.5 Standard Error of the Least Squares Slope Coefficient

The variance of $b_2$ depends on $\sigma^2_u$, the variance of the error term, which is unknown.

Since $\sigma^2_u = \mathrm{E}[(y - \beta_1 - \beta_2 x)^2]$, an obvious estimate of $\sigma^2_u$ is the average of $(y_i - \widehat{y}_i)^2$. Under assumptions 1-4 it can be shown that an unbiased estimated of $\sigma^2_u$ is obtained if the divisor in the average is $(n-2)$. Thus we estimate $\sigma^2_u$ by $s^2_e$ where

$$s_e = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \widehat{y}_i)^2},$$

introduced in Chapter 8.6, is called the **standard error of the regression** or the **root mean square error**.

Replacing $\sigma_u^2$ by $s_e^2$ in the formula for $\text{Var}[b_2]$, and then taking the square root, the estimated standard deviation of $b_2$, called the **standard error of the slope coefficient** is

$$se(b_2) = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Under assumptions 1-4, $\text{E}[se(b_2)^2] = \text{Var}[b_2]$, so $se(b_2)^2$ is unbiased for $\text{Var}[b_2]$.

As an example, consider the artificial data on a sample of size five, introduced in Appendix 8.A. Then the regression line was $\widehat{y} = 0.8 + 0.4x$, $s_e = 0.365148$, and $\sum_{i=1}^n (x_i - \bar{x})^2 = 10$. It follows that $se(b_2) = 0.365148/\sqrt{10} = 0.115$.

### 9.3.6 When is the Slope Coefficient Precisely Estimated?

The standard error $se(b_2)$ measures the precision of $b_2$ as an estimate of $\beta_2$. Precision improves as $se(b_2)$ gets smaller. Since $se(b_2) = s_e/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$, it follows that the precision of estimation of $\beta_2$ is better

1. the closer the data are to the true regression line, as then $s_e$ is small;

2. the larger is the sample size $n$, as then $\sum_{i=1}^n (x_i - \bar{x})^2$ has more terms and is larger;

3. the more widely scattered are the regressors $x$ about their mean $\bar{x}$, as then $\sum_{i=1}^n (x_i - \bar{x})^2$ is larger.

The second property leads to an inverse square root rule for precision. If the sample size $n$ quadruples, the sum $\sum_i (x_i - \bar{x})^2$ approximately quadruples, its reciprocal is approximately one-fourth as large and, taking the square root, $se(b_2)$ is approximately halved. More generally, if the sample is $m$ times larger then the standard error of $b_2$ is $1/\sqrt{m}$ times as large. Note that this is similar to the result for the sample mean which had standard error $s/\sqrt{n}$.

The third property suggests that precision is best if the regressors are widely scattered. This is used in disciplines where the researcher can determine the regressor values, such as in the design of experiments.

**Remark 68** *The standard error of $b_2$, the estimated standard deviation of $b_2$, is denoted $se(b_2)$ and equals $s_e/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$ under assumptions 1-4, where $s_e$ is the standard error of the regression. Under assumptions 1-4 $se(b_2)^2$ has mean equal to the variance of $b_2$. Bigger samples are better – if the sample is $m$ times larger then $se(b_2)$ is approximately $1/\sqrt{m}$ times as large. And wider dispersion of the regressor is better.*

### 9.3.7 Normal Distribution and the Central Limit Theorem

From the left panels of Figures 9.4 and 9.5, the slope coefficients appear to be approximately normally distributed. This is the case more generally, provided the sample size is sufficiently large.

The preceding results for the mean and variance of $b_2$ imply that under assumptions 1-4

$$b_2 \sim (\beta_2, \sigma^2_{b_2}).$$

Standardizing, by subtracting the mean and dividing by the standard deviation, yields a random variable with mean 0 and variance 1. Denoting this random variable by Z, we have

$$Z = \frac{b_2 - \beta_2}{\sigma_{b_2}} \sim (0, 1).$$

If additionally we assume that the sample size $n \to \infty$, then $Z$ has the **standard normal distribution** by the **central limit theorem** (see Appendix 5.B). Thus

$$Z \sim N(0, 1) \text{ as } n \to \infty.$$

The least squares slope coefficient $b_2$ itself is **asymptotically normally distributed**, with mean $\beta_2$ and standard deviation $\sigma_{b_2}$, where the term "asymptotically normal" means normally distributed as $n \to \infty$.

How large should $n$ be to use this result? In the univariate case it is felt that $n > 30$ is sufficient. For regression $n$ should be larger, but there is no clear guide.

**Remark 69** *Under assumptions 1-4 the standardized random variable $Z = (b_2 - \beta_2)/\sigma_{b_2}$ is standard normally distributed as the sample size goes to infinity.*

### 9.3.8 The Least Squares Intercept Coefficient

The preceding results consider the slope coefficient. Similar results, including the following, can be obtained for the intercept.

Under assumptions 1-2 the intercept estimate $b_1 = \bar{y} - b_2\bar{x}$ has the property that $\mathrm{E}[b_1] = \beta_1$. So $b_1$ is unbiased for $\beta_1$.

Given assumptions 1-4 the variance of $b_1$ is $\mathrm{Var}[b_1] = \sigma_u^2 \sum_{i=1}^n {x_i}^2 / [n \sum_{i=1}^n (x_i - \bar{x})^2]$. The standard error of the intercept estimate $b_1$ is then

$$se(b_1) = \sqrt{\frac{s_e^2 \sum_{i=1}^n {x_i}^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}},$$

where $s_e$ is the standard error of the regression.

If additionally $n \to \infty$ then $Z = (b_2 - \beta_2)/\sigma_{b_2} \sim N(0, 1)$, where $\sigma^2_{b_2}$ is the variance of $b_2$.

## 9.4 t-Statistic

The development of the $t$-statistic for inference on $\beta_2$ is qualitatively similar to development of the $t$-statistic for inference on $\mu$ given in Chapter 5.5.

The result that $Z = (b_2 - \beta_2)/\sigma_{b_2} \sim N(0,1)$ given assumptions 1-4 and $n \to \infty$ requires knowledge of $\sigma_{b_2}$, the population standard deviation of $b_2$. In practice this is not known, as it depends in part on the standard deviation of the error $\sigma_u$.

Instead $\sigma_{b_2}$ is replaced by its estimate $se(b_2)$. This adds noise, leading to a statistic that is approximated by the $t$ distribution, a distribution that has fatter tails than the standard normal.

### 9.4.1 t-Statistic

Replacing $\sigma_{b_2}$ with its estimate $se(b_2)$ yields the **$t$-statistic**. We write

$$t = \frac{b_2 - \beta_2}{se(b_2)} \sim T(n-2),$$

where $T(n-2)$ denotes the $t$ distribution with $n-2$ degrees of freedom. More formally, $t$ is the realization of a random variable that is $T(n-2)$ distributed. The $t$ distribution is discussed in some detail in Chapter 6.2.

It can be shown that this result is exact if assumptions 1-4 hold and additionally the errors are normally distributed. If the errors are nonnormal then the $T(n-2)$ distribution is a good approximation provided $n$ is sufficiently large, but there is no clear-cut rule on how large $n$ needs to be.

A less common alternative uses the standard normal distribution for $t$. This has the advantage of simplicity. It coincides with using the $T(n-2)$ distribution for very large $n$. For small and moderate $n$, however, using the standard normal distribution leads to confidence intervals that may be too narrow and to tests that may over-reject.

**Remark 70** *Under assumptions 1-4 the t-statistic $t = (b_2 - \beta_2)/se(b_2)$ is the realization of a randomly variable that is $T(n-2)$ distributed, exactly if data are normally distributed and approximately for nonnormal data if $n$ is sufficiently large.*

### 9.4.2 Summary for the Least Squares Slope Coefficient

A summary given assumptions 1-4 is the following.

1. Individual $y_i$ given $x_i$ are assumed to have conditional mean $\beta_1 + \beta_2 x_i$ and conditional variance $\sigma_u^2$.

2. The slope coefficient $b_2$ from a sample of size $n$ has mean $\beta_2$ and variance $\sigma_b^2 = \sigma_u^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.

3. The standardized statistic $Z = (b_2 - \beta_2)/\sigma_b$ has mean 0 and variance 1.

4. As sample size $n \to \infty$, $Z$ is standard normal distributed by the central limit theorem.

5. Replacing the unknown $\sigma_b$ by $s_b$, the standard error of the slope coefficient, leads to (approximately) a $t$ distribution with $(n-2)$ degrees of freedom, denoted $T(n-2)$.

6. The sample $t$-statistic $t = (b_2 - \beta_2)/s_b$ is a realization of a random variable that is approximately $T(n-2)$ distributed.

7. For normally distributed data this last result is exact. For nonnormal data the approximation improves as $n$ gets larger.

## 9.5 Point Estimation

Desirable properties of the **point estimates** $b_1$ and $b_2$ are that they be centered on the population parameters $\beta_1$ and $\beta_2$, and that they be as precise as possible, meaning that their variance be as small as possible. These considerations are qualitatively similar to those presented in detail in Chapter 5.6 for using the sample mean to estimate the population mean.

The least squares estimates have the desirable properties of being unbiased under assumptions 1-2 and consistent under additional assumptions, such as assumptions 3-4. Under assumptions 1-4 they are the best estimates, in a qualified sense defined below.

### 9.5.1 Unbiased and Consistent Estimator

Under assumptions 1-2, $\mathrm{E}[b_1] = b_1$ and $\mathrm{E}[b_2] = b_2$, so the least squares estimates are **unbiased estimators**. This means that if we obtain many samples and hence many estimates $b_1$ and $b_2$, on average these estimates will equal $\beta_1$ and $\beta_2$.

With a sample of very large size, ideally estimation can be so precise that the distributions of $b_1$ and $b_2$ collapse on the points $\beta_1$ and $\beta_2$. More precisely, $b_2$ is **consistent** for $\beta_2$ if the probability that $|b_2 - \beta_2|$ exceeds $\varepsilon$ goes to zero as the sample size goes to infinity, no matter how small $\varepsilon > 0$ is chosen to be. A sufficient condition for this to happen is that $b_2$ be unbiased and the variance of $b_2$ goes to zero as $n \to \infty$. This is the case under assumptions 1-4, provided additionally that the regressors are such that $\sum_{i=1}^n (x_i - \bar{x})^2$ goes to infinity as progressively more observations are added, since then $\mathrm{Var}[b_2] = \sigma_u^2 / [\sum_{i=1}^n (x_i - \bar{x})^2]$ goes to zero as $n \to \infty$. Similarly $b_1$ is consistent for $\beta_1$ under these assumptions.

**Remark 71** *The OLS estimators $b_1$ and $b_2$ are unbiased for $\beta_1$ and $\beta_2$ under assumptions 1-2. And $b_1$ and $b_2$ are consistent for $\beta_1$ and $\beta_2$ under assumptions 1-2 plus any additional assumptions, such as assumptions 3-4, that ensure that the variances of the estimates exist and go to zero as $n \to \infty$.*

### 9.5.2 Best Estimator

The least squares estimates $b_1$ and $b_2$ are not the only unbiased and consistent estimates of $\beta_1$ and $\beta_2$. As in Chapter 5.6, ideally we should use the **best estimator** or **efficient estimator,** which is the estimator with **minimum variance** in the class of estimators under consideration, such as unbiased estimators or consistent estimators.

First, restrict attention to estimators that are weighted averages of the dependent variable $y$, of the form

$$b_2 = \sum\nolimits_{i=1}^{n} w_i y_i.$$

The least squares slope coefficient is a special case with $w_i = (x_i - \bar{x}) / \sum_{j=1}^{n} (x_j - \bar{x})^2$. It can be shown that under assumptions 1-4, the least squares estimator has the smallest variance of all such weighted average estimators. The weighted average is a linear combination of $y_1, ..., y_n$. So under assumptions 1-4 the least squares estimator $b_2$ is the **best linear unbiased estimator** (or **BLUE**) of $\beta_2$, a result called the **Gauss-Markov Theorem**. A similar result holds for $b_1$ as an estimator of $\beta_1$.

Second, if we make assumptions 1-4 and the additional assumption that the errors are normally distributed then the least squares estimates of $b_1$ and $b_2$ can be shown to have the lowest variances of any unbiased estimators of $\beta_1$ and $\beta_2$. So under assumptions 1-4 and the assumption of normal errors the least squares estimators are the **best unbiased estimators** (or **BUE**) .

Third, if we consider large samples and make assumptions 1-4 then the least squares estimators $b_1$ and $b_2$ are the **best estimators** in the following sense. They have the smallest variance of all estimators of $\beta_1$ and $\beta_2$ that are consistent and asymptotically normal distributed. This essentially means that in large sample OLS is best given assumptions 1-4, as limiting attention to estimators that are asymptotically normally distributed is not a big restriction.

**Remark 72** *The OLS estimators $b_1$ and $b_2$ are best linear unbiased under assumptions 1-4, and best unbiased if additionally the more errors are normally distributed. In large samples $b_1$ and $b_2$ are best among consistent and asymptotically normal estimators under assumptions 1-4.*

### 9.5.3 Least Squares in Practice

Assumptions 1-2 are essential for the least squares estimators to be unbiased and consistent. These assumptions are sufficient for unbiasedness, and not much more is needed for consistency. If these assumptions fail then alternative estimation methods are needed. Examples include failure of assumption 1 due to a nonlinear relationship, and failure of assumption 2 due to regressors correlated with the error.

This introductory text focuses on settings where at least assumptions 1-2 hold, so least squares is a valid procedure. But the formulas presented for the standard errors, and hence the values of the resultant confidence intervals and test statistics, depend on the additional assumptions 3-4. In Chapter 10.7 we discuss alternative formulas for the standard errors under variations of assumptions 3-4 that are often more reasonable assumptions for the economics data analyzed in practice.

## 9.6 Key Concepts

1. The conditional mean of $Y$ given $X = x$, denoted, $\mathrm{E}[Y|X = x]$ denotes the conditional mean of $Y$ given $X = x$, the probability-weighted average of all possible values of $Y$ in the population given $X = x$.

2. The simpler notation $\mathrm{E}[y|x]$ is typically used in the econometrics literature.

3. For linear regression $\mathrm{E}[y|x] = \beta_1 + \beta_2 x$.

4. Statistical inferential methods extrapolate the sample OLS estimates $b_1$ and $b_2$ of the regression line to the population parameters $\beta_1$ and $\beta_2$.

5. The error term $u$ leads to randomness of $y$ around the population line $\beta_1 + \beta_2 x$.

6. The sample must be such that there is some variation in the regressors.

7. The key population assumptions are 1 and 2 that $y = \beta_1 + \beta_2 x + u$ and $\mathrm{E}[u|x] = 0$.

8. Additional population assumptions are needed to obtain the standard error and distributions of the estimates.

9. The simplest additional assumptions are assumptions 3-4 that $\mathrm{Var}[u|x] = \sigma_u^2$ does not vary with $x$ (homoskedasticity) and that the errors $u$ are statistically independent.

10. Under assumptions 1 and 2, $b_1$ and $b_2$ are unbiased and, with additional assumptions such as assumptions 3 and 4, consistent estimates of $\beta_1$ and $\beta_2$.

11. Under assumptions 1-4, least squares is the best linear unbiased estimator. If additionally the errors are normally distributed then least squares is the best estimator.

12. The slope coefficient standard error $se(b_2)$ is smaller the better the regression line fits the data, the larger the sample, and the greater the variability of the regressors in the sample.

13. Statistical inference on $\beta_2$ is based on the $t$-statistic $T = (\beta_2 - b_2)/se(b_2)$.

14. The $t$-statistic is treated as being $T(n-2)$ distributed, exactly if data are normally distributed and approximately for nonnormal data if $n$ is sufficiently large.

15. Key Terms: Population line; parameter; conditional mean; conditional variance; error term; disturbance term; assumptions 1-4; homoskedastic errors; ordinary least squares; standard error of the slope coefficient; standard error of the regression; $t$-statistic; $t$ distribution; degrees of freedom; unbiased; best linear unbiased; BLUE; best unbiased; consistent; asymptotically normal.

## 9.7 Exercises

To come.

## 9.8 Appendix 9.A: Conditional Mean and Variance

This appendix provides a more detailed definition of conditional mean and conditional variance, introduced in Chapter 9.1.

### 9.8.1 Joint Probabilities

Consider the statistical relationship between two random variables $X$ and $Y$ that are both discrete random variables. Their joint probability of occurrence is defined by the **joint probability mass function**

$$\Pr[X = x, Y = y], \quad x = x_1, x_2, ..., \quad y = y_1, y_2, ...$$

where upper case denotes the random variable and lower case denotes the values that the random variable might take.

Given knowledge of the joint probabilities of $X$ and $Y$ we can obtain the separate probabilities for $X$ and for $Y$. For example, $\Pr[X = x]$ equals the sum over all possible values of $y$ of the joint probability $\Pr[X = x, Y = y]$.

### 9.8.2 Example

Throughout this appendix we consider the following example:

$$\Pr[X = x, Y = y] = \begin{cases} 0.1 & \text{for } x = 1,\ y = 50 \\ 0.1 & \text{for } x = 1,\ y = 30 \\ 0.2 & \text{for } x = 0,\ y = 30 \\ 0.6 & \text{for } x = 0,\ y = 10. \end{cases}$$

Note that these probabilities sum to one.

For this example, $\Pr[X = 1] = \Pr[X = 1, Y = 50] + \Pr[X = 1, Y = 30] = 0.1 + 0.1 = 0.2$ and $\Pr[X = 0] = 0.2 + 0.6 = 0.8$. So $X = 1$ with probability 0.2 and $X = 0$ with probability 0.8. Thus

$$\Pr[X = x] = \begin{cases} 0.2 & \text{for } x = 1 \\ 0.8 & \text{for } x = 0. \end{cases}$$

By similar summation, now over the possible values of $X$, $\Pr[Y = 50] = 0.1$, $\Pr[Y = 30] = 0.1 + 0.2 = 0.3$, and $\Pr[Y = 10] = 0.6$. So

$$\Pr[Y = y] = \begin{cases} 0.1 & \text{for } y = 50 \\ 0.3 & \text{for } y = 30 \\ 0.6 & \text{for } y = 10. \end{cases}$$

### 9.8.3 Conditional Distribution

A very important result is **Bayes Theorem** that states that for any two events $A$ and $B$, the probability that event $A$ happens, given that $B$ happens, is the joint probability that $A$ and $B$ happen divided by the probability that $B$ happens:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]},$$

where $A \cap B$ means the intersection of events $A$ and $B$.

For example, consider the probability of getting a three on the second toss of a six-sided die (event $A$), given that the combined sum of the first two tosses is five (event $B$). The only way that both $A$ and $B$ can occur is if the first toss was a 2 and the second was a 3. The two tosses can sum to five in four ways, with tosses $(1, 4)$, $(2, 3)$, $(3, 2)$, and $(4, 1)$. All these outcomes are equally likely, so the conditional probability of event $A$ given $B$ is one in four, or 0.25.

More generally, the **conditional probability** of $Y$ given $X$

$$\Pr[Y = y|X = x] = \frac{\Pr[Y = y, X = x]}{\Pr[X = x]}.$$

Consider the earlier numerical example and condition on $X = 1$. Then $Y = 50$ or $Y = 30$ with $\Pr[X = 1, Y = 50] = 0.1$ and $\Pr[X = 1, Y = 30] = 0.1$. Also $\Pr[X = 1] = 0.2$. It follows that the conditional distribution of $Y$ given $X = 1$ is $\Pr[Y = 30|X = 1] = 0.1/0.2 = 0.5$ and $\Pr[Y = 50|X = 1] = 0.1/0.2 = 0.5$. By similar reasoning, since $\Pr[X = 0] = 0.8$, the conditional distribution of $Y$ given $X = 0$ is $\Pr[Y = 10|X = 0] = 0.6/0.8 = 0.75$ and $\Pr[Y = 30|X = 0] = 0.2/0.8 = 0.25$.

The conditional probabilities for $Y$ given $X = 1$ are therefore

$$\Pr[Y = y|X = 1] = \begin{cases} 0.5 & \text{for } y = 50 \\ 0.5 & \text{for } x = 30, \end{cases}$$

and for $Y$ given $X = 0$:

$$\Pr[Y = y|X = 0] = \begin{cases} 0.25 & \text{for } y = 30 \\ 0.75 & \text{for } y = 10. \end{cases}$$

If $X$ and $Y$ are **statistically independent** then the probability of $Y$ taking a particular value is unaffected by the value taken by $X$. In that case the conditional probability $\Pr[Y = y|X = x]$ reduces to the unconditional probability $\Pr[Y = y]$.

In the example of this appendix $X$ and $Y$ are **not** statistically independent.

### 9.8.4 Conditional Mean

The **conditional expected value** of a function $g(Y)$ given $X = x$ is an extension of the usual unconditional expected value of $g(Y)$, except that the values $g(y)$ are weighted by the conditional probabilities of $Y|X = x$ rather than the unconditional probabilities of $Y$. Thus

$$\mathrm{E}[g(Y)|X = x] = g(y_1) \times \Pr[Y = y_1|X = x] + g(y_2) \times \Pr[Y = y_2|X = x] + \cdots$$

The **conditional mean** of $Y$ given $X$ is the mean of the conditional distribution. This is the weighted sum of the possible values of $Y$ where the weighting is by the conditional probability of $Y$ given $x$. So

$$\mathrm{E}[Y|X = x] = \sum\nolimits_y \Pr[Y = y|X = x] \times y.$$

For the example of this appendix begin with the conditional mean of $Y$ given $X = 1$. When $X = 1$, $Y$ takes value 30 with $\Pr[Y = 30|X = 1] = 0.5$, and $Y$ takes value 50 with $\Pr[Y = 50|X = 1] = 0.5$. It follows that $\mathrm{E}[Y|X = 1] = 0.5 \times 30 + 0.5 \times 50 = 40$. By similar calculation $\mathrm{E}[Y|X = 0] = \Pr[Y = 10|X = 0] \times 10 + \Pr[Y = 30|X = 0] \times 30 = 0.75 \times 10 + 0.25 \times 30 = 15$. So

$$\mathrm{E}[Y|X = x] = \begin{cases} 40 & \text{for } x = 1 \\ 15 & \text{for } x = 0. \end{cases}$$

In this example the conditional mean of $Y$ given $X$ varies with the value of $X$.

For linear regression it is assumed that the **conditional mean** $\mathrm{E}[Y|X=x]$ **is a linear function of** $x$. Then the population relationship between $Y$ and $X$ is a line with

$$\mathrm{E}[Y|X=x] = \beta_1 + \beta_2 x.$$

For example, suppose that $\mathrm{E}[Y|X=x] = 3+2x$. Then when $X$ takes the values 1, 2, and 3, for example, the corresponding conditional means are $\mathrm{E}[Y|X=1] = 5$, $\mathrm{E}[Y|X=2] = 7$, and $\mathrm{E}[Y|X=3] = 9$.

The conditional mean function is not necessarily linear. For example, suppose $\mathrm{E}[Y|X=1] = 5$, $\mathrm{E}[Y|X=2] = 7$, and $\mathrm{E}[Y|X=3] = 12$. Then the conditional mean function is nonlinear in $X$ since it increases by 2 from $X=1$ to $X=2$ but increases by 5 from $X=2$ to $X=3$. Chapter 12 presents more general models that relax the assumption of a conditional mean that is linear in $X$.

### 9.8.5 Conditional Variance

The **conditional variance** of $Y$ given $X$ measures the variation in $Y$ around the conditional mean $\mathrm{E}[Y|X=x]$, where the deviation is squared and is weighted by the conditional probabilities. Then

$$\mathrm{Var}[Y|X=x] = \mathrm{E}[(Y - \mathrm{E}[Y|X=x])^2|X=x],$$

is the probability-weighted average of all possible values of $(Y - \mathrm{E}[Y|X=x])^2$ when $X=x$.

For the example of this appendix, we have already calculated that $\Pr[Y=50|X=1] = 0.5$, $\Pr[Y=30|X=1] = 0.5$, and $\mathrm{E}[Y|X=1] = 40$. It follows that $\mathrm{Var}[Y|X=1] = (50-40)^2 \times .5 + (30-40)^2 \times .5 = 100$. Similarly, $\mathrm{Var}[Y|X=0] = (30-15)^2 \times .25 + (10-15)^2 \times .75 = 75$. Note that the conditional variance here differs according to whether we condition on $X=0$ or on $X=1$.

Assumption 3 (homoskedastic errors) that $\mathrm{Var}[u|X=x]$ does not depend on $x$ implies that $\mathrm{Var}[Y|X=x]$ does not depend on $x$. Alternative assumptions regarding $\mathrm{Var}[u|X=x]$ lead to different ways to estimate the precision of the least squares estimates; see Chapter 10.7.

## 9.9 Appendix 9.B: Population Covariance and Correlation

In this appendix the population covariance and correlation, the population analogs of the sample covariance and correlation introduced in Chapter 8.4, are defined.

### 9.9.1 Population Covariance

The **population covariance** of $Y$ and $X$ measures the joint variation of $X$ and $Y$ around their respective means. Let $\mu_X$ denote $\mathrm{E}[X]$ and $\mu_Y$ denote $\mathrm{E}[Y]$. Then

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathrm{E}[(X-\mu_x)(Y-\mu_Y)] \\ &= \sum_x \sum_y \Pr[Y=y|X=x] \times (X-\mu_x)(Y-\mu_Y).\end{aligned}$$

For the example in Appendix 9.A, $\mu_X = 0.2 \times 1 + 0.8 \times 0 = 0.2$ and $\mu_Y = 0.1 \times 50 + 0.3 \times 30 + 0.6 \times 10 = 20$. Then $\mathrm{Cov}[X,Y] = 0.1 \times (1-0.2) \times (50-20) + 0.1 \times (1-0.2) \times (30-20) + 0.2 \times (0-0.2) \times (30-20) + 0.6 \times (0-0.2) \times (10-20) = 4$.So $\mathrm{Cov}[X,Y] = 4$.

### 9.9.2 Population Correlation

The **population correlation coefficient** of $Y$ and $X$ standardizes the covariance to lie between $-1$ and 1. We have

$$\mathrm{Cor}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X] \times \mathrm{Var}[Y]}}.$$

For the example in Appendix 9.A, $\mathrm{Var}[X] = 0.2 \times (1-0.2)^2 + 0.8 \times (0-0.8)^2 = 0.16$ and $\mathrm{Var}[Y] = 0.1 \times (50-20)^2 + 0.3 \times (30-20)^2 + 0.6 \times (10-20)^2 = 180$. So $\mathrm{Cor}[X,Y] = 4/\sqrt{0.16 \times 180} = 0.745$. In this example $X$ and $Y$ are quite highly positively correlated.

## 9.10 Appendix 9.C Derivation of Properties of OLS

This advanced appendix derives the mean and variance of the OLS slope coefficient given in Chapter 9.3 under Assumptions 1-4.

### 9.10.1 A model without intercept

For simplicity we consider a model with no intercept, so $y = \beta x + u$. Then the OLS estimator $b$ of $\beta$, obtained from regressing $y$ on $x$ without an intercept, is

$$b = \left(\sum_{i=1}^n x_i^2\right)^{-1} \sum_{i=1}^n x_i y_i.$$

Throughout it is assumed that $\sum_i x_i^2 \neq 0$, since otherwise $b$ is not defined.

There is no loss in generality in dropping the intercept if the data $y$ and $x$ are demeaned data. To see this, note that OLS regression of the demeaned variable $(y_i - \bar{y})$ on the demeaned variable $(x_i - \bar{x})$ using the above formula yields $b = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$.

But this is just the usual formula for the OLS estimator. The algebra is much less unwieldy, however, in the demeaned case.

Additionally the regressor $x_i$ is assumed to be fixed. In that case there is no need to take conditional expectations with respect to $x$, simplifying the notation.

Assumptions 1-4 in Chapter 9.3 become: (1) $y_i = \beta x_i + u_i$; (2) $\mathrm{E}[u_i] = 0$; (3) $\mathrm{Var}[u_i] = \sigma_u^2$; and (4) $u_i$ and $u_j$ are statistically independent.

### 9.10.2 A convenient expression for $b$

Given assumption 1 it is always the case that

$$b = \beta + (\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i u_i).$$

Properties of the OLS estimator will therefore be determined primarily by assumptions made about the errors $u_i$.

To obtain this result, note that

$$\begin{aligned} b &= (\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i y_i) \\ &= (\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i(\beta x_i + u_i)) \\ &= (\textstyle\sum_i x_i^2)^{-1}(\sum_i \beta x_i^2 + x_i u_i) \\ &= (\textstyle\sum_i x_i^2)^{-1}(\sum_i \beta x_i^2) + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i) \\ &= \beta + (\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i u_i), \end{aligned}$$

where the second equality uses assumption 1 that $y_i = \beta x_i + u_i$.

### 9.10.3 The Mean of $b$

Given the preceding result,

$$\begin{aligned} \mathrm{E}[b] &= \mathrm{E}\left[\beta + (\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i u_i)\right] \\ &= \mathrm{E}[\beta] + \mathrm{E}\left[(\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i u_i)\right], \end{aligned}$$

since the expected value of a sum is the sum of the expected values.

For the first term, $\mathrm{E}[\beta] = \beta$ as the expected value of a constant is the constant. For the second term, with fixed regressor $x_i$, $(\sum_i x_i^2)^{-1}$ is a constant and the expected value of a constant times a random variable is the constant times the expected value of the random variable. So

$$\mathrm{E}[b] = \beta + (\textstyle\sum_i x_i^2)^{-1} \times \mathrm{E}[\sum_i x_i u_i].$$

It follows that $\mathrm{E}[b] = \beta$ if $\mathrm{E}[\sum_i x_i u_i] = 0$. Now the expected value of a sum is the sum of the expected values:

$$\mathrm{E}[\textstyle\sum_i x_i u_i] = \sum_i \mathrm{E}[x_i u_i].$$

So we need $\mathrm{E}[x_i u_i] = 0$. This is the case under assumption 2 that $\mathrm{E}[u_i] = 0$ since then $\mathrm{E}[x_i u_i] = x_i \mathrm{E}[u_i] = x_i \times 0 = 0$.

Thus $\mathrm{E}[b] = \beta$ under assumptions 1 and 2. (If the regressor $x_i$ is instead stochastic a modification of the above yields that OLS is unbiased if $\mathrm{E}[u_i|x_i] = 0$, i.e. if the error has mean 0 conditional on the regressor.)

### 9.10.4 The variance of $b$

Given $b = \beta + (\sum_i x_i^2)^{-1}(\sum_i x_i u_i)$ and $\mathrm{E}[b] = \beta$ it follows that under assumptions 1-2

$$b - \mathrm{E}[b] = (\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i u_i).$$

Now $\mathrm{Var}[b] = \mathrm{E}[(b - \mathrm{E}[b])^2] = \mathrm{Var}[(b - \mathrm{E}[b])]$. So

$$\mathrm{Var}[b] = \mathrm{Var}\left[(\textstyle\sum_i x_i^2)^{-1}(\sum_i x_i u_i)\right].$$

Now $(\sum_i x_i^2)^{-1}$ is a constant while $(\sum_i x_i u_i)$ is a random variable. In general the variance of a constant times a random variable is the constant squared times the variance of the random variable, so

$$\mathrm{Var}[b] = \left[(\textstyle\sum_i x_i^2)^{-1}\right]^2 \times \mathrm{Var}[\textstyle\sum_i x_i u_i].$$

This general result holds given just assumptions 1-2.

Now add in assumptions 3-4. Given assumption 4 that $u_i$ are statistically independent the terms $x_i u_i$ are statistically independent. But the variance of a sum of independent variables equals the sum of the variances. So

$$\mathrm{Var}[b] = (\textstyle\sum_i x_i^2)^{-2} \times \sum_i \mathrm{Var}[x_i u_i].$$

Now in general $\mathrm{Var}[x_i u_i] = x_i^2 \times \mathrm{Var}[u_i]$, for $x_i$ a fixed regressor. So given assumption 3 that $\mathrm{Var}[u_i] = \sigma_u^2$, we have $\mathrm{Var}[x_i u_i] = x_i^2 \sigma_u^2$. It follows that

$$\mathrm{Var}[b] = (\textstyle\sum_i x_i^2)^{-2} \times \sum_i x_i^2 \sigma_u^2 = \sigma_u^2 (\sum_i x_i^2)^{-1}.$$

Upon converting to data before demeaning, we replace $x_i$ by $x_i - \bar{x}$ and obtain $\mathrm{Var}[b_2] = \sigma_u^2 (\sum_i (x_i - \bar{x})^2)^{-1}$ as presented in Chapter 9.3.

### 9.10.5 Distribution of $b$

The OLS estimator is a weighted linear combination of the errors $u_i$, since we have shown that

$$b = \beta + \textstyle\sum_i w_i u_i \text{ where } w_i = (\sum_i x_i^2)^{-1} x_i.$$

Suppose it is assumed that, in addition to assumptions 1-4, the errors $u_i$ are normally distributed. Then it follows immediately that $b$ is normally distributed because linear combinations of the normal are normal. So $b$ is normally distributed with mean $b$ and variance $\sigma_u^2(\sum_i x_i^2)^{-1}$.

This result cannot be used directly because $\sigma_u^2$ is not known. An advanced result from statistics theory shows that if we replace $\sigma_u^2$ with $s_e^2$, where $s_e$ is the standard error of the residual, then $T = (b - \beta)/s_b$, where $s_b^2 = s_e^2\ (\sum_i x_i^2)^{-1}$, is $t$ distributed under assumptions 1-4 and the assumption of normally distributed errors.

In practice the assumption of normally distributed errors is relaxed. Then the $t$ distribution continues to be used for the $t$-statistic but is now only an approximation, one that improves as the sample size gets larger.

### 9.10.6 Stochastic Regressors

For simplicity it was assumed that the regressors $x$ are fixed. This would be the case for an experiment where $x$ is set by the experimenter. For example, $x$ may be the amount of fertilizer applied to a field or the dosage level of a medicine given to a patient.

More often in social sciences such as economics the data are observational, in which case the regressors are random since the sample observations $(y, x)$ are randomly chosen. Then, as in Chapter 9.3, the assumptions on the errors are conditional on $x$, so $u|x$ must have mean zero. For statistical inference it is additionally assumed, for example, that $u|x$ has constant variance.

For random regressors, the results of this appendix still hold immediately if analysis is viewed as being conditional on the value of the stochastic regressors. If instead the analysis is viewed as unconditional on the regressors, then we ultimately use the same methods as in the conditional case, though there are some changes in the intermediate results. In particular, now $\text{Var}[b] = \sigma_u^2(\sum_i \text{E}[x_i^2])^{-1}$. Also the $t$-statistic is no longer exactly $t$ distributed, even if errors are normally distributed.

# Chapter 10

# © A. Colin Cameron: Statistical Inference for Bivariate Regression

As explained in Chapter 9, the OLS intercept and slope coefficients are random outcomes – different samples lead to different estimates. The sample at hand is viewed as being one from a population, where the population conditional mean is $\beta_1 + \beta_2 x$.

The goal is especially to make inference on the population slope parameter given the estimated slope parameter. For example, is the view that house prices increase with house size consistent with slope estimate of 73.77 obtained for the house price and size dataset HOUSE analyzed in Chapter 8.

The chapter continues directly from the previous chapter. For readers who bypassed the details in Chapter 9, the key result for statistical inference on $\beta_2$ is reproduced in Chapter 10.2.

## 10.1 Example: House Price and Size

Table 10.1 presents key output for statistical inference following regression of house price (in dollars) on size (in square feet) using data in dataset HOUSE.

The slope and intercept coefficients (73.77 and 115017) were discussed in Chapter 8.5. The standard errors of the coefficients, introduced in Chapter 9.3, are $se(b_2) = 11.17$ and

Table 10.1: House price and size: Regression estimates with default standard errors.

| **Variable** | **Coefficient** | **Standard Error** | **t-statistic** | **p-value** | **95% confidence interval** | |
|---|---|---|---|---|---|---|
| *Size* | 73.77 | 11.17 | 6.60 | 0.000 | 50.84 | 96.70 |
| *Intercept* | 115017.30 | 21489.36 | 5.35 | 0.000 | 70924.76 | 159109.8 |

$se(b_2) = 21489$. These are the estimated standard deviations of $b_2$ and $b_1$.

The table description uses the term **default** standard errors because they are computed under assumptions 1-4. Chapter 10.7 introduces alternative standard errors that are obtained when assumptions 3 and 4 are relaxed.

The remainder of the table is explained briefly here, with detailed explanation provided in the rest of the chapter.

The $t$-statistic for the slope coefficient is for test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$. The reason for testing specifically whether $\beta_2 = 0$ is that if this is indeed the case, then there is no (linear) association between house price and house size. Given $\beta_2 = 0$, the $t$-statistic simplifies to $t = b_2/se(b_2) = 73.77/11.17 = 6.60$. The $p$-value is the probability of observing a $T(n-2)$ distributed random variable that is larger than the observed $t$-statistic in absolute value; here $p = \Pr[|T_{27}| > |6.60|] = 0.0000004$ from $t$ tables or statistical package. Since $p < 0.05$ we reject $H_0$ at significance level 0.05.

The 95% confidence interval for $\beta_2$ is (50.84, 96.70). There is a 0.95 probability that this interval includes the unknown $\beta_2$.

The sample size is $n = 29$, a small sample size so the $T(n-2)$ distribution may not necessarily be a good approximation. This approximation may nonetheless be reasonable since the model fits the data well with $R^2 = 0.618$.

## 10.2 t Statistic

Statistical inference on the population slope parameter is based on the $t$-statistic, a transformation of the estimated slope coefficient, rather than the estimated slope coefficient itself. In this section we repeat the results given in Chapter 9.3-9.4.

The sample $(x_1, y_1), ..., (x_n, y_n)$ is assumed to be such that there is some variation across observations in the values of the regressors. OLS regression of $y$ on $x$ then yields estimated intercept $b_1$ and slope $b_2$.

The population model is assumed to satisfy the assumptions detailed in Chapter 9.3: (1) $y = \beta_1 + \beta_2 x + u$; (2) $u$ is an error that has mean zero conditional on $x$; (3) $u$ has constant variance (denoted $\sigma_u^2$) conditional on $x$; and (4) errors for different observations are independent.

A remarkable result, due to the central limit theorem, is that if the sample is of large size then $b_1$ and $b_2$ are normally distributed, regardless of the distribution of the error term (and hence the distribution of $y$). This distribution, however, depends in part on the unknown standard deviation of the error $\sigma_u$, so we replace $\sigma_u$ by its estimate $s_e$, the standard error of the regression. From Chapter 9.4, this leads to a $t$-statistic that is $T(n-2)$ distributed.

The following result considers the slope parameter. A similar result exists for the intercept.

**Remark 73** *From a sample $(x_1, y_1), ..., (x_n, y_n)$ calculate the OLS intercept and slope coefficients $b_1$ and $b_2$, the standard error of the regression, $s_e$, and hence the standard error of $b_2$, $se(b_2) = s_e/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Given assumptions 1-4, the* ***t-statistic***

$$t = \frac{b_2 - \beta_2}{se(b_2)}$$

*is a realization of a random variable that is approximately $T(n-2)$ distributed, where $T(n-2)$ denotes the $t$ distribution with $n-2$ degrees of freedom.*

The $T(n-2)$ approximation is exact if additionally the data are normally distributed or if $n \to \infty$.

The $t$ distribution, discussed in detail in Chapter 6.2, is bell-shaped and coincides with the standard normal distribution as $n \to \infty$. One departure from Chapter 6 is that for bivariate regression the $T(n-2)$ distribution is used rather than the $T(n-1)$ for the mean. The intuition is that two degrees of freedom are lost due to estimation of two parameters ($\beta_1$ and $\beta_2$), whereas only one parameter ($\mu$) was estimated in the univariate case.

To form the $t$-statistic the only sample statistics needed are the fitted slope $b_2$ and its standard error $se(b_2)$. Additionally the $t$-statistic depends on the population slope parameter $\beta_2$, which is unknown. The knowledge that the $t$-statistic is $T(n-2)$ distributed is used to make statistical inference on $\beta_2$, as detailed in subsequent sections of this chapter.

## 10.3 Confidence Intervals

Different samples yield regression lines with different slopes. A confidence interval for $\beta_2$ gives a range of values that, with a specified probability or level of confidence, includes $\beta_2$.

### 10.3.1 Confidence Interval for the Slope Parameter

Construction of a confidence interval for $\beta_2$ mimics that for the population mean given in Chapter 6.3 and the generalization given in Chapter 7.2.

In general, a $t$-statistic that is $t_v$ distributed yields a $100(1-\alpha)$ percent confidence interval for a parameter that is of form *estimate* $\pm t_{v,\alpha/2} \times$ *standard error*. Applying this general result to $\beta_2$ estimated by $b_2$ yields the following.

**Remark 74** *A* $\mathbf{100(1-\alpha)}$ ***percent confidence interval for the slope parameter*** $\beta_2$ *is*

$$b_2 \pm t_{n-2,\alpha/2} \times se(b_2),$$

*where $b_2$ is the OLS slope coefficient, $t_{n-2,\alpha/2}$ is that value such that a $T(n-2)$ distributed random variable exceeds it in absolute value with probability $\alpha$, and $se(b_2)$ is the standard error of the slope coefficient estimate $b_2$.*

The confidence interval is appropriate if the sample size is sufficiently large, or if the sample is small and the error terms $u$ are normally distributed.

The confidence interval is centered around $b_2$, the estimate of $\beta_2$. It is narrower the more precisely $b_2$ estimates $\beta_2$, as $se(b_2)$ is then smaller. Thus the confidence interval should narrow as the sample size gets larger.

A $100(1-\alpha)\%$ confidence interval for $\beta_2$ is one that with probability $1-\alpha$ includes the true (and unknown) population slope coefficient $\beta_2$. Higher levels of confidence arise if the confidence interval is wider; in the formula above $T(n-2)$ becomes larger as $\alpha$ becomes smaller.

### 10.3.2 What Level of Confidence?

There is no best choice for the confidence level. The trade-offs between high level of confidence and narrow confidence interval are the same as those discussed at length for confidence intervals for the population mean; see Chapter 6.3.

The standard choice is to form a 95% confidence interval, so $\alpha = 0.05$. This is the default level used in output from a regression package. Most regression packages allow the user the option to select a different level of confidence. The other most common choices are 90% and 99% confidence intervals.

From Table 6.2, $t_{n-2,.025} \simeq 2$ for most values of $n > 30$. Then an approximate 95% confidence interval is $b_2 \pm 2 \times se(b_2)$.

**Remark 75** *It is most common, though arbitrary, to use a 95% confidence interval.* ***An approximate 95% confidence interval*** *for $\beta_2$ is the* ***two-standard error interval****, the estimate $b_2$ plus or minus two times the standard error $se(b_2)$.*

In any published work or in assignments, however, use the more precise interval $b_2 \pm t_{n-2,.025} \times se(b_2)$.

### 10.3.3 Example: House Price and Size

As an example, consider the house price and size data. Regression on a sample of size $n = 29$ yields coefficient $b_2 = 73.77$ and standard error $se(b_2) = 11.17$. And $t_{n-2,\alpha/2} = t_{27,.025} = 2.052$ from tables or a statistical package. So the 95% confidence interval for $\beta_2$ is

$$b_2 \pm t_{n-2,\alpha/2} \times se(b_2) = 73.77 \pm 2.052 \times 11.17 = 73.77 \pm 22.93 = (50.84, 96.70).$$

This is the confidence interval given in Table 10.1.

### 10.3.4 Example: Artificial Data

As a second example, consider artificial data on a sample of size $n = 5$, where $y$ is the number of cars in a household and $x$ is the number of people in the household. The sample values of $(x, y)$ are $(1, 1)$, $(2, 2)$, $(3, 2)$, $(4, 2)$ and $(5, 3)$.

From Appendix 8.A, $b_1 = 0.8$ and $b_2 = 0.4$, and from Chapter 9.3 the standard error of $b_2$ is $se(b_2) = 0.115$. For a 95% confidence interval from a sample of size 5 we use critical value $t_{3,.025} = 3.182$. A 95% confidence interval for the slope coefficient $\beta_2$ is

$$b_2 \pm t_{3,.025} \times se(b_2) = 0.4 \pm 3.182 \times 0.115 = 0.4 \pm 0.367 = (0.033, 0.767).$$

Given the small sample size this inference requires the additional assumption that the error term $u$ is normally distributed.

### 10.3.5 Interpretation of Confidence Intervals

Interpretation of confidence intervals is conceptually difficult. For this reason the discussion in Chapter 6.3 for the population mean is repeated, with adaptation to the regression case.

Consider the house price example with 95% confidence interval $(50.84, 96.70)$ for the population slope $\beta_2$. The correct interpretation of this confidence interval is that **the calculated 95 percent confidence interval** from this sample **includes $\beta_2$ with probability 0.95.**

To understand this interpretation it is necessary to imagine that there are many separate samples of the population, each of size $n = 29$ in the house price and size example. For concreteness suppose the true population slope coefficient is $\beta_2 = 70.0$. The first sample may yield a confidence interval of, say, $(40.7, 75.8)$. A second sample will have a different slope estimate and associated standard error and hence a different confidence interval of, say, $(56.7, 95.3)$. And so on. Some samples will have unusually large or small estimated values of the slope coefficient. For example, one sample might have an unusually high value $b_2 = 98.7$ with standard error $se(b_2) = 11.3$, say, so that $t_{27,.025} \times se(b_2) = 23.2$, leading to a 95% confidence interval of $(75.5, 121.9)$. This confidence interval, unlike the preceding ones, does not include the population slope parameter which is 70.0. On average we expect 95 out of 100 of these 95% confidence intervals to include the unknown true value of $\beta_2$.

Thus the confidence interval will sometimes fail to include the population value $\beta_2$, due to the randomness inherent in sampling. A 95% confidence interval has the property that if we were able to obtain many separate random samples, 95 percent of the resulting confidence intervals will include $\beta_2$, and 5 percent will not.

In fact we have only one sample, and we say that the calculated 95% confidence interval from this sample includes the true population slope coefficient $\beta_2$ with probability 0.95. Note that it is wrong to instead interpret this confidence interval as meaning that with probability

0.95 the population parameter $\beta_2$ lies inside $(50.84, 96.70)$ and with probability 0.05 it lies outside this range.

**Remark 76** *A calculated 95 percent confidence interval for the population slope parameter $\beta_2$ is an interval that includes $\beta_2$ with probability 0.95.*

## 10.4 Tests of Statistical Significance

Hypothesis tests are used to decide whether a specified value of $\beta_2$ is plausible, given the sample estimate $b_2$. Before providing a general treatment of hypothesis testing for the slope coefficient, building on Chapter 6.4 for tests on the population mean, we briefly consider the most common test, that of whether or not $\beta_2 = 0$.

### 10.4.1 Tests of Statistical Significance

For linear regression, the most common hypothesis test is a test of whether or not the population slope parameter equals zero.

This test is called a test of the statistical significance of a regressor. It answers the question of whether or not there is an association between $x$ and $y$. If there is no association, then clearly $\beta_2 = 0$ and the population regression model $y = \beta_1 + \beta_2 x + u$ reduces to $y = \beta_1 + u$, so that $y$ bounces around a mean value of $\beta_1$.

**Remark 77** *A **two-sided test** or **two-tailed test of the statistical significance of a regressor** for the population mean $\mu$ is a test of the null hypothesis*

$$H_0 : \beta_2 = 0,$$

*against the alternative hypothesis*

$$H_a : \beta_2 \neq 0.$$

The test of whether $\beta_2 = 0$ is based on how far the sample estimate $b_2$ is from zero. If $b_2$ was very precisely estimated, then even small departures of $b_2$ from 0 lead to the conclusion that $\beta_2 \neq 0$, while if $b_2$ is very imprecisely estimated then large departures of $b_2$ are needed to conclude $\beta_2 \neq 0$. The precision or imprecision of $b_2$ is controlled for by using the $t$-statistic, $T = (b_2 - \beta_2)/se(b_2)$, evaluated here at the null hypothesis value of $\beta_2 = 0$. Thus we use the $t$-statistic, also called the $t$-ratio $t = \frac{b_2}{se(b_2)}$. $H_0$ is rejected if the observed $t$ is far from zero.

**Remark 78** *The $t$-**test statistic** for test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ is the* **t-*ratio***

$$t = b_2/se(b_2).$$

*Under $H_0$, and given assumptions 1-4, $t$ is the realization of a random variable that is approximately $T(n-2)$ distributed.*

From Chapter 6.4 the $p$-**value** is the probability of observing a $t$-statistic at least as large in absolute value as that obtained in the current sample. Here $p = \Pr[|T_{n-2}| \geq |t|]$. This $p$-value is compared to the **significance level** $\alpha$ that is a preset cutoff value for the $p$-value.

Regression packages automatically print out both the $t$-statistic for test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ and the associated $p$-value. Since the $p$-value is the probability that the observed $t$ occurs by chance, if in fact $\beta_2 = 0$, small $p$-values are evidence against the null hypothesis and in favor of the alternative hypothesis.

**Remark 79** *Let $p = \Pr[|T_{n-2}| \geq |t|]$ be the p-value for test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$. If $p < \alpha$ then $H_0$ is rejected at significance level $\alpha$, and the regressor is said to be* ***statistically significant at significance level*** *$\alpha$. If instead $p > \alpha$ then $H_0$ is not rejected and the regressor is said to be* ***statistically insignificant at level*** *$\alpha$.*

For example if $p = 0.03$ then the regressor is statistically significant at significance level 0.05, since $0.03 < 0.05$. But the regressor is statistically insignificant at significance level 0.01, since $0.03 \not< 0.01$. The most common choice of $\alpha$ is 0.05, followed by 0.01 and 0.10.

The **critical region** approach, see Chapter 6.4, provides an alternative approach to using $p$-values that leads to the same conclusion. This defines a range of values for the $t$-statistic that leads to rejection at the specified significance level $\alpha$. For regression with a single regressor and test at significance level $\alpha$, the **critical value** is $c = t_{n-2,\alpha/2}$ and we reject $H_0$ if $|t| > t_{n-2,\alpha/2}$.

As demonstrated in Table 6.2, $t_{n-2,.025} \simeq 2$ for $n > 30$. So, approximately, a test will conclude that the regressor is statistically significant at level 0.05, the most common choice of $\alpha$, if $|t| > 2$. This leads to the following:

**Remark 80** *A regressor is (approximately) statistically significant at level* 0.05 *if $|b_2|$ exceeds two times the standard error of $b_2$.*

### 10.4.2 Example: House Prize and Size

For the house price and size data, Table 10.1 gives $t = 6.60$ with $p = 0.000$ to three decimal places. (More precisely, $p = \Pr[|T_{27}| > 6.60] = 0.000000441$). So, at the usual significance level of $\alpha = 0.05$, $H_0$ is rejected and there is a statistically significant relationship between *Price* and *Size*. Indeed, here *Size* is a highly statistically significant regressor since $p$ is so low.

### 10.4.3 Economic Significance versus Statistical Significance

For a regressor to be useful in explaining the dependent variable it should be both statistically significant and economically significant.

**Remark 81** *A regressor is of* ***economic significance*** *if its coefficient is of large enough value for it to matter in practice. Economic significance depends directly on $b_2$, while statistical significance depends directly on $t$ which is the ratio $b_2/se(b_2)$.*

Determining economic significance requires subjective interpretation that varies with the application. Suppose house price (in dollars) is regressed on number of bedrooms. A slope coefficient of $b_2 = 100$ has little economic significance, as then another bedroom is associated with only a \$100 increase in price, while a slope coefficient of 100,000 has great economic significance. Even this interpretation might change if the data pertain to Afghanistan, rather than the U.S., as Afghanistan has per capita GDP (on a purchasing power parity basis) of only US \$1,000.

The fact that a regressor is statistically significant does not necessarily mean that it is of economic significance. With enough data any regressor that is even very mildly associated with the dependent variable is likely to be statistically significant, since the standard error of $b_2$ falls as the sample size increases.

Going the other way, the fact that a regressor is statistically insignificant does not necessarily mean that it is of no economic significance. It may just be that imprecise estimation, often due to a small sample size or little variation in the regressor, leads to a standard error that is too large for the variable to be statistically significant.

It is easy to be overly reliant on tests of statistical significance. For a two-sided test at level 0.05 statistical significance is determined, approximately, according to whether or not $|b_2| > 2 \times se(b_2)$. Economic significance is instead determined by whether $|b_2|$ is "small" or "large", a subjective judgement that is context specific.

### 10.4.4 Tests Based on the Correlation Coefficient

The **population correlation coefficient** $\rho_{xy}$, defined in Appendix 9.B, measures the correlation between $X$ and $Y$ in the population. An alternative way to test whether there is a relationship between $y$ and $x$ is to test whether $\rho_{xy} = 0$, based on the sample coefficient coefficient $r_{xy}$ defined in Chapter 8.4.

Some algebra shows that if $se(b_2)$ is computed under assumptions 1-4, then $t = b_2/se(b_2)$ can be re-expressed in terms of $r_{xy}$ as

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}}}.$$

So the two tests are related, and the $t$-statistic is larger in absolute value the larger in absolute value is the sample correlation between $y$ and $x$.

In some disciplines association between $y$ and $x$ is tested using the sample correlation coefficient rather than the regression slope coefficient. In large samples tests of statistical

significance based on $b_2/se(b_2)$ lead to the same conclusions as tests of $H_0 : \rho_{xy} = 0$ against $H_a : \rho_{xy} = 0$ based on $r_{xy}$. In smaller samples the conclusions can differ somewhat.

This link between the $t$-statistic and the sample correlation coefficient provides two insights. First, the correlation coefficient measures only linear association, since the regression model is linear. Second, statistically significant correlation between two variables means merely that the association is statistically significant; it need not be economically important.

Economists test association between two variables using regression based tests of $\beta_2 = 0$ rather than tests of zero correlation. One reason for doing so is that tests of correlation require that assumptions 1-4 hold, whereas tests based on $t = b_2/se(b_2)$ can relax assumptions 3-4 leading to different ways to compute $se(b_2)$.

## 10.5 Two-Sided Hypothesis Tests

The preceding section presented a simple way to test whether or not $\beta_2 = 0$. More generally we may wish to test whether $\beta_2$ equals other values. Then regression packages no longer automatically print out the relevant test statistic and its $p$-value. Instead some additional manual computation is needed, unless the regression package provides a separate command to perform such tests following regression.

Extension of hypothesis testing from $\mu$ to $\beta_2$ is almost immediate. To reinforce the methods, however, a fairly complete treatment is given here, despite obvious duplication of Chapter 6.4.

Let $\beta_2^*$ denote a hypothesized value for $\beta_2$, such as $\beta_2^* = 0$ or $\beta_2^* = 90$.

**Remark 82** *A **two-sided test** or **two-tailed test** for the population slope $\beta_2$ is a test of the null hypothesis*

$$H_0 : \beta_2 = \beta_2^*,$$

*where $\beta_2^*$ is a specified value for $\beta_2$, against the alternative hypothesis*

$$H_a : \beta_2 \neq \beta_2^*.$$

The term two-sided is used as the alternative hypothesis covers both the case that $\beta_2 < \beta_2^*$ and the case that $\beta_2 > \beta_2^*$.

The following discussion builds on Chapter 6.4 for tests on the population mean.

### 10.5.1 Example: House Price and Size

As an example, consider the claim that the population mean sale price of a house increases by \$90 per additional square foot in size, so $\beta_2^* = 90$. This would be a natural value to test if

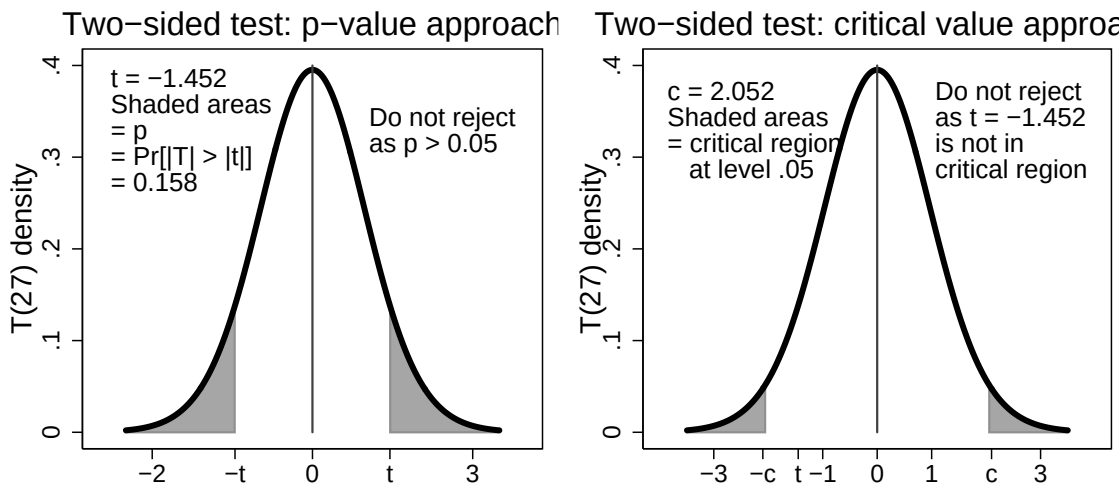


Figure 10.1: Hypothesis tests: $p$-value approach and critical region approach.

enlarging an existing house costs \$90 per square foot. From Table 10.1, the regression slope coefficient $b_2 = 73.77$ for a sample of $n = 29$ houses. Is this far enough away from 90 to reject the hypothesis that $\beta_2 = 90$? Or could this difference from 90 be merely an artifact of sampling variability?

To test $H_0 : \beta_2 = 90$ we evaluate the $t$-statistic when $\beta_2 = 90$, yielding

$$t = (b_2 - 90)/se(b_2) = (73.77 - 90)/11.17 = -1.452.$$

If indeed $\beta_2 = 90$ this is a draw from the $T(27)$ distribution, since $n - 2 = 27$.

How likely are we to obtain a draw from the $T(27)$ distribution equal to or greater than $-1.45$ in absolute value? This probability is the $p$-value

$$p = \Pr[|T_{27}| \geq |-1.452|] = 0.158.$$

The first panel of Figure 10.1 illustrates the $p$-value. There is a somewhat low probability (0.158) of observing a $t$-statistic value as large or larger than $-1.45$ in absolute value than if in fact $\beta_2 = 90$, suggesting that the null hypothesis that $\beta_2 = 90$ may be wrong. If we use the common convention to reject $H_0$ if this probability is less than 0.05, however, then we do not reject $H_0 : \beta_2 = 90$ since $p = 0.158 > 0.05$.

An alternative equivalent method bases rejection or non-rejection directly on the value of the $t$-statistic. Suppose we require that $p < 0.05$ in order to reject $H_0$. This would require that $|t| > 2.052$ since $\Pr[|T_{27}| \geq 2.052] = 0.05$. The second panel of Figure 10.1 illustrates the rejection region, which does not include $t = -1.45$. Again $H_0 : \beta_2 = 0$ is not rejected at significance level 0.05.

### 10.5.2 Test Statistic

The test of whether $\beta_2 = \beta_2^*$ is based on the closeness of the sample estimate $b_2$ to the population slope $\beta_2^*$. But $b_2$ may differ from $\beta_2^*$ for two reasons that lead to polar opposite conclusions. It may be that $\beta_2 \neq \beta_2^*$, so the null hypothesis should be rejected. Or it may be that $\beta_2 = \beta_2^*$ but $b_2 \neq \beta_2^*$ due to sampling variability in $b_2$. Then the null hypothesis should not be rejected.

The obvious decision rule is to reject $H_0 : \beta_2 = \beta_2^*$ if the OLS slope estimate $b_2$ is far from the hypothesized $\beta_2^*$. It is convenient to transform from this difference $(b_2 - \beta_2^*)$ to $(b_2 - \beta_2^*)/se(b_2)$, since the latter is known to have a $t$ distribution.

**Remark 83** *The t-test statistic for test of $H_0 : \beta_2 = \beta_2^*$ against $H_a : \beta_2 = \beta_2^*$ is $t = (b_2 - \beta_2^*)/se(b_2)$. Under $H_0$ and assumptions 1-4, $t$ is the realization of a random variable that is approximately $T(n-2)$ distributed.*

### 10.5.3 Rejection using $p$-values

The $p$-**value** is the probability of observing a t-test statistic at least as large in absolute value as that obtained in the current sample, under the assumption that $H_0$ is correct so that $\beta_2 = \beta_2^*$.

**Remark 84** *The $p$-**value** is the probability of observing a t-test statistic at least as large in absolute value as that obtained in the current sample. For a two-sided test of $H_0 : \beta_2 = \beta_2^*$ against $H_a : \beta_2 \neq \beta_2^*$ the p-value is $p = \Pr[|T_{n-2}| \geq |t|]$. $H_0$ is rejected at significance level $\alpha$ if $p < \alpha$, and is not rejected otherwise.*

The standard testing approach is to take the conservative stance of rejecting $H_0$ only if the $p$-value is low. The most common choice of significance level is $\alpha = .05$. The next most common choices are .10 and .01.

### 10.5.4 Rejection using Critical Regions

The $p$-value approach requires access to a computer, in order to precisely compute the $p$-value for any given value of $t$. An alternative approach requires only tables of the $t$ distribution for selected values of $\alpha$, and was the method used before the advent of ready access to computers.

This alternative approach defines a **critical region** or **rejection region** that gives the range of values of the $t$-statistic that lead to rejection of $H_0$ at the specified significance level $\alpha$. $H_0$ is rejected if the computed value of $t$ falls in this range.

Table 10.2: Hypothesis Tests: Type I and Type II Errors.

| **Decision** | **Truth** | |
|---|---|---|
| | $H_0$ really true: $\beta_2 = \beta_2^*$ | $H_0$ really false: $\beta_2 \neq \beta_2^*$ |
| Do not reject $H_0$ : Find $\beta_2 = \beta_2^*$ | Correct decision | Type II error (false positive) |
| Reject $H_0$ : Find $\beta_2 \neq \beta_2^*$ | Type I error (false negative) | Correct decision |

**Remark 85** *For a two-sided test of $H_0 : \beta_2 = \beta_2^*$ against $H_a : \beta_2 = \beta_2^*$, and for specified $\alpha$, the* ***critical value*** *$c$ is such that $c = t_{n-2,\alpha/2}$; equivalently $\Pr[|T_{n-2}| \geq c] = \alpha$. $H_0$ is rejected at significance level $\alpha$ if $|t| > c$, and is not rejected otherwise.*

Many textbooks provide tables to compute $c$ for selected values of $\alpha$, as already noted. Alternatively one can use a function provided in a statistical package.

### 10.5.5 Relationship to Confidence Interval

Two-sided tests can be implemented using confidence intervals.

**Remark 86** *If the null hypothesis value $\beta_2^*$ falls inside the $100(1-\alpha)$ percent confidence interval then do not reject $H_0$ at significance level $\alpha$. Otherwise reject $H_0$ at significance level $\alpha$.*

For the house price data, the 95% confidence interval for $\beta_2$ is $(50.84, 96.70)$. Since this interval includes 90 we do not reject $H_0 : \beta_2 = 90$ at significance level 0.05.

### 10.5.6 Test Size and Power

The usual choice of significance level is $\alpha = 0.05$. This is a convention – there is no clear best choice of $\alpha$. As the significance level $\alpha$ gets smaller, stronger evidence is required to reject $H_0$, making it less likely that $H_0$ is rejected erroneously. This is good news, but it comes at the expense of making it more likely to fail to reject $H_0$ when it should be rejected. There is a trade-off between the two types of error.

From Chapter 6.5 the following terminology is used.

**Remark 87** *A* ***type I error*** *(or false negative) occurs if $H_0$ is rejected when $H_0$ is true. A* ***type II error*** *(or false positive) occurs if $H_0$ is not rejected when $H_0$ is false.*

The possibilities are summarized in Table 10.2 where, for example, a false negative is rejecting $H_0 : \beta_2 = \beta_2^*$ when in fact $\beta_2 \neq \beta_2^*$.

Ideally both type 1 and type II errors occur with low probability. But unfortunately decreasing one means increasing the other. For example, we can reduce the probability of a type I error by being very conservative and rarely rejecting, but then the type II error probability will rise as we then are unlikely to reject even if we should.

From Chapter 6.5 the following terms are used.

**Remark 88** ***Test size*** *is the* ***probability of a type I error. Test power*** *is* ***one minus the probability of a type II error.***

Test power is ideally high as then the probability of a type II is small. The best tests, those with the most power for a given test size, are called **most powerful tests**. Intuitively, the more precisely $\beta_2^*$ the less likely it is to make a type II error for a specified probability of a type I error. This is indeed the case, and for this reason it is optimal to base tests on the unbiased estimator (or consistent estimator) of $\beta_2$with minimum variance.

For tests of hypotheses on the population slope parameter, the $t$ test is the most powerful test if assumptions 1-4 hold and additionally the data are normally distributed. More generally the power is higher the more accurately $b_2$ is estimated.

The standard choices of $\alpha$ are 0.05 or 0.01 or 0.10, reflecting a reluctance to make the error of unwarranted rejections of the null hypothesis. But this comes with the potential cost of a high probability of failing to reject the null hypothesis when we should have rejected. Furthermore there is an asymmetry. As sample size gets larger, leading to more precise estimation, it is possible to simultaneously improve both test size and power. Instead, standard practice is to keep test size at the usual 0.05 level, and only test power improves.

In general the choice of significance level $\alpha$ is not clear. The attraction of reporting $p$-values, rather than critical values, is that the reader can then easily test using his or her own preferred value of $\alpha$. For this reason the $p$-value method is preferred. The critical value method was developed for an earlier time when reliance on printed tables made it difficult to accurately calculate $p$-values.

### 10.5.7 Summary: House Price Summary

A summary of the house price hypothesis test example is the following.

| | |
|---|---|
| Hypotheses | $H_0 : \beta_2 = 90$, $H_a : \beta_2 \neq 90$ |
| Significance level | $\alpha = 0.05$ |
| Data | $b_2 = 73.77$, $se(b_2) = 11.17$, $n = 29$ |
| Test statistic | $t = (73.77 - 90)/11.17 = -1.452$ |
| (1) p-value approach | $p = \Pr[\lvert T_{27}\rvert \geq \lvert -1.452\rvert] = 0.158$ |
| (2) Critical value approach | $c = t_{27,.025} = 2.052$ |
| Conclusion | Do not reject $H_0$ at level .05 as (1) $p > .05$ or (2) $\lvert t\rvert < c$. |

The p-value and critical value approaches are alternative methods that, for given significance level $\alpha$, always lead to the same conclusion.

## 10.6 One-Sided Hypothesis Tests

A one-sided hypothesis test is used to test a one-sided claim such as that the population slope parameter in the house sale example exceeds 90. A one-sided test uses the same test statistic as a two-sided test but the $p$-value and critical value are calculated differently, leading to a different rejection region. This section adapts to regression the lengthier treatment in Chapter 7.1.

### 10.6.1 One-Sided Hypotheses

For a **one-sided hypothesis test** care needs to be used in specifying the null and alternative hypotheses as the conclusion can differ according to which hypothesis is set up as the null and which is the alternative.

**Remark 89** *For one-sided tests the statement being tested is specified to be the alternative hypothesis. And if a new theory is put forward to supplant an old, the new theory is specified to be the alternative hypothesis.*

For example, if we wish to test the claim that the population slope parameter exceeds 90, we should test $H_0 : \beta_2 \leq 90$ against $H_a : \beta_2 > 90$. By contrast, to test the claim that the population slope parameter is less than 90, we should test $H_0 : \beta_2 \geq 90$ against $H_a : \beta_2 < 90$.

**Remark 90** *An **upper one-tailed alternative test** is a test of $H_0 : \beta_2 \leq \beta_2^*$, where $\beta_2^*$ is a specified value for $\beta_2$, against $H_a : \beta_2 > \beta_2^*$. A **lower one-tailed alternative test** is a test of $H_0 : \beta_2 \geq \beta_2^*$ against $H_a : \beta_2 < \beta_2^*$.*

Some texts use a different notation for the null hypothesis in one-sided tests, defining it to be $H_0 : \beta_2 = \beta_2^*$ rather than $H_0 : \beta_2 \leq \beta_2^*$ (or $H_0 : \beta_2 \leq \beta_2^*$). This alternative notation makes no difference to the subsequent analysis as the alternative hypothesis remains the same.

### 10.6.2 P-values and Critical Regions

Inference for both types of one-sided test is based on the same calculated test statistic

$$t = (b_2 - \beta_2^*)/se(b_2),$$

as used for two-sided hypothesis tests. As usual this statistic is viewed as being the realization of a $T(n-2)$ distributed random variable. What differs in the one-sided case is calculation of the $p$-values and critical values.

For an **upper** one-tailed alternative test large positive values of $t$ are grounds for rejection of $H_0$, since then $b_2$ (the estimate of $\beta_2$) is much larger than $\beta_2^*$. For a **lower** one-tailed alternative test large negative values of $t$ lead to rejection of $H_0$, since then $b_2$ is much smaller than $\beta_2^*$.

**Remark 91** *Let $t$ be the usual $t$-test statistic. For an upper one-tailed alternative test the $p$-value is $p = \Pr[T_{n-2} \geq t]$, the critical value at significance level $\alpha$ is $c = t_{n-1,\alpha}$, and we reject $H_0$ if $p < \alpha$ or, equivalently, if $t > c$. For a lower one-tailed alternative test the $p$-value is $p = \Pr[T_{n-2} \leq t]$, the critical value at significance level $\alpha$ is $c = -t_{n-2,\alpha}$, and we reject $H_0$ if $p < \alpha$ or, equivalently, if $t < c$.*

Note that for a one-sided test $p = \Pr[T_{n-2} \leq t]$ which is one-half of $p = \Pr[|T_{n-2}| \leq t]$ for a two-sided test. We are therefore more likely to reject, at given significance level $\alpha$, if we use a one-sided test.

**Remark 92** *Provided $t > 0$ for an upper one-tailed alternative test, or $t < 0$ for an lower one-tailed alternative test, the $p$-value for a two-sided test can be calculated as one-half the $p$-value for a two-sided test.*

### 10.6.3 Example: House Price and Size

Suppose the claim is made that house prices increase by less than \$90 per additional square foot. Then the appropriate test is a lower one-sided test of $H_0 : \beta_2 \geq 90$ against $H_a : \beta_2 < 90$, as we make the claim the alternative.

The $t$-statistic is the same as in the two-sided case, so $t = -1.45$. Now $p = \Pr[T_{27} \leq -1.45] = 0.079$. $H_0$ is not rejected at significance level 0.05, since $p > 0.05$. Equivalently, the critical value $c = t_{27,.05} = -1.70$. $H_0$ is again not rejected since $t = -1.45 > -1.70$. We therefore do not reject $H_0$ at significance level $\alpha = .05$. The data are not inconsistent with the claim that house prices increase by less than \$90 per additional square foot.

Note that compared to the similar two-sided test at level $\alpha = 0.05$ the $p$-value is lower in the one-sided test (.079 versus .158), making rejection of $H_0$ more likely. This is because it is easier to determine that $\beta_2$ lies in the narrower alternative hypothesis region $\beta_2 < 90$ than in the broader region $\beta_2 \neq 90$.

### 10.6.4 One-sided Tests of Statistical Significance

When testing statistical significance of a regressor there may be a prior belief that the slope coefficient is positive or that it is negative. In that case an appropriate one-sided test should

Table 10.3: Hypothesis tests: Summary for tests on the slope parameter.

| | **Two-sided Test** | **One-sided Upper alternative** | **One-sided Lower alternative** |
|---|---|---|---|
| Null hypothesis | $H_0 : \beta_2 = \beta_2^*$ | $H_0 : \beta_2 \leq \beta_2^*$ | $H_0 : \beta_2 \geq \beta_2^*$ |
| Alternative Hypothesis | $H_a : \beta_2 \neq \beta_2^*$ | $H_a : \beta_2 > \beta_2^*$ | $H_a : \beta_2 < \beta_2^*$ |
| t-statistic | $t = (b_2 - \beta_2^*)/se(b_2)$ | $t = (b_2 - \beta_2^*)/se(b_2)$ | $t = (b_2 - \beta_2^*)/se(b_2)$ |
| p-value | $p = \Pr[\lvert T_{n-2}\rvert \geq t]$ | $p = \Pr[T_{n-2} \geq t]$ | $p = \Pr[T_{n-2} \leq t]$ |
| Rejection rule | $p < \alpha$ | $p < \alpha$ | $p < \alpha$ |
| Critical value | $c = t_{n-2,\alpha/2}$ | $c = t_{n-2,\alpha}$ | $c =$-$t_{n-2,\alpha}$ |
| Rejection region | $\lvert t\rvert > c$ | $t > c$ | $t < c$ |

be used. For example if we believe the coefficient is positive we test $H_0 : \beta_2 \leq 0$ against $H_a : \beta_2 > 0$. This requires halving the printed $p$-value for a two-sided test. In addition one needs to verify that $b_2 > 0$ for tests against $H_a : \beta_2 > 0$ or that $b_2 < 0$ for tests against $H_a : \beta_2 < 0$.

Unfortunately there can be ambiguity in the statement that "the regressor is statistically significant at significance level 0.05" as it will not always be clear whether a one-sided or two-sided test was performed. Usually a two-sided test is reported. If a one-sided test is used it is clearer is to say that a regressor is **positively statistically significant** if the prior belief that $\beta_2$ is positive is supported by an upper one-tailed alternative test, and a regressor is **negatively statistically significant** if the prior belief that $\beta_2$ is negative is supported by a lower one-tailed alternative test.

### 10.6.5 Summary for One-Sided and Two-Sided Tests

Table 10.3 summarizes one-sided and two-sided hypothesis tests on the population slope coefficient. The $p$-value approach is the simplest testing approach, though some care is needed for one-sided tests.

The standard error $se(b_2)$ may be computed in several different ways. To date only default standard errors. Next we present alternative standard errors that need to be used if assumptions 3-4 are relaxed.

## 10.7 Robust Standard Errors

Regression analysis is often used in settings where assumptions 3-4, that model errors are homoskedastic and statistically independent, are inappropriate. Throughout we assume that OLS is consistent; if assumption 4 is relaxed this requires some strengthening of assumption 2 as discussed below.

There are two ways to proceed. We can continue to use OLS estimation but base inference on alternative formulas for the standard errors. Or we can use estimation methods other than OLS that potentially lead to more precise estimation of slope coefficients. The latter approach, deferred to Chapter 14.7, is used less frequently and we consider the first approach here.

Standard errors for the least squares coefficients obtained under assumptions weaker than assumptions 3-4 are called **robust standard errors**, while the formula for the standard error of $b_2$ under assumptions 3-4 is called the **default standard error**.

Given a robust standard error of $b_2$, denoted $se_{rob}(b_2)$, confidence intervals and hypothesis tests are based on using

$$T = \frac{b_2 - \beta_2}{se_{rob}(b_2)}.$$

The distribution of $T$ is approximated by the $T(n-2)$ distribution, except in the case of cluster-robust standard errors in which case one uses the $T(G-1)$ distribution, where $G$ is the number of clusters.

There are several different formulas for robust standard errors that vary according to the type of data being analyzed. Most statistical packages, though not Excel, will compute these alternative standard errors as an option to the regression command.

### 10.7.1 Heteroskedastic Robust

For **cross-section data** it is often reasonable to assume that model errors are independent across observations (assumption 4). But model errors may have variance whose size varies across observations. For example, the error variance may increase as the value of the regressor $x$ increases. Then the errors are said to be **heteroskedastic** and assumption 3 is replaced with the assumption that $\text{Var}[u_i|x_i] = \sigma_i^2$, $i = 1, ..., n$, where $\sigma_i^2$ varies with $i$.

Provided assumptions 1-2 and 4 still hold, the OLS estimates remain unbiased and consistent. The one change is that the usual formula for the standard error of $b_2$, called the **default standard error**, is replaced by the **heteroskedasticity-robust standard error**

$$se_{het}(b_2) = \sqrt{\frac{n}{n-2}} \times \frac{\sqrt{\sum_{i=1}^n e_i(x_i - \bar{x})^2}}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where $e_i$ is the OLS residual. Heteroskedasticity-robust standard errors are often within 20% of default standard errors, sometimes larger and sometimes smaller.

Most statistical packages provide this as an option for OLS regression. These standard errors are often called **White robust standard errors**, after the econometrician Halbert White who introduced them in a seminal 1980 research article.

For the house price example the default standard errors for the slope and intercept coefficients are, respectively, 11.17 and 21489. The heteroskedastic-robust standard errors

are, respectively, 11.33 and 20298, so the difference is very small in this example. Then a 95% confidence interval for $\beta_2$ becomes $73.77 \pm t_{.025;27} \times 11.33 = (50.53, 97.01)$. And for test of $H_0 : \beta_2 = 0$, $t = 73.77/11.17 = 6.51$, rather than 6.60 using default standard errors. Again $p = \Pr[|T_{27}| > 6.51] = 0.000$ and we reject the null hypothesis at level 0.05.

**Remark 93** *Heteroskedastic-robust standard errors are used if assumptions 1-2 and 4 hold, but model errors are heteroskedastic rather than homoskedastic. These are routinely used for cross-section data and some special cases of time series data.*

### 10.7.2 HAC Robust

**Time series data** are often correlated across observations, with correlation that fades as observations become further apart in time. It is likely that the model errors are also autocorrelated, so assumption 4 no longer holds.

To ensure that OLS remains consistent assumption 2 is strengthened to the assumption that $\mathrm{E}[u_t|x_t, x_{t-1}, ..., x_1] = 0$ for all $t$. Now the error must be uncorrelated with not just the current value of the regressor but also with past regressors. As explained in Chapter 20 this means that HAC standard errors cannot be used if the lagged dependent variable ($y_{t-1}$) is a regressor.

If model errors are autocorrelated only up to $m$ periods apart, then one uses **heteroskedasticity- and autocorrelation consistent (HAC) robust standard errors**.

Most statistical packages provide HAC standard errors as an option for OLS regression. The key ingredient is what value of $m$ to use. Some specialized software determines this automatically, but usually the user must specify a value for $m$. For moderately correlated errors with $\mathrm{Cor}[u_t, u_{t-1}] < 0.5$, rules of thumb include letting $m$ equal the first integer larger than $0.75T^{1/3}$ or larger than $T^{1/3}$. If the error correlation is higher then larger $m$, possibly much larger $m$, is needed. A useful guide is to let $m + 1$ be the lag at which the autocorrelations (see Chapter 8.4) of the OLS residual become small, say less than 0.1 or 0.2 in value.

As an example, return to the price inflation example of Chapter 6.6. For data in dataset INFLATION the sample mean inflation rate was 3.359 with standard error of the mean equal to $2.378/\sqrt{259} = .148$. This sample standard deviation was based on the assumption that observations were independent, when in fact inflation this quarter is correlated with inflation in preceding quarters. Since the sample mean can be obtained by regression of $y$ on just an intercept, we regress inflation on just an intercept. As expected, the estimated intercept is 3.359, equal to the sample mean. And the default standard error is .148, equal to the standard sample standard deviation.

To compute the HAC standard error we need to specify $m$. The OLS residual for this intercept-only regression is $e_t = y_t - \bar{y}$ and is highly autocorrelated. For example, the first four autocorrelations exceed 0.68. We set $m = 31$ as the autocorrelations at longer lags are

less than 0.2. Then the HAC standard error is .637, four times larger than the default .148, and the 95% confidence interval for population mean inflation is $(2.10, 4.61)$ compared to $(3.07, 3.65)$ using default standard errors.

**Remark 94** *Heteroskedastic- and autocorrelation consistent (HAC) robust standard errors are used if assumptions 1-2 hold, but model errors are autocorrelated and possibly heteroskedastic.*

In some time series applications it may be reasonable to assume that model errors are independent. A notable example is models for financial returns data, in which case errors may be independent if financial markets are efficient. In that case one may use heteroskedastic robust standard errors, the special case of HAC with $m = 0$.

### 10.7.3 Cluster Robust

A common setting for **panel data** on $N$ individuals observed over $T$ time periods is that model errors are independent across individuals but correlated over time for a given individual. So assumption 4 no longer holds.

A similar type of error correlation can arise in some **cross-section data** examples. For example, individuals may be clustered in villages with model errors correlated within villages but independent across villages.

In both settings the data are clustered or grouped so that model errors are correlated within cluster but are uncorrelated across clusters. Then, to ensure consistency of OLS, assumption 2 needs to be strengthened to the assumption that the error term for a given observation in time is uncorrelated with the regressor for all observations in the same cluster.

For inference one can use **cluster-robust standard errors** provided the number of clusters is large.

**Remark 95** *Cluster-robust standard errors are used if assumptions 1-2 hold, but model errors are correlated within clusters, uncorrelated across clusters, and possibly heteroskedastic. The $t$-statistic is then treated as $T(G-1)$ distributed, rather than $T(n-2)$ distributed, where $G$ is the number of clusters and $G$ needs to be large.*

## 10.8 Presentation of Regression Results

Published articles can differ in the method of presentation of regression results. The intercept and slope coefficients $b_1$ and $b_2$ are always reported, and $R^2$ is usually reported. But there can be great variation in the extent to which combinations of the standard error, $t$-statistic (for test that the population coefficient equals zero), and its associated $p$-value are reported. Given knowledge of one of these three, and knowledge of the slope coefficient, it is always

possible to compute the other two. For example, given $b_2$ and $se(b_2)$, we can compute $t = b_2/se(b_2)$ and $p = \Pr[|T| \leq t]$ where $T \sim T(n-2)$. Similarly, given $b_2$ and $t$ we can compute $se(b_2) = b_2/t$. Finally confidence intervals can be computed given $b_2$ and $se(b_2)$.

It is easiest if all four of $b$, $se(b)$, $t$ and $p$ are reported, along with confidence intervals. Indeed these are all given in typical computer output. But for space reasons, especially if there are several different models estimated or if the models have additional regressors, it is quite common for published studies to report only $b$ and one of $se(b)$, $t$ and $p$.

Thus for the house price regression, using the results in Table 10.1 with default standard errors, we might report the coefficients and standard errors

$$Price = \underset{(21489)}{115017} + \underset{(11.17)}{73.77} \times Size \qquad R^2 = 0.618.$$

Alternatively we may report the coefficients, along with $t$-statistics for whether the population coefficients equal zero

$$Price = \underset{(5.35)}{115017} + \underset{(6.60)}{73.77} \times Size \qquad R^2 = 0.618.$$

Or just the coefficients and $p$-values (for test of $\beta_1 = 0$ and $\beta_2 = 0$) may be reported

$$Price = \underset{(0.000)}{115017} + \underset{(0.000)}{73.77} \times Size \qquad R^2 = 0.618.$$

A fourth possibility is to report the 95% confidence intervals, possibly also with the coefficients

$$Price = \underset{(70925,159110)}{115017} + \underset{(50.84,96.70)}{73.77} \times Size \qquad R^2 = 0.618.$$

And yet another possibility, one that takes less space, is to report just coefficients along with asterisks that indicate the level of statistical significance. Then one, two, or three asterisks are used for statistical significance levels of, respectively, 10%, 5%, and 1%. Thus

$$Price = 115017^{***} + 73.77^{***} \times Size \qquad R^2 = 0.618.$$

Table 10.4 presents the same results reported using these various methods of presentation. Since there are various ways to compute the standard errors it is good practice to additionally include a table footnote stating the method used.

Using any of these alternatives we can verify that the slope coefficient is statistically significant at level 0.05. And while we have focused on the slope coefficient it is clear from this output that the intercept is also statistically significant at level 0.05.

Table 10.4: Regression estimates: Various ways to report the results.

| In parentheses: | Results 1 Standard errors | Results 2 t-statistics | Results 3 p-values | Results 4 95% Conf. int. | Results 5 |
|---|---|---|---|---|---|
| *Size* | 73.77 | 73.77 | 73.77 | 73.77 | 73.77*** |
| | (11.17) | (6.60) | (0.000) | (50.84,96.70) | |
| *Intercept* | 115017 | 115017 | 115017 | 115017 | 115017*** |
| | (21489) | (5.35) | (0.000) | (70925,159110) | |
| $R^2$ | 0.618 | 0.618 | 0.618 | 0.618 | 0.618 |
| n | 29 | 29 | 29 | 29 | 29 |

Note: Default standard errors are reported.

## 10.9 Prediction

From Chapter 8, the fitted value $\widehat{y} = b_1 + b_2 x^*$ can be used to predict $y$ for a given value of $x$, say $x^*$.

What is not made clear in Chapter 8 is that this prediction can be used in two quite different ways, and that these two ways have quite different precision. First we can estimate $\mathrm{E}[y|x] = \beta_1 + \beta_2 x^*$, the population conditional mean of $y$ given $x = x^*$. Second, we can predict the actual value of $Y$ given $x = x^*$. Since $y|x^* = \beta_1 + \beta_2 x^* + u$ this is more difficult to estimate as the error is unpredictable.

### 10.9.1 Prediction of the Conditional Mean

The **conditional mean** of $y$ for a given value of $x$, say $x^*$, is

$$\mathrm{E}[y|x^*] = \beta_1 + \beta_2 x^*.$$

The obvious estimate is

$$\widehat{y}_{cm} = b_1 + b_2 x^*,$$

which replaces the unknown population parameters by their least squares estimates. The subscript $cm$ is used to make clear that we are using $\widehat{y}$ to predict the conditional mean.

**Remark 96** *$\widehat{y}_{cm} = b_1 + b_2 x^*$ is an estimate of $E[y|x^*] = \beta_1 + \beta_2 x^*$, the conditional mean of $y$ given $x = x^*$.*

Since $\widehat{y}_{cm}$ is a linear combination of $b_1$ and $b_2$, its statistical properties follow from those of $b_1$ and $b_2$. This leads to $\widehat{y}_{cm}$ being unbiased and consistent for $\mathrm{E}[y|x^*]$ under assumptions 1-4, and $(\widehat{y}_{cm} - \mathrm{E}[y|x^*])/\sigma_{\widehat{y}_{cm}} \sim N(0,1)$ as $n \to \infty$. Here $\sigma^2_{\widehat{y}_{cm}} = \mathrm{Var}[\widehat{y}_{cm}]$ depends in part on $\sigma_u$. Replacing $\sigma_u$ by its estimate leads to a $t$-statistic. Some algebra leads to the following result.

**Remark 97** *When model assumptions 1-4 hold, so default standard errors for the OLS estimates are used, the* ***standard error*** *for the predicted conditional mean is*

$$se(\widehat{y}_{cm}) = s_e \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

*where* $s_e = \frac{1}{n-2}\sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2$ *is the standard error of the regression. Inference is based on* $t = (\widehat{y}_{cm} - E[y|x^*])/se(\widehat{y}_{cm})$ *which is the realization of a random variable that is approximately* $T(n-2)$ *distributed.*

The formula for $se(\widehat{y}_{cm})$ implies that the predicted conditional mean is more precise in the following situations.

1. The data are less scattered around the regression line, as then $s_e$ is smaller.
2. The greater the variation in regressors, as then $\sum_{i=1}^n (x_i - \bar{x})^2$ is larger.
3. The closer $x^*$ is to the sample mean as then $(x^* - \bar{x})^2$ is smaller.
4. The larger the sample size, as then $1/n$ is smaller and $(x^* - \bar{x})^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ is smaller.

### 10.9.2 Point Prediction of Actual Value

The actual value of $y$ given $x = x^*$ is $\beta_1 + \beta_2 x^* + u^*$, which adds the error term $u^*$ to the conditional mean $\beta_1 + \beta_2 x^*$. Due to this additional error term means it is much more difficult to predict the actual value than it is to predict the conditional mean.

The obvious estimate of the **actual value** of $y$ given $x = x^*$ is

$$\widehat{y}_f = b_1 + b_2 x^*,$$

since the best estimate of the error term $u^*$ is its conditional mean of zero. This estimate is called a **forecast**. The subscript $f$ is used to make clear that we are using $\widehat{y}$ to forecast $y$.

**Remark 98** $\widehat{y}_f = b_1 + b_2 x^*$ *is an estimate of* $y|x^* = \beta_1 + \beta_2 x^* + u$*, the actual value of* $y$ *given* $x = x^*$.

The forecast $\widehat{y}_f$ equals $\widehat{y}_{cm}$, but they are used for different purposes. The forecast $\widehat{y}_f$ is used to predict $\beta_1 + \beta_2 x + u$, not just $\beta_1 + \beta_2 x$. Since $u$ is pure noise that is unpredictable and is uncorrelated with $b_1$ and $b_2$, it follows that the variance of the forecast is

$$\text{Var}[\widehat{y}_f] = \text{Var}[\widehat{y}_{cm}] + \text{Var}[u].$$

Now $\text{Var}[\widehat{y}_{cm}]$ is estimated by $se(\widehat{y}_{cm})^2$, defined earlier, and $\text{Var}[u]$ is estimated by $s_e^2$, the standard error of the regression. Taking the square root gives $se(\widehat{y}_f)$.

**Remark 99** *When model assumptions 1-4 hold, so default standard errors for the OLS estimates are used, the* ***standard error for the predicted conditional mean*** *is*

$$se(\widehat{y}_f) = \sqrt{se(\widehat{y}_{cm})^2 + s_e^2}.$$

*Inference is based on* $t = (\widehat{y}_f - (y|x^*))/se(\widehat{y}_f)$ *which is the realization of a random variable that is approximately* $T(n-2)$ *distributed.*

Note that $se(\widehat{y}_{cm})$ is necessarily at least $s_e$, regardless of the sample size. This makes sense since the model error is unforecastable and has standard deviation of $\sigma_u$ that is estimated by $s_e$. By contrast $se(\widehat{y}_{cm})$ can approach zero as the sample size gets larger and $\beta_1$ and $\beta_2$ are more precisely estimated.

### 10.9.3 Example: House Price given House Size

Consider prediction of the conditional mean price of a house that is 2000 square feet in size.

$$\widehat{y}_{cm} = b_1 + b_2 x^* = 115017 + 73.771 \times 2000 = 262559.$$

The conditional mean price is predicted to be \$262,559. The standard error of $\widehat{y}_{cm}$ when assumptions 1-4 hold is computed as follows. From regression output, $n = 29$ and $s_e = 23551$, and from summary statistics $\bar{x} = 1882.76$ and $s_x = 389.2721$. It follows that $\sum_i (x_i - \bar{x})^2 = (n-1)s_x^2 = 28 \times 398.2721^2 = 4441379$. Then

$$se(\widehat{y}_{cm}) = 23551 \times \sqrt{\frac{1}{29} + \frac{(2000 - 1882.76)^2}{4441379}} = 4565.$$

Next consider prediction of the actual value of the price of a house that is 2000 square feet in size. Then $\widehat{y}_f = b_1 + b_2 x^* = 262559$, with standard error

$$se(\widehat{y}_f) = \sqrt{4565 + 23551^2} = 23989.$$

While both the conditional mean and actual value of a house of size 2000 square feet are estimated to be \$262,559, the actual value is much less precisely estimated with a standard error that is five times that of the estimate for the conditional mean.

The resulting 95% confidence interval for $\mathrm{E}[y|x = 2000]$ is $262559 \pm 2.052 \times 4565.3 = (253192, 271927)$, since $t_{.025,27} = 2.052$. The resulting 95% confidence interval for the actual value of $y$ given $x = 2000$ is $262559 \pm 2.052 \times 23989 = (213338, 311781)$. The confidence interval for $y$ given $x = 2000$ is of width almost \$100,000, whereas the confidence interval for $\mathrm{E}[y|x = 2000]$ was only of width about \$20,000. It is much more difficult to forecast accurately the actual value.

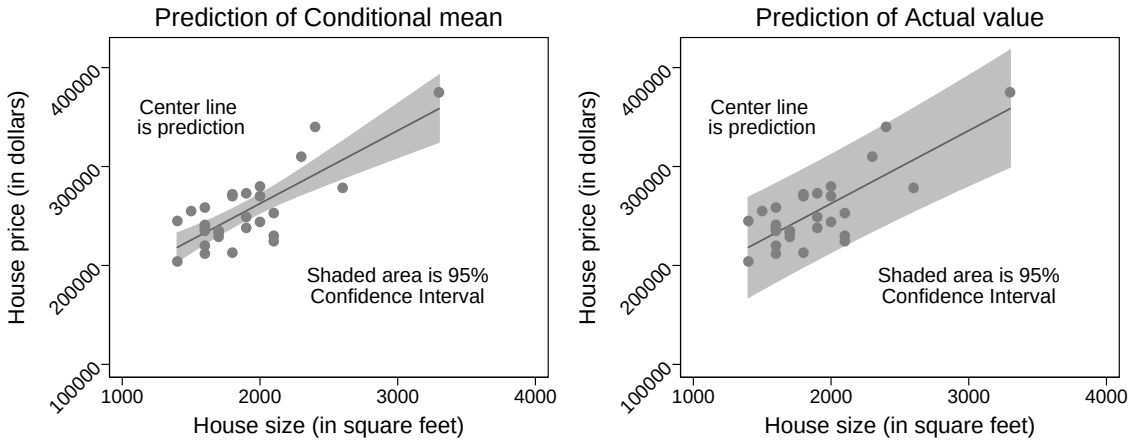


Figure 10.2:

Figure 10.2 presents predictions from the fitted model of house price for house size ($x^*$) ranging from 1400 to 3300 square feet. The center line in each panel gives the point prediction, $\widehat{y} = b_1 + b_2 x^* = 115017 + 73.771x^*$.

The first panel gives the 95 percent confidence interval for the prediction of the conditional mean. The 95% confidence interval is narrowest at the sample mean $\bar{x} = 1883$ and widens as the regressor value $x^*$ moves away from $\bar{x}$.

The second panel gives the 95% confidence interval for the actual price of the house given house size. In this example the interval is so wide that it is of little use for someone trying to select an appropriate price for a particular house using information on size alone.

### 10.9.4 Prediction with Robust Standard Errors

If only assumptions 1-2 hold then the predictions $\widehat{y}_{cm}$ and $\widehat{y}_f$ remain unbiased and consistent. But the associated standard error formulas change.

Most statistical packages have a command that will compute robust standard errors for $\widehat{y}_{cm}$ that in turn are based on the appropriate robust standard errors for the OLS coefficients $b_1$ and $b_2$.

For $\widehat{y}_f$ we need an estimate of $\text{Var}[\widehat{y}_f] = \text{Var}[\widehat{y}_{cm}] + \text{Var}[u|x^*]$, where there is now the extra complication that errors are now heteroskedastic as assumption 3 does not hold. For $\text{Var}[\widehat{y}_{cm}]$ we use the robust estimate $s_e^2(\widehat{y}_{cm})$. Estimating $\text{Var}[u|x^*]$ is now more problematic as it varies with $x^*$. The simplest solution is to assume that the error variance for $u|x^*$ equals the sample average error variance, in which we use $s_e^2$.

### 10.9.5 Discussion

The confidence interval for forecasts of an individual value can be wide. This will happen whenever the variance of the model error is large, so that the model does not explain the dependent variable well.

To see this, recall that $se(\widehat{y}_f) \geq s_e$, so a 95% confidence interval is, approximately, at least $\widehat{y}_f \pm 2 \times s_e$ where $s_e$ is the standard error of the regression, which can be quite large.

### 10.9.6 Are Poor Forecasts a Problem?

Econometric models, especially those estimated using cross-section data, can have low $R^2$, and consequently relatively large error variance. This can lead to very poor forecasts of the actual value for an **individual observation**. But the predictions can still be very informative about **average behavior**. Recall that the standard error for the conditional mean is much smaller than that for a forecast, and can approach zero as the sample size gets large.

For example, many studies find that on average education has a statistically significant impact on earnings, and furthermore this impact is large in magnitude. This is the case even though for an individual the confidence interval for forecast earnings given years of education is very wide. Knowing that on average high levels of education are predicted to lead to high earnings is very useful information for policy makers, even though we cannot predict with much certainty that a given person with a high level of education will have high earnings.

## 10.10 Examples

To come.

## 10.11 Key Concepts

1. Confidence intervals and hypothesis tests on the population slope $\beta_2$ are based on the $t$-statistic $t = (b_2 - \beta_2)/se(b_2)$.

2. The $t$-statistic is treated as being $T(n-2)$ distributed. This is an approximation that improves as $n \to \infty$. If assumptions 1-4 hold and additionally the errors are normally distributed then the $t$-statistic is exactly $T(n-2)$ distributed.

3. A $100(1-\alpha)\%$ confidence interval for $\beta_2$ is $b_2 \pm t_{n-2,.\alpha/2} \times se(b_2)$. This interval will include $\beta_2$ with probability $1-\alpha$.

4. It is most common to use 95%, 90% and 99% confidence intervals, corresponding to $\alpha$ equal to .05, .10 and .01.

5. An approximate 95% confidence interval for $\beta_2$ is $b_2 \pm 2se(b_2)$.

6. A two-sided hypothesis test of $H_0 : \beta_2 = \beta_2^*$ against $H_a : \beta_2 \neq \beta_2^*$ can be computed using either $p$-values or critical values.

7. A test of statistical significance is the special case that $\beta_2^* = 0$.

8. A statistically significant variable can be of little economic significance, and a statistically insignificant variable can still be of considerable economic significance.

9. For a two-sided test, the $p$-value is the probability of observing a $T(n-2)$ distributed random variable at least as large as the observed $t$ in absolute value, i.e., $p = \Pr[|T_{n-2}| \geq |t|]$. $H_0$ is rejected at significance level $\alpha$ if $p < \alpha$.

10. For a two-sided test at significance level $\alpha$, the critical value $c$ is that value for which there is probability $\alpha$ of observing a $T(n-2)$ distributed random variable at least as large as $c$, so $c$ is such that $\Pr[|T_{n-2}| \geq c] = \alpha$. We reject $H_0$ at level $\alpha$ if $|t| < c$.

11. An approximate two-sided test at level 5% is to reject $H_0$ if $|t| > 2$, or equivalently if $|b_2 - \beta_2^*| > 2se(b_2)$.

12. A type I error occurs if $H_0$ is rejected when $H_0$ is true. Test size is the probability of a type I error.

13. A type II error occurs if $H_0$ is not rejected when $H_0$ is false. Test power is one minus the probability of a type II error.

14. The test size is set to be low, such as $\alpha = 0.05$, while the test method is ideally one for which test power is high given test size.

15. For one-sided tests we make the claim the alternative hypothesis.

16. An upper one-tailed test is a test of $H_0 : \beta_2 \leq \beta_2^*$ against $H_a : \beta_2 > \beta_2^*$.

17. A lower one-tailed test is a test of $H_0 : \beta_2 \geq \beta_2^*$ against $H_a : \beta_2 < \beta_2^*$.

18. Robust standard errors are used when assumptions 3 and/or 4 do not hold.

19. Heteroskedasticity-robust standard errors are used when model errors are heteroskedastic and independent (assumption 3 does not hold).

20. Heteroskedasticity- and autocorrelation-consistent (HAC) robust standard errors are used when model errors are heteroskedastic and autocorrelated over time (assumptions 3 and 4 do not hold).

21. Cluster-robust standard errors are used when model errors are heteroskedastic, correlated within cluster and uncorrelated across clusters (assumptions 3 and 4 do not hold).

22. Regression results given in published articles usually give the estimated coefficients along with just one of the following: standard error, $t$-statistic, $p$-value and 95% confidence interval.

23. The conditional mean of $y$ given $x = x^*$ is estimated by $\widehat{y}_{cm} = b_1 + b_2 x^*$ .

24. The actual value of $y$ given $x = x^*$ is estimated by $\widehat{y}_f = b_1 + b_2 x^*$. In fact $se(\widehat{y}_{cm}) = \sqrt{se(\widehat{y}_{cm})^2 + s_e^2}$.

25. The actual value of $y$ given $x = x^*$ is much more noisily estimated than is the conditional mean. In fact

26. Key Terms: t-statistic; confidence interval; null hypothesis; alternative hypothesis; significance level; statistical significance; economic significance; two-sided test; one-sided test; rejection; $p$-value; critical value; critical region; rejection region; type I error; type II error; test size; test power; most powerful test; heteroskedasticity; robust standard errors; heteroskedastic-robust standard errors; HAC standard errors; cluster-robust standard errors; presentation of regression results; prediction; conditional mean; actual value.

# Chapter 11

# © A.Colin Cameron: Bivariate Case Studies

This chapter provides four detailed examples of bivariate data analysis. The first example analyses monthly data on the returns on Coca-Cola stock compared to returns on the market portfolio, using the capital asset pricing model. The second example analyses the relationship between real GDP growth and unemployment using annual macroeconomic data for the U.S. The third and fourth examples analyze cross-country data on, respectively, health outcomes and health expenditures.

In these examples the initial analysis uses default standard errors for the regression coefficients. Then consideration is given to using alternative robust standard errors, introduced in Chapter 10.7. The health examples estimate models in natural logarithms in addition to models in levels.

## 11.1 Capital Asset Pricing Model

The **capital asset pricing model** (CAPM) is a workhorse model for finance. It formalizes the relationship between the return on a given stock and the return on the market portfolio.

### 11.1.1 Theory of CAPM

We consider the relationship between the return on a risk-free asset, the return on the market portfolio, and the return on a (risky) investment asset.

Let $RF$ denote the **risk-free interest rate**. In the example below $RF$ is the interest rate on a one-month U.S. Treasury bill. And let $RM$ denote the **overall market return** on stocks. In the example below $RM$ is the value-weighted return on all stocks listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX) and NASDAQ.

The difference $(RM - RF)$ is called the **market excess return** or the **equity market premium** as it gives the additional return received by investing in the market compared to holding a riskless asset.

Now consider individual investments. Let $RA$ denote the return on the **investment asset** $A$, such as an individual stock or a stock portfolio. In the example below we focus on stock in Coca-Cola.

The CAPM links the returns on individual investments to the market return, according to the relation

$$\mathrm{E}[RA_t - RF_t] = \beta_A \mathrm{E}[RM_t - RF_t],$$

where the subscript $t$ denotes time. The expected return on an individual investment in excess of the risk-free rate, $\mathrm{E}[RA - RF]$, is just a multiple of the expected excess return in the market. The investment-specific coefficient $\beta_A = \mathrm{Cov}[RA_t, RM_t]/\mathrm{Var}[RA_t]$ and is a measure of the correlation of the investment with the market portfolio.

The multiple $\beta_A$ varies across investments. In aggregate individual investments sum to the market so on average $\beta_A$ equals one. Some stocks have $\beta_A > 1$ and are often called **growth stocks** as they will outperform the market when the equity premium $RM_t - RF_t$ is positive. At the same time they will underperform the market when the equity premium is negative. Other stocks have $\beta_A < 1$ and are often called **value stocks**. Utilities are an example. Cash has $\beta_A = 0$. An investment with $\beta_A < 0$ is viewed as a hedge or insurance policy.

### 11.1.2 Estimation of CAPM

The population values of $\mathrm{E}[RA_t - RF_t]$, $\mathrm{E}[RM_t - RF_t]$ and $\beta_A$ are unknown. Using time series data we instead estimate by OLS for investment $A$ the model

$$RA_t - RF_t = \alpha_A + \beta_A(RM_t - RF_t) + u_t,$$

where the error term $u_t$ is assumed to have zero mean and to be uncorrelated with the regressors, so that $\mathrm{E}[u_t|(RM_t - RF_t)] = 0$.

The estimated slope coefficient $b_A$ is called the stock's **beta.** The estimated intercept $a_A$ is called the stock's **alpha.** According to the CAPM $\alpha_A = 0$. If instead $\alpha_A > 0$, then the investment has performed better than is to be expected given its variability and correlation with the market portfolio as measured by $\beta_A$. Warren Buffett's company Berkshire Hathaway has for many years had an alpha that is both positive and high, though this is not a guarantee of future performance.

### 11.1.3 Application

The dataset CAPM has monthly data on the returns to holding stock in Coca-Cola, Target and Walmart from May 1983 to October 2013. It also includes the one-month U.S. Treasury

Table 11.1: CAPM data: Variable definitions and summary statistics

| Variable | Definition | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *RM* | Market return | .0091 | .0456 | -.2254 | .1285 |
| *RF* | One-month U.S. Treasury Bill rate | .0035 | .0022 | .0000 | .0100 |
| *RKO* | Return on Coca-Cola | .0137 | .0618 | -.1909 | .2266 |
| *RTGT* | Return on Target | .0138 | .0842 | -.4781 | .2673 |
| *RWMT* | Return on Walmart | .0156 | .0703 | -.2698 | .2644 |
| *RM-RF* | Excess Market Return (equity premium) | .0055 | .0456 | -.2314 | .1243 |
| *RKO-RF* | Excess Return on Coca-Cola | .0102 | .0616 | -.1952 | .2188 |
| *RTGT-RF* | Excess Return on Target | .0103 | .0842 | -.4841 | .2629 |
| *RWMT-RF* | Excess Return on Walmart | .0121 | .0702 | -.2758 | .2612 |

bill rate and a market return that is the value-weighted return on all stocks listed on the NYSE, AMEX and NASDAQ. The data are summarized in Table 11.1.

The market return and individual stock returns substantially exceed the risk-free rate, reflecting a reward for holding a riskier asset. The average market excess return is .55% per month, an annual compounded return of 6.8% ($= 1.0055^{12}$). The average returns on individual stocks are higher than the market return, reflecting the greater risk due to less diversification – the standard deviations of returns on the three individual stocks are 35% to 65% higher than the standard deviation of the market return.

The first panel of Figure 11.1 is a time series plot of the market return and the return to Coca-Cola stock from January 2007 to September 2013, the last 20% of the sample period. It is clear that the Coca-Cola stock return does follow the market return, though not perfectly. The financial crash in late 2008 is also apparent, with market monthly returns of -9.8%, -18.5% and -8.6% in, respectively, September, October and November for a cumulative loss of 33% over three months ($= 100 \times (1 - 0.902 \times 0.815 \times 0.914)$).

The second panel of Figure 11.1 displays a scatter plot of the Coca-Cola return against the market return. A nonparametric regression line estimated by local linear methods, see Chapter 8.11, is very close to linear, with departure from linearity only for the very extreme low and high excess returns. So a linear model, implied by the CAPM model, seems appropriate for these data.

OLS regression gives fitted CAPM model for Coca-Cola (RKO)

$$(RKO - RF) = \underset{(0.00295)}{0.00681} + \underset{(0.0644)}{0.6063} \times (RM - RF), \quad R^2 = 0.201, \quad s_e = 0.055, \quad n = 354,$$

where default standard errors are given in parentheses.

The slope coefficient, the stock's beta, is statistically different from zero, since $t = 0.6063/0.0644 = 9.41 > t_{0.025,352} = 1.967$. The slope coefficient is also statistically different from one, since $t = (0.6063 - 1)/0.0644 = -6.11$ and $|-6.11| > t_{0.025,352} = 1.967$.

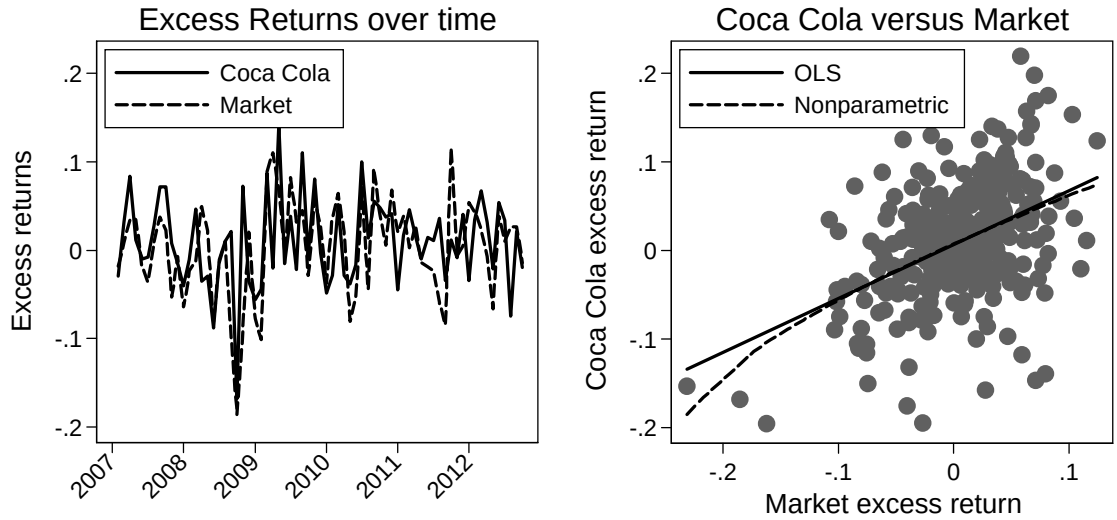


Figure 11.1: CAPM Model: Coca Cola and market excess returns

Since Coca-Cola's beta lies between 0 and 1 it is viewed to be a value stock. Large companies such as Coca-Cola generally move less than the market as a whole, leading to $\beta < 1$.

The intercept coefficient, the stock's alpha, is a risk-adjusted measure of stock performance that measures the return in excess of that expected given the riskiness of the stock. The CAPM model in its purest form restricts $\alpha = 0$. This restriction is rejected here at significance level 0.05, since $t = 0.00681/0.00295 = 2.31 > t_{0.025,352} = 1.967$. Furthermore the alpha is large in magnitude. The monthly return for Coca-Cola stock is 0.68 basis points higher than expected given its beta, leading to a compounded annual return that is 8.5 percentage points ($= 100 \times (1.0068^{12} - 1)$) higher. Such high alpha arises in part due to survivor bias. A company whose stock has a large negative alpha will eventually face bankruptcy or a takeover. Since on average alpha is expected to be zero, companies such as Coca Cola that do survive for a long time can be expected to have a positive alpha in a time series analysis such as this.

Analyses of the monthly excess returns of stock in Target and Walmart are left as exercises.

### 11.1.4 Robust Standard Errors

The preceding statistical inference uses default standard errors based on model assumptions 1-4.

It is likely that when the regressor $(RM - RF)$ is large in absolute value the error in the linear model for the dependent variable $(RKO - RF)$ is also large in absolute value. In that case the errors are heteroskedastic rather than homoskedastic. When heteroskedastic robust standard errors are computed there is little change in the standard error for the intercept. The standard error of the slope coefficient increases from 0.0644 to 0.0770, so again the estimated beta is statistically significantly different from both zero and one. For these data there is little difference between default and heteroskedastic-robust standard errors.

For time series data the model errors may be correlated over time, in which case model assumption 4 is incorrect. For financial data, however, excess returns are intrinsically not forecastable if markets are efficient. So the error term should be uncorrelated. This is the case for these data. Let $e_t$ be the fitted residual from the original OLS regression. Then the correlation between $e_t$ and $e_{t-1}$ is $-0.039$, very close to zero.

To guard against possible model error autocorrelation we use heteroskedastic- and autocorrelation-consistent (HAC) standard errors, see Chapter 10.7, allowing for the errors to be potentially correlated for up to 13 months. Then there is little change in the standard error for the intercept, while the HAC standard error of the slope coefficient is 0.0885, fifteen percent higher than the heteroskedastic-robust standard error of 0.0770.

More generally for data on financial returns it is often sufficient to use heteroskedastic robust standard errors.

## 11.2 Output and Unemployment

Over the business cycle there is a negative relationship between unemployment and GDP growth rates, with unemployment falling as GDP growth rises.

### 11.2.1 Okun's Law

Furthermore, unemployment fluctuates less than GDP growth. Each percentage point increase in the unemployment rate is associated with an approximate two percentage point decrease in the GDP growth rate. This phenomenon is called **Okun's law**, after Okun who first proposed it in a 1962 journal article. A better term is "Okun's rule-of-thumb" as it is an empirical relationship rather than an ironclad law.

### 11.2.2 Application

We investigate this relationship using annual U.S. data from 1948 to 2013. The dataset GDPUNEMPLOY has data on *Growth*, the annual percentage growth in real GDP, and on *URATEchange*, the annual change in the percentage unemployment rate for the civilian

Table 11.2: Output and unemployment data: Variable definitions and summary statistics

| Variable | Definition | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *Growth* | Annual percentage growth in real GDP | 3.255 | 2.401 | -2.802 | 8.715 |
| *URATEchange* | Annual change in the unemployment rate | 0.053 | 1.089 | -2.100 | 3.500 |

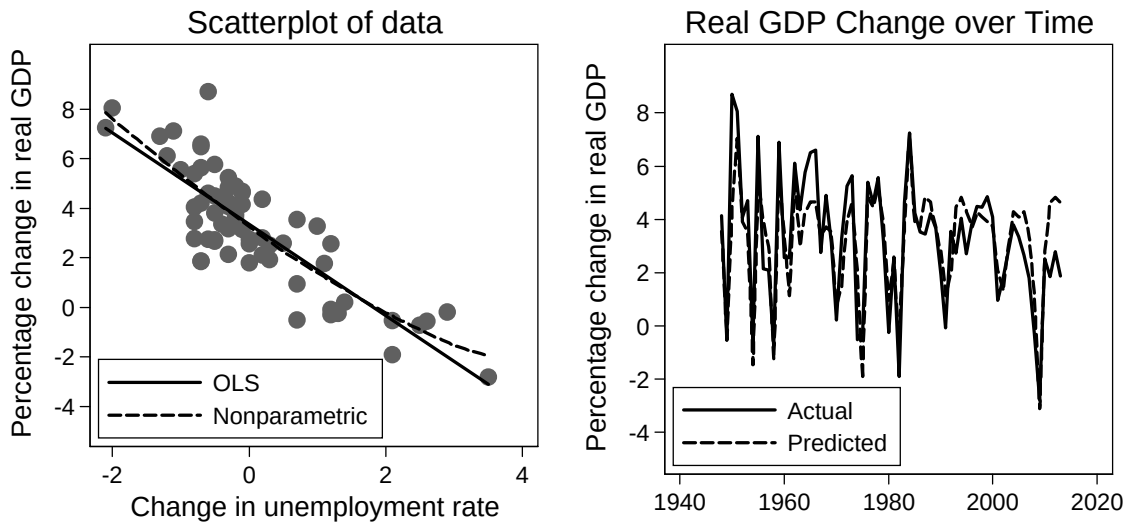


Figure 11.2: Okun's Law: Percentage change in output and change in unemployment rate

population aged 16 years and older. For example, if the unemployment rate expressed as a percentage increases from 5.3% to 6.5% then *URATEchange* equals 1.2.

Table 11.2 summarizes the data. Over this period real GDP grew on average 3.25% per year and was approximately twice as volatile as the annual change in the unemployment rate, with standard deviation 2.40 compared to 1.09.

The first panel of Figure 11.2 presents a scatter plot of the data. A nonparametric regression line estimated by local linear methods, see Chapter 8.11, is very close to linear.

OLS regression yields

$$Growth = \underset{(0.162)}{3.352} - \underset{(0.150)}{1.858} \times URATEchange, \quad R^2 = 0.703, \quad s_e = 1.319, \quad n = 66,$$

where default standard errors are given in parentheses. A useful approximation is that in the post World War II period, U.S. real GDP growth has averaged 3.35 percent per year, and

declines by 1.86 percentage points for each 1 percentage point increase in the unemployment rate. The slope coefficient of 1.86 is close to Okun's estimate of 2.

The slope coefficient is highly statistically significant with $t = -1.858/.150 = -12.31$. A more interesting test is whether the data confirm Okun's law. For a test of $H_0 : \beta_2 = -2.0$ against $H_a : \beta_2 \neq -2.0$, $t = (1.858 - 2.0)/0.150 = -0.95$, so $p = \Pr[|T_{23}| \geq 0.95] = 0.36$. The null hypothesis is not rejected at significance level 0.05, so Okun's law is not rejected by the data.

### 11.2.3 Prediction

How well does this estimated model fit the economic downturn following the global financial crisis? In the two years 2007 to 2009 the unemployment rate rose from 4.6 to 9.3 percent. The model predicts that real GDP would change over these two years by $2 \times 3.352 - 1.858 \times (9.3 - 4.6) = -2.0$ percent. This is reasonably close to the actual fall in real GDP from 2007 to 2009 of 3.1 percent.

The model does not do so well for the subsequent recovery. From 2009 to 2013 the unemployment rate fell from 9.3% to 7.4%. The model predicts that real GDP would change over these four years by $4 \times 3.352 - 1.858 \times (7.4 - 9.3) = 16.9$ percent. But in fact real GDP has only grown by 9.3 percent from 2007 to 2011.

The second panel of Figure 11.2 plots both the actual change in real GDP growth rate and the change predicted from the regression model. The predictions track the actual changes reasonably well, but there is considerable overprediction of the real GDP growth rate in the most recent years, confirming the preceding numerical analysis. Okun's rule of thumb performs poorly in explaining the economic recovery following the global financial crisis. One reason is that the global financial crisis has seen job loss not only through higher unemployment rates, the only explanatory variable in the model estimated here, but also reduced labor force participation, as detailed in Chapter 3.3.

### 11.2.4 Robust Standard Errors

For time series data the model errors may be correlated over time, in which case model assumption 4 is incorrect.

For a model with the dependent variable measured as a change rather than a level, however,this error autocorrelation is greatly reduced. For example, for these data the autocorrelation between real GDP and previous year's real GDP is 0.96, while the correlation between real GDP growth and previous year's real GDP growth in the previous year is only 0.14. And the autocorrelation in the residual from regression of *Growth* on *URATEchange* with the previous year's residual is 0.22.

Due to the low residual autocorrelation, HAC robust standard errors are likely to be similar to the default standard errors in this specific time series application. The HAC

Table 11.3: Health data: Variable definitions and summary statistics

| Variable | Definition | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *Hlthpc* | Annual health expenditure per capita (in US $) | 3256 | 1494 | 923 | 7990 |
| *Lifeexp* | Life expectancy at birth of males (in years) | 76.7 | 2.94 | 69.8 | 79.9 |
| *InfMort* | Infant mortality per 1,000 live births | 4.44 | 2.72 | 1.8 | 14.7 |

robust standard error of the coefficient, allowing for the errors to be potentially correlated out to five years, is 0.129. This is a bit lower than the default standard error of 0.150, and is similar to the heteroskedastic robust standard error of 0.133.

For macroeconomics data it is safest to use HAC standard errors. This can make a big difference when the dependent variable is measured in levels. It makes a smaller difference when the dependent variable is a change, as was the case for this example.

## 11.3 Health Outcomes

The two most widely-used measures of country health outcomes, available for most countries, are life expectancy and infant mortality. These are expected to improve as more resources are devoted to health care, measures by health expenditures per capita. Qualitatively similar results are obtained if instead the regressor is health expenditure as a percentage of GDP or the regressor is GDP per capita.

### 11.3.1 Application

Dataset HEALTH2009 has 2009 data for the 34 wealthy and relatively wealthy nations in the Organization of Economic and Community Development (OECD). The countries, ordered alphabetically, are Australia, Austria, Belgium, Canada, Chile, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom, and United States.

Table 11.3 provides some summary statistics. There is wide variation in annual health expenditures per capita, ranging from $923 in Mexico to $7,990 in the United States. Life expectancy at birth of males ranges from 69.8 years in Estonia to 79.9 years in Switzerland. Infant mortality (death in the first year of life) ranges from 1.8 per one thousand live births in Iceland to 14.7 in Mexico.

A useful way to compare variation across different series, measured on different scales, is to use the coefficient of variation. This normalizes by the average of the series, since it

equals the standard deviation divided by the sample mean. This makes clear that there is much less variation in life expectancy ($CV = 0.04$) across countries than there is in the other two measures.

### 11.3.2 Linear Regression

**Life expectancy at birth** in 2009 is calculated by assuming that someone born in 2009 will face the same death rates as they age as the death rates computed for each year of age in 2009.

OLS regression yields

$$\mathit{Lifeexp} = \underset{(71.36)}{73.08} + \underset{(3.88)}{0.00111} \times \mathit{Hlthpc}, \quad R^2 = 0.320, \quad s_e = 2.46, \quad n = 34,$$

where $t$-statistics based on default standard errors are given in parentheses. Here life expectancy is for males, but qualitatively similar results are obtained for females or for both genders combined.

The relationship between health spending and life expectancy is quantitatively significant. A \$1,000 increase in per capita health spending, a two-thirds of a standard deviation change, is associated with an increase in life expectancy of 1.11 years.

The relationship is highly statistically significant, as $t = 3.88$. Since the prior belief is that $\beta_2 > 0$ in this example, it is most appropriate to perform a one-sided test of $H_0 : \beta_2 \leq 0$ against $H_a : \beta_2 > 0$. Then $c = t_{.05;32} = 1.69$ and we reject $H_0$ at significance level 0.05 since $t = 3.88 > c$. The $p$-value, not listed, is $p = \Pr[T_{32} > 3.88] = 0.000$ to three decimal places.

The first panel of Figure 11.3 plots the data and regression line. The country codes are generally the first three letters of the country name, exceptions being AUT for Austria, CHL for Chile, NZ for New Zealand, SLR for Slovak Republic, UK for the United Kingdom, and USA for the United States. The scatter plot suggests that the relationship is curvilinear rather than linear, and that the U.S. has much lower life expectancy (six years lower) than expected given its high level of health spending.

### 11.3.3 Robust Standard Errors

With cross-section data, model errors may be heteroskedastic and it is now common econometric practice to use heteroskedastic-robust standard errors.

For the linear regression the slope coefficient has default standard error 0.000287 and heteroskedastic-robust standard error 0.0004637, so the $t$-statistic falls considerably from 3.88 to 2.40 when heteroskedastic-robust standard errors are used. The regressor *Hlthpc* nonetheless remains statistically significant at level 0.05.

The observed substantial difference between default and heteroskedastic robust standard errors (60% larger) is unusual. It arises because the USA residual $e_i$ is very large, see the first

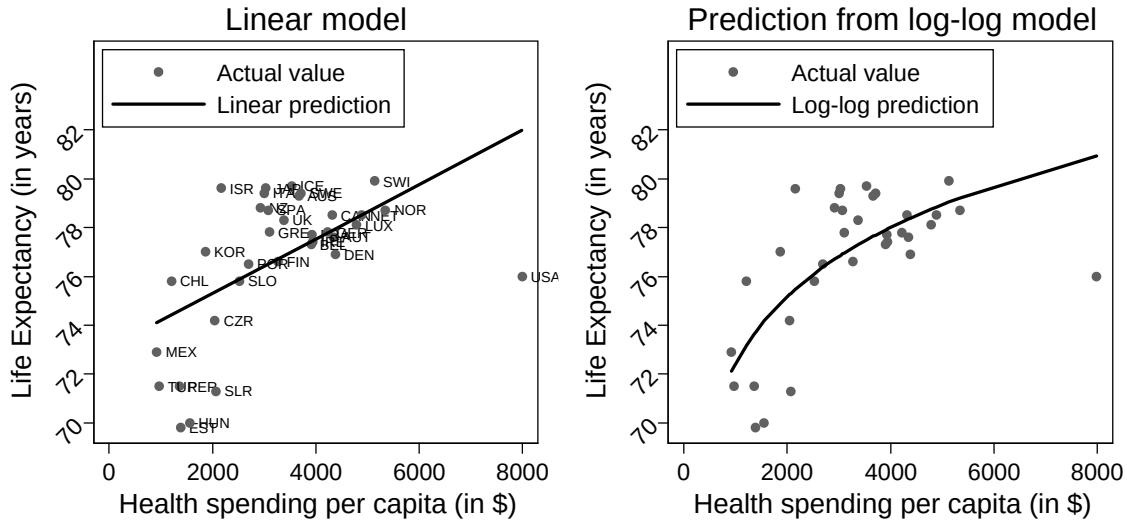


Figure 11.3: Health Outcomes: Life expectancy and health spending per capita

panel of Figure 11.3. As a result $e_i^2(x_i - \bar{x})^2$ is unusually large for this observation and hence $se_{het}(b_2) = \sqrt{\sum_{i=1}^n e_i^2(x_i - \bar{x})^2} / \sum_{i=1}^n (x_i - \bar{x})^2$ is large. Dropping the U.S. observation, the default standard error is 0.000300 and the heteroskedastic-robust standard error is 0.000301, essentially the same.

### 11.3.4 Log-Log Model

This subsection estimates a log-log model, introduced in the next chapter (Chapter 12). It illustrates some of the issues analyzed in more detail in Chapter 12.

The scatter plot in Figure 11.3 clearly indicates that the relationship is nonlinear rather than linear. One way to capture the nonlinear relationship is to perform OLS regression where the dependent variable and the regressor are first transformed by taking the natural logarithm. Thus we estimate the model $\ln y = b_1 + b_2 \ln x$, called a log-log model.

OLS regression in natural logarithms yields

$$\ln(LifeExp) = \underset{(50.75)}{3.91} + \underset{(5.53)}{0.053} \times \ln(Hlthpc), \quad R^2 = 0.488, \quad s_e = 0.028, \quad n = 34,$$

where $t$-statistics based on default standard errors are given in parentheses.

As discussed in Chapter 12.6, the slope coefficient in a log-log model can be interpreted as an elasticity. So a 1% increase in health spending per capita is associated with a 0.053% increase in life expectancy, and a 10% increase with a 0.53% increase in life expectancy.

The relationship is highly statistically significant, as $t = 5.53$. Since the prior belief is that $\beta_2 > 0$ in this example, it is most appropriate to perform a one-sided test of $H_0 : \beta_2 \leq 0$ against $H_a : \beta_2 > 0$. Then $c = t_{.05;32} = 1.69$ and we reject $H_0$ at significance level 0.05 since $t = 5.53 > c$.

### 11.3.5 Prediction

In this subsection we compare the predictions of life expectancy from the linear model with those from the log-log model.

Prediction of *Lifeexp* from the linear model is straight-forward, with $\widehat{y} = 73.08 + 0.00111x$, where $y$ is *Lifeexp* and $x$ is *Hlthpc.*

Prediction from the log-log model is not so straightforward. The model directly predicts $\ln y$, but we want a prediction of $y$. For the log-log model $\widehat{\ln y} = b_1 + b_2 \ln x$ it is natural to take the exponential of both sides and use $\widehat{y} = \exp(b_1 + b_2 \ln x)$. As explained in Chapter 12.5, however, these predictions should be multiplied by $\exp(s_e^2/2)$, where $s_e$ is the standard error of the regression from the log-log model. This multiple is the exact multiple to use if the log-log model errors satisfy assumptions 3-4 and are normally distributed, and may be a good approximation in other situations.

For the current model the multiple is $\exp(s_e^2/2) = \exp(0.028^2/2) = 1.0004$. So we use $\widehat{y} = 1.000040 \times \exp(3.91 + 0.053 \ln x)$; in other examples the correction factor can be more substantial.

The right panel of Figure 11.3 presents the in-sample predictions of life expectancy from the log-log model. The log-log model picks up the curvature evident in the scatter plot. The U.S. is still an outlier, however. Life expectancy is four years less than predicted by the log-log model, compared to six years less than predicted from the original linear model.

The accuracy of the predictions from the two models can be compared using the correlation $r_{\widehat{y}y}$ between predicted and actual value. For the linear model this is $R = \sqrt{0.488} = 0.70$, since for linear regression $R^2 = r_{\widehat{y}y}^2$; see Chapter 8.7. For the log-log model, additional calculation finds that the correlation coefficient between the predictions plotted in the second panel of Figure 11.3 and the actual values equals 0.69. So there is little difference between the two models on this measure. Nonetheless the log-log model is preferred as it better captures the curvature of the relationship.

### 11.3.6 Discussion

For **infant mortality** qualitatively similar results are obtained. The relationship is nonlinear, and a log-log model fits the data better than a linear model. This model finds that a 10% increase in health spending per capita is associated with a substantial 4.58% decrease in infant mortality. Again the U.S. is an outlier. Even in the raw data it has the fourth highest infant mortality rate, despite relatively large per capita health expenditure. The predicted

value for the U.S. from the log-log model is 2.7 deaths per 1,000 live births, much less than the observed 6.4 deaths.

This cross-country comparison example indicates that two key health outcome measures improve as health expenditures are increased. These improvements in outcomes are much greater for samples that included much poorer countries, a result consistent with the curvilinear relationship given in the second panel of Figure 11.3. Especially for low-income countries, increased health expenditures can have a great impact on basic outcomes such as infant mortality and life expectancy. While Mexico is the poorest country in this sample of OECD countries, it should be noted that its per capita GDP exceeds the world average and exceeds that of the median country. The world's poorest countries have GDP per capita of $500, life expectancy of 50 years and infant mortality of 100 per one thousand live births.

The analysis also indicates that the U.S. is a clear outlier among OECD countries. From a statistical perspective, should we include this outlier in the analysis? If the U.S. observation is dropped the slope coefficient in the linear model increases substantially from 0.00111 to 0.00169. The difference is much less in the log-log model, with slope coefficient increasing from 0.0534 to 0.0619, another reason for preferring the log-log model. The practice in economic analysis is to not drop outliers, unless there is a compelling reason for doing so. In the current setting it is standard to include the U.S. in the analysis, though perhaps it might be mentioned as an aside that it is an outlying observation.

From an economic policy perspective it is clear that health outcomes in the U.S. are much lower than predicted given the high level of health expenditures. Health economists have determined that these high expenditures are particularly due to the high price of health goods and services, rather than the volume of health care. Explanations for the poorer than expected health outcomes include the greater heterogeneity in income, race and immigrant status of the U.S. population compared to many countries, the lack of universal health insurance cover, and a higher rate of teenage pregnancy which increases infant mortality rates. Many studies have looked at this issue.

## 11.4 Health Expenditures

If health is a normal good, then health expenditures rise with income. And if health is a superior good, then health expenditures rise more than proportionately with income, so a 1% increase in income is associated with a more than 1% increase in health expenditures.

Here we analyze the relationship between health expenditures and income using aggregate data for a number of developed nations. Similar analysis can also be done using individual level data, if this is available.

Table 11.4: Health data: Variable definitions and summary statistics

| Variable | Definition | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *Gdppc* | GDP per capita (in US $) | 33054 | 12918 | 13807 | 82901 |
| *Hlthpc* | Health expenditure per capita (in US $) | 3256 | 1494 | 923 | 7990 |

### 11.4.1 Application

The analysis uses the dataset HEALTH2009 introduced in Chapter 11.3. This has 2009 data for the 34 wealthy and relatively wealthy nations in the Organization of Economic and Community Development (OECD).

Table 11.4 provides some summary statistics. Health expenditure is measured per capita, and income is measured using GDP per capita. There is considerable variation in GDP per capita, measured in current US dollars at current exchange rates, ranging from $13,807 for Mexico to $82,901 for Luxembourg, a small European country with population of half a million.

### 11.4.2 Linear Regression

If health expenditure is a normal good, then there is a positive relationship between per capita health expenditures (*Hlthpc)* and per capita GDP *(Gdppc).*

OLS regression yields

$$Hlthpc = \underset{(0.63)}{285} + \underset{(6.99)}{0.0899} \times Gdppc, \quad R^2 = 0.604, \quad s_e = 954, \quad n = 34,$$

where $t$-statistics based on default standard errors are given in parentheses.

The slope coefficient estimate implies that an extra $1,000 in GDP per capita is associated with an $89.90 increase in per capita health expenditures. The relationship is highly statistically significant, as $t = 6.99$. Since the prior belief is that $\beta_2 > 0$ in this example, it is most appropriate to perform a one-sided test of $H_0 : \beta_2 \leq 0$ against $H_a : \beta_2 > 0$. Then $c = t_{.05;32} = 1.69$ and we reject $H_0$ at significance level 0.05 since $t = 6.99 > c$.

The first panel of Figure 11.4 plots the data and regression line. There are two notable outlying observations. First the U.S. spends almost two times as much on health given its per capita GDP as predicted by the OLS regression. Second, Luxembourg has unusually high per capita GDP while its health spending per capita is similar to that of the other wealthiest European countries.

These two observations are so far from the line that they warrant further investigation, but a check reveals that the data have not been entered incorrectly. It is tempting to drop these two outliers from the analysis, but the econometrics practice is to retain all observations

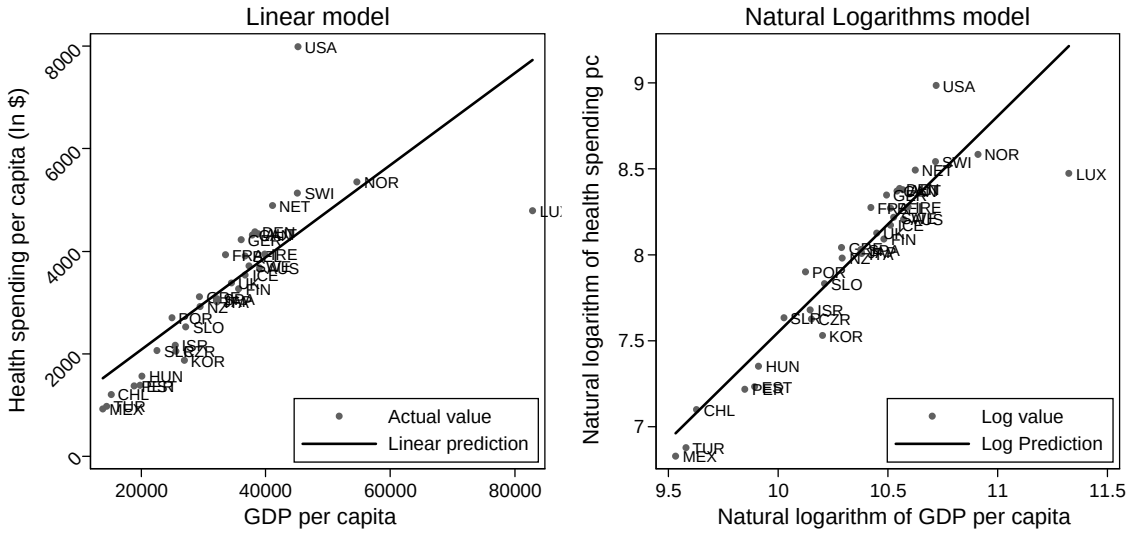


Figure 11.4: Health Expenditures: Relationship with GDP per capita

unless the data are clearly miscoded. We prefer that the analysis be of all OECD countries rather than an analysis of OECD countries excluding the U.S. and Luxembourg.

### 11.4.3 Robust Standard Errors

With cross-section data, model errors may be heteroskedastic and it is now common econometric practice to use heteroskedastic-robust standard errors. For the original linear regressor the heteroskedastic-robust standard error for the slope coefficient increases substantially from 0.0129 to .0293, so the $t$-statistic falls considerably from 6.99 to 3.08.

This substantial difference between default and heteroskedastic robust standard errors is very unusual. It arises because for Luxembourg and the USA the residuals $e_i$ are very large, see the first panel of Figure 11.4, so $e_i^2(x_i - \bar{x})^2$ is large and hence $se_{het}(b_2) = \sqrt{\sum_{i=1}^n e_i^2(x_i - \bar{x})^2} / \sum_{i=1}^n (x_i - \bar{x})^2$ is large. Dropping these two countries the slope coefficient is now 0.1267, the default standard error is 0.0064 and the heteroskedastic-robust standard error is 0.0076, only modestly higher.

### 11.4.4 Log-Log Model

This subsection estimates a log-log model, introduced in Chapter 12.

The OLS regression confirms that health is a **normal good**, since $\beta_2 > 0$. A more

interesting question is whether health is a **superior good**, meaning that the income elasticity of health exceeds unity. On average the income elasticity of a good is unity, since if income doubles so too will consumption on average (ignoring savings). Here we ask whether health is a superior good, with income elasticity in excess of one.

This is important to policy makers as if health is a superior good, and GDP rises over time, then we expect that health expenditures as a fraction of GDP will rise over time (in the case of the U.S. from an already high 17.7% of GDP).

In Chapter 12.6, it is shown that this elasticity can be obtained from OLS regression where the dependent variable and regressor are first transformed to natural logarithm (a **log-log model**). Specifically, if $\widehat{\ln y} = b_1 + b_2 \ln x$, then $b_2$ measures the elasticity of $\widehat{y}$ with respect to $x$.

OLS regression in natural logarithms yields

$$\ln(Hlthpc) = \underset{(-5.27)}{-5.01} + \underset{(13.66)}{1.256} \times \ln(Gdppc), \quad R^2 = 0.853, \quad s_e = 0.199, \quad n = 34,$$

where $t$-statistics based on default standard errors are given in parentheses. The estimated income elasticity of health spending is 1.256, so a 10% increase in GDP per capita is associated with a 12.56% increase in health spending per capita.

A superior good is one for which the income elasticity exceeds one. So we test $H_0 : \beta_2 \leq 1$ against $H_a : \beta_2 > 1$. From output not give $se(b_2) = .0920$, so $t = (1.256 - 1)/.0920 = 2.78$. We reject $H_0$ at level 0.05 since $p = \Pr[T_{32} > 2.68] = 0.006 < 0.05$, or since $c = t_{.05,32} = 1.69$ and $t = 2.78 > c$. The data support the view that health is a superior good.

As an aside, note that $se(b_2)$ can actually be computed from the output given above. Since the reported $t$-statistic is for test of $\beta_2 = 0$, $t = b_2/se(b_2)$. It follows that $se(b_2) = t/b_2 = 1.256/13.66 = 0.092$.

The data, in natural logarithms, and the regression line are illustrated in the second panel of Figure 11.4. The model fit appears to be good, as already known since $R^2 = 0.853$.

Comparing the first and second panels of Figure 11.4, the residuals for the U.S. and Luxemburg are not as large (relatively) in the log-log model. As a result, in the log-log model the heteroskedastic-robust standard error of 0.0117 is much closer to the default standard error of 0.0097 than was the case following linear regression.

## 11.5 Exercises

To come. Repeat Okun with quarterly data or data from another country.

# Chapter 12

# © A. Colin Cameron: Bivariate Data Transformation

The standard linear regression model has the attraction of simplicity. The slope coefficient is readily interpretable as it gives the change in the dependent variable as the regressor changes by one unit.

In this chapter we consider models where it is still easy to run a bivariate OLS regression, but where interpretation of the results is more difficult because the dependent variable and/or the regressor have been transformed before regression.

The chapter begins with OLS regression on an indicator variable that takes just the two values 0 or 1. In that case the slope coefficient can be interpreted as measuring the difference in means of the two categories formed by the indicator variable.

The remainder of the chapter considers nonlinear models that permit the relationship between $y$ and $x$ to be a curve, rather than a straight line. There are many possible nonlinear relationships. Here we just consider those that transform $y$ and/or $x$ in such a way that a bivariate linear regression model can still be estimated, albeit one with transformed variables. Most often the transformation used is the natural logarithm.

## 12.1 Example: Earnings and Age

The data in dataset EARNINGSAGE are on 494 full-time male workers in 2010 aged 25 to 65 years. Summary statistics for the data are given in Table 12.1. Mean earnings are \$63,476 per year and mean age is 42.84 years. Table 12.2 additionally includes transformations of variables *Earnings* and *Age* that are used in this chapter.

By bivariate least squares regression of *Earnings* ($y$) on *Age* ($x$) yields fitted model

$$\widehat{y} = \underset{(3.21)}{36609} + \underset{(2.43)}{627.2}x, \quad R^2 = .012,$$

Table 12.1: Annual Earnings of male full-time workers aged 25-65 in 2010: Variable definitions and summary statistics.

| **Variable** | **Definition** | **Mean** | **Standard Deviation** | **Min** | **Max** |
|---|---|---|---|---|---|
| *Earnings* | Annual earnings in $ | 63476 | 61713 | 5000 | 504000 |
| *Age* | Age in years | 42.84 | 10.72 | 25 | 65 |
| *Dage* | = 1 if *Age* >= 50 years | 0.30 | 0.46 | 0 | 1 |
| *Lnearnings* | Natural logarithm of *Earnings* | 10.78 | 0.72 | 8.52 | 13.13 |
| *Lnage* | Natural logarithm of *Age* | 3.72 | 0.26 | 3.22 | 4.17 |
| n | 494 | | | | |

where $t$-statistics based on default standard errors are given in parentheses. (The $t$-statistic based on heteroskedastic-robust standard errors is 2.55 for the slope coefficient). The linear fitted model predicts that earnings increase with age at all ages by the same $627 per year of aging.

The first panel of Figure 12.1 presents a scatter plot of *Earnings* against *Age*, along with the fitted OLS line and a nonparametric regression line fitted by Lowess; see Chapter 8.11. For readability the scatter plot only plots observations with earnings less than $150,000, but OLS and Lowess were estimated using all the data. The fitted Lowess curve suggests a nonlinear relationship between *Earnings* and *Age*.

## 12.2 Indicator Variable Model

Regression on an intercept and an indicator variable provides an estimate of the difference in means of the dependent variable across the two categories defined by the indicator variable.

### 12.2.1 Indicator Variable Regression

An **indicator variable** or **dummy variable** or **categorical variable** is a variable that takes just two values, for simplicity 0 and 1. For example an indicator variable may take value 1 if a person is female and 0 if a person is male.

Linear regression of $y$ on an intercept and the indicator variable $d$ yields fitted values

$$\widehat{y} = b_1 + b_2 d.$$

Then

$$\widehat{y} = \begin{cases} b_1 + b_2 & \text{if } d = 1 \\ b_1 & \text{if } d = 0. \end{cases}$$

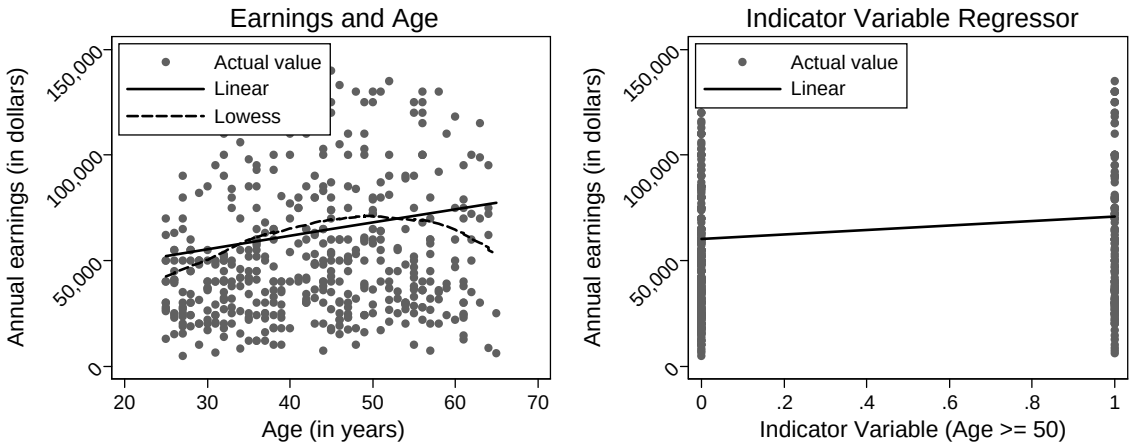


Figure 12.1: Earnings OLS Regressions: Age as regressor and indicator variable as regressor.

It follows that the marginal effect of changing $d$ from 0 to 1, a one unit increase in $d$, is to change $\widehat{y}$ by $b_2$. Thus the slope coefficient measures the difference between fitted earnings when $d = 1$ and fitted earnings when $d = 0$.

Let $\bar{y}_1$ denote the sample average of $y$ for observations with $d = 1$ and let $\bar{y}_0$ denote the sample average of $y$ for observations with $d = 0$. Then it can be shown that the OLS estimated coefficients are

$$b_1 = \bar{y}_0 \text{ and } b_2 = \bar{y}_1 - \bar{y}_0.$$

Furthermore, a test for statistical significance of the indicator variable provides a test of the difference in means.

**Remark 100** *The slope coefficient from OLS regression of $y$ on an intercept and an indicator variable $d$ equals the difference in the mean of $y$ across the two categories defined by $d$.*

Economists view tests of the difference in means as just a special case of OLS regression. Then a test of the difference in means is implemented as a $t$ test of $H_0 : \beta_2 = 0$ after OLS regression. In several other areas of applied statistics, especially ones that do not extensively use regression methods, the methods presented in Chapter 7.3 are instead used. The two approaches give slightly different answers due to different calculation of standard errors.

### 12.2.2 Example: Earnings and Age

Consider the earnings data and create an indicator variable for whether or not a person is more than fifty years old. Then

$$d = \begin{cases} 1 & \text{if } Age \geq 50 \\ 0 & \text{if } Age < 50 \end{cases}$$

The resulting variable *Dage* equals one for 29.5% of the sample and equals zero for 70.5% of the sample.

The second panel of Figure 12.1 presents a scatter plot of *Earnings* against *Dage*, along with the fitted OLS line. For readability the scatter plot only plots observations with earnings less than \$150,000, but the OLS regression was fitted using all the data.

OLS regression of *Earnings* on an intercept and *Dage* yields fitted value

$$\widehat{y} = \underset{(18.29)}{60383} + \underset{(1.72)}{10467} \times d, \quad R^2 = .006,$$

where $t$-statistics based on default standard errors are given in parentheses. So the marginal effect of being 50 or older, compared to being less than 50, is for annual earnings to be \$10,467 higher. This difference is not statistically significant using a two-sided test at level 0.05 since $|t| = 1.72 < c = t_{.025;492} = 1.965$.

Next consider using the difference in means test of Chapter 7.3. For the 348 individuals aged less than 50 years (so $d = 0$) the sample mean of *Earnings* is 60383 and the standard deviation is 57085. For the 147 individuals aged 50 years or more (so $d = 1$) the mean and standard deviation are, respectively, 70850 and 71225. Mean earnings when $d = 0$ equals the OLS intercept $b_1$. The difference in the two means, $70850 - 60383 = 10467$ equals the OLS slope coefficient $b_2$. The standard deviation of this difference is $\sqrt{(57084^2/348) + (71225^2/146)} = 6641$. So $t = 10467/6641 = 1.58$ compared to 1.72 using OLS regression with default standard errors (or $t = 1.58$ using heteroskedastic standard errors). The methods lead to the same estimate of the mean difference, with somewhat different standard errors and $t$-statistics.

## 12.3 Nonlinear Relationships

The most commonly-used nonlinear relationships in the bivariate case involve variables transformed to natural logarithms.

### 12.3.1 Examples of Nonlinear Models

Figure 12.2 presents scatter plots of $y$ against $x$ for generated data from three nonlinear models that involve the natural logarithm transformation, as well as fitted lines obtained

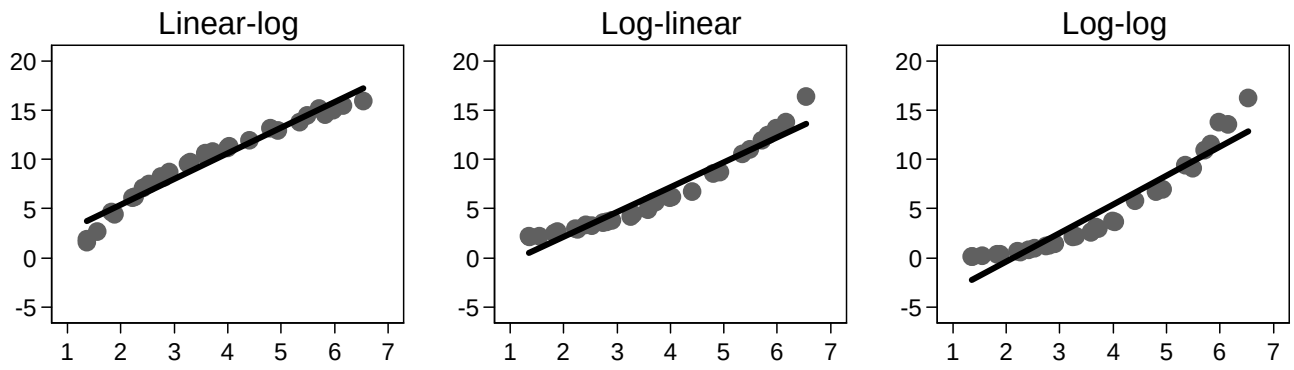


Figure 12.2: Nonlinear model examples

from least squares regression of $y$ on $x$. The scatter plots approximate the true relationship closely, as the data are generated with a relatively small error term. In all cases $y$ increases as $x$ increases, though for the nonlinear relationships this increase is not at a constant rate.

The first panel of Figure 12.2 is generated by the model $y = -1 + 9 \ln x + u$, where $u \sim N(0, 0.2^2)$. This model with $\ln x$ as a regressor is an example of a **linear-log model**. In this case the increase in $y$ as $x$ increases is larger at lower levels of $x$ than it is at higher levels of $x$, and the regression line overpredicts $y$ for the low and high values of $x$ and underpredicts $y$ for the middle values of $x$.

The second panel of Figure 12.2 is generated by the model $\ln y = 0.2 + 0.4x + u$, where $u \sim N(0, 0.04^2)$. This model with $\ln y$ as the dependent variable is an example of a **log-linear model**. Given $\ln y$, the variable $y = \exp(\ln y)$ is computed, plotted and regressed against $x$. It is clear that the increase in $y$ as $x$ increases is smaller at lower levels of $x$ than it is at higher levels of $x$. Also the regression line underpredicts $y$ for the low and high values of $x$ and overpredicts $y$ for the middle values of $x$.

The third panel of Figure 12.2 is generated by the model $\ln y = 0.06 + 3 \ln x + u$, where $u \sim N(0, 0.06^2)$. This model with $\ln y$ as the dependent variable and $\ln x$ as regressor is an example of a **log-log model**. Given $\ln y$, the variable $y = \exp(\ln y)$ is computed, plotted and regressed against $x$. The pattern of nonlinearity is similar to that for the log-linear model.

The linear-log, log-linear and log-log models are estimated in later sections of this chapter.

### 12.3.2 Marginal Effects in a Nonlinear Model

In many studies interest lies in the effect on the dependent variable of a change in the regressor. This is called the marginal effect.

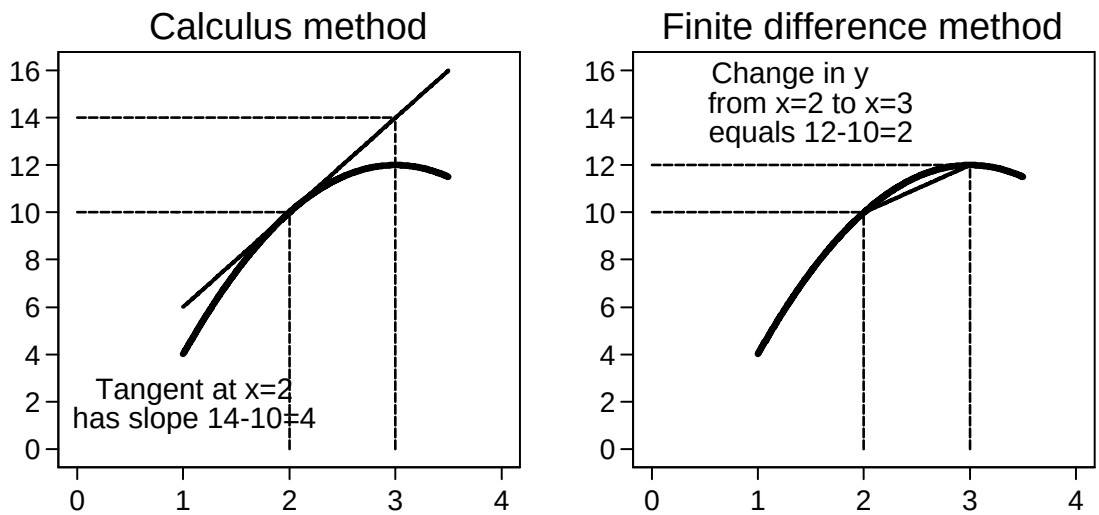


Figure 12.3: Marginal effect computation: Calculus method and finite difference method

**Remark 101** *The **marginal effect (ME)** on the predicted value of the dependent variable of a change in a regressor is defined as the ratio of the change in $\widehat{y}$ to the change in $x$, so ME= $\Delta\widehat{y}/\Delta x$.*

For the linear regression model the ME is simply the slope coefficient $b_2$.

For a nonlinear model, however, this is no longer the case. Instead, the marginal effect depends in part on $b_2$ but it also varies with the evaluation point $x$. Furthermore, there is more than one way to calculate the marginal effect at a given point $x$.

Figure 12.3 plots a nonlinear curve whose slope (the marginal effect) clearly varies with $x$. The two panels present two different ways to compute the marginal effect at $x = 2$.

The first panel of Figure 12.3 evaluates the marginal effect at $x = 2$ as being the **slope of the tangent** to the curve at $x = 2$. Since the tangent line passes through $y = 10$ at $x = 2$ and $y = 14$ at $x = 3$, the slope of the tangent equals $14 - 10 = 4$. This method is called the **calculus method** as the slope of the tangent is the **derivative** of the curve. It computes the ratio $\Delta y/\Delta x$ for a very small change in $x$ (the tangent to the curve) and then extrapolates this to a one unit change in $x$. For those familiar with calculus, see Appendix 12.A, the plotted curve $y = 12 - 2 \times (x - 3)^2$ has derivative $dy/dx = 12 - 4x$ which equals 4 at $x = 2$.

The second panel of Figure 12.3 evaluates the marginal effect at $x = 2$ as being the change in $y$ when $x$ increases by one unit, from $x = 2$ to $x = 3$. Since $y = 10$ at $x = 2$ and

$y = 12$ at $x = 3$, the marginal effect equals $12 - 10 = 2$. This method is called the **finite difference method**. It computes the ratio $\Delta y/\Delta x$ for a one unit change in $x$.

Economists most often use the first method, the calculus method. When the finite difference method is used it is most often used for an indicator variable regressor, such as a variable for gender.

**Remark 102** *Marginal effects at a given value of $x$ are usually computed exactly at $x$ (the **calculus method**). In some cases marginal effects at $x$ are instead computed for a discrete change from $x$ such as a one-unit change (the **finite difference method**).*

Using either method the marginal effect will change as $x$ changes. For example, using the finite difference the marginal effect at $x = 1$ equals 6, since $y = 4$ at $x = 1$ and $y = 10$ at $x = 2$, a change of $10 - 4 = 6$. By contrast the ME at $x = 2$ equalled 2.

There are three commonly-used methods to summarize the marginal effect on $y$ of a change in $x$.

1. Average marginal effect (AME): the average over the sample of the marginal effect for each individual (so evaluate the ME at each $x_i$ and then average).

2. Marginal effect at the mean (MEM): the marginal effect evaluated at the sample mean $\bar{x}$.

3. Marginal effect at a representative value (MER): the marginal effect evaluated at a representative value $x^*$.

For regressions involving transformations of earnings and age, $\bar{x} = 42.84$ and, for example, $x^* = 50$ if computing the marginal effect for a fifty-year old.

**Remark 103** *Three alternative measures of the marginal effect are the average marginal effect (AME), the marginal effect at the mean (MEM), and the marginal effect at a representative value (MER).*

### 12.3.3 Marginal Effect for the Natural Logarithm

This chapter emphasizes models involving the natural logarithm. From Chapter 4.1, the change in the natural logarithm of $x$ approximately equals the proportion change in $x$, for small changes in $x$. That is, $\Delta \ln x \simeq \Delta x/x$ for small $\Delta x$.

Rearranging this result $\Delta \ln x/\Delta x \simeq 1/x$ for small $\Delta x$.

**Remark 104** *If $y = \ln x$ then the marginal effect on $y$ of a change in $x$ is $1/x$ using calculus methods. That is, $\Delta \ln x/\Delta x \simeq 1/x$ for small $\Delta x$.*

### 12.3.4 Nonlinear Models in Practice

Several issues that arise when the relationship is nonlinear, in addition to computation of marginal effects, are studied in the remainder of this chapter.

The first issue is how to estimate a nonlinear relationship. In all the Figure 12.2 examples the coefficients in the model still appear linearly – it is only the variables $x$ and/or $y$ that appear nonlinearly. In that case it is possible to estimate the nonlinear relationship using standard least squares regression, after appropriate transformation of the variables. For example, for the linear-log model in the second panel we have $\mathrm{E}[y|x] = \beta_1 + \beta_2 \ln x$ which can be estimated by least squares regression of $y$ on $\ln x$, rather than $y$ on $x$.

For some other nonlinear models, however, such as $\mathrm{E}[y|x] = e^{\beta_1} e^{\beta_2 x}$ standard least squares estimation is not possible as the parameters do not appear linearly. Then alternative more advanced estimation methods are needed; see Appendix 19.B.

A second issue is whether a direct interpretation of slope coefficients is possible. In some cases it is.

A third issue is prediction of $y$ when $y$ is transformed before regression. For example, if the model is $\ln y = \beta_1 + \beta_2 x + u$, then how do we retransform to predict the level of $y$?

Finally, how do we choose the appropriate nonlinear model? For a bivariate model, the focus of this chapter, we can plot $y$ against $x$ and get some idea of the relationship. This is more difficult for a multivariate model that includes additional regressors. In many applications economic theory and the experiences of other studies that use similar data can provide a good guide.

## 12.4 Transformation of the Regressor

Consider the following bivariate model with transformation of the regressor $x$, so

$$y = \beta_1 + \beta_2 g(x) + u$$

where $g(x)$ is a specified function of $x$, such as $\ln x$, and $y$ is not transformed.

This model can be estimated by least squares regression of $y$ on $g(x)$, and the statistical significance of $x$ can be determined by $t$-test of whether $\beta_2 = 0$. The challenge is to measure the effect on $y$ of a change in $x$, rather than a change in $g(x)$.

### 12.4.1 Linear-Log Model

The leading bivariate example is the **linear-log model**, also called the **natural logarithm model**. Then the fitted model is

$$\widehat{y} = b_1 + b_2 \ln x,$$

where $\ln x$ is the natural logarithm of $x$ defined in Chapter 4.1.This model is appropriate only for positive $x$, since $\ln x$ is defined only if $x > 0$.

For this model the marginal effect of a change in $x$ is usually calculated using calculus methods since the linear-log model is typically used when $x$ is a continuous variable. For example in a regression explaining hours worker per year it is common to include the natural logarithm of wage as a regressor.

**Remark 105** *The marginal effect in the linear-log model equals $b_2/x$ using calculus methods.*

This result is obtained using the result from Chapter 12.3 that $\Delta \ln x/\Delta x = 1/x$ for small $\Delta x$. Since $\widehat{y} = b_1 + b_2 \ln x$ multiplies $\ln x$ by $b_2$ it follows that $\Delta\widehat{y}/\Delta x = b_2/x$. Note that the marginal effect varies with the point of evaluation $x$.

### 12.4.2 Linear-Log Example: Earnings and Age

For the earnings data the fitted linear-log model is

$$\widehat{y} = \underset{(1.10)}{-43590} + \underset{(2.70)}{28745} \ln x, \quad R^2 = .015,$$

where $t$-statistics based on default standard errors are in parentheses.

The marginal effect of aging is then $28745/x$. The marginal effect is always positive, and declines from \$1,150 at age 25 years to \$422 at age 65 years. For this model the AME= $\frac{1}{n}\sum_{i=1}^{n} 28745/x_i = \$717$, while the MEM= $28745/\bar{x} = 28745/42.84 = \$671$.

By contrast the ME in the linear model is \$627 and does not vary with age. The log-linear model fits the data slightly better than the linear model as $R^2$ has increased from 0.012 to 0.015.

## 12.5 Transformation of the Dependent Variable

Consider the following bivariate model with transformation of the dependent variable, so

$$g(y) = \beta_1 + \beta_2 x + u,$$

where $g(y)$ is a specified function of $y$, and to begin with $x$ is not transformed. The most common examples are $g(y) = \ln y$ and, for time series data, the changes $g(y) = y_t - y_{t-1}$ and $g(y) = \ln y_t - \ln y_{t-1}$.

The model can be estimated by least squares regression of $g(y)$ on $x$, and the statistical significance of $x$ can be determined by $t$-test of whether $\beta_2 = 0$. The challenge is to measure the effect on $y$, rather than on $g(y)$, of a change in $x$. Transforming the dependent variable also complicates post-regression prediction of $y$ and the use of $R^2$.

### 12.5.1 Log-linear Model

Taking the natural logarithm can eliminate right-skewness in data such as individual income, earnings or prices. This is illustrated for earnings data in Figure 4.3 in Chapter 4.1. For this reason, and because the resulting coefficients can be interpreted as semi-elasticities (see Chapter 12.6) it is common to take the natural logarithm.

Then $g(y) = \ln y$ and the fitted **log-linear model** is

$$\widehat{\ln y} = b_1 + b_2 x,$$

where $\ln y$ is the natural logarithm of $y$. This model can be used only if $y > 0$ since only then is $\ln y$ defined.

**Remark 106** *The marginal effect in the linear-log model equals $b_2\widehat{y}$ using calculus methods and equals $\widehat{y} \times [\exp(b_2) - 1]$ for a one-unit change using the finite difference method.*

To obtain this result note that for this OLS regression the slope $b_2$ gives the marginal effect on $\ln y$ of a change in $x$, that is $\Delta \ln y/\Delta x = b_2$. Now $\Delta \ln y = \Delta y/y$ for small $\Delta y$ from Chapter 4.1. Making this substitution, $(\Delta y/y)/\Delta x = b_2$. Rearranging yields $\Delta y/\Delta x = b_2 y$. The finite difference result is left as an exercise. Note that the marginal effect varies with the point of evaluation $x$ since $\widehat{y}$ varies with $x$.

The log-linear model is sometimes called the **exponential** model, since taking the exponential of both sides leads to $\widehat{y} = \exp(b_1 + b_2 x)$. In fact, however, as explained in the next subsection it is better to use

$$\widehat{y} = \exp(s_e^2/2)\exp(b_1 + b_2 x),$$

where $s_e$ is the standard error of the residual in the log-linear regression.

### 12.5.2 Retransformation Bias

For the log-linear model the population model is

$$\ln y = \beta_1 + \beta_2 x + u.$$

In this model the conditional mean of $\ln y$ given $x$ is $\beta_1 + \beta_2 x$. However, interest lies in predicting the conditional mean of $y$ given $x$, not $\ln y$ given $x$.

The log-linear model implies $y = \exp(\beta_1 + \beta_2 x + u)$, so the conditional mean of $y$ given $x$

$$\mathrm{E}[y|x] = \mathrm{E}[\exp(\beta_1 + \beta_2 x + u)|x] = \exp(\beta_1 + \beta_2 x) \times \mathrm{E}[\exp(u)|x].$$

The problem is that the condition $\mathrm{E}[u|x] = 0$, required for consistent estimation of the log-linear model, does not imply that $\mathrm{E}[\exp(u)|x] = 1$. As a result $\mathrm{E}[y|x] \neq \exp(\beta_1 + \beta_2 x)$, so it is wrong to use the obvious retransformed prediction $\widehat{y} = \exp(b_1 + b_2 x)$.

In the special case that errors are normally distributed and homoskedastic, so $u|x \sim N(0, \sigma_u^2)$, it can be shown with some difficulty that $\mathrm{E}[\exp(u)|x] = \exp(\sigma_u^2/2)$. It follows that in this special case

$$\mathrm{E}[y|x] = \exp(\sigma_u^2/2)\exp(\beta_1 + \beta_2 x).$$

Then $\mathrm{E}[y|x]$ is predicted using

$$\widehat{y} = \exp(s_e^2/2)\exp(b_1 + b_2 x),$$

where $s_e$ is the standard error of the log-linear regression.

For example, suppose $s_e = 0.4$. This means that the standard error of the residual is 0.4 on a log scale, which is quite large. Then $\exp(s_e^2/2) = \exp(0.08) = 1.083$ and we multiply $\exp(b_1 + b_2 x)$ by 1.083.

In general regression of $g(y)$ on $x$ with fitted value $\widehat{g(y)} = b_1 + b_2 x$ leads to unbiased prediction of $g(y)$ but biased (and inconsistent) prediction of $y$. For a model with $\ln y$ as the dependent variable and with normally distributed errors it is possible to provide a simple formula for a correct prediction. For other models, however, this is usually not the case.

**Remark 107** ***Retransformation bias*** *arises whenever a model for $g(y)$ is estimated but prediction of $y$ is desired. For a log-linear model the obvious prediction* $\exp(b_1 + b_2 x)$ *is multiplied by* $\exp(s_e^2/2)$ *where $s_e$ is the standard error of the log-linear regression.*

### 12.5.3 Log-Linear Example: Earnings and Age

For the earnings data the fitted log-linear model is

$$\widehat{\ln y} = \underset{(78.08)}{10.38} + \underset{(3.08)}{0.0093}x, \quad R^2 = .019, \quad s_e = 0.717,$$

where $t$-statistics based on default standard errors are in parentheses. So the marginal effect of aging is $0.0093\widehat{y}$. The marginal effect is always positive, and increases with age since ME= $0.0093\widehat{y}$ and $\widehat{y}$ increases with age.

Here $s_e = 0.717$, and $\exp(0.717^2/2) = 1.293$. So we use as prediction $\widehat{y} = 1.293 \times \exp(10.38 + 0.0093x)$. This substantial correction by multiplication by 1.293 relies on the assumption of normally distributed errors and is therefore not perfect, but it is much better than making no correction.

In this sample the residuals from the log-linear model appear to be approximately normally distributed, from visual inspection of their histogram and kernel density estimate, and because the skewness statistic is 0.19, close to zero, and the kurtosis statistics is 3.81, close to 3. As a result the correction does a good job of predicting average earnings. For these data the sample mean of variable *Earnings* is \$63,476, the average of the in-sample

predictions without correction is \$48,120, a large under-estimate, while the average of the in-sample predictions with correction is \$62,215, close to \$63,476.

The marginal effect of aging is $0.0093\widehat{y}$. Using predictions with the retransformation bias correction, the marginal effect is always positive, and increases from \$486 at age 25 years to \$704 at age 65 years. The AME= $\frac{1}{n}\sum_{i=1}^{n} 0.00093\widehat{y}_i$ = \$576, while the MEM= $28745/\bar{x} = 28745/42.84$ = \$573. By contrast the ME in the linear model is \$627 and does not vary with age.

### 12.5.4 Log-Log Model

The **log-log model** applies the natural logarithm transformation to both $y$ and $x$, leading to fitted model

$$\widehat{\ln y} = b_1 + b_2 \ln x.$$

This model can be used only if both $y > 0$ and $x > 0$.

**Remark 108** *The marginal effect in the log-log model equals $b_2\widehat{y}/\widehat{x}$ using calculus methods.*

To obtain this result note that for this OLS regression the slope $b_2$ gives the marginal effect on $\ln y$ of a change in $\ln x$, that is $\Delta \ln y/\Delta \ln x = b_2$. Now $\Delta \ln y = \Delta y/y$ for small $\Delta y$, and $\Delta \ln y = \Delta y/y$ for small $\Delta x$, from Chapter 4.1. Making these substitutions, $(\Delta y/y)/(\Delta x/x) = b_2$. Rearranging yields $\Delta y/\Delta x = b_2 y/x$.

The log-linear model is sometimes called the **power model** , since taking the exponential of both sides and manipulating leads to $\widehat{y} = e^{b_1}x^{b_2}$. In fact, however, we should additionally allow for retransformation bias and use

$$\widehat{y} = \exp(s_e^2/2)\exp(b_1 + b_2 \ln x),$$

where $s_e$ is the standard error of the residual in the log-log regression.

### 12.5.5 Log-Log Example: Earnings and Age

For the earnings data the fitted log-log model is

$$\widehat{\ln y} = \underset{(19.80)}{9.19} + \underset{(3.44)}{0.427} \ln x, \quad R^2 = .023, \quad s_e = 0.715,$$

where $t$-statistics based on default standard errors are in parentheses.

The residuals appear to be approximately normally distributed, as was the case for the log-linear model. Since $s_e = .715$, and $\exp(0.715^2/2) = 1.291$, we use the corrected prediction $\widehat{y} = 1.291 \times \exp(9.19 + 0.427 \ln x)$. For these data the sample mean of variable *Earnings* is \$63,476, the average of the in-sample predictions without correction is \$48,174, a large

under-estimate, while the average of the in-sample predictions with correction is \$62,211, close to \$63,476.

The marginal effect of aging is $0.427\widehat{y}/x$. For these data the marginal effect is always positive, so increases with age. Using predictions with the retransformation bias correction, the marginal effect is always positive, and decreases from \$851 at age 25 years to \$493 at age 65 years. The AME equals \$644, while the MEM equals \$625. By contrast the ME in the linear model is \$627 and does not vary with age.

### 12.5.6 R-squared with Transformed Dependent Variable

Recall that $R^2$ in regression of $y$ on $x$ measures the fraction of the variation in $y_i$, around the mean $\bar{y}$, that is explained by the regressor. By contrast $R^2$ in regression of $g(y)$ on $x$ explains the fraction of the variation in $g(y_i)$, around its mean $\overline{g(y)}$, that is explained by the regressor. These two measures are not comparable.

**Remark 109** *It is meaningless to use $R^2$ to compare the fit of models with different transformations of $y$ as the dependent variable.*

For data $y$ that are right-skewed, the model with $\ln y$ as dependent variable generally has higher $R^2$ than the model with $y$ as dependent variable, since the logarithmic transformation brings in large outlying values that can lead to poor model fit.

For persistent time series data such as GDP data there is much less variation in $y_t$ around $\bar{y}$ than there is variation in the one-period change $\Delta y_t = y_t - y_{t-1}$ around its mean. So $R^2$ in a model in regression with $\Delta y_t$ as the dependent variable can be much lower than $R^2$ in a model in regression with $y_t$ as the dependent variable.

Recall that $R^2$ in the linear regression model equals the squared correlation coefficient between the fitted and actual values of $y$. More generally we can compare models with transformations of $y$ as the dependent variable on the basis of the correlation between $\widehat{y}$ and $y$, provided it is possible to compute predictions $\widehat{y}$ that correct for retransformation bias.

## 12.6 Elasticities and Semi-elasticities

A marginal effect can be interpreted as how many units does $y$ change by when $x$ changes by one unit. In some applications it can be more useful to consider proportionate changes. These are measured using elasticities, widely used in economics and seldom used in many other areas of applied statistics. The natural logarithm transformation greatly simplifies computation of elasticities.

For example, in modelling earnings it can be more natural to think that one more year of education or one more year of aging leads to a certain percentage increase in earnings,

rather than a certain dollar increase in earnings. As discussed below, this estimate is a semi-elasticity and can be directly estimated by OLS regression in a nonlinear model.

### 12.6.1 Definition of Elasticity

Economists often use elasticities, the ratio of proportionate changes.

**Remark 110** *The **elasticity of** $y$ **with respect to** $x$ is the ratio of the proportionate change in $y$ to the proportionate change in $x$ :*

$$Elasticity_{yx} = \frac{\Delta y/y}{\Delta x/x}.$$

*Multiplying both numerator and denominator by 100, the elasticity equals the percentage change in $y$ when $x$ changes by one percent.*

For example, if the price elasticity of demand for a good is $-2.0$, then a one percent increase in price leads to a 2.0 percent decrease in demand. And if the income elasticity of demand for a good is 0.4, then a one percent increase in income leads to a 0.4 percent increase in demand.

The elasticity can also be written as $(\Delta y/\Delta x) \times x/y$, so it equals the slope or marginal effect $\Delta y/\Delta x$ multiplied by $x/y$. For the fitted linear model $\widehat{y} = b_1 + b_2 x$ the slope $\Delta\widehat{y}/\Delta x = b_2$ so the elasticity equals $b_2 x/\widehat{y}$. This varies with $x$.

A related measure is a semi-elasticity.

**Remark 111** *The **semi-elasticity of** $y$ **with respect to** $x$ is the ratio of the proportionate change in $y$ to the level of change in $x$ :*

$$Semi\text{-}elasticity_{yx} = \frac{\Delta y/y}{\Delta x}.$$

For example, if the semi-elasticity of earnings with respect to years of schooling is 0.08, then one more year of schooling is associated with a 0.08 proportionate increase in earnings, or an 8% increase in earnings.

### 12.6.2 Models with Natural Logarithms

Elasticities and semi-elasticities can be directly obtained by OLS regression of models that first transform variables to natural logarithms.

From Chapter 4.1, $\Delta \ln y \simeq \Delta y/y$ and $\Delta \ln x \simeq \Delta x/x$ for small changes $\Delta y$ and $\Delta x$. It follows that the elasticity, the ratio of proportionate changes, equals the ratio $\Delta \ln y/\Delta \ln x$ using calculus methods.

Table 12.2: Summary for models involving natural logarithms

| **Model** | **Specification** | **Marginal Effect** | **Interpretation of $b_2$** | **Elasticity** |
|---|---|---|---|---|
| Linear | $\widehat{y} = b_1 + b_2 x$ | $b_2$ | Slope: $\Delta\widehat{y}/\Delta x$ | $b_2 x/\widehat{y}$ |
| Linear-log | $\widehat{y} = b_1 + b_2 \ln x$ | $b_2/x$ | Semi-elasticity: $\Delta\widehat{y}/(\Delta x/x)$ | $b_2/x$ |
| Log-Linear | $\widehat{\ln y} = b_1 + b_2 x$ | $b_2\widehat{y}$ | Semi-elasticity: $(\Delta\widehat{y}/\widehat{y})/\Delta x$ | $b_2/\widehat{y}$ |
| Log-log | $\widehat{\ln y} = b_1 + b_2 \ln x$ | $b_2\widehat{y}/x$ | Elasticity: $(\Delta\widehat{y}/\widehat{y})/(\Delta x/x)$ | $b_2$ |

Now consider regression of $\ln y$ on $\ln x$. The slope coefficient is then $\Delta \ln y/\Delta \ln x$, but this is just the elasticity. So the slope coefficient in the fitted model $\widehat{\ln y} = b_1 + b_2 \ln x$ directly yields the estimated elasticity. Furthermore this elasticity is constant, whereas the elasticity obtained from the linear model varies with the regressor value.

**Remark 112** *The elasticity of $y$ with respect to $x$ equals the ratio of changes in natural logarithms:*

$$\textit{Elasticity}_{yx} = \frac{\Delta y/y}{\Delta x/x} \simeq \frac{\Delta \ln y}{\Delta \ln x},$$

*The slope coefficient from OLS regression of $\ln y$ on $\ln x$ (log-log model) directly estimates this elasticity.*

If only one of $y$ or $x$ is transformed then a **semi-elasticity** is obtained.

**Remark 113** *The elasticity of $y$ with respect to $x$ equals the ratio of changes in natural logarithms:*

$$\textit{Elasticity}_{yx} = \frac{\Delta y/y}{\Delta x/x} \simeq \frac{\Delta \ln y}{\Delta x}.$$

*The slope coefficient from OLS regression of $\ln y$ on $x$ (log-linear model) directly estimates this semi-elasticity.*

For the linear-log model $\widehat{y} = b_1 + b_2 \ln x$, $b_2 = \Delta\widehat{y}/\Delta \ln x$ equals the number of units that $\widehat{y}$ changes by for a one unit proportionate change in $x$.

Table 12.2 provides a summary a summary for models where one or both of $y$ and $x$ are transformed by taking the natural logarithm. The final column gives the formula to compute the elasticity $(\Delta\widehat{y}/\widehat{y})/(\Delta x/x)$ for each of these models.

The results in Table 12.2 are based on calculus methods to compute the marginal effect. For the log-linear model with a regressor that is an indicator variable a discrete one-unit change might be considered instead. From Chapter 12.3, the marginal effect in the log-linear model is then $\Delta\widehat{y}/\Delta x = \widehat{y} \times [\exp(b_2) - 1]$. It follows that the semi-elasticity $(\Delta\widehat{y}/\widehat{y})/\Delta x =$

Table 12.3: Regression of Earnings (y) on Age (x) for various transformation models of the variables.

| **Model** | **Estimates** | $R^2$ | **AME** | **Min ME** | **Max ME** |
|---|---|---|---|---|---|
| Linear | $\widehat{y} = 36609 + 627.2x$ | 0.012 | 627 | 627 | 627 |
| Linear-log | $\widehat{y} = -43590 + 28745 \ln x$ | 0.015 | 717 | 442 | 1150 |
| Log-linear | $\widehat{\ln y} = 10.38 + 0.0093x$ | 0.019 | 576 | 486 | 704 |
| Log-log | $\widehat{\ln y} = 9.19 + 0.427 \ln x$ | 0.023 | 646 | 500 | 897 |
| Indicator: Linear | $\widehat{y} = 60383 + 10467 \times d$ | 0.006 | - | - | - |
| Indicator: Log-linear | $\widehat{\ln y} = 10.74 + 0.138 \times d$ | 0.008 | - | - | - |

$\exp(b_2) - 1$ when $\Delta x = 1$. Thus, for example, if we estimate $\widehat{\ln y} = 2.0 + 0.1d$ then when $d = 1$ rather than $d = 0$ the proportionate change in $\widehat{y}$ equals $\exp(0.1) - 1 = 0.105$, or a 10.5% increase.

### 12.6.3 Elasticity Example: Earnings and Age

From results given earlier for the earnings data, the fitted log-log model is $\widehat{\ln y} = 9.19 + 0.427 \ln x$, so a one percent increase in age is associated with a 0.427% increase in earnings.

The fitted log-linear model is $\widehat{\ln y} = 10.38 + 0.0093x$, so a one year increase in age is associated with a 0.0093 proportionate change, or a 0.93% change, in earnings. For this application this is perhaps the most natural metric to use – what is the percentage change in earnings when age increases by one year?

## 12.7 Comparison of Models

Table 12.3 summarizes key results for the various models discussed in this chapter applied to the earnings data. The indicator variable model uses as regressor the variable $d = 1$ if $x \geq 50$ and $d = 0$ if $x < 50$.

Standard errors are not provided but for all models the regressors were statistically significant at significance level 0.05. The economic significance of the coefficients is also substantial. For example, the linear model suggests that aging ten years is associated with an increase of \$6,270 in annual earnings ($= 10 \times b_2$).

The fit of all models is poor, with low $R^2$. $R^2$ can be compared across models with the same dependent variable and with the same number of regressors. So the linear-log model fits better than the linear model or indicator variable model. And the log-log model provides a better fit than the log-linear model.

Table 12.3 also reports average marginal effects for the models with *Age* as regressor, as well as the minimum and maximum marginal effect over the sample values of age (from 25 to 65), for the various models. For computing $\widehat{y}$ in models with $\ln y$ as the dependent variables, the correction factor $\exp(s_e^2/2)$ was used to correct for retransformation bias under the assumption of normally distributed errors. The AME ranges from \$576 to \$717 across the models. There is substantial variation in the range of values taken by the ME across observations for all but the linear model, a consequence of the other models being nonlinear.

Results for the two models with indicator variable *Dage* as regressor are interpreted as follows. For the linear model earnings are on average \$10,467 higher for those aged 50 years or more. For the log-linear model earnings are on average $100 \times (e^{0.138} - 1) = 14.8$ percent higher for those aged 50 years or more.

Which model with *Age* as regressor is best for these data? Since some models involve $\ln y$ rather than $y$ the models are compared on the basis of correlation between $\widehat{y}$ and $y$, where the predictions $\widehat{y}$ from models for $\ln y$ are made correcting for retransformation bias. Then the correlations of $\widehat{y}$ with $y$ for the linear, linear-log, log-linear and log-log models are, respectively, 0.109, 0.121, 0.104, and 0.116. On this basis the linear-log model is preferred, though there is not much difference between the models – all models with just the one regressor predict earnings poorly.

More generally economic theory and policy can guide in determining which model to use. For example, a production function of Cobb-Douglas form can be estimated using a log-log model; see Chapter 15.2. And for right-skewed data it is natural to have $\ln y$ as the dependent variable. For example, for earnings the log-linear model is often used, as then the least squares coefficients give the proportionate change in $y$ as $x$ changes by one unit.

## 12.8 Key Concepts

1. The slope coefficient from OLS regression of $y$ on an intercept and an indicator variable $d$ equals the difference in the mean of $y$ across the two categories defined by $d$.

2. The marginal effect (ME) on the predicted value of the dependent variable of a change in a regressor is defined as the ratio of the change in $\widehat{y}$ to the change in $x$, so ME= $\Delta\widehat{y}/\Delta x$.

3. In a linear model the ME equals the slope coefficient and does not vary with $x$.

4. In a nonlinear model the ME varies with $x$.

5. The finite difference method calculates the ME for a discrete change $\Delta x$ in $x$.

6. The calculus method calculates the ME for an infinitesimally small change $\Delta x$. It equals the ME from the finite difference method as $\Delta x \to 0$.

7. Three alternative measures of the marginal effect are the average marginal effect (AME), the marginal effect at the mean (MEM), and the marginal effect at a representative value (MER).

8. For $y = \ln x$ the marginal effect equals $1/x$ using calculus methods. That is, $\Delta \ln x/\Delta x \simeq 1/x$ for small $\Delta x$.

9. The marginal effect in the linear-log model equals $b_2/x$ using calculus methods.

10. The marginal effect in the log-linear model equals $b_2\widehat{y}$ using calculus methods, and equals $\widehat{y} \times [\exp(b_2) - 1]$ for a finite one-unit change.

11. The marginal effect in the log-log model equals $b_2\widehat{y}/x$ using calculus methods.

12. Models with transformed dependent variable have the complications of retransformation bias and noncomparability of R-squared.

13. The elasticity of $y$ with respect to $x$ equals $(\Delta y/y)/(\Delta x/x)$, the ratio of the proportionate change in $y$ to the proportionate change in $x$.

14. The semi-elasticity of $y$ with respect to $x$ equals $(\Delta y/y)/\Delta x$, the ratio of the proportionate change in $y$ to the level of change in $x$.

15. The log-log model gives a direct estimate of the elasticity and the log-linear model gives a direct estimate of the semi-elasticity.

16. Key Terms: Data transformation; indicator variable; dummy variable; difference in means; marginal effect; finite-difference method; one-unit change; calculus method; derivative; average marginal effect (AME); marginal effect at mean (MEM); marginal effect at representative value (MER); natural logarithm; linear-log model; log-linear model; retransformation bias; log-log model; elasticity; semi-elasticity.

## 12.9 Exercises

1. Suppose $\widehat{y} = \ln x$ and that $x$ changes by $\Delta x$. Use the properties of the natural logarithm given in Chapter 4.1 to answer the following.

   (a) Show that $\Delta\widehat{y} = \ln(x + \Delta x) - \ln x$.

   (b) Hence show that $\Delta\widehat{y} = \ln[(x + \Delta x)/x]$.

   (c) Hence show that $\Delta\widehat{y} = \ln[1 + \Delta x/x]$.

   (d) Hence show that $\Delta\widehat{y} \simeq \Delta x/x$ for small $\Delta x/x$.

   (e) Hence conclude that $\Delta\widehat{y}/\Delta x = 1/x$ as $\Delta x \to 0$.

## 12.10 Appendix 12.A: Calculus and Derivatives

**Differential calculus** is occasionally used in this book, notably in this chapter and in Chapter 16.

For those not familiar with calculus, it is sufficient to interpret the derivative $dy/dx$ as the slope $\Delta y/\Delta x$, the ratio of the change in $y$ to the change in $x$, for very small changes in $x$.

**Remark 114** *The* ***derivative*** $\frac{dy}{dx}$ *equals* $\frac{\Delta y}{\Delta x}$ *for very small changes* $\Delta x$ *(formally as* $\Delta x \to 0$*).*

For example, if $dy/dx = 2$ then, for a given very small change in $x$, $y$ changes by twice as much. For example, if $x$ changes by 0.1 the $y$ changes by approximately 0.2.

In some special cases it is straightforward to obtain the formula for the derivative. As an example, suppose $y = x^2$. If we change $x$ by $\Delta x$ to $x + \Delta x$ then $y$ changes from $x^2$ to $(x + \Delta x)^2 = x + 2\Delta x + (\Delta x)^2$. It follows that for $y = x^2$

$$\begin{aligned}
\frac{\Delta y}{\Delta x} &= \frac{[x^2 + 2x\Delta x + (\Delta x)^2] - [x^2]}{\Delta x} \\
&= \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \\
&= 2x + \Delta x \\
&\to 2x \text{ as } \Delta x \to 0.
\end{aligned}$$

We have shown that the derivative of $y = x^2$ with respect to $x$ equals $2x$.

Table 12.4 presents the derivatives of the functions used in this book. The derivatives for $e^x$ and $\ln x$ are obtained using more advanced methods than that just used to show that $dx^2/dx = 2x$.

Most often **linear relationships** of the form $y = a + bx$ are estimated in this book. Then $\Delta y = b\Delta x$, so $\Delta y/\Delta x = b$, regardless of the size of $\Delta x$, as does $dy/dx = b$. For linear relationships the derivative equals the (constant) slope of the line.

**Nonlinear relationships**, notably quadratic relationships and relationships involving the natural logarithm, are studied in Chapters 12 and 16. Then the derivative is the slope of the tangent to the curve. This varies with the value of $x$ that we evaluate the derivative at.

Furthermore, for nonlinear relationships the derivative does not exactly equal the finite difference $\Delta y/\Delta x$ unless $\Delta x \to 0$. For example, we have shown that $dx^2/dx = 2x$, while $\Delta x^2/\Delta x = 2x + \Delta x$ which equals $2x$ only if $\Delta x \to 0$.

Table 12.4: Derivatives: Leading Examples.

| Function | Derivative |
|---|---|
| $y = f(x)$ | $\frac{dy}{dx} = \frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x)-f(x)}{\Delta x}$ for very small $\Delta x$ |
| $y = a$ | $\frac{dy}{dx} = 0$ |
| $y = x$ | $\frac{dy}{dx} = 1$ |
| $y = x^2$ | $\frac{dy}{dx} = 2x$ |
| $y = x^n$ | $\frac{dy}{dx} = nx^{n-1}$ |
| $y = a + bx + cx^2$ | $\frac{dy}{dx} = b + 2cx$ |
| $y = e^x$ | $\frac{dy}{dx} = e^x$ |
| $y = \ln x$ | $\frac{dy}{dx} = \frac{1}{x}$ |

# Chapter 13

# © A. Colin Cameron: Data Summary with Multiple Regression

This chapter presents the leading statistical method used in analysis of economics data, multiple regression of one variable on an intercept and several other variables. For example, we consider how house sale price is related to various features of the house such as house size, number of bedrooms, lot size and age.

This chapter considers summary of the relationship between variables in the sample. Chapter 14 presents statistical inference for the population.

## 13.1 Example: House Price and Characteristics

We consider the same dataset HOUSE for 29 houses as that analyzed in Chapters 8-10, except now additional variables are included in the analysis.

Table 13.1 presents variable descriptions and summary statistics for these data on houses sold in a homogeneous community in a single year. A half bathroom is a lavatory without bath or shower. Precise data on lot size are unavailable; instead the lot size is coded as 1 for small, 2 for medium and 3 for large.

Interest lies in the extent to which house price is explained by the other variables. For this example the dataset is small, so we can make some preliminary conclusions by simply listing the data.

Table 13.1: House price data: Variable definitions and summary statistics

| Variable | Definition | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|
| *Price* | Sale Price in dollars | 253910 | 37391 | 204000 | 375000 |
| *Size* | House size in square feet | 1883 | 398 | 1400 | 3300 |
| *Bedrooms* | Number of bedrooms | 3.79 | 0.68 | 3 | 6 |
| *Bathrooms* | Number of bathrooms | 2.21 | 0.34 | 2 | 3 |
| *Lotsize* | Size of lot (1, 2 or 3) | 2.14 | 0.69 | 1 | 3 |
| *Age* | House age in years | 36.4 | 7.12 | 23 | 51 |
| *Month Sold* | Month of year house was sold | 5.97 | 1.68 | 3 | 8 |

Table 13.2 gives a listing, ordered by house price, and suggests that price increases with house size as already noted in Chapter 8. House price increases with number of bedrooms as, for example, the four highest priced houses have one more bedroom than the four lowest priced houses. There appears to be little relationship with number of bathrooms, lot size, age of house and month of the year in which the house was sold.

A more systematic analysis is needed, for several reasons. First, for many datasets the data listing is so large as to make interpretation impossible. Second, it is desirable to quantify the relationship between the different series. For example, what is the estimated change in house price associated with an increase in house size of one hundred square feet? Third, we wish to extrapolate from the sample to the population, in which case confidence intervals and hypothesis tests are used. All of these points have already been raised in the preceding chapters on bivariate regression.

Multiple regression departs from bivariate regression by controlling for several variables simultaneously. This allows isolation of the effect of one regressor after controlling for other regressors.

For example, consider the relationship between house price and both number of bedrooms and house size. Bivariate regression with *Bedrooms* as the only regressor yields fitted model

$$\widehat{Price} = 164138 + 23667 \times Bedrooms.$$

This suggests a large association between price and number of bedrooms, with an increase in price of \$23,667 per bedroom. If instead we perform multiple regression with both *Bedrooms* and *Size* as regressors, then

$$\widehat{Price} = 111691 + 1553 \times Bedrooms + 72.41 \times Size.$$

Now an extra bedroom is associated with only a \$1,553 increase in the price of the house. So bedrooms actually have relatively small association with house price, once house size is controlled for.

Table 13.2: House Price data: Complete listing.

| Price | Size | Bedrooms | Bathrooms | Lot Size | Age | Month Sold |
|---|---|---|---|---|---|---|
| 375,000 | 3,300 | 4 | 2.5 | 2 | 39 | 3 |
| 340,000 | 2,400 | 4 | 3 | 2 | 34 | 6 |
| 310,000 | 2,300 | 4 | 2.5 | 2 | 28 | 5 |
| 279,900 | 2,000 | 4 | 2 | 2 | 31 | 7 |
| 278,500 | 2,600 | 6 | 2 | 3 | 38 | 8 |
| 273,000 | 1,900 | 5 | 2 | 2 | 37 | 7 |
| 272,000 | 1,800 | 4 | 2.5 | 2 | 46 | 3 |
| 270,000 | 2,000 | 4 | 2.5 | 3 | 39 | 5 |
| 270,000 | 1,800 | 4 | 2 | 3 | 31 | 3 |
| 258,500 | 1,600 | 3 | 2 | 1 | 39 | 8 |
| 255,000 | 1,500 | 4 | 2 | 3 | 47 | 7 |
| 253,000 | 2,100 | 4 | 2 | 3 | 47 | 6 |
| 249,000 | 1,900 | 4 | 3 | 3 | 37 | 6 |
| 245,000 | 1,400 | 4 | 2 | 2 | 30 | 8 |
| 244,000 | 2,000 | 4 | 2 | 1 | 29 | 7 |
| 241,000 | 1,600 | 4 | 2 | 2 | 34 | 8 |
| 239,500 | 1,600 | 3 | 2 | 3 | 34 | 6 |
| 238,000 | 1,900 | 4 | 2 | 2 | 29 | 7 |
| 236,500 | 1,600 | 3 | 2 | 3 | 23 | 8 |
| 235,000 | 1,600 | 3 | 2 | 3 | 35 | 5 |
| 235,000 | 1,700 | 4 | 2 | 2 | 29 | 7 |
| 233,000 | 1,700 | 3 | 2 | 1 | 40 | 6 |
| 230,000 | 2,100 | 4 | 2 | 2 | 34 | 8 |
| 229,000 | 1,700 | 4 | 2.5 | 2 | 35 | 3 |
| 224,500 | 2,100 | 4 | 2.5 | 2 | 47 | 6 |
| 220,000 | 1,600 | 3 | 2 | 1 | 49 | 4 |
| 231,000 | 1,800 | 3 | 2 | 2 | 51 | 4 |
| 212,000 | 1,600 | 3 | 3 | 2 | 33 | 5 |
| 204,000 | 1,400 | 3 | 2 | 1 | 31 | 7 |

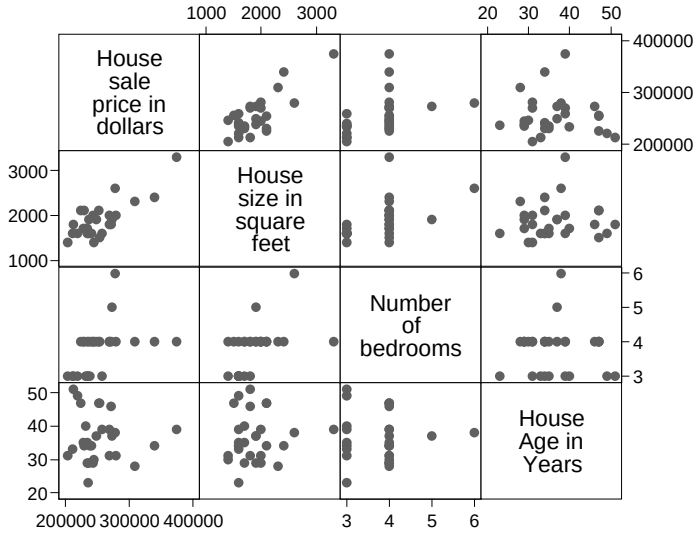


Figure 13.1: House price data: Two-way scatterplots

## 13.2 Two-way Scatterplots

Graphical methods for displaying the data are necessarily restricted to plots in two or three dimensions. In practice two-way scatter plots are used much more than three-dimensional graphs (or surface plots).

Figure 13.1 presents an **array of separate two-way scatter plots** for four of the variables in the house price dataset. The top row has house price on the vertical axis and shows that, respectively, price has little relationship with house age, a modest positive relationship with number of bedrooms, and a strong positive relationship with size. The second row has house size on the vertical axis and shows that size has, respectively, a strong positive relationship with house size (this is the reverse regression of the price-size regression given in the first row), little relationship with age and increasing relationship with number of bedrooms. And so on. The scatter plots also show no data points that are unusually large outliers.

In general for $k$ series there are potentially $k \times (k-1)$ separate plots, where for each pair of series there are two plots depending on which variable is on the vertical axis, and there is no point in plotting a series against itself. It can be simpler to read a variant of Figure 13.1 that provides only one-half of the possible scatter plots by not including scatter plots with the axes reversed.

Twoway scatter plots are especially useful for plotting related series where separate two-

way comparisons are of intrinsic interest. For example, this can be the case for related financial data series such as interest rates at various maturities. And the scatter plots may show unusual outlying data points, possibly due to data coding error.

## 13.3 Correlation

Pairwise correlations can be useful for exploratory data analysis.

### 13.3.1 Pairwise Correlation

From Chapter 8.4, the **sample correlation coefficient** between variables $y$ and $x$ is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $-1 \leq r_{xy} \leq 1$. The sign of $r_{xy}$ gives the direction of the association, strictly speaking the linear association, between $x$ and $y$, and the association is stronger the larger is $|r_{xy}|$.

In multivariate analysis several variables are considered. Then we can construct **pairwise correlations** that are correlations for each unique pair of variables. If there are $k$ variables then there are potentially $k^2$ pairwise correlations. Not all of the correlations need to be computed and reported, however, since $r_{xx} = 1$ and $r_{xy} = r_{yx}$ as the correlations are symmetric.

### 13.3.2 Example: House Price

Table 13.3 presents pairwise correlations for the data summarized in Table 13.1, with asterisks denoting that the sample correlation coefficient is statistically different from zero at significance level 0.05. The table lists correlations only in the lower half – the upper half is just the mirror image as $r_{yx} = r_{xy}$. For example the entry for row *Price* and column *Size* would be 0.79.

Since we ultimately wish to explain *Price*, the first column of Table 13.3 is of greatest relevance. The most highly correlated variable with *Price* is *Size* ($r = .79$), followed by *Bedrooms* ($r = .43$). From the second column, however, *Size* and *Bedrooms* are fairly highly correlated ($r = .52$) so, for example, *Bedrooms* may not add much more explanatory power than *Size* alone. Regression of *Price* on *Size* and *Bedrooms* permits determination of the independent effect on price of the number bedrooms, controlling for house size.

Table 13.3: House price data: correlations for 29 houses sold in Central Davis

| | *Price* | *Size* | *Lot Size* | *Bedrooms* | *Bathrooms* | *Age* | *Month Sold* |
|---|---|---|---|---|---|---|---|
| *Price* | 1 | | | | | | |
| *Size* | .79* | 1 | | | | | |
| *Bedrooms* | .43* | .52* | 1 | | | | |
| *Bathrooms* | .33 | .32 | .04 | 1 | | | |
| *Lot Size* | .15 | .11 | .29 | .10 | 1 | | |
| *Age* | -.07 | .08 | -.02 | -.03 | .04 | 1 | |
| *Month Sold* | -.21 | -.21 | -.06 | .18 | -.39* | -.37 | 1 |

## 13.4 Regression Line

The regression line for multiple regression is similar to that for bivariate regression, except additional variables are added. The least squares estimates again minimize a sum of squared residuals. The key difference is in the interpretation of the slope coefficients.

### 13.4.1 Regression Line

The **regression line** from regression of $y$ on several variables $x_2, ..., x_k$ is denoted

$$\widehat{y} = b_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k.$$

The terminology is the same as that for bivariate regression. The variable $y$ is called the dependent variable, $\widehat{y}$ is the predicted value or fitted value, $x_2, .., x_k$ are regressor (or independent or explanatory) variables or covariates, $b_1$ is the estimated y-axis intercept and $b_2, ..., b_k$ are the estimated slope parameters.

Note that the first regressor variable is denoted $x_2$ rather than $x_1$; implicitly $x_1 = 1$ for the intercept. There are $k$ regression coefficients, including the intercept, and hence $n - k$ degrees of freedom. Some other texts denote the first regressor after the intercept as $x_1$ and specify $\widehat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$. Then there are $k + 1$ regression parameters, and hence $n - k - 1$ degrees of freedom. In that case wherever $n - k$ appears below, these other texts would use $n - k - 1$.

The fitted model for the $i^{th}$ observation is written as

$$\widehat{y}_i = b_1 + b_2x_{2i} + b_3x_{3i} + \cdots + b_kx_{ki},$$

where the first subscript for regressors denotes the variable and the second subscript denotes the observation.

### 13.4.2 Least Squares Estimation

For multiple regression the **residual**, the difference between actual value of $y_i$ and fitted values, is

$$e_i = y_i - \widehat{y}_i = y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} - \cdots - b_k x_{ki}.$$

As for bivariate regression, **least squares regression** computes the regression coefficients, here $b_1, ..., b_k$, to make as small as possible the **residual sum of squares**,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - \cdots - b_k x_{ki})^2.$$

The **ordinary least squares (OLS) estimates** $b_1, ..., b_k$ are the solution to $k$ equations in $k$ unknowns. These equations, called the **normal equations** or **first-order conditions** are that

$$\sum_{i=1}^{n} x_{ji}(y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} - \cdots - b_k x_{ki}) = 0, \quad j = 1, ..., k,$$

where $x_{ji}$ is the $j^{th}$ regressor and $x_{1i} = 1$ for all $i$.

The sample size must exceed the number of regressors (including the intercept), since there are $k$ equations in $k$ unknowns. Additionally there must be enough **variation in the sample regressors** so that the regressors are not perfectly correlated with each other, or with linear combinations of each other. For example, if $x_3 = 2x_2$ then it is not possible to regress $y$ on $x_2$ and $x_3$ and disentangle the combined influence of $x_2$ and $x_3$.

**Remark 115** *The multiple regression line is $\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k$. The OLS estimates $b_1, b_2, ..., b_k$ are obtained by minimizing the sum of squared residuals where the residual $e = y - \widehat{y}$ is the difference between the actual and predicted value of $y$.*

Given the definition of the residual $e_i$, the normal equations can be rewritten as

$$\sum_{i=1}^{n} x_{ji} e_i = 0, \quad j = 1, ..., k.$$

So the sum over observations of the regressor times the residual equals zero. The mathematical terminology for this property is that **each regressor is orthogonal to the residual**.

The intercept term is $x_{1i} = 1$ for all $i$, so the preceding result with $j = 1$ implies that $\sum_{i=1}^{n} e_i = 0$. Thus the residuals sum to zero, and their in-sample average is zero. Since $e_i = y_i - \widehat{y}_i$, it follows that the sample mean of the fitted values equals the sample mean of the dependent variable.

**Remark 116** *A property of OLS regression is that residuals are orthogonal to the regressors, so that the sum over observations of the cross product of any regressor with the residuals equals zero. Furthermore, if an intercept is included in the regression then the residuals sum to zero and average zero. This in turn implies that the average of the in-sample fitted values is equal to the average of the sample values of the dependent variable.*

### 13.4.3 Least Squares Estimates

The complete formula for the OLS estimates from multiple regression cannot be expressed directly in summation notation, unlike bivariate regression.

The simplest representation is the following. Consider the coefficient $b_j$ of the $j^{th}$ regressor $x_j$. Let $\widetilde{x}_{ji}$ denote the residual from regressing $x_{ji}$ on an intercept and all regressors other than $x_{ji}$. Thus $\widetilde{x}_{ji} = x_{ji} - \widehat{x}_{ji}$ where $\widehat{x}_{ji}$ is obtained from the auxiliary multiple regression

$$\widehat{x}_{ji} = b_1 - b_2 x_{2i} - \cdots - b_{j-1} x_{j-1,i} - b_{j+1} x_{j+1,i} - \cdots - b_k x_{ki}.$$

Then it can be shown that $b_j$ can be calculated by bivariate regression of $y_i$ on the residual $\widetilde{x}_{ji}$.

Thus the coefficient $b_j$ of the $j^{th}$ regressor equals

$$b_j = \frac{\sum_{i=1}^n \widetilde{x}_{ji}(y_i - \bar{y})}{\sum_{i=1}^n \widetilde{x}_{ji}{}^2},$$

where $\bar{x}_j$ is the sample average for the $j^{th}$ regressor and we have $\overline{\widetilde{x}_j} = 0$, since $\widetilde{x}_{ji}$ is a residual and OLS residuals have zero average.

This representation of the OLS slope coefficient $b_j$ from multiple regression makes clear that $b_j$ measures the effect of $x_j$, after controlling for the other regressors in the regression.

**Remark 117** *The OLS coefficient $b_j$ can be calculated by bivariate regression of $y$ on $\widetilde{x}_j$, where $\widetilde{x}_j = x_j - \widehat{x}_j$ is the residual from regressing $x_j$ on an intercept and all regressors other than $x_j$.*

In more advanced courses that can assume matrix algebra it is simple to give the formula for the OLS estimates from multiple regression. For completeness the matrix formula is given in Appendix 14.A in the next chapter.

## 13.5 Estimated Partial Effects

Interest lies especially in the slope coefficients. Consider for simplicity $b_2$, the coefficient of regressor $x_2$. An important distinction is that between estimated partial effect and estimated total effect.

### 13.5.1 Estimated Partial Effect

Suppose $x_2$ is changed by $\Delta x_2$, while holding all other regressors at their current values. Then the predicted value of $y$ becomes

$$\begin{aligned}
\widehat{y}_{new} &= b_1 + b_2(x_2 + \Delta x_2) + b_3 x_3 + \cdots + b_k x_k \\
&= b_2 \Delta x_2 + b_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k \\
&= b_2 \Delta x_2 + \widehat{y}_{old}.
\end{aligned}$$

It follows from subtraction that $\Delta\widehat{y} = b_2\Delta x_2$, so that $\Delta\widehat{y}/\Delta x_2 = b_2$.

This result is the **estimated partial effect** of $x_2$ on $\widehat{y}$, and is expressed as

$$\left.\frac{\Delta\widehat{y}}{\Delta x_2}\right|_{x_3,\ldots,x_k} = b_2,$$

where the notation means that only $x_2$ is changed, while the other regressors $x_3, ..., x_k$ are held constant. The effect is also called the effect of $x_2$ **ceteris paribus**, where the Latin phrase "ceteris paribus" means all other things equal or held constant.

For convenience the simpler notation $\Delta\widehat{y}/\Delta x_2 = b_2$ may be used in place of the more formal notation $\Delta\widehat{y}/\Delta x_2|_{x_3,\ldots,x_k} = b_2$.

**Remark 118** *The slope coefficient $b_2$ in the multiple regression $\widehat{y} = b_1 + b_2x_2 + \cdots + b_kx_k$ gives the partial effect on the predicted value of $y$ when $x_2$ changes by one unit, holding the remaining regressors $x_3, ..., x_k$ constant.*

### 13.5.2 Estimated Total Effect

In the house price example, the partial effect corresponds to an experiment where one feature of the house is changed in isolation, while all other features are held constant.

In practice, however changes in one feature of the house may also be associated with changes in other features of the house. For example, an extra bedroom may be associated with a larger house – both features may effect the price of the house. The **total effect** on $\widehat{y}$ of a change in $x_2$ includes all the secondary effects through $x_2, ..., x_k$ also changing.

For simplicity, consider multiple regression on just two regressors $x_2$ and $x_3$, with fitted model $\widehat{y} = b_1 + b_2x_2 + b_3x_3$. Suppose changing $x_2$ by $\Delta x_2$ is associated with a change in $x_3$ of $\Delta x_3$. Then the total effect on $y$ of changing $x_2$ by $\Delta x_2$ equals $\Delta\widehat{y} = b_2\Delta x_2 + b_3\Delta x_3$. Dividing by $\Delta x_2$, it follows that the total effect of a change in $x_2$ on $\widehat{y}$ equals $\Delta\widehat{y}/\Delta x_2 = b_2 + b_3\Delta x_3/\Delta x_2$. As an aside for those familiar with advanced differential calculus, the partial effect is the partial derivative of $\widehat{y}$ with respect to $x_2$, denoted $\partial\widehat{y}/\partial x_2$. The total effect is the total derivative of $\widehat{y}$ with respect to $x_2$, denoted $d\widehat{y}/dx_2$.

This leads to the following more general result.

**Remark 119** *For multiple regression of $y$ on an intercept and $x_2, ..., x_k$, the estimated total effect of a change in $x_2$ allows for the secondary effect that the remaining regressors change when $x_2$ changes. The estimated total effect is*

$$\frac{\Delta\widehat{y}}{\Delta x_2} = b_2 + b_3\frac{\Delta x_3}{\Delta x_2} + \cdots + \frac{\Delta x_k}{\Delta x_2}.$$

### 13.5.3 Estimated Partial and Total Effects for OLS

There is a mechanical relationship between estimated partial and total effects when models are estimated by OLS regression.

The slope coefficient from bivariate regression of $x_3$ on $x_2$ is an estimate of $\Delta x_3/\Delta x_2$. And in general the slope coefficient from bivariate regression of $x_j$ on $x_2$ is an estimate of $\Delta x_j/\Delta x_2$.

Using these $k-2$ bivariate slope estimates for the $k-2$ changes $\Delta x_j/\Delta x_2$ in the preceding formula yields a value for the estimated total effect $\Delta\widehat{y}/\Delta x_2$ that, remarkably, exactly equals the slope coefficient from bivariate regression of $y_2$ on $x_2$ alone.

**Remark 120** *When regression is by OLS, the total effect on the predicted value of $y$ when $x_2$ changes by one unit from a multivariate regression simply equals the sloe coefficient from bivariate regression of $y$ on $x_2$ alone.*

### 13.5.4 Example: House Price

As an example, consider regression of house sale price on two regressors, number of bedrooms, size and age. This yields

$$\widehat{Price} = 1358323 + 913 \times Bedrooms + 73.89 \times Size - 673 \times Age.$$

The slope coefficient for *Bedrooms* gives the partial effect of adding a bedroom, while controlling for the size and age of the house. This *partial effect* is an increase in house price of \$913, a relatively small effect.

What is the total effect on house sale price of adding a bedroom? To compute this from the preceding multiple regression we additionally need to know the effect of changing bedrooms on the remaining regressors, *Size* and *Age*. Now bivariate regression yields $\widehat{Size} = 724+305.4\times Bedrooms$, so $\Delta Size/\Delta Bedrooms = 305.4$. And $\widehat{Age} = 37.46-0.2758\times Bedrooms$, so $\Delta Age/\Delta Bedrooms = -0.2758$.

Thus the *total effect* on house sale price of adding a bedroom

$$\frac{\Delta\widehat{Price}}{\Delta Bedrooms} = b_2+b_3\frac{\Delta Size}{\Delta Bedrooms}+b_4\frac{\Delta Size}{\Delta Age} = 913+73.89\times 305.4-673\times(-0.2758) = 23667.$$

The total effect is an increase in house price of \$23,667, much larger than the partial effect, controlling for size and age, of \$1,553.

Furthermore, if we simply regress price on bedrooms we obtain

$$\widehat{Price} = 164038 + 23667 \times Bedrooms.$$

Thus the slope coefficient from bivariate regression of *Price* on *Bedrooms* alone equals the total effect on *Price* of *Bedrooms* from multiple regression of *Price* on *Bedrooms*, *Size* and *Age*.

### 13.5.5 Partial versus Total Effects

Regression of price on bedrooms only and regression of price on bedrooms and size led to quite different slope coefficients for the bedrooms variable. In the former case the slope coefficient gives the total effect. In the latter case the slope coefficient gives the partial effect after controlling for size, and this effect is much smaller after size is controlled for.

In many regression applications it is the partial effect that we are interested in. For example, in looking at earnings difference by educational level, interest lies in the size of any earnings difference after controlling for individual characteristics such as age, gender and socioeconomic background.

### 13.5.6 Causation

OLS regression measures association between $y$ and the regressors $x_2, ..., x_n$, but not necessarily causation.

For example, regression of earnings on education and other regressors such as age and gender may find that higher earnings are associated with more education. But it is not necessarily the case that more education causes higher earnings. Instead it may be that the positive association is due to unobserved ability, not included as a regressor, that causes both higher earnings and more education.

As a second example, a regression of a measure of individual health status on income and other regressors such as age and gender may find that lower income is associated with poorer health. But it may be that it is poor health that leads to lower income due to limiting an individual's ability to work, rather than lower income that leads to poorer health.

For this reason, unless there is clear a priori information about the direction of causality one should merely say that in a fitted multiple regression model a one unit change in $x_j$ is **associated with** a $b_j$ change in $\widehat{y}$, holding all other regressors constant. Methods to determine causation are presented in Chapter 18.

**Remark 121** *In a fitted multiple regression model a one unit change in $x_j$ is associated with a $b_j$ change in $\widehat{y}$, holding all other regressors constant. The association need not imply causation.*

## 13.6 Model Fit

The bivariate regression measures of model fit, $s_e$ and $R^2$, extend naturally to the multiple regression setting. Additionally there are modifications to these measures that provide a penalty to models that include more variables in the regression.

### 13.6.1 Standard Error of the Regression

For multiple regression the **standard error of the regression** is

$$s_e = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}.$$

Now division is by $n-k$, rather than $n-2$ in the bivariate case, as $k$ degrees of freedom are lost since computation of $\widehat{y} = b_1 + b_2 x + \cdots + b_k x_k$ is based on the $k$ estimates $b_1, ...., b_k$.

Another name for $s_e$ is the **root mean squared error (MSE) of the residual**. It is also sometimes called the **standard error of the residual**.

Lower values of $s_e$ indicate better fit. As regressors are added to the model, $s_e$ can either rise or fall. Adding regressors cannot increase the sum of squared residuals, since if another regressor is added we can always leave the sum of squared residuals unchanged by giving the new regressor a coefficient of zero and leaving the coefficients of the other regressors unchanged. In practice the sum of squared residuals falls, so $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ falls and $s_e$ falls. At the same time $k$ is now higher, so the term $1/(1-k)$ is higher and $s_e$ rises.

**Remark 122** *The standard error of the regression measures the standard deviation of the residual, and hence the variability of the dependent variable around the regression line. As regressors are added to a model $s_e$ may decrease or may increase, though when $n$ is large it is more likely to increase.*

### 13.6.2 R-Squared

As in the bivariate case the total sum of squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$, denoted *TSS,* can be decomposed as

$$TSS = ExpSS + \ RSS,$$

the explained sum of squares ($ExpSS$) equals $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$ and the residual sum of squares ($RSS$) equals $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ .

The coefficient of determination, or **R-squared**, is defined exactly as in the bivariate case as

$$R^2 = \frac{ExpSS}{TSS} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

The only change is that $\widehat{y}_i$ is the prediction from a model with additional regressors. $R^2$ takes a maximum value of 1 if the residual sum of squares is zero, in which case the actual data $y_i$ are perfectly fit by the regression line. And $R^2 \geq 0$ provided an intercept term is included in the regression – if there is no intercept in the regression then $R^2$ is a meaningless concept and $R^2 < 0$ is possible.

As in the bivariate case, $R^2$ can also be expressed as

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

It can be shown that $R^2$ equals the squared correlation coefficient between the fitted values $\widehat{y}_i$ and the actual values $y_i$.

**Remark 123** *$R^2$ measures the fraction of the variation in the dependent variable explained by the regressor. Provided the regression includes an intercept, $R^2$ lies between 0 and 1. $R^2$ can equivalently be measured as the squared correlation coefficient between $y$ and $\widehat{y}$, and as one minus the residual sum of squares divided by the total sum of squares.*

### 13.6.3 Adjusted R-squared

A weakness of $R^2$ is that it necessarily increases, or at worse remains unchanged, as more regressors are added to a model. This is because the sum of squared residuals will at most be unchanged, so $R^2 = 1 - RSS/TSS$ is at the least unchanged.

To compare the fit across models with different numbers of regressors it is better to use a measure that penalizes larger models. The **adjusted R-squared**, denoted $\bar{R}^2$, is defined by

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2/(n-k)}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)}.$$

The motivation is that $k$ regression coefficients are estimated to form $\widehat{y}_i$, so the residual sum of squares is divided by the degrees of freedom, $n-k$. For similar reason the total sum of squares is divided by $n-1$ due to the estimate $\bar{y}$.

Given the above definition of $\bar{R}^2$, it follows immediately that

$$\bar{R}^2 = 1 - \frac{s_e^2}{s_y^2},$$

where $s_e^2/s_y^2$ is the ratio of the squared standard error of the regression to the sample variance of $y$. It follows that as regressors are added to the model, $\bar{R}^2$ will increase or decrease according to whether or not the standard error of the regression decreases as regressors are added.

$\bar{R}^2$ is related to $R^2$ by the formula

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2).$$

Clearly $\bar{R}^2 \leq R^2$. It can be shown that the difference between $R^2$ and $\bar{R}^2$ is greater the larger is the number of regressors ($k$) and the smaller is the sample size ($n$).

**Remark 124** *The adjusted $R^2$, denoted $\bar{R}^2$, is less than $R^2$ due to a modest penalty for the number of regressors in the model.*

While $\bar{R}^2$ is better than $R^2$, it may still err on the side of favoring models with many regressors. In particular, the penalty that $\bar{R}^2$ gives to larger models is not as large as the penalty imposed using hypothesis tests at the standard significance levels; see Chapter 14.5.

### 13.6.4 Information Criteria

The adjusted $R$-squared provides a penalty for larger models. Similarly the sum of squared residuals can be penalized as more regressors are added to the model.

Let $\widehat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{y}_i)^2$ be the sample average of the squared residuals. This is similar to $s_e^2$ except there is no degrees of freedom correction, so division is by $n$ rather than $n-k$.

The three commonly-used **information criteria** (IC) that penalize for model size are the following, in order of increasing penalty for model size.

| **Criteria** | | **General formula** |
|---|---|---|
| Akaike IC | $AIC$ | $= n \times \ln \widehat{\sigma}_e^2 + n(1 + \ln 2\pi) + 2k$ |
| Bayesian (or Schwarz) IC | $BIC$ | $= n \times \ln \widehat{\sigma}_e^2 + n(1 + \ln 2\pi) + k \times \ln(n)$ |
| Hannan-Quinn IC | $HQIC$ | $= n \times \ln \widehat{\sigma}_e^2 + n(1 + \ln 2\pi) + 2k \times \ln(\ln(n))$. |

Some statistical packages drop the constant $n(1+\ln 2\pi)$ in computing the information criteria. Models with low information criterion are preferred.

The information criteria favor models with smaller sum of squared residuals, but with a penalty for models with more regressors. $AIC$ provides a penalty for model size that is too light in practice. $BIC$ provides a larger penalty and is most often used.

For basic multiple regression the information criteria are used much less than $R^2$ and $\bar{R}^2$.

## 13.7 Computer Output Following Multiple Regression

As an example we consider regression of house sale price on several regressors. The example is illustrative. In particular with seven regressors, including the intercept, but only 29 observations the $t_{22}$ distribution may not be a good to the distribution of the $t$-statistic, unless the errors are normally distributed. And Chapter 16.3 presents a better way to incorporate the variable *Lotsize* that takes the three arbitrary values 1, 2 and 3.

The middle part of Table 13.4 presents the key regression output. We have $b_2 = 68.37$ and so on, with fitted model

$$\begin{aligned}\widehat{Price} &= 137791 + 68.37 \times Size + 2685 \times Bedrooms + 6833 \times Bathrooms \\ &\quad +2302 \times Lot\ Size - 833 \times Age - 2089 \times Month\ Sold.\end{aligned}$$

Table 13.4: House price: Computer output from multiple regression.

| **ANOVA Table** | | | | | |
|---|---|---|---|---|---|
| **Source** | **SS** | **df** | **MS** | **F** | **p-value** |
| Explained | $2.546\times10^{10}$ | 6 | $4.244\times10^{9}$ | 6.83 | 0.0003 |
| Residual | $1.3679\times10^{10}$ | 22 | $0.6218\times10^{9}$ | | |
| Total | $3.9146\times10^{10}$ | 28 | $0.1398\times10^{9}$ | | |

| **Dependent Variable** *Price* | | | | | | |
|---|---|---|---|---|---|---|
| **Regressor** | **Coefficient** | **Standard Error** | **t statistic** | **p value** | **95% conf. int.** | |
| *Size* | 68.37 | 15.39 | 4.44 | 0.000 | 36.45 | 101.29 |
| *Bedrooms* | 2685 | 9193 | 0.29 | 0.773 | -16379 | 21749 |
| *Bathrooms* | 6833 | 15721 | 0.43 | 0.668 | -25771 | 39437 |
| *Lot Size* | 2303 | 7227 | 0.32 | 0.753 | -12684 | 17290 |
| *Age* | -833 | 719 | -1.16 | 0.259 | -2325 | 659 |
| *Monthsold* | -2089 | 3521 | -0.59 | 0.559 | -9390 | 5213 |
| *Intercept* | 137791 | 61464 | 2.24 | 0.036 | 10321 | 265261 |
| **Summary Statistics** | | | | | | |
| Observations | 29 | | | | | |
| F(6,22) | 6.83 | | | | | |
| p-value for F | 0.0003 | | | | | |
| R-squared | 0.651 | | | | | |
| Adjusted $R^2$ | 0.555 | | | | | |
| St. error of regression | 24936 | | | | | |

Controlling for other likely determinants of house price, the estimated partial effect of an increase in the size of the house by one square foot is an increase in house price of \$68.37. This is not very different from the earlier bivariate regression slope estimate of \$73.77, which gives the estimated total effect.

The $t$-statistics and $p$-values given are for tests of statistical significance of each regressor. Since only regressor *Size* has $p < .05$, it is the only regressor that is statistically significant at level 0.05. Hypothesis testing for multivariate regression is presented in Chapter 14.

We conclude that house size is the primary determinant of house price in this market. One reason for the lack of importance of the other regressors may be that, aside from size, the houses are quite homogeneous as they are of similar vintage and are in a small part of a small city. A real estate maxim is "location, location, location", and here location has already been controlled for by choosing to focus on a fairly homogenous region. A second reason may be the small sample size. A third reason may be that in fact collectively the house attributes other than size may matter, even if individually each attribute is not statistically significant.

The bottom part of Table 13.4 presents the second most important output. The $F$-statistic and its associated $p$-value are used to test the joint statistical significance of all the regressors; see Chapter 14.5. $R^2 = .651$, an increase compared to $R^2 = .618$ for the model with just *Size* as regressor.

The adjusted $R^2$ can be computed as $\bar{R}^2 = .651 - (6/22) \times .349 = .555$. This is actually lower than $\bar{R}^2 = .603$ for the regression with just *Size* as regressor, suggesting that the additional regressors have not added enough to the model fit to compensate for the extra five slope coefficients that are estimated.

The standard error of the regression equals 24936. This is substantially smaller than the standard deviation of variable *Price* which from Table 8.2 equals 37391. At the same time it is larger than $s_e = 23551$ for the regression with just *Size* as regressor, again confirming that the additional regressors added little to model fit.

The top part of Table 13.4 gives the analysis of variance table, given the definition of $R^2$, we have $R^2 = 2.546 \times 10^{10}/(3.9146 \times 10^{10}) = 0.651$. And given the definition of the standard error of the regression, $s_e = \sqrt{1.3679 \times 10^{10}/22} = 24936$. These values are those given at the bottom of Table 13.4.

## 13.8 Inestimable Models

It is not always possible to estimate all $k$ regression coefficients in the regression of $y$ on an intercept and regressors $x_2, ..., x_k$. In such cases computer output such as that in the middle of Table 13.4 will have no entries for one or more regressors, and may include the word omitted. When not all coefficients can be estimated the coefficients are said to be **not identified.**

This situation is referred to as one of **perfectly collinear regressors** or of **data matrix of less than full rank.** It may arise due to inadequate variation in the data in a well specified model, or due to a poorly specified model.

### 13.8.1 Inadequate Variation in the Data

Suppose we wish to estimate how average earnings vary by gender, after controlling for years of schooling. To do so we fit the model $\widehat{y} = b_1 + b_2 d + b_3 x$ where $y$ is earnings, $d$ is an indicator variable equal to 1 if female and 0 if male, and $x$ is years of schooling. Then $b_2$ measures the difference in average earnings by gender after controlling for years of schooling.

Now suppose that we do not realize that the sample includes women only. Now $d = 1$ for all observations, so the model reduces to $\widehat{y} = (b_1 + b_2) + b_3 x$. We can only estimate an intercept and the coefficient of years of schooling. The problem here is that there is a perfectly linear relationship between the intercept and the regressor $d$, with $1 - d = 0$ in this sample.

As a second example, suppose we believe that medical expenditures rise with the number of chronic conditions a person has, such as heart disease, cancer and diabetes. However, we think the relationship is nonlinear, so we fit a quadratic model of the form $\widehat{y} = b_1 + b_2 x + b_3 x^2$. So we regress medical expenditures on an intercept, variable *Chronic* and variable *ChronicSq*, where *ChronicSq*= *Chronic*$^2$.

Now suppose that in our sample people had at most one chronic condition. Then variable *Chronic* takes only the values 0 or 1, in which case variable *ChronicSq* exactly equals variable *Chronic*. Then

$$\begin{aligned}\widehat{y} &= b_1 + b_2 \times \mathit{Chronic} + b_2 \times \mathit{ChronicSq} \\ &= b_1 + b_2 \times \mathit{Chronic} + b_2 \times \mathit{Chronic} \\ &= b_1 + (b_2 + b_3) \times \mathit{Chronic}.\end{aligned}$$

From the last equation, we can obtain estimates $b_1$ and $(b_2 + b_2)$ but we will not be able to separately estimate all three of $b_1$, $b_2$, and $b_3$. The problem here is that there is a perfect linear relationship between two of the regressors, with *Chronic*− *ChronicSq*= 0.

### 13.8.2 Poorly Specified Model

Individual earnings vary with age, education and work experience. Suppose we have data on age (variable *Age*) and years of schooling *(School)*. Data on years of work experience is not available, so we construct a variable *Exper*= *Age*− *School*− *6*, assuming that people enter school at age 6 and start working as soon as they leave school. Then

$$\begin{aligned}\widehat{y} &= b_1 + b_2 \times \mathit{Age} + b_3 \times \mathit{School} + b_4 \times \mathit{Exper} \\ &= b_1 + b_2 \times \mathit{Age} + b_3 \times \mathit{School} + b_4 \times (\mathit{Age} - \mathit{School} \ - \ 6) \\ &= (b_1 - 6b_4) + (b_2 + b_4) \times \mathit{Age} + (b_3 - b_4) \times \mathit{School}.\end{aligned}$$

From the last equation, we can obtain estimates $(b_1 - 6b_4)$, $(b_2 + b_4)$ and $(b_3 - b_4)$ but we will not be able to separately estimate all four of $b_1$, $b_2$, $b_3$, and $b_4$. So we can only include two out of *Age*, *School* and *Exper* in the regression.

As a second example, suppose that individual earnings vary with the type of worker, where all workers fall into exactly one of the following three categories of worker – self-employed, employed in the private sector and employed in the government sector. To include these in a regression model we create three indicator variables. Variable *dself* equals 1 if self-employed and equals 0 otherwise; variable *dpriv* equals 1 if employed in the private sector and equals 0 otherwise; and variable *dgovt* equals 1 if employed in the government sector and equals 0 otherwise. We then regress earnings on an intercept, *dself*, *dpriv*, *dgovt* and, say, regressor *Age*.

A problem arises because the three categories of worker are mutually exclusive. Because a worker is in exactly one of the categories, the sum *dself*+ *dpriv*+ *dgovt*= 1 for every worker.

For example, for a self-employed person we have $1+0+0=1$. So *dgovt*$=1-$ *dself*$-$ *dpriv*. Then

$$\begin{aligned}\widehat{y} &= b_1 + b_2 \times dself + b_3 \times dpriv + b_4 \times dgovt + b_5 \times Age \\ &= b_1 + b_2 \times dself + b_3 \times dpriv + b_4 \times (dgovt - dself - dpriv) + b_5 \times Age \\ &= b_1 + (b_2 - b_4) \times dself + (b_3 - b_4) \times dpriv + b_5 \times Age.\end{aligned}$$

We can obtain estimates $b_1$, $(b_2 - b_4)$, $(b_3 - b_4)$ and $b_5$, but we will not be able to separately estimate all five of $b_1$, $b_2$, $b_3$, $b_4$ and $b_5$. The solution is to drop as regressors one of *dself*, *dpriv* or *dgovt*, or to drop the intercept. This is an example of the **dummy variable trap**, discussed further in Chapter 16.3.

### 13.8.3 Solution

When there is inadequate variation in the data the solution is to either find data that has sufficient variation, or if that is not possible drop one or more of the regressors.

When the model is poorly specified in the first place, one or more regressors will need to be dropped.

Note that the problem of collinear regressors effects only estimation of the regressors that are collinear. For example, in the last example even though we could not separately estimate the coefficients of *dself*, *dpriv* or *dgovt*, the coefficient of *Age* can be estimated (even in the model with all three of *dself*, *dpriv* or *dgovt* included as regressors).

## 13.9 Key Concepts

1. Least squares regression with many regressors is no more difficult to perform than regression with just one regressor.

2. The multiple regression line is $\widehat{y} = b_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k$.

3. The residual $e = y - \widehat{y}$ is the difference between the actual and fitted values of the dependent variable.

4. The least squares coefficients $b_1, ..., b_k$ minimize the sum of squared residuals.

5. The least squares estimates can always be computed, provided there is sufficient variation in the regressor values across observations.

6. OLS residuals are orthogonal to the regressors, so for the $j^{th}$ regressor, $\sum_{i=1}^{n} x_{ji} e_i = 0$.

7. If an intercept is included in the regression then OLS residuals sum to zero, average to zero, and the average of in-sample fitted values equals the sample average of the dependent variable.

8. The coefficient $b_j$ in a multiple regression can be calculated by bivariate regression of $y$ on $\widetilde{x}_j$, where $\widetilde{x}_j = x_j - \widehat{x}_j$ is the residual from regressing $x_j$ on an intercept and all regressors other than $x_j$.

9. The slope coefficient $b_j$ measures the partial effect of a change in the fitted value of $y$ when $x_j$ changes by one unit and all other regressors are unchanged.

10. The total effect on the predicted value of $y$ when $x_j$ changes by one unit additionally allows for the secondary effect that the remaining regressors change when $x_j$ changes. When regression is by OLS, this total effect simply equals the slope coefficient from bivariate regression of $y$ on $x_j$.

11. OLS regression measures association or correlation, rather than causation. Without further information we can only say that a one-unit increase in $x_j$ is associated with, rather than is caused by, a $b_j$ increase in $y$, holding all other variables constant.

12. The standard error of the regression ($s_e$) measures the standard deviation of the residual, and hence the variability of $y$ around the regression line.

13. R-squared ($R^2$) measures the fraction of the variation of $y$ (around the sample mean $\bar{y}$) that is explained by the regressor. It should only be used if the regression includes an intercept.

14. R-squared equals the squared correlation between $y$ and $\widehat{y}$.

15. The adjusted $R^2$, denoted $\bar{R}^2$, is less than $R^2$ due to a modest penalty for the number of regressors in the model.

16. As regressors are added, $R^2$ cannot decrease, $\bar{R}^2$ will often increase but can also decrease, and $s_e$ will often decrease but may also increase.

17. Information criteria, used less often than $\bar{R}^2$, apply a penalty for model size to the sample average of the squared residuals.

18. If regressors are perfectly collinear then not all coefficients are identified.

19. Key terms: Two-way scatter plot; correlation; regression line; dependent variable; ordinary least squares regression; residual; fitted value; slope coefficient; intercept coefficient; causation; standard error of the residual; R-squared; adjusted R-squared; total sum of squares; residual (or error) sum of squares; explained (or regression or model)

sum of squares; information criteria; inestimable model; not identified model; perfectly collinear regressors; dummy variable trap.

## 13.10 Exercises

1. Consider tests of overall statistical significance. Then $F = [\textit{Explained SS}/(k-1)]/[\textit{Residual SS}/(n-k)]$. Show that this implies $F = [R^2/(k-1)]/[(1-R^2)/(n-k)]$. Hint: Divide numerator and denominator by *Total SS* and use the definitions of $R^2$.

# Chapter 14

# © A. Colin Cameron: Inference for Multiple Regression

Different samples will lead to different fitted regression lines, due to different random departures in the data from the population conditional mean. Statistical inferential methods control for this randomness.

This chapter presents the statistical properties of the least squares estimates and uses these properties to construct confidence intervals and perform hypothesis tests on the population parameters.

## 14.1 Properties of the Least Squares Estimates

The assumptions and their consequences for multiple regression are essentially the same as those for bivariate regression detailed in Chapter 9.1-9.4, except the population model now includes additional regressors and conditioning is on these extra regressors. A more abbreviated treatment is given here for multiple regression.

### 14.1.1 Data Assumptions

Throughout it is assumed that all the OLS coefficients can be computed. This requires that the sample size exceeds the number of regressors (including the intercept) and that no regressor is perfectly correlated with another regressor or a linear combination of the other regressors.

If this condition is not met then regression output from most statistical packages will not report coefficients for all regressors. Instead the computer output will indicate that one or more regressors (or the intercept) are omitted, or will simply have no entry for one or more regressors. This problem of perfect collinearity among regressors is discussed in Chapter 18.

The sampling process can be simple random sampling, where $(y_i, x_{2i}, ..., x_{ki})$ are jointly sampled from the population. Or the $x's$ may be fixed in advance, as is the case in a controlled experiment, with just the $y's$ randomly determined.

### 14.1.2 Population Line or Population Conditional Mean

The population relationship between $y$ and $x_2, ..., x_n$ is defined by $\mathrm{E}[Y|X_2 = x_2, ..., X_k = x_k]$, the **population conditional mean** of $Y$ given $X_2 = x_2, ..., X_k = x_k$. This is the probability-weighted average of all possible values of $Y$ in the population given $X_2 = x_2, ..., X_k = x_k$. In the econometrics literature it is standard to write the population conditional mean more simply as $\mathrm{E}[y|x_2, ..., x_k]$.

For linear regression it is assumed that **conditional mean** is a **linear** function of the regressors. Thus we write that the **conditional mean** or **conditional expectation function** or **population line** is

$$\mathrm{E}[y|x_2, ..., x_k] = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

The goal is to estimate the population parameters $\beta_1, \beta_2, ..., \beta_k$ as then we can make predictions, and we can estimate the effect on the conditional mean of changing one or more of the regressors.

**Remark 125** *$E[y|x_2, ..., x_k]$ denotes the conditional mean of $Y$ given $X_2 = x_2, ..., X_k = x_k$, the probability-weighted average of all possible values of $Y$ in the population given $X_2 = x_2, ..., X_k = x_k$. For linear regression $E[y|x_2, ..., x_k] = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.*

### 14.1.3 Population Assumptions

To obtain the statistical properties of the OLS coefficients we need to make assumptions about the population model and the sampling process that yielded the sample $(x_1, y_1), ...., (x_n, y_n)$.

Standard assumptions are that:

1. The **population model** is

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i \text{ for all } i.$$

2. The **error for the $i^{th}$ observation has zero mean conditional on all regressors**:

$$\mathrm{E}[u_i|x_{2i}, ..., x_{ki}] = 0 \text{ for all } i.$$

3. The **error for the $i^{th}$ observation has constant variance conditional on the regressors**:

$$\mathrm{Var}[u_i|x_{2i}, ..., x_{ki}] = \sigma_u^2 \text{ for all } i.$$

4. The **errors for different observations are statistically independent**

$$u_i \text{ is independent of } u_j, \text{ for al } i \neq j.$$

Assumptions 1-2 are the crucial assumptions that ensure that the population relationship between $y$ and the regressors is a linear relationship. Specifically the population mean

$$\mathrm{E}[y_i|x_{2i}, ..., x_{ki}] = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki},$$

so that the data are generated by a linear population relationship. In that case it makes sense to fit the data with a regression line.

Assumptions 3 and 4 are additional assumptions that are used in determining the precision and distribution of the estimates $b_1$ and $b_2$. Assumption 3 implies that the error term is homoskedastic, so the conditional variance of $y$ is the same across observations, with

$$\mathrm{Var}[y_i|x_{2i}, ..., x_{ki}] = \sigma_u^2.$$

More detailed discussion of these assumptions was given in Chapter 9.3.

### 14.1.4 Mean and Variance of a Least Squares Slope Coefficient

Let the typical regressor be the $j^{th}$ regressor $x_j$. Then inference on its coefficient $\beta_j$ is based on the OLS estimate $b_j$. The following results are simply stated here. Proofs in the case of a single regressor are given in Appendix 9.C.

The **mean of the least squares estimate** $b_j$ is

$$\mathrm{E}[b_j] = \beta_j, \quad j = 1, ...k,$$

given assumptions 1-2. So $b_j$ is unbiased for $\beta_j$. If many samples were available, yielding many estimates $b_j$, the average of the $b_j$ equals $\beta_j$.

If we additionally make assumptions 3-4 then the **variance of the least squares slope coefficient** $b_j$ can be shown to be

$$\mathrm{Var}[b_j] = \sigma_{b_j}^2 = \frac{\sigma_u^2}{\sum_{i=1}^n \widetilde{x}_{ji}{}^2},$$

where $\widetilde{x}_{ji}$ denotes the residual from regressing $x_{ji}$ on an intercept and all regressors other than $x_{ji}$.

**Remark 126** *Under assumptions 1-2 the estimated slope coefficient $b_l$ has mean equal to the population slope parameter $\beta_l$. Under assumptions 1-4 the estimated slope coefficient $b_j$ has variance $\sigma_{b_j}^2 = \sigma_u^2 / \sum_{i=1}^n \widetilde{x}_{ji}{}^2$, where $\widetilde{x}_{ji}$ denotes the residual from regressing $x_{ji}$ on an intercept and all regressors other than $x_{ji}$.*

### 14.1.5 Standard Error of a Least Squares Slope Coefficient

The variance of $b_j$ depends in part on $\sigma_u^2$, the variance of the error term, which is unknown. Under assumptions 1-4, an unbiased estimator of $\sigma_u^2$ is the **standard error of the regression**

$$s_e^2 = \frac{1}{n-k}\sum\nolimits_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

The **standard error of** $b_j$, the estimated standard deviation of $b_j$ is obtained by replacing $\sigma_u^2$ in the formula for $\sigma_{b_j}^2$ by $s_e^2$ and then taking the square root. Then

$$se(b_j) = \frac{s_e}{\sqrt{\sum_{i=1}^{n}\widetilde{x}_{ji}{}^2}}.$$

### 14.1.6 When is a Slope Coefficient Precisely Estimated?

Given the formula for $se(b_2)$ the precision of estimation of $\beta_2$ is better the closer the data are to the true regression line (then $s_e$ is small) and the larger is the sample size $n$ (then there are more terms in the sum). This is the same as in bivariate regression.

Additionally $se(b_j)$ is smaller the **less** $x_j$ is explained by the other regressors $x_2, ..., x_{j-1}, x_{j+1}, ..., x_k$. The simple intuition is that the coefficient of $x_j$ will be more precisely estimated the more $x_j$ adds an independent piece of information, rather than duplicating information already contained in the other regressors.

Algebraically, $se(b_j)$ is small if $|\widetilde{x}_{ji}|$ is large since then $\sum_{i=1}^{n}\widetilde{x}_{ji}{}^2$ is large. Since $\widetilde{x}_{ji} = x_{ji} - \widehat{x}_{ji}$ where $\widehat{x}_{ji}$ is the fitted value from regressing $x_{ji}$ on an intercept and all regressors other than $x_{ji}$, $|\widetilde{x}_{ji}|$ is large when $\widehat{x}_{ji}$ does a poor job of predicting $x_{ji}$.

**Remark 127** *The standard error of $b_j$ is $se(b_j) = s_e/\sqrt{\sum_{i=1}^{n}\widetilde{x}_{ji}{}^2}$, where $s_e$ is the standard error of the regression. Under assumptions 1-4 $se(b_j)^2$ has mean equal to the variance of $b_j$. Bigger samples are better – if the sample is $m$ times larger then $se(b_j)$ is approximately $1/\sqrt{m}$ times as large. And wider dispersion of $x_j$, after controlling for the other regressors,*

### 14.1.7 The $t$-Statistic

The $t$-statistic for multiple regression is the exact analog of that for bivariate regression, with the one change that the degrees of freedom are now $n-k$ rather than $n-2$.

Under assumptions 1-4, $b_j \sim (\beta_j, \sigma_{b_j}^2)$, so the standardized statistic $Z = (b_j - \beta_j)/\sigma_{b_j} \sim (0,1)$. By the central limit theorem, $Z$ is standard normal distributed as $n \to \infty$. However, $\sigma_{b_j}$ depends on the unknown parameter $\sigma_u^2$. Replacing $\sigma_{b_j}$ by its estimate $se(b_j)$ leads to the $t$-statistic.

The **$t$-statistic**

$$t = \frac{b_j - \beta_j}{se(b_j)} \sim T(n-k),$$

where $T(n-k)$ denotes the $t$ distribution with $n-k$ degrees of freedom. More formally, $t$ is the realization of a random variable that is $T(n-k)$ distributed. The $t$ distribution is discussed in some detail in Chapter 6.2.

This result is exact if assumptions 1-4 hold and additionally the errors are normally distributed. If the errors are nonnormal then the $T(n-k)$ distribution is a good approximation provided $n$ is sufficiently large, but there is no clear-cut rule on how large $n$ needs to be.

**Remark 128** *Under assumptions 1-4 the t-statistic $t = (b_j - \beta_j)/se(b_j)$ is the realization of a randomly variable that is $T(n-k)$ distributed, exactly if data are normally distributed and approximately for nonnormal data if $n$ is sufficiently large.*

### 14.1.8 How Large Should the Sample Be?

The $T(n-k)$ distribution is exact under the strong assumptions 1-4 with normal errors. Generally these assumptions are too strong, as model errors are not exactly normally distributed and often assumptions 3-4 are relaxed (then alternative methods given in Chapter 14.6 are used to compute $se(b_j)$).

In these more realistic settings the $T(n-k)$ is an approximation, one that gets more reasonable as the sample size gets large. Unfortunately, there is no simple way to determine whether the sample size for a particular data set and regression is adequate for the $T(n-k)$ approximation to be good, though more advanced analysis using simulations can provide some guidance. Thus there is no hard and fast rule for how large the sample size should be.

For bivariate regression with a single regressor many authors feel that there should be at least thirty observations, though this is by no means sufficient for all types of data. For example, suppose the dependent variable $y$ takes only two values, such as a variable for whether or not an individual is employed. Then a larger sample size will be needed since it is very unlikely that the error term $u$ has a continuous normal distribution when the dependent variable $y$ is discrete and takes only two values.

For multiple regression, as more regressors are added the sample size should be even larger.

If the sample size is too small then the distribution of the $t$-statistic generally has fatter tails than the $T(n-k)$. As a consequence, confidence intervals based on the $T(n-k)$ will be too narrow, and hypothesis tests will tend to reject the null hypothesis too often.

The house price example used in several chapters of this book has only 29 observations. This has the pedagogical advantage of, for example, making it easy to list the complete data. In practice, however, it would be better to have a larger sample.

Aside from considerations of approximating the distribution of the $t$-statistic, another reason for not using samples that are too small is that estimation may be very imprecise, so that statistical analysis is too noisy to be useful.

## 14.2 Point Estimation

The first goal of inference is to obtain estimates of the population parameters $\beta_1, ..., \beta_k$. This is called **point estimation**, to distinguish it from interval estimation using confidence intervals.

### 14.2.1 Optimal Properties of OLS Estimators

The desirable properties of OLS in bivariate regression, discussed at some length in Chapter 9.5, carry over to multiple regression.

First, the estimated OLS coefficients $b_j$ are **unbiased** for the population slope parameters $\beta_j$, provided the population model satisfies assumptions 1 and 2 since then $\mathrm{E}[b_j] = \beta_j$.

Second, the estimated OLS coefficients $b_j$ are **consistent** for the population slope parameters $\beta_j$, if additionally $\mathrm{Var}[b_j]$ goes to zero as $n \to \infty$. This is the case under the additional assumptions 3 and 4, and under alternative assumptions such as those given in Chapter 14.6.

Third, under assumptions 1-4, the OLS estimates $b_j$ are **best linear unbiased (BLUE)**. They have smallest variance among unbiased estimators that are a weighted average of $y_i$ of the form $b_j = \sum_{i=1}^{n} w_i y_i$ where the weights will depend on the regressors. This result is called the **Gauss-Markov Theorem**.

Fourth, if additionally the errors are normally distributed then the OLS estimates $b_j$ are **best unbiased**, meaning the have minimum variance among unbiased estimators. OLS is also the best estimator among those that are consistent and asymptotically normal, even if the errors are not normally distributed.

**Remark 129** *The OLS estimators $b_1, ..., b_k$ are unbiased for $\beta_1, ..., \beta_k$ under assumptions 1-2, best linear unbiased under assumptions 1-4, and best unbiased if additionally the more errors are normally distributed. In large samples the OLS estimators $b_1, ..., b_k$ are consistent for $\beta_1, ..., \beta_k$ under assumptions 1-2, plus assumptions(s) such assumptions 3-4 that ensure the estimators have finite variance, and are best among consistent and asymptotically normal estimators under assumptions 1-4.*

### 14.2.2 Failure of Assumptions 1-4

The slope parameter estimates $b_j$ will usually be **biased** and **inconsistent** if assumptions 1 and/or 2 are not satisfied. In that case alternative analysis is needed.

Often interest lies in the coefficient of a single regressor, say as the coefficient of years of schooling in an earnings-schooling regression that includes additional control variables. Alternative analysis that may lead to consistent estimation of the key parameter of interest may entail using a different functional form for the conditional mean, including additional

control variables in the regression, using data from a different source, and using an estimation method other than OLS. These complications are considered at length in Chapters 17 and 18.

Assumptions 3-4 generally effect only the estimation of the precision of $b_j$. These assumptions can be relaxed. First, one can continue to estimate by OLS but use an alternative formula to compute the standard errors of the OLS coefficient estimates; see Chapter 14.6. Second, one can use alternative estimation methods that may lead to coefficient estimates than are more precise than the OLS estimates; see Chapter 14.7.

## 14.3 Confidence Intervals

Different samples lead to different estimates of the regression model parameters $\beta_1, ..., \beta_k$. A confidence interval for the $j^{th}$ parameter, $\beta_j$, gives a range of values that includes $\beta_j$ with a specified probability or level of confidence.

### 14.3.1 Confidence Intervals

As usual, the confidence interval is the estimate plus or minus the relevant critical $t$ value times the standard error of the estimate.

**Remark 130** *A* $100(1-\alpha)$ ***percent confidence interval for the slope parameter*** $\beta_j$ *is*

$$b_j \pm t_{n-k,\alpha/2} \times se(b_j),$$

*where* $b_j$ *is the OLS slope coefficient,* $t_{n-k,\alpha/2}$ *is that value such that a* $T(n-k)$ *distributed random variable exceeds it in absolute value with probability* $\alpha$*, and* $se(b_j)$ *is the standard error of the slope coefficient estimate* $b_j$*.*

This confidence interval is exact only if assumptions 1-4 hold and the errors are normally distributed.

The interpretation of the confidence interval is the same as that detailed at length in Chapter 10.3 for bivariate regression. A 95% confidence interval for $\beta_j$ is interpreted as an interval that has probability 0.95 of including $\beta_j$. That is, if we had many repeated samples leading to many different confidence intervals, 95% of these confidence intervals will include the true unknown value of $\beta_j$.

### 14.3.2 Example: House Price and Size

As an example, consider the house price data in dataset HOUSE. Output from regression of house price on an intercept and six other regressors was presented in Table 13.4. For regressor *Size*, the output lists the 95 percent confidence interval to be $(36.45, 100.29)$.

This confidence interval can be computed from first principles as follows. The output gives $b_{Size} = 68.37$, from the coefficient column, and $se(b_{Size}) = 15.39$, from the standard error column. And here $t_{n-k,\alpha/2} = t_{22;.025} = 2.052$. Then a 95% confidence interval for $\beta_{Size}$ is

$$b_{Size} \pm t_{n-k,\alpha/2} \times se(b_{Size}) = 68.37 \pm 2.074 \times 15.39 = 68.37 \pm 31.92 = (36.45, 100.29).$$

## 14.4 Hypothesis Tests on a Single Parameter

Hypothesis tests on individual parameters in a multiple regression model are the same as those for the slope parameter i bivariate regression, except that the $T(n-k)$ distribution is used rather than the $T(n-2)$ distribution, where $k$ is the number of regressors including the intercept, rather than $(n-2)$ degrees of freedom.

The presentation here is therefore brief. Chapter 10.4-10.5 provides a more detailed discussion of the issues involved in hypothesis testing.

### 14.4.1 Tests on Individual Parameters

Consider tests on the $j^{th}$ regression parameter $\beta_j$. A **two-sided test** or **two-tailed test** on the parameter $\beta_j$ is a test of $H_0 : \beta_j = \beta_j^*$ against $H_a : \beta_j \neq \beta_j^*$, where $\beta_j^*$ is a specified value for $\beta_j$.

The null hypothesis is rejected when $b_j$ is far from $\beta_j^*$. Equivalently the null hypothesis is reject if the $t$-statistic $t = (b_j - \beta_j^*)/se(b_j)$ is large. Under the null hypothesis that $\beta_j = \beta_j^*$ and under assumptions 1-4, the $t$-statistic is a draw from the $T(n-k)$ distribution. The null hypothesis is rejected if the observed value $|t|$ is so large that it is very unlikely to have observed this value, if indeed $H_0$ was true so that the $t$-statistic was indeed a draw from the $T(n-k)$.

Similar to the bivariate case, this leads to the following.

**Remark 131** *For a **two-sided test** of $H_0 : \beta_j = \beta_j^*$ against $H_a : \beta_j \neq \beta_j^*$ the $t$-**statistic***

$$t = \frac{b_j - \beta_j^*}{se(b_j)}$$

*is a draw from the $T(n-k)$ distribution, approximately, if $H_0$ is true. The $p$-**value** is $p = \Pr[|T_{n-k}| \geq |t|\ ]$ and $H_0$ is rejected if $p < \alpha$, where $\alpha$ is the desired **significance level** of the test. The **critical value** $c$ is such that $c = t_{n-k,\alpha/2}$, equivalently $\Pr[|T_{n-k}| \geq c] = \alpha$, and $H_0$ is rejected at significance level $\alpha$ if $|t| > c$.*

It is also possible to perform a **one-sided test** or **one-tailed test.** In that case the claim to be tested is set as the alternative hypothesis. Table 14.1 summarizes one-sided and two-sided hypothesis tests on the population parameter slope coefficient.

Table 14.1: Summary for hypothesis tests on the slope parameter in multiple regression.

| | Two-sided Test | One-sided Upper alternative | One-sided Lower alternative |
|---|---|---|---|
| Null hypothesis | $H_0 : \beta_j = \beta_j^*$ | $H_0 : \beta_j \leq \beta_j^*$ | $H_0 : \beta_j \geq \beta_j^*$ |
| Alternative Hypothesis | $H_0 : \beta_j \neq \beta_j^*$ | $H_0 : \beta_j > \beta_j^*$ | $H_0 : \beta_j < \beta_j^*$ |
| t-statistic | $t = (b_j - \beta_j^*)/se(b_j)$ | $t = (b_j - \beta_j^*)/se(b_j)$ | $t = (b_j - \beta_j^*)/se(b_j)$ |
| p-value | $p = \Pr[\lvert T_{n-k}\rvert \geq t]$ | $p = \Pr[T_{n-k} \geq t]$ | $p = \Pr[T_{n-k} \leq t]$ |
| Rejection rule | $p < \alpha$ | $p < \alpha$ | $p < \alpha$ |
| Critical value | $c = t_{n-k,\alpha/2}$ | $c = t_{n-k,\alpha}$ | $c = -t_{n-k,\alpha}$ |
| Rejection region | $\lvert t\rvert > c$ | $t > c$ | $t < c$ |

### 14.4.2 Relationship between $t$ test and Adjusted R-Squared

It can be shown that the adjusted R-squared, $\bar{R}^2$, increases as one regressor is added to the model if and only if a test for statistical significance of this regressor, based on default standard errors, yields $|t| > 1$.

By contrast a hypothesis test at the usual significance level of 0.05 has a much higher threshold such as $|t| > 1.96$ for a large sample size. It follows that an increase in $\bar{R}^2$ does not necessarily imply statistical significance of the additional regressor(s) at conventional levels of significance such as $\alpha = 0.05$. While $\bar{R}^2$ provides a penalty for large model size, the penalty is too weak.

### 14.4.3 Test of Individual Statistical Significance

A special case of tests on an individual parameter is a **test of statistical significance**, in which case the hypothesized value of $\beta_j$ is $\beta_j^* = 0$.

Regression packages print out the necessary statistics to automatically test this hypothesis, including both the $t$-statistic and the $p$-value for the test of $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$. An example is given in Table 13.4. $H_0$ is rejected at statistical significance level $\alpha$ if $p < \alpha$. If instead $p > \alpha$ we do not reject $H_0$, and conclude that there is no statistically significant relationship or equivalently, that the regressor is statistically insignificant. The most common choice of $\alpha$ is 0.05, followed by 0.01 and 0.10.

It is most common to perform two-sided tests of statistical significance, so computer output reports the $p$-value for a two-sided test. If we have prior beliefs about the sign of $\beta_j$, however, then this prior belief can be employed by performing a one-sided test. The prior belief is setup as the alternative hypothesis, as explained in Chapter 7.1. For example, since we believe that house price increases with size, a test of significance of house size may be a test of $H_0 : \beta_j \leq 0$ against $H_a : \beta_j > 0$. In that case we halve the printed $p$-value, provided the estimate $b_j$ is positive.

There can be ambiguity in the statement that "the regressor is statistically significant at significance level 0.05", as it is not always made clear whether a one-sided or two-sided test was performed. Most often tests of statistical significance are performed as two-sided tests, even if there are strong prior beliefs on the direction of the relationship.

Statistical significance by itself does not imply automatically that the regressor has **economic significance**. An economically significant regressor has coefficient $b_j$ that is large enough that changing $x_j$ is associated with meaningful changes in the dependent variable. Statistical significance is determined by the size of the $t$-statistic which is the coefficient $b_j$ scaled by $se(b_j)$. Economic significance is instead determined only by by whether the coefficient $b_j$ is "small" or "large". This is a subjective judgement that is context specific.

### 14.4.4 Example: House Prices

As an example we consider regression of house sale price on several regressors using dataset HOUSE. The example is illustrative. In particular with seven regressors including the intercept but only 29 observations the $T(22)$ distribution may not be a good to the distribution of the $t$-statistic, unless the errors are normally distributed.

Results are given in Table 14.2, which repeats Table 13.4.

The only regressor that is statistically significant at significance level 0.05 is *Size* with $p = 0.0002$. The next most statistically significant regressor is *Age* with a $p$-value of .259 that is much higher than .05.

Suppose that instead of a test of statistical significance we wish to test whether an increase in house size of one square foot is associated with a \$50 increase in house price. For test of $H_0 : \beta_{Size} = 50$ against $H_0 : \beta_{Size} \neq 50$, $t = (68.37 - 50)/15.39 = 1.194$. Then $p = \Pr[|T_{22}| > |1.194|] = 0.245$ so we do not reject $H_0$ at significance level 0.05.

Looking at the magnitude of the coefficients, the estimate $b_{Size} = 68.37$ means that a 100 square foot increase in size, equivalent to a small room that is ten feet by ten feet, is associated with a \$6,837 increase in house price, an economically meaningful effect. An additional bathroom is also associated with a substantial increase in house price, but this estimate is so imprecise (the 95% confidence interval is $-$\$25,771 to \$39,437) that no attention should be paid to it.

We conclude that house size is the primary determinant of house price in this market.

One reason for the lack of importance of the other regressors may be that, aside from size, the houses are quite homogeneous as they are of similar vintage and are in a small part of a small city. A real estate maxim is "location, location, location", and here location has already been controlled for by choosing to focus on a fairly homogenous region.

A second reason may be the small sample size.

A third reason may be that in fact collectively the house attributes other than size may matter, even if individually each attribute is not statistically significant. This can be tested using the $F$ test presented below.

Table 14.2: House price: Computer output from multiple regression.

| ANOVA Table | | | | | | |
|---|---|---|---|---|---|---|
| Source | SS | df | MS | F | p-value | |
| Explained | $2.546\times10^{10}$ | 6 | $4.244\times10^{9}$ | 6.83 | 0.0003 | |
| Residual | $1.3679\times10^{10}$ | 22 | $0.6218\times10^{9}$ | | | |
| Total | $3.9146\times10^{10}$ | 28 | $0.1398\times10^{9}$ | | | |
| Dependent Variable *Price* | | | | | | |
| Variable | Coefficient | Standard Error | t statistic | p value | 95% conf. int. | |
| *Size* | 68.37 | 15.39 | 4.44 | 0.000 | 36.45 | 101.29 |
| *Bedrooms* | 2685 | 9193 | 0.29 | 0.773 | -16379 | 21749 |
| *Bathrooms* | 6833 | 15721 | 0.43 | 0.668 | -25771 | 39437 |
| *Lot Size* | 2303 | 7227 | 0.32 | 0.753 | -12684 | 17290 |
| *Age* | -833 | 719 | -1.16 | 0.259 | -2325 | 659 |
| *Monthsold* | -2089 | 3521 | -0.59 | 0.559 | -9390 | 5213 |
| *Intercept* | 137791 | 61464 | 2.24 | 0.036 | 10321 | 265261 |
| Observations | 29 | | | | | |
| F(6,22) | 6.83 | | | | | |
| p-value for F | 0.0003 | | | | | |
| R-squared | 0.651 | | | | | |
| Adjusted $R^2$ | 0.555 | | | | | |
| St. error of regression | 24936 | | | | | |

## 14.5 Joint Hypothesis Tests

For multiple regression we may perform joint hypothesis tests on several parameters, rather than on a single regression parameter. A simple example is test of the null hypothesis that both $\beta_2 = 0$ and $\beta_3 = 0$. A more complicated example is test that both $\beta_2 = -\beta_3$ and $2\beta_4 + \beta_6 = 9$.

Joint hypothesis tests use the $F$ distribution, rather than the $t$ distribution.

### 14.5.1 $F$ Distribution

The $F$ distribution is a continuous right-skewed distribution for a random variable that takes only positive values. The distribution depends on two parameters, called the first and second degrees of freedom. This can be denoted $F_{v_1,v_2}$ where the degrees of freedom $v_1$ and $v_2$ are positive integers. Note that the order of the degrees of freedom matters, as $F_{v_1,v_2} \neq F_{v_2,v_1}$. The mean exists if $v_2 > 2$ and equals $v_2/(v_2-2)$, so the mean approaches one as $v_2$ gets large. When $v_1 = 1$ the $F$ distribution reduces to the square of the $t$ distribution, i.e. $F_{1,v_2} = (T_{v_2})^2$.

Figure 14.1 presents the probability density function of the $F$ distribution for $(v_1, v_2) =$

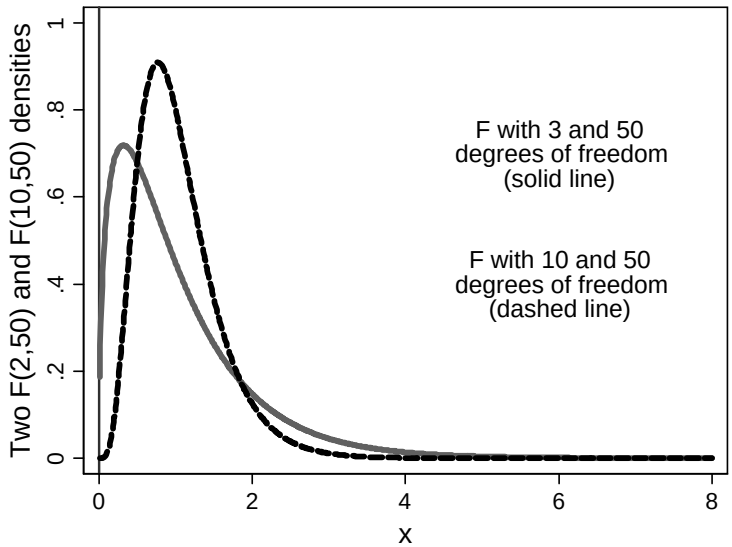


Figure 14.1: F distribution: F(3,50) and F(10,50)

$3, 50$ and $(v_1, v_2) = (10, 50)$. In most economics applications $v_1$ is small $v_2 = n - k$, the regression degrees of freedom.

**Remark 132** *The F($v_1$,$v_2$) distribution is a right-skewed distribution, with degrees of freedom $v_1$ and $v_2$, for a random variable that takes only positive values. $F_{v1,v2,\alpha}$ denotes the value for which the area in the right tail of the distribution equals $\alpha$.*

### 14.5.2 $F$-Statistic

Consider comparison of two models that are nested in each other.

The more general model, called the **unrestricted model** or **complete model**, is a model with $k$ regressors, so

$$y = \beta_1 + \beta_2 x + \beta_3 x_3 + \cdots + \beta_k x_k + u.$$

The **restricted model** or **reduced model** is a model that places restrictions on the parameters $\beta_1, \beta_2, ..., \beta_k$. Most often this restricted model is one that omits some of the regressors, so that some of the $\beta's$ are set to zero.

Let $q$ denote the number of restrictions imposed by the restricted models. Examples studied below are

1. All regressors but the intercept are dropped in the restricted model, so $q = k - 1$.

2. Only a subset of $g$ regressors, including the intercept, are included in the restricted model, so $q = k - g$.

3. Just one regressor is dropped in the restricted model, so $q = 1$.

Intuitively if the fit of the restricted model is very close to that of the unrestricted model we should favor the restricted model since it is a smaller model, whereas if the difference is great then the cost of imposing the restrictions is too high and we should favor the unrestricted model.

The $F$ test formalizes this, using the residual sum of squares, $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, as a measure of model fit. Let $RSS_r$ denote the residual sum of squares in the restricted model, and $RSS_u$ denote the residual sum of squares in the unrestricted model. Then $RSS_r > RSS_u$ since OLS minimizes the residual sum of squares and this minimum will be lower in the more general model. A large value for $RSS_r - RSS_u$ indicates that the restricted model has much worse fit, so we should not use the restricted model. The $F-$statistic uses an appropriate rescaling of $RSS_r - RSS_u$.

**Remark 133** *Define the* ***F-statistic***

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k)},$$

*where RSS denotes residual sum of squares, subscripts $r$ and $u$ denote the restricted and unrestricted models, $q$ is the number of parameter restrictions, $n$ is the number of observations, and $k$ is the number of regressors in the unrestricted model. If the restricted model is the appropriate model and assumptions 1-4 hold then $F$ is the realization of a random variable that is approximately $F(q, n-k)$ distributed.*

The $F(q, n-k)$ approximation is exact if additionally the data are normally distributed or if $n \to \infty$.

### 14.5.3 $F$ Tests

In some special cases the $F$-statistic can be used to form a confidence region. For example, a confidence interval for $\beta_2$, say, can be generalized to a **joint confidence region** for $\beta_2$ and $\beta_3$ (which using the $F$-statistic turns out to be an ellipse). Such confidence regions are rarely used. Instead the main use of the $F$-statistic is in hypothesis testing.

**Remark 134** *An $F$* ***test*** *is a* ***two-sided test of***

$$H_0: \text{ The parameter restrictions implied by the restricted model are correct,}$$

*against the alternative hypothesis*

$$H_a: \text{ The parameter restrictions implied by the restricted model are incorrect.}$$

Table 14.3: F distribution with q and n-k degrees of freedom: test critical values.

| Test size | | $q=1$ | $q=2$ | $q=3$ | $q=10$ | $q=20$ |
|---|---|---|---|---|---|---|
| 10% | $n-k=30$ | 2.88 | 2.49 | 2.28 | 1.82 | 1.67 |
| | $n-k=\infty$ | 2.71 | 2.30 | 2.08 | 1.60 | 1.42 |
| 5% | $n-k=30$ | 4.17 | 3.32 | 2.92 | 2.16 | 1.93 |
| | $n-k=\infty$ | 3.84 | 3.00 | 2.61 | 1.83 | 1.57 |
| 1% | $n-k=30$ | 7.56 | 5.39 | 4.51 | 2.98 | 2.55 |
| | $n-k=\infty$ | 6.64 | 4.61 | 3.78 | 2.32 | 1.88 |

The $F$-statistic is necessarily positive since, as already explained, $RSS_r > RSS_u$. Large values of the $F$-statistic lead to rejection of $H_0$.

**Remark 135** *The $p$-**value** for the $F$ test is $p = \Pr[F_{q,n-k} \geq F\ ]$ and $H_0$ is rejected if $p < \alpha$, where $\alpha$ is the desired **significance level** of the test. The **critical value** $c$ is such that $c = F_{q,n-k,\alpha}$, equivalently $\Pr[|F_{q,n-k}| \geq c] = \alpha$, and $H_0$ is rejected at significance level $\alpha$ if $|t| > c$.*

Note that the $F$ test is a **two-sided test**. It is true that the $F$ test $p$-value or critical value is calculated using only **one tail**, the right tail, of the $F$ distribution. But the test itself is a two-sided test because in the case of regressor exclusion restrictions, for example, the alternative hypothesis can only be that parameters are not equal to zero rather than, for example, greater than zero. A one-sided $F$ test is not possible.

Table 14.3 gives examples of the critical values for the $F(q, n-k)$ distribution with different degrees of freedom and for tests at significance level 0.10, 0.05 and 0.01. The critical values decrease as the number of restrictions ($q$) increases, the degrees of freedom in the unrestricted model ($n-k$) increases, and as test significance level ($\alpha$) increases.

The $F$ critical values are quite complicated and it is much easier to rely on $p$-values computed by a statistical package. If these are unavailable note that if $n-k > 10$ then $H_0$ is always rejected at level 0.05 if $F > 5$, regardless of the size of $q$.

The leading examples of the $F$ test are now presented.

### 14.5.4 Test of Overall Significance

A **test of overall significance** simultaneously tests whether all of the regressors, aside from the intercept, are statistically significant.

**Remark 136** *A **test of overall significance** is a **two-sided test of***

$$H_0 : \beta_2 = 0, ..., \beta_k = 0,$$

*against the alternative hypothesis*

$$H_a: \text{ At least one of } \beta_2 \neq 0, ..., \beta_k \neq 0.$$

*Then the general formula simplifies to* $F = [(ExpSS)/(k-1)]/RSS/(n-k)$*, where ExpSS and RSS are, respectively, the explained and residual sum of squares in the unrestricted model.*

Here $q = k-1$ since the restricted model is obtained by setting $k-1$ slope coefficients to zero. To obtain the simpler formula for $F$, note that the restricted model is simply regression on an intercept. This yields $b_1 = \bar{y}$, so $RSS_r = \sum_{i=1}^{n}(y_i - \bar{y})^2 = TSS$, where $TSS$ is the total sum of squares in the unrestricted model. Then $RSS_r - RSS_u = TSS - RSS = ExpSS$, since the total sum of squares can be decomposed as $TSS = ExpSS + RSS$.

Most statistical packages automatically print out the $F$-statistic for test of overall significance, along with its associated $p$-value.

Yet another equivalent formula for test of overall statistical significance is $F = [R^2/(k-1)]/[(1-R^2)/(n-k)]$. Thus $F$ increases as $R^2$ increases.

$$F = \frac{(ExpSS)/(k-1)}{RSS/(n-k)},$$

This quantity can be directly calculated from an ANOVA table – it is the ratio of the first two terms in the last column in Table 14.2.

Another equivalent formula for test of overall statistical significance is $F = [R^2/(k-1)]/[(1-R^2)/(n-k)]$. This implies that $F$ necessarily increases as $R^2$ increases.

### 14.5.5 Example: Test of Overall Significance

Consider the house price example results given in Table 14.2. The entry $F(6, 22) = 6.83$ is the $F$-statistic for test of overall significance, since six regression coefficients are then set to zero. Since $p = 0.0003 < 0.05$ we reject $H_0$ and assume that the regressors are jointly statistically significant.

This $F$-statistic can be computed from the ANOVA table given in Table 14.2. From the last column of that table, $(ExpSS)/(k-1) = 4.244 \times 10^9$ and $(RSS)/(n-k) = 0.6218 \times 10^9$. The ratio is then 6.83.

Finally, since $R^2 = 0.6506$, the formula for $F$ in terms of $R^2$ yields $F = (.6506/6)/(.3494/22) = 6.83$.

Note that the $F$ test is merely a test of whether at least one regressor is statistically significant, not that all regressors are statistically significant. In this example from Table 14.2 only one of the regressors, *Size*, is individually statistically significant at level 0.05.

### 14.5.6 Test of Subsets of Regressors

**Tests of a subset of regressors** compare two nested models that include a common set of regressors, and test whether the additional regressors in the unrestricted model are jointly statistically significant.

The unrestricted model or complete model with $k$ regressors is specified to be

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k + u.$$

The restricted model or reduced model is specified to includes only the first $g$ regressors so

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_g x_g + u.$$

The restricted model is therefore obtained by setting $k - g$ slope coefficients to zero.

**Remark 137** *Let the last $q = k - g$ regressors be omitted in the restricted model. Then a* ***test of a subset of regressors*** *is a* ***two-sided test of***

$$H_0 : \beta_{g+1} = 0, ..., \beta_k = 0,$$

*against the alternative hypothesis*

$$H_a : \text{ At least one of } \beta_{g+1} \neq 0, ..., \beta_k \neq 0.$$

In this case we use the original definition of the $F$-statistic, where $q = k - g$ since there are $k - g$ restrictions. So we need to obtain the residual sum of squares in both the complete and the reduced models.

Some statistical packages provide a post-estimation test command that computes the $F$ test for a user-specified subset of regressors. In that case the command may additionally enable test of more complicated hypotheses, such as a test of $H_0 : \beta_2 + \beta_3 = 1$, $\beta_4 = 2$ against $H_0$ : At least one of $\beta_2 + \beta_3 \neq 1$, $\beta_4 \neq 2$.

### 14.5.7 Example: Test of Subsets of Regressors

Return to the house price example. We have shown that house size is the only one of the six regressors that are statistically insignificant. A natural question to ask is whether the five regressors statistically insignificant regressors when taken together are statistically significant, or whether a model with just house size as a regressor is adequate.

From Table 14.2, $RSS_u = 1.3679 \times 10^{10}$ in the unrestricted model that includes all six regressors. The restricted model is a bivariate regression of house price on house size. From Table 8.5 in Chapter 8.8, $RSS_r = 1.4975 \times 10^{10}$. Then

$$F = \frac{(1.4975 \times 10^{10} - 1.3679 \times 10^{10})/5}{1.3679 \times 10^{10}/22} = 0.417.$$

Since $p = \Pr[F_{5,22} > 0.417] = 0.832 > 0.05$, we do not reject $H_0 : \beta_3 = 0, ..., \beta_7 = 0$ at significance level 0.05.

We conclude that the additional five regressors are jointly statistically insignificant. It is best to just include *Size* as a regressor.

### 14.5.8 Test of Single Regressor

The $F$ test for subsets of regressors can be used to test whether a single regressor is statistically significant, in which case $q = 1$. This $F$ test yields exactly the same result as that from a two-sided $t$ test of statistical significance.

First, it can be shown algebraically that the $F$ test statistic is the square of the usual $t$ test statistic, so $F = t^2$. Second, a random variable with $T(n-k)$ distribution when squared is $F(1, n-k)$ distributed.

It follows that when $q = 1$, the $F$-test critical value equals the square of the critical value for a two-sided $t$ test. So a two-sided $t$-test critical value of 1.96 corresponds to an $F$-test critical value of $1.96^2 = 3.84$.

**Remark 138** *The $F$ test for statistical significance of a single regressor is equivalent to a two-sided $t$ test, since $F = t^2$ and $\Pr[|T_{n-k}| > t] = \Pr[|F_{1,n-k}| > t^2]$.*

### 14.5.9 Relationship between $F$ test and Adjusted R-Squared

It can be shown that the adjusted R-squared, $\bar{R}^2$, increases as one or more regressors are added to the model if and only if the $F$-statistic for these additional regressors exceeds one.

By contrast an $F$ test at the usual significance level of 0.05 has a much higher threshold than $F > 1$. It follows that an increase in $\bar{R}^2$ does not necessarily imply statistical significance of the additional regressors at conventional levels of significance such as $\alpha = 0.05$. While $\bar{R}^2$ provides a penalty for large model size, the penalty is too weak.

**Remark 139** *As regressors are added the adjusted R-squared increases if and only if $F > 1$ where $F$ is the test statistic for test of joint statistical significance of the additional regressors.*

### 14.5.10 Chi-squared Distribution

The $t$ distribution goes to the standard normal distribution as the degrees of freedom go to infinity, that is the $T(\infty)$ distribution is the $N(0, 1)$ distribution.

A similar limit result exists for the $F$ distribution. A random variable that has $F(q, \infty)$ distribution when multiplied by $q$ has the chi-squared distribution with $q$ degrees of freedom, denoted $\chi^2(q)$. The $\chi^2(q)$ distribution mean $q$ and variance $2q$.

Some statistical packages report results for the $\chi^2$ distribution rather than the $F$ distribution. In that case note that a $\chi^2(v)$ random variable divided by its degrees of freedom $q$ has the $F(q, \infty)$ distribution.

## 14.6 Inference with Robust Standard Errors

Model assumptions 1-4 lead to standard errors for the regression coefficients that are called **default standard errors ,**

When model assumptions 3 and 4 are relaxed, statistical inference needs to be based on alternative standard errors, called **robust standard errors**, that were introduced in Chapter 10.7. This leads to different numerical values for standard errors and the associated confidence intervals, $t$-statistics and $F$-statistics, but they are then used in the same way as previously.

**Remark 140** *Standard errors for the least squares coefficients obtained under assumptions weaker than assumptions 3-4 are called* ***robust standard errors****.*

The three leading examples of robust standard errors are given in Table 14.5.

Table 14.4: Robust standard errors: Leading examples.

| Complication | Robust Standard Error Type | Data Type |
|---|---|---|
| Heteroskedasticity: Error variance varies over i | Heteroskedasticity robust | Most Cross section |
| Autocorrelation: Errors correlated over time | Heteroskedasticity and autocorrelation (HAC) robust | Most Time Series<br>Some Time Series |
| Clustered: Errors in same cluster are correlated | Cluster robust | Some Cross section<br>Most Short Panel |

All statistical packages other than Excel provide heteroskedasticity-robust standard errors. Statistical packages vary in the extent to which they provide HAC and clustered standard errors.

### 14.6.1 How To Use Robust Standard Errors

For statistical inference on a **single parameter** the main change is to replace the default standard error with the appropriate robust standard error. Thus given a robust standard error of $b_j$, denoted $se_{rob}(b_j)$, confidence intervals and hypothesis tests are based on

$$t = \frac{b_j - \beta_j}{se_{rob}(b_j)}.$$

The distribution of $t$ is approximated by the $T(n-k)$ distribution in most cases. For clustered errors, defined below, one instead uses the $T(G-1)$ distribution where $G$ is the number of clusters and $G$ must be large, rather than the usual requirement that $n$ is large.

For tests on **several parameters** with robust standard errors the formulas for the $F$-statistic given in the preceding section no longer hold – the $F$-statistic is no longer simply determined by the residual sums of squares. The adjusted formulas for the robust $F$-statistic use matrix algebra, that is not presented here. The robust $F$-statistic is treated as being $F(q, n-k)$ distributed, or as $F(q, G-1)$ distributed in the case of clustered errors with $G$ clusters. A statistical package that has a robust option will calculate and report the appropriate $F$-statistic for joint statistical significance of the regressors and, if available, a post-estimation test command will also compute the correct $F$-statistic.

For statistical inference on the **population mean** using univariate data, one should also use robust standard errors if observations are from an autocorrelated time series or from a clustered sample. This is rarely done or mentioned in textbooks, and statistical packages do not have an obvious command to do this, but it should be done. A robust standard error for the sample mean $\bar{y}$ can be easily computed by OLS regression of $y$ on just an intercept, since then the intercept estimate $b_1 = \bar{y}$, with standard errors computed as HAC-robust or cluster-robust. Failure to make this correction can lead to large underestimation of the standard error of the sample mean. This correction is not necessary when there is heteroskedasticity as in the intercept only case heteroskedastic-robust standard errors equal default standard errors.

**Remark 141** *If assumptions 3 and 4 do not hold then the appropriate robust standard errors should be computed and used to calculate confidence intervals, $t$-statistics and $F$-statistics.*

### 14.6.2 Heteroskedastic-Robust Standard Errors

We replace assumption 3 with the assumption of heteroskedastic errors

3.($Het$) The **error for the $i^{th}$ observation has varying variance conditional on the regressors**:

$$\mathrm{Var}[u_i|x_{2i}, ..., x_{ki}] = \sigma_i^2, \quad i = 1, ..., n.$$

The OLS estimates are unchanged, while the standard errors are computed by a different method, leading to different $t$-statistics, $p$-values and confidence intervals.

For bivariate regression the formula for the heteroskedastic-robust standard error of the OLS slope coefficient was given in Chapter 10.7. It requires that the sample size goes to infinity.

Table 14.5 presents results for the house price example when heteroskedastic-robust standard errors are used rather than default standard errors. The estimated coefficients are the

Table 14.5: House price data: OLS regression with heteroskedastic-robust standard errors.

| **Variable** | **Coefficient** | **Standard Error** | **t statistic** | **p value** | **95% conf. int.** | |
|---|---|---|---|---|---|---|
| *Size* | 68.37 | 15.36 | 4.44 | 0.000 | 36.52 | 100.22 |
| *Bedrooms* | 2685 | 8286 | 0.32 | 0.749 | -14498 | 19868 |
| *Bathrooms* | 6833 | 19284 | 0.35 | 0.726 | -33159 | 46825 |
| *Lot Size* | 23020 | 5329 | 0.43 | 0.670 | -8748 | 13355 |
| *Year Built* | -833 | 763 | -1.09 | 0.287 | -2415 | 749 |
| *Age* | -2089 | 3738 | -0.56 | 0.582 | -9841 | 5664 |
| *Intercept* | 137791 | 65545 | 2.10 | 0.047 | 1856 | 273723 |
| **Summary Statistics** | | | | | | |
| n | 29 | | | | | |
| F(6,22) | 6.41 | | | | | |
| p-value for F | 0.0005 | | | | | |
| $R^2$ | 0.651 | | | | | |
| St. error | 24936 | | | | | |

same as those given in Table 14.2, while the standard errors increase or decrease by as much as 30 percent compared to the default standard errors given in Table 14.2. In this example, despite these changes the variable *Size* remains the only statistically significant regressor at significance level 5%.

The $F$-statistic for overall statistical significance of all regressors is now 6.41 with $p = 0.0005$, compared to 6.83 with $p = 0.0003$ when default standard errors are used. Again the regressors are jointly statistically significant at significance level 5%.

From analysis not presented in Table 14.5, the $F$-statistic for tests that all coefficients other than the variable *Size* have coefficient equal to zero is now 0.46 with $p = 0.8038$, compared to 0.42 with $p = 0.8320$ when default standard errors are used. Again these regressors are jointly statistically insignificant at significance level 5%.

### 14.6.3 Heteroskedasticity- and Autocorrelation-Robust Standard Errors

Assumption 4 is that all errors for different observations are independent. For time series data, we might expect that the error in one time period is correlated with the error in the preceding time period, in which case the errors are said to be **serially correlated**. For example, if $y$ given regressors is unusually high in one year, so $u > 0$, then we might expect that $y$ given regressors is also likely to be unusually high in the next year, so again $u > 0$. This situation is common for time-series data; see Chapter 20 for additional discussion.

To ensure that OLS remains consistent assumption 2 needs to be strengthened to assume that the error is also uncorrelated with past values of the regressors. As explained in Chapter

20 this means that HAC standard errors cannot be used if the lagged dependent variable ($y_{t-1}$) is a regressor.

The errors are additionally allowed to be heteroskedastic. So, replacing subscript $i$ by subscript $t$ for time-series data, assumptions 3-4 are replaced by

2.($HAC$) The **error for the** $t^{th}$ observation **has zero mean conditional on current and past regressors values**

$$\mathrm{E}[u_t|x_{2t}, x_{2t-1}, ..., x_{21}, ...x_{kt}, x_{kt-1}, ...x_{k1}] = 0, \quad t = t, ..., T.$$

3.($HAC$) The **error for the** $t^{th}$ observation **has varying variance conditional on the regressor**:

$$\mathrm{Var}[u_t|x_{2t}, ..., x_{kt}] = \sigma_t^2, \quad t = t, ..., T.$$

4.($HAC$) The **errors for observations are correlated up to** $m$ **periods apart**

$$\mathrm{Cor}[u_t, u_s|x_{2t}, ..., x_{kt}, x_{2s}, ..., x_{ks}] \begin{cases} \neq 0 & |t-s| \leq m \\ = 0 & |t-s| > m. \end{cases}$$

The resulting standard errors, denoted $se_{HAC}(\cdot)$, are called **heteroskedasticity- and autocorrelation-robust (HAC) standard errors.** The formula is presented in Chapter 20 and assumes that the sample size goes to infinity.

Econometrics packages that have a HAC option usually require the user to provide $m$, the maximum number of time periods for which the errors are autocorrelated. This depends on both $T$ and the extent of correlation in the errors. If the correlation in the errors is high then $m$ may need to be quite large; for choice of $m$ see the discussion in Chapter 10.7.

HAC standard errors are usually larger than default or heteroskedastic-robust standard errors. The intuition is that for data correlated over time, each additional observation provides less than a complete piece of new information. This information loss, which HAC standard errors control for, leads to less precise estimation and hence larger standard errors. The HAC standard errors simplify to heteroskedasticity-robust standard errors if $m = 0$.

### 14.6.4 Cluster-Robust Standard Errors

Another way that assumption 4 can be violated is if observations can be grouped into clusters where the error for observations in the same cluster are correlated with each other while errors for observations in different clusters are independent. In that case model errors are said to be **clustered**. For example, individuals may be grouped into villages, with errors for individuals in the same village being correlated, while errors for individuals in different villages are independent. See Chapter 19.3, and also Chapter 21.2 on panel data, for additional discussion.

To ensure that OLS remains consistent assumption 2 needs to be strengthened to assume that the error is also uncorrelated with the regressors of other observations in the cluster.

We again also allow for the errors to be heteroskedastic. So assumptions 3-4 are replaced by

2.($Clu$) The **error for the** $t^{th}$ observation **has zero mean conditional on all regressors in the same cluster:**

$$\mathrm{E}[u_i|x_{2i},...,x_{ki},x_{2j},...,x_{kj}] = 0, \quad i \text{ and } j \text{ in same cluster.}$$

3.($Clu$) The **error for the** $i^{th}$ observation **has varying variance conditional on the regressor**:

$$\mathrm{Var}[u_i|x_{2i},...,x_{ki}] = \sigma_i^2, \quad i = 1,...,N.$$

4.($Clu$) The **errors for observations in the same cluster are correlated:**

$$\mathrm{Cor}[u_i,u_j|x_{2i},...,x_{ki},x_{2j},...,x_{kj}] \begin{cases} \neq 0 & i \text{ and } j \text{ in same cluster} \\ = 0 & i \text{ and } j \text{ in different clusters.} \end{cases}$$

The resulting standard errors, presented in Chapter 19.3 and denoted $se_{clu}(b_2)$, are called **cluster-robust standard errors**, though more precisely they are both heteroskedasticity-robust and cluster-robust. The cluster-robust standard errors simplify to heteroskedasticity-robust standard errors if there is just one observation per cluster.

Cluster-robust standard errors are usually larger than default or heteroskedastic-robust standard errors. This difference can be exceptionally large, so it can be especially important that cluster-robust standard errors be used if the model errors are clustered; see Chapter 19.3. The intuition is that for data correlated within cluster, each additional observation in a cluster provides less than a complete piece of new information. This information loss, which cluster-robust standard errors control for, leads to less precise estimation and hence larger standard errors.

Econometrics packages that have a cluster-robust option require the user to provide a variable that identifies which of the $G$ clusters each observation falls in. In this case it is required that the number of clusters is large, not just that the number of observations is large, as the theory assumes $G \to \infty$. The distribution of the $t$-statistic is approximated by the $T(G-1)$ distribution, and $F$ tests should use the $F(1,G-1)$ distribution.

### 14.6.5 When to Use Robust Standard Errors

When model errors are independent, so assumption 4 holds, it is now standard in econometrics practice to use heteroskedastic-robust standard errors rather than default standard

errors, even if it might be felt reasonable to assume that errors are homoskedastic. Model errors are often independent, though not always independent, when cross-section data is used. And in some time series applications with financial returns data, model errors may be independent (see Chapter 11.1).

HAC standard errors need to be used in most, though not all, time series regressions; see Chapter 20. HAC standard errors should not be used if the lagged dependent variable is a regressor. HAC standard errors may not be necessary for some financial returns regressions.

Clustered standard errors are commonly-used in two settings. One is cross-section data where individuals are grouped, for example into villages, and model errors are correlated within group; see Chapter 19.3. A second setting is panel data on individuals (people or firms or county) over time, where errors are independent across individuals but are autocorrelated over time for a given individual; see Chapter 21.2.

The last column of Table 14.5 provides a brief summary of these points.

## 14.7 More Precise Estimation

When assumptions 3-4 are relaxed the OLS is no longer the best linear unbiased estimator (BLUE), so more precise estimation than OLS is possible. The best estimator then varies with how assumptions 3-4 are relaxed.

### 14.7.1 Feasible Generalized Least Squares Estimator

More precise estimates of the model parameters may be obtained if instead of OLS we use the **feasible generalized least squares (FGLS) estimator**. In particular if we can propose a model for the error variances (and error covariances if relevant) that is correctly specified, and we can consistently estimate the parameters of this model, then the feasible GLS estimator is BLUE and hence is more precise than OLS.

Different feasible GLS estimators are used for different types of data that have different types of model error variances and correlations. For cross-section data with independent heteroskedastic errors the FGLS estimatoion method is weighted least squares; see Appendix 19.A. For time series data the most commonly-used FGLS estimator is the Cochrane-Orcutt estimator; see Appendix 20.A. For cross-section data with clustered errors and short panel data the most commonly-used FGLS estimator is the random effects estimator; see Chapters 19.3 and 21.3.

**Remark 142** *If assumptions 3-4 do not hold then feasible generalized least squares estimation may lead to more precise estimates than OLS regression. Different FGLS estimators are used in different settings.*

### 14.7.2 Robust Standard Errors for FGLS

FGLS is guaranteed to be more efficient than OLS only if the model for the error variances (and error covariances if relevant) is correctly specified. If instead this model is incorrectly specified then FGLS remains consistent in the case of weighted least squares and random effects estimation, but not Cochrane-Orcutt. In the first two cases one should use appropriate robust standard errors after GLS estimation to ensure valid inference if the model for the errors is misspecified. Then it is no longer guaranteed that FGLS is fully efficient, but FGLS is still likely to be more efficient than OLS.

In practice OLS, with appropriate robust standard errors, is used more often than FGLS. This is partly because the efficiency gains to FGLS can be small in some applications, notably for weighted least squares. But in other applications there can be considerable efficiency gains to using FGLS.

## 14.8 Nonrandom Samples

Samples used in regression are not always simple random samples. Instead, some units in the population may be oversampled or undersampled. Then, as discussed in Chapter 5.7, the population mean of a single variable needs to be estimated using a weighted mean, rather than an unweighted mean.

In this section the regression case is considered. An important distinction is whether the nonrandom sampling is on variables that are regressors or is on the dependent variable.

### 14.8.1 Nonrandom Sampling on Regressors

If the nonrandom sampling is solely on variables that are regressors then OLS can still be used, provided that enough regressors are included in the regression model and the model is sufficiently flexible so that the remaining error in the model has zero mean conditional on the regressors. Then model assumptions 1-2 are still satisfied.

For example, for the house price dataset small houses will be oversampled, since small houses are sold more frequently than large houses. As a result, the sample mean price understates the population mean price for all houses in the community. By contrast, the regression model for house price includes house size as a regressor, thereby controlling for the unrepresentative sample.

More specifically, the sample mean corresponds to the regression model $y = \beta_1 + u$ and in this example we feel $\mathrm{E}[u] < 0$, so assumption 2 does not hold, because our sample oversamples smaller homes that will on average sell for less. A weighted mean needs to be used. But in the regression model $y = \beta_1 + \beta_2 \times \mathit{Size} + u$ it is reasonable to assume that $\mathrm{E}[u|\mathit{Size}] = 0$, so now assumption 2 holds and OLS can be used.

As a second example, consider an earnings regression based on data from a complex survey such as the American Community Survey. If smaller states and minority populations are deliberately oversampled, then this will also lead to oversampling of people with low earnings, since earnings tend to be lower in smaller states and for minorities. So we should use a weighted mean in estimating population mean earnings. For regression analysis, however, we can include sufficient controls for minority status and state of residence and estimate using regular OLS.

Economic studies using survey data often include many regressors, in addition to the regressor(s) of intrinsic interest, to yield a model that satisfies

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$
$$\mathrm{E}[u|x_2, ..., x_k] = 0,$$

when nonrandom sampling is solely on the regressors. Then estimation can be by OLS.

**Remark 143** *For nonrepresentative samples with nonrandom sampling on variables other than the dependent variable, OLS estimation is possible if appropriate regressors are included as controls.*

### 14.8.2 Nonrandom Sampling on the Dependent Variable

If the nonrandom sampling is on variables that are dependent variables then OLS is inconsistent and cannot be used.

For example, suppose high-priced houses are over-sampled and we estimate the model $y = \beta_1 + \beta_2 \times$ *Size*$+u$. Then a high price arises because size is large, given $\beta_2 > 0$. But it can also arise due to unusually large values of the error term. But this implies that $\mathrm{E}[u|Size] > 0$ so Assumption 2 does not hold.

As a second example, consider an earnings regression when people with low earnings are deliberately oversampled. y Then people with negative values of the error term will have been over-sampled, so $\mathrm{E}[u|x_2, ..., x_k] < 0$.

Sampling on the dependent variable, rather than just the regressor variables, is called **sample selection**. This is much more problematic as the nonrandom sampling is partly on the error term, and no data are available for the error term. Models for sample selection exist, but they generally rely on strong assumptions about the error process and are inconsistent if any of these strong assumptions do not hold.

**Remark 144** *For nonrepresentative samples with nonrandom sampling on the dependent variable, OLS cannot be used.*

In some cases population weights are available. In that case, considered next, analysis is straightforward.

### 14.8.3 Weighted Least Squares

It is not always possible to ensure that $\mathrm{E}[u|x_2, ..., x_k] = 0$ when the sample is not representative of the population. One example is nonrandom sampling on the dependent variable. A second example is nonrandom sampling on variables other than the dependent variable, but data on these variables is not available. In the house price example it may be that houses that were substantially remodelled in the past ten years were oversampled, but this variable is not known to the data analyst.

If $\mathrm{E}[u|x_2, ..., x_k] \neq 0$, unbiased estimation is still possible if sample weights are available. Let $\pi_i$ denote the **probability** that the $i^{th}$ observation is included in the sample, where $\pi_i \neq 1/n$ if the sample is not purely random. Higher values of $\pi_i$ reflect oversampling. Oversampled observations are then downweighted by the **sample weight** $w_i = 1/\pi_i$. The **weighted least squares estimates** minimize the weighted sum of squares

$$\sum\nolimits_{i=1}^{n} w_i(y_i - b_1 - b_2x_{2i} - \cdots - b_kx_{ki})^2.$$

This downweights the squared residual or oversampled observations by $w_i = 1/\pi_i$.

Statistical packages that have weighted least squares as an option will compute the correct standard errors, $t$-statistics, $p$-values and confidence intervals. But care is needed in providing the correct weights as, for example, it is easy to specify that $w_i = \pi_i$ rather than $1/\pi_i$.

**Remark 145** *For nonrepresentative samples with sample weights provided, population weighted least squares can be used.*

## 14.9 Prediction

The issues detailed in Chapter 10.9 for prediction from bivariate regression carry over to multiple regression. In particular, it can be much easier to predict average behavior (the conditional mean) than behavior of any given individual.

Multiple regression offers the potential to greatly improve the precision of prediction, since it controls for additional variables.

### 14.9.1 Prediction of Conditional Mean

As in the bivariate case interest may lie in predicting the **conditional mean** of $y$ given specified values of the regressors:

$$\mathrm{E}[y|x_2^*, ..., x_k^*] = \beta_1 + \beta_2x^* + \cdots + \beta_kx_k^*.$$

Then the **prediction of the conditional mean** is

$$\widehat{y}_{CM} = b_1 + b_2x^* + \cdots + b_kx_k^*.$$

This prediction is unbiased under assumptions 1-2.

In the multivariate case there is no simple way to express the formula for $se(\widehat{y}_{CM})$, the standard error of $\widehat{y}_{CM}$, even in the simplest case where default standard errors are used. A $100(1-\alpha)\%$ confidence interval for the conditional mean is

$$\mathrm{E}[y|x_2^*, ..., x_k^*] \in \widehat{y}_{CM} \pm t_{n-k,\alpha/2} \times se(\widehat{y}_{CM}).$$

With many observations the regression parameters can be precisely estimated, leading to $se(\widehat{y}_{CM})$ being potentially quite small and hence quite narrow confidence intervals for the conditional mean when the sample size is large.

### 14.9.2 Prediction of Actual Value

Alternatively interest may lie in predicting the **actual value** of $y$ given specified values of the regressors:

$$y|x_2^*, ..., x_k^* = \beta_1 + \beta_2 x^* + \cdots + \beta_k x_k^* + u,$$

which has an additional term, the error term about which little is known aside from its mean of zero given assumption 2. The **prediction of the actual value**, or **forecast**, is

$$\widehat{y}_f = b_1 + b_2 x^* + \cdots + b_k x_k^*.$$

While $\widehat{y}_f = \widehat{y}_{CM}$ it should be clear that it is being used for a different and more difficult purpose of predicting the actual value.

Under assumptions 1-4

$$se(\widehat{y}_f) = \sqrt{se(\widehat{y}_{CM}) + s_e^2},$$

where $s_e$ is the standard error of the regression. (Under assumptions 1-2 we need to replace $s_e^2$ by an estimate of $\mathrm{Var}[u|x_2^*, ..., x_k^*]$). A $100(1-\alpha)\%$ confidence interval for the actual value is then

$$y|(x_2^*, ..., x_k^*) \in \widehat{y}_f \pm t_{n-k,\alpha/2} \times se(\widehat{y}_f).$$

With a large amount of data the regression parameters can be precisely estimated, leading to $se(\widehat{y}_{CM})$ being potentially quite small. But $se(\widehat{y}_f) > s_e$ always, and a 95% confidence interval for the actual value of $y$ is at least $\widehat{y}_f \pm 1.96 \times s_e$.

Statistical packages differ in the extent to which they enable statistical inference for prediction. The key requirement is that there be a way to compute $se(\widehat{y}_{CM})$. Under assumptions 1-4, given $se(\widehat{y}_{CM})$ and $s_e^2$ we can easily compute $se(\widehat{y}_f)$ and manually implement confidence intervals and hypothesis tests on the actual value of $y$. Under assumptions 1-2 we need a robust version of $se(\widehat{y}_{CM})$ and an estimate of $\mathrm{Var}[u|x_2^*, ..., x_k^*]$.

### 14.9.3 Example: House Price

Suppose we wish to make predictions for a 2000 square foot house with medium lot size, four bedrooms and two bathrooms that is forty-years old and is sold in June.

Then the predicted value is

$$\widehat{y} = b_1 + 2000b_2 + 2b_3 + 4b_4 + 2b_5 + 40b_6 + 6b_7 = 257691.$$

First, consider prediction of the conditional mean. Using statistical software that includes commands for prediction after OLS regression, we find that $se(\widehat{y}_{CM}) = 648$, using default standard errors. Since $t_{22,.025} = 2.074$, the 95% confidence interval for the conditional mean house price, given the regressor values stated above, ranges from \$244,235 to \$271,146.

If instead heteroskedastic-robust standard errors are used then $se(\widehat{y}_{CM}) = 6631$ and the confidence interval is \$243,939 to \$271,442.

Second, consider forecasting the actual value assuming assumptions 1-4 hold. Here $s_e = 24936$, so $se(\widehat{y}_f) = \sqrt{6488^2 + 24936^2} = 25766$. The 95% confidence interval for the actual house price, given the regressor values stated above, ranges from \$204,255 to \$311,126, a much broader range.

If instead heteroskedasticity is a concern we additionally need an estimate of $\text{Var}[u|x^*, ..., x_k^*]$. It is simplest to again use $s_e^2 = 24936^2$. Then $se(\widehat{y}_f) = \sqrt{6631^2 + 24936^2} = 25803$.

### 14.9.4 Data Science and Big Data

There has been a recent explosion in the availability of enormous amounts of data and the computer power to store and analyze these data. Economic examples include shopper scanner data, web click data, real-time financial trading data, and Google search terms.

This has led to terms such as the following. **Data science** or **data analytics** is the science of discerning patterns in data. **Machine learning** is a branch of artificial intelligence that attempts to algorithmically learn from data. In the popular media it is common to refer to almost any data analysis as big data, but to data scientists the term **big data** refers to datasets that are enormously large, so large that the commonly-used software tools such as the standard statistical and econometric packages are inadequate and more specialized tools are needed.

Often the goal of big data is **prediction**. Large amounts of data can lead to improved prediction as it enables estimation of models with many explanatory variables and models that can be very complex. And with large amounts of data not all the data needs to be used in determining the model; some of the data can be set aside for out-of-sample testing.

The use of big data methods to predict is ubiquitous. For example, the first few letters typed in a Google search automatically lead to several possible completions of the search; these are instantaneous predictions.

There are many commercial applications where improved forecasts can be very profitable. Often these are settings where one is repeatedly making predictions; an extreme example is high frequency trading. In that case a profit can be made by on average predicting better than the competition. Similarly web browsing history is used to predict what banner ads an individual is exposed to.

But one should not lose sight of the fact that any one individual prediction may still be very imprecise. For example, with more data and richer models a major league sporting team may be better able to predict **on average** how players in the draft may perform once selected. And there is great benefit to predicting on average better than do other teams. But there will still be considerable imprecision in the forecast for any given player.

Predicting whether a baseball player should be drafted appears to be a positive and benign use of data. But similar tools can be used to predict, for example, the probability that a prisoner will re-offend upon release from prison, given observable prisoner characteristics such as neighborhood they will return to and closeness to family. A good prediction model obtained by an algorithmic machine learning method will be a black box that is impossible to explain. The model may predict average behavior well but will most likely have low R-squared, so that individual behavior is poorly predicted. Yet if the model predicts that a given individual has a relatively high probability of re-offence then early release is ruled out and the person may remain in prison for many more years.

As another example, an algorithm may determine whether or not someone receives an expensive life-saving medical treatment that is the only available treatment for their medical condition.

Is society willing to allow such major decisions to be made by a black box? The era of data science opens up many such societal questions.

## 14.10 Key Concepts

1. The methods of statistical significance are generally similar to that in the bivariate case, except the population model includes additional regressors and conditioning is on these additional regressors.

2. It is assumed that there is sufficient variation in the regressor values across observations to enable computing the OLS coefficients.

3. For linear regression $\mathrm{E}[y|x] = \beta_1 + \beta_2 x$.

4. Population Assumptions 1-4 now condition on $x_2, ..., x_k$ and not just $x$.

5. Under assumptions 1 and 2, $b_1, ..., b_k$ are unbiased and, with additional assumptions such as assumptions 3 and 4, consistent estimates of $\beta_1, ..., \beta_k$.

6. Under assumptions 1-4, least squares is the best linear unbiased estimator. If additionally the errors are normally distributed then least squares is the best estimator.

7. The slope coefficient standard error $se(b_j)$ is smaller the better the regression line fits the data, the larger the sample, and the greater the variability of the regressors in the sample, after controlling for the other regressors.

8. Statistical inference on $\beta_j$ is based on the $t = (\beta_j - b_j)/se(b_j)$ that is treated as being $T(n-k)$ distributed, approximately.

9. Confidence intervals for $\beta_j$ and hypothesis tests on $\beta_j$ are similar to the bivariate case except the $T(n-k)$ distribution is used.

10. Joint hypothesis tests on several parameters can be implemented using an $F$ test.

11. The general formula is $F = [(RSS_r - RSS_u)/q]/RSS_u/(n-k)]$ where $RSS$ is the residual sum of squares, $q$ is the number of restrictions, and subscripts $r$ and $u$ denote, respectively, the restricted and unrestricted models.

12. $F$ is $F(q, n-k)$ distributed under the null hypothesis that the restricted model is correct.

13. A test of overall significance is a test of whether $\beta_2 = 0, ..., \beta_k = 0$.

14. A test of subset of regressors is a test of whether a subset of $\beta_2, ..., \beta_k$ is equal to zero.

15. Robust standard errors are used when assumptions 3-4 do not hold.

16. Heteroskedasticity-robust standard errors are used when model errors are heteroskedastic and independent (assumption 3 does not hold).

17. Heteroskedasticity- and autocorrelation-consistent (HAC) robust standard errors are used when model errors are heteroskedastic and correlated over time (assumptions 3 and 4 do not hold).

18. Cluster robust standard errors are used when model errors are heteroskedastic, correlated within cluster and uncorrelated across clusters (assumptions 3 and 4 do not hold).

19. Generalized least squares estimators may be more precise than OLS when assumptions 3-4 do not hold.

20. For nonrepresentative samples with nonrandom sampling on variables other than the dependent variable, OLS estimation is possible if appropriate regressors are included as controls.

21. For nonrepresentative samples with nonrandom sampling on the dependent variable, OLS cannot be used.

22. For nonrepresentative samples with sample weights provided, population weighted least squares can be used.

23. As in the bivariate case distinction needs to be made between prediction of the conditional mean and prediction of the actual value, or forecasting. The latter is more imprecise.

24. Key terms: population model; error term; disturbance term; assumptions 1-4; standard error of the regression coefficient; standard error of the regression; $t$-statistic; $t$ distribution; degrees of freedom; parameter; unbiased; best unbiased; best linear unbiased; confidence interval; confidence region; null hypothesis; alternative hypothesis; one-sided test; two-sided test; rejection; $p$-value; critical value; critical region; $F$ distribution; $F$ test; test of overall significance; test of subsets of regressors; robust standard errors; heteroskedastic-robust standard errors; HAC standard errors; cluster-robust standard errors; feasible generalized least squares; nonrepresentative sample; sample selection; sample weights; weighted least squares; prediction of conditional mean; prediction of actual value.

## 14.11 Exercises

1. Consider tests of overall statistical significance. Then $F = [\textit{Explained SS}/(k-1)]/[\textit{Residual SS}/(n-k)]$. Show that this implies $F = [R^2/(k-1)]/[(1-R^2)/(n-k)]$. Hint: Divide numerator and denominator by *Total SS* and use the definitions of $R^2$.

## 14.12 Appendix 14.A: OLS with Matrix Algebra

This section presumes familiarity with vectors and matrices, taught in a linear algebra course.

The preceding results considered only one regressor (and no intercept) for notational simplicity. When there are several regressors, usually the case, it is no longer possible to express results simply using summation notation. Instead matrix notation is used. For completeness we present the matrix results.

Let the $i^{th}$ equation be

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i.$$

In vector notation this can be written as

$$y_i = \begin{bmatrix} 1 & x_{2i} & \cdots & x_{ki} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + u_i.$$

Stacking all $n$ equations for the $n$ observations into vectors and matrices yields

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2i} & \cdots & x_{ki} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix},$$

or

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k)}{\mathbf{X}} \underset{(k\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{u}},$$

defining the $n \times 1$ vectors $\mathbf{y}$ and $\mathbf{u}$, the $n \times k$ matrix $\mathbf{X}$, and the $k \times 1$ vector $\boldsymbol{\beta}$.

The OLS estimator minimizes the sum of squared residuals $\sum_{i=1}^{n} u_i^2 = \mathbf{u}'\mathbf{u}$, since

$$\mathbf{u}'\mathbf{u} = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = u_1^2 + \cdots u_n^2.$$

It can be shown that the OLS estimator is the solution to the so-called normal equations $\mathbf{X}'\mathbf{u} = \mathbf{0}$ or, equivalently,

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}.$$

Solving for $\boldsymbol{\beta}$ yields the following formula for the OLS estimator:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{b}$ is a $k \times 1$ vector with entries $b_1, b_2, ..., b_k$. This step requires that the inverse of $\mathbf{X}'\mathbf{X}$ exists, which requires that the regressor matrix $\mathbf{X}$ has full column rank $k$.

Under assumptions 1-4 given in Chapter 14.1, including the assumption of independent homoskedastic errors,

$$\mathrm{Var}[\mathbf{b}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1},$$

which is estimated by

$$\widehat{\mathrm{Var}}[\mathbf{b}] = s_e^2(\mathbf{X}'\mathbf{X})^{-1},$$

where $s_e^2 = \frac{1}{n-k}\sum_{i=1}^n \widehat{u}_i^2 = \frac{1}{n-k}\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ where $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}$.

As an example, we apply these formulas to the data of Appendix 8.A. There are five sample observations with $y_1 = 1$, $y_2 = 2$, $y_3 = 2$, $y_4 = 2$ and $y_5 = 3$. The regressors are an intercept, equal to 1 for all five observations, and a variable that takes values of, respectively, 1, 2, 3, 4, and 5. Stacking these data yields

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 3 \end{bmatrix}; \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}.$$

It follows that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}.$$

Matrix inversion yields

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}^{-1} = \frac{1}{275-225}\begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix}.$$

Also

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 10 \\ 34 \end{bmatrix}.$$

Combining,

$$\mathbf{b} = \begin{bmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{10}\begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix}\begin{bmatrix} 10 \\ 34 \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 8 \\ 4 \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.4 \end{bmatrix},$$

so the intercept is 0.8, the slope coefficient is 0.4, and the predicted value is $\widehat{y}_i = 0.8 + 0.4x_i$.

Some algebra, not given, yields $\sum_{i=1}^n \widehat{u}_i^2 = 0.4$, so $s_e^2 = 0.4/3$. It follows that

$$\widehat{\mathrm{Var}}[\mathbf{b}] = s_e^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{0.4}{3} \times \frac{1}{10}\begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} \frac{.44}{3} & -.04 \\ -.04 & \frac{.4}{30} \end{bmatrix}.$$

In particular, taking the square root of the $(2,2)$ element of $\widehat{\mathrm{Var}}[\mathbf{b}]$, the standard error of the slope coefficient equals $\sqrt{.4/30} = 0.11547$.

# Chapter 15

# © A. Colin Cameron: Multivariate Case Studies

This chapter presents three case studies. These illustrates the multiple regression methods presented in Chapters 13 and 14, as well as the earlier univariate analysis and bivariate regression. The second example, on the Cobb Douglas production function uses regression with natural logarithms, for which the material in Chapter 12 provides sufficient background.

## 15.1 School Academic Performance

California schools are assessed using the Academic Performance Index (API), introduced in 1999. The API is a score for each school that is determined by the performance of all students at the school in a standardized test, the STAR test, that varies by grade level. The lowest possible API score is 200 and the highest possible score is 1000. The goal is for schools to attain a score of at least 800.

Dataset API99 has data for 807 high schools in California in 1999 on the API and on factors that may affect school performance on the API. The data are summarized in Table 15.1. The variables *Edparent*, *Meals* and *Englearn* measure the socioeconomic background of students – variable *Meals* gives the percentage of students in a free or reduced price lunch program which is available only to students from poorer families. The variables *Credteach* and *Emerteach* measure the extent to which teachers have passed state teacher credentialing requirements and may be a crude measure of teacher quality.

### 15.1.1 Univariate Analysis

The first panel of Figure 15.1 gives a histogram for *Api*, along with the kernel density estimate. *Api* appears to be normally distributed. Often scaled scores such as the API are

Table 15.1: Academic Performance Index data: Variable definitions and summary statistics (n=807)

| Variable | Definition | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|
| *Api* | Academic Performance Index | 620.94 | 107.44 | 355 | 966 |
| *Edparent* | Average years of schooling of parents | 12.84 | 1.23 | 9.62 | 16 |
| *Meals* | % of students in lunch program | 21.92 | 23.67 | 0 | 98 |
| *Englearn* | % of students English learners | 14.00 | 12.79 | 0 | 66 |
| *Yearround* | = 1 if multi-track year-round school | 0.02 | 0.15 | 0 | 1 |
| *Credteach* | % of teachers with full credentials | 89.84 | 8.44 | 33 | 100 |
| *Emerteach* | % of teachers with emergency credentials | 10.47 | 8.21 | 0 | 56 |

constructed in such a way as to follow the normal distribution bell curve. As an aside we pursue the question of normality further.

The second panel of Figure 15.1 gives both the kernel density estimate and a normal density with mean and standard deviation equal to the sample mean and standard deviation (620 and 107). The data do appear to be quite close to being normally distributed. In output not given, the symmetry statistic for *Api* is 0.121, close to 0, the median is 620, close to the mean of 620.9, and the kurtosis statistic is 2.61, close to 3 for the normal. Finally, if *Api* is normally distributed then we expect 5% of the sample, or 40 observations, to be more than 1.96 standard deviations from the mean. In fact 35 observations lie outside this range.

The standard error for the sample mean is $107.44/\sqrt{807} = 3.78$. A 95 percent confidence interval for the population mean school API is $620.94 \pm 1.96 \times 3.78 = (613, 638)$.

### 15.1.2 Bivariate Regression

Student performance is known to be highly correlated with the educational attainment of their parents.

OLS regression of *Api* on *Edparent* yields estimates

$$\widehat{Api} = \underset{(16.08)}{-400.31} + \underset{(1.25)}{79.53} \times \ Edparent, \ s_e = 43.674, \ R^2 = 0.835, \ \bar{R}^2 = 0.834,$$

where default standard errors are given in parentheses.

*Edparent* is highly statistically significant, with $t = 79.53/1.246 = 63.82$. Furthermore it is economically very significant. A one year increase in average parental schooling is associated with an 80 point increase in *Api*, which is a three-quarter of a standard deviation in *Api* ($= 79.53/107.44$).

Figure 15.2 presents a scatter plot of *Api* against *Edparent* and the fitted regression line. This confirms visually the strength of the relationship, with $R^2 = 0.83$. A fitted

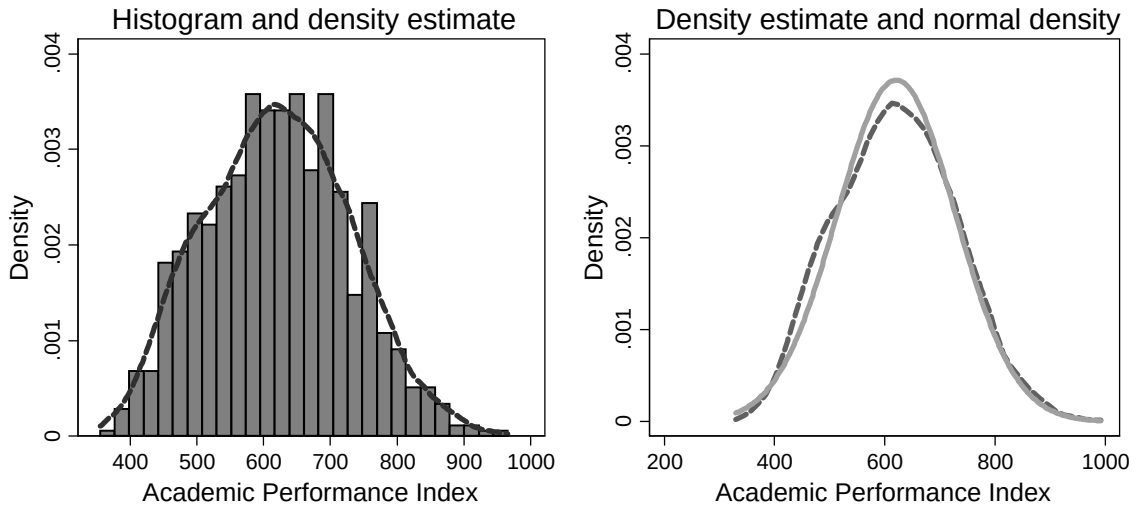


Figure 15.1: Academic Performance Index: Histogram and Density

nonparametric regression, obtained using local linear regression, is almost indistinguishable from the OLS fitted line, confirming that the relationship is linear.

Deviations of sample observations from the fitted line appear unrelated to the level of the regressor *Edparent*, so there is no real sign of heteroskedasticity. And in fact the heteroskedastic-robust standard error for $b_{Edparent}$ is 1.22, very close to the default standard error of 1.25.

### 15.1.3 Multiple Regression

It is possible that parent education level is simply picking up the effect of other variables associated with parent education level, including other socioeconomic variables and school input variables such as teacher quality.

Table 15.2 presents pairwise correlations for the data. The asterisk indicates statistical significance at significance level 0.05 – here all correlations are statistically significantly different from zero.

From the first column of Table 15.2, school API is positively correlated with parents' education, negatively associated with children who receive free and reduced price lunches or are English learners, positively associated with teachers having full credentials and negatively correlated with teachers having emergency credentials. All these correlations have the expected sign and are at least of moderate size, aside from variable *Yearround* with

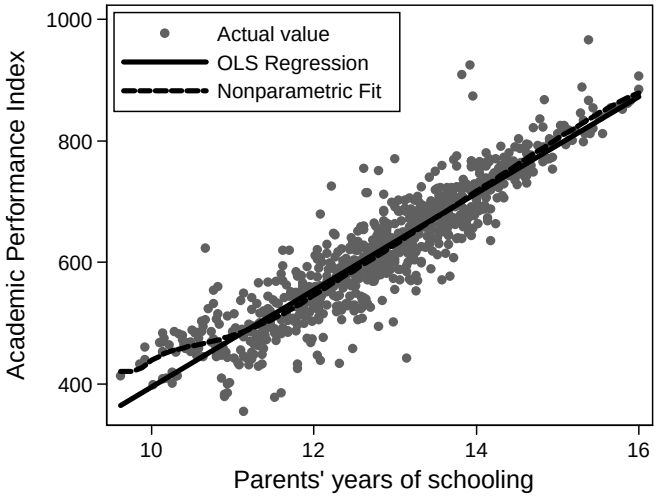


Figure 15.2: Academic Perfomance Index: Regression on Parents' Education

correlation 0.19.

The second column of Table 15.2 indicates that parents' education is reasonably strongly correlated with the other socioeconomics and school teacher variables. Perhaps the strong association of parents' education with API is merely picking up the effect of these other variables.

Table 15.3 lists results from multiple regression of *Api* on *Edparent* and the other variables. Even including the other variables, *Edparent* has a large impact, with coefficient lowered only slightly from 79.5 in bivariate regression to 73.9. There is a modest improvement in adjusted $R^2$ from .834 to .852. The regressors are jointly highly statistically significant since the overall $F$ test $p = 0.000$.

Table 15.2: Academic Performance Index data: pairwise correlations

| | *Api* | *Edparent* | *Meals* | *Englearn* | *Yrrd* | *Cred* | *Emer* |
|---|---|---|---|---|---|---|---|
| *Api* | 1 | | | | | | |
| *Edparent* | .91* | 1 | | | | | |
| *Meals* | -.54* | -.60* | 1 | | | | |
| *Englearn* | -.66* | -.71* | .56* | 1 | | | |
| *Yearround* | -.19* | -.25* | .29* | .22* | 1 | | |
| *Credteach* | .46* | .40* | -.27* | -.26* | -.18* | 1 | |
| *Emerteach* | -.45* | -.37* | .22* | .20* | .09* | -.82* | 1 |

Table 15.3: Academic Performance Index: Results from multiple regression.

| ANOVA Table | | | | | | |
|---|---|---|---|---|---|---|
| Source | SS | df | MS | F | p-value | |
| Explained | 7932923 | 6 | 1322154 | 771.4 | 0.000 | |
| Residual | 1371154 | 800 | 1714 | | | |
| Total | 9304076 | 806 | | | | |
| Dependent Variable *Price* | | | | | | |
| Variable | Coefficient | Standard Error | t statistic | p value | 95% conf. int. | |
| *Edparent* | 73.942 | 1.883 | 39.27 | 0.000 | 70.246 | 77.638 |
| *Meals* | 0.079 | 0.081 | 0.98 | 0.327 | -0.080 | 0.238 |
| *Englearn* | -0.358 | 0.167 | -2.14 | 0.032 | -0.685 | -0.030 |
| *Yearround* | 25.956 | 10.185 | 2.55 | 0.011 | 5.963 | 45.949 |
| *Credteach* | 0.287 | 0.311 | 1.25 | 0.213 | -0.222 | 0.997 |
| *Emerteach* | -1.470 | 0.315 | -4.67 | 0.000 | -2.088 | -0.853 |
| *Intercept* | -345.328 | 39.954 | -8.84 | 0.000 | -423.752 | -277.900 |
| Observations | 807 | | | | | |
| F(6,22) | 771.4 | | | | | |
| p-value for F | 0.000 | | | | | |
| R-squared | 0.853 | | | | | |
| Adjusted $R^2$ | 0.852 | | | | | |
| St. error of regression | 41.4 | | | | | |

The regressors *Englearn* and *Credteach* are statistically significant at 5%, and have the expected negative sign. The coefficient of *Englearn* implies that, after controlling for the other regressors, a one standard deviation change in the proportion of English learners is associated with a five point decrease ( $0.358 \times 14.00$) in API, which is a relatively small effect. Similarly a one standard deviation change in the percentage of teachers with emergency credential is associated with a modest twelve point decrease in the API.

The joint statistical significance of the additional regressors is a test of $H_0 : \beta_3 = 0, ..., \beta_7 = 0$ against $H_a$ :at least one of $\beta_3, ..., \beta_7 \neq 0$. The residual sum of squares (RSS) in the bivariate regression model test statistic is 1535468. (While RSS was not listed in the bivariate regression output, $s_e$ was listed, and we can use $RSS = (n-2) \times s_e^2 = 805 \times 43.674^2$). The remaining numbers are provided in Table 15.3. Then $F = [(1535468 - 1371154)/(7 - 2)]/1714 = 19.2$. The additional five regressors are jointly highly statistically significant since $p = \Pr[|F_{5,800}| > 19.2] = 0.000$ (or since $19.2 > F_{.05;5,800} = 2.23$).

We conclude that performance of a high school on the Academic Performance Index is very highly correlated with the average educational attainment of the parents of the school children. This association is of large economic magnitude. The association remains high

even after controlling for several variables that measure student socioeconomic background and teacher quality.

## 15.2 Cobb-Douglas Production Function

A production function models output as a function of capital and labor, and possibly additionally inputs such as land. The benchmark functional form is the Cobb-Douglas production function, the subject of this section.

Specifically let $Q$ be the number of units of output, $K$ be the number of units of capital and $L$ be the number of units of output. Then the **Cobb-Douglas production function** specifies

$$Q = \alpha K^{\beta_2} L^{\beta_3}.$$

Of particular interest is testing whether the returns to scale are constant, so that doubling both inputs leads to exactly doubling output. For the Cobb-Douglas this is the case if $\beta_2 + \beta_3 = 1$. Returns to scale are increasing if $\beta_2 + \beta_3 > 1$ and are decreasing if $\beta_2 + \beta_3 < 1$. For example, we might have $Q = 10K^{0.4}L^{0.5}$. Then returns to scale are decreasing since $0.4 + 0.5 = 0.9 < 1$.

### 15.2.1 Natural Logarithm Transformation

The model for $Q$ is nonlinear in $K$ and $L$, making multiple regression seemingly impossible. But standard regression is possible as the model for $\ln Q$ is linear in $\ln K$ and $\ln L$. Taking the natural logarithm of both sides, some algebra yields

$$\begin{aligned}
\ln Q &= \ln(AK^{\beta_2}L^{\beta_3}) \\
&= \ln A + \ln(K^{\beta_2}) + \ln(L^{\beta_3}) \\
&= \ln A + \beta_2 \ln K + \beta_3 \ln L \\
&= \beta_1 + \beta_2 \ln K + \beta_3 \ln L,
\end{aligned}$$

where $\beta_1 = \ln \alpha$. This result uses the properties of natural logarithm that $\ln(a \times b) = \ln a + \ln b$ and $\ln a^b = b \ln a$.

It follows that the Cobb-Douglas production function can be estimated by OLS regression of $\ln Q$ on $\ln K$ and $\ln L$. This is a special case of the log-log model with two regressors; see Chapter 12.5 for the bivariate case.

### 15.2.2 Example: Original Cobb-Douglas Study

Dataset COBBDOUGLAS has U.S. aggregate data on manufacturing for the 24 years from 1899 to 1922. These data come from, respectively, Tables VI, II and III of the article by

Table 15.4: Cobb-Douglas Data: Output (Q), Capital (K) nd Labor (L) for U.S. from 1899 to 1922 (n=24).

| Year | Q | K | L | Year | Q | K | L |
|---|---|---|---|---|---|---|---|
| 1899 | 100 | 100 | 100 | 1911 | 153 | 216 | 145 |
| 1900 | 101 | 107 | 105 | 1912 | 177 | 226 | 152 |
| 1901 | 112 | 114 | 110 | 1913 | 184 | 236 | 154 |
| 1902 | 122 | 122 | 118 | 1914 | 169 | 244 | 149 |
| 1903 | 124 | 131 | 123 | 1915 | 189 | 266 | 154 |
| 1904 | 122 | 138 | 116 | 1916 | 225 | 298 | 182 |
| 1905 | 143 | 149 | 125 | 1917 | 227 | 335 | 196 |
| 1906 | 152 | 163 | 133 | 1918 | 223 | 366 | 200 |
| 1907 | 151 | 176 | 138 | 1919 | 218 | 387 | 193 |
| 1908 | 126 | 185 | 121 | 1920 | 231 | 407 | 193 |
| 1909 | 155 | 198 | 140 | 1921 | 179 | 417 | 147 |
| 1910 | 159 | 208 | 144 | 1922 | 240 | 431 | 161 |

C.W. Cobb and P.H. Douglas (1928), "A Theory of Production," The American Economic Review," pages 139-165.

Table 15.3 presents the data on output ($Q$), capital ($K$) and labor ($L$) that are normalized to equal 100 in 1899.

Over this period output more than doubled, capital input quadrupled, and labor input less than doubled. The capital/output ratio ($K/Q$) increased, the labor/output ratio ($L/Q$) decreased, and the capital/labor ratio ($K/L$) increased (it more than doubled).

### 15.2.3 Regression Results

The regression results (with standard errors in parentheses) are

$$\widehat{\ln Q} = \underset{(.434)}{-.177} + \underset{(.064)}{.233} \times \ln K + \underset{(.145)}{.807} \times \ln K, \;\; s_e = 0.0581, \;\; R^2 = 0.957,$$

where default standard errors are given in parentheses.

The model fits the data very well, with high $R^2$, and the coefficients of $\ln K$ and $\ln L$ are reasonably precisely estimated and are highly statistically significant at level 0.05.

Cobb and Douglas did not estimate this model by linear regression, but instead set $\beta_2 = .25$ and $\beta_3 = .75$. The estimated coefficients, $b_2 = 0.233$ and $b_3 = 0.807$, are individually not statistically different from these values. For example, for test of $H_0 : \beta_3 = .75$, the test statistic $t = (.75 - .807)/.145 = -.393$ with $p = .698$ so the null hypothesis is not rejected at significance level $\alpha$.

### 15.2.4 Test of the Specified Parameter Values

An interesting question is whether or not the joint test of $H_0 : \beta_2 = .25$, $\beta_3 = .75$ against $H_0$ : at least one of $\beta_2 \neq .25$, $\beta_3 \neq .75$ leads to rejection of the parameter values specified by Cobb and Douglas.

This $F$ test can be immediately implemented using a specialized post estimation test command, should one be available.

Instead we implement this $F$ test manually. We need the residual sum of squares (RSS) from the restricted model. In the restricted model $\ln Q = \beta_1 + 0.25 \times \ln K + 0.75 \times \ln L + u$, so $\ln Q - 0.25 \times \ln K + 0.75 \times \ln L = \beta_1 + u$. So RSS in the restricted model can be obtained as RSS from regression of $\ln Q - 0.25 \times \ln K + 0.75 \times \ln L$ on just an intercept, yielding $RSS_r = .07168$. And, from output not listed, for the unrestricted model $RSS_u = .07098$. So $F = [(.07168 - .07098)/2]/(.07098/21) = 0.103$ and $p = \Pr[F_{2,21} > .103] = .903$. The restrictions are not rejected at significance level 0.05. The data are consistent with $\beta_2 = .25$ and $\beta_3 = .75$.

### 15.2.5 Test of Constant Returns to Scale

A second test is whether returns to scale are constant. We test $H_0 : \beta_2 + \beta_3 = 1$ against $H_a : \beta_2 + \beta_3 \neq 1$. Since $b_3 + b_3 = .233 + .807 = 1.040$ the estimates sum to close to one. The test of a single restriction can also be immediately implemented using a specialized post estimation test command, should one be available.

Instead we implement this test of constant returns manually. We need the residual sum of squares (RSS) from the restricted model. Now if $\beta_2 + \beta_3 = 1$ then $\beta_3 = 1 - \beta_2$. Making this substitution $\ln Q = \beta_1 + \beta_2 \times \ln K + (1 - \beta_2) \times \ln L + u$, so $\ln Q - \ln L = \beta_1 + \beta_2 \times (\ln K - \ln L) + u$. So RSS in the restricted model can be obtained as RSS from regression of $\ln Q - \ln L$ on an intercept and $\ln K - \ln L$, yielding $RSS_r = .07164$. And, again, $RSS_u = .07098$. So $F = [(.07164 - .07098)/1]/(.07098/21) = 0.195$ and $p = \Pr[F_{1,21} > .195] = .663$. The restrictions are not rejected at significance level 0.05. The data are consistent with constant returns to scale.

### 15.2.6 Predicted Output

In the bivariate case we found that if $\widehat{\ln y} = b_1 + b_2 x$ then we use $\ln \widehat{y} = \exp(s_e^2/2) \exp(b_1 + b_2 x)$, where $s_e$ is the standard error of the $\ln y$ regression and the additional term $\exp(s_e^2/2)$ arises due to transformation from predicting $\ln y$ to predicting $y$. The same result applies when

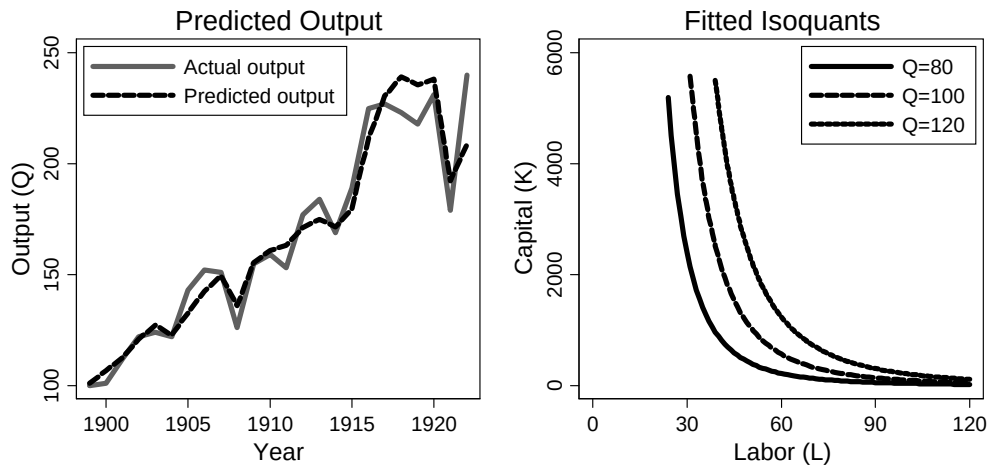


Figure 15.3: Cobb-Douglas Production Function: Predicted Output and Fitted Isoquants

additional regressors are added. It follows that

$$\begin{aligned}\widehat{Q} &= \exp(s_e^2/2) \times \exp(-.177 + .233 \ln K + .807 \times \ln K) \\ &= \exp(.0581^2/2) \times \exp(-.177) \times K^{.233} \times L^{.807} \\ &= .839 \times K^{.233} L^{.807}.\end{aligned}$$

Using the estimated coefficients yields $\widehat{Q}$ with mean 166.0, quite close to $Q$ with mean 165.9.

The first panel of Figure 15.3 plots actual $Q$ and predicted $Q$ against time. The fit is quite good, though in the final year (1922) actual output was much higher than that predicted given the 1922 levels of capital and labor.

### 15.2.7 Fitted Isoquants

An isoquant plots capital ($K$) as a function of labor ($L$) for a given level of output ($Q$).

In general the isoquants can be obtained from the Cobb-Douglas production function using

$$\begin{aligned} & Q &= \alpha K^{\beta_2} L^{\beta_3} \\ \Rightarrow\ & K^{\beta_2} &= Q/(\alpha L^{\beta_3}) \\ & &= \alpha^{-1} Q L^{-\beta_3} \\ \Rightarrow\ & K &= \alpha^{-1/\beta_2} Q^{1/\beta_2} L^{-\beta_3/\beta_2}.\end{aligned}$$

For the estimated Cobb-Douglas function this yields $K = 2.140 \times Q^{4.29} \times L^{-3.46}$. (This ignores the log transformation bias, for simplicity. The bias here is small as $\exp(.0581^2/2) = 10.0017$ is close to 1.)

The second panel of Figure 15.3 plots three isoquants with output set to $Q = 80$, $Q = 100$, and $Q = 120$, and with labor varying from $L = 30$ to $L = 120$. As expected the isoquants do not cross so that increasing both capital and labor necessarily increases output.

### 15.2.8 Robust Standard Errors

With time series data the errors may be autocorrelated, in which case inference should be based on HAC standard errors. The first three autocorrelations of the residuals are $\widehat{\rho}_1 = 0.10$, $\widehat{\rho}_2 = -0.13$, and $\widehat{\rho}_3 = -0.13$, so the residuals appear to be uncorrelated. There is no need to use HAC standard errors.

And the errors may be heteroskedastic. The heteroskedastic-robust standard errors of $b_2$ and $b_3$ are .105 and .216, about 50% higher than the default standard errors of .064 and .145. This substantial difference suggests that heteroskedastic robust standard errors should be used. In that case the $F$ tests for joint tests on $\beta_2$and $\beta_3$ cannot be based on the residual sum of squares. Instead a more specialized post-estimation test command needs to be used, one that uses heteroskedastic-robust standard errors. For test of $\beta_2 = .25$ and $\beta_3 = .75$, this yields $F = .109$ compared to $F = .103$ with default standard errors. And for test of $\beta_2 + \beta_3 = 1$, the heteroskedastic-robust $F = .119$ compared to $F = .196$ with default standard errors. In both cases we again do not reject the null hypothesis at significance level .05.

## 15.3 Phillips Curve

The **Phillips curve** is a curve that plots price inflation against unemployment. The curve was proposed in A. W. Phillips (1958), "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957", Economica, 1861-1957. Phillips found a negative relationship. The explanation is that an increase in money supply may stimulate the economy in the short-run, leading to lower unemployment accompanied by some increase in wages. While Phillips considered wage inflation it subsequently became standard to analyze the relationship between price inflation and unemployment.

The relationship observed by Phillips had a strong effect on macroeconomic policy, especially in the 1960's, as it offered the possibility of lowering unemployment at the mild expense of somewhat higher price inflation. At the same time it spawned a fierce debate as to whether this relationship held in the long-run.

### 15.3.1 Example: U.S. Price Inflation

Table 15.4 presents variable definitions and summary statistics for dataset PHILLIPS that has annual U.S. data from 1949 to 2014. We focus on inflation measured by the GDP implicit

Table 15.5: Phillips curve data: Variable definitions and summary statistics

| Variable | Definition | Obs | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|---|
| *Urate* | Civilian unemployment rate (percentage) | 66 | 5.87 | 1.63 | 2.70 | 10.80 |
| *Inflation* | Annual inflation rate in GDP implicit price deflator | 66 | 3.20 | 2.32 | -1.97 | 10.51 |
| *Expinflation* | Forecast of one-year ahead *Inflation* | 45 | 3.31 | 2.05 | 1.14 | 8.67 |
| *Pastinflation* | Average of *Inflation* over past 4 years | 63 | 3.65 | 2.04 | 1.48 | 9.37 |

price deflator. The more extended analysis uses expectations of future price inflation. Two measures are used. The first is the one-year head expectations of inflation measured by the GDP implicit price deflator. This comes from the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia, and is available only from 1970. The second is an ad hoc measure that is the weighted average of inflation over the past four years ($\dot{p}_t^e = 0.4\dot{p}_{t-1} + 0.3\dot{p}_{t-2} + 0.2\dot{p}_{t-3} + 0.1\dot{p}_{t-1}$), where $\dot{p}_t$ denotes the inflation rate in year $t$.

### 15.3.2 Bivariate Regression - Phillips Curve pre-1970

Figure 15.4 presents both a time series plot (first panel) and a scatter plot of CPI inflation and unemployment (second panel) for the period 1949 to 1969. It is clear that there is a negative relationship between the two series.

OLS regression using data from 1949 to 1969 yields

$$\widehat{Inflation} = \underset{(5.69)}{7.111} - \underset{(-3.97)}{1.030} \times Urate,\ s_e = 1.32,\ R^2 = 0.454,\ n = 21,$$

where $t$-statistics given in parentheses are based on default standard errors. There is a negative relationship between inflation and the unemployment, and the regressor is statistically significant at level 0.05 since $|t| = 3.97 > t_{.025,19} = 2.09$.

The errors in this time series model are potentially correlated over time, in which case model assumption 4 is incorrect. However the residuals are not highly correlated, as the first three autocorrelations are $\widehat{\rho}_1 = 0.11$, $\widehat{\rho}_2 = -0.18$, and $\widehat{\rho}_3 = 0.04$. When HAC robust standard errors are computed, allowing for the errors to be potentially correlated out to three years, the $t$-statistic for the slope coefficient is $-3.17$. And if heteroskedastic-robust standard errors are used this $t$-statistic is $-2.91$. In all cases *Urate* is statistically significant at significance level 0.05.

A lowering in the unemployment rate of one percentage point is associated with a 1.03 percentage point increase in annual inflation, which might be viewed as a reasonable trade-off.

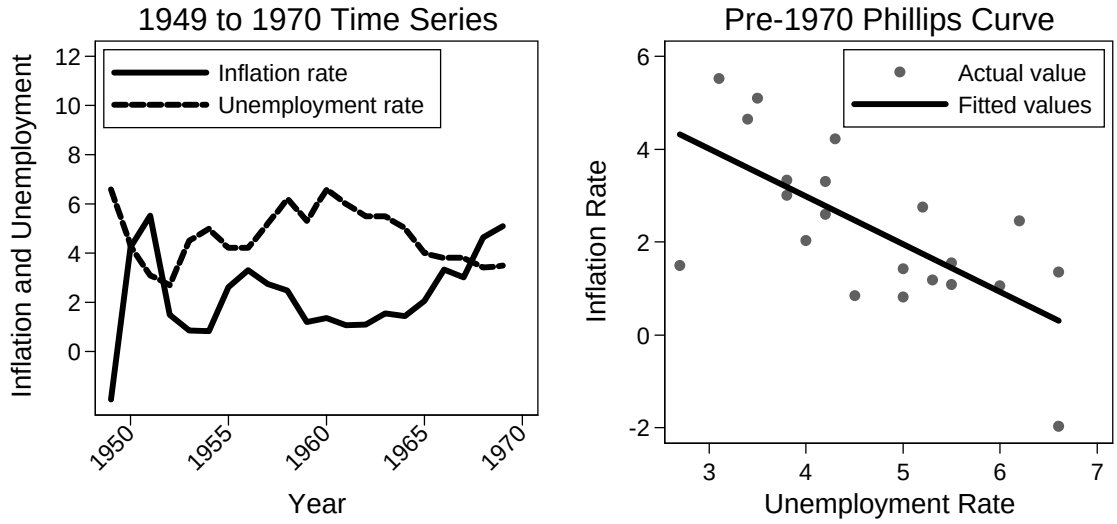


Figure 15.4: Inflation and unemployment rates from 1949 to 1969 and the associated Phillips curve.

### 15.3.3 Bivariate Regression - Phillips Curve post-1970

The pre-1970 estimates suggest that by running an expansionary monetary policy one can enjoy low unemployment at the cost of a modest increase in price inflation. The monetarist economist Milton Friedman in a 1968 paper objected to this line of reasoning, arguing that people would eventually adapt their expectations to higher price inflation so that the expansionary monetary policy would have no real effect. His predictions were borne out and led to richer models.

The first panel of Figure 15.5 presents time series plots for the period 1970 to 2014. If anything it appears that the two series often move in the same direction.

OLS regression using data from 1970 to 2014 yields

$$\widehat{Inflation} = \underset{(1.92)}{1.923} + \underset{(1.11)}{0.266} \times Urate, \; s_e = 2.44, \; R^2 = 0.258, \; n = 45,$$

where $t$-statistics given in parentheses are based on default standard errors. There is a now a positive relationship between inflation and the unemployment, though the regressor is statistically significant at level 0.05 since $|t| = 1.11 > t_{.025,43} = 2.02$.

The residuals from this model are highly correlated over time, with the first three autocorrelations being $\widehat{\rho}_1 = 0.84$, $\widehat{\rho}_2 = 0.62$, and $\widehat{\rho}_3 = 0.53$. When HAC robust standard errors are computed, allowing for the errors to be potentially correlated out to three years,

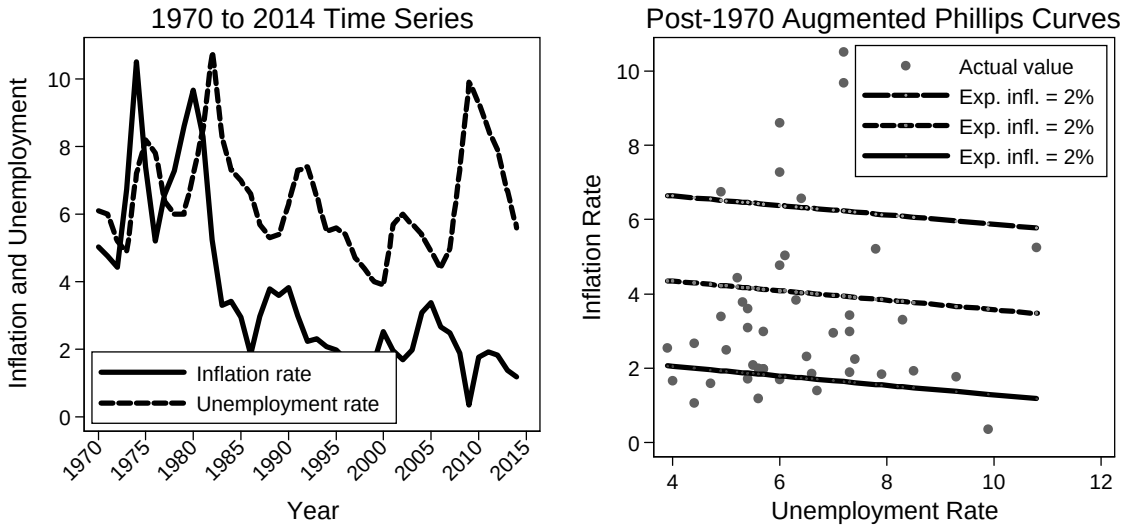


Figure 15.5: Inflation and unemployment rates from 1970 to 2014 and the associated Phillips curve.

the $t$-statistic for the slope coefficient falls to 0.72, so *Urate* becomes even more statistically insignificant.

### 15.3.4 Multiple Regression - Augmented Phillips Curve

Clearly the Phillips curve relationship broke down, beginning in the 1970's.

One possible explanation is that the stimulatory effect of an increase in money supply varies according to whether inflationary expectations are high or low. If people expect high price inflation, then an increase in money supply of 5%, say, would merely support these expectations and provide no stimulatory effect. But if people expect little price inflation, then an increase in money supply of 5% may have a major stimulatory effect with a lower increase in price inflation and a decrease in the unemployment rate.

The **augmented Phillips curve** includes an additional regressor, the **expected inflation rate**, in the regression of the inflation rate on the unemployment rate.

The practical problem is to obtain a measure of expected inflation. One way to determine this is through a survey of inflationary expectations of randomly chosen individuals or of professional business cycle forecasters. Such surveys are done, though generally they are available in more recent years. One of the oldest available measures is the series on one-year ahead expectations of inflation measured by the GDP implicit price deflator. This forecast

comes from the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia, and is available from 1970.

OLS regression using data from 1970 to 2014 yields

$$\widehat{Inflation} = \underset{(0.47)}{0.265} - \underset{(1.46)}{0.128} \times Urate + \underset{(17.36)}{1.147} \times Expinflation, \; s_e = 0.86, \; R^2 = 0.881, \; n = 45,$$

where $t$-statistics given in parentheses are based on default standard errors.

First, notice that the coefficient of expected price inflation is close to one, so there is a one-to-one relationship between expected inflation and inflation. The standard error of $b_3$ equals 0.066, so a 95% confidence interval for $b_3$ is $1.147 \pm t_{.025,42} \times .066 = (1.01, 1.28)$. This is close to containing the value one, but since it does not we marginally reject $H_0 : \beta_3 = 1$ against $H_0 : \beta_3 \neq 1$ at significance level 5% .

Second, after controlling for expected price inflation the coefficient of the unemployment rate is now negative, as in the original Phillips curve. But the coefficient is not statistically significant at 5%, as $|t| = 1.46 < t_{.025,43} = 2.02$.

The residuals have first three autocorrelations of $\widehat{\rho}_2 = 0.60$, $\widehat{\rho}_2 = 0.38$, $\widehat{\rho}_3 = 0.41$ and $\widehat{\rho}_4 = 0.35$, so we should use HAC robust standard errors. Doing so, the $t$-statistic for *Urate* is little changed but the $t$-statistic for *Expinflation* falls to 13.45, and now the 95% confidence interval is $(0.97, 1.32)$ which includes one.

The breakdown in the Phillips curve can be attributed to the omission of expected price inflation as a regressor. After controlling for inflationary expectations, there is an inverse relationship between price inflation and the unemployment rate.

### 15.3.5 Predictions

The augmented Phillip curve relationship can be represented by a series of curves, where each curve represents a different expected inflation rate. For example, for an expected inflation rate of 2.0% we have

$$\widehat{Inflation} = 0.265 - 0.128 \times Urate + 1.147 \times 2 = 2.559 - 0.128 \times Urate.$$

The second panel of Figure 15.5 plots augmented Phillips curves for expected inflation rates of 2%, 4% and 6%. It is clear that as inflationary expectations increase, the Phillips curve shifts up and out.

### 15.3.6 Omitted Variables Bias

The observed sign reversal for the coefficient of *Urate* is a classic example of omitted variables bias. The true model is

$$Inflation = \beta_1 + \beta_2 \times Urate + \beta_3 \times Expinflation + u_t.$$

Mistakenly running a regression of *Inflation* on *Urate* alone yields estimate *Inflation*= $b_1 + b_2 \times$ *Inflation*. Omitted variables bias is presented in Chapter 17, where it is stated that

$$\mathsf{E}[b_2] = \beta_2 + \beta_3\gamma,$$

where $\gamma$ is the coefficient of *Urate* in a regression of *Expinflation* on *Urate*.

From the multivariate regression we have estimates of $\beta_2$ and $\beta_3$ of $-.128$ and 1.147. Separate bivariate regression of *Expinflation* on *Urate*, not detailed here, yields a slope of .343. Then $\widehat{\mathrm{E}[b_2]} = -.128 + 1.147 \times 1.147 = 0.266$, which is the estimated coefficient of *Urate* from regression of *Inflation* on *Urate*.

## 15.4 From Raw Data to Final Data

This text, like most texts, analyzes and provides datasets that can be immediately used for data analysis.

In practice data come in a raw form. Going from this raw form to a final dataset ready for analysis can require considerable work, a task that has been recently labelled as **data carpentry.**

The first task is **reading** any sort of data into a statistical package. Easy-to-read examples include Excel spreadsheets (with extension .xls or .xlsx), a plain text file with comma-separated values or character-separated values (with extension .csv), or a data file formatted for a commonly-used statistical package. Somewhat less convenient is data that appear in a table in a PDF document (with extension .pdf). Hardcopy data may be scanned and digitized using an optical character recognition program such as that in Adobe Acrobat. Or web data may be obtained using a web scraping program.

The second task can be **combining data** that come from multiple sources. Merging data requires care to ensure that data from different sources are correctly matched to the same individual. Merging can be especially tricky if names are misspelled or have variations (such as with or without middle initial), or if the data are time series observed at different frequencies and with different formats for recording dates.

The third task is **cleaning** the data. This entails recoding data and detecting data that are in error.

This section provides a simple example. The data are downloaded from the web in a form that can be immediately read into a statistical package. There is just a single dataset. The emphasis is on then cleaning the data so that it is ready for analysis.

### 15.4.1 Downloaded Data in Fixed Format

The data come from the Michigan Panel Survey of Income Dynamics (PSID) which tracks thousands of people who are surveyed every year for many years. The PSID is a major data

Table 15.6: Survey data: Format of raw data file.

| Code | Columns | Variable Name |
|---|---|---|
| V30001 | 1-4 | 1968 INTERVIEW NUMBER |
| V30002 | 5-7 | PERSON NUMBER |
| V30809 | 8-9 | AGE OF INDIVIDUAL 93 |
| V30820 | 10-11 | G90 HIGHEST GRADE COMPLETED 93 |
| V30821 | 12-17 | TOTAL LABOR INCOME 93 |
| V30823 | 18-21 | 1992 ANNUAL WORK HOURS 93 |
| V32000 | 22 | SEX OF INDIVIDUAL |
| V32022 | 23-24 | LIVE BIRTHS TO THIS INDIVIDUAL |
| V32049 | 25 | LAST KNOWN MARITAL STATUS |

set for research by academic economists especially in labor economics and public economics.

The data are on age, number of children, education, annual hours worked, annual labor income (earnings), marital status and sex for women from aged 30 to 50 years from the PSID. These data were directly downloaded using a web interface provided by the PSID. This interface has changed since the data were downloaded, so the many steps used to obtain the data are not repeated here. The download created two files.

The data are in file psid3050.dat that has fixed width data on 4856 observations on nine variables. The first two lines of the date are

```
4 43912 7725029402 21
4 63512 1200020402 24
```

These data could be mistakenly viewed as being on just four variables, separated by spaces. Instead there are nine variables in fixed format, without separators. Especially for large datasets this is common as the size of the data file is reduced by not having separators. To read the data we need to know the formatting.

The second file, psidf3050.sas, is a SAS command file to read the data using SAS. If we are using SAS we can immediately run this program to read in file psidf3050.dat. If we are using other software, we can adapt the information in file psidf3050.sas to read in the file. For example, file psid3050.sas includes the lines

```
V30809 8-9
V30809="AGE OF INDIVIDUAL 93"
```

which means that the 8th and 9th columns in each line of file psidf3050.dat contain data on variable V30809 which is the person's age. Table 15.5 summarizes the variables in file psidf3050.dat and the file format.

Table 15.7: Summary statistics: Survey data before cleaning.

| Name | Obs | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *Intnum* | 4856 | 4958.10 | 2761.97 | 4 | 9306 |
| *Persnum* | 4856 | 59.21 | 79.75 | 1 | 205 |
| *Age* | 4856 | 38.46 | 5.59 | 30 | 50 |
| *Educatn* | 4856 | 16.38 | 18.45 | 0 | 99 |
| *Earnings* | 4856 | 14244.51 | 15985.45 | 0 | 240000 |
| *Hours* | 4856 | 1235.33 | 947.18 | 0 | 5160 |
| *Sex* | 4856 | 2.00 | 0.00 | 2 | 2 |
| *Kids* | 4856 | 4.48 | 14.49 | 0 | 99 |
| *Married* | 4856 | 1.92 | 1.50 | 1 | 9 |

### 15.4.2 Read Data into Statistical Package

Using the information in Table 15.6 the data were read into a statistical package, the variables were renamed, and the results saved as dataset PSIDRAW. Summary statistics are given in Table 15.7.

There are 4856 observations. Most variables have sensible descriptive statistics, except that variables *Educatn*, *Kids* and *Married* have unusually high mean values, due to the unrealistic maximum values of 99 and 9.

Variable *Age* is between 30 and 50 as selected. Variable *Earnings* has a maximum value of 240000, so presumably this is annual earnings in dollars. Variable *Hours* has a maximum of 5160 which is 14 hours per day seven days a week, a value that is unusually high but physically possible. Variable *Sex* is always 2, which presumably is the code for women.

### 15.4.3 Read the Codebook

Variables *Educatn*, *Kids* and *Married* have the unrealistic maximum values of 99 and 9. This is most likely due to missing values for these variables.

Not all questions in an interview survey are answered, and responses can include don't know or a deliberate nonresponse. A given question may deliberately be asked of just a subset of individuals. Even when questions are answered, they may need to be recoded. For example, it is standard to define a variable equal to 1 if married and 0 otherwise, but the survey data may be more detailed than this with responses such as widowed.

Tabulation of the variables in dataset PSIDRAW reveals that variable *Educatn* takes values 0-17 and then 98 and 99; variable *Kids* takes values 0-10 and then 98 and 99; and variable *Married* takes values 1-5 and then 8 and 9. We need to find out what values 98 and 99 mean for variables *Educatn* and *Kids*, and what the values for variable *Married* mean.

This requires looking at the **codebook** for the data, available from the PSID website.

We find that for *Educatn* the value 98 corresponds to "don't know" and 99 corresponds to "inappropriate" , for variable *Kids* the value 98 means "don't know" or "inappropriate" and 99 means "no birth history"; and for variable *Married* value 8 corresponds to "don't know" and 9 corresponds to "no history".

### 15.4.4 Missing Data

Missing data are often given a distinctive value to indicate that the data are missing. The current example used values 8, 9, 98 and 99. Commonly-used values to indicate missing data include -999, -99, -9, NA, N/A, * and even a blank entry. Yet another missing data code can be 0, as in some situations a value of 0 may simply mean that the respondent was not asked the question. When characters are used for missing values this can make numeric data harder to read in as it is now a mix of numbers and characters.

Once a data point is viewed to be missing, a command is given to assign a missing value to that data point. Statistical packages usually assign a particular value for missing and have a particular way of displaying the missing data. For example, Stata displays missing data as "." , Eviews uses "NA" and Gretl uses " ". Excel does not provide this option.

The missing data point is given a particular value internally, often a very large number. This can cause problems when transforming data. For example a command to recode all data values greater than 1000 to 1000 will also erroneously recode missing values to 1000 if the missing value is stored internally as a number greater than 1000.

It is best to code missing data as missing, rather than drop an observation entirely, as this enables analysis using as much of the data as is available. Suppose there are 100 observations and some are missing so that there are 80 nonmissing observations on $x$, 70 nonmissing observations on $y$, and only 60 observations have nonmissing data for both $x$ and $y$. Then deleting all 40 observations that are incompletely observed is not a good idea if analysis is only on $y$, for which 80 observations available. Going the other way, if interest lies in regression of $y$ on $x$ that uses only the 60 observations for which data on both $x$ and $y$ are available then it may be best to first report descriptive statistics on $x$ and $y$ using just the same 60 observations, rather than all observable data.

When data are missing it is possible to impute the missing value. This yields a larger sample but imputation relies on strong assumptions about the process generating the data. Imputation methods are beyond the scope of this book. The data provided in large national surveys such as the Current Population Survey may include imputed values for key variables such as income that some respondents refuse to answer.

### 15.4.5 Recode Data

The main variable in this example that needs recoding is variable *Married*. From the codebook, in addition to the missing values (8 and 9) already encountered, we have that 1 =

married, 2 = never married, 3 = widowed, 4 = divorced or annulment, 5 = separated. Here we recode 2-5 as not married (0), so the marital status variable here equals one if currently married and not separated and 0 otherwise.

The variable *Educatn* is topcoded at 17, which corresponds to one year of graduate school since 16 corresponds to completed four years of college. The value 17 could possibly be recoded to 18 as some master's degrees take two years and some of the people will have received a Ph.D. which is 20 years of schooling. This is not done here.

Analysis of earnings will be influenced by outlying observations. For example, should a value of $240,000 for annual earnings be included, dropped on grounds of being suspiciously high (perhaps due to a coding error) or should it be top-coded at a value of, say $100,000. By using other years of data for the same individual one can ascertain in most cases whether the data appears reasonable. Here we do not recode the variable.

### 15.4.6 Transform Data

Some new variables may be created, such as the natural logarithm of earnings or hourly wage which here is calculated as annual earnings divided by annual hours . Before doing such transformations, however, one needs to be sure that the variables fall into the allowable range. For example, one cannot take the logarithm of zero or a negative number.

Here we define variable *Wage* to be the hourly wage, equal to *Earnings* divided by *Hours*. The variable *Wage* is set to missing if the person had zero earnings. Many studies using self-reported data to construct wage will delete observations that have unusually low or unusually high wage.

### 15.4.7 Inspect the Data

After coding missing values, recoding variable *Married* and creating variable *Wage* we have dataset PSIDFINAL.

Table 15.8 presents summary statistics. Note that the number observations varies across variables, due to missing values or, for the last two variables, no earnings.

Before proceeding to analysis we should ask for each variable whether the summary statistics make sense? In particular, are the sample mean, minimum and maximum reasonable? How does the sample size compare to the original sample size? Does it seem that too many observations were lost for reasons such as missing data?

For the data set here the variables mostly make sense, except perhaps that only 63% of women between age 35 and 39 are married. The minimum of 0 year of schooling is very low. The hourly wage varies from $0.01, which seems too low, to $364.23 which may be too high.

A box plot can also be a useful aid in detecting outliers. While a complete listing of the data is not practical for large datasets, it can be good practice to at least inspect a partial listing, such as listing all variables for several of the observations.

Table 15.8: Summary statistics: Survey data after cleaning.

| **Name** | **Obs** | **Mean** | **St. Dev.** | **Min** | **Max** |
|---|---|---|---|---|---|
| *Intnum* | 4856 | 4958.10 | 2761.97 | 4 | 9306 |
| *Persnum* | 4856 | 59.21 | 79.75 | 1 | 205 |
| *Age* | 4856 | 38.46 | 5.60 | 30 | 50 |
| *Educatn* | 4630 | 12.37 | 3.05 | 0 | 17 |
| *Earnings* | 4856 | 14244.51 | 15985.45 | 0 | 240000 |
| *Hours* | 4856 | 1235.33 | 947.18 | 0 | 5160 |
| *Sex* | 4856 | 2.00 | 0.00 | 2 | 2 |
| *Kids* | 4738 | 2.14 | 1.43 | 0 | 10 |
| *Married* | 4804 | 0.63 | 0.48 | 0 | 1 |
| *Wage* | 3652 | 13.05 | 17.77 | 0.01 | 364.23 |
| *Lnwage* | 3652 | 2.22 | 0.84 | -5.07 | 5.90 |

In some cases data are obviously miscoded. For example, an age of 178 years is clearly miscoded. In other cases it is a judgement call. For example, unusually low and unusually large values of hourly wage may be dropped, but this requires determining what is unusual. Additional information may help in determining what is unusual. For example, an hourly-paid employee in a job covered by the minimum wage law should not have a wage of $1, whereas this is possible for a self-employed person. When data are viewed as miscoded the most common thing to do is to treat the variable as missing.

### 15.4.8 Data Analysis

Some variables are missing data, and in analysis we would need to decide whether to restrict analysis to a dataset where data are available for all the variables. This throws away data but has the advantage that the sample size and mix does not vary according tow which variables are analyzed.

Large data sets such as the PSID are not random samples of the U.S. population. For reasons of cost they are instead clustered samples from select geographic areas, to reduce interview costs. Furthermore these data sets oversample groups of particular interest to policy-makers, namely minorities and low-income groups. A thorough analysis would recognize this and do analysis weighted by the so-called sampling weights, possibly dropping observations from the deliberately over-sampled groups. We ignore this complication here.

For these data regression of *Lnwage* on *Educatn* yields $b_{Educatn} = 0.11$, so one additional year of education is associated with an 11% increase in hourly wage.

# Chapter 16

# © A. Colin Cameron: Multivariate Data Transformation

This chapter presents a range of data transformations that arise in multiple regression, building on material in Chapter 12 on data transformation in bivariate regression.

First regression models often include regressors that are indicator variables. These indicator variables may appear individually, such as gender, or as sets of mutually exclusive indicator variables, such as three indicator variables for education level that is, respectively, less than high school, high school or more than high school.

Second, variables may appear nonlinearly in a model. A simple nonlinear model is the quadratic model, easily estimated by OLS regression of $y$ on an intercept, $x$ and $x^2$. But interpretation of the coefficients is more difficult as the slope coefficient for $x$ no longer gives the change in the fitted value of $y$ as $x$ changes by one unit.

Third, regression models may include interactions of variables as regressors. For example, we may regress $y$ on $x$, $z$ and $x \times z$. Again interpretation of coefficients can be difficult.

Fourth, many economics models include as dependent variable, or as regressor(s), variables that have been transformed to natural logarithms.

## 16.1 Example: Earnings, Age, Education and Type of Worker

This chapter uses the same dataset EARNINGSAGE as that used in Chapter 12.

Table 16.1 presents summary statistics for the dependent variable *Earnings* and the other variables used in this chapter. Mean earnings are \$63,476 per year, mean age is 42.84 years and mean education is 13.72 years of schooling.

The type of worker is broken into three categories – self-employed, employed in private

Table 16.1: Annual Earnings of male full-time workers aged 25-65 in 2010: Variable definitions and summary statistics.

| **Variable** | **Definition** | **Mean** | **Standard Deviation** | **Min** | **Max** |
|---|---|---|---|---|---|
| *Earnings* | Annual earnings in $ | 63476 | 61713 | 5000 | 504000 |
| *Age* | Age in years | 42.84 | 10.72 | 25 | 65 |
| *Education* | Years of schooling | 13.72 | 3.05 | 0 | 20 |
| *d1* | = 1 if self-employed | 0.117 | 0.322 | 0 | 1 |
| *d2* | =1 if private sector employee | 0.747 | 0.435 | 0 | 1 |
| *d3* | =1 if government sector employee | 0.136 | 0.343 | 0 | 1 |
| *Agesq* | *Age* squared | 1949.9 | 938.9 | 625 | 4225 |
| *Educbyage* | *Education* times *Age* | 327.69 | 133.35 | 0 | 704 |
| *Hours* | Usual hours worked per week | 45.50 | 9.20 | 35 | 99 |
| *Lnhours* | Natural logarithm of *Hours* | 3.81 | 0.18 | 3.56 | 4.60 |
| *Lnearnings* | Natural logarithm of *Earnings* | 10.78 | 0.72 | 8.52 | 13.13 |
| n | 494 | | | | |

sector and employed in government sector. Then three indicator variables are defined

$$d1 = \begin{cases} 1 & \text{if self-employed} \\ 0 & \text{otherwise,} \end{cases}$$

$$d2 = \begin{cases} 1 & \text{if employed in private sector} \\ 0 & \text{otherwise.} \end{cases}$$

$$d3 = \begin{cases} 1 & \text{if employed in government sector} \\ 0 & \text{otherwise.} \end{cases}$$

Every individual in the sample (restricted to workers) falls into exactly one of the three categories. Thus for any individual observation only one of the three indicator variables takes value 1, while the remaining two indicator variables take value 0. It follows that

$$d1 + d2 + d3 = 1.$$

As a result the sum of the sample means of the three indicators variables $0.117 + 0.747 + 0.136 = 1$.

## 16.2 Single Indicator Variable

From Chapter 12.2, an **indicator variable** or **dummy variable** or **categorical variable** takes only two possible values, usually 0 or 1. This section extends Chapter 12.2 by allowing for additional regressors to be included in the model.

### 16.2.1 Single Indicator Variable

Consider a regression model where one of the regressors is an indicator variable. The population model is

$$y = \beta_1 + \beta_2 x + \alpha d + u,$$

where $\alpha$ is the coefficient of the indicator variable

$$d = \begin{cases} 1 & \text{if in the first category} \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity only one regressor other than $d$ is included, but results carry through immediately if $\beta_2 x$ is replaced with $\beta_2 x_2 + \cdots + \beta_k x_k$.

It follows that in the fitted model

$$\widehat{y} = \begin{cases} b_1 + b_2 x + a & \text{if in the first category} \\ b_1 + b_2 x & \text{otherwise.} \end{cases}$$

Thus the coefficient $a$ of the indicator variable regressor measures the change in $\widehat{y}$ when in the first category rather than the other category, holding all other regressor(s) constant.

For example, consider measuring and testing the difference in earnings by gender. Using just data on earnings one can simply use the test of difference in means given in Chapter 7.3 without need for regression methods. But for comparison of earnings by gender after controlling for other factors that also determine earnings, such as education, age, occupation of job, and industry of job, it is necessary to use regression methods.

Thus earnings ($y$) are regressed on the indicator variable $d$ and the control variables. Then $a$ measures the amount that female income exceeds male income on average, holding the other regressors constant. For example, if $a = -2000$ then female earnings are \$2,000 less than male earnings, once we control for the other regressors. A test of the claim that women earn less than men, controlling for the other regressors, is a test of $H_0 : \alpha \geq 0$ against $H_a : \alpha < 0$.

The first panel of Figure 16.1 illustrates that the indicator variable shifts the intercept in the regression according to whether $d = 1$ or $d = 0$, while keeping the slope coefficients the same.

**Remark 146** *In multiple regression the coefficient on an indicator variable acts as an intercept shifter.*

### 16.2.2 Dummy Variable Trap

Suppose the model includes an intercept and an indicator variable $d1$ that equals 1 if female and 0 otherwise. Then there is no need to include a second indicator variable $d2$ which equals 1 if male and 0 otherwise.

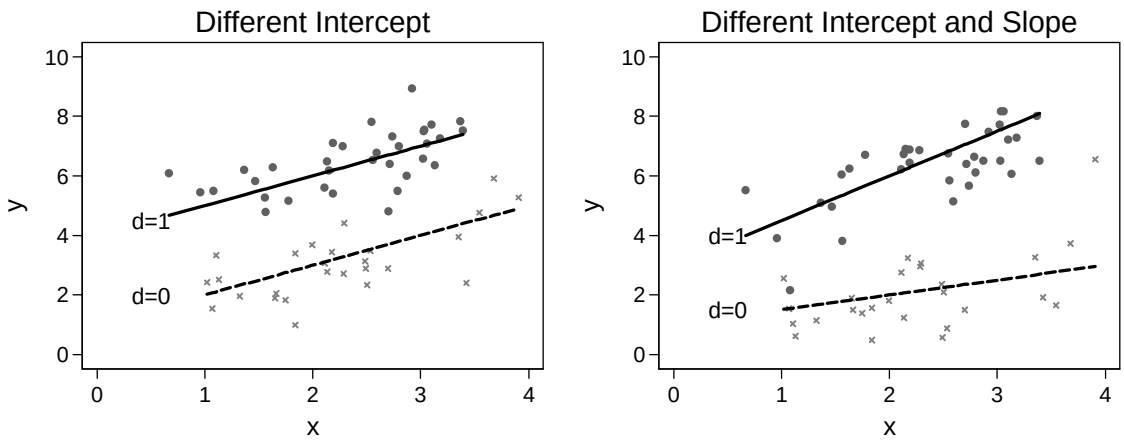


Figure 16.1: Indicator variable (left panel) and Interacted Indicator Variable (right panel) as Regressor.

To see this, observe that the model

$$y = \beta_1 + \beta_2 x + \alpha_1 d1 + \alpha_2 d2 + u,$$

can be re-expressed, using $d2 = 1 - d1$ since by construction $d1 + d2 = 1$, as

$$\begin{aligned} y &= \beta_1 + \beta_2 x + \alpha_1 d1 + \alpha_2(1 - d1) + u \\ &= (\beta_1 + \alpha_2) + \beta_2 x + (\alpha_1 - \alpha_2) d1 + u. \end{aligned}$$

Estimates of $(\beta_1 + \alpha_2)$, $\beta_2$ and $(\alpha_1 - \alpha_2)$ can be obtained, but these three estimates are insufficient to obtain estimates of all four parameters $\beta_1$, $\beta_2$, $\alpha_1$ and $\alpha_2$.

The potential problem of including too many indicator variables is called the **dummy variable trap**. There are two simple solutions that lead to equivalent estimates of the difference in $y$ across the two categories, and equivalent slope coefficients of other regressors.

The most common solution is to omit one of the mutually exclusive indicator variables, $d1$ or $d2$, as was done in the preceding earnings and gender example. Suppose $d2$ is omitted, so $\alpha_2 = 0$. Then, as before, the coefficient of $d1$ measures the amount by which female earnings ($d1 = 1$) exceed male earnings ($d1 = 0$).

The second solution is to omit the intercept, so $\beta_1 = 0$, and regress $y$ on $x$, $d1$ and $d2$ without an intercept. Then $a_1 - a_2$ measures the amount that female earnings ($d1 = 1$) exceeds male earnings ($d2 = 1$). A test of the claim that women earn less than men, controlling for the other regressors, is then a test of $H_0 : \alpha_1 - \alpha_2 \geq 0$ against $H_a : \alpha_1 - \alpha_2 < 0$.

**Remark 147** *The dummy variable trap arises if two indicator variables are included and they exhaust all possibilities so that their sum equals one for every observation. Then the intercept or one of the indicator variables need to be dropped. These solutions lead to equivalent estimates of the difference in $y$ across the two categories, and equivalent slope coefficients of other regressors.*

### 16.2.3 Interacted Indicator Variables

An **interacted indicator variable** is a regressor that is the product of an indicator variable and another regressor. Consider the model

$$y = \beta_1 + \beta_2 x + \alpha_1 d + \alpha_2 d \times x + u,$$

where $d$ is an indicator variable taking values 0 or 1 and $x$ is another regressor. Then in the fitted model

$$\widehat{y} = \begin{cases} (b_1 + a_1) + (b_2 + a_2)x & \text{if } d = 1 \\ b_1 + b_2 x & \text{if } d = 0. \end{cases}$$

As before $a_1$ measures the difference in the intercept term between the case $d = 1$ and the case $d = 0$. Additionally $a_2$ gives the difference in the slope coefficient of the regressor $x$ between the case $d = 1$ and the case $d = 0$.

Returning to the earnings and gender example, we may wish to ask whether the returns to schooling differ by gender. Then we regress earnings on an intercept, a gender indicator ($d$), earnings ($x$), the gender indicator times earnings ($d \times x$), and any additional regressors. The coefficient of the interacted regressor $d \times x$ measures the higher return to women, compared to men, of an additional year of schooling.

The second panel of Figure 16.1 that if regressors include both an indicator variable and the same indicator variable interacted with $x$ then both the intercept and the slope coefficient of $x$ differ according to whether $d = 1$ or $d = 0$.

**Remark 148** *An interacted indicator variable is a regressor that is the product of an indicator variable and another regressor. This enables slope coefficients to vary according to the value of the indicator variable.*

### 16.2.4 Dependent Variable is an Indicator Variable

In some applications the dependent variable is an indicator variable. For example, we may wish to model whether or not an individual is employed, in which case $y = 1$ if a person works and $y = 0$ if a person does not.

Regression models with dependent variable that is an indicator variable are presented in detail in Chapter 19.4. If estimation is by OLS regression of $y$ on regressors then

heteroskedastic-robust standard errors must be used and the OLS slope coefficients can only be interpreted as giving average marginal effects. It is better to use logit or probit models presented in Chapter 19.4.

## 16.3 Sets of Indicator Variables

In some cases a categorical variable may include several values.

If these categories are **ordered** it can be convenient to create a single regressor with a distinct value for each category. For example, in the house price example of Chapter 13 the lot sizes of small, medium and large were given values of 1, 2 and 3. This method restricts the predicted change in house price when moving from a small lot to a medium lot to equal the predicted change when moving from a medium lot to a large lot.

Furthermore this approach is not possible when the categories are **unordered**, such as a variable for color that has categories blue, red and green.

Instead when a categorical variable includes several categories we form a **set of indicator variables**, one for each category where each indicator variable takes value 0 or 1. These indicator variables are included as regressors, with one of the indicator variables omitted to avoid the dummy variable trap.

### 16.3.1 Sets of Indicator Variables

As an example of sets of indicator variables, consider a categorical variable that can take three values, so we create three indicator variables, one for each category. An example, given in Chapter 16.1, is to form three indicator variables when type of worker has three categories – self-employed ($d1 = 1$), employed in private sector ($d2 = 1$) and employed in government sector ($d3 = 1$).

A set of indicator variables is **mutually exclusive** if any individual in the sample falls into exactly one of the categories. Then for any individual observation only one of the indicator variables takes value 1, while the remaining indicator variables take value 0. Thus the indicator variables sum to one

$$d1 + d2 + d3 = 1.$$

Due to this relationship between the three indicators not all three indicators can be included in the regression if an intercept is also included. The reason is similar to that given in the case of a single indicator variable and again erroneous inclusion of the complete set of indicator variables is called the **dummy variable trap.**

The usual approach is to drop one of the indicator variables. Variable $d1$ can be dropped

by substituting in $d1 = 1 - d2 - d3$, so

$$\begin{aligned} y &= \beta_1 + \beta_2 x + \alpha_1 d1 + \alpha_2 d2 + \alpha_3 d3 + u \\ &= \beta_1 + \beta_2 x + \alpha_1(1 - d2 - d3) + \alpha_2 d2 + \alpha_3 d3 + u \\ &= (\beta_1 + \alpha_1) + \beta_2 x + (\alpha_2 - \alpha_1)d2 + (\alpha_3 - \alpha_1)d3 + u. \end{aligned}$$

It follows that if $d1$ is dropped then the coefficient $(\alpha_2 - \alpha_1)$ of $d2$ measures the difference between earnings for a private sector worker ($d2 = 1$) and a self-employed worker ($d1 = 1$) after controlling for the other regressors.

Similarly the coefficient $(\alpha_3 - \alpha_1)$ of $d3$ measures the difference between earnings for a government sector worker ($d3 = 1$) and a self-employed worker ($d1 = 1$) after controlling for other regressors. So the coefficients of the included indicator variables measure the marginal effect of being in those categories compared to the omitted category.

The omitted category is called the **base category** or **reference category**. Interpretation of coefficients is simplest if the base category is a meaningful category. For example, in the type of worker example it might be most natural to use employment in the private sector as the base category, as most workers are in this category.

**Remark 149** *Suppose a categorical variable has C categories. Form a set of C mutually exclusive indicator variables d1, d2,..., dC. To avoid the dummy variable trap drop one of the indicator variables, called the omitted or base category. The coefficient of an included indicator variable measures the marginal effect of being in that categories compared to the base category, after controlling for the other regressors.*

Care is needed in interpreting hypothesis tests. When $d1$ is the omitted category the coefficient of $d2$ measures $(\alpha_2 - \alpha_1)$, so a test of statistical significance of $d2$ is a test of $H_0 : \alpha_2 = \alpha_1$ against $H_a : \alpha_2 \neq \alpha_1$. In the earnings example it is a test of whether private sector workers earn the same on average as self-employed workers after controlling for the other regressors. It is important to note that it is not a test of whether or not $\alpha_2 = 0$. And if instead $d3$ was dropped, say, then by similar reasoning the coefficient of $d_2$ measures $(\alpha_2 - \alpha_3)$.

**Remark 150** *A t test of the statistical significance of a single indicator variable is a test of whether the marginal effect of that category differs from that for the base category. It is not a test of whether the marginal effect of that category is zero.*

To test whether collectively the set of indicator variables is statistically significant we use the $F$ test for a subset of regressors. If $d1$ is the base category then we test whether the coefficients of $d2$ and $d3$ are jointly statistically significant. Similarly if $d2$ is the base category, say, then we test whether the coefficients of $d1$ and $d3$ are jointly statistically significant.

**Remark 151** *An F test of the joint statistical significance of the C-1 included indicator variables is a test of whether the set of indicator variables is statistically significant. This joint F test leads to the same result regardless of the category that is dropped.*

### 16.3.2 Example: Earnings and Type of Worker

Table 16.2 presents the results of several regressions that vary in how the sets of indicators are introduced in the model. Default standard errors are given in parentheses and $t$-statistics are given in square brackets.

Table 16.2: Sets of indicator variables: Earnings regression with default standard errors in parentheses and t-statistics in square brackets.

| **Variable** | **No Indicators** | **Drop d1** | **Drop d2** | **Drop d3** | **Drop intercept** |
|---|---|---|---|---|---|
| *Age* | 565 | 475 | 475 | 475 | 475 |
| | (243) | (244) | (244) | (244) | (244) |
| | [2.33] | [1.95] | [1.95] | [1.95] | [1.95] |
| *Education* | 6857 | 6921 | 6921 | 6921 | 6921 |
| | (855) | (851) | (851) | (851) | (851) |
| | [8.02] | [8.13] | [8.13] | [8.13] | [8.13] |
| *d1 (self-employed)* | | | 20259 | 25489 | -33232 |
| | - | - | (8204) | (10379) | (17683) |
| | | | [2.47] | [2.46] | [-1.88] |
| *d2 (private sector)* | | -20259 | | 5231 | -53490 |
| | - | (8204) | - | (7638) | (15586) |
| | | [-2.47] | | [0.68] | [-3.43] |
| *d3 (government sector)* | | -25489 | -5231 | | -58721 |
| | - | (10379) | (7638) | - | (16978) |
| | | [-2.46] | [-0.68] | | [-3.46] |
| *Intercept* | -54783 | -33232 | -53490 | -58721 | |
| | (15643) | (17683) | (15586) | (16978) | - |
| | [-3.50] | [-1.88] | [-3.46] | [-3.46] | |
| F(2,n-k) for indicators | - | 3.68 | 3.68 | 3.68 | 3.68 |
| p-value for F | - | .0260 | .0260 | .0260 | .0260 |
| n | 494 | 494 | 494 | 494 | 494 |
| $R^2$ | .126 | .139 | .139 | .139 | .582 |
| Adjusted $R^2$ | .123 | .132 | .132 | .132 | .578 |
| St. error | 57800 | 57496 | 57496 | 57496 | 61412 |
| Overall F | 35.51 | 19.74 | 19.74 | 19.74 | 136.21 |
| p-value for overall F | .000 | .000 | .000 | .0000 | .0000 |
| Subset F (indicators) | - | 3.60 | 3.60 | 3.60 | 3.60 |
| p-value for subset F | - | .028 | .028 | .028 | .028 |

The first regression simply regresses earnings on education and age, without any of the

indicator variables. Education has a large effect, one more year of schooling is associated with a \$6,857 increase in annual earnings, and is highly statistically significant. Variable *Age* is also statistically significant at level 0.05.

The next three columns of Table 16.1 present results when the indicator variables for type of worker are included as regressors and the base category is, in turn, self-employed ($d1 = 1$), private sector ($d2 = 1$) and government sector ($d3 = 1$).

From columns 2-4, many of the entries in Table 16.1 are invariant to the choice of omitted category. The coefficients, standard errors and $t$-statistics for the other regressors, *Education* and *Age*, do not change. The goodness of fit measures – $R^2$, $\bar{R}^2$, $s_e$ and the $F$-statistic for overall significance – do not change. Additionally the $F$-statistic for the joint significance of the two included indicator variables, given at the bottom of Table 16.2, does not change. These $F(2, 48)$ tests all have $F = 3.60$ and $p = .028$ so we conclude that the type of worker is statistically significant at significance level 0.05.

From columns 2-4, change in the omitted category does lead to change the estimated intercept and the estimated coefficients for the included indicator variables. From the second column, when $d1$ is omitted $d2$ has estimated coefficient $-20259$ and $d3$ has estimated coefficient $-25489$. So after controlling for age and education, private sector workers earn \$20,259 less than self-employed workers and government workers earn \$25,489 less than self-employed workers.

From column 3, with $d2$ instead omitted, $d1$ has estimated coefficient 20259, so self-employed sector workers earn \$20,259 more than those in the private sector. This is consistent with the results with $d1$ omitted that found that private sector workers earn \$20,259 less than self-employed workers. And $d3$ has estimated coefficient $-5231$, so government workers earn \$5,231 less than those in the private sector. This is consistent with the results with $d1$ omitted that found that private sector workers earn \$20,259 less than self-employed workers and government workers earn \$25,489 less than self-employed workers, since $20259 - 25489 = -5230$ (the minor difference reflects rounding error).

The $t$ tests of individual coefficients differ according to which indicator variable is omitted since they are tests of the significance compared to the reference group. For example, using the base category of private worker, so $d2$ is omitted, we see from column 3 that the difference between earnings of self-employed workers and private sector workers is statistically significant at level 0.05, since $t = 2.47$. But the difference between earnings of government workers and private sector workers is not statistically significant since $t = -0.68$.

The choice of omitted category is one of convenience. If, for example, we are interested in measuring earnings of workers relative to those in the private sector then drop the indicator variable for the private sector.

### 16.3.3 Dropping the Intercept

Another way to avoid the dummy variable trap is to include all the indicator variables but drop the intercept.

The final column of Table 16.1 gives results using this alternative method. Again the coefficients, standard errors and $t$-statistics for the other regressors, *Age* and *Education*, do not change.

The coefficients of the indicator variables now directly estimate the coefficients $\alpha_1$, $\alpha_2$ and $\alpha_3$ (at the expense of being unable to estimate the intercept $\beta_1$). Since $a_1 = -33232$ and $a_2 = -53490$, the difference in earnings between self-employed workers and private sector workers equals $a_1 - a_2 = -33232 - (-53490) = -20528$. This equals the estimated coefficient of $d2$ in the model with $d1$ omitted, given in column 3 of Table 16.2.

From the last column of Table 16.2, dropping the intercept has led to a change in the overall goodness of fit measures – $R^2$, $\bar{R}^2$, $s_e$ and the $F$-statistic for overall significance. The reason is that these statistics are inappropriate in regressions that do not include an intercept. Recall that $R^2$ measures explained variation relative to the total sum of squares, which is regression with just an intercept. Similarly the $F$ test presumes that the restricted model includes an intercept. This is one reason for omitting an indicator variable rather than the intercept.

### 16.3.4 Difference in Means

Suppose we simply want to test whether earnings vary across the type of worker, without inclusion of any controls. In areas of applied statistics that do not use regression methods, this is tested using **analysis of variance (ANOVA) methods** that are an extension of the $t$ test for difference in two means.

The same test can be implemented using regression methods. Regressing earnings on an intercept, $d2$ and $d3$ yields

$$\widehat{y} = \underset{(10.30)}{83022} - \underset{(-2.49)}{21589}d2 - \underset{(-2.29)}{25215}d3,$$

where $t$-statistics based on default standard errors are given in parentheses. Then earnings are \$83,022 for self-employed workers, the omitted category, and are \$21,589 less than this for private sector workers and \$25,215 lower for government sector workers. The $F$-statistic for joint statistical significance of $d2$ and $d3$ equals 3.43 with $p = 0.033$ so at significance level 0.05 there is a statistically significant difference in earnings across the three types of workers.

Note that the OLS estimates imply that earnings are on average \$83,022 for the self-employed, \$61,433 ($= 83022 - 21589$) for government workers, and \$57,807 ($= 83022 - 25215$) for private sector workers. These quantities equal the sample mean of earnings in each of

the three categories of worker, and could also be directly obtained from OLS regression of *Earnings* on *d1*, *d2* and *d3* with the intercept omitted.

### 16.3.5 More Complex Indicators and Interactions

More complex models can include several sets of indicator variables. Then it is simplest to include the intercept while one indicator variable is dropped from each of the sets of indicator variables.

The sets of indicator variables may be interacted with other regressors. For example, to allow for the returns to education to vary with type of worker we regress earnings on an intercept, *Age*, *Education*, $d2$, $d3$, $d2$ times *Education*, and $d3$ times *Education.*

As increasingly more complex models are estimated, care is needed to avoid inadvertently running into the dummy variable trap. For example, suppose the regressors include an indicator for female, an indicator for married, and an interaction between these two indicators. Then problems will arise if there are, for example, no married men in the sample.

When the dummy variable trap does arise a statistical package will drop one of the regressors, or more than one regressor if the dummy variable trap occurs more than once. This can change the way that the coefficients of indicator variables are interpreted. Special care is needed if a statistical package drops one or more regressors.

## 16.4 Quadratic Model and Polynomial Models

A leading example of a nonlinear relationship is a model where $y$ is a quadratic function of $x$ rather than a linear function of $x$. More generally polynomial models include powers of $x$ as regressors.

### 16.4.1 Quadratic Models

A **quadratic model** is the model

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + u.$$

This is a polynomial model of degree two, and reduces to a linear model if $\beta_3 = 0$.

The model defines a parabola, though often only one side of the parabola appears for the sample range of $x$. Figure 16.2 presents various possible shapes for the relationship between $y$ and $x$ for a quadratic model.

The top three panels of Figure 16.2 show examples where $\beta_3$, the coefficient of $x^2$, is positive. For the parabola in the top left panel it can be shown that the minimum value of $y$ occurs when $x = -\beta_2/(2\beta_3)$. For the middle panel in the top row $x < -\beta_2/(2\beta_3)$, so the

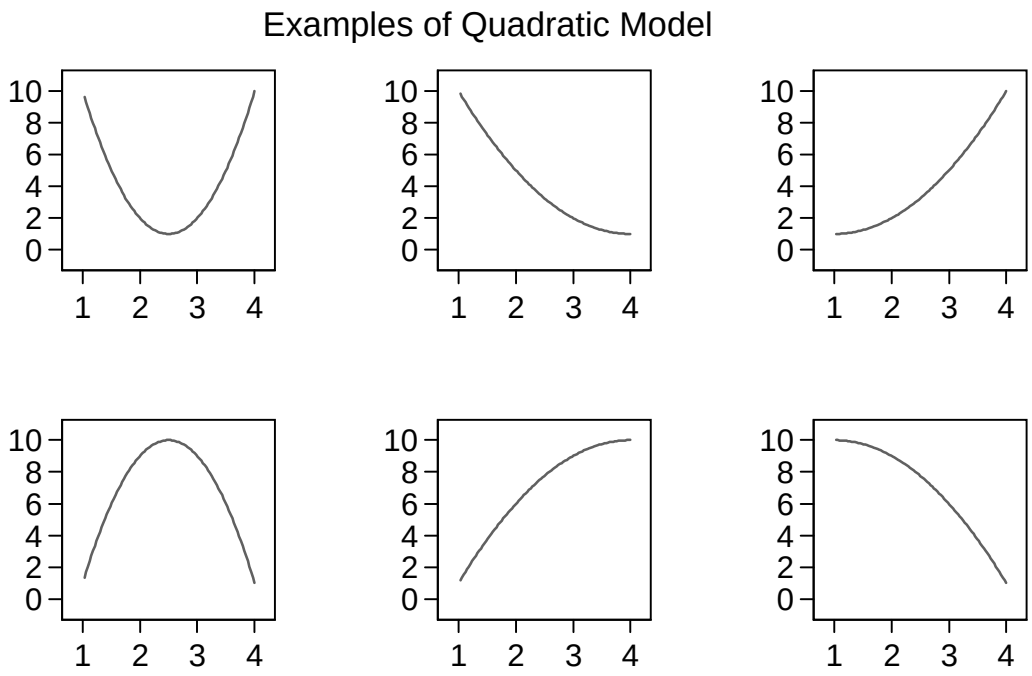


Figure 16.2: Various quadratic models: The coefficient of $x^2$ is positive for the top three panels and negative for the lower three panels.

data include only the left half of the parabola, while for the right panel in the top row $x > -\beta_2/(2\beta_3)$.

The bottom three panels of Figure 16.2 show examples where $\beta_3 < 0$. The parabola in the left panel $y$ takes its maximum value when $x = -\beta_2/(2\beta_3)$, for the middle panel $x < -\beta_2/(2\beta_3)$ and for the right bottom panel $x > -\beta_2/(2\beta_3)$.

A quadratic model is estimated as follows. First, create a variable called, say, $xsq$, where $xsq = x^2$. Then do least squares regression of $y$ on $x$ and $xsq$. The only change from bivariate regression is that an estimation command such as *regress y x* becomes *regress y x xsq*.

**Remark 152** *The parabola $y = \beta_1 + \beta_2 x + \beta_3 x^2$ has turning point $-\beta_2/(2\beta_3)$ at a minimum if $\beta_3 > 0$ and at a maximum if $\beta_3 < 0$.*

### 16.4.2 Marginal Effects: Quadratic Model Example

Marginal effects in nonlinear models were discussed in Chapter 12, with emphasis on models in natural logarithms. Here the quadratic model is considered.

Begin with fitted model

$$\widehat{y} = b_1 + b_2 x + b_3 x^2 + b_4 z,$$

where an additional regressor $z$ has been added.

If $x$ increases by the amount $\Delta x$, so now evaluation is at $x + \Delta x$, then $\widehat{y}$ becomes

$$\begin{aligned} \widehat{y}_{new} &= b_1 + b_2(x + \Delta x) + b_3(x + \Delta x)^2 + b_4 z \\ &= b_1 + b_2 x + b_2 \Delta x + b_3 x^2 + 2b_3 x \Delta x + b_3(\Delta x)^2 + b_4 z. \end{aligned}$$

On subtraction of $\widehat{y}_{old} = b_1 + b_2 x + b_3 x^2 + b_4 z$ it follows that $\Delta\widehat{y} = b_2 \Delta x + 2b_3 x \Delta x + b_3(\Delta x)^2$, so the **marginal effect** on $y$ of changing $x$ by $\Delta x$, computed by the **finite-difference method** is

$$\text{ME} = \frac{\Delta\widehat{y}}{\Delta x} = b_2 + 2b_3 x + b_3 \Delta x.$$

Note that the size of ME varies with both $x$, through the term $2b_3 x$, and $\Delta x$, through the term $b_3 \Delta x$.

For a finite change, the usual choice is to consider a **one-unit change** in $x$, so $\Delta x = 1$ and ME$= b_2 + 2b_3 x + b_3$.

It is more common to use the **calculus method** that computes the ME when $\Delta x$ is very small. When $\Delta x \to 0$ the third term $b_3 \Delta x \to 0$ and ME$= b_2 + 2b_3 x$. For those familiar with calculus this result can be obtained immediately, as the derivative of $\widehat{y} = b_1 + b_2 x + b_3 x^2$ is $d\widehat{y}/dx = b_2 + 2b_3 x$.

**Remark 153** *The marginal effect in the fitted quadratic model $\widehat{y} = b_1 + b_2 x + b_3 x^2$ equals $b_2 + 2b_3 x$ using calculus methods.*

For nonlinear models three alternative measures of the marginal effect are the average marginal effect (AME), the marginal effect at the mean (MEM), and the marginal effect at a representative value (MER); see Chapter 12.3. For the quadratic model the AME is

$$\text{AME} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(b_2 + 2b_3x_i) = b_2 + 2b_3 \times \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i = b_2 + 2b_3\bar{x},$$

which is just the MEM. For other nonlinear models the AME and MEM differ.

### 16.4.3 Quadratic Model Example: Earnings and Age

Begin with the model given in the first column of Table 16.2 and add a quadratic term in age.

The fitted quadratic model, obtained by multivariate least squares regression of *Earnings* ($y$) on *Age* ($x$), *Agesq* ($x^2$), and *Education* ($z$), is

$$\widehat{y} = \underset{(-3.38)}{-147573} + \underset{(2.53)}{5153x} - \underset{(-2.27)}{52.78x^2} + \underset{(7.98)}{6795z}, \quad R^2 = .136, \quad n = 494,$$

where the $t$-statistics given in parentheses are based on default standard errors. A quadratic term is warranted as it is statistically significant at level 0.05 since $|t| = 2.27 > t_{.025;490} - 1.965$.

In general the turning point for the quadratic is at $x = -b_2/2b_3$. Here the turning point is at *Age*$= -5153/(2 \times -(52.78)) = 48.8$. Controlling for education, earnings on average increase with age to a maximum at 48.8 years and then decline with additional age.

The estimated ME$= 5153 - 2 \times 52.78x - 52.78\Delta x$ by the finite difference method. Using calculus methods the ME$= 5153 - 105.56x$, while for a one-unit change, with $\Delta x = 1$, ME$= 5000 - 105.56x$. In this particular example there is little difference between the two methods since aging one year ($\Delta$*Age*$= 1$) is a relatively small change given that age in the sample ranges from 25 to 65.

Using calculus methods the ME is as high as \$2,515 at age 25 years and as low as -\$1,707 at age 65 years. The MEM$= 5153 - 105.56\bar{x} = 5153 - 105.56 \times 42.84 = 631$, compared to \$565 for the linear model. As already noted, for the quadratic model this also equals the AME.

### 16.4.4 Polynomial Model

A generalization of the quadratic model is a **polynomial model of degree** $p$ that includes powers of $x$ up to $x^p$. Then the fitted model is

$$\widehat{y} = b_1 + b_2x + b_3x^2 + \cdots + b_{p+1}x^p.$$

This model has up to $p - 1$ turning points. The order of the polynomial can be determined by progressively adding terms $x^2$, $x^3$, ... in a multivariate regression until additional terms are no longer statistically significant.

By calculus methods it can be shown that as $\Delta x \to 0$, the marginal effect is

$$\text{ME} = b_2 + 2b_3x + 3b_4x^2 + \cdots + pb_{p+1}x^{p-1},$$

which again will vary with the point of evaluation $x$.

## 16.5 Interacted Regressors

An **interacted regressor** is a regressor that is the product of other regressors that appear in the model. Estimation by OLS regression remains straightforward, but adjustment is needed in testing statistical significance of a variable and in interpreting estimated coefficients.

### 16.5.1 Interacted Regressors

An example of a model with an interacted regressor is

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 x \times z + u.$$

Here the regressor $x \times z$ is an interacted regressor.

Estimation of the model is straightforward – create a variable $xz$, say, that equals $x \times z$ and run an OLS regression of $y$ on an intercept, $x$, $z$ and $xz$. Then the fitted model is

$$\widehat{y} = b_1 + b_2 x + b_3 z + b_4 xz,$$

where $xz = x \times z$.

Interpretation of regressors is more difficult, however, as the variable $x$ is a component of two regressors, the regressor $x$ and the regressor $xz$. Similarly the variable $z$ appears in the two regressors $z$ and $xz$.

First, to compute the marginal effect (ME) on $\widehat{y}$ of a change in $x$, holding $z$ constant, we need to consider the combined effect of change in any regressor that depends on $x$, here the regressors $x$ and $xz$. Then

$$\text{ME}_x = \frac{\Delta \widehat{y}}{\Delta x} = b_2 + b_4 z, \text{ and } \text{ME}_z = \frac{\Delta \widehat{y}}{\Delta z} = b_3 + b_4 x.$$

To see this, note that if $x$ changes by $\Delta x$ to $x + \Delta x$, then $\widehat{y}$ changes to $\widehat{y} = b_1 + b_2(x + \Delta x) + b_3 z + b_4(x + \Delta x) \times z$, which is an increase of $\beta_2 \Delta x + \beta_4 \Delta x \times z$. Dividing by $\Delta x$ yields $\text{ME}_x$. This result can also be obtained by directly taking the derivative of $\widehat{y}$ with respect to $x$, holding $z$ constant. The marginal effect varies with the level of variable $z$. Similarly algebra yields the expression for $\text{ME}_z$.

Second, to test whether variable $x$, say, is statistically significant we need to perform a test of the joint statistical significance of all regressors that depend on the variable $x$, here

the regressors $x$ and $xz$. Thus we perform an $F$ test of $H_0 : \beta_2 = 0$, $\beta_4 = 0$ against $H_a$ : at least one of $\beta_2 \neq 0$, $\beta_4 \neq 0$.

Third, to test whether there is an interaction between $x$ and $z$ we perform a standard $t$ test of $H_0 : \beta_4 = 0$ against $H_a : \beta_4 \neq 0$.

**Remark 154** *When a variable appears directly as a regressor and additionally as an interacted regressor the marginal effect of the variable includes an effect through the interaction term. A test of the statistical significance of the variable is an $F$ test of the joint statistical significance of the regressor and the interaction term. A $t$ test of statistical significance of the interacted regressor is a test of whether the interaction term is needed.*

### 16.5.2 Interactions Example: Earnings, Education and Age

The first column of Table 16.2 lists results for OLS regression of *Earnings* on *Age* ($x$) and *Education* ($z$).

Now suppose we wish to allow for the possibility that the returns to schooling vary with age. To test this we introduce an interaction term *AgebyEduc* ($x \times z$) that equals *Age* times *Education*. The fitted model is now

$$\widehat{y} = \underset{(1.04)}{41198} - \underset{(-1.59)}{1048}\, x - \underset{(-0.11)}{311}\, z + \underset{(2.63)}{218}\, x \times z, \quad R^2 = .138, \quad n = 494,$$

where $t$-statistics based on default standard errors are given in parentheses.

Compared to the model without the interaction term, the regressors *Age* and *Education* are now statistically insignificant at significance level 0.05 and only the interaction term, with $t = 2.63$, is statistically significant.

At first glance it appears that now education does not have a statistically significant effect on earnings, since regressor *Education* has $t = -0.11$. But this conclusion would be wrong. A joint test of the statistical significance of the regressor *Education* ($z$) and the education-age interaction term ($x \times z$), here a test of $H_0 : \beta_3 = 0$, $\beta_4 = 0$ yields $F = 36.02$ with $p = 0.000$. So education remains highly statistically significant. Similarly the $F$-statistic for joint statistical significance of the regressor *Age* ($x$) and the education-age interaction term ($x \times z$) has $F = 6.20$ with $p = 0.002$, so age is still statistically significant.

Why have the variables *Education* and *Age* become statistically insignificant? The interaction variable *AgebyEduc* is quite highly correlated with *Education* ($\widehat{\rho} = 0.77$) and *Age* ($\widehat{\rho} = 0.62$). When regressors are highly correlated with each other their individual contributions are measured much less precisely. Thus the standard errors of *Education* and *Age* given in the first column of Table 16.2 are 855 and 243. These more than double to 2854 and 659 with inclusion of variable *AgebyEduc*. As already noted, joint $F$-tests that include the interaction term show that education and age remain statistically significant. This increase in imprecision when regressors are highly correlated is an example of multicollinearity, studied in Chapter 17.

Now consider interpretation of the regression coefficients, allowing for inclusion of the interaction term. The marginal effect of one more year of schooling is

$$\text{ME}_{Ed} = -311 + 218 \times Age.$$

So the returns to education increase as one ages. For example the marginal effect is \$5,139 at age 25 years and \$13,859 at age 65 years. The average marginal effect (AME) is $\text{AME}_{Ed} = \frac{1}{n}\sum_{i=1}^{n}[-311 + 218x_i] = -311 + 218\bar{x} = 9029$.This equals the marginal effect at the mean (MEM), a result that holds more generally whenever the interaction term is of multiplicative form $x \times z$. The AME is quite different from the slope coefficient of 6857 in the original OLS regression without the interaction term.

Some statistical packages provide a command that directly computes the AME or MEM of a variable that appears as in interaction in the regression, allowing for the complication that the variable appears in more than one regressor, as in the preceding example. Such a command will also compute the standard error of the marginal effect, which in the previous example equals $\sqrt{se(b_3)^2 + 2\bar{x} \times \widehat{cov}(b_3, b_4) + \bar{x}^2 se(b_4)^2}$, where $\widehat{cov}(b_3, b_4)$ is the estimated covariance between $b_3$ and $b_4$. In this example the resulting standard error of $\text{AME}_{Ed}$ is 1185, so the AME is highly statistically significant at significance level 0.05 as $t = 9029/1185 = 7.62$.

## 16.6 Natural Logarithm Transformations

Transformation models using $\ln y$ as dependent variable or $\ln x$ as regressor have been studied in Chapters 12.4-12.6 for the bivariate case, where their interpretation in terms of semi-elasticities and elasticities was highlighted. In the multivariate case the interpretations are similar, so only a brief presentation is given here.

**Remark 155** *The interpretation of models with dependent variable or regressor(s) in natural logarithms is similar to the bivariate case. Coefficients are most easily interpreted as semi-elasticities or elasticities.*

The only real complication is that a mix of regressor types may appear – in levels or as natural logarithm – and that indicator variables and interactions may also appear as regressors. Two examples are presented – one with the dependent variable in levels and the other with dependent variable in logs.

### 16.6.1 Dependent Variable in Levels

It is natural to include age and education as regressors in levels, as interest lies in the effect on earnings of one more year of aging or one more year of education. If interest lies instead in

the effect on earnings of a proportionate or percentage change in a variable then the variable can be included in natural logarithms, as the slope coefficient can then be interpreted as giving the result of a proportionate change; see Chapter 12.6.

Here we estimate the main model of Table 16.2 with the inclusion of the additional regressor *Lnhours*, the natural logarithm of usual hours worked per week. The fitted OLS regression is

$$\begin{aligned}\widehat{Earnings} &= \underset{(-6.51)}{-364986} + \underset{(2.12)}{501} \times Age + \underset{(7.55)}{6276} \times Education + \underset{(1.62)}{13117} \times Dself \\ &\quad - \underset{(-0.43)}{3169} \times Dgovt + \underset{(5.77)}{84138} \times Lnhours, \qquad R^2 = .194, \quad n = 494,\end{aligned}$$

where the omitted category for type of worker is private worker ($Dpriv = 1$). Here $t$-statistics based on default standard errors are given in parentheses.

The coefficients of the first four regressors are interpreted in the usual way. Aging one year is associated with an increase in earnings of \$501, one more year of schooling is associated with an increase in earnings of \$6,276, self-employed workers earn \$13,117 more than private sector workers and government workers earn \$3,169 less than self-employed workers, after controlling for the other regressors.

The coefficient of *Lnhours* (which equal $\ln(Hours)$) can be most easily interpreted as a semi-elasticity. A 1% increase in usual hours worked is associated with an \$842 increase in annual earnings ($= 0.01 \times 84138$).

### 16.6.2 Dependent Variable in Natural Logarithms

This chapter analyzed the level of earnings for ease of exposition of sets of indicator variables and interacted regressors. It is more common to model the natural logarithm of earnings, since these are right-skewed data.

The fitted OLS regression with dependent variable *Lnearnings*, the natural logarithm of earnings, is

$$\begin{aligned}\widehat{Lnearnings}) &= \underset{(8.67)}{5.014} + \underset{(3.28)}{0.009} \times Age + \underset{(9.78)}{0.091} \times Education - \underset{(-0.96)}{0.086} \times Dself \\ &\quad + \underset{(0.87)}{0.072} \times Dgovt + \underset{(6.66)}{1.089} \times \ln(Hours), \qquad R^2 = .262, \quad n = 494,\end{aligned}$$

where $t$-statistics based on default standard errors are given in parentheses.

The coefficients of the first four regressors are most easily interpreted as semi-elasticities. Aging one year is associated with a 0.92% increase in earnings ($= 100 \times 0.009$), one more year of schooling is associated with a 9.1% increase in earnings, self-employed workers earn 8.6% less than private sector workers, and government workers earn 7.2% more than private sector workers, after controlling for the other regressors.

For the indicator variables it is more natural to consider a one-unit change than use calculus methods. Then for regressor *Dself*, going from self-employed (*Dself*$= 1$) to a private sector worker (*Dself*$= 0$) is associated with an $\exp(-0.086) - 1 = -0.082$ proportionate change in earnings. Thus, more precisely, self-employed people earn 8.2% less than private sector workers, controlling for education, age and usual hours worked.

The coefficient of $\ln(Hours)$ can be interpreted directly as an elasticity. A 1% increase in usual hours worked is associated with a 1.089% increase in annual earnings.

A leading example of a model with both dependent and regressor variables in natural logarithms is the Cobb-Douglas production function, presented in Chapter 15.2.

### 16.6.3 Retransformation Bias

The OLS regression with $\ln y$ as dependent variable leads to unbiased prediction of $\ln y$, provided model assumptions 1-2 hold. But it leads to biased prediction of $y$, due to **retransformation bias**.

Suppose we regress $\ln y$ on $x$ and $z$ and obtain the fitted value $\widehat{\ln y} = b_1 + b_2 x + b_3 z$. The obvious predicted value for $y$ is $\widehat{y} = \exp(\widehat{\ln y})$, but this will underpredict $y$ as explained in Chapter 12.5. If the model errors in the regression for $\ln y$ are homoskedastic and normally distributed then unbiased predictions of $y$ can be obtained by rescaling by $\exp(s_e^2/2)$ where $s_e$ is the standard error of the regression with $\ln y$ as the dependent variable. That is, we use $\widetilde{y} = \exp(s_e^2/2) \times \exp(\widehat{\ln y})$.

Consider the current log-earnings regression example. If we form the incorrect prediction $\widehat{y} = \exp(\widehat{\ln y})$ for each observation in the sample, then the sample average of $\widehat{y}$ equals \$51,214 which is much less than the sample average of earnings which equals \$63,476. The log-earnings regression has $s_e = 0.624$ so the correction factor $\exp(s_e^2/2) = 1.215$. Multiplying $\widehat{y}$ by this yields the corrected prediction $\widetilde{y}$ with sample average of \$62,223, much closer to the sample average of earnings.

The in-sample correlation of earnings with the corrected prediction from the log-earnings regression is 0.452. By contrast, if we predict from the regression with the level of earnings as the dependent variable then the in-sample correlation of the prediction with earnings is 0.440, which is not as high.

**Remark 156** *Retransformation bias arises if predictions from a regression with the dependent variable in natural logarithms are used to predict the level of the dependent variable. A correction is possible if it is assumed that model errors are homoskedastic and normally distributed.*

## 16.7 Key Concepts

1. In multiple regression the coefficient on an indicator variable acts as an intercept shifter.

2. An interacted indicator variable is a regressor that is the product of an indicator variable and another regressor. This enables slope coefficients to vary according to the value of the indicator variable.

3. For a categorical variable that takes one of several possible values we form a set of indicator variables, one for each category.

4. The dummy variable trap arises if two or more indicator variables are included and they exhaust all possibilities so that their sum equals one for every observation. Then the intercept or one of the indicator variables need to be dropped. These solutions lead to equivalent estimates of the difference in $y$ across the two categories, and equivalent slope coefficients of other regressors.

5. The coefficient of an included indicator variable measures the marginal effect of being in that category compared to the base category, the omitted category, after controlling for the other regressors.

6. A $t$ test of the statistical significance of a single indicator variable is a test of whether the marginal effect of that category differs from that for the base category. It is not a test of whether the marginal effect of that category is zero.

7. An $F$ test of the joint statistical significance of the included indicator variables is a test of whether the set of indicator variables is statistically significant. This joint $F$ test leads to the same result regardless of the category that is dropped.

8. The marginal effect in the quadratic model $b_1+b_2x+b_3x^2$ equals $b_2+2b_3x$ using calculus methods.

9. When a variable appears directly as a regressor and additionally as an interacted regressor the marginal effect of the variable includes an effect through the interaction term. A test of the statistical significance of the variable is an $F$ test of the joint statistical significance of the regressor and the interaction term. A $t$ test of statistical significance of the interacted regressor is a test of whether the interaction term is needed.

10. The coefficient of the indicator variable then measures the marginal effect of changing from one category to the other.

11. Key Terms: indicator variable; dummy variable; interacted indicator variable; categorical variable; sets of indicator variables; mutually exclusive; dummy variable trap; base category; reference category; no intercept; difference in means; joint $F$ test; difference in means; quadratic model; marginal effect; polynomial model; interacted regressor; natural logarithm; semi-elasticity; elasticity; retransformation bias.

# Chapter 17

# © A. Colin Cameron: Data and Model Misspecification

In this chapter we present several complications that can arise in multiple regression analysis.

The estimates may be unduly influenced by one or more observations, possibly due to miscoding of the data. The variation in the regressors may not be enough to enable precise estimation. The model specified may be a poor model for the data. And even if the model is the appropriate mode, default standard errors are incorrect if model errors are not homoskedastic and uncorrelated.

## 17.1 Data Example: Democracy and Growth

Dataset DEMOCRACY is used to illustrate use of various model diagnostics. This dataset comes from the article by Acemoglu, Johnson, Robinson, and Yared (2008) that investigates whether increased income in a country leads to increased democracy. The OLS regression analyzed is the regression for democracy in the very long run given in the fourth column of Table 8A of the article. (Daron Acemoglu, Simon Johnson, James A. Robinson, and Pierre Yared (2008), "Income and Democracy," American Economic Review, Vol.98, pp. 808-42.)

Table 17.1 presents variable definitions and summary statistics for the data on 131 countries. The original dataset had 173 observations. Only 135 observations had data on the dependent variable *Democracy*, and an additional four observations are lost due to missing data on other variables.

The level of democracy in 2000 is based on a Polity Composite Democracy Index that is normalized to be between zero and one, with higher values corresponding to greater democracy. The level of democracy in 1500 for many countries is zero, including those countries that did not exist in 1500, though some countries have non-zero democracy if there was some constraint on the executive. The variable *Democracy*, measuring the 500-year change in the

Table 17.1: Democracy and Growth: Variable definitions and summary statistics.

| Variable | Definition | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|---|
| *Democracy* | 500 year democracy change (1500-2000) | 0.647 | 0.3310 | 0.0 | 1.0 |
| *Growth* | 500 year income per capita change (1500-2000) | 1.916 | 1.108 | -0.089 | 4.253 |
| *Constraint* | Constraint on the executive at independence | 0.372 | 0.3622 | 0.0 | 1.0 |
| *Indcent* | Year of independence / 100 | 19.044 | 0.677 | 18.00 | 19.77 |
| *Catholic* | Catholics proportion of population in 1980 | 0.305 | 0.355 | 0.0 | 0.969 |
| *Muslim* | Muslim proportion of population in 1980 | 0.250 | 0.371 | 0.0 | 0.997 |
| *Protestant* | Protestant proportion of population in 1980 | 0.127 | 0.213 | 0.0 | 0.978 |
| *Other* | Other religion proportion of population in 1980 | 0.320 | 0.320 | 0.001 | 1.000 |
| n | 131 | | | | |

level of democracy, is between zero and one as no country saw a reduction in the level of democracy (Qatar and Saudi Arabia remained at zero) while 20 countries went from zero to one.

The variable *Growth* measures the 500-year change in log GDP per capita. If the natural logarithm changes by amount $g$ then the level has multiplied by amount $e^g$. For example, a two-unit increase in variable *Growth* means that GDP per capita is $e^2 = 7.389$ times higher in 2000 than in 1500.

Bivariate OLS regression yields fitted model

$$\widehat{Democracy} = \underset{(0.046)}{0.397} + \underset{(0.019)}{0.131}\,Growth, \quad R^2 = .192, \quad n = 131,$$

where heteroskedastic-robust standard errors are given in parentheses. Figure 17.1 presents a scatter plot with the fitted regression line.

The variable *Growth* is highly statistically significant, with $t = .131/.024 = 5.46$. The magnitude of the coefficient is difficult to gauge as the variable *Democracy* is an index rather than inherently measurable variable. One way to interpret the result is to note that $R^2$ is the squared correlation coefficient $r^2$ as this is a bivariate regression. Since $r^2 = 0.192$, a one standard deviation change in *Growth* is associated with a $\sqrt{.192} = 0.44$ standard deviation change in *Democracy*, a large effect.

The variable *Constraint* takes a range of values between zero and one, with the most common values 0, 1/3 and 1 for, respectively, 38, 29 and 24 of the 131 observations. The four religious affiliation variables sum to one so one category, in this case other religions, needs to be omitted.

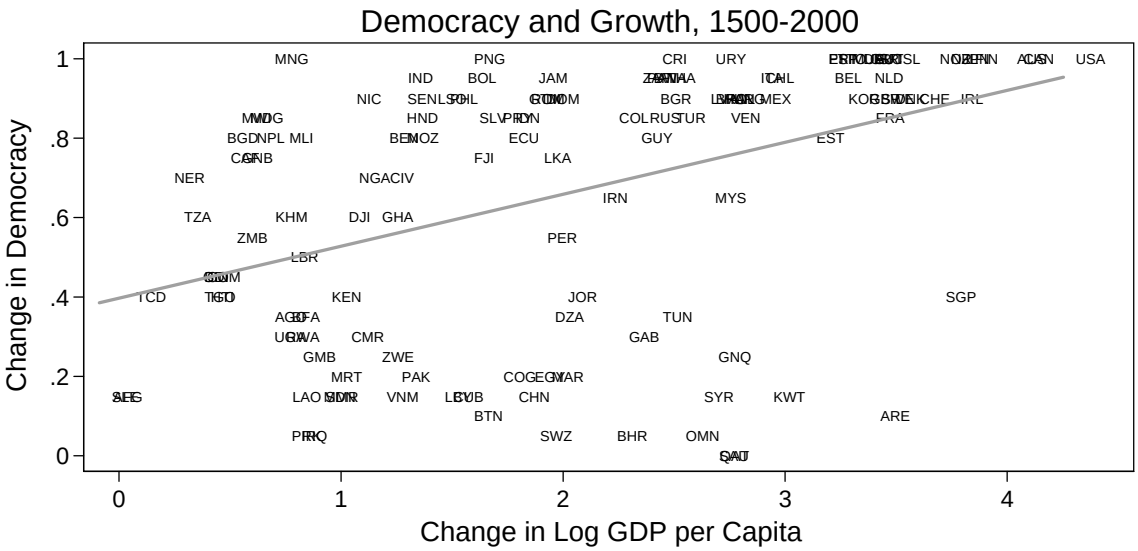


Figure 17.1: Democracy and Growth: Scatter plot

The complete multivariate fitted model is

$$\begin{aligned}\widehat{Democracy} &= \underset{(0.870)}{3.031} + \underset{(0.025)}{0.047}\,Growth + \underset{(0.072)}{0.164}\,Constraint - \underset{(0.050)}{0.133}\,Indcent \\ &\quad + \underset{(0.089)}{0.117}\,Catholic - \underset{(0.101)}{0.233}\,Muslim + \underset{(0.180)}{0.180}\,Protestant, \qquad R^2 = .192, \quad n = 131,\end{aligned}$$

where heteroskedastic-robust standard errors are given in parentheses. Default standard errors are within twenty percent of the default standard errors.

Now *Growth* is much less statistically significant, with $t = .047/.025 = 1.87$. More importantly the coefficient of variable *Growth* is now one-third as large. Controlling for institutional features such as religion is important, the main point of the article.

## 17.2 Model Diagnostics

THIS SECTION IS INCOMPLETE

A discussion of general issues is given before presenting several methods for detecting potential data problems and model misspecification.

### 17.2.1 Outliers, Influential Observations and Misspecified Model

An **outlying observation** is one whose value is unusual given the rest of the data. For a single variable an outlier can be detected using a box plot or box-and-whisker plot presented in Chapter 2.1. With data on several variables detection of an outlier is less straightforward. It can be more difficult to detect an outlier and, going the other way, what seems to be an outlier when one variable is considered in isolation may not be an outlier given data on the other variables. For example, someone with unusually high earnings might be viewed as an outlier given only earnings data. But if the person is a heart surgeon, and regressors include detailed occupation dummies, then the observation would not be viewed as an outlier. Going the other way, if the same heart surgeon had earnings similar to those of a fast food worker then the observation would be an outlier, but looking at earnings alone would not detect this.

A **misspecified model** is one that does not satisfy the population assumptions necessary for consistent estimation and correct standard errors. The essential requirements are model Assumptions 1-2 that the model has the correct functional form with errors that are uncorrelated with the regressors. Assumptions 3-4 that the model errors are homoskedastic and independent are less important if inference is based on appropriate robust standard errors.

An **influential observation** is one that has a relatively large effect on the results of regression analysis predicted values of the dependent variable, or on the estimated OLS coefficients.

There is no clearly best method to diagnose these problems. The methods given in the remainder of this section are the methods most commonly used. They are generally developed to identify one of the complications, but can also shed light on other complications.

If diagnostics reveal that an observation is miscoded then it should obviously be coded correctly or dropped if this is not possible.

If the data are valid but an influential observation then the standard procedure in economic data analysis is to still include the observation(s) in the analysis, though perhaps also do analysis without the observation(s) excluded as a robustness check.

If the model is determined to be misspecified then more flexible models may be considered, by adding quadratic terms, interaction terms or additional regressors. Or a different functional form may be considers, such as through natural logarithm transformation.

### 17.2.2 Scatter Plots against the Fitted Values

In bivariate regression a useful diagnostic plot is a scatter plot of $y$ against $x$ superimposed with the fitted regression line, such as that given in Figure 17.1. Outlying observations are those a long way from the regression line, and influential observations are likely to be those that are a long way from both $\bar{x}$ and $\bar{y}$.

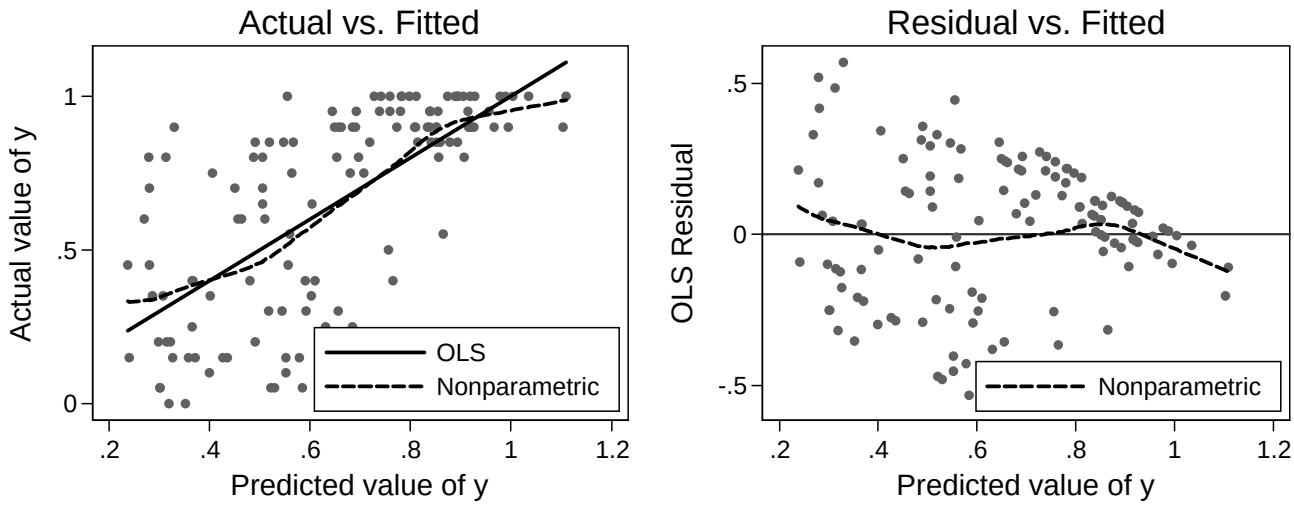


Figure 17.2: Democracy regression: Actual values plotted against fitted values and OLS residuals plotted against fitted values

For multivariate regression such a plot is no longer possible, as there is then more than one regressor. Instead, commonly-used diagnostic plots are to plot the actual value of $y$ against the fitted value $\widehat{y}$, or to plot the the residual $y - \widehat{y}$ against the fitted value $\widehat{y}$.

The first panel of Figure 17.2 plots the actual value of democracy against the fitted value of democracy from the preceding OLS multiple regression that has six regressors. The solid line is obtained by OLS regression of $y$ on $\widehat{y}$. The dashed line is obtained by nonparametric regression of $y$ on $\widehat{y}$, using Lowess. There seem to be no outliers and the Lowess curve suggests that the relationship between actual and fitted values is linear, so there is no sign of misspecification of the conditional mean.

The second panel of 17.1 plots the residual from the earnings regression against the fitted value of earnings. This scatter plot is a rotation of the scatter plot in the first panel, since it is a plot of $y - \widehat{y}$ against $\widehat{y}$ whereas the first panel is a plot of $y$ against $\widehat{y}$. Since an intercept was included in the OLS regression the residuals are on average zero, and regressing $e$ on $\widehat{y}$ yields a line with intercept zero and slope zero. The dashed line is obtained by nonparametric regression of $e$ on $\widehat{y}$, using Lowess. The second panel suggest that the errors are heteroskedastic, with more variability for low values of $\widehat{y}$. It is best to use heteroskedastic-robust standard errors for this model, though as noted earlier these are within twenty percent of default standard errors.

### 17.2.3 Residual Plots

In addition to histogram or kernel density plot of residuals, do a qqplot of residuals against the normal distribution.

And can plot residuals against each regressor.

### 17.2.4 Partial Regression Plots

A **partial regression plot** is a plot of the dependent variable against the $j^{th}$ regressor after taking account of the effect of the other regressors.

Let $\widetilde{x}_{ji}$ denote the residual from regressing $x_j$ on an intercept and all regressors other than $x_j$. Thus $\widetilde{x}_{ji} = x_{ji} - \widehat{x}_{ji}$ where $\widehat{x}_{ji}$ is obtained from the auxiliary multiple regression of $x_j$ on an intercept, $x_2, ..., x_{j-1,}, x_{j+1}, ..., x_k$. And let $\widetilde{y}_i$ denote the residual from regressing $y$ on an intercept and all regressors other than $x_j$. Thus $\widetilde{y}_i = y_i - \widehat{y}_i$ where $\widehat{y}_i$ is obtained from the auxiliary multiple regression of $y$ on an intercept, $x_2, ..., x_{j-1,}, x_{j+1}, ..., x_k$. Then a **partial regression plot** or **added variable plot** is a plot of $\widetilde{y}$ on $\widetilde{x}_j$. It can be shown that if we fit this scatter plot with a bivariate regression line then the slope coefficient equals that of $x_j$ from the original multiple OLS regression of $y$ on an intercept and $x_2, ..., x_k$. And the residuals from the two methods coincide.

### 17.2.5 Component plus Residual Plots

A **component plus residual plot** or **partial regression plot** is a rotation of the plot of the residual against the regressor. Thus for the $j^{th}$ regressor plot $pr_{ji}$ against $x_{ji}$ where $pr_{ji} = b_j x_{ji} + e_i$, $x_{ji}$ is the $i^{th}$ observation for regressor $j$, $b_j$ is the OLS coefficient and $e_i$ is the OLS residual for the $i^{th}$ observation.

### 17.2.6 Influential Observations

DFITS to detect observations that has big influence on the predictions $\widehat{y}$.

DFBETA to detect observations that has big influence on the OLS slope coefficients $b_j$.

## 17.3 Model Specification Tests

THIS SECTION IS INCOMPLETE

The **RESET test** is a simple test for whether the model is correctly specified is to add powers of the OLS predictions as extra regressors. For example, let $\widehat{y}$ be the predicted value of $y$ from regression of $y$ on an intercept and $x_2, ..., x_k$. Then add $\widehat{y}^2$ and $\widehat{y}^3$ as regressors and estimate

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \alpha_1 \widehat{y}^2 + +\alpha_2 \widehat{y}^3 + u.$$

Test $H_0 : \alpha_2 = 0, \alpha_3 = 0$. Rejection of $H_0$ is interpreted as model misspecification. Note that $\widehat{y}$ is not included because it is a linear combination of the other regressors so it is perfectly collinear with the other regressors.

Various tests for heteroskedastic errors exist; see Chapter 19.2.

## 17.4 Data Problems: Multicollinearity

OLS estimation requires that there is sufficient variation in the regressors to enable computation of the regression coefficients. This section expands on the initial discussion of this topic in Section 13.8.

### 17.4.1 Perfect Collinearity

**Perfect collinearity** arises when one (or more) of the regressors can be expressed as an exact linear combination of the other regressors. A simple example is the dummy variable trap, see Chapter 16.3. With three mutually exclusive categories, for example, the three indicator variables satisfy $d1 + d2 + d3 = 1$, so $d1$ can be expressed as a linear combination of $d2$ and $d3$ since $d1 = 1 - d2 - d3$. A second more subtle example is given in Chapter 13.8. If regressors are perfectly correlated then one or more of the regressors needs to be dropped.

Perfect collinearity effects estimation of only the coefficients of the regressors that are collinear, and usually these are just a subset of the regressors, such as in the dummy variable trap example.

### 17.4.2 Multicollinearity

A less extreme situation arises when one (or more) of the regressors is very close to equalling a linear combination of the other regressors. For example, in an earnings regression data may be available on a persons age (*Age*), years of education (*Education*) and years of work experience (*Experience*). Then we expect that

$$Experience \simeq Age - Education - 6,$$

assuming that education begins at age six and work experience begins immediately after the end of education. In this case it will be difficult to disentangle the separate roles of age, education and years of work experience.

**Multicollinearity** is the situation where the regressors are close to being linearly dependent. Then least squares regression is still valid, but it is difficult to separately identify the impact of each individual regressor. Then individual coefficients may be statistically insignificant due to relatively large standard errors. and estimates of individual coefficients

can be numerically unstable so that adding another observation can lead to quite different estimated coefficients.

Signs of multicollinearity are high standard errors, low $t$-statistics and "wrong" signs. But these are also indications of a poor model even in the absence of multicollinearity. How do we determine that our problems are those of multicollinearity? The problem is likely to be one of multicollinearity if overall $F$ tests are statistically significant at the same time that individual $t$ tests are insignificant. A simple diagnostic method is to regress one regressor on the remaining regressors. IF the $R^2$ is very high then multicollinearity is a problem (and if $R^2 = 1$ then there is perfect collinearity).

Multicollinearity is not a problem if it is confined to regressors that are included only as controls and does not spill over to regressors of intrinsic interest. And it is not a problem if interest lies solely in predicting the dependent variable, rather than in the role of individual regressors.

Solutions to multicollinearity include the following. If only a subset of regressors are highly correlated, do joint inference ($F$ tests) on this subset. Put restrictions on the coefficients of variables, usually by dropping variable(s), in recognition that the data are being "pushed" too far. For experimental data design the experiment to minimize the correlation of regressors.

Most fundamentally, get more data. The problem with multicollinearity is not that OLS is biased or inconsistent. The problem is a lack of estimator precision, and precision improves with additional data.

Multicollinearity is a problem particularly in time series analysis using aggregate data which tend to move together. The preferred solution is to get more data, since the precision of estimation increases with more observations. Failing this, the usual solution is to realize that there are too many variables in the model, the data are being "pushed" too far, and to drop some of these variables.

## 17.5 Failure of Model Assumptions

Statistical inference is based on assumptions about the population model that generates the data. These assumptions lead to the distribution of the regression coefficients (sampling theory). This distribution is then used for confidence intervals and hypothesis testing.

Statistical inference for multiple regression is based on the following assumptions

1. The population model is $y = \beta_1 + \beta_2 x + \beta_3 x_3 + \cdots + \beta_k x_k + u$.

2. The error for the $i^{th}$ observation has zero mean conditional on all regressors and conditional on all regressors.

3. The error for the $i^{th}$ observation has constant variance $\sigma_u^2$ conditional on the regressors.

4. The errors for different observations are statistically independent.

When any of these assumptions are not met the standard analysis needs to be modified to permit valid inference. In some cases OLS is inconsistent, a major problem. In other cases OLS is still consistent, but $t$-statistics and $p$-values are incorrect.

To see the issues involved consider OLS regression with just one regressor in addition to the intercept; a more detailed analysis is given in Appendix 9.C.

Suppose assumption 1 holds. Then after some algebra it can be shown that the OLS slope coefficient $b_2$ can be re-expressed as

$$b_2 = \beta_2 + \sum\nolimits_{i=1}^{n} w_i u_i,$$

where the weight $w_i = ((x_i - \bar{x})) \,/\, (\sum_{i=1}^{n}(x_i - \bar{x})^2)$ depends only on the regressors.

This representation is very useful as it makes clear that, given assumption 1, the OLS estimator $b_2$ equals $\beta_2$ plus a weighted sum of the error terms $u_i$. The properties of $b_2$ as an estimator of $\beta_2$ will therefore depend crucially on assumptions about the error terms.

For OLS to be unbiased, i.e. $\mathrm{E}[b_2] = \beta_2$, the second term on the right-hand side must have expected value 0. This requires assumption 1, already used to obtain the above expression for $b_2$, and assumption 2, so that $\mathrm{E}[u_i|x_i] = 0$. Going the other way, the OLS estimator is biased and inconsistent if either assumption 1 or 2 fail.

Now consider the variance of the OLS estimator. If assumptions 3 and 4 hold, in addition to assumptions 1 and 2, then the $u_i$ are independent of each other with constant variance $\sigma_u^2$, and some considerable algebra yields $\mathrm{V}[b_2] = \sigma_u^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$. Failure of either assumption 3 or 4 leads to a different expression for $\mathrm{V}[b_2]$.

Finally, consider the distribution of the OLS estimator. From the equation it is $\beta_2$ plus a weighted sum of the $u_i$. If errors $u_i$ are normally distributed, then the weighted sum and hence $\beta_2$ is also normally distributed. If instead errors are nonnormal, the usual case, we rely on a central limit theorem so that if the sample is large then the normal distribution still provides a good approximation.

The remainder of this chapter considers in turn the failure of assumptions 1 through 4. If either assumption 1 or 2 fails, OLS is inconsistent. This is a fundamental problem that is not easily overcome. If assumptions 3 or 4 fail, OLS is unbiased but the usual estimate of the variance of $b_2$, and hence the standard error, $t$-statistic and $p$-value, and appropriate robust standard errors need to be used.

## 17.6 Incorrect Population Model

Assumption 1 specifies that

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u.$$

This specification may be wrong due to omission of regressors, unnecessary inclusion of regressors, or incorrect functional form of the model.

### 17.6.1 Omitted Variables

Suppose for example that the true population model has an additional regressor $x_{k+1}$, so that

$$y = \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_k x_k + \alpha_{k+1} x_{k+1} + u.$$

It can be shown that if we regress $y$ on $x_2, ..., x_k$ and omit $x_{k+1}$, then the least squares coefficient estimates $b_2, ..., b_k$ are all biased, unless the omitted variable $x_{k+1}$ is uncorrelated with the included regressors $x_2, ..., x_k$.

In fact it can be shown that

$$\mathrm{E}[b_j] = \alpha_j + \gamma_j \alpha_{k+1},$$

where $\gamma_j$ is the coefficient of $x_j$ from OLS regression of the omitted regressor $x_{k+1}$ on the included regressors $x_2, ..., x_k$. For example, if $x_j$ and $x_{k+1}$ are both positively related with $y$ and with each other the true effect $\alpha_j$ of $x_j$, is being overstated since $\gamma_j \alpha_{k+1} > 0$. Such bias is called **omitted variables bias**.

As a concrete example, suppose that the true model is

$$Earnings = \alpha_1 + \alpha_2 Education + \alpha_3 Ability + u,$$

where *Education* is years of schooling and *Ability* is a measure of raw ability. Studies usually regress *Earnings* on only *Education*, because a measure of ability such as an IQ score is not available. That is we estimate

$$Earnings = b_1 + b_2 Education.$$

Since *Earnings* and *Ability* are both positively related with *Education* and with each other, we get the result that a regression of *Earnings* on *Education* alone will overstate the effect of being in school. This accords with intuition - people with more schooling on average have higher raw ability and this ability effect is being ascribed to length of time in school.

Omitted variables is clearly a serious problem, unless the omitted variable is uncorrelated with the variables included. It is possible in some cases to know the direction of the bias, however, as in the above example. If $b_2 = 500$, for example, then we know that an extra year of schooling leads to at most an additional \$500 of earnings.

### 17.6.2 Irrelevant Variables

Now the problem is that the true population model should omit the regressor $x_k$, so that

$$y = \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{k-1} x_{k-1} + u.$$

Then it can be shown that if we regress $y$ on $x_2, ..., x_k$, erroneously including the **irrelevant variable** $x_k$, then the least squares coefficient estimates $b_2, ..., b_{k-1}$ are unbiased but are not as precisely estimated.

Thus the only problem is a loss of precision in estimation – the standard errors will be higher and the $t$ statistics will be lower than in the model that correctly excludes irrelevant regressors.

Combining the effects of omitted variables and irrelevant variables, it is best to be on the cautious side, including questionable regressors, though bearing in mind that including too many regressors decreases the precision of estimates and may lead to problems with multicollinearity. For this reason regression models for economics data tend to include many regressors.

### 17.6.3 Wrong Functional Form

Now the problem is that

$$y = f(x_2, x_3, x_k, u),$$

but this relationship need not be linear. Then the least squares estimates from regression of $y$ on $x_2, ..., x_k$ are biased, since a model with **incorrect functional form** has been estimated.

As an example it may be that we estimate a linear regression of $y$ on $x_2, ..., x_k$, but instead we should have estimated a log-linear model regressing $\ln y$ on $x_2, ..., x_k$.

If the functional form $f(\cdot)$ is known, and is not as simple as models such as log-linear or linear-log, estimation is still possible using the method of nonlinear least squares. Tests of whether the correct function $f(\cdot)$ is specified are available. And even if $f(\cdot)$ is not specified nonparametric regression methods such as kernel regression permit estimation of $f(\cdot)$, though these methods are most practical when there is only one regressor. Such estimation methods are usually studied in graduate level courses.

## 17.7 Regressors Correlated with Errors

Assumption 2 is that the errors are unrelated with the regressor, so

$$\mathrm{E}[u|x_2, ..., x_k] = 0.$$

Correlation of regressors with errors is the most serious problem. It leads to biased and inconsistent estimates. Methods to control for this problem are presented in Chapter 18. These methods require additional assumptions that in many applications cannot be justified, and in many instances cannot be tested. So if errors are correlated with regressors it is possible to adapt methods or the model to get consistent estimates in only a subset of the applications.

A general term for correlation of regressors with errors is **endogeneity** or **endogenous regressors**, where the term "endogenous" means caused by factors inside the system. Regressors that are uncorrelated with the error are instead called exogenous regressors, where "exogenous" means outside the system. The key notion is that while we may consider estimation of a single equation, the data generating process may be more complex than this single equation. Instead the data may be generated by a system of equations. We now present several leading examples of such systems.

First, **omitted variables** is an example of the endogenous regressor problem that has already been discussed. The error term will be correlated with regressors since the error term will include the omitted regressor and the omitted regressor is correlated with included regressors.

Second, (unmodelled) **feedback** from the dependent variable to the regressors also leads to correlation between the regressors and error term. The leading example in econometrics is **simultaneous equations bias**. A simple example is the consumption-income relationship using national aggregate data. Increases in income lead to increases in consumption, but these increases in consumption in turn lead to an increase in income since income in a closed-economy is the sum of consumption, investment and government spending.

A more complicated example of feedback is the estimation of a demand relationship, with price as a function of quantity demanded, using aggregate data. An increase in demand leads to an increase in price, but an increase in price will then lead to an increase in supply which will then depress the price.

Third, if some regressors are measured with error the error term can be shown to be correlated with the regressors leading to **measurement error bias**.

Fourth, if the sample is selected on the basis of values of taken by the dependent variable then it can be shown that the error term will be correlated with regressors. This is called **sample selection bias**.

Fifth, models with lagged dependent variable and autocorrelated error can be shown to have regressor correlated with the error.

This type of misspecification can occur in many regression applications. A distinguishing feature of econometrics is the concern about this possible misspecification, and the use of methods that are nonetheless valid. These methods are presented in Chapter 18.

## 17.8 Nonconstant Error Variance

Assumption 3, the assumption that the error variance is constant for different observations and does not depend on regressors, is called the assumption of **homoskedastic errors**. This assumption is often not satisfied with cross-section data. When this assumption is not satisfied, i.e. the error variances differ across observations, we have **heteroskedastic errors**.

As an example, if we regress earnings on schooling we expect more variability around the regression line for people with high levels of schooling, such as a postgraduate degree, than for people with low levels of schooling, such as a high school only education. Thus the errors are heteroskedastic, with variance that increases with the schooling regressor.

The least squares coefficient estimates are still unbiased. But they are not estimated as precisely as they would be if the heteroskedasticity was controlled for. And more importantly, the standard errors and $t$ statistics reported by the usual computer output will be incorrect, unless heteroskedasticity is controlled for.

Heteroskedasticity most often arises with cross-section data. The simplest solution is to continue to use OLS but to obtain heteroskedastic-robust standard errors; see Chapter 14.6.

## 17.9 Correlated Errors

Assumption 4 is the assumption that the errors for different observations are uncorrelated. When this assumption fails but assumptions 1 and 2 still hold the OLS estimates remain unbiased and consistent, but the standard errors and resulting statistical inference are incorrect.

There are two common ways in which this assumption fails.

First, consider time series regression, such as a regression model for U.S. gross domestic product (GDP) using quarterly data. If GDP is overpredicted in one quarter, i.e. actual GDP is unusually low, then it is likely that GDP will be overpredicted in the subsequent quarter. Similarly for underprediction. But then the error terms in different quarters is positively correlated, and assumption 4 on the error term no longer holds. Time series correlated errors are called **autocorrelated errors** or **serially correlated errors**. The simplest solution is to continue to use OLS estimation and base inference on heteroskadastic and autocorrelation consistent (HAC) robust standard errors; see Chapter 14.6. This requires that the regressors do not include lagged values of the dependent variable.

Second, suppose that the error correlation structure is one of clustering where errors are correlated for observations in the same cluster but uncorrelated for observations in different clusters. One example is panel data on individuals, with errors uncorrelated across individuals but correlated over time for a given individual. The simplest solution is to continue to use OLS estimation and base inference on cluster robust standard errors; see Chapter 14.6. This requires that the regressors do not include lagged values of the dependent variable.

Assumption 4 can also fail with panel data. Solutions are discussed in Chapter 21.

The least squares coefficient estimates are still unbiased. But they are not estimated as precisely as they would be if the error correlation was controlled for. And more importantly, the standard errors and $t$-statistics reported by the usual computer output will be incorrect, unless error correlation is controlled for.

## 17.10 Conclusion

There are many potential pitfalls in using the regression methods presented in this course. The regression methods are always useful for data summary. The problems outlined in this chapter are ones of performing valid statistical inference.

The first step is to ensure that the dataset is free of error.

Next The first fundamental question is whether a valid model is being estimated. In particular do we have the correct model, with correct functional form and no omitted regressors, and are the regressors uncorrelated with the error.

The second question is whether correct standard errors, $t$ statistics and $p$ values are being used. Modifications to the usual ordinary least squares output in the presence of heteroskedastic errors or correlated errors are easily implemented in standard econometrics packages, though not in Excel. Statistical tests also exist to confirm the presence of heteroskedasticity and correlated errors.

The first consideration becomes of critical importance if we wish to go beyond summarizing correlation to ascribing causation. For example, we may regress the natural logarithm of earnings on years of schooling and find that an increase of one year of schooling *is associated with* a ten percent increase in earnings. This is a very useful summary measure of the association between earnings and education. But if we want to make the much stronger statement that one more year of schooling *leads*, on average, to a ten percent increase in earnings, much more care needs to be paid to ensure, for example, that other relevant variables such as ability are controlled for.

Students in this course should be aware of the limitations of their training. Hopefully some will be encouraged to take further courses in econometrics, to tackle the complications raised in this chapter. The issues are very important. And they are practically relevant as they are germane to studies using observational data.

## 17.11 Summary

Key Terms: Outlier; influential observation; influential observation; perfect collinearity; multiple collinearity; Omitted variables bias; irrelevant variables; incorrect functional form; homoskedastic error; heteroskedastic error; autocorrelated error; serially correlated error; endogenous variable; endogeneity; instrumental variables; feedback; simultaneous equations; measurement errors; sample selection bias; perfectly correlated; multicollinearity.

# Chapter 18

# © A. Colin Cameron: DRAFT Causality

**Chapter Outline**

1. Experiments
2. Regression
3. Differences in differences
4. Panel data
5. Matching
6. Weighting
7. Regression discontinuity design
8. Instrumental variables

Economic policy makers are often interested in how an outcome or outcomes change in response to a policy change. For example, interest may lie in how individual earnings change in response to participation in a training program. If the earnings response is large enough then it becomes economically worthwhile for government to subsidize the training program.

An obvious way to proceed is to use observational data and compare the earnings of those who received the training to those who did not, after using regression methods to control for observed individual characteristics such as age. This method is likely to overstate the benefits of training as those who selected to receive the training are likely to be those who felt that they would most benefit from the training. While regression can measure association, with observational data it will not necessarily measure the causative effect.

This chapter presents a range of methods that measure the causative effect. It is important to note that while many of the methods presented here are simple extensions of linear regression, the topic is an advanced one that entails many subtleties. Each method is reliant on assumptions, results may be limited in their generalizability, and the method can be employed depends very much on data availability.

## 18.1 Experiments

Lalonde data use nsw_dw.dta to compare treatment and control.

## 18.2 Regression

Use nsw_dw.dta treatments and psid_controls.dta for controls Why does this not have unemployment ?

Raw differences. Add in controls.

## 18.3 Differences in Differences

To come

## 18.4 Panel Data

To come.

## 18.5 Matching

Use distance from college example.

## 18.6 Weighting

To come

## 18.7 Regression discontinuity design

To come

## 18.8 Instrumental Variables

To come

# Chapter 19

# © A. Colin Cameron: Cross-section Data

Cross-section data are data observed on a number of individuals, such as firms or people or countries, observed at a single point in time. This chapter presents adaptation of linear regression to the most common issues that arise with cross-section data.

The most common complication is that model errors are usually heteroskedastic rather than homoskedastic. This complication is easily controlled for using heteroskedastic-robust standard errors, introduced in Chapters 10.7 and 14.6.

A second complication that arises for some data sets is that model errors may not be independent. A common example is cross-section data on individuals from several regions, such as village or state, with regression model errors correlated for individuals in the same region though uncorrelated for people in different regions. This form of error correlation is called clustering, and is controlled for using cluster-robust standard errors, introduced briefly in Chapters 10.7 and 14.6. It is important to make this adjustment if clustering is present, as default and heteroskedastic—robust standard errors can be much too small compared to the correct cluster-robust standard errors.

Regression methods are often applied to outcomes that are not continuously distributed. The most common example is a binary outcome that takes only two values, such as whether or not a person is employed. For such categorical data it is better to use a logit model or a probit model, the leading examples of a nonlinear model, rather than a linear regression model. More generally a wide range of models have been proposed for various different types of cross-section data; these are reviewed in an appendix to this chapter. A further appendix to the chapter presents a test for association between two variables when both variables are categorical.

Finally cross-section data often come from complex surveys that introduce the complication of nonrandom sampling.

Table 19.1: Annual Earnings of male full-time workers aged 25-65 in 2010: Variable definitions and summary statistics

| Variable | Definition | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|
| *Earnings* | Annual earnings | 56368.69 | 51516.69 | 4000 | 504000 |
| *Education* | Years of Schooling | 13.85 | 2.88 | 0 | 20 |
| *Age* | Age in years | 43.31 | 10.68 | 25 | 65 |
| *Agesq* | Age squared | 1989.67 | 935.69 | 625 | 4225 |
| *Gender* | Equals 1 if female and 0 if male | 0.16 | 0.37 | 0 | 1 |

Table 19.2: Earnings: OLS estimates with default standard errors.

| Variable | Coefficient | Default Standard Error | t statistic | p value | AME |
|---|---|---|---|---|---|
| *Education* | 5922 | 561 | 10.56 | 0.000 | 5921 |
| *Age* | 3226 | 1262 | 2.56 | 0.011 | 584 |
| *Agesq* | -30.5 | 14.4 | -2.12 | 0.034 | |
| *Gender* | -18979 | 3258 | -5.83 | 0.000 | -18979 |
| *Intercept* | -96446 | 27207 | -3.54 | 0.000 | |
| n | 872 | | | | |
| $R^2$ | 0.152 | | | | |

## 19.1 Data Example

As an example, consider regression models for annual earnings of 872 full-time workers in 2010 aged 25 to 65 years.

Table 19.1 presents summary statistics and variable definitions for the data in dataset EARNINGS_CH19. Mean earnings are \$56,369 per year and mean education is 13.85 years of schooling.

Table 19.2 presents results from OLS regression of earnings on the other variables. Earnings increase with education. Earnings increase with age to age 53 and then are decreasing in age (the turning point using Chapter 16.4 results is $-3226/(2 \times (-30.5)) = 53$). Women earn \$18,979 less than men per year, after controlling for education and age. The fitted model has $R^2 = 0.152$, so the model explains 15.2% of the variation in earnings.

The regressors are statistically significant at 0.05. And *Age* and *Agesq* are jointly highly statistical significant as a joint test, not given in Table 19.2, has a $p$-value of 0.000.

The final column gives the average marginal effect (AME), the sample-average of each individual's response to a change in, respectively, education, age and gender. For regressors

that appear linearly this is just the OLS coefficient. For variable *Age* that appears quadratically, from Chapter 16.4 the AME is $b_{Age}+2b_{Agesq}\overline{Age} = 3226-2\times30.5\times43.31 = 584$. Change in each of the regressors is associated with a substantial change in earnings. The AME for education is roughly ten times that for age, so aging ten years is similar to having one more year of education. And the AME for gender is roughly three times that for education, with opposite sign, so being female is similar to having three less years of education.

Table 19.2 uses default standard errors based on the assumption that the regression model errors are homoskedastic and independent, assumptions that are now relaxed.

## 19.2 Heteroskedastic Errors

Consider the linear regression model

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i,$$

where the error $u_i$ has conditional mean zero, so $\mathrm{E}[u_i|x_{2i}, ..., x_{ki}] = 0$, and is uncorrelated over $i$. The complication is that the error is heteroskedastic with

$$\sigma_i^2 = \mathrm{Var}[u_i|x_{2i}, ..., x_{ki}] \neq \sigma^2.$$

Thus assumptions 1, 2 and 4 are assumed to hold but not assumption 3.

Under these assumptions the OLS estimates are still unbiased, but default standard errors and the consequent $p$-values and $t$-statistics given in usual computer output will be incorrect. Instead, inference should be based on heteroskedastic-robust standard errors.

Provided the models errors are uncorrelated across observations the heteroskedastic-robust standard errors are consistent if model errors are either heteroskedastic or homoskedastic. It is now standard to report heteroskedastic-robust standard errors after OLS regression, even if there is no reason to suspect that errors are heteroskedastic.

### 19.2.1 Robust Inference with Heteroskedastic Errors

The standard response to heteroskedasticity is the simplest – continue to use the same OLS estimates, but base inference on heteroskedasticity-robust standard errors.

For bivariate OLS regression on an intercept and a single regressor the heteroskedasticity-robust standard error of the slope coefficient is

$$se_{het}(b_2) = \sqrt{\frac{n}{n-2}}\frac{\sqrt{\sum_{i=1}^n e_i^2(x_i-\bar{x})^2}}{\sum_{i=1}^n (x_i-\bar{x})^2},$$

where $\sqrt{n/(n-2)}$ is a degrees of freedom adjustment and $e_i$ is the OLS residual.

Table 19.3: Earnings: OLS regression with heteroskedastic-robust standard errors.

| Variable | Coefficient | Heteroskedastic-Robust Standard Error | t statistic | p value |
|---|---|---|---|---|
| Education | 5922 | 648 | 9.13 | 0.000 |
| Age | 3226 | 1089 | 2.96 | 0.003 |
| Agesq | -30.5 | 12.5 | -2.44 | 0.015 |
| *Gender* | -18979 | 3115 | -6.09 | 0.000 |
| Intercept | -96446 | 24341 | -3.96 | 0.000 |
| n | 872 | | | |
| $R^2$ | 0.152 | | | |

This method extends to multiple regressors. The formula, omitted as it requires use of matrix notation, then uses $\sqrt{n/(n-k)}$ as the degrees of freedom adjustment. The standard errors are often called more simply "robust" but it is best to call them heteroskedastic-robust as there are other forms of robust standard errors.

The $t$-statistic for test of $H_0 : \beta_j = \beta_j^*$ is $t = (b_j - \beta_j^*)/se_{het}(b_j)$ and is treated as being $T(n-k)$ distributed under the null hypothesis. As noted in Chapter 14.6, joint hypothesis tests are no longer based on sums of squared residuals. Instead specialized commands for hypothesis testing need to be used.

Table 19.3 presents inference for OLS with heteroskedastic-robust standard errors. The heteroskedastic-robust standard errors for the coefficients of *Education*, *Age*, *Agesq* and *Gender* are, respectively, 1.16, 0.86, 0.87 and 0.96 times the default standard errors given in Table 19.3. In general the heteroskedastic-robust standard errors may be larger or smaller than the default, and in many applications they are within 30% of the default. The $p$-values in Table 19.3 are based on the $T(857)$ distribution, since $n - k = 857$.

### 19.2.2 Detecting Heteroskedasticity

For completeness ways to detect heteroskedasticity are presented, even though this step is usually skipped and heteroskedastic-robust standard errors are routinely used following OLS regression.

It is well-known that earnings are more variable at higher levels of education. This is illustrated in the first panel of Figure 19.1, which plots earnings against education. In principle such variability may disappear after controlling for the regressors.

The second panel of Figure 19.1 plots the residual, from the regression given in Table 19.2, against years of education. A similar pattern is observed. This is not surprising as in this application the $R^2$ is low so that the regressors explained only a small part of the variation in earnings. The residual is more variable at higher levels of one of the regressors,

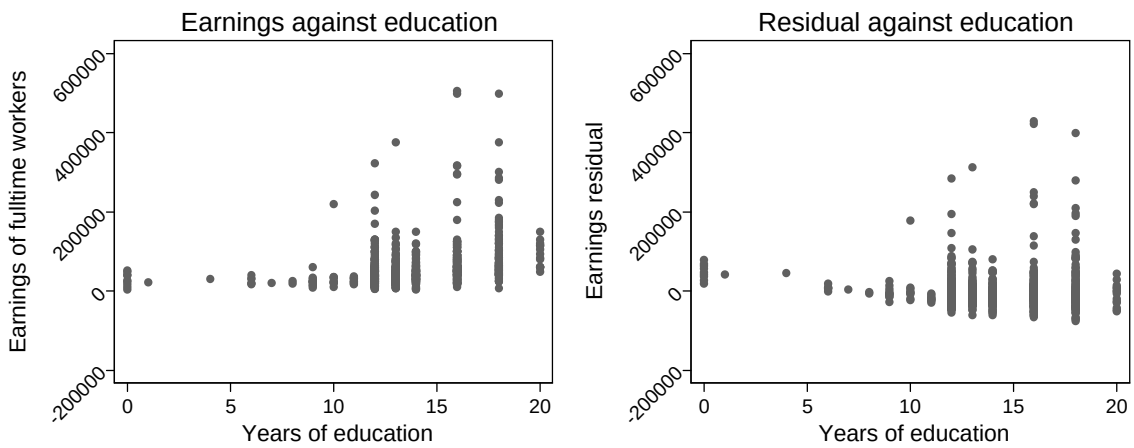


Figure 19.1: Heteroskedastic data: Scatter plots of earnings and earnings residual against education.

which suggests strongly that the model error is heteroskedastic.

Some care is needed in interpreting graphs such as Figure 19.1 as they visually underestimate the variability of $y$ in regions of $x$ values for which there are few observations, since there are then few observations on $y$ to vary. Similarly they will visually overestimate the variability of $y$ in regions of $x$ values for which there are many observations. In the current application, the right panel of Figure 19.1 visually underestimates the variability in earnings at low levels of schooling, since there are relatively few observations with less than 12 years of schooling. Nonetheless it is indeed the case that there is less variability in earnings at lower levels of schooling. For example, the standard deviation of earnings is 33,734 for those with 12 years or less of schooling and is a much higher 58,037 for those with more than 12 years of schooling. (When the regressors are included the corresponding values for the residual are 34,480 and 53,747.)

A formal test for heteroskedasticity is the following. Suppose assumption 3 is replaced with the assumption that

$$\sigma_i^2 = \text{Var}[u_i|x_{2i}, ..., x_{ki}] = h(\gamma_1 + \gamma_2 z_{2i} + \cdots + \gamma_m z_{mi}),$$

where $h(\cdot)$ is a continuously increasing function that need not be specified. Examples are $h(x) = x$ and $h(x) = \exp(x)$. The variables $z_2, ..., z_m$ in this model are usually the same as those in the original regression. Calculate $e_i^2/s_e^2$ where $e_i$ is the residual from the original OLS regression of $y$ on an intercept $x_2, ..., x_k$, and $s_e$ is the standard error of the residual from that regression. Regress $e_i^2/s_e^2$ on an intercept and $z_2, ..., z_k$ and perform an $F$ test of the joint statistical significance of $z_2, ..., z_m$. Large values of the $F$-statistic lead to rejection

of the null hypothesis of homoskedastic errors. This test is one of several variations of the original test proposed by Breusch and Pagan.

For the earnings regression, with heteroskedasticity posited to depend on all four regressors in the earnings regression, the auxiliary OLS regression for $e_i^2/s_e^2$ yields $F = 4.30$. The test statistic is $F(4, 867)$ distributed under the null hypothesis of homoskedasticity. Since $p = \Pr[F_{4,867} > 4.30] = 0.0019 < 0.05$, the null hypothesis is strongly rejected. The model errors are found to be heteroskedastic.

### 19.2.3 More Efficient Inference with Heteroskedastic Errors

If model errors are heteroskedastic then OLS is no longer the best estimator, as it is only fully efficient if the error $u_i$ is homoskedastic so that model assumptions 1-4 are satisfied.

Instead, more efficient estimation is possible using feasible generalized least squares estimation, introduced in Chapter 14.7. For heteroskedastic errors this leads to weighted least squares (WLS) estimation, presented in Appendix 19.A, where the weights are the inverse of the error variance.

WLS estimation is not often used, mainly because it typically does not lead to much improvement in estimator precision. As noted in Chapter 4.1, economics data are often right-skewed. Regression models for such data often do feature considerable heteroskedasticity. But it is standard for such data to do analysis in natural logarithms, in which case heteroskedasticity is greatly reduced.

**Remark 157** *If model errors are heteroskedastic then it is standard to continue to estimate by OLS, but heteroskedastic-robust standard errors need to be used. Formal tests for heteroskedasticity exist, and more efficient weighted least squares estimation is possible, but these methods are not used very often.*

## 19.3 Clustered Errors

A second complication arises if model errors are correlated in the following specific manner. Observations can be grouped into clusters, where model errors are correlated for observations in the same cluster, but are uncorrelated for observations in different clusters. For example, it may be that if a regression model overpredicts (or underpredicts) for one individual in a household or village or region then it is also likely to overpredict (or underpredict) for other individuals in the same household or village or region.

When errors are clustered, default OLS standard errors and heteroskedastic-robust OLS standard errors can be downwards biased, in some settings greatly so. The reason is that standard errors that assume model errors are uncorrelated assume that each new observation is a new piece of information. With errors that are positively correlated within cluster,

by contrast, a new observation in the cluster provides less than an independent piece of information.

This complication occurs in a minority of cross-section data applications. Its occurrence can be subtle and easily missed. It most often arises when a key regressor varies little within cluster. For example, the regressor may be a policy variable that only varies across regions, so is the same for all individuals in the same region. If additionally model errors are correlated within each region, even mildly so, then inference needs to use methods that control for clustered errors.

### 19.3.1 Regression with Clustered Errors

We again consider linear regression of $y$ on an intercept and $x_2, ..., x_k$. Assumption 2 that the error has conditional mean zero now needs to condition on all observations in the same cluster, so assumption 2 becomes

$$\mathrm{E}[u_i|x_{2i}, ..., x_{ki}, x_{2j}, ..., x_{kj}] = 0, \text{ for all } i \text{ and } j \text{ in same cluster.}$$

The key assumption is that errors for observations in different clusters are independent, but errors in the same cluster are potentially correlated. Thus assumption 4 is relaxed to the following:

$$\sigma_{ij} = \mathrm{Cor}[u_i, u_j|x_{2i}, ..., x_{ki}, x_{2j}, ..., x_{kj}] \begin{cases} \neq 0 & i \text{ and } j \text{ in same cluster} \\ = 0 & i \text{ and } j \text{ in different clusters.} \end{cases}$$

The error variances $\sigma_i^2$ and covariances $\sigma_{ij}$ may depend on the value of the regressors, so errors are also potentially heteroskedastic.

**Remark 158** *Regression model errors are said to be clustered if they can be grouped into clusters, with errors uncorrelated across clusters but potentially correlated within cluster.*

The OLS coefficient estimates remain unbiased and consistent, as the error still has zero mean conditional on the regressors. But the standard errors and $t$-statistics reported by the usual computer output will be incorrect, unless clustering is controlled for. Furthermore, better (more efficient) estimators than OLS may be possible.

How wrong are the OLS default standard errors? A rough guide is that if model errors are clustered then for the OLS coefficient of the $k^{th}$ regressor the default standard error should be inflated by

$$\tau_k \simeq \sqrt{1 + \rho_{x_k}\rho_u((n/G) - 1)},$$

where $\rho_{x_k}$ is a measure of the within-cluster correlation of $x_k$, $\rho_u$ is the within-cluster error correlation, and $n/G$ is the average number of observations per cluster. **Within-cluster correlation** of a variable is also called **intragroup correlation** or **intraclass correlation**.

The difference is especially great if $\rho_{x_k}$ is large. In the extreme case a regressor may take the same value for everyone in the same cluster, in which case $\rho_{x_k} = 1$. For example, suppose $\rho_{x_k} = 1$, $\rho_u = 0.1$ and $n/G = 81$. Then $\tau_k \simeq \sqrt{1 + 1 \times 0.1 \times 80} = 3$. So with 81 observations per cluster and a regressor that does not vary within cluster the correct standard error of the regressor's coefficient is approximately three times the default standard error, even though the errors were only weakly correlated within cluster ($\rho_u = 0.01$).

**Remark 159** *If model errors are clustered then default standard errors can be much too small. This downwards bias can be especially large if model errors are correlated within cluster, regressors are highly correlated within cluster, and there are many observations within each cluster.*

### 19.3.2 Cluster-Robust Standard Errors

It is convenient to define an indicator variable $\delta_{ij}$ that indicates whether or not two observations are in the same cluster:

$$\delta_{ij} = \begin{cases} = 1 & i \text{ and } j \text{ in same cluster} \\ = 0 & i \text{ and } j \text{ in different clusters.} \end{cases}$$

Then model errors are correlated if $\delta_{ij} = 1$ and are uncorrelated if $\delta_{ij} = 0$.

For OLS regression with an intercept and a single-regressor the **cluster-robust standard error** of the slope coefficient equals

$$se_{clu}(\widehat{\beta}_2) = \sqrt{\frac{G}{G-1}} \frac{\sqrt{\sum_{i=1}^n \sum_{j=1}^n \delta_{ij} e_i e_j (x_i - \bar{x})(x_j - \bar{x})}}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where $G$ is the number of clusters, $\sqrt{G/(G-1)}$ is a degrees of freedom adjustment, and $e_i$ is the OLS residual. The standard errors are called "cluster-robust" though they are additionally heteroskedastic-robust. The formula is an extension of that for heteroskedastic errors. With heteroskedastic errors $\delta_{ij}$ only equals one if $i = j$ so the term $\delta_{ij} e_i e_j (x_i - \bar{x})(x_j - \bar{x})$ simplifies to $e_i^2 (x_i - \bar{x})^2$.

The formula can be extended to multiple regressors but is omitted as it requires use of matrix notation. Some packages use $\sqrt{(n-1)G/\{(n-k)(G-1)\}}$ rather than $\sqrt{G/(G-1)}$ for degrees of freedom adjustment.

The $t$-statistic for test of $H_0 : \beta_j = \beta_j^*$ is $t = (b_j - \beta_j^*)/se_{clu}(b_j)$. It is best to treat the $t$-statistic as being $T(G-1)$ distributed, rather than standard normal distributed, especially if there are few clusters.

It is important to note that the underlying statistical theory requires that the number of clusters should be large, say greater than 30. It is not enough that the number of observations is large. If there are few clusters then tests based on cluster-robust standard errors tend to

reject the null hypothesis too often and confidence intervals are too narrow. But it is still much better to use cluster-robust standard errors, if there is clustering, rather than default of heteroskedastic-robust standard errors.

For statistical packages that provide cluster-robust standard errors one needs to define an additional variable that identifies which cluster the observation is in. For example, a variable for region may be used that takes a distinct value for each region.

**Remark 160** *If model errors are clustered then it is still okay to estimate by OLS, but cluster-robust standard errors need to be used. Confidence intervals and $t$-statistics then use the $T(G-1)$ distribution where $G$ is the number of clusters.*

### 19.3.3 Cluster-Robust Standard Errors Example: Earnings Data

For illustrative purposes, consider clustering on state in the current example where individuals are in 49 different states. The motivation is that earnings are systematically higher in some states, for example New York, and lower in other states, for example, Alabama. Thus some states may have earnings that on average are higher (or lower) than predicted by the regression model, meaning that the model errors for individuals in that state tend to be positive (or negative). This implies that there is positive correlation of the model error within state.

For the data in the current example there is actually very little within state correlation of the errors or regressors. Specifically, the within-state correlation for variables *Education*, *Age*, and *Gender*, and for the residual $e$ from the OLS regression, are all less than 0.2. On average there are $n/G = 872/49 \simeq 18$ observations per cluster. The standard error inflation formula is then at most $\sqrt{1 + 0.2 \times 0.2(18-1)} \simeq 1.30$, so we expect that standard errors will increase by at most 30% when we control for clustering. In other applications there can be a much greater difference.

Table 19.4 gives OLS estimates with cluster-robust standard errors and $p$-values based on the $T(48)$ distribution as $g - 1 = 48$. In this particular example the cluster-robust standard errors are within 25% of the default heteroskedastic-robust standard errors. In other applications the difference can be much, much larger.

### 19.3.4 Cluster-Specific Random Effects and Fixed Effects

For clustered data the standard procedure in economic data analysis is to use OLS estimates with cluster-robust standard errors. We briefly describe some alternative ways to handle clustering.

More efficient feasible GLS estimation is possible if a specific model for the within-cluster error correlation is specified. The standard feasible GLS estimator for clustered data is the **random effects estimator**, and can lead to greater efficiency improvement than occurs

Table 19.4: Earnings: OLS regression with cluster-robust standard errors.

| Variable | Coefficient | Heteroskedastic-Robust Standard Error | t statistic | p value |
|---|---|---|---|---|
| *Education* | 5922 | 481 | 12.32 | 0.000 |
| *Age* | 3226 | 1249 | 2.58 | 0.013 |
| *Agesq* | -30.5 | 14.3 | -2.13 | 0.038 |
| *Gender* | -18979 | 3415 | -5.56 | 0.000 |
| *Intercept* | -96446 | 26131 | -3.69 | 0.001 |
| n | 872 | | | |
| $R^2$ | 0.152 | | | |

with weighted least squares when errors are heteroskedastic. Detailed presentation of the random effects estimator, originally proposed for clustered data, is deferred to Chapter 21.3 for panel data. Here the cluster corresponds to the individual unit in the panel setting, and the individuals within the cluster correspond to the different time periods in panel data. The efficiency improvement in using feasible GLS can be more substantial with clustering than with heteroskedasticity.

The random effects model does impose a particular model for the within-cluster correlations, so to ensure valid inference one should use cluster-robust standard errors after random-effects estimation. Alternatively richer error correlation models, called hierarchical linear models, may be used. These are not used much in economics but are extensively used in other areas of applied statistics such as biostatistics and educational statistics.

The **fixed effects estimator**, presented in Chapter 21.4 for panel data, can also be applied to clustered data. As is the case for panel data, the main reason for using the fixed effects estimator is that it actually permits a limited form of correlation of the model error with the regressors – the error can be correlated with a component of the error that is common to all individuals in the cluster. See Chapter 21.4 for discussion of this important point.

**Remark 161** *The random effects and fixed effects estimators, presented in Chapter 21 for panel data, can also be applied to clustered data. Cluster-robust standard errors should be used after random effects estimation or fixed effects estimation to control for richer error correlation than specified in the model.*

## 19.4 Models for Binary Outcome Data

This introductory text focuses on linear regression models. To the extent that models are nonlinear the nonlinearity is introduced in a modest way, by transforming data to natural

logarithms or raising variables to a power. Now we consider a model that is more intrinsically nonlinear.

The most commonly-used nonlinear regression models are those developed specifically for dependent variable that is a **binary outcome**, so $y$ takes only two possible values. Examples are whether a person is employed or not employed, or whether a worker commutes to work by car or commutes by other means. Such data are categorical, see Chapter 1, and we model the **probability** that $y$ falls into one category or the other as a function of the regressors.

Binary outcome data are typically coded as $y$ taking values of one or zero. The goal for regression analysis is to model $\Pr[y = 1|x_2, ..., x_k]$. Since probabilities sum to one, $\Pr[y = 0|x_2, ..., x_k] = 1 - \Pr[y = 1|x_2, ..., x_k]$.

There are two standard models for such data, the logit model and the probit model. These models are presented, followed by a discussion of the usefulness of OLS regression for such data.

**Remark 162** *A binary outcome dependent variable takes one of two possible values. The two commonly-used models for binary outcomes are the logit and probit models.*

### 19.4.1 Example: Generated Data

Figure 19.2 provides scatter plots of a binary outcome $y$, coded to take values of 1 or 0, against a single regressor $x$, using generated data. Clearly there is a positive relationship.

The first panel also includes a fitted OLS regression. This line leads to predictions for some observations that exceed one or are less than zero, whereas a probability should lie between zero and one.

The second panel of Figure 19.2 provides the same scatter plot, along with a fitted logit model that now predicts probabilities to lie between zero to one. Because this curve is nonlinear, however, interpretation of estimated coefficients is more difficult than those from OLS regression.

**Remark 163** *A binary outcome model is a model of the probability that the outcome falls into one of the two categories, conditional on regressors. Then one minus this probability gives the probability of falling into the other category.*

### 19.4.2 Logit Model

For simplicity assume that the binary outcome is coded to take values one or zero. The **logit model** specifies that

$$\Pr[y = 1|x_2, ..., x_k] = \frac{\exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}.$$

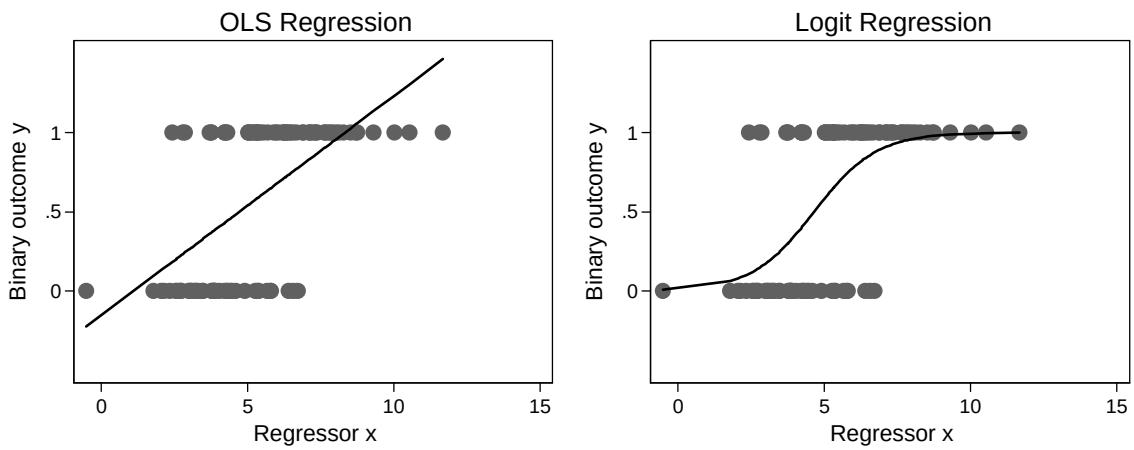


Figure 19.2: Binary outcome data: predictions from logit, probit and OLS models.

This model guarantees that the probability lies between zero and one, since the ratio on the right-hand side is of the form $a/(1+a)$ where $a > 0$. The model is also called the **logistic model**.

Most regression packages include a logit command that obtains the estimated coefficients $b_1, ...., b_k$ by a method called maximum likelihood estimation, along with their standard errors. For the logit model it can be shown that there is little difference between using default standard errors or heteroskedastic-robust standard errors. If observations are clustered, so that they are correlated within cluster but uncorrelated across clusters, then cluster-robust standard errors should be used.

Unlike OLS regression, the logit model coefficients do not equal the marginal effects. Letting $\widehat{p}$ denote the predicted probability obtained by replacing $\beta_1, ..., \beta_k$ in the logit model formula with $b_1, ..., b_k$. Then it can be shown that the marginal effect on the predicted probability that $y = 1$ of a one-unit change in the $j^{th}$ regressor is

$$\text{ME}_j = \frac{\Delta \widehat{p}}{\Delta x_j} = \widehat{p}(1 - \widehat{p})b_j.$$

Some computer packages provide a command to compute the marginal effect after logit estimation. Most often this is the average marginal effect (AME) that averages over individuals, with $\text{AME}_j = \frac{1}{n}\sum_{i=1}^{n} \widehat{p}_i(1 - \widehat{p}_i)b_j$.

There are several ways to interpret logit model coefficients without computing the marginal effects. First, a positive coefficient means a positive effect, since $\widehat{p}(1 - \widehat{p}) > 0$ always implies that $\text{ME}_j > 0$ if $b_j > 0$. Second, if one coefficient is twice another coefficient, then the corresponding marginal effect is twice as large. Third, the marginal effect is at most

0.25 times the estimated coefficient $b_j$, since $0 < \widehat{p}(1-\widehat{p}) < 0.25$ given $0 < \widehat{p} < 1$. Finally, the estimated coefficient $b_j$ exactly equals the change in the log-odds ratio when $x_j$ equals one unit, where the odds ratio is $\widehat{p}/(1-\widehat{p})$. This last interpretation is frequently used in the biostatistics literature where the odds ratio may be the odds of death compared to survival.

Assessing the goodness of fit for a logit model is difficult. Although some statistical packages report an $R^2$ or pseudo-$R^2$ this does not have the same interpretation as $R^2$ following least squares estimation. One diagnostic that is used is a classification table. This predicts $\widehat{y} = 1$ if $\widehat{p} \geq 0.5$ and $\widehat{y} = 0$ if $\widehat{p} > 0.5$ and compares these predictions to the actual value of $y$. It is not unusual to find that a logit model does a poor job of classification, however, especially if most of the sample has $y = 0$ (or $y = 1$).

**Remark 164** *For a logit model the sign of the estimated slope coefficient equals the sign of the marginal effect. If one regressor has coefficient that is twice as big as the other then the marginal effect is twice as big.*

### 19.4.3 Logit Model Example: High Earnings

As an example we use dataset EARNINGS_CH19 and consider the binary outcome of whether or not annual earnings exceed \$80,000, the case for 16.4% of observations in the sample.

Logit model estimates are given in the first column of Table 19.5. The signs of the coefficients imply that the probability of earnings exceeding \$80,000 increases with education, decreases with gender, and increases to age fifty and then decreases (the turning point is $-0.327/(2 \times -0.0033) \simeq 50$). Since the coefficient of variable *Gender* is minus 3.5 times that of variable *Education*, being female has similar association with the probability of earnings exceeding \$80,000 female as does having 3.5 less years of education.

The $t$-statistics are based on default standard errors. (The heteroskedastic default standard errors, not necessary for the logit model, are all within 5% of the default). All regressors are statistically significant at 5%.

Interest lies in $\widehat{p}$, the predicted probability that $y = 1$ given the regressors, and how this changes as the regressors change. For these data the predicted probabilities for each individual range from .002 to .779 with a mean of .164. The second column of Table 19.5 gives the average marginal effects. One more year of education, for example, is associated with a .0427 increase in the probability that earnings exceed \$80,000, or a 4.27 percentage point increase.

Table 19.6 presents a classification table based on comparing the predictions $\widehat{y} = 1$ if $\widehat{p} \geq 0.5$ and $\widehat{y} = 0$ if $\widehat{p} > 0.5$ to to the actual value of $y$. The logit model does a poor job of predicting ones, with only 34 of the 143 values of $y = 1$ correctly classified. This is not unusual. If most of the sample has $y = 0$ (or $y = 1$) then most or even all observations will be predicted to have $y = 0$ (or $y = 1$), unless the model does a great job of discriminating.

Table 19.5: Earnings: OLS, random effects and fixed effects estimates with default and cluster-robust standard errors.

| **Variable** | **Coefficients, t-statistic and Average marginal effect** | | | | | |
|---|---|---|---|---|---|---|
| | **Logit** | | **Probit** | | **OLS** | |
| | **Coeff / t** | **AME** | **Coeff / t** | **AME** | **Coeff / t** | **AME** |
| *Education* | 0.382 | .0427 | 0.214 | .0433 | 0.039 | .0393 |
| | (8.95) | | (9.10) | | (8.80) | |
| *Age* | 0.327 | .0035 | 0.174 | .0036 | 0.029 | .0045 |
| | (3.47) | | (3.45) | | (3.61) | |
| *Agesq* | -0.0033 | | -0.0017 | | -0.0003 | |
| | (-3.13) | | (-3.08) | | (-3.00) | |
| *Gender* | -1.143 | -.1278 | -0.628 | -.1270 | -0.135 | -.1348 |
| | (-5.17) | | (-5.27) | | (-5.86) | |
| *Intercept* | -14.570 | | -7.991 | | -1.007 | |
| | (-6.66) | | (-6.92) | | (-5.74) | |
| Predicted probabilities range | (0.002, 0.779) | | (0.000, 0.747) | | (-0.467, 0.514) | |
| n | 872 | 872 | 872 | 872 | 872 | 872 |
| $R^2$ | 0.152 | 0.152 | 0.152 | 0.152 | 0.212 | 0.212 |

Table 19.6: High earnings: Classification table folloiwng logit regression.

| | True | | |
|---|---|---|---|
| Classification | $y = 1$ | $y = 0$ | Total |
| $\widehat{y} = 1$ | 34 | 20 | 54 |
| $\widehat{y} = 0$ | 109 | 709 | 818 |
| Total | 143 | 729 | 862 |

This is the case for earnings – a large component of individual earnings is not explained by a regression model.

### 19.4.4 Probit Model

The logit model is not the only model that ensures that probabilities lie between 0 and 1. The other commonly-used model for binary outcomes is the probit model.

Again for simplicity assume that the binary outcome is coded to take values one or zero. The **probit model** specifies

$$\Pr[y = 1|x_2, ..., x_k] = \Phi(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The

formula for $\Phi(\cdot)$ is complicated, so is not given here, but $\Phi(\cdot)$ satisfies $0 < \Phi(\cdot) < 1$ so that the probability lies between zero and one. Estimation is again by maximum likelihood.

For the probit model it can be shown, using calculus methods, that for small $\Delta x_j$,

$$\mathrm{ME}_j = \phi(\widehat{p})b_j,$$

where $\phi(\cdot)$ is the standard normal density function. So a positive coefficient means a positive marginal effect, since $\phi(\cdot) > 0$. Again if one coefficient is twice as large as another then the marginal effect is twice as large. And for the probit model the average marginal effect is $\mathrm{AME}_j = \frac{1}{n}\sum_{i=1}^{n}\phi(\widehat{p}_i)$. Since $\phi(\cdot)$ takes a minimum value of $1/\sqrt{2\pi} \simeq 0.40$, it follows that $|\mathrm{ME}_j| \leq 0.40 \times |b_j|$.

Note that probit model coefficients are scaled differently from logit coefficients. In fact logit slope coefficients are very approximately 1.6 times probit slope coefficients.

**Remark 165** *The probit model leads to similar predicted probabilities and marginal effects to those from the logit model, but the coefficients are scaled differently.*

### 19.4.5 Probit Model Example: High Earnings

The third and fourth columns of Table 19.5 give results for probit regression using the same data as previously estimated by logit. The results are remarkably similar to those for logit, with similar $t$-statistics, AME, and range of predicted probabilities. The predicted probabilities from the two models are very similar – the correlation coefficient of the logit and probit predicted probabilities equals 0.9987.

The one notable difference is in the size of the probit and logit regression coefficients. This is because the two models are scaled quite differently. For example, it is not meaningful to compare coefficients in a model with $\mathrm{E}[y|x] = \beta_1 + \beta_2 x$ to those from a model with $\mathrm{E}[y|x] = \exp(\beta_1 + \beta_2 x)$, a quite different functional form for the conditional mean. Similarly, it is not meaningful to compare the coefficients from two quite different models for $\Pr[y = 1|x]$. In this example the logit slope coefficients are around 1.8 times the probit slope coefficient.

### 19.4.6 OLS Estimation

An alternative, simpler, method for binary outcomes is OLS regression.

Again assume that the binary outcome is coded to take values zero or one. In that case the conditional mean equals the conditional probability since, from Appendix 5.A, a Bernoulli random variable with probability $p$ has mean $p$. So the linear regression model with conditional mean $\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ can be viewed as a binary outcome model with linear conditional probability $\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. For this reason, when $y$ is a binary outcome the linear regression model is called the **linear probability model**. By contrast, the logit and probit models transform this linear function.

The last two columns of Table 19.5 give results from OLS regression for these data. The $t$-statistics, which for OLS applied to binary outcome data must be based on heteroskedastic-robust standard errors, and AME are similar to those obtained by logit and probit estimation. The coefficients are quite differently scaled to those from logit or probit. The logit estimates are almost ten times larger than OLS.

The limitation of OLS is evident in the predicted probabilities. These are as low as -0.467 and are negative for more than 5% of the sample. OLS should not be used for individual prediction, and will also do poorly in estimating the marginal effect at a given value of the regressors. OLS does produce an in-sample average marginal effect that is similar to that from logit or probit. So it can be good for a first pass analysis of the data and for summarizing average behavior.

**Remark 166** *For binary outcomes, OLS regression is a poor model for prediction of individual probabilities, but does provide a reasonable estimate of average marginal effects.*

For binary outcome data one should use a logit or probit model. The two models yield similar results. Many areas of applied statistics most often use the logit model. Economists more frequently use the probit model.

**Remark 167** *The logit and probit models can also be used for proportions data, such as the employment-population ratio in various regions, that are continuous between 0 and 1.*

## 19.5 Survey Data

Cross-section data often come from large surveys such as the American Community Survey that are not simple random samples.

Instead, for example, smaller states and minority populations are oversampled. Then, as discussed at the end of Chapter 5.7, the population mean of a single variable needs to be estimated using a sample-weighted mean, rather than an unweighted mean. Complex surveys provide the necessary sample weights.

The regression case is presented in Chapter 14.8. If the nonrandom sampling is only on regressors then it may be possible to continue to use OLS estimation, provided enough regressors are included in the model and the model is sufficiently flexible so that the remaining error in the model has zero mean conditional on the regressors. However, OLS cannot be used if the nonrandom sampling is on the dependent variable. If sample weights are available, then the weighted least squares estimator can be used.

It is important to note that there are two distinct variations of weighted least squares. The first, discussed in Chapter 19.2 and detailed in Appendix 19.A is that OLS is fine but weighting by the inverse of the error variance leads to more efficient estimation. The second, given in this section, is that OLS is actually inconsistent, due to nonrepresentative sampling,

but weighting by the inverse of the probability of inclusion in the sample leads to consistent estimation.

Another feature of complex surveys is that individuals are sampled in clusters, such as households or blocks of houses. This can substantially reduce survey costs, but means that the regression error is likely to be correlated for individuals in the same cluster. Then one should use cluster-robust standard errors, presented earlier in this chapter.

Specialized statistical software exists for complex survey data that enables weighted estimation and controls for clustered sampling and stratified sampling, an additional feature of complex surveys that can lead to somewhat more precise estimation. Such software is rarely used in economics. Instead economists use flexible regression models that control for nonrepresentative sampling of the regressors and base inference on cluster-robust standard errors. This approach is valid provided the nonrepresentative sampling is not directly on the dependent variable.

## 19.6 Key Concepts

1. If model errors are heteroskedastic (though still uncorrelated across observations) then it is standard to continue to estimate by OLS, but heteroskedastic-robust standard errors need to be used.

2. Formal tests for heteroskedasticity exist, and more efficient weighted least squares estimation is possible, but these methods are not used very often.

3. Regression model errors are said to be clustered if they can be grouped into clusters, with errors uncorrelated across clusters but potentially correlated within cluster.

4. If model errors are clustered then it is standard to continue to estimate by OLS, but cluster-robust standard errors need to be used.

5. For clustered errors the random effects and fixed effects estimators, presented in Chapter 21 for panel data, may also be used.

6. A binary outcome dependent variable takes one of two possible values.

7. A binary outcome model is a model of the probability that the outcome falls into one of the two categories, conditional on regressors. Then one minus this probability gives the probability of falling into the other category.

8. The two commonly-used models for binary outcomes are the logit and probit models.

9. For logit and probit models the sign of the estimated slope coefficient equals the sign of the marginal effect. If one regressor has coefficient that is twice as big as the other then the marginal effect is twice as big.

10. The logit and probit models lead to similar predicted probabilities and marginal effects, but the coefficients are scaled differently.

11. For binary outcomes, OLS regression is a poor model for prediction of individual probabilities, but does provide a reasonable estimate of average marginal effects.

12. Survey data are based on nonrepresentative sampling. In some cases OLS estimation is still possible provided sufficient controls are included in the regression. In other cases weighted least squares estimation using sample weights may be necessary.

13. For surveys that use clustered sampling, cluster-robust standard errors need to be used.

14. Key terms: heteroskedastic error; heteroskedastic-robust; weighted least squares; Breusch-Pagan test; clustered errors; within cluster correlation; cluster-robust; random effects; fixed effects; complex survey; binary outcome; conditional probability model; logit; probit; average marginal effect; OLS; linear probability model; survey data; nonrepresentative sample; weighted least squares.

## 19.7 Appendix 19.A: Weighted Least Squares

This appendix presents weighted least squares, a special case of feasible generalized least squares, to obtain an estimator that is more efficient than OLS if the errors are heteroskedastic. A quite different use of weighted least squares, to ensure consistent estimation, is presented in Chapter 14.8.

To begin with, suppose the variance of the model error is known up to a multiple, so that

$$\sigma_i^2 = \mathrm{V}[u_i|x_{2i}, ..., x_{ki}] = \delta^2 w_i$$

where $w_i$ is known. For example, $w_i$ may be one of the regressors. Then dividing all terms in the original linear model by $\sqrt{w_i}$ yields

$$\frac{y_i}{\sqrt{w_i}} = \beta_1 \frac{1}{\sqrt{w_i}} + \beta_2 \frac{x_{2i}}{\sqrt{w_i}} + \cdots + \beta_k \frac{x_{ki}}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}}.$$

or

$$y_i^* = \beta_1 x_{1i}^* + \beta_2 x_{2i}^* + \cdots + \beta_k x_{ki}^* + u_i^*.$$

The error $u_i^*$ in the transformed model is now homoskedastic, as $u_i^* = u_i/\sqrt{w_i}$ has variance $\delta^2$ (equals the variance of $u_i$ divided by the square of $\sqrt{w_i}$, or $\delta^2 w_i/(\sqrt{w_i^2})^2 = \delta^2$). Since

Assumptions 1-4 now all hold in the transformed model, OLS estimation in this model is fully efficient.

The **weighted least squares** (WLS) estimator is the OLS estimator in this transformed model. Note that the intercept has been transformed to $x_{1i}^* = 1/\sqrt{w_i}$, which is no longer a constant. So $y^*$ is regressed on the $k$ regressors $x_1^*$, $x_{2i}^*, ..., x_{ki}^*$ in a regression without an intercept.

Now suppose $w_i$ is unknown. A commonly-used model is to suppose that the variance depends on the same variables as appear in the regression model and let

$$\sigma_i^2 = \mathrm{V}[u_i|x_{2i}, ..., x_{ki}] = \exp(\alpha_1 + \alpha_2 x_{2i} + \cdots + \alpha_k x_{ki}),$$

where the exponential is taken to ensure a positive variance. Estimates of $\alpha_1, ..., \alpha_k$ can be obtained by OLS regression of $\ln e_i^2$ on an intercept and $x_{2i}, ..., x_{ki}$, where $e_i$ is the residual from initial OLS estimation of the model for $y_i$. Then $w_i = \widehat{\sigma}_i$ where $\widehat{\sigma}_i^2 = \exp(\widehat{\alpha}_1 + \widehat{\alpha}_2 x_{2i} + \cdots + \widehat{\alpha}_k x_{ki})$.

For statistical inference there are two ways to proceed. First, the default standard errors following OLS regression in the transformed model may be used. These will be correct if the model specified for $\sigma_i^2$ is correct.

Second, heteroskedastic-robust standard errors following OLS regression in the transformed model may be used. These will be correct even if the model specified for $\sigma_i^2$ is incorrect. In this second case, one that is more realistic, there is no guarantee that weighted least squares will be more efficient than the original OLS.

When this method is applied to the earnings example, assuming that the model error variance is a multiple of variable *Education*, the WLS coefficient estimates were generally within 20% of those for OLS, and the heteroskedastic-robust standard errors were sometimes larger and sometimes smaller than those following OLS.

## 19.8 Appendix 19.B: Nonlinear Models

Ordinary least squares regression allows the variables to enter nonlinearly by, for example, taking the natural logarithm or the square of a variable, but restricts the parameters to enter linearly. Thus the conditional mean is specified to be $\beta_1 + \beta_2 x_2^* + \cdots + \beta_k x_k^*$ where $x_2^*, ..., x_k^*$ may be nonlinear transformations of the underlying variables.

**Nonlinear models** have the additional complication that the parameters may enter nonlinearly. For example, the conditional mean specification $\exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$ may be used to ensure that the conditional mean is always positive.

Different types of data lead to use of different nonlinear models. For example, for binary outcome data the logit and probit models are used. This section provides an overview of the most commonly-used nonlinear models. Mechanically implementing these models in a statistical package is generally straightforward. Understanding the appropriate and wise use

of these models is more difficult and requires consulting a reference more advanced than this text.

### 19.8.1 Estimation of Nonlinear Models

The parameter estimates in most nonlinear models are obtained by methods other than least squares estimation. Most often **maximum likelihood estimates** (MLE) are used. ML estimates are the values of the model parameters that make the probability of observing the data in the sample as large as possible. For linear regression with independent homoskedastic errors that are normally distributed ML estimation simplifies to OLS estimation. For nonlinear models, this is no longer the case.

Furthermore, for nonlinear models there is no explicit solution for the parameter estimates. Instead the estimates are defined by a system of equations that are solved by **iterative methods** – initial estimates are used to obtain an improved second-round estimate and so on. This process, automatically done by a statistical package, continues until there is little change in the estimates. For the standard models introduced here such iterative methods work well and usually it takes less than ten iterations to obtain the estimates.

Once estimates are obtained, along with standard errors of the estimates, confidence intervals and hypothesis tests can be conducted in similar manner to the linear regression model. One variation is that some statistical packages continue to use the $T(n-k)$ and $F(q, n-k)$ distributions, while others use the standard normal and $\chi^2(q)$ distributions. The standard normal coincides with $T(\infty)$ and the $\chi^2(q)$ is $q$ times the $F(q, \infty)$ distribution.

The magnitude of the coefficient estimates is generally gauged using marginal effects, as was the case for the logit and probit models.

The MLE requires complete specification of the process generating the data. For some models the MLE remains valid even if some of these assumptions are relaxed. For example the MLE for the linear regression model with independent normal homoskedastic errors makes model assumptions 1-4 plus the assumption of normally distributed errors. But the estimator is actually a consistent estimator under much weaker assumptions, since it equals the OLS estimator and essentially only assumptions 1-2 (that the conditional mean is correctly specified) are needed for consistency.

Some of the nonlinear models surveyed here have similar robustness. In that case one should always base inference on appropriate robust standard errors – heteroskedastic-robust given independence across observations or cluster-robust given clustering – as this enables valid inference under the weaker assumptions.

But many other models, especially those with censoring or truncation, are very fragile. Any departure from the specified assumptions leads to estimator inconsistency. If these models are nonetheless used this fragility needs to be kept in mind.

Table 19.7 provides a summary of different data types commonly encountered in cross-section studies and the simplest models or estimators that are used for such data.

Table 19.7: Nonlinear models: Leading examples for different data types.

| Data Type for Dependent variable y | Model and/or Estimator |
|---|---|
| Continuous in interval $(-\infty, \infty)$ | Ordinary least squares, nonlinear least squares |
| Binary (two categories) | Logit, probit |
| Multinomial (m categories) | Multinomial logit |
| Count taking values 0, 1, 2, .... | Poisson, negative binomial |
| Duration taking values in $(0, \infty)$ | Cox proportional hazards, Weibull |
| Nonnegative continuous in interval $[0, \infty)$ | Nonlinear least squares, Poisson |
| Nonnegative with bunching at zero | Tobit, two-part, selection |
| y is incompletely observed | Tobit, censored, truncation, selection |

### 19.8.2 Multinomial Data

For **multinomial data** the dependent variable $y$ falls into one of $m$ categories. A multinomial model then specifies $m$ separate models for how the probability of falling into each of the $m$ categories varies with the regressors, while these models need to be specified in such a way as to ensure that the probability of being in any one category lies between zero and one, and that the sum of the probabilities across the categories equals one.

In some cases there is no natural ordering of the categories. For example, the outcome may be whether an individual regularly commutes to work by car, by bus or by train. The simplest multinomial model for unordered outcomes is the **multinomial logit model**.

In other cases there is a natural ordering. For example, the outcome may be whether an individual's health status is excellent, good, fair or poor. In that case standard models are the **ordered logit** and the **ordered probit model**.

### 19.8.3 Count Data

For **count data** the dependent variable takes nonnegative integer values 0, 1, 2, ... An example is individual-level data on the number of doctor visits in the past month.

For these data the standard model is the **Poisson regression model**. This model specifies the conditional mean to be of **exponential form**, so

$$\mathrm{E}[y|x_2, ..., x_k] = \exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k).$$

The Poisson model is called the **loglinear model** in the statistics literature as the natural logarithm of the conditional mean is then linear in the regressors. Note that this is quite different from the loglinear model of Chapter 12.4 that specified a model for $\ln y$ as a linear function of regressors.

The Poisson MLE has robustness properties similar to those for OLS in the linear regression model. In particular, the estimator is consistent provided only that the conditional

mean is correctly specified. It is standard to always use heteroskedastic robust standard errors for this model, as the Poisson assumption is usually too strong and it can be shown that default MLE standard errors can be much too small.

Poisson regression can be applied to any data, not just count data, provided the conditional mean is of exponential form. In particular, for right-skewed dependent variable it may be used instead of the log-linear model. The loglinear model directly predicts $\ln y$ rather than $y$, so predicting the level of $y$ introduces of retransformation bias, see Chapter 12.5. The Poisson with exponential conditional mean does not have this problem as it directly models the level of $y$. And the Poisson model can be applied when $y = 0$, while $\ln y$ cannot be calculated when $y = 0$.

### 19.8.4 Duration Data

For **duration data** the dependent variable is the length of time until an event happens. For example, the outcome may be the length of time that an individual participating in a medical study survives.

For such data the outcome is usually incompletely observed. Thus the medical study may last only five years, so not all people in the medical study will have died by the end of the study. This makes modelling the conditional mean difficult as calculation of a mean requires complete data, unless strong parametric assumptions are made such as those for the **Weibull model** of durations.

Fewer assumptions are needed to model the hazard rate, where the **hazard rate** at time $t$ is the probability of death at time $t$ conditional on having survived to time $t$. The standard model for duration data is the **Cox proportional hazards model** as it relies less on strong parametric assumptions.

### 19.8.5 Nonlinear Least Squares

Nonlinear least squares can be used whenever a particular functional form for the conditional mean is specified. An example is the Cobb-Douglas production function for output $y$ and inputs $x_2$ and $x_3$, with conditional mean function $\mathrm{E}[y|x_2, x_3] = \beta_1 x_2^{\beta_2} x_3^{\beta_3}$. Then **nonlinear least squares** minimizes the sum of squared residuals $\sum_{i=1}^{n} e_i^2$, where $e_i = y_i - \beta_1 x_{2i}^{\beta_2} x_{3i}^{\beta_3}$ in the Cobb-Douglas example. The nonlinear least squares estimator is consistent provided the model for the conditional mean is correctly specified.

Note that for specific types of data other estimation methods are preferred to nonlinear least squares. For example, the logit model can be estimated by nonlinear least squares, but ML estimation leads to more precise estimates. Similarly for count data with exponential conditional mean, Poisson regression is better than nonlinear least squares.

### 19.8.6 Bunching at Zero

Individual cross-section data can be disproportionately bunched at zero. This is particularly the case for data on expenditure by individuals on a particular type of good or service, such as cars or hospitalization.

Several quite different models for such data have been proposed. It is often not clear which model is the best model, and results can be very dependent on very strong distributional assumptions.

First, it may be assumed that the zeroes arise because negative values are not possible. Then it is common to use the **Tobit model**. This model assumes that there is an underlying observed variable $y^*$ for, say, desired expenditures that comes from a linear regression model with independent homoskedastic normally-distributed errors. If $y^*$ is positive then we observe $y = y^*$ and if $y^*$ is negative then we observe $y = 0$. This model is exceptionally fragile. For instance if either the normality assumption or homoskedasticity assumption is wrong then the Tobit MLE is inconsistent.

Second, it may be assumed that the zeroes occur naturally. Then an exponential conditional mean may be assumed with estimation by nonlinear least squares or by Poisson regression.

Third, a **two-part model** specifies that two processes are at work. For example, annual hospital expenditures are zero for all but roughly ten percent of the population. A logit model may be used to model whether or not someone is hospitalized during the year. And a separate log-linear OLS regression may be estimated for hospital expenditures (which are very right-skewed) using data on just those individuals with positive hospital expenditures.

Fourth, a **selection model on unobservables model** generalizes the two-part model to allow the error terms in the two parts to be correlated. This richer model is needed in some applications, but then results rely heavily on the assumed model for the error correlation in the two parts.

### 19.8.7 Truncated and Censored Data

Data are **truncated** if neither $y$ nor the regressors are observed for some values of $y$. For example, if a survey is conducted of car purchasers, then no data at all are available for non-purchasers.

Data are **censored** if $y$ is not completely observed, but regressors are always observed. For example, if a survey is conducted of the population then regressors are observed for all individuals, but the amount spent on car purchase is only observed for those who purchased a car. The Tobit model is an example.

The models developed for such data are generally very fragile as consistency requires very strong distributional assumptions. A notable exception for censored data is the already-mentioned Cox proportional hazards model for duration data.

## 19.9 Appendix 19.C: Categorical Data

Suppose we want to do bivariate analysis where both $y$ and $x$ are categorical. For example, an opinion survey wants to see whether intention to vote for a particular political party (Democrat, Republican or other) is associated with race (White, Hispanic, ...).

Then OLS regression and correlation methods are not appropriate, especially if there is no natural ordering of the categories. Instead, whether or not there is relationship between the two variables is determined by comparing the observed frequency of each combination of the two variables (a cell) and comparing this to the frequency expected if the two variables were independent. An example of computation of expected frequencies is given in Chapter 8.2.

Suppose the two variables have, respectively, $m$ categories and $n$ categories, so there are $m \times n$ cells, and let subscript $j$ denote the $j^{th}$ cell, $j = 1, ..., m \times n$. Then **Pearson's chi-squared goodness-of-fit statistic** for a two-way tabulation is

$$\text{CGF} = \sum_{j=1}^{m\times n} \frac{Observed_j - Expected_j}{Expected_j}.$$

Under the null hypothesis that the two categorical variables are independent the statistic CGF is chi-squared with $(m-1)(n-1)$ degrees of freedom; see Chapter 14.5 for the chi-squared distribution. If the two categories are indeed independent then CGF is small, since then observed values are close to expected values. So the null hypothesis is rejected if CGF is large. The $p$-value for the test is $p = \Pr[\chi^2_{(m-1)(n-1)} > \text{CGF}]$, and we reject at level $\alpha$ if $p < \alpha$.

As an example, consider the two-way cross tabulation example of Chapter 8.2, with house price in two categories (low and high) and house size in three categories (small, medium and large). From Table 8.4 there were six cells, with observed cell frequencies of respectively, 11, 6, 0, 2, 7 and 3, and expected frequencies if the two variables were statistically independent of, respectively, 7.62, 7.62, 1.76, 5.38, 5.38 and 1.24. Then

$$\text{CGF} = \frac{(11-7.62)^2}{7.62} + \frac{(6-7.62)^2}{7.62} + \frac{(0-1.76)^2}{1.76} + \frac{(2-5.38)^2}{5.38} + \frac{(7-5.38)^2}{5.38} + \frac{(3-1.24)^2}{1.24} = 8.70.$$

Here $(m-1)(n-1) = (2-1)\times(3-1) = 2$ and $p = \Pr[\chi^2_2 > 8.70] = 0.011$. Since $p < .05$ we reject the null hypothesis of statistical independence. There is a statistically significant relationship at level .05 between house price and size.

Pearson's statistic is widely used in many areas of applied statistics that often analyze data with both variables categorical, but is used infrequently in empirical economics. Economics data are most often ordered numerical data. Then interest lies in determining the magnitude of any relationship, not just whether a relationship exists, and one uses OLS regression methods, the focus of this book. And if just the dependent variable is categorical then logit, probit or multinomial regression models can be used.

# Chapter 20

# © A. Colin Cameron: Time Series Data

Superficially, regression analysis with time series data seems straightforward. Simply perform standard OLS regression, with the modification that inference is based on heteroskedastic- and autocorrelation-consistent (HAC) standard errors, introduced in Chapter 10.7. HAC standard errors are used as model errors are likely to be correlated over time, so population assumption 4 of independent errors no longer holds.

Unfortunately, this approach is fraught with pitfalls.

First, if regressors include lags of the dependent variable, then OLS is inconsistent if the model errors are serially correlated. A simple example is regressing this period's value on last period's value.

Second, using standard regression methods it is quite easy to falsely discover relationships between time series that are in fact unrelated. This possibility can be reduced by including appropriate time trends in the regression model. But if variables are highly correlated with their values in previous periods, including time trends is not enough. This situation arises, for example, with key macroeconomic variables such as GDP, prices and interest rates.

For these reasons it is very important to evaluate the persistence of individual time series, and persistence of regression model errors. And this chapter deliberately uses data that are highly persistent.

The first half of the chapter considers time series models for a single series and emphasizes evaluating the extent of such persistence, and the consequences of high persistence. The remainder of the chapter considers regression models for time series data. More advanced methods are relegated to Appendices 20.B and 20.C.

Table 20.1: Ten-year and one-year U.S.Treasury Notes from January 1982 to January 2015: Variable definitions and summary statistics

| Variable | Definition | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| *GS10* | 10-year Treasury Note Constant Maturity Rate | 6.186 | 2.878 | 1.53 | 14.59 |
| *GS1* | 1-year Treasury Note Constant Maturity Rate | 4.691 | 3.283 | 0.10 | 14.73 |
| $\Delta$*GS10* | Monthly change in 10-year Treasury Note Rate | -0.032 | 0.274 | -1.43 | 0.78 |
| $\Delta$*GS1* | Monthly change in 1-year Treasury Note Rate | -0.036 | 0.288 | -1.81 | 0.76 |

## 20.1 Example: Ten-year and One-year Interest Rates

Throughout this chapter we analyze monthly data in dataset TIMESERIES on ten-year U.S. Treasury notes (*GS10*) and one-year U.S. Treasury note (*GS1*) from January 1982 to January 2015. The rates given are annualized percentage nominal interest rates, and are yields at constant maturity rate.

Table 20.1 summarizes the data. On average the 10-year rate exceeds the one-year rate by 1.5 percentage points, reflecting a reward for greater exposure to risk in holding longer term bonds. The one-year rate is a little more volatile over this period, as *GS10* and *GS1* have standard deviations of, respectively, 2.88 and 3.28. Over this period there was considerable variation in interest rates, from a high of close to 15% per annum to as low as 0.10% for the one-year rate.

Table 20.1 additionally summarizes data on the monthly changes in the rates. This indicates that rates were on average declining over this period. For example, the ten-year rate was declining by around 0.4 percentage points per year ($\simeq 12 \times (-0.032)$).

The first panel of Figure 20.1 presents a time series plot of the data. There is strong persistence for both series, as the current period rates are closely related to those in the immediate preceding periods. Both rates show a pronounced downward trend, due to the decline in inflation, and even more importantly the decline in inflationary expectations, from highs in the 1970's and early 1980's to moderate levels in the 1990's. The most recent continued decline is a consequence of the 2007 global financial crisis. Note that the one-year interest rate at the end of the sample period is very close to the zero interest rate lower bound – it is difficult, though not impossible, to have negative nominal interst rates. The two interest rates tend to move together, with the one-year rate fluctuating more.

The second panel of Figure 20.1 presents a scatterplot of the ten-year rate against the one-year rate. The relationship between the two rates appears to be linear with a slope of roughly $(14-3)/(14-0) \simeq 0.8$.

Modelling of economic time series data should be guided by economic theory, in additional to the statistical properties of the data. Economic theory applies best to interest rates after correction for expected price inflation, expected changes in exchange rates and expectations

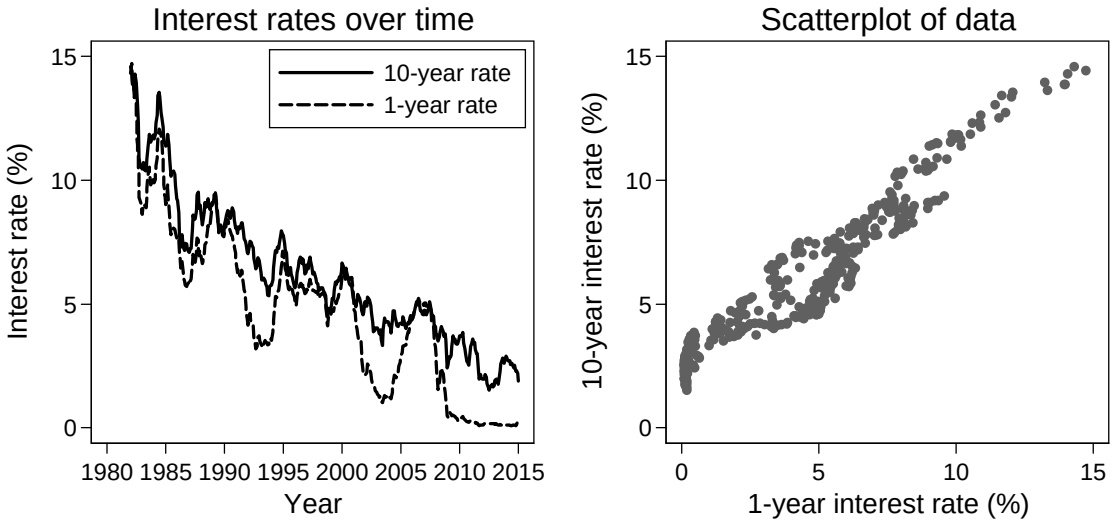


Figure 20.1: Annualized rate of return at constant maturity on 1-year and 10-year U.S. Treasury bills from January 1982 to January 2015.

of default. For example, we expect that the expected return to a U.S. investor of a U.S. bond should equal the expected return in a U.S. dollars on a Japanese bond of similar duration and smilar risk after accounting for expected price inflation in each country and expected changes in the U.S.-Japan exchange rate.

For U.S. Treasuries the most important consideration is expected price inflation, which is unknown. We could instead model real interest rates, the nominal interest rate less the observed price inflation rate, rather than nominal rates. This does not make too much difference for these data. Price inflation did not vary enormously during this period (the inter-quartile range for the CPI was 2.3% to 3.9%). And for bonds of different maturities the difference in real rates equals the difference in nominal rates, so that price inflation mostly nets out when modelling the relationship between bonds of different maturities. The various theories of the interrelationship between bonds of different maturity is not presented here.

## 20.2 Single Time Series: Persistence

Statistical inference links sample estimates to population parameters. To date the sample has been viewed as being a subsample of a finite population or as outcomes from an experiment. The dataset on U.S. Treasury interest rates fall into neither category. Instead the data are viewed as generated by a process that evolves over time, and the sample is used to estimate

the parameters of this process.

The methods employed to model and perform inference on individual series, and on regression models, depend greatly on how this time series process is modelled. This in turn depends on the extent and nature of the persistence in the time series data. This lengthy section analyzes persistence.

The following general terminology and notation is used. A **time series** is a collection of observations taken at different points in time that are assumed here to be equidistant in time. The observations are ordered by time, with $y_t$ denoting the time series variable of interest. Then $y_{t-1}$ denotes $y$ in the preceding period and $y_{t-j}$ denotes $y$ **lagged** $j$ periods. For example, if $y_t$ is the rate in March 2002 then $y_{t-2}$ is the rate in January 2002. The first observed value of $y$ is $y_1$ and the sample of size $T$ is $y_1, y_2, ..., y_T$. At times **first differenced** data $\Delta y_t$, where $\Delta y_t = y_t - y_{t-1}$, are used. For example, $\Delta GS10_t$ denotes the monthly change in the ten-year rate. First differencing leads to loss of the first observation and the sample is $\Delta y_2, \Delta y_3, ..., \Delta y_T$.

### 20.2.1 Stationary Processes

The building block for analysis of a single time series is a **covariance stationary process**. This greatly simplifies analysis by restricting the mean and (finite) variance of $y_t$ to be the same regardless of the point in time $t$ considered, and restricts the covariance between $y_t$ and $y_s$ to depend only on the time separation $|t - s|$.

**Remark 168** *For a **covariance stationary process** (1) $E[y_t]$ is the same for all $t$; (2) $Var[y_t]$ is finite and is the same for all $t$; and (3) $Cov[y_t, y_s]$ depends only on the time separation $|t - s|$ and not on the time period $t$.*

The simplest example of a stationary process, and the building block for many time series models for a stationary process is a white noise process.

**Remark 169** *A **white noise process**, often denoted $\varepsilon_t$, has mean zero, constant variance $\sigma^2_\varepsilon$ and is uncorrelated across time, so $Cov[\varepsilon_t, \varepsilon_s] = 0$ for $t \neq s$. In that case the population autocorrelations are zero.*

### 20.2.2 Data Transformation

Time series data are often transformed to a stationary process before any analysis. In particular, many macroeconomic variables such as real GDP are trending upward over time. The series are nonstationary as the mean and variance are increasing over time.

If such growth is exponential then the first step is to take the **natural logarithm** as then the growth is linear in natural logarithms; see Chapter 4.1. For example, macroeconomic data such as real GDP are typically analyzed in natural logarithm.

After such transformation the data often have a linear trend. Consider the **linear time trend model**, with

$$y_t = \alpha + \delta t + u_t,$$

where $\alpha$ is "alpha", the Greek letter $a$, and $\delta$ is "delta", the Greek letter $d$. The error $u_t$ is assumed to be stationary. Then on average $y_t = \alpha + \delta t$ if $\mathrm{E}[u_t] = 0$. The linear time trend model specifies that on average $y$ increases by $\delta$ each period, so the data are nonstationary.

There are several ways to handle this nonstationarity. Analysis may be based on **detrended data**. Thus regress $y$ on an intercept and a time trend and analyse the residual $y_t - \widehat{\alpha} - \widehat{\delta} t$. Analysis may be based on **first-differenced data** $\Delta y$, as the time trend model implies $\Delta y_t = \delta + \Delta u_t$ which is stationary given $u_t$ is stationary. Or **regression analysis** may be based on $y_t$, provided a linear trend is included in the model.

### 20.2.3 Sample Autocorrelations

Persistence in a time series is gauged using sample autocorrelations, introduced in Chapter 8.4, that measure the correlation between the current value and lagged values.

The **sample autocorrelation at lag** $j$ an estimate of the correlation coefficient between $y_t$ and $y_{t-j}$. This can be computed in several ways, including

$$\widehat{\rho}_j \simeq \frac{\frac{1}{T-j}\sum_{t=j+1}^{T}(y_t - \overline{y})(y_{t-j} - \overline{y})}{\frac{1}{T}\sum_{t=j+1}^{T}(y_t - \overline{y})^2}, \quad -1 \leq \widehat{\rho}_j \leq 1.$$

where $\overline{y} = \frac{1}{T-j}\sum_{t=j+1}^{T} y_t$. The sum in the numerator is only over $(T-j)$ terms as $y_{t-j}$ can only be computed from time $t = j+1$ on (since $y_{t-j} = y_1$ when $t = j+1$). Some computer packages use a refinement of this formula that is given in Chapter Appendix 20.A.

For a white noise process, it can be shown that the autocorrelation estimate $\widehat{\rho}_j$ has standard error equal to $1/\sqrt{T}$, while for autocorrelated process the standard error is larger. So a 95% confidence interval for the population autocorrelation coefficient $\rho_j$ is at least $(\widehat{\rho}_j - 2\sqrt{T}, \widehat{\rho}_j + 2\sqrt{T})$.

Specialized software such as Stata computes sample autocorrelations automatically, along with confidence intervals, for the first $m$, say, autocorrelations. Excel does not provide a function or tool to compute autocorrelations, but they are easy to compute using one of the methods given in Appendix 20.A.

**Remark 170** *The sample autocorrelation at lag $j$, denoted $\widehat{\rho}_j$, measures the correlation between $y_t$ and $y_{t-j}$. Persistence is higher the closer $\widehat{\rho}_j$ is close to* 1.

### 20.2.4 Ljung-Box Test for No Autocorrelation

It can be useful to test whether or not a time series is serially correlated. Here the time series may be the original time series, a transformed series, or the residual from an OLS regression provided the regression does not include lagged dependent variables.

Reasons for wanting to test for serial correlation include the following. A method may be valid only if errors are serially correlated. If the error is serially correlated then it may be possible to develop a better model for the data. And economic theory such as a no-arbitrage conditoin under the assumption of efficient markets may imply that the series is not serially correlated.

**Ljung-Box portmanteau test statistic** for test of whether the first $m$ autocorrelations are all zero is

$$\begin{aligned} Q &= T(T+2)\textstyle\sum_{j=1}^{m} \dfrac{1}{T-j}\widehat{\rho}_j^2 \\ &\simeq T \times (\widehat{\rho}_1^2 + \widehat{\rho}_2^2 + \cdots + \widehat{\rho}_m^2), \end{aligned}$$

where the user specifies $m$. The test statistic is $\chi^2(m)$ distributed under $H_0 : \rho_1 = 0, ...\rho_m = 0$, and $H_0$ is rejected at level $\alpha$ if $Q > \chi^2_\alpha(m)$.

An older test statistic for autocorrelation is the **Durbin-Watson (DW) test statistic** for first-order autocorrelation that approximately equals $2 \times (1 - \widehat{\rho}_1)$. This test has been superseded by the Ljung-Box test.

The Ljung-Box test can be applied to the residuals from a regression model, provided the model does not include lagged dependent variables. If instead lagged dependent variables are included then the Breusch-Godfrey test given later in this section should be used.

### 20.2.5 Example: Interest Rates

Table 20.2 presents autocorrelations for the two interest rate series, changes in these series, and the residuals from regression models.

From the first two rows, there is extraordinarily high persistence for both the ten-year and one-year rates. Even 36 months or three years out the autocorrelation is 0.56 for *GS10* and 0.42 for *GS1*. The Ljung-Box test statistics at 12 lags are $Q = 3797.6$ for *GS10* and $Q = 3693.0$ for *GS1*. In both cases $p = 0.000$, so the null hypothesis of no autocorrelation is strongly rejected for both series.

For the ten-year rate the fitted linear time trend model is

$$\widehat{GS10}_t = 10.858 - 0.023 \times t, \; s_e = 1.015.$$

There is an appreciable time trend, with the ten-year rate dropping by 0.023 percentage points per month, or 0.28 percentage points per year, over this time period.

Table 20.2: Interest rates: Autocorrelations and residual autocorrelations.

| **Variable** | **Autocorrelation at lag j** ($\widehat{\rho}_j$) | | | | | | | Ljung-Box | |
|---|---|---|---|---|---|---|---|---|---|
| | j=1 | j=2 | j=3 | j=4 | j=5 | j=18 | j=36 | Q (m=12) | p-value |
| *GS10* | 0.98 | 0.96 | 0.94 | 0.92 | 0.90 | 0.75 | 0.56 | 3797.6 | 0.000 |
| *GS1* | 0.98 | 0.96 | 0.94 | 0.92 | 0.90 | 0.69 | 0.42 | 3693.0 | 0.000 |
| $\Delta$*GS10* | 0.33 | 0.00 | 0.02 | -0.01 | -0.04 | -0.07 | 0.01 | 51.6 | 0.000 |
| $\Delta$*GS1* | 0.42 | 0.18 | 0.12 | 0.05 | 0.09 | -0.02 | -0.08 | 97.1 | 0.000 |
| Residual: GS10 time trend model | 0.95 | 0.87 | 0.80 | 0.72 | 0.65 | 0.10 | 0.03 | 1781.0 | 0.000 |
| Residual: GS1 time trend model | 0.97 | 0.92 | 0.87 | 0.82 | 0.76 | 0.18 | -0.40 | 2432.9 | 0.000 |
| Residual: GS10 AR(1) model | 0.33 | 0.00 | 0.01 | -0.01 | -0.05 | -0.07 | 0.01 | - | - |
| Residual: GS1 AR(1) model | 0.41 | 0.18 | 0.12 | 0.05 | 0.08 | -0.02 | -0.08 | - | - |

The fifth row of Table 20.2 gives the autocorrelations for the residual $e_t = GS10_t - 10.858 + 0.023 \times t$ from the time trend regression. The persistence in the residual is less than that in the original series, but there is still considerable persistence that takes 18 lags or so to disappear. The sixth row gives the autocorrelations for the residual from OLS regression of *GS1* on the time trend and even after 18 months the autocorrelation is still 0.18.

For these data a time trend reduces the persistence, but by nowhere near as much as first differencing.

### 20.2.6 Autoregressive Model

Another obvious model for persistence is to specify this period's value to be a multiple of last year's value plus random noise.

An **autoregressive model of order 1 (AR(1)) model** specifies

$$y_t = \delta + \phi y_{t-1} + u_t, \ |\phi| < 1.$$

where $\phi$ is "phi" the Greek letter $f$. The term autoregressive is used as $y$ is regressed on a past value of itself. The model can be extended to an **AR(p) model** that includes the first p lags of $y$ as regressors.

If the error $u_t$ in the AR(1) model is a white noise error then it can be shown that

$$\rho_j = \text{Cor}[y_t, y_{t-j}] = \phi^j.$$

So the series $y_t$ is very persistent if $\phi$ is close to 1.

The sample autocorrelations for *GS10* given in the first row of Table 20.2 suggest that a good model for *GS10* is an AR(1) model with $\phi \simeq 0.98$. This model predicts the autocorrelation at lag $j$ to be $0.98^j$ which yields first five autocorrelations of approximately 0.98,

0.96, 0.94, 0.92 and 0.90, $\rho_{18} = 0.98^{18} = 0.70$ and $\rho_{36} = 0.98^{36} = 0.54$. (In fact $\phi \simeq 0.984$ does better as, for example, $0.984^{36} = 0.75$).

OLS regression yields estimated AR(1) model for *GS10* of

$$\widehat{GS10}_t = 0.046 + 0.987 \times GS10_{t-1},\ s_e = 0.285.$$

This model fits these data much better than the linear time trend model, as $s_e$ is much lower.

The seventh row of Table 20.2 gives the autocorrelations of the residuals $e_t = GS10_t - 0.046 - 0.987 \times GS10_{t-1}$ from this AR(1) regression. The residuals are essentially uncorrelated after the first lag. And from the eighth row of Table 20.2 the residuals from the AR(1) model for *GS1* are correlated for only the first two lags. The Ljung-Box statistic is not reported in Table 20.2 since the residuals are from a regression with a lagged dependent variable as a regressor.

The autocorrelations for the AR(1) model residuals are very similar to those for $\Delta GS10$ and $\Delta GS1$ since an AR(1) model with $\phi = 1$ yields a first difference, and in this case the estimated AR(1) models had $\widehat{\phi}$ close to one.

**Remark 171** *Two common models for a persistent time series are the linear time trend model, $y_t = \alpha + \delta t + u_t$, and the autoregressive moving average of order one (AR(1)) model, $y_t = \delta + \phi y_{t-1} + u_t$.*

## 20.3 Nonstationary Time Series

Great complications arise when data are too highly persistent, as then the standard statistical properties of estimators no longer hold.

As a result the methods of the first nineteen chapters may be the wrong methods to use!

### 20.3.1 Stationary Processes

Loosely speaking, a stationary process is one with a finite variance and a nonstationary process is one with possibly infinite variance.

A specialization is a **covariance stationary process**. Then the mean and (finite) variance of $y_t$ is the same regardless of the point in time $t$ considered, and the covariance between $y_t$ and $y_s$ depends only on the time separation $|t - s|$.

**Remark 172** *For a **covariance stationary process** (1) $E[y_t]$ is the same for all $t$; (2) $Var[y_t]$ is finite and is the same for all $t$; and (3) $Cov[y_t, y_s]$ depends only on the time separation $|t - s|$ and not on the time period $t$.*

A simple example of a stationary process is a white noise process. Another example is an AR(1) process with white noise error, so $y_t = \delta + \phi y_{t-1} + \varepsilon_t$, where $\varepsilon_t$ is white noise and $|\phi| < 1$.

### 20.3.2 Random Walk Processes

Random walk processes are the leading examples of a nonstationary process.

A **random walk with drift** is the process

$$y_t = \delta + y_{t-1} + \varepsilon_t,$$

where $\varepsilon_t$ is a white noise process with zero mean and variance $\sigma_\varepsilon^2$. A **random walk without drift** is the special case $\delta = 0$, so then $y_t = y_{t-1} + \varepsilon_t$.

A random walk without drift is an obvious model for highly persistent data – this period's observation is simply last period's observation plus some purely random noise. A random walk with drift additionally allows an increment of $\delta$ each period.

A random walk (with or without drift) has the serious complication that it is a nonstationary process, with variability that increases over time.

**Remark 173** *A random walk with drift process is the model $y_t = \delta + y_{t-1} + \varepsilon_t$ where $\varepsilon_t$ is white noise. Then $E[y_t] = y_0 + \delta t$, a time trend, and $Var[y_t] = t\sigma_\varepsilon^2$ so that the variance is explosive. A random walk without drift sets $\delta = 0$.*

This result is obtained as follows. By recursive substitution

$$\begin{aligned}
y_t &= y_{t-1} + \delta + \varepsilon_t \\
&= [y_{t-2} + \delta + \varepsilon_{t-1}] + \delta + \varepsilon_t \\
&= [(y_{t-3} + \delta + \varepsilon_{t-2}) + \delta + \varepsilon_{t-1}] + \delta + \varepsilon_t \\
&= y_{t-3} + 3\delta + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\
&= \quad \vdots \qquad \vdots \\
&= y_0 + \delta t + \textstyle\sum_{j=0}^{t-1} \varepsilon_{t-j},
\end{aligned}$$

where the second equality lags the model once so $y_{t-1} = y_{t-2} + \delta + \varepsilon_{t-1}$, the third equality uses $y_{t-2} = y_{t-3} + \delta + \varepsilon_{t-2}$, and so on. So $y_t$ has mean $y_0 + \delta t$, since the third term $\sum_{j=0}^{t-1} \varepsilon_{t-j}$ has mean zero if $\varepsilon_t$ is white noise. The nonstationarity arises because $\mathrm{Var}[y_t] = \mathrm{Var}\left[\sum_{j=0}^{t-1} \varepsilon_{t-j}\right] = \sum_{j=0}^{t-1} \mathrm{Var}[\varepsilon_{t-j}] = t\sigma_\varepsilon^2$ which increases with $t$.

### 20.3.3 Unit Root or I(1) Processes

A random walk model, with or without drift, is a special case of a **unit root process**. The term unit root is used as the random walk is the special case of the AR(1) model with $\phi = 1$ (**unity**).

A unit root process is also called a process that is **integrated of order one**, denoted I(1). A process $y_t$ is integrated of order one if $y_t$ is nonstationary but $\Delta y_t$ is covariance

stationary. The term integrated is used as $y_t$ can be viewed as the integral of the difference $\Delta y_t = y_t - y_{t-1}$, just as in calculus $y(t)$ is the integral of $dy/dt$. A covariance stationary process $y_t$ is **integrated of order zero**, denoted I(0).

A random walk with drift is often called a **stochastic trend model**, since it has mean $y_0 + \delta t$. The linear time trend model of Chapter 20.2 that has a stationary error, by contrast, is called a **deterministic trend model**.

**Remark 174** *A process $y_t$ is a unit root process, also called integrated of order one, denoted I(1), if $y_t$ is nonstationary but $\Delta y_t$ is covariance stationary. A random walk is a leading example of a nonstationary process.*

### 20.3.4 Why does it Matter if there is a Unit Root?

There are several important distinctions between unit root and non-unit root processes.

The implications for policy analysis are quite different, as the explosiveness of a random walk model means that a temporary shock can have a permanent effect. For example, if GDP follows a random walk (with or without drift), then a temporary one-period negative shock to GDP will permanently lower GDP, not just lower it in the next few periods.

Forecasts from a unit root process will be much noisier as they will have variance that increases with the time horizon.

And the statistical properties of estimators presented in the first nineteen chapters of this book no longer hold if a unit root is present. Thus if conventional statistical methods are used it is easy, for example, to falsely discover a relationship between two series when there is in fact no relationship. This is the subject of Chapter 20.4.

If unit roots are present the simplest and safest thing to do is to transform the data and do analysis in differences, assuming that the differenced data are stationary. This is the approach emphasized in this chapter. If instead regression is in levels then more advanced methods need to be used; see Chapter 20.7.

### 20.3.5 Unit Root Tests

If data are not highly persistent, then the possibility of a unit root can be ignored. One rule of thumb is that a unit root is very unlikely if the sample autocorrelation at lag one is less than 0.8 and autocorrelations after two or three lags are below 0.5.

If instead a series has high persistence then it is important to test whether the persistence is so high that a unit root is present. For example, a unit root is likely to be present if low-order autocorrelations exceed 0.9 and are slow to die out.

The autocorrelation function provides only a guide, albeit a valuable one. A formal test for unit roots, the augmented Dickey-Fuller test, is presented in chapter Appendix 20.B.

### 20.3.6 Example: Interest Rates

From Table 20.2 the autocorrelations for both *GS10* and *GS1* indicate very high persistence. For example, the autocorrelations at lags 1-3 are, respectively, 0.98, 0.96 and 0.94. The data are either unit root processes or very close to unit root processes. The formal tests in Appendix 20.B give somewhat mixed results, but lean towards rejecting the null hypothesis of a unit root.

The empirical macroeconomic literature finds very high persistence in both nominal and real interest rates. Formal statistical tests provide mixed results on whether or not this persistence is high enough to be a unit root.

At the same time economic theory does not support unit roots in interest rates. And if interest rates really were unit root processes then we would expect to find many more periods of very high interest rates than observed historically.

The safest approach, however, is to assume a unit root and perform regression analysis in first differences. This is the approach emphasized in this chapter. The reason for doing so is to avoid running a spurious regression.

## 20.4 Spurious Regression

A **spurious regression** is one that detects a relationship between series when in fact there is no relationship. This can arise with persistent data in a couple of ways.

First, if time trends or other deterministic trends that should be included are omitted, then omitted variables bias may lead to erroneously finding a relationship.

Second, and more subtly, if a unit root series is regressed on a unit root series then standard inference methods following OLS regression of $y_t$ on $x_t$ are invalid and are greatly biased towards finding a relationship even when does not exist.

A commonly-used example of a spurious regression is a time series regression of GDP of a country such as the U.S. on the cumulative rainfall in the country's capital. Rainfall is not driving GDP, but a strong statistically significant relationship will be found as both series are trending upwards over time.

### 20.4.1 Time Trends (Deterministic)

A simple example of spurious regression is regression of $y_t$ on $x_t$ without a time trend when $y_t$ and $x_t$ are generated by unrelated linear time trends.

Specifically, suppose

$$\begin{aligned} y_t &= \alpha + \beta t + u_t \\ x_t &= \gamma + \delta t + v_t, \end{aligned}$$

where $u_t$ and $v_t$ are errors that are independent of each other. Then there is no relationship between $y_t$ and $x_t$.

Nonetheless OLS regression of $y_t$ on $x_t$ will suggest a relationship, since solving the two equations (by eliminating $t$) yields

$$y_t = (\alpha - \beta\gamma/\delta) + (\beta/\delta)x_t + (u_t - (\beta/\delta)v_t).$$

So $y_t$ is linearly related with $x_t$ with slope coefficient equal to $\beta/\delta$.

In this particular case inclusion of a time trend as a regressor will avoid finding a spurious regression, since the original equation for $y_t$ implies that

$$y_t = \alpha + \beta t + 0 \times x_t + u_t,$$

so the OLS coefficient of $x_t$ will be close to zero once a time trend is included.

Alternatively, running the regression in changes rather than levels will also reveal no relationship, since first differencing yields $\Delta y_t = \beta + \Delta u_t$ so

$$y_t = \beta + 0 \times x_t + \Delta u_t.$$

### 20.4.2 Unit Root Trends (Stochastic)

A more subtle example of spurious regression arises when $y_t$ and $x_t$ are generated by unrelated unit root processes.

Specifically, suppose $y_t$ and $x_t$ are generated by random walks without drift

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t \\ x_t &= x_{t-1} + \eta_t, \end{aligned}$$

where $\varepsilon_t$ and $\eta_t$ are errors that are independent of each other, and $\eta$ or "eta" is the Greek letter $h$.. Again there is no relationship between $y_t$ and $x_t$.

Now consider OLS regression of the model $y_t = \beta_1 + \beta_2 x_t + u_t$, where $\beta_2 = 0$ in this example. The OLS slope estimate converges to a random variable, rather than 0, as the sample size gets large. This and other unusual properties arise because the error term $u_t$ in this model has variance that is increasing over time. Specifically, a random walk can be re-expressed as $y_t = y_0 + \sum_{j=0}^{t-1} \varepsilon_{t-j}$, see Chapter 20.3. This is the model $y_t = \beta_1 + \beta_2 x_t + u_t$ with $\beta_1 = y_0$, $\beta_2 = 0$ and $u_t = \sum_{j=0}^{t-1} \varepsilon_{t-j}$, so $\text{Var}[u_t] = t\sigma_\varepsilon^2 \to \infty$ as $t \to \infty$.

Furthermore, these unusual properties result in the standard $t$-statistic with normal or $t$ critical values frequently finding a relationship between the two variables, even though by construction there is no such relationship. Formal proof requires very advanced statistical methods. Similar problems arise for random walks with drift.

**Remark 175** *Spurious regression can arise because of omitted time trends, or because the data are unrelated unit root processes.*

Before presenting ways to rule out spurious regressions, a simulation is presented to demonstrate the high probability of finding spurious relationships when data are unrelated random walks.

### 20.4.3 Simulation

Ten thousand samples are generated from the following data generating process

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0,1), \quad t = 1, ..., 50, \\ x_t &= x_{t-1} + \eta_t, \quad \eta_t \sim N(0,1), \quad t = 1, ..., 50. \end{aligned}$$

In each sample $y_t$ is regressed on $x_t$ and the usual $t$-statistic, the slope coefficient divided by the default standard error, is computed. By definition, a test at significance level 5% should erroneously find a relationship in 5% of the samples (i.e. in 500 of the 10,000 samples).

In the current setting with unit root data, the $t$-statistic does not have the usual $t$ distribution. Instead of being centered on zero with a standard deviation of approximately one, the distribution is centered on zero with a standard deviation of approximately five, and the critical values for a 5% test are approximately $\pm 10$. If the usual $t$ critical values of $\pm t_{.025,48} = \pm 2.01$ are nonetheless used, then in this simulation a relationship is (erroneously) found 67% of the time rather than 5% of the time.

This problem of over-rejection still exists if HAC standard errors (with $m = 12$) are used rather than default standard errors. Repeating the simulation but using HAC standard errors, the $t$-statistic has standard deviation of approximately four and critical values for a 5% test of approximately $\pm 8$. The usual $t$ critical values of $\pm t_{.025,48} = \pm 2.01$ lead to erroneously finding a relationship 50% of the time.

Similar over-rejection occurs even if a time trend is included in the regression.

### 20.4.4 Tests for Spurious Regression

Running the regression in changes rather than levels will detect whether or not there is a relationship. If the two series are unrelated random walks then $\Delta y_t = a + \varepsilon_t$, so $\Delta y_t = a + 0 \times \Delta x_t + \varepsilon_t$. So estimate by OLS the model

$$\Delta y_t = \beta_1 + \beta_2 \Delta x_t + u_t,$$

and test $H_0 : \beta_2 = 0$ using the usual $t$-test, where the HAC robust standard error of $b_2$ is used.

A second method is to add first order lags of $y_t$ and $x_t$ as regressors, and estimate by OLS the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \beta_4 x_{t-1} + u_t.$$

Then test $H_0 : \beta_2 = 0$ using a $t$-statistic based on HAC standard errors. Note that if the error $u_t$ is serially correlated, often the case, then this $t$-statistic is only a rough guide as OLS is inconsistent in a model with both a lagged dependent variable and a serially correlated error.

### 20.4.5 Cointegration

If $y_t$ and $x_t$ are unit root processes and a relationship exists between $y_t$ and $x_t$, then $y_t$ and $x_t$ are said to be **cointegrated**. A formal definition is the following.

**Remark 176** *Two series that are integrated of order one (I(1)) are said to be* ***cointegrated*** *if there is a linear combination of the series that is integrated of order zero I(0). Thus $y_t$ and $x_t$ are cointegrated if $y_t - \beta_1 - \beta_2 x_t$ is I(0).*

Formal tests for spurious regression are called **cointegration tests**. One such test is detailed in Appendix 20.B.

If two series are cointegrated, then OLS regression of $y_t$ on $x_t$ (in levels) leads to consistent estimation of the slope parameter. However, because $y_t$ and $x_t$ are I(1) processes, the distribution of the OLS estimates is nonstandard, and HAC standard errors are not the correct standard errors. Instead HAC standard errors can only be used as a guide. Estimation methods for cointegrated series are briefly discussed in Chapter 20.7.

### 20.4.6 Example: Interest Rates

We begin by running the regression in levels, with a time trend included. Then

$$\widehat{GS10}_t = \underset{(8.88)}{5.910} - \underset{(-5.79)}{0.011} \times t + \underset{(6.28)}{0.508} \times GS1_t,\ s_e = 0.665,\ R^2 = 0.947,$$

where $t$-statistics given in parentheses are based on HAC standard errors with 24 lags; see the discussion of Table 20.3 for this choice of lag length. Even with a time trend included there appears to be a strong relationship. Note, however, that the HAC standard errors can only be used as a guide if interest rates are I(1) processes.

OLS regression in changes yields

$$\widehat{\Delta GS10}_t = \underset{(-0.62)}{-0.006} + \underset{(16.37)}{0.720} \times \Delta GS1_t,\ s_e = 0.180,\ R^2 = 0.570,$$

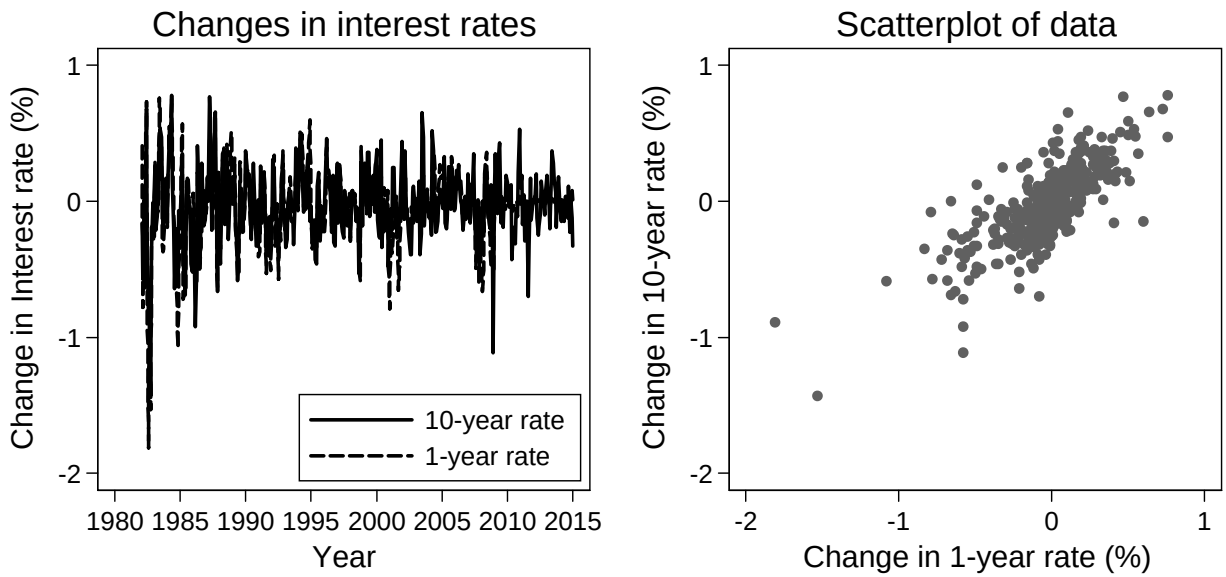


Figure 20.2: Changes in annualized rate of return on 1-year and 10-year U.S. Treasury bills from February 1982 to January 2015.

where $t$-statistics given in parentheses are based on HAC standard errors with 3 lags. This indicates a very strong relationship between the two rates. The coefficient is large – a one percentage point change in *GS1* is associated with a 0.72 percentage point change in *GS10* – and the $t$-statistic equals 16.37.

Figure 20.2 plots the monthly changes in interest rates over time (first panel) and provides a scatterplot of the changes (second panel). There does appear to be a relationship.

Running an OLS regression in levels with first-order lags included as regressors yields

$$\widehat{GS10}_t = \underset{(1.50)}{0.047} + \underset{(16.85)}{0.719} \times GS1_t + \underset{(110.37)}{0.978} \times GS10_{t-1} - \underset{(-16.41)}{0.701} \times GS1_{t-1},\ s_e = 0.180,\ R^2 = 0.996,$$

where $t$-statistics given in parentheses are based on HAC standard errors with 12 lags. The coefficient on $GS1_t$ remains highly statistically significant and meaningfully large, so we conclude that the relationship is not spurious. Note, however, that the $t$-statistic can only be a guide as the error is serially correlated and the model includes a lagged dependent variable.

The relationship between *GS10* and *GS1* does appear to be real. In this chapter the interest rates are viewed as having unit roots, so regression analysis is in first differences. This procedure is valid, though there are better more advanced methods that enable richer

modelling of the data; see Chapter 20.7.

**Remark 177** *If two series are integrated of order one then doing OLS regression in first differences leads to valid statistical inference and avoids spurious regression.*

## 20.5 Static Regression

By **static regression** we mean that the regressors do not include lags of the dependent variable. Dynamic or autoregressive models, ones that do include lags of the dependent variable as a regressor, raise additional issues and are presented in Chapter 20.6.

Additionally it is assumed that there is no unit root present in the dependent variable or the regressors. From the previous section, this can be determined by inspection of autocorrelations or by formal test, or by following conventions established in the economics literature.

In the applications in this section, $y_t$, $x_t$ and $z_t$ are changes in interest rates. Changes rather than levels of the interest rate are analyzed because the first differences are definitely stationary.

**Remark 178** *Chapter 20.5 considers only static regression, OLS regression when lagged dependent variables are* ***not*** *included in the regression.*

### 20.5.1 Model Assumptions

When model errors are independent across observations, model assumptions 1-2 given in Chapter 14.1 are sufficient for OLS to be unbiased, and little more needs to be assumed for OLS to be consistent. In particular, assumption 2 is that the errors and regressors for the current observation are uncorrelated. For time series models, where errors and/or regressors are correlated over time, this assumption needs to be strengthened.

We consider regression of $y_t$ on one or more regressors denoted $z_{2t}, ..., z_{kt}$. Then

$$y_t = \beta_1 + \beta_2 z_{2t} + \cdots + \beta_k z_{kt} + u_t,\ t = 1, ..., T.$$

While the regressors are dated to be observed at time $t$, they could instead be observed at earlier times. In particular in a finite distributed lag model the regressors are contemporaneous variables and their lags: $x_t, x_{t-1}, ..., x_{t-q}$. But in this section the regressors cannot include lags of the dependent variable $y_t$.

Let $\mathbf{z}_t = (z_{2t}, ..., z_{kt})$ denote the regressors included in the model for the $t^{th}$ time period. Assumption 2 is that errors at time $t$ are uncorrelated with the regressors at time $t$. In the time series context this assumption is called one of **contemporaneous exogeneity** or **present exogeneity** with

$$\mathrm{E}[u_t|\mathbf{z}_t] = 0.$$

If model errors are correlated over time, however, this assumption needs to be strengthened. It is usually sufficient to make the assumption, called **weak exogeneity**, that the error term at time $t$ is uncorrelated with all regressors in the current and in past time periods, so

$$\mathrm{E}[u_t|\mathbf{z}_t, \mathbf{z}_{t-1}, ..., \mathbf{z}_1] = 0.$$

The assumption is also called that of **present and past exogeneity**. A regressor satisfying this assumption is called a **predetermined**.

The weak exogeneity assumption fails if the regressors include lagged dependent variables and the errors are serially correlated; see Chapter 20.6 for explanation. For this reason models with lagged dependent variables, deferred to this later section, usually include enough lags so that the errors are serially uncorrelated.

For some estiamtors an even stronger assumption is needed. Under **strict exogeneity** the error term at time $t$ is uncorrelated with all regressors in all time periods, so

$$\mathrm{E}[u_t|\mathbf{z}_T, ..., \mathbf{z}_{t+1}, \mathbf{z}_t, \mathbf{z}_{t-1}, ..., \mathbf{z}_1] = 0, \text{ for all } t.$$

**Remark 179** *For time series models, model assumption 2 in Chapter 14.1 is replaced with the stronger assumption, called* ***weak exogeneity*** *or* ***present and past exogeneity****, that the error term is uncorrelated with present and past values of all regressors.*

### 20.5.2 Robust Inference with Serially Correlated Errors

We consider inference when model errors are serially correlated, under the assumption that regressors satisfy the contemporaneous and past exogeneity assumption. As already noted, this then rules out lagged dependent variables as regressors.

Correct standard errors can be obtained using **heteroskedastic and autocorrelation consistent (HAC)** standard errors, introduced in Chapter 10.7. Specialized software is needed to compute HAC errors. Several methods are proposed; the most commonly-used estimate is due to Newey and West (1987). In the simplest case of regression on a single regressor, so $y_t = \beta x_t + u_t$, the HAC variance estimate assuming that the error term $u_t$ is correlated for no more than $m$ lags, is

$$\widehat{\mathrm{Var}}_{HAC}[b] = \left\{\sum_{t=1}^T x_t^2 e_t^2 + 2\sum_{j=1}^m \tfrac{m+1-j}{m+1}\left(\sum_{t=j+1}^m x_t x_{t-j} e_t e_{t-j}\right)\right\} / (\sum_{t=1}^T x_t^2)^2.$$

If $m = 0$ this reduces to $\left\{\sum_{t=1}^T x_t^2 e_t^2\right\} / (\sum_{t=1}^T x_t^2)^2$ which is the heteroskedastic-robust estimate.

The HAC standard errors differ with the value of $m$, which should be chosen to be no larger than is necessary but large enough that model errors more than $m$ periods apart are uncorrelated. For moderately correlated errors with $\mathrm{Cor}[u_t, u_{t-1}] < 0.5$, rules of thumb

Table 20.3: Interest rates: HAC standard error of slope coefficient for various choices of lag length m

| **Regression** | **HAC standard error of slope coefficient for m lags** | | | | | |
|---|---|---|---|---|---|---|
| | m=0 | m=3 | m=6 | m=12 | m=24 | j=36 |
| $\Delta$*GS10* on $\Delta$*GS1* | 0.447 | 0.440 | 0.449 | 0.448 | 0.423 | 0.402 |
| *GS10* on *GS1* | 0.0114 | 0.0221 | 0.0283 | 0.0364 | 0.0450 | 0.0494 |

include letting $m$ equal the first integer larger than $0.75T^{1/3}$ or larger than $T^{1/3}$. If the error correlation is higher then larger $m$, possibly much larger $m$, is needed. The autocorrelations of the OLS residual provide a guide.

**Remark 180** *For static regression OLS can be used if errors are serially correlated, with inference based on heteroskedastic and autocorrelation consistent (HAC) standard errors.*

### 20.5.3 Example: Interest Rates

Table 20.3 lists HAC standard errors of the slope coefficient in two regression models, for various choices of the maximum lag length $m$.

The first row gives HAC standard errors for regression of $\Delta$*GS10* on an intercept and $\Delta$*GS1*. The autocorrelations of the OLS residual at lags 1-4 are 0.26, $-0.04$, 0.06 and 0.02. Since the autocorrelations are small and die out quickly, the HAC standard errors differ little with choice of $m$. The rule of thumb yields $m = 6$ since $0.75 \times 397^{1/3} = 5.51$. For these data the choice $m = 3$ appears reasonable.

For comparison, the second row gives HAC standard errors for regression of *GS10* on an intercept and *GS1*. These should only be used as a guide since, from the previous section, *GS10* and *GS1* are cointegrated so that standard inference results do not apply. The autocorrelations of the OLS residual at lags 1-4 are 0.95, 0.88, 0.83 and 0.77, and even at lag 24 the residual autocorrelation is 0.11. Since the autocorrelations are large and die out slowly the HAC standard errors vary greatly with choice of $m$ and a large value for $m$, say $m = 24$, is warranted.

### 20.5.4 More Efficient Inference with Serially Correlated Errors

When errors are serially correlated, HAC robust standard errors enable inference based on OLS estimation. But OLS is no longer the best estimator, as it is only fully efficient if the error $u_t$ is a white noise error so that model assumptions 1-4 are satisfied.

Instead, more efficient estimation is possible using feasible generalized least squares (FGLS) estimation, introduced in Chapter 14.7. The leading example for time series data

is the Cochrane-Orcutt estimator when the model error follows an AR(1) process, so $u_t = \rho u_{t-1} + \varepsilon_t$,where $\varepsilon_t$ is white noise and $|\rho| < 1$. This model implies that $\text{Cor}[u_t, u_{t-j}] = \rho^j$.

This estimator is included in many econometrics packages. Alternatively it can be implemented by the following procedure. First, perform usual OLS regression and obtain the OLS residual $e_t$. Second, estimate $\rho$ by $\widehat{\rho} = \Sigma_t e_t e_{t-1} / \Sigma_t e_{t-1}^2$. Third estimate by OLS regression the transformed model

$$y_t - \widehat{\rho} y_{t-1} = \beta_1(1 - \widehat{\rho}) + \beta_2(x_{2t} - \widehat{\rho} x_{2t-1}) + \cdots + \beta_k(x_{kt} - \widehat{\rho} x_{kt-1}) + v_t,\ t = 2..., T.$$

The usual standard errors can be used in this equation, assuming the AR(1) model is the correct model for the error process, even though $\widehat{\rho}$ is estimated. There are several variations of this method leading to estimates that differ by a small amount in finite samples.

While there are potential efficiency gains if the model error is indeed AR(1), it is more common practice to just use OLS, with HAC standard errors.

### 20.5.5 Finite Distributed Lag Model

The full effect on $y$ of a change in regressor $x$ may take several periods, rather than being concentrated in only the current period. Thus $y_t$ will depend not only on current $x_t$, but also immediate past values of $x_t$.

A finite distributed lag adds lagged values of the regressor $x_t$ to the regression. So

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_q x_{t-q} + u_t,\ t = q + 1, ..., T.$$

Here the regressors are not lagged dependent variables, so a serially correlated error is permitted.

For a model with lags in $x_t$, distinction is made between the **initial impact** ($\beta_0$) on $y_t$ of a change in $x_t$, and the **cumulative effect** ($\beta_0 + \beta_1 + \cdots + \beta_q$) over all $q$ time periods on $y_t$ of a change in $x_t$.

**Remark 181** *A finite distributed lag adds lagged values of the regressor to the regression.*

### 20.5.6 Example: Interest Rate Data

For the interest rate data with two lags added we obtain

$$\begin{aligned}\widehat{\Delta GS10}_t &= \underset{(-0.73)}{-0.008} + \underset{(15.23)}{0.759} \times \Delta GS1_t - \underset{(-1.42)}{0.058} \times \Delta GS1_{t-1} \\ &\quad - \underset{(-1.04)}{0.040} \times \Delta GS1_{t-2},\ s_e = 0.172,\ R^2 = 0.580.\end{aligned}$$

where $t$-statistics computed using HAC standard errors ($m = 3$) are given in parentheses. There is only a slight improvement in $R^2$, from 0.570 to 0.580, and a joint test finds that $\Delta GS1_{t-1}$ and $\Delta GS1_{t-2}$ are jointly statistically insignificant ($p = 0.31$).

The impact of a change in $\Delta GS1$ of one percentage point is a 0.759 percentage point change in $\Delta GS10$. The cumulative effect is a 0.661 percentage point change ($= 0.759 - 0.058 - 0.040$).

## 20.6 Dynamic Regression

Dynamic regressions again consider regression of the model

$$y_t = \beta_1 + \beta_2 z_{2t} + \cdots + \beta_k z_{kt} + u_t, \text{ for all } t.$$

Now, however, the regressors may include lags of the dependent variable, such as $y_{t-1}$.

This allows richer models for $y_t$, but also means that OLS can be inconsistent.

### 20.6.1 Inconsistency of OLS if Errors are Serially Correlated

Consider OLS estimation in the following model

$$\begin{aligned} y_t &= \beta_1 + \beta_2 y_{t-1} + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t, \end{aligned}$$

where $\varepsilon_t$ is white noise. This model features both a lagged dependent variable as a regressor and a serially correlated error.

The error $u_t$ is correlated with $u_{t-1}$, given the second equation. But $u_{t-1}$ is correlated with $y_{t-1}$, since the first equation implies that $u_{t-1} = y_{t-1} - \beta_1 - \beta_2 y_{t-2}$. Since $u_t$ is correlated with $u_{t-1}$ and $u_{t-1}$ is correlated with $y_{t-1}$, it follows that the error $u_t$ is correlated with the regressor $y_{t-1}$. But this violates assumption 2, that the error is uncorrelated with regressors. So OLS is inconsistent.

More generally, OLS is inconsistent in a model with lagged dependent variable(s) as regressors and a serially correlated error. One possibility is to use alternative estimation methods. More often, sufficient lags are added to the model to ensure that the errors are serially uncorrelated.

**Remark 182** *For dynamic regression OLS can only be used if sufficient lags of the dependent variable are included to yield errors that are serially uncorrelated.*

### 20.6.2 Test for Serial Correlation in Errors

It is important therefore to assess whether errors in a dynamic model are serially uncorrelated. In a static model the Ljung-Box Portmanteau test of Chapter 20.2 can be applied to the residuals after OLS regression. But this test is no longer valid if the residuals are from a dynamic model with lagged dependent variables as regressors.

Instead the **Breusch-Godfrey Lagrange multiplier** (LM) test is used. The null hypothesis is that the first $p$ autocorrelations of the error $u_t$ are all zero. First, obtain the residual $e_t$ after OLS regression of $y_t$ on an intercept and $z_{2t}, ..., z_{kt}$, where these regressors may include lagged dependent variables. Then estimate by OLS the auxiliary regression

$$e_t = \alpha_1 + \alpha_2 z_{2t} + \cdots + \alpha_k z_{kt} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-p} + v_t,\ t = 1, ..., T,$$

where missing lagged values of $e_{t-j}$ in the initial periods are replaced by zeroes. Then $T$ times $R^2$ from this regression is $\chi^2(p)$ distributed under $H_0$. If $TR^2 > \chi^2_{0.05}(p)$ then the null hypothesis of no correlation in the errors is rejected at significance level 0.05. An alternative version of the test, equivalent in large samples, is an $F$-test of $H_0 : \theta_1 = 0, ..., \theta_p = 0$.

While dynamic models provided the motivation for this test, the same test can be used to test error serial correlation in static models with regressors that do not include lagged dependent variables.

**Remark 183** *The Breusch-Godfrey Lagrange multiplier (LM) test is used to test whether model errors are serially correlated. It should be used instead of the Ljung-Box test if regressors include lagged dependent variables.*

### 20.6.3 Autoregressive model

The simplest dynamic model includes as regressors only lagged values of the dependent variable.

An **autoregressive model of order p** (an **AR(p)** model) is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t,\ t = p+1, ..., T,$$

where the error term $\varepsilon_t$ is white noise. To ensure stationarity the $\alpha$ coefficients need to satisfy a restriction, not given here, that generalizes the condition that $|\phi| < 1$ in the AR(1) model.

Valid statistical inference after OLS regression requires that sufficient lags in $y$ are included to ensure that $\varepsilon_t$ is white noise. If instead the error is serially correlated, then OLS is inconsistent as already noted.

**Remark 184** *An autoregressive model of order p (an AR(p) model) includes p lags of y as regressors.*

### 20.6.4 Autoregressive distributed lag model

The more general **autoregressive distributed lag model of orders p and q** (an **ADL(p,q)** model) additionally includes current and lagged values of $x_t$ as regressors. Then

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} \\ &\quad + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_q x_{t-q} + \varepsilon_t,\ t = (\max(p,q)+1), ..., T, \end{aligned}$$

Table 20.4: Static and Dynamic Models: Change in ten-year interest rates.

| **Variable** | **Model** | | |
|---|---|---|---|
| | Distributed Lag DL(2) | Autoregressive AR(2) | Autoregressive Distributed Lag ADL(2,2) |
| $\Delta GS10_{t-1}$ | | 0.374 | 0.283 |
| | | (6.27) | (5.17) |
| $\Delta GS10_{t-2}$ | | -0.121 | -0.102 |
| | | (-1.96) | (-1.86) |
| $\Delta GS1_t$ | 0.759 | | 0.740 |
| | (15.23) | | (17.10) |
| $\Delta GS1_{t-1}$ | -0.058 | | -0.259 |
| | (-1.42) | | (-3.99) |
| $\Delta GS1_{t-2}$ | -0.405 | | 0.051 |
| | (-1.05) | | (0.92) |
| *Intercept* | -0.007 | -0.023 | -0.007 |
| | (-0.73) | (-1.78) | (-0.80) |
| n | 394 | 394 | 494 |
| $R^2$ | 0.580 | 0.125 | 0.611 |
| St. error of regression | 0.178 | 0.256 | 0.172 |
| Breusch-Godfrey (12 lags) | 37.14 | 10.13 | 15.18 |
| p-value for BG Test | 0.000 | 0.604 | .232 |
| Standard errors calculation | HAC(3 lags) | Het-robust | Het-robust |

where it is assumed that enough lags are included so that the error $\varepsilon_t$ is white noise.

### 20.6.5 Example: Interest Rates

Table 20.4 presents OLS regression results for several models for $\Delta GS10$. Appropriate $t$-statistics, explained below, are given in parentheses.

The first model is the distributed lag model with up to two lags of $\Delta GS1$ as regressors. Estimates from this model have already been presented in Chapter 20.5. Since none of the regressors are lags of the dependent variable, OLS is consistent even if the error is serially correlated. The Breusch-Godfrey test, for autocorrelation of the error up to 3 lags, has $p = 0.000$ so the null hypothesis of no correlation in the error is clearly rejected. Due to this error correlation, the $t$-statistics are based on HAC standard errors, here with $m = 3$. The lagged values of $\Delta GS1$ are statistically insignificant at level 0.05, and further analysis reveals that the preferred distributed lag model has only $\Delta GS1_t$ as a regressor.

The second model in Table 20.4 is an AR(2) model for $\Delta GS10$. OLS is now inconsistent if the error is serially correlated, since lags of the dependent variable are regressors. The Breusch-Godfrey test has $p = 0.604$, so the the null hypothesis of no correlation in the error

is clearly not rejected, and OLS estimates are consistent. While the Breusch-Godfrey test is based on the assumption of white noise errors under the null hypothesis, the OLS estimator is consistent under the weaker assumption of heteroskedastic errors, provided the errors are independent. So the reported $t$-statistics are based on heteroskedastic–robust standard errors. Both lags of $\Delta GS10$ are statistically significant at level 0.05, and further analysis reveals that two lags are sufficient. The AR(2) model has much worse fit than the DL(2) model, with $R^2$ falling from 0.580 to 0.125.

The third model in Table 20.4 is an ADL(2,2) model that combines the previous two models. The model errors are serially uncorrelated, as was the case for the AR(2) model. The model fit is improved somewhat compared to the DL(2) model, with $R^2$ increased from 0.580 to 0.611. The reported $t$-statistics based on heteroskedastic–robust standard errors show that $\Delta GS1_{t-2}$ could be dropped from the model.

### 20.6.6 Forecasting

Forecasting is the prediction of future values of $y_t$ on the basis of past data and a dynamic model linking variables observed over time. The model is estimated using the sample, and the forecasts are out-of-sample predictions.

If the purpose of regression is forecasting, rather than inference, then the regression model may be selected on the basis of its forecast performance, rather than whether sufficient lags are added to ensure that the error is white noise. The model should not be chosen on the basis of $R^2$ alone, as $R^2$ provides no penalty for large models. A commonly-used method is to choose lag length to minimize the Bayesian information criterion (BIC), defined in Chapter 13.6.

Forecasting from a dynamic model is not straightforward. Consider forecasting from an estimated AR(2) model, with $\widehat{y}_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2}$. We wish to forecast in periods $T+1$, $T+2, ...$ given model estimates using data available up to time $T$. These forecasts are denoted $\widehat{y}_{T+1|T}$, $\widehat{y}_{T+2|T}, ....$ The forecast for period $T+1$ is a one-step ahead forecast that is straightforward as all necessary data are available. We obtain $\widehat{y}_{T+1|T} = b_0 + b_1 y_T + b_2 y_{T-1}$. The two-step ahead forecast for $T+2$ is more problematic, as data for $y_{T+1}$ is not available. Instead we estimate $y_{T+1}$ by the one-step ahead forecast $\widehat{y}_{T+1|T}$. We obtain $\widehat{y}_{T+2|T} = b_0 + b_1 \widehat{y}_{T+1|T} + b_2 y_T$. More generally, for an $h$-step ahead forecast from an AR(p) model, actual data are used for time period $T$ and earlier, while forecast values are used after time period $T$. The forecast is called a **recursive forecast**.

If the model is estimated in changes, then this method gives forecasts in changes. These can be converted to forecasts of levels. For example if $y_T = 3.2$ and the model yields forecast $\widehat{\Delta y}_{T+1|T} = 0.4$, then $\widehat{y}_{T+1|T} = y_T + \widehat{\Delta y}_{T+1|T} = 3.2 + 0.4 = 3.6$.

Forecasting from models with variables $x_t$ other than lags of $y_t$ is more problematic, as this requires forecasts of $x_t$. In some examples we may be willing to replace $x_{T+1}$ by a forecast $\widehat{x}_{T+1}$ obtained as an educated estimate or from an econometric model. More often

in this case a systems model is used that models both $y_t$ and $x_t$.

**Remark 185** *Forecasts of future values from a time series model are recursive forecasts.*

## 20.7 Key Concepts

1. Time series analysis is complicated by persistence in the data.
2. The sample autocorrelation at lag $j$ measures the correlation between $y_t$ and $y_{t-j}$. Persistence is higher the closer the sample autocorrelations are to one.
3. A white noise process has mean zero, constant variance and is uncorrelated across time.
4. Two common models for a persistent time series are the linear time trend model and the autoregressive moving average of order one (AR(1)) model.
5. A covariance stationary process is one that has the same mean and variance for all time and has covariance between two observations that depends only on the time separation between the observations.
6. A unit root process, also called integrated of order one, denoted I(1), is one that is nonstationary but whose first difference is stationary. A random walk is a leading example.
7. Spurious regression can arise because of omitted time trends, or because the data are unrelated unit root processes.
8. If two series are integrated of order one then doing OLS regression in first differences leads to valid statistical inference and avoids spurious regression.
9. A regression is said to be static if lagged dependent variables are not included in the regression.
10. For static regression OLS can be used if errors are serially correlated, with inference based on heteroskedastic and autocorrelation consistent (HAC) standard errors.
11. A regression is said to be dynamic if lagged dependent variables are included in the regression.
12. For dynamic regression OLS can only be used if sufficient lags of the dependent variable are included to yield errors that are serially uncorrelated.
13. The Breusch-Godfrey Lagrange multiplier (LM) test is used to test whether model errors are serially correlated.

14. An autoregressive model of order p (an AR(p) model) includes p lags of $y$ as regressors.

15. An autoregressive distributed lag model of orders p and q (an ADL(p,q) model) includes p lags of $y$ and the current and q lagged values of $x$ as regressors.

16. Forecasts of future values from a time series model are recursive forecasts.

17. Key words: persistence; autocorrelation; white noise process; Ljung-Box portmanteau test; Durbin-Watson test statistic; time trend model; autoregressive model; covariance stationary process; random walk with drift; random walk without drift; nonstationary process; unit root process; integrated of order one; explosive variance; spurious regression; static regression; heteroskedastic and autocorrelation consistent (HAC) standard errors; finite distributed lag model; dynamic regression; Breusch-Godfrey Lagrange multiplier test; autoregressive model of order p (AR(p)); autoregressive distributed lag model of orders p and q (ADL(p,q)); recursive forecast.

## 20.8 Exercises

Newey-west using Wooldridge method.

Exercise: Real interest rates.

Exercise: compute autocorrelations

## 20.9 Appendix 20.A: Autocorrelations

The **population correlation coefficient** between $y_t$ and $y_{t-j}$ equals $\text{Cov}[y_t, y_{t-j}]/\sqrt{\text{Var}[y_t]\text{Var}[y_{t-j}]}$. For a **stationary process** $\text{Var}[y_t] = \text{Var}[y_{t-j}]$, so the population correlation coefficient is

$$\rho_j = \frac{\text{Cov}[y_t, y_{t-j}]}{\text{Var}[y_t]}.$$

Then $\rho_0 = 1$, $-1 \leq \rho_j \leq 1$, and for a white noise process $\rho_j = 0$ for $j \neq 0$. Persistence is very high if $\rho_j$ is close to 1 for the first few lags.

There are several ways to estimate $\rho_j$. These yield the same estimate in large samples but differ in finite samples. A commonly-used refinement of the formula given in Chapter 20.2 is to compute the **sample autocorrelation** at lag $j$ as

$$\widehat{\rho}_j = \frac{\frac{1}{T-j}\sum_{t=j+1}^{T}(y_t - \overline{y}^*)(y_{t-j} - \overline{y}^{**})}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2},$$

where $\overline{y}^* = \frac{1}{T-j}\sum_{t=j+1}^{T} y_t$ and $\overline{y}^* = \frac{1}{T-j}\sum_{t=j+1}^{T} y_{t-j}$. Other estimates, ones that do not require specialized software, are the sample correlation coefficient between $y_t$ and $y_{t-j}$, and the slope coefficient from OLS regression of $y_t$ on an intercept and $y_{t-j}$.

## 20.10 Appendix 20.B: Unit Root and Cointegration Tests

This appendix contains more advanced material that is relevant only for highly persistent data.

### 20.10.1 Dickey-Fuller Test

The nonstationary random walk model can be viewed as an AR(1) model with $\phi = 1$. If instead $|\phi| < 1$ then the AR(1) model is stationary, as some algebra shows that then $y_t$ has constant mean $\delta/(1-\phi)$ and constant variance $\sigma_\varepsilon^2/(1-\phi)$.

A test for stationarity estimates the AR(1) model

$$y_t = \alpha + \delta t + \phi y_{t-1} + \varepsilon_t$$

and tests whether $\phi = 1$ against the alternative $\phi < 1$. The linear time trend term is added to allow for the possibility of a time trend under the alternative $\phi < 1$ (recall that the random walk plus drift model has mean $y_0 + \delta t$).

It is convenient to subtract $y_{t-1}$ from both sides of the AR(1) model and equivalently estimate by OLS the transformed model

$$\begin{aligned} \Delta y_t &= \alpha + \delta t + (\phi - 1) y_{t-1} + \varepsilon_t \\ &= \alpha + \delta t + \gamma y_{t-1} + \varepsilon_t, \end{aligned}$$

where $\gamma = \phi - 1$ and $\varepsilon_t$ is assumed to be white noise.

Then we test whether $\gamma$ equals zero or is less than zero, on the basis of the usual $t$-statistic. The test is called the Dickey-Fuller test after the two statisticians who proposed this test and obtained its critical values.

**Remark 186** *The Dickey-Fuller test is a* ***one-sided test*** *of the null hypothesis $H_0 : \phi = 1$, a nonstationary process, against the alternative hypothesis $H_a : \phi < 1$, a stationary process. The test statistic is obtained by OLS regression in the model $\Delta y_t = \alpha + \delta t + \gamma y_{t-1} + \varepsilon_t$ and the test statistic is*

$$t_{DF} = \frac{\widehat{\gamma}}{se(\widehat{\gamma})}.$$

*The critical values and p-values are obtained from special tables and vary with the sample size. For example, if the sample size is very large we reject $H_0$ at level 0.05 if $t_{DF} < -3.41$.*

Note that the critical values are not based on the standard normal distribution or the $t$ distribution. If the usual standard normal critical value of $-1.645$ (for $\alpha = 0.05$) is used, then $H_0$ will be erroneously rejected much of the time.

### 20.10.2 Augmented Dickey-Fuller (ADF) Test

The critical values of the Dickey-Fuller test are obtained under the assumption that the error $\varepsilon_t$ in the AR(1) model is white noise. In practice this assumption is often violated as the error is often serially correlated. The inclusion of lags of $\Delta y_t$ in the model can lead to an error that is white noise.

The **augmented Dickey-Fuller** (**ADF**) test estimates by OLS

$$\Delta y_t = \alpha + \delta t + \gamma y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p} + \varepsilon_t,$$

where sufficient lags are added so that $\varepsilon_t$ is serially uncorrelated. One criteria used to determined the lag length is AIC, defined in Chapter 13.6. The test based on $t_{ADF} = \widehat{\gamma}/se(\widehat{\gamma})$ uses the same critical values as those given earlier for the Dickey-Fuller test.

In some cases even the differenced series $\Delta y_t$ may still be nonstationary. In that case we consider the second difference $\Delta^2 y_t = \Delta(\Delta y_t) = \Delta y_t - \Delta y_{t-1}$. This possibility can be tested by applying the same ADF test procedure to $\Delta y_t$, estimating by OLS the model $\Delta^2 y_t = \alpha + \delta t + (1-\phi)\Delta y_{t-1} + \gamma_1 \Delta^2 y_{t-1} + \cdots + \gamma_1 \Delta^2 y_{t-p} + \varepsilon_t$, where $\Delta^2 y_t = \Delta(\Delta y_t) = \Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$. If $\Delta y_t$ is nonstationary and $\Delta^2 y_t$ is stationary then $y_t$ is integrated of order two (I(2)).

### 20.10.3 Alternative Unit Root Tests

A weakness of unit root tests is that they can have low power, especially if the time series is short. In the current context this means that there can be low probability of rejecting nonstationarity ($H_0$) even when the series is stationary ($H_a$). This is not desirable as analysis is simpler if the data can be viewed as stationary.

There are several variants of the DF and ADF test depending on inclusion of the intercept and time trend (or even time trend squared). In particular a version of the test that omits the time trend ($\delta t$), and then has different critical values to those above, has greater power provided it is correct to omit the trend. Omitting the time trend may be reasonable for non-trending series such as unemployment rates. In this chapter only the version with a time trend is presented as it is valid for either the trending or nontrending case.

Many alternative methods for testing for unit roots have been proposed. One advantage of these other tests is that they may have higher power. A second advantage is that some tests do not require specifying a lag length, unlike the ADF test. The most commonly-used tests, aside from the ADF test, are a serially correlated error correction of the DF test proposed by Phillips and Perron (1998) and a GLS version of the DF test proposed by Elliott, Rothenberg and Stock (1996).

The ADF test is popular as it is both easy to implement and easy to understand. In particular, even in the absence of a formal test, if the ADF regression yields $\widehat{\gamma} = \widehat{\phi - 1} >$

$-0.1$, say, so $\widehat{\phi} > 0.9$, then persistence may be strong enough to warrant treating $y_t$ as a unit root process.

### 20.10.4 Example: Interest Rates

We apply the augmented Dickey Fuller test to the ten-year rate. The test includes a trend. Three lagged changes are sufficient as subsequent lags have very small $t$-statistics.

OLS regression yields

$$\begin{aligned}\widehat{\Delta GS10}_t &= \underset{(4.35)}{0.637} - \underset{(-4.14)}{0.0014} \times t - \underset{(-4.70)}{0.063} \times \ GS10_{t-1} + \underset{(8.03)}{0.397} \times \Delta GS10_{t-1} \\ &\quad - \underset{(-2.34)}{0.123} \times \Delta GS10_{t-2} + \underset{(1.83)}{0.091} \times \ \Delta GS10_{t-3},\end{aligned}$$

where $t$-statistics are given in parentheses. Then $t_{ADF} = -4.70$, the $t$-statistic on $GS10_{t-1}$.

From computer output or tables (with $T = 393$) the critical values for the ADF test with trend are $-3.984$ at 1%, $-3.424$ at 5% and $-3.130$ at 10%. So the null hypothesis, that there is a unit root, is rejected even at significance level 1%, since $t_{ADF} = -4.70 < -3.984$. We conclude that there is not a unit root. For *GS1* a similar test yields $t_{ADF} = -4.10$, so again nonstationarity is rejected.

While the ADF test rejects nonstationarity, the estimate $\widehat{\gamma} = -0.063$ implies $\widehat{\phi} = 0.937$ which is close to one. And an alternative test, the GLS version of the DF test obtains a test statistic of $-2.012$ (with three lags) that exceeds even the 10% critical value of $-2.589$, so with this alternative test the null hypothesis of nonstationarity is not rejected even at 10%.

The evidence is mixed. It is safest to treat the interest rate series as nonstationary, and to do regression analysis in changes in the interest rates.

### 20.10.5 Cointegration Test

The unit root series $y_t$ and $x_t$ are cointegrated if the linear combination $y_t - \beta_1 - \beta_2 x_t$ is I(0). Since $\beta_1$ and $\beta_2$ are not observed, OLS estimates $b_1$ and $b_2$ are obtained for the model

$$y_t = \beta_1 + \beta_2 x_t + u_t.$$

Then test whether the residual $e_t = y_t - b_1 - b_2 x_t$ is stationary using an ADF test, where the no-trend version of the ADF test is used. The null hypothesis is that $u_t$ is I(1), in which case $y_t$ and $x_t$ are not cointegrated, and the alternative hypothesis is that $u_t$ is I(0), so that $y_t$ and $x_t$ are cointegrated. The critical values are the same as those for ADF test of $y_t$ alone.

Alternative tests for cointegration have been developed. The ADF test of residuals has the advantage of simplicity.

### 20.10.6 Example: Interest Rates

We test whether the ten-year and one-year rates are cointegrated. The residual from OLS regression in levels is $e_t = GS10_t - 2.264 - 0.836 \times GS1_t$. Three lagged changes are sufficient for these data. The ADF test on the residuals yields

$$\widehat{\Delta e}_t = \underset{(0.10)}{-0.001} - \underset{(-2.89)}{0.0294} \times e_{t-1} + \underset{(6.35)}{0.318} \times \Delta e_{t-1} - \underset{(-1.99)}{0.103} \times \Delta e_{t-2} + \underset{(2.87)}{0.43} \times \Delta e_{t-3},$$

where $t$-statistics are given in parentheses. Then $t_{ADF} = -2.89$, the $t$-statistic on $e_{t-1}$.

From computer output or tables (with $T = 393$) the critical values for the ADF test with no trend are $-3.449$ at 1%, $-2.874$ at 5% and $-2.570$ at 10%. So the null hypothesis, that there is a unit root, is rejected at significance level 5%, since $t_{ADF} = -2.89 < -2.874$. We conclude that the two series are cointegrated.

## 20.11 Appendix 20.C: Additional Time Series Models

Many different models, with associated acronyms, are used in time series analysis. For completeness this appendix presents a brief overview of methods that are quite advanced compared to the rest of this book.

### 20.11.1 ARMA Models

A **moving average error model of order $q$ (MA(q))** specifies the model error $u_t$ to be a weighted average over time (a moving average) of a white noise error. Thus

$$u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\varepsilon_t$ is white noise.

An **autoregressive moving average error model of orders $p$ and $q$ (MA(p,q))** combines the AR and MA models, so

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

An ARMA model has the attraction that it may enable modelling the persistence in $y_t$ with fewer parameters than an AR model. But MA and ARMA models require specialized estimation procedures, whereas an AR model can be estimated by OLS, provided sufficient lags are included to ensure that the error $u_t$ is serially uncorrelated. For this reason it is common to use AR models.

### 20.11.2 ARCH model

Most regression analysis models the conditional mean of the dependent variable, with the conditional variance viewed as a nuisance that needs to be properly accounted for to ensure valid statistical inference.

For financial data, however, there is great interest in additionally modelling the conditional variance. Financial data experience periods of relatively high volatility and periods of relatively low volatility. Periods of greater volatility, usually measured by the variance, mean greater risk. An investor may want to be compensated for this higher risk with a higher return. And options-pricing models require an estimate of the variance of the underlying asset.

The class of models called **autoregressive conditional heteroskedasticity (ARCH) models** are regression models that model both the conditional mean and the conditional variance. The starting point is the model

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t \\ \varepsilon_t &\sim [0, \sigma_t^2]. \end{aligned}$$

A range of models are obtained depending on how the error variance $\sigma_t^2$ evolves over time.

The **ARCH(m) model** specifies that the variance depends on the squared error in the previous $m$ periods, so

$$\sigma_t^2 = \gamma_0 + \gamma_1 \varepsilon_{t-1}^2 + \cdots + \gamma_m \varepsilon_{t-m}^2.$$

In applications $m$ can be quite large.

The **generalized ARCH model of orders 1 and 1** (**GARCH(1,1)**) specifies that

$$\sigma_t^2 = \gamma_0 + \gamma_1 \varepsilon_{t-1}^2 + \delta_1 \sigma_{t-1}^2,$$

so that the conditional variance depends on both the squared error and the error variance in the preceding period. This model often involves less parameters than an ARCH(m). The model can be generalized to a GARCH(m,n) model, but in practice a GARCH(1,1) model is often adequate.

Estimation of ARCH and GARCH models requires specialized software, due to the need to estimate the parameters in the model for $\sigma_t^2$.

### 20.11.3 Systems Models

In many macroeconomic settings, such as interest rates, the variables $y_t$ and $x_t$ are jointly determined, whereas to date we have modelled $y_t$ as being determined by $x_t$. Modelling the two jointly has the potential to lead to more efficient estimation and better forecasts.

The **vector autoregressive model** treats $y_t$ and $x_t$ symmetrically, as components of the vector $(x_t, y_t)$. Then $y_t$ depends on lags of $y_t$ and $x_t$, while $x_t$ depends on lags of $x_t$ and $y_t$. A **vector autoregressive model of order** $p$ (**VAR(p)**) for $y_t$ and $x_t$ is

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} + \varepsilon_t \\ x_t &= \gamma_0 + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + \delta_1 x_{t-1} + \cdots + \delta_p x_{t-p} + \eta_t. \end{aligned}$$

Sufficient lags need to be included to ensure that the errors $\varepsilon_t$ and $\eta_t$ are serially uncorrelated, though they may be contemporaneously correlated with each other. More precisely, $(\varepsilon_t, \eta_t)$ is a generalization of white noise with means zero, variances $\sigma_\varepsilon^2$ and $\sigma_\eta^2$, and covariance $\sigma_{\varepsilon\eta}$ and $(\varepsilon_t, \eta_t)$ independent of $(\varepsilon_s, \eta_s)$ for $t \neq s$.

The two equations are like those for an ADL(p,p) model, except that no current period variables appear as regressors. This does not mean that $y_t$ and $x_t$ are unrelated, however, as $y_t$ and $x_t$ depend on errors $\varepsilon_t$ and $\eta_t$ that may be contemporaneously correlated with each other.

The two equations are estimated separately by OLS. Usually if model errors are correlated across equations, the case here, jointly estimation the two equations leads to more efficient parameter estimates. But there is no such gain if the two equations have exactly the same set of regressors, which is the case here. So equation by equation OLS is fine.

### 20.11.4 Dynamic OLS for Cointegrated Series

Consider two series $y_t$ and $x_t$ that are cointegrated. Specifically, $y_t$ and $x_t$ are unit root processes, $\Delta y_t$ and $\Delta x_t$ are stationary processes, and there is a linear combination $y_t - \beta_1 - \beta_2 x_t$ that is stationary.

To date we have considered only regression in first differences for such data. Now consider regression in levels. Then the coefficients of OLS regression of $y_t$ on $x_t$ are consistent, but the usual theory does not apply so that neither the default OLS standard errors nor HAC standard errors are correct.

Default OLS standard errors can be used if estimation is by the **dynamic OLS estimator** that adds many leads and lags of changes in $x_t$ as regressors. Thus estimate by OLS the model

$$y_t = \beta_0 + \beta_2 x_t + \gamma_{-p}\Delta x_{t-p} + \gamma_{-p-1}\Delta x_{t-p+1} + \cdots + \gamma_p \Delta x_{t+p} + u_t.$$

Then inference on $\beta_2$ is based on HAC standard errors.

### 20.11.5 Error Correction Model

Again consider two series $y_t$ and $x_t$ that are cointegrated. Modelling the data in differences ignores any long-term relationship between the series. Incorporating the long-run relationship should improve model fit and forecasts.

Begin with a VAR(2) model in levels for $y_t$, so

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \varepsilon_t.$$

In the long-run (with $y_t$ constant and $x_t$ constant), this model implies $y = \alpha_0 + \alpha_1 y + \alpha_2 y + \beta_1 x + \beta_2$, or $y = \lambda_0 + \lambda_1 x$ where $\lambda_0 = \alpha_0/(1-\alpha_1-\alpha_2)$ and $\lambda_1 = (\beta_1+\beta_2)/(1-\alpha_1-\alpha_2)$.

The VAR(2) model can be re-expressed to a model that is partly in changes and partly in levels, with

$$\Delta y_t = \alpha_0 + (\alpha_1+\alpha_2-1)y_{t-1} + \alpha_2 \Delta y_{t-1} + (\beta_1+\beta_2)x_{t-1} + \beta_2 \Delta x_{t-1} + \varepsilon_t.$$

This model can in turn be re-expressed as

$$\Delta y_t = (\alpha_1+\alpha_2-1)(y_{t-1} - \tfrac{\alpha_0}{1-\alpha_1-\alpha_2} - \tfrac{\beta_1+\beta_2}{1-\alpha_1-\alpha_2}x_{t-1}) + \alpha_2 \Delta y_{t-1} + \beta_2 \Delta x_{t-1} + \varepsilon_t,$$

or more simply as

$$\Delta y_t = \lambda_2(y_{t-1} - (\lambda_0 + \lambda_1 x_{t-1})) + \alpha_2 \Delta y_{t-1} + \beta_2 \Delta x_{t-1} + \varepsilon_t,$$

where $\lambda_2 = \alpha_1+\alpha_2-1$, $\lambda_0 = \alpha_0/(1-\alpha_1-\alpha_2)$ and $\lambda_1 = (\beta_1+\beta_2)/(1-\alpha_1-\alpha_2)$. Note that $\lambda_0$ and $\lambda_1$ are the coefficients in the long-run relationship derived earlier.

This final model is a VAR(2) model in first differences, augmented by the term $y_{t-1} - (\lambda_0 + \lambda_1 x_{t-1})$ that is called an error correction term as it measures the deviation of last period's value $y_{t-1}$ from its equilibrium value of $\lambda_0 + \lambda_1 x_{t-1}$. The model is then called an **error correction model** (ECM).

More generally a VAR(p) model in levels can be re-expressed as a VAR(p−1) model in differences plus an error correction term. The Engle-Granger two-step estimator first computes $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$ by OLS regression of $y_t$ on $x_t$. Then estimate a VAR(p−1) model of $\Delta y_t$ on $\Delta x_t$, with additional inclusion of the fitted error correction term $y_{t-1} - (\widehat{\lambda}_0 + \widehat{\lambda}_1 x_{t-1})$.

The Engle-Granger two-step estimator is convenient as it can be implemented using only OLS regression. But it has several limitations, including estimators that are biased in finite samples and a lack of uniqueness in computing the error correction term (it differs depending on whether we regress $y_t$ on $x_t$ or regress $x_t$ on $y_t$).

Instead it is standard to use Johansen's procedure that directly estimates a vector error correction model while imposing restrictions on the parameters in this model to ensure that variables are cointegrated. Furthermore, these restrictions are testable, allowing tests for whether or not variables are cointegrated. This requires specialized software that is included in econometrics packages such as Stata and Eviews.

# Chapter 21

# © A. Colin Cameron: Panel Data

Panel data or longitudinal data are data on a number of individuals observed at each of a number of time periods. Examples include data over time on individual persons, firms or countries.

This chapter focuses on the type of panel data and the panel data methods most often used by economists. The data are assumed to be observed at regular intervals of time, such as daily, monthly, quarterly or annually, rather than intermittently such as medical data obtained at periodic doctor visits.

Short panels, with many individuals and relatively few time periods, are emphasized in this chapter. Long panels, with relatively few individuals and many time periods, are discussed briefly at the end of the chapter. Analysis of short panels is simpler as time series methods are not required.

The simplest models continue to use the OLS estimator, though inference should be based on cluster-robust standard errors, introduced in Chapter 19.3, that adjust for the likely correlation of model errors over time while assuming independence over individuals. For many purposes this is sufficient, especially if one only wants a descriptive noncausal summary of the relationship between the dependent variable and regressors.

Additionally, richer models become estimable with panel data. In particular, the random effects and fixed effects models and their associated estimators allow the intercept to vary across individuals. In principle these estimators can be implemented by OLS estimation of a suitably transformed model, but coefficient estimates and their associated standard errors are much more easily obtained, and without potential error, if one uses software that includes a panel methods component. Such software also enables distinguishing between variation of variables across individuals and variation over time.

Table 21.1: NBA Team revenue: Variable definitions and summary statistics (n=29, T=10).

| Variable | Definition | Mean | Standard deviation | Min | Max |
|---|---|---|---|---|---|
| *Revenue* | Team revenue in millions of 1999 $ | 95.714 | 24.442 | 58.496 | 187.721 |
| *Lnrevenue* | Natural logarithm of team revenue | 4.532 | 0.236 | 4.069 | 5.235 |
| *Wins* | Number of wins including playoff games | 41.04 | 12.438 | 9 | 67 |
| *Playoff* | =1 if made playoffs in previous season | 0.545 | 0.499 | 0 | 1 |
| *Champ* | =1 if champion in previous season | 0.035 | 0.184 | 0 | 1 |
| *Allstars* | Number of players voted Allstars | 0.860 | 0.871 | 0 | 4 |
| *Lncitypop* | Natural log of city population in millions | 1.301 | 0.801 | -0.0152 | 2.940 |
| *Teamid* | Team identifier | 14.86 | 8.35 | 1 | 29 |
| *Season* | Season identifier | 5.54 | 2.87 | 1 | 10 |

## 21.1 Data Example: NBA Team Revenue

We consider the relationship between the current season revenue of National Basketball Association (NBA) teams and the number of wins (including playoff games) in the previous season.

### 21.1.1 Data Summary

The data on twenty-nine teams for the ten seasons 2001-02 to 2010-11 are in dataset NBA.

A **balanced panel** is one for which all $N$ individuals are observed in all $T$ time periods. In this example the Charlotte Bobcats expansion team was only established in 2004, so the panel is (mildly) unbalanced, with 286 observations rather than $29 \times 10 = 290$ if it was balanced. The formulae given below are generally for balanced panels, for simplicity. Most methods can be adapted to unbalanced panels.

A **short panel** is one for which there are many individuals and relatively few time periods. The panel can have as few as two observations over time for an individual. Dataset NBA, with 29 individuals and 10 time periods, is treated as being a short panel.

The dataset is arranged so that each line contains all data for a team-season pair. Alternatively the dataset could have been arranged with each line containing all data for all seasons for a given team, or with each line containing all data for all teams in a given season. It may be necessary to rearrange data to get it into the format required by a statistical package. And the dataset should have a variable that identifies the team and a variable that identifies the season.

Table 21.1 presents variable definitions and summary statistics. The average team revenue was $96 million and the average number of wins was 41. There is considerable variation in both revenue and wins.

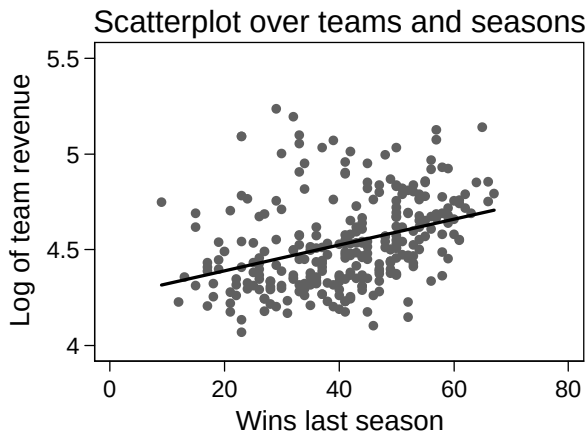


Figure 21.1: Natural logarithm of team earnings plotted against number of wins in previous season for 29 teams over 10 seasons.

Figure 21.1 presents a scatter plot of the natural logarithm of team revenue (in millions of dollars) against number of wins last season. There is a positive relationship, with correlation coefficient of 0.36 and fitted OLS regression line with slope coefficient 0.0068. This coefficient in a log-linear model is a semi-elasticity, see Chapter 12.6, so one more win last season is associated with a 0.68% increase in team earnings.

The observed positive relationship may arise for several reasons. It may be that some teams are successful year after year, with both high revenue and high number of wins – like the New York Yankees in baseball – while others are unsuccessful year after year. Or it may be that there is a relatively level playing field, with teams enjoying higher revenues following seasons with more wins. The analysis in this chapter seeks to identify these different mechanisms.

**Remark 187** *A **balanced panel** is one for which data are observed for all individuals in all time periods.*

**Remark 188** *A **short panel** is one for which there are many individuals and model errors are uncorrelated across individuals. A short panel can have as few as two observations per individual.*

### 21.1.2 Within and Between Variation

The summary statistics and Figure 21.1 summarize the **overall variation** in the data. This combines variation over time for a given individual, called **within variation**, and variation across teams, called **between variation**.

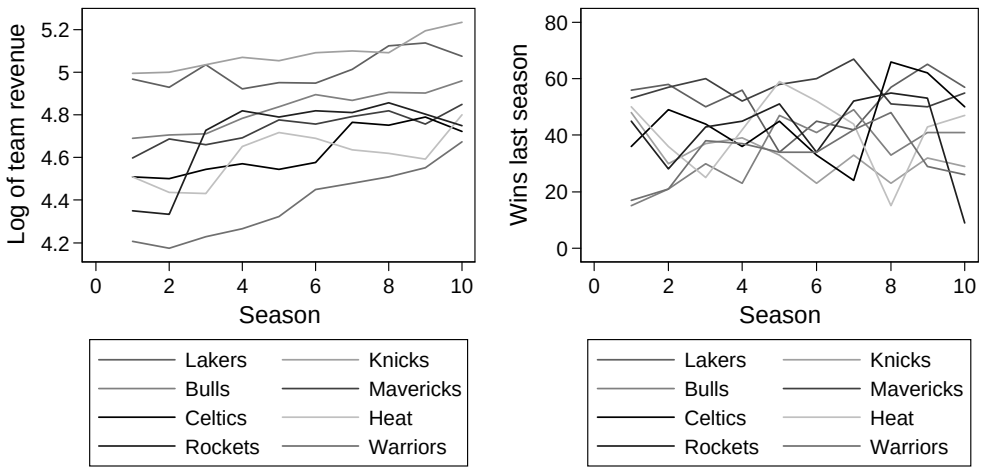


Figure 21.2: Panel data: (1) Natural logarithm of team earnings plotted against season for 8 teams. (2) Number of wins least season earnings plotted against season for 8 teams.

Figure 21.2 presents time series plots of the natural logarithm of team revenue (first panel) and wins per season (second panel) for eight of the twenty-nine teams. Team revenues are clearly trending upward over time, even after the adjustment for inflation, and the regressions presented below control for these time trends. The number of wins does not trend up over time since the regular season is of fixed length over this time and the playoff format is unchanging. Each time series plot shows the within variation, while the difference in the plots across teams shows the between variation. It appears that between variation is most important for team revenue while within variation is more important for wins.

Let $x_{it}$ denote variable $x$ observed for individual $i$ at time $t$, let $\bar{x} = \frac{1}{NT}\sum_i \sum_t x_{it}$ denote the average value of $x$ where averaging is over both individuals and time, called the **grand mean**. And let $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$ denote the average value of $x_{it}$ over time for individual $i$, called an **individual mean**.

The variability of $x_{it}$ around the grand mean $\bar{x}$, the **total sum of squares**, can be decomposed into the variation of $x_{it}$ around the individual means $\bar{x}_i$, called the **within sum of squares**, and the variation of the individual means $\bar{x}_i$ around the grand mean $\bar{x}$, called the **between sum of squares**. Algebraically

$$\begin{aligned}\sum_{i=1}^{T}\sum_{t=1}^{T}(x_{it}-\bar{x})^2 &= \sum_{i=1}^{T}\sum_{t=1}^{T}[(x_{it}-\bar{x}_i)+(\bar{x}_i-\bar{x})]^2 \\ &= \sum_{i=1}^{T}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2+\sum_{i=1}^{T}\sum_{t=1}^{T}(\bar{x}_i-\bar{x})^2,\end{aligned}$$

as the cross-product term $2\sum_i\sum_t(x_{it}-\bar{x}_i)(\bar{x}_i-\bar{x})$ can be shown to equal zero.

It is more common to present this decomposition in terms of standard deviations. We present formulas for the case of balanced panels. The usual formula for the standard deviation yields the **overall standard deviation** $s_o$, where $s_o^2 = \frac{1}{NT-1}\sum_i \sum_t (x_{it} - \bar{x})^2$. The **within standard deviation** is denoted $s_w$, where $s_w^2 = \frac{1}{NT-N}\sum_i \sum_t (x_{it} - \bar{x}_i)^2$. The **between standard deviation** is denoted $s_b$, where $s_b^2 = \frac{1}{N-1}\sum_i (\bar{x}_i - \bar{x})^2$. It can be shown that

$$s_o^2 \simeq s_w^2 + s_b^2.$$

The lower is within variation, then the more highly correlated are the regressors over time for a given individual. Going the other way, the within correlation of regressors is high when most of the variation is between variation.

For the natural logarithm of team revenues, $s_o^2 = .056$, $s_b^2 = .045$ and $s_w^2 = .017$, so substantially more of the variation in revenue comes from between variation. Thus some teams year after year have relatively high revenues while others have low revenues. In fact average team revenue over the ten seasons ranged from \$70 million for the Bucks to \$162 million for the Knicks.

For the number of wins, $s_o^2 = 154.7$, $s_b^2 = 49.6$ and $s_w^2 = 107.3$, so more of the variation in wins comes from within variation. Thus teams on average experience both good and bad seasons over the ten years. At the extremes, the Mavericks had between 50 and 67 wins, while the Rockets had between 9 and 50 wins.

The distinction between the various types of variation is important because different estimators use different sources of variation. Thus, for example, if there is little within variation in the data then an estimator that uses only within variation will be very imprecise.

**Remark 189** *The **overall variation** in the data can be decomposed into variation over time for a given individual, called **within variation**, and variation across teams, called **between variation**. Different estimators emphasize different sources of variation.*

## 21.2 Pooled OLS Estimation

The **pooled OLS estimator** is the OLS estimator in the model

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \cdots + \beta_k x_{kit} + u_{it},\ t = 1, ..., T,\ i = 1, ..., n.$$

Thus simply regress $y$ on an intercept and $x_2, x_3, ..., x_k$ using all available observations. The term pooled is used as the cross-section and time series variation in the data have been pooled or combined.

### 21.2.1 Cluster-Robust Standard Errors

For a short panel it is usually assumed that the model errors $u_{it}$ are independent across individuals $i$ but correlated over time $t$ for a given individual. This is an example of **clustered errors**, discussed at some length in Chapter 19.3.

So inference should be based on cluster-robust standard errors, where clustering is on the individual. Assumption 2 that the error has conditional mean zero now needs to condition on all observations in the same cluster. Let $\mathbf{X}_i$ denote all data on the regressors for individual $i$, so $\mathbf{X}_i = x_{2i1}, ..., x_{ki1}, ..., x_{2iT}, ..., x_{kiT}$. Then assumption 2 becomes

$$\mathrm{E}[u_i|\mathbf{X}_i] = 0, \text{ for individual } i.$$

The key assumption is that errors for a given individual (the cluster unit) in different clusters are independent, but errors over time for a given individual are potentially correlated. Thus assumption 4 is relaxed to the following:

$$\sigma_{ijts} = \mathrm{Cor}[u_{it}, u_{js}|\mathbf{X}_i, \mathbf{X}_j] \begin{cases} \neq 0 & i = j \\ = 0 & i \neq j. \end{cases}$$

The error variances $\sigma_{it}^2$ and covariances $\sigma_{ijts}$ may depend on the value of the regressors, so errors are also potentially heteroskedastic.

For short panels **cluster-robust standard errors** are implemented by clustering on the individual. Ideally there are at least thirty individuals, as the underlying statistical theory assumes the number of individuals goes to infinity. Also recall from Chapter 19.3 that different statistical packages can give slightly different values for the cluster-robust standard errors due to different degrees of freedom corrections. In the panel context these standard errors are also called **panel-robust standard errors**.

It is essential in any panel application to control for possible error correlation over time. Intuitively, adding one more time period of data for each individual does not bring in completely independent additional information. For example, if the model overpredicts (or underpredicts) revenue given wins using ten seasons of data it is also likely to overpredict (or underpredict) wins if data on an eleventh season is added. The default and heteroskedastic-robust standard errors overstate estimator precision, and understate the standard errors, as they erroneously treat the additional data as independent information.

In some applications one may cluster at a more aggregated level than an individual. For example the panel data may be observations over time for individuals in various regions and, as in Chapter 19.3, model errors may be uncorrelated across regions but are correlated for individuals in the same region. In that case the pooled OLS standard errors should be obtained by clustering on region, rather than the individual.

**Remark 190** *Standard OLS estimation with panel data is called pooled OLS as data across individuals and over time are pooled or combined. For a short panel inference is based on cluster-robust standard errors.*

Table 21.2: NBA Team Revenue: Pooled OLS with various standard error estimation methods.

| Variable | Coefficient | Default | | Het robust | | Cluster robust | |
|---|---|---|---|---|---|---|---|
| | | St. Error | t | St. Error | t | St. Error | t |
| *Wins* | .00681 | .00102 | 6.65 | .00099 | 6.88 | .00190 | 3.59 |
| *Season* | .01824 | .00443 | 4.11 | .00450 | 4.05 | .00331 | 5.51 |
| *Intercept* | 4.15162 | .05060 | 82.04 | .05075 | 81.81 | .09658 | 42.99 |
| Observations | 286 | | | | | | |
| $R^2$ | .176 | | | | | | |

### 21.2.2 Example: NBA Team Revenue

Table 21.2 presents OLS estimates from regression of the natural logarithm of team revenues on an intercept, *Wins*, and the variable *Season* which equals 1 in the first season, 2 in the second season, and so on. The variable *Season* acts as a linear time trend, which is evident in the first panel of Figure 21.2.

The estimated slope coefficients indicate that team revenue increases by 0.68% with each extra win, and team revenue is increasing (in real terms) by 1.82% per year.

Table 21.2 presents three different standard errors and associated $t$-statistics. The default and heteroskedastic robust standard errors are very similar. By contrast the cluster-robust standard error for the key variable, *Wins*, is 1.86 times as large (= .00190/.00102). Using these preferred standard errors variable *Wins* is nonetheless still highly statistically significant.

### 21.2.3 Further Details on Cluster-Robust Inference

Adapting the formula in Chapter 19.3 to the panel case, for model errors correlated over time for a given individual the default standard error of the OLS coefficient of the $k^{th}$ regressor should be inflated by approximately

$$\tau_k \simeq \sqrt{1 + \rho_{x_k}\rho_u((obs/n) - 1)},$$

where $\rho_{x_k}$ is a measure of the within-cluster correlation of $x_k$, $\rho_u$ is the within-cluster error correlation, *obs* denotes the total number of observations and $n$ is the number of individuals.

Separate analysis finds that $\widehat{\rho}_{x_k} = 0.237$ for variable *Wins* and $\widehat{\rho}_{\widehat{u}} = 0.831$ for the OLS residual, while $obs/n = 286/29$. The formula then yields an approximate variance inflation factor of 1.66, compared to 1.86 actually found.

Table 21.3: NBA Team Revenue: Pooled OLS with various models and cluster-robust t-statistics.

| **Variable** | **OLS coefficients and t-statistics** | | | |
|---|---|---|---|---|
| *Wins* | .0068 | .0068 | .0068 | .0049 |
| | (3.59) | (3.58) | (3.53) | (3.32) |
| *Season* | .0182 | .0355 | | .0180 |
| | (5.51) | (3.16) | | (5.45) |
| *Season^2* | | -.0016 | | |
| | | (1.60) | | |
| *Playoff* | | | | .0306 |
| | | | | (0.68) |
| *Champion* | | | | .1089 |
| | | | | (2.17) |
| *Allstars* | | | | .0353 |
| | | | | (1.99) |
| *Lncitypop* | | | | .1440 |
| | | | | (2.41) |
| *Intercept* | 4.1516 | 4.1179 | 4.1656 | 3.9945 |
| | (42.99) | (41.47) | (43.33) | (67.04) |
| Time Dummies | No | No | Yes | No |
| Observations | 286 | 286 | 286 | 286 |
| $R^2$ | 0.176 | 0.178 | 0.180 | 0.456 |
| Adjusted $R^2$ | 0.170 | 0.170 | 0.150 | 0.445 |

### 21.2.4 Additional Control Variables

Table 21.3 presents OLS estimates for a range of models, along with cluster-robust standard errors.

The first three models are different models for the trend – linear as in Table 21.2, quadratic and time dummy variables. For the second model the quadratic term in the trend has a negative coefficient, indicating that the rate of increase in team revenues each season is decreasing over time, though the term is statistically insignificant at level 0.05. The third model instead adds a time dummy for each year, with the first season dummy omitted to avoid the dummy variable trap. The dummy variables for seasons two through ten are, respectively, -.00, .04, .08, .09, .12, .12, .13, .13, and .15. As for the quadratic trend, team revenues are increasing over time, but at a decreasing rate. The season dummies are jointly statistically significant with an $F_{9,28}$ distributed statistic equal to 5.14 and a p-value of .0004. At the same time, however, the adjusted $R^2$ is substantially lower than that for the linear or quadratic trends, and the coefficient of *Wins* is unchanged in the first two significant digits. We conclude that a linear time trend as in the first model is sufficient.

The fourth model adds some additional regressors. Being in the playoffs in the previous

season is not statistically significant. Note that the regressor variable *Wins* already includes wins in playoff games. Being the champion in the previous year has a statistically significant effect that is large in magnitude, with a 13% increase in revenues. Few players are selected as NBA All-Stars – only 24 players per season. Each All-Star player on a team in the current season is associated with a 3.6% increase in team revenues. As expected team revenues are higher in larger. The coefficient is an elasticity in this log-log specification, so a 1% increase in city population is associated with a .14% increase in team revenues.

## 21.3 Random Effects Estimation

The preceding analysis used OLS estimation, with the only panel adjustment being to use cluster-robust standard errors. We now consider alternative estimators that are used with panel data.

Pooled OLS is not the most efficient estimator as the model errors in a short panel are clustered, rather than independent and homoskedastic. Feasible GLS estimation based on a more realistic model for the model errors should lead to more precise estimation. The most commonly-used FGLS estimator is the random effects estimator.

### 21.3.1 Random Effects Model

The regression model is the same as for pooled OLS, with

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \cdots + \beta_k x_{kit} + u_{it}.$$

The **random effects model** or **error-components model** decomposes the model error $u_{it}$ into two components, one time-invariant and one time-varying, so

$$u_{it} = \alpha_i + \varepsilon_{it},$$

where $\alpha_i \sim (0, \sigma_\alpha^2)$ is an **individual-specific error** that is constant over all time periods for a given individual, and $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$ is called an **idiosyncratic error** that varies over both time and individuals.

These assumptions can be shown to imply that the combined errors $u_{it} = \alpha_i + \varepsilon_{it}$ are homoskedastic and independent across individuals. For a given individual, the errors in different periods are **equicorrelated**. This means that the correlation of $u_{it}$ and $u_{is}$ is restricted to be the same for all $s \neq t$.

**Remark 191** *The random effects model or error components model decomposes the model error into an individual-specific component that is time-invariant and an idiosyncratic component that varies over time and individuals. It restricts the within individual errors to be equicorrelated.*

### 21.3.2 Random Effects Estimator

For the random effects model pooled OLS remains a consistent estimator, but it is not as efficient as feasible generalized least squares. In the current setting the feasible GLS estimator is the **random effects estimator** which can be obtained from OLS estimation in the transformed model

$$y_{it} - \widehat{\lambda}_i \bar{y}_i = (1 - \widehat{\lambda}_i)\beta_1 + (x_{2it} - \widehat{\lambda}_i \bar{x}_{2i})'\beta_2 + \cdots + (x_{kit} - \widehat{\lambda}_i \bar{x}_{ki})'\beta_k + u^*_{it},$$

where $\lambda_i = 1 - \sigma_\varepsilon / \sqrt{T_i \sigma^2_\alpha + \sigma^2_\varepsilon}$ and $\widehat{\lambda}_i$ replaces $\sigma^2_\alpha$ and $\sigma^2_\varepsilon$ by consistent estimates. Here $T_i$ is the number of time periods for which data are available for the $i^{th}$ individual. In a balanced panel $T_i = T$ for all $i$.

The motivation for this estimator is that it can be shown, with considerable difficulty, that the transformed error $u^*_{it} = u_{it} - \lambda_i \bar{u}_i$ is independent and identically distributed over both $i$ and $t$. So OLS is efficient in a model where every variable, and hence the error, is transformed by subtracting $\lambda_i$ times its individual-specific mean. For example, $x_{2it}$ is transformed to $x_{2it} - \lambda_i \bar{x}_{2i}$.

Computer packages with panel commands or with a random effects command can be used to obtain the random effects estimator, without need to manually perform the regression. There are several methods for consistently estimating $\sigma^2_\alpha$ and $\sigma^2_\varepsilon$. In finite samples these different methods lead to somewhat different numerical estimates of $\sigma^2_\alpha$ and $\sigma^2_\varepsilon$, and hence of $\lambda_{it}$ and $\beta_1, ..., \beta_k$. So different statistical packages can give somewhat different random effects estimates.

### 21.3.3 Cluster-Robust Standard Errors

When the random effects estimator is applied to panel data, inference should always be based on cluster-robust standard errors, with clustering on the individual. This is valid provided there is independence across individuals and the number of individuals is large.

The reason for doing so is the following. The random effects model makes assumptions that are too strong in practice. In particular, the correlation between $u_{it}$ and $u_{is}$ is likely to decrease as the time difference $|t - s|$ increases rather then be constant as implied by the random effects model. While the random effects estimator can then no longer be assumed to be fully efficient, it is still likely to be more efficient than OLS which is based on the even stronger model that errors are independent over time as well as over individuals.

**Remark 192** *Random effects estimation is likely to be more efficient than pooled OLS, though one should still use cluster-robust standard errors.*

Table 21.4: NBA Team Revenue: Pooled OLS with various models and cluster-robust standard errors.

| **Variable** | **Estimator, coefficients and standard errors** | | | | | |
|---|---|---|---|---|---|---|
| | **Pooled OLS** | | **Random Effects** | | **Fixed Effects** | |
| | **Het-robust** | **Robust** | **Default** | **Robust** | **Default** | **Robust** |
| *Wins* | .0049 | .0049 | .0024 | .0024 | .0027 | .0027 |
| | (.0014) | (.0015) | (.0008) | (.0008) | (.0007) | (.0007) |
| *Season* | .0180 | .0180 | .0188 | .0188 | .0200 | .0200 |
| | (.0035) | (.0033) | (.0017) | (.0033) | (.0017) | (.0029) |
| *Playoff* | .0306 | .0306 | .0385 | .0385 | .0362 | .0362 |
| | (.0359) | (.0447) | (.0176) | (.0200) | (.0167) | (.0209) |
| *Champion* | .1089 | .1089 | .0118 | .0118 | .0052 | .0052 |
| | (.0331) | (.0473) | (.0316) | (.0163) | (.0300) | (.0167) |
| *Allstars* | .0353 | .0353 | .0372 | .0372 | .0356 | .0356 |
| | (.0127) | (.0178) | (.0075) | (.0066) | (.0071) | (.0068) |
| *Lncitypop* | .1440 | .1440 | .0196 | .0196 | -.2021 | -.2021 |
| | (.0196) | (.0598) | (.0315) | (.0872) | (.0491) | (.0632) |
| *Intercept* | 3.9945 | 3.9945 | 4.2477 | 4.2477 | 4.5222 | 4.5222 |
| | (.0491) | (.0596) | (.0560) | (.1076) | (.0649) | (.0957) |
| Observations | 286 | 286 | 286 | 286 | 286 | 286 |

### 21.3.4 Example: NBA Team Revenue

In this section we estimate the same model as that estimated in the final column of Table 21.3, one with a linear time trend and several regressors. Three estimators are given – pooled OLS, random effects and (discussed later) fixed effects. For each estimator two standard errors are given – default and cluster-robust with clustering on the individual.

Columns 3 and 4 give the random effects estimates. For this application $\widehat{\lambda}_i = .83$ for the Bobcats, an expansion team for which only six seasons of data are available, and $\widehat{\lambda}_i = .87$ for all other teams. The coefficient for variable *Wins* falls from .0068 for pooled OLS to .0046 for the random effects estimator. The coefficients of *Champion* and *Lncitypop* are now much smaller and statistically insignificant.

Inference should be based on cluster-robust standard errors, given in column 4. In this example they are sometimes larger and sometimes smaller than the default standard errors, given in column 3, that assume $\alpha_i$ and $\varepsilon_{it}$ are homoskedastic and independent.

The random effects estimates are much more precise than the pooled OLS estimates in this example. The cluster-robust standard errors for variables *Wins*, *Playoff*, *Champion* and *Allstars* given in column 4 are generally less than half those given in column 2 for pooled OLS.

### 21.3.5 Alternative FGLS Estimators

The random effects estimator is the most commonly-used feasible GLS estimator for short panel data. Other feasible GLS estimators exist, based on alternative assumptions about the model error $u_{it}$.

In particular, many studies assume that the error term follows that error process $u_{it} = \rho u_{it-1} + \varepsilon_{it}$, where $|\rho| < 1$ and $\varepsilon_{it}$ is a zero-mean homoskedastic independent error. Thus for a given individual the error this period is a multiple of last period's error plus an independent shock. This model implies that $u_{it}$ and $u_{is}$ have correlation $\rho^{|t-s|}$ which decreases as the time separation $|t-s|$ increases, unlike the random effects model which keeps this error correlation constant.

Clearly a wide range of models is possible. Advanced econometrics and statistical packages provide efficient estimates for these models. In short panels one should again guard against possible misspecification of the model error process by using cluster-robust standard errors.

## 21.4 Fixed Effects Estimation

The preceding regression results indicate that more wins are associated with higher team revenue. One possible reason for this is that there are teams with a high "X-factor" that season after season both win and have relatively high revenue, while there are other teams without the "X-factor" that both lose a lot and have relatively low revenue. Wins and revenue are associated but the relationship is noncausal. Instead the result is driven by a third variable.

A second explanation is that regardless of whether a team is strong or weak, increasing the number of wins leads to higher revenue. This second explanation is a causal explanation.

If the "X-factor" depends on observable measures, such as size of the team's market, then this can be controlled for by inclusion of these variables as regressors. But these variables may not be known or may not be observable. Fixed effects estimation enables controlling for the "X-factor" without needing to include the variables that determine the "X-factor". The resulting estimates can be given a causal interpretation.

### 21.4.1 Fixed Effects Model

The **fixed effects model** specifies the intercept to vary across individuals while the slope coefficients are the same for each individual. So

$$y_{it} = \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}.$$

Note that as we will explicitly estimate or control for the $n$ individual intercepts $\alpha_1, ..., \alpha_n$, the overall intercept $\beta_1$ is dropped to avoid the dummy variable trap.

This model superficially looks like the random effects model, as both have an **individual specific effect** $\alpha_i$. The difference is that the random effects model treats $\alpha_i$ as purely random, and hence unrelated to the regressors, whereas the fixed effect $\alpha_i$ can potentially be correlated with the regressors. While both models have an unobserved "X-factor" ($\alpha_i$) that is constant over time, the "X-factor" in the random effects is assumed to be unrelated to observables, whereas in the fixed effects model the "X-factor" may be correlated with other regressors in the model such as wins and measures of market size.

If regressors are correlated with $\alpha_i$ then both pooled OLS and random effects estimators are inconsistent, while the fixed effects estimator is consistent.

**Remark 193** *The fixed effects model decomposes the model error into a time-invariant individual-specific component that may be correlated with the regressors and an idiosyncratic component that is assumed to be uncorrelated with the regressors.*

### 21.4.2 Fixed Effects Estimator

There are as many fixed effects $\alpha_1, ..., \alpha_n$ as there are individuals in the sample. Estimation is greatly simplified by transforming to a model that eliminates the many fixed effects and leaves only $\beta_2, ..., \beta_k$ to be estimated.

Averaging the fixed effects model over time for each individual yields

$$\bar{y}_i = \beta_2 \bar{x}_{2i} + \cdots + \beta_k \bar{x} + \alpha_i + \bar{\varepsilon}_i,$$

since $\bar{\alpha} = \frac{1}{T}\sum_{t=1}^{T} \alpha_i = \alpha_i$. Subtracting this averaged model from the original fixed effects model eliminates $\alpha_i$, since $\alpha_i - \alpha_i = 0$. The resulting model is called the **within model** or **mean-differenced model**

$$y_{it} - \bar{y}_i = \beta_2(x_{2it} - \bar{x}_{2i}) + \cdots + \beta_k(x_{kit} - \bar{x}_{ki}) + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

The **fixed effects estimator** or **within estimator** of $\beta_2, ..., \beta_k$ is the OLS estimator in this model. The estimator is consistent even if regressors are correlated with $\alpha_i$, since the error $\varepsilon_{it} - \bar{\varepsilon}_i$ in the within model does not involve $\alpha_i$, and we assume that $\varepsilon_{it}$ is uncorrelated with the regressors.

To understand this estimator, consider bivariate regression of *Revenue* on *Wins*. The fixed effects slope coefficient estimate will be positive if seasons with above-average wins, that is seasons in which a team has more wins than its ten-year average number of wins, are associated with above-average revenue in the following season.

The fixed effects estimator uses only within variation of the data. This implies that it is not possible to estimate the coefficients of variables that take the same value in all time periods, as they have no within variation. For example, if we include as a regressor the number of championships a team has won before the 2001-02 season, the first year of

these data, then while this coefficient can be estimated by pooled OLS and random effects, it cannot be estimated by the fixed effects estimator. And while estimation is possible for variables that have relatively little within variation, the estimation may be very imprecise.

The transformation for random effects estimation subtracts $\widehat{\lambda}_i$ times individual means, while for fixed effects the individual mean is subtracted. So the random effects estimator will be close to the fixed effects estimator if $\widehat{\lambda}_i$ is close to one.

**Remark 194** *The fixed effects estimator or within estimator is obtained by OLS estimation in the mean-differenced model that eliminates the individual-specific effect. For a short panel inference is based on cluster-robust standard errors.*

### 21.4.3 Least Squares Dummy Variable Estimator

An alternative way to estimate the fixed effects model is to directly estimate the individual intercepts as well as the common slope coefficients. Define $n$ dummy variables $d1_{it}, d2_{it}, ..., dn_{it}$, where $d1_{it}$ equals one if $i = 1$ and equals zero otherwise, $d2_{it}$ equals one if $i = 2$ and equals zero otherwise, and so on. The **least squares dummy variable (LSDV) model** is

$$y_{it} = \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_1 d1_{it} + \cdots + \alpha_n dn_{it} + \varepsilon_{it}.$$

Note that for the $i^{th}$ individual $di_{it} = 1$ in all time periods while all the other indicator variables equal zero, so the model implies

$$y_{it} = \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}.$$

The **LSDV estimator** is the OLS estimator in the LSDV model. In principle it is not possible in this model to obtain consistent estimates, in a short panel with small $T$, as there are too many parameters to estimate. There are $n + k - 1$ parameters to estimate using $nT$ observations. Increasing $n$ is self defeating as the benefit of increased data is negated by the need to estimate more intercept parameters. It turns out, however, that while the intercepts $\alpha_1, ..., \alpha_n$ cannot be consistently estimated, with small $T$, the slopes $\beta_2, ..., \beta_k$ can be consistently estimated, assuming large $n$.

Just as remarkably it turns out that the LSDV estimates of $\beta_2, ..., \beta_k$ equal the fixed effects estimates. The LSDV estimate of the intercept for the $j^{th}$ individual is $\widehat{\alpha}_j = \bar{y}_j - \widehat{\beta}_2 \bar{x}_{2j} + \cdots + \widehat{\beta}_k \bar{x}_{kj}$.

**Remark 195** *The least squares dummy variable (LSDV) model includes an indicator variable for each individual in the sample. The LSDV estimator and the fixed effects estimator are equivalent.*

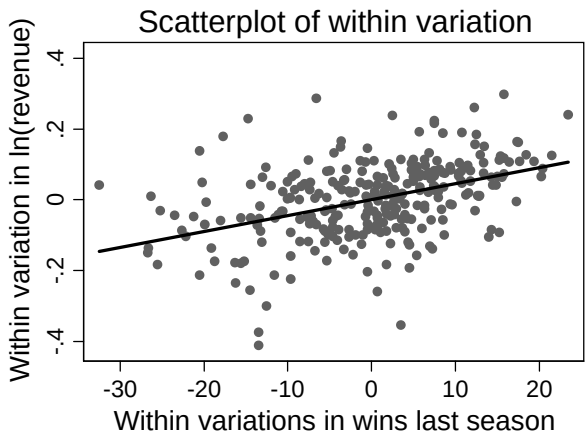


Figure 21.3: Within variation in natural logarithm of team earnings plotted against wintin variation in number of wins in previous season for 29 teams over 10 seasons.

### 21.4.4 Cluster-Robust Standard Errors

The LSDV estimator uses OLS so it is fully efficient if the error $\varepsilon_{it}$ is zero-mean homoskedastic and independent. Since fixed effects estimation coincides with LSDV estimation, it follows that the fixed effects estimator is also fully efficient if $\varepsilon_{it}$ is zero-mean homoskedastic and independent.

In practice model errors are correlated over time for a given individual even with inclusion of the fixed effect $\alpha_i$ (though $\alpha_i$ can reduce the extent of this correlation). And, additionally, errors may be heteroskedastic. So cluster-robust standard errors should always be used after fixed effects estimation.

### 21.4.5 Example: NBA Team Revenue

The fixed effects estimator uses only within variation in the data. For bivariate regression the within variation can be viewed using a scatter plot of the data.

Figure 21.3 presents a plot of $y_{it} - \bar{y}_i$ against $x_{it} - \bar{x}_i$ where $y$ is the natural logarithm of team revenue and $x$ is the number of wins. There is considerable within variation for both variables, and there is a positive relationship with correlation coefficient of 0.43.

The last two columns of Table 21.4 present the results of fixed effects estimation. The estimated coefficients are very close to the random effects estimates (the coefficient of variable *Lncitypop* changes sign but is highly statistically insignificant). The small difference is because the random effects multiplier $\widehat{\lambda}_i = .87$ (or .83 for the Bobcats) is quite close to one. In other applications the difference can be much greater. The table includes an intercept

estimate as the results are from regression that adds back in the grand mean, so $y_{it} - \bar{y}_i + \bar{y}$ is regressed on an intercept and $x_{2it} - \bar{x}_{2i} + \bar{x}_2, ..., x_{kit} - \bar{x}_{ki} + \bar{x}_k$.

Inference should be based on cluster-robust standard errors, given in column 4. In this example they are sometimes larger and sometimes smaller than the default standard errors, given in column 3, that assume $\alpha_i$ and $\varepsilon_{it}$ are homoskedastic and independent.

The results presented in Table 21.4 were obtained using software with a fixed effects estimator command. Exactly the same slope coefficient estimates are obtained by OLS estimation of the within model or by OLS regression of the LSDV model. The intercept, however, can differ with the model estimated and the particular panel software used. The intercept equals zero after OLS estimation of the within model, and after OLS estimation of the LSDV model if all $n$ dummies are includes as regressors and a no-intercept option is used in estimation. As explained above the results in Table 21.4 have a nonzero intercept as the grand means of $y$ and of the regressors are added back in to the within model.

Computation of the standard errors is more complicated if one does not have access to a fixed effects estimator command. For default standard errors one should use those from OLS estimation of the LSDV model, as the degrees of freedom used in OLS estimation of the within model fail to include allowance for the $n$ intercepts that are differenced out. For robust standard errors, one should instead use those from the OLS estimation of the within model.

### 21.4.6 When to Use Fixed Effects

The fixed effects model is used when there is concern that regressors may be correlated with the model error, but this correlation is only with a time-invariant component of the error. Specifically, the regressors may be correlated with the model error $\alpha_i + \varepsilon_{it}$, but only with $\alpha_i$ and not with $\varepsilon_{it}$. Under these assumptions the pooled OLS and random effects estimators are inconsistent, whereas the fixed effects estimator is consistent because it eliminates $\alpha_i$ by mean-differencing.

In the current example the time-invariant component $\alpha_i$ of the error was called an "X-factor." In a panel data regression of earnings on schooling, or some other measure of training that changes over time, the unobserved component $\alpha_i$ could be interpreted as unobserved ability.

Given these assumptions the fixed effects estimates can be given a casual interpretation. From Table 21.4 the estimate of .0027 means that one more win leads to a .27% increase in team revenue.

Economists often use the fixed effects estimator as they seek to obtain estimates that can be given a causal interpretation. In other areas of applied statistics the random effects estimator is used much more. In settings with experimental data this is fine as there is no issue regarding causality, while with observational data not all disciplines are as concerned with causal inference as are economists.

With observational data it is safest to always use the fixed effects estimator. But there is a cost to doing so. There can be a great loss in precision as the fixed effects estimator uses only within variation of the data, whereas the pooled OLS and random effects estimators use both within and between variation. And in some extreme cases a key regressor may have no within variation at all. Finally, the assumption that regressors are correlated only with $\alpha_i$ and additionally with $\varepsilon_{it}$ is a nontestable assumption.

### 21.4.7 Testing for Fixed Effects

Due to the loss of efficiency if fixed effects estimation it useful to test whether fixed effects estimation is necessary.

If regressors are uncorrelated with the model error then the random effects and fixed effects estimators are both consistent so they should be similar in magnitude. A substantial difference between the two is interpreted as a need to perform fixed effects estimation (though it could be due to a misspecified model).

This test is called a Hausman test. Statistical packages that include this test generally use a version of the test that assumes that the error $\varepsilon_{it}$ in the fixed effects model is homoskedastic and independent. In practice, however, $\varepsilon_{it}$ is correlated over time for a given individual. Failure to make this adjustment means that the commonly-used version of the Hausman test tends to over-reject, so that in some cases it suggests that fixed effects estimation is necessary when this is not the case.

Instead the following cluster-robust version of the Hausman test should be used. Estimate by OLS the original pooled model that is supplemented by the individual mean of the regressors, so

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \pi_2 \bar{x}_{2i} + \cdots + \pi_k \bar{x}_{ki} + \varepsilon_{it},$$

where averages for regressors that have no within variation are necessarily omitted. An overall test of whether fixed effects estimation is necessary is a test of whether the additional regressors are jointly statistically, i.e., test $H_0 : \pi_2 = 0, ..., \pi_k = 0$. Often studies simply test whether just the coefficient of the key regressor of interest is statistically significant. For example, test $H_0 : \pi_j = 0$ if the $j^{th}$ regressor is of interest. In both cases the test is based on cluster-robust standard errors.

For the current application the individual coefficients for mean (over time for each team) of all regressors except *Lncitypop* are statistically insignificant at level 0.05, suggesting that it is not necessary to use the fixed effects estimator. At the same time, the joint test that all six coefficients are zero yields an $F_{6,28}$ statistic of 12.18 with $p = .0000$, so the two coefficients are jointly statistically significant, indicating a need for fixed effects estimation.

**Remark 196** *The need to estimate by fixed effects can be assessed using a cluster-robust version of the Hausman test.*

## 21.5 Long Panels

The discussion to date has focused on panel data from short panels with many independent individuals. The methods used are often panel extensions of methods for cross-section data.

An alternative form of panel data can be time series data on a number of macroeconomic units such as countries. An example is the Penn World Tables. In a long panel there are many time periods but potentially data on few individuals. This section briefly presents some of the issues that arise in analyzing such data.

**Remark 197** *A **long panel** is one for which there are many time periods. A long panel can have as few as two individuals.*

### 21.5.1 Heterogeneous Models

With many time periods it is possible to specify and estimate separate models, or **heterogeneous models**, for each individual. Thus we can separately estimate for each individual $i$ the model

$$y_{it} = \beta_{1i} + \beta_{2i}x_{2it} + \cdots + \beta_{ki}x_{kit} + u_{it}.$$

Then the time series methods of Chapter 20 are directly applicable. For example, if the regressors do not include lagged dependent variables and there are no unit root complications, we may estimate by OLS and obtain heteroskedastic and autocorrelation consistent (HAC) standard errors that control for correlation in $u_{it}$ over time. Or we may specify a model for the error, such as $u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it}$ with $\varepsilon_{it} \sim (0, \sigma_i^2)$, and estimate by feasible GLS.

### 21.5.2 Pooled Models

With many individuals there will be many $(n \times k)$ parameter estimates. It is simpler to restrict some or all parameters to be the same for different individuals, and such restrictions will increase estimator precision.

In the simplest case all parameters are the same across individuals, leading to the **pooled model**

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}.$$

In addition to simplicity this should lead to more precise parameter estimates, since there are fewer parameters to estimate.

Inference is more complicated, however. The error $u_{it}$ is likely to be correlated over time for a given individual and heteroskedastic. HAC standard errors cannot be used as the data are no longer a pure time series once there is more than one individual. Cluster-robust standard errors can be used assuming independence across individuals, but only if there are many individuals in the long panel. Instead a model for the errors needs to be specified. For

example, we might assume that $u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it}$ with $\varepsilon_{it} \sim (0, \sigma_i^2)$, or more simply that $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$ with $\varepsilon_{it} \sim (0, \sigma^2)$. Inference is then dependent on correct specification of the model for the errors.

### 21.5.3 Additional Complications

In many long panel applications it may not be reasonable to assume that different individuals are uncorrelated. For example, the error for one country may be correlated with that for nearby countries. Or countries economies may be subject to a common global shock. And in some settings the data may be nonstationary, so that the time series complications of unit roots and cointegration become relevant.

Analysis for short panels is relatively straightforward as the assumption of many independent individuals makes it easy to control for time series correlation of the errors. With long panels and few individuals this is no longer the case, and a range of methods exist depending on how one models the errors.

**Remark 198** *Long panels introduce additional complications and are generally not as straightforward as short panels.*

## 21.6 Key Concepts

1. A balanced panel is one for which data are observed for all individuals in all time periods.

2. A short panel is one for which there are many individuals and model errors are uncorrelated across individuals. A short panel can have as few as two observations per individual.

3. The overall variation in the data can be decomposed into variation over time for a given individual, called within variation, and variation across teams, called between variation.

4. Standard OLS estimation with panel data is called pooled OLS as data across individuals and over time are pooled or combined. For a short panel inference is based on cluster-robust standard errors.

5. The random effects model or error components model decomposes the model error into an individual-specific component that is time-invariant and an idiosyncratic component that varies over time and individuals. It restricts the within individual errors to be equicorrelated.

6. Random effects estimation is likely to be more efficient than pooled OLS, though one should still use cluster-robust standard errors.

7. The fixed effects model decomposes the model error into a time-invariant individual-specific component that may be correlated with the regressors and an idiosyncratic component that is assumed to be uncorrelated with the regressors.

8. The fixed effects estimator or within estimator is obtained by OLS estimation in the mean-differenced model that eliminates the individual-specific effect. For a short panel inference is based on cluster-robust standard errors.

9. The least squares dummy variable (LSDV) model includes an indicator variable for each individual in the sample. The LSDV estimator and the fixed effects estimator are equivalent.

10. The fixed effects estimator allows causal inference, assuming regressors are correlated only with a time-invariant component of the model error.

11. The need to estimate by fixed effects can be assessed using a cluster-robust version of the Hausman test.

12. A long panel is one for which there are many time periods. A long panel can have as few as two individuals.

13. Long panels introduce additional complications and are generally not as straightforward as short panels.

14. Key terms: balanced panel; long panel; short panel; grand mean; individual mean; overall variation; within variation; between variation; pooled OLS; cluster robust; panel robust; random effects; error components; individual-specific error; idiosyncratic error; equicorrelation; fixed effects; individual-specific effect; random effects; heterogeneous model.

## 21.7 Additional Models and Estimators

This section presents additional estimators for short panels that are less often used by economists, and briefly discusses nonlinear models for panel data.

### 21.7.1 Between Estimator

The between estimator is the OLS estimator of the between model obtained by averaging the data over all time periods for each individual. Thus we estimate by OLS the model

$$\bar{y}_i = \beta_1 + \beta_2 \bar{x}_{2i} + \cdots + \beta_k \bar{x} + \bar{u}_i,$$

where there are now $n$ observations rather than $nT$ observations.

The between estimator is less efficient than the pooled OLS estimator as it uses only between variation in the data, whereas OLS uses both within and between variation. Like the pooled OLS and random effects estimator, the between estimator is inconsistent if fixed effects are present.

### 21.7.2 First Difference Estimator

Consider the fixed effects model, with error $u_{it} = \alpha_i + \varepsilon_{it}$. Lagging the model by one period and subtracting leads to the first difference model

$$\Delta y_{it} = \beta_2 \Delta x_{2it} + \cdots + \beta_k \Delta x_{kit} + \Delta \varepsilon_{it},$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and $\alpha_i$ drops out since $\Delta \alpha_i = \alpha_i - \alpha_i = 0$.

The first difference estimator is the OLS estimator in this model. Since the fixed effects $\alpha_i$ have dropped out this estimator is consistent in the fixed effects model. If there are exactly two periods of data for every individual the estimator is the same as the fixed effects estimator. Otherwise the two estimators differ, and the fixed effects estimator is preferred as it is generally more efficient than the first difference estimator.

### 21.7.3 Dynamic Models

The short panel models have implicitly been static models, with current period $y_{it}$ depending on current period $x_{it}$. There is no problem in including lags of $x_{it}$ as regressors. If lags of the dependent variable are included as a regressor then sufficient lags need to be included to ensure that the error $u_{it}$ is not correlated over time, an additional restriction not needed in the discussion to date. Otherwise alternative estimators, notably instrumental variables estimators, need to be used.

To see the complication of a lagged dependent variable as a regressor, consider the model

$$y_{it} = \beta_1 + \beta_2 y_{i,t-1} + \beta_3 x_{it} + u_{it}.$$

Recall that the OLS estimator is consistent if the error $u_{it}$ has zero mean conditional on regressors. In the current model, if $u_{it}$ is correlated with $u_{i,t-1}$, then the regressor $y_{i,t-1}$ is correlated with $u_{it}$ since $y_{i.t-1} = \beta_1 + \beta_2 y_{i,t-2} + \beta_3 x_{i,t-1} + u_{i,t-1}$ depends on $u_{i,t-1}$ which

we assumed to be correlated with $u_{it}$. So OLS is inconsistent in this model if $u_{it}$ is serially correlated.

Furthermore, even if sufficient lags are added to ensure $u_{it}$ is uncorrelated over time, if fixed effects are present the fixed effects estimator is no longer consistent. Instead a more advanced estimator, the Arellano-Bond estimator, needs to be used.

### 21.7.4 Nonlinear Models

For simplicity consider short panel versions of the logit and probit models introduced in Chapter 19.4.

The simplest estimators are pooled logit and pooled probit. Simply perform usual logit or probit estimation, but compute standard errors by clustering on the individual, to control for correlation over time for each individual. In the biostatistics literature pooled estimates are called population-averaged estimates.

Random effects versions of the logit and probit model exist. Consider the probit case and a single regressor for simplicity. Then the random effects probit model specifies

$$\Pr[y_{it} = 1|x_{it}, \alpha_i] = \Phi(\beta_1 + \beta_2 x_{it} + \alpha_i),$$

where, as before, $\Phi(\cdot)$ is the standard normal cumulative distribution function. The random effect $\alpha_i$ is assumed to be normally distributed with mean 0 and variance $\sigma_\alpha^2$. Note that compared to the linear model the stronger assumption of a normal distribution is needed. Furthermore, consistency of the random effects model requires that this assumption is correct. Introducing the random effects leads to a rescaling of the regression parameters $\beta_1$ and $\beta_2$, so they are not comparable with those obtained from pooled probit estimation. Instead marginal effects should be compared.

A fixed effects probit model treats $\alpha_i$ as being potentially correlated with the regressor $x_{it}$. There is no way to consistently estimate this model if the panel is a short panel. For example, suppose we had only two periods of data and a balanced panel of $n$ individuals. Then there are $2n$ observations and $n+1$ parameters $(\beta_2, \alpha_1, ..., \alpha_n)$ to estimate ($\beta_1$ is dropped to avoid the dummy variable trap). There is no way to consistently estimate all these parameters, unless there are many observations for each individual. For the linear model it was possible to eliminate $\alpha_1, ..., \alpha_n$ by mean differencing, leaving only $\beta_2$ to estimate. There is no similar method for the probit model.

# Appendix A

# © A. Colin Cameron: Computing

The best way to get started in any statistical package is hands-on use. Open the statistical package, open a dataset, some are often included in the statistical package, and use the pull-down menus.

This chapter begins with general issues that arise regardless of the statistical package used. It then introduces several statistical packages often used by economists and economics students: the commercial packages Stata and Eviews and the free packages Gretl and R. Additionally the spreadsheet Excel is considered.

For lengthier and more advanced analysis it is best to use commands collected into a script or program file. Examples are provided here for each of the statistical packages. The book website provides scripts that reproduce the results for the entire book.

## A.1 General Issues

This section addresses the following general issues that arise regardless of what statistical package is used.

- How to obtain help.
- How to input commands.
- How to stop command execution.
- How to use a working directory.
- How to read in data.
- How to check that the data are correctly read in.
- How to save results.

- How to incorporate these results into a word processor.

### A.1.1 Help

Statistical packages generally have built-in help that provides an overview of commands and a short description of specific commands. This can be obtained through the help menu. For some packages help on a specific command can also be obtained by directly typing in the command window.

Reference manuals for statistical packages can generally be downloaded as free pdf files through the help menu. And statistical package websites usually provide tutorials, often available as powerpoints or short video clips.

### A.1.2 Inputting Commands

Statistical packages have a **windowing interface** or **graphical user interface** (GUI) that enables inputting commands by using **drop-down menus**. For example, to obtain descriptive statistics in Stata, choose the menu for Statistics, the sub-menu for Summary, Tables, and Tests, the subsequent sub-menu Summary and Descriptive Statistics, the further sub-menu Summary Statistics, and within that choose the relevant variable(s), the range of observations, and additional options such as whether to display additional statistics beyond those reported by default. This is the best way to begin using a package.

Once you have familiarity with the package, however, it is often quicker to directly type in the relevant command in the statistical package's **command window**. In Stata, for example, the command `summarize Price` gives descriptive statistics for the variable *Price*. For highly customized graphs, however, it may be easier to use the windowing interface.

For analysis requiring many commands, or lengthier commands, it is best to collect all the commands into a **script**, also called a **program**, that is then stored as a plain text file. Then execute or run the entire script. Double-clicking on the script file, provided your computer recognizes that the extension used for your script file is associated with your statistical package, opens the package's editor and in some cases automatically executes the script. Alternatively enter the statistical package and read in the script from the File menu.

A script provides a permanent record of how results were obtained, and makes it easy to modify and redo the analysis if necessary. Placing comments in the script can greatly improve its readability. Script files can be created using a text editor or, for some packages, using an editor window within the package. In the latter case one can run selected portions of the script by highlighting the desired part before execution.

Script files are given different file-name extensions by different packages. These are listed in Table A.1. An example script for each specific statistical package is presented in the relevant section.

Table A.1: Statistical packages: File-name extensions for script files and data files.

| Package | Script file | Data file | Results also in Data File? |
|---|---|---|---|
| Stata | .do or .ado | .dta | No |
| Eviews | .prg | .wf1 | Yes |
| Gretl | .inp | .gdt and .gretl | Yes for .gretl |
| R | .R of .r | .Rdata or .rda | Yes |
| Excel | – | .xls or .xlsx | Yes |

### A.1.3 Halting Commands

In some cases a command takes a long time to execute or produces so much output that you wish to break execution of the command, without having to exit the statistical package.

The command script window for most statistical packages includes a break or stop button to stop execution of the script. Additionally there may be a hot key that also stops execution. For example, in R one can use the escape key.

### A.1.4 Working Directory

It is simplest to keep datasets, script files and result files in the same directory, called a **working directory**. This avoids the need to use lengthier filenames that include the complete directory path for the file.

If the statistical package is opened directly from the working directory, for example by double-licking a script file or dataset that is in the working directory, and your computer associates the file extension with the appropriate statistical package, then the statistical package may automatically open in the desired working directory.

In other cases, after opening your statistical package you may need to use drop-down menus or type a command to change to the desired working directory.

### A.1.5 Reading in Data

It can be remarkably difficult to read in data as data come in many formats. A major distinction is between text files, sometimes called ASCII files, that can be read by a simple text editor such as Windows Notepad, and formatted files such as a Stata dataset or an Excel spreadsheet, that can only be read by more specialized software that can interpret that specific format.

For **text files** the standard format is to have all variables for one observation on the one line, and separate each variable by a character. It is most common to use a comma to separate the values, so such files are usually called **comma-separated values (csv)** files

and given the extension **.csv**. Other characters can be used as a separator, though if a space is used then care is needed if the data include text data with spaces, such as a variable for street address. The first line of the csv file may give the names for each of the variables.

An alternative form for a text file is a **fixed-width file** where, for example, the first variable is stored in columns 1-3, the second in columns 4-7, and so on. In that case one needs to tell the statistical package the formatting used, along with the corresponding variable names. A commonly-used extension for such files is **.txt**, though other extensions are also used.

**Formatted data files** are specific to each statistical package. They have the advantage over text files of being immediately readable by the statistical packages and potentially including additional information beyond the data, such as descriptions of each of the variables in the dataset and in some cases results. The standard file-name extensions used for data files produced by the different statistical packages are given in Table A.1.

Some of these package-specific files, notably those with extension **.wf1**, **.gretl**, **.rda** or **.Rdata** and **.xls** or **.xlsx** can additionally include the results from statistical analysis. Files with file-name extension **.dta** and **.gdt** contain only data.

The datasets for this book are provided in three formats. Files with **extension .DTA** are saved in Stata version 11 format (Stata 11 was introduced in 2008). These can be read in Stata (version 11 and higher), and can be imported into Eviews, Gretl and R. Note, however, that the latter three statistical packages will not necessarily be able to read Stata datasets saved in the most recent version of Stata. The .DTA files for this book usually include variable descriptions, called variable labels. The same datasets are also saved in Excel 1997-2003 format, with **extension .XLS**, and in a comma-separated values file, with **extension .CSV**.

### A.1.6 Checking Data

Once data are read in one should list some of the observations and obtain summary statistics to check that the data are successfully read in.

A common pitfall is dealing successfully with **missing values**. A code such as -999 or 999 may be used to denote a missing value. In that case care is needed to not misinterpret this as being the actual value of the variable. If a missing value is left as a blank, then the statistical package may mistakenly interpret this value as being a zero. And if an asterisk is used for missing values, then problems may arise as this is a character whereas the remaining data are numerical.

The datasets used in this book do not have these complications. They also generally have less than 1,000 observations so that student versions of statistical packages can be used.

### A.1.7 Saving Results

It is important to save results, along with the commands used to obtain these results.

Stata saves text results, though not graphs, in a results file, called a **log file**. For Stata create the log file as soon as you start analysis as only results from that point on are saved. For Eviews, Gretl, R and Excel both data and results can be saved in a special format file with filename extension, respectively, **.wf1**, **.gretl**, **.rda** and **.xls**. Results in these files can be cut and pasted into a word processor. For Eviews, Gretl and R it is also possible to send results to a log file.

Statistical packages often store graphics internally in a special format that can only be read by that package. To read the graph into a word processor graphs need to be saved in an appropriate graphics file format. Common **graphics file formats** are Windows Metafile (extension **.wmf**) and extended Windows Metafile (extension **.emf**) if working in a PC Windows environment, Portable Network Graphics (extension **.png**) and Encapsulated Postscript (extension **.eps**). The format to save in will vary with the word processor and the computer platform that you use. For instance, for Microsoft Word use .wmf, and for Latex use .eps or .png formats.

### A.1.8 Reporting Results in a Word Processor

Descriptive statistics and regression output are provided in a **monospace font** where each character takes up the same space, so that output table columns will be properly aligned. When such output is copied into a word processor or text editor one should therefore change to a monospace font, such as Courier New or Lucinda Console.

Output from a statistical package can be voluminous. For reports and assignment answers only the essential output should be included.

## A.2 Stata

**Stata** is a commercial statistical package that is used mostly in economics and related social sciences, especially political science and sociology, and in biostatistics. A student version is available that, at the time of writing, is restricted to 1,200 observations on up to 99 variables. The Stata website is `stata.com` and a helpful website is `http://www.ats.ucla.edu/stat/stata/`.

Example datasets can be obtained by selecting File > Example Datasets ... A common Stata manual dataset to use is file `auto.dta`. To read in your own data it is best to first use File > Change Working Directory ... to change to the directory that has the dataset.

It is easiest to learn Stata by initially using the Stata drop-down menus. The resulting output is preceded by the actual Stata command, however, so one can potentially learn and remember the direct commands for future use.

To save results, create a log file immediately upon opening Stata by choosing menu File > Log > Begin ... in the Results Window.

Basic Stata commands, such as those needed for this book, are relatively straightforward. Note that like many statistical packages, Stata is case sensitive. Commands can be directly typed in at the Command Window, a small window that is usually beneath the Results Window. Alternatively, a script can be stored in a file with file-name extension **.do**. The script file can be opened from outside Stata by double-clicking on the file, or within Stata by choosing menu File > Do... and selecting the desired file.

The base installation of Stata covers most methods in this book. In some rare cases, however, it is useful to additionally use user-written Stata programs. These can be downloaded from the internet and added permanently to your installation of Stata. For example, to download the `freduse` command mentioned in the introduction to Chapter 3, in Stata give command `findit freduse`, to locate the command and then hit the install button. Command `help freduse` then explains how to use the command.

An example Stata script, with many of the key commands needed for this book, is the following.

```
* aed_example.do   Stata example script written 4/1/2014
* Create a plain text log file that contains commands and results
log using Stataresults.txt, text replace
* Read in Stata dataset HOUSE.DTA; first clear any existing data in memory
clear
use HOUSE.DTA
* Example of help command - how to import datasets in different formats
help import
* Descriptive statistics for all variables
summarize
* List first five observations for all variables
list in 1/5
* Example of creating a new variable
generate PriceperSqfoot = Price/Size
* Two-way scatterplot with regression line saved in .png graphic format
scatter Price Size || lfit Price Size, saving(graphstata.gph, replace)
graph export graphstata.png, replace
* Descriptive statistics for two variables
summarize Price Size
* More detailed descriptive statistics for one variable
summarize Price, detail
* Confidence intervals for a single variable
mean Price
```

```
* T-test that population mean Price = 250000
ttest Price == 250000
* Multivariate OLS regression - same as chapter 13
regress Price Size Lotsize Bedrooms Bathrooms Age Monthsold
* Same regression with heteroskedastic-robust standard errors
regress Price Size Lotsize Bedrooms Bathrooms Age Monthsold, vce(robust)
* Save the Stata dataset
save Stataexample.dta, replace
* Close the log file
log close results.txt
* Exit Stata
exit
```

Lines preceded by an asterisk (*) and any text after two forward slashes (//) are interpreted as comments. Stata commands include options, given after a comma. For example, in the first `log` command the option `text` means that the output is saved as a plain text file and the option `replace` permits any previous version of the same file to be overwritten. Unlike most other packages, many Stata commands and options can be abbreviated. For example, `summarize` can be shortened to `sum` or even `su`. The script requires that file HOUSE.DTA is in the working directory, and not in some other directory. As already noted, graphs may be saved in several different graphics formats – the PNG format used here will not work in all word processors. The command `exit` results in automatically exiting Stata once the script is run.

Figure A.1 gives some of the output from Stata. This is the form that the output takes both on the screen and in the log file. Commands are preceded by a period, while results from the command are not.

The first set of results in Figure A.1 are descriptive statistics – the number of non-missing observations for each variable, the sample mean, the standard deviation, and the minimum and maximum sample values of the variable. More complete summary statistics, including the median, quartiles and skewness and kurtosis statistics, can be obtained by using option `detail` of the `summarize` command. For variable *Price* this yields skewness statistic 1.5608 and kurtosis statistic 5.6127.

The second set of results in Figure A.1 gives a 95 percent confidence interval for the population mean price that is (239688, 268133). It also gives the standard error of the sample mean, which is the standard deviation 37390.71 divided by $\sqrt{29}$.

The third set of results in Figure A.1 are regression results that are identical to those given in Table 13.4 and discussed in Chapters 13.7 and 14.5. Additionally the top left-hand corner of the regression output includes an analysis of variance table that gives the explained sum of squares, which Stata calls the model sum of squares, the residual sum of squares and the total sum of squares. Note that Stata automatically includes an intercept, and this

```
. * Descriptive statistics for two variables
. summarize Price Size

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       Price |         29    253910.3    37390.71     204000     375000
        Size |         29    1882.759    398.2721       1400       3300

.
. * Confidence intervals for a single variable
. mean Price

Mean estimation                     Number of obs    =      29

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       Price |   253910.3   6943.281      239687.7      268133
--------------------------------------------------------------

.
. * Multivariate OLS regression - same as Chapter 13
. regress Price Size Lotsize Bedrooms Bathrooms Age Monthsold

      Source |       SS       df       MS              Number of obs =      29
-------------+------------------------------           F(  6,    22) =    6.83
       Model |  2.5466e+10     6  4.2444e+09           Prob > F      =  0.0003
    Residual |  1.3679e+10    22   621790812           R-squared     =  0.6506
-------------+------------------------------           Adj R-squared =  0.5552
       Total |  3.9146e+10    28  1.3981e+09           Root MSE      =   24936

------------------------------------------------------------------------------
       Price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Size |   68.36942   15.38947     4.44   0.000     36.45361    100.2852
     Lotsize |   2303.221   7226.535     0.32   0.753     -12683.7    17290.14
    Bedrooms |   2685.315   9192.526     0.29   0.773    -16378.82    21749.45
   Bathrooms |    6832.88   15721.19     0.43   0.668    -25770.88    39436.64
         Age |  -833.0386   719.3345    -1.16   0.259    -2324.847    658.7699
   Monthsold |  -2088.504   3520.898    -0.59   0.559    -9390.399    5213.392
       _cons |   137791.1   61464.95     2.24   0.035     10320.56    265261.6
------------------------------------------------------------------------------
```

Figure A.1: Stata: Sample output.

is reported as the final regression coefficient, not the first. Additional regression output, such as standardized regression coefficients, can be obtained as options to the `regress` command. And to learn how to obtain still further output give command `help regress postestimation`.

## A.3 Eviews

**Eviews** is a commercial statistical package that is used mostly in economics, though is more broadly especially useful for regression analysis of time series data. A student version is available that, at the time of writing, can analyze large datasets but is restricted to saving and exporting datasets with no more than 1,500 observations per variable and 15,000 total observations. The Eviews website is `eviews.com`. A good introductory document is Eviews Illustrated, available at `http://www.eviews.com/illustrated/illustrated.html`.

Eviews stores both data and output as objects in a container called a **workfile**, with file-name extension **.wf1**. Some output is automatically saved in the workfile, while other output such as that from regression needs to be explicitly named as an **object** in order for it to be saved in the workfile. Storing results in a workfile is qualitatively similar to storing results in an Excel worksheet. And if desired there can be more than one worksheet, which Eviews calls a **page**. As for Excel spreadsheet analysis, results are linked to the data. So if the data is changed then the results also change. If instead one wants the results to be permanent, the command `freeze` is used.

An example workfile can be obtained by selecting File > New > Workfile, going to the Eviews directory and within this selecting subdirectory Example Files / EV8 Manual Data / Chapter 2 - A Demonstration / and opening file demo.wf1. This includes dataset variables, which Eviews calls series, and results from two estimated OLS equations.

To read in your own data select File > Open, and then select either Foreign Data as Workfile or Eviews Workfile, depending on whether or not the data is already in an Eviews workfile. Once you have browsed to find the dataset or workfile check the option Update default directory to set the working directory. This will remain the working directory in subsequent uses of Eviews, until you elect to change it. The working directory can also be set using the `cd` command.

Eviews does not automatically create a log file. Instead the default is to store results in the workfile. These results can then be cut and pasted into a word processor. Additionally one can direct output to a log file, discussed after the script below.

Eviews does not provide confidence intervals and $t$-tests in the univariate case. One can construct these manually, see the code below. Alternatively, a confidence interval for a single variable can be obtained following OLS regression of the variable on just an intercept and then using the confidence interval option within regression.

Eviews commands can be directly entered in the Command Window, located immediately

below the menu bar. Eviews, unlike many other packages, is not case sensitive. Alternatively, an Eviews command script can be stored in a file with file-name extension **.prg**. Unlike Stata and Gretl, however, using the Eviews windows interface does not lead to a listing of the corresponding Eviews command. This makes it harder to learn the Eviews command language, and Eviews is much easier to use interactively. The script can be opened from outside by double-clicking on the file, or within Eviews by choosing menu File > Run... and selecting the desired file. Chapter 12 of the Eviews 8 Command and Programming Reference summarizes Eviews basic commands.

An example Eviews script, with many of the commands needed for this book, is the following.

```
' aed_example.prg   Eviews example script written 4/1/2014
' Create a log file that contains just results in two steps
' First send all output to the printer - command pon
' Second redirect printer output to an RTF file - command output
pon
output(r) Eviewsresults.rtf
' Read in Stata dataset HOUSE.DTA
open HOUSE.DTA
' Create an object named gr0 that contains the variables Price and Size
group gr0 Price Size
' List the data for the two variables in object gr0
show gr0
' Descriptive statistics for the two variables
gr0.stats
' Example of creating a new variable
genr PriceperSqfoot = Price/Size
' Two-way scatterplot with regression line saved in .emf graphic format
freeze(grapheviews) gr0.scat(r)
grapheviews.save(t=wtf) grapheviews
' 95% confidence interval for one variable
scalar stderror = @stdev(price)/@sqrt(@obs(price))
scalar lb = @mean(price) - @qtdist(.975,@obs(price)-1)*stderror
scalar ub = @mean(price) + @qtdist(.975,@obs(price)-1)*stderror
show lb
show ub
' Two-sided T-test that population mean Price = 250000
scalar tstat = (@mean(price)-250000) / (@stdev(price)/@sqrt(@obs(price)))
scalar p = 2*(1-@ctdist(tstat,@obs(price)-1))
show tstat
```

```
show p
' OLS regression - same as chapter 13 - results are not saved in an object
ls Price c Size Lotsize Bedrooms Bathrooms Age Monthsold
' Same regression with results saved in equation object eq1
equation eq1.ls Price c Size Lotsize Bedrooms Bathrooms Age Monthsold
' 95% confidence interval for parameters
eq1.cinterval(p) .95
' Same regression with heteroskedastic robust standard errors - saved as eq1rob
equation eq1rob.ls(cov-White) Price c Size Lotsize Bedrooms Bathrooms Age Monthsold
' Save in an Eviews workfile the preceding analysis
save Eviewsexample.wf1
' Close output file
poff
' Exit Eviews
exit
```

Lines preceded by a single-quote (') are interpreted as comments, and command options appear in parentheses.

The command `output(r)` sends any text and graphics output that would normally be sent to the printer to instead be sent to a **rich text format** (**RTF**) file. Alternatively `output(t)` sends only text output, but not graph output, to a text file. Command `pon` sends output from each command to a printer. The commands `pon` and `poff` do not work in interactive use. If giving commands interactively, to send output to a file we again use the preceding `output` command, but now the option `(p)` needs to be used for each command that we want to obtain output for. For example the commands become `gr0.stats(p)`, `show(p)` and `eq1.ls(p)`. Output can also be sent to a **spool** file – this is not covered here. Note that if the file named in the `output` command already exists then Eviews appends the new output to the end of the file.

Eviews interprets a file with extension `.DTA` as a Stata dataset. The script requires that file HOUSE.DTA is in the working (default) directory, and not in some other directory. Command `freeze` preserves a graph or table as an object in the workfile. This Eviews script includes examples of commands to manually compute confidence intervals and perform a $t$-test. The command `ls` leads to OLS regression, and option `(cov_White)` leads to heteroskedastic-robust standard errors. By defining the equation `eq.ls`, the OLS results will be preserved in the workfile as the object `eq1`. The command `exit` results in automatically exiting Eviews once the script is run.

The resulting workfile includes data on each variable, the computed scalars such as `ub`, and the results for the named objects such as `gr0.stats`, `grapheviews` and `eq1.ls`. The log file includes all results, but not the commands that produced the results.

Figure A.2 gives some of the output from Eviews. Note that the output file does not

include the command that produced the output.

The first set of results in Figure A.2 are descriptive statistics. These include the Jarque-Bera test for whether the data are normally distributed with common mean and variance. Since $p < .05$ the null hypothesis of normality is rejected at significance level 0.05.

The second set of results in Figure A.2 are regression results that are identical to those given in Table 13.4 and discussed in Chapters 13.7 and 14.5. Unlike some of the other packages, the regression output does not include an analysis of variance table as econometrics analysis does not use this table. The regression output includes the Schwarz, Akaike and Hannan-Quinn criteria for model fit, where these criteria defined in Chapter 13.6 have been divided by the number of observations, here 29. The Durbin-Watson statistic is for test of whether the error term in a time series model is autocorrelated; see Chapter 20.

## A.4 Gretl

**Gretl** is a free statistical package developed specifically for economics. It covers considerably more econometrics methods than those presented in this book, though not as many methods as the commercial packages. The Gretl website `gretl.sourceforge.net` provides the package for free download as well as documentation.

Example datasets can be obtained by selecting File > Open data > Sample File ... To read in your own data it is best to first use File > Working Directory ... to change to the directory that has the dataset.

Gretl stores data in a data file with extension .gdt. It also stores both data and output as objects in a container called a **session file** with file-name extension **.gretl**. In order for output to be saved in the session file, one selects File > Save to session from the relevant results window. Then before exiting from Gretl choose from the main Gretl window File > Sessions Files > Save session. Storing results in a session file is qualitatively similar to storing results in an Excel worksheet.

It is easiest to learn Gretl by initially using the Gretl drop-down menus. The basic commands, though not necessarily subsequent options chosen through the Windows interface, can be viewed under Tools > Command log. This log file is called session.inp and can be saved, under a different name, by choosing the "Save as" option once viewing the log file.

Basic Gretl commands, such as those needed for this book, are relatively straightforward. Example Gretl scripts can be obtained by selecting File > Script Files > Practice file ... Note that like many other statistical packages, Gretl is case sensitive. Commands can be given from the command window, opened by choosing Tools > Gretl console. Or a script can be stored in a file with extension **.inp**. The script can be opened from outside Gretl by double-clicking on the file, or within Gretl by choosing menu File > Script Files > User file ... and selecting the desired file.

Output can be directed to a text file using command `outfile`. This output includes

| | PRICE | SIZE |
|---|---|---|
| Mean | 253910.3 | 1882.759 |
| Median | 244000.0 | 1800.000 |
| Maximum | 375000.0 | 3300.000 |
| Minimum | 204000.0 | 1400.000 |
| Std. Dev. | 37390.71 | 398.2721 |
| Skewness | 1.560837 | 1.727665 |
| Kurtosis | 5.612707 | 6.742809 |
| Jarque-Bera | 20.02340 | 31.35374 |
| Probability | 0.000045 | 0.000000 |
| Sum | 7363400. | 54600.00 |
| Sum Sq. Dev. | 3.91E+10 | 4441379. |
| Observations | 29 | 29 |

Dependent Variable: PRICE
Method: Least Squares
Date: 01/22/14 Time: 12:02
Sample: 1 29
Included observations: 29

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 137791.1 | 61464.95 | 2.241783 | 0.0354 |
| SIZE | 68.36942 | 15.38947 | 4.442610 | 0.0002 |
| LOTSIZE | 2303.221 | 7226.535 | 0.318717 | 0.7529 |
| BEDROOMS | 2685.315 | 9192.526 | 0.292119 | 0.7729 |
| BATHROOMS | 6832.880 | 15721.19 | 0.434629 | 0.6681 |
| AGE | -833.0386 | 719.3345 | -1.158068 | 0.2593 |
| MONTHSOLD | -2088.504 | 3520.898 | -0.593174 | 0.5591 |

| | | | |
|---|---|---|---|
| R-squared | 0.650553 | Mean dependent var | 253910.3 |
| Adjusted R-squared | 0.555249 | S.D. dependent var | 37390.71 |
| S.E. of regression | 24935.73 | Akaike info criterion | 23.29250 |
| Sum squared resid | 1.37E+10 | Schwarz criterion | 23.62253 |
| Log likelihood | -330.7412 | Hannan-Quinn criter. | 23.39586 |
| F-statistic | 6.826098 | Durbin-Watson stat | 1.259370 |
| Prob(F-statistic) | 0.000342 | | |

Figure A.2: Eviews: Sample output.

comments, commands and resulting text output.

Gretl does not automatically provide confidence intervals after regression. When using the Windows interface, after regression (menu Model > Ordinary Least Squares ...), within the regression output window select Analysis > Confidence intervals for coefficients. This method can also be used to obtain confidence intervals for a single variable – simply regress the variable on an intercept. When running Gretl using a script, however, there appears to be no way to obtain confidence intervals other than through saving coefficients and standard errors and writing code to compute the resultant confidence intervals.

An example Gretl script, with many of the key commands needed for this book, is the following.

```
# aed_example.do   Gretl example script written 4/1/2014
# Output file that contains both commands and results
outfile Gretlresults.txt -- write
# Read in the Stata dataset HOUSE.DTA
open HOUSE.DTA
# Example of help command - how to import datasets in different formats
help open
# Descriptive statistics for all variables
summary
# List data for a single variable
Price
# Example of creating a new variable
genr PriceperSqfoot = Price/Size
# Two-way scatterplot with regression line saved in .png graphic format
graph1 <- gnuplot Price Size --output=graphgretl.png
# 95% Confidence interval for one variable
summary Price
scalar lb = mean(Price) - critical(t,$nobs-1,0.025) * sd(Price)/sqrt($nobs)
scalar ub = mean(Price) + critical(t,$nobs-1,0.025) * sd(Price)/sqrt($nobs)
# Two-sided t-test that population mean Price = 250000
summary Price
scalar tstat = (mean(Price) - 250000) / (sd(Price)/sqrt($nobs))
scalar p = 2*pvalue(t,$nobs-1,abs(tstat))
# Descriptive statistics for two variables
summary Price Size
# Multivariate OLS regression - same as chapter 13
model1 <- ols Price const Size Lotsize Bedrooms Bathrooms Age Monthsold
# 95% confidence interval for regressor Size
scalar lb = $coeff(Size) - critical(t,$df,0.025) * $stderr(Size)
```

```
scalar ub = $coeff(Size) + critical(t,$df,0.025) * $stderr(Size)
printf "The 95%% confidence interval is (%.4f, %.4f)\n",lb,ub
# Same regression with heteroskedastic robust standard errors
model2 <- ols Price const Size Lotsize Bedrooms Bathrooms Age Monthsold --robust
# Save in a Gretl data file the dataset
store Gretldata.gdt
# Close output file
outfile ---close
# Exit Gretl
quit
```

Lines preceded by a hash mark (#) are interpreted as comments. The script requires that file HOUSE.DTA is in the working (default) directory, and not some other directory. Gretl treats a file with extension .DTA (or .dta) as a Stata format dataset. This Gretl script includes examples of commands to manually perform a $t$-test and to compute confidence intervals, since there is no Gretl command to do so. The command for regression requires the user to explicitly include the intercept, if one is desired, using the Gretl constructed variable called `const`. Leading text such as `model1 <-` is only necessary if you want to include the results in the session file, in this case as model1.

The output file includes results from running the script, aside from graphs, as well as the commands and comments. In principle a session file can be saved after running this script. But this seems possible only if commands are entered from the dropdown menus or from the Command Console, rather than through running an entire script.

Figure A.3 gives some of the output from Gretl. This is the form that the output takes in the script output window and in the corresponding text file if it is saved. Commands are preceded by a question mark, while results from the command are not.

The first set of results in Figure A.3 provides a wide range of descriptive statistics. The kurtosis statistic given is that for excess kurtosis.

The second set of results in Figure A.3 are regression results that are identical to those given in Table 13.4 and discussed in Chapters 13.7 and 14.5. Unlike some of the other packages, the regression output does not include an analysis of variance table as econometrics analysis does not use this table. The regression output includes the Schwarz, Akaike and Hannan-Quinn criteria for model fit; see Chapter 13.6.

## A.5 R

**R** is a free programming language for statistical and graphical analysis of data. It is used more in the biological and physical sciences than it is in economics. The R website `http://www.r-project.org/` provides the package for free download as well as links to manuals, and a helpful website is `http://www.ats.ucla.edu/stat/r/`.

```
# Descriptive statistics for two variables
? summary Price Size

                     Mean         Median        Minimum        Maximum
Price         2.5391e+005    2.4400e+005    2.0400e+005    3.7500e+005
Size               1882.8         1800.0         1400.0         3300.0

                Std. Dev.           C.V.       Skewness   Ex. kurtosis
Price              37391.        0.14726         1.5608         2.6127
Size               398.27        0.21154         1.7277         3.7428

                 5% perc.      95% perc.       IQ range   Missing obs.
Price         2.0800e+005    3.5750e+005         39500.              0
Size               1400.0         2950.0         450.00              0

# Multivariate OLS regression - same as chapter 13
? ols Price const Size Lotsize Bedrooms Bathrooms Age Monthsold

Model 1: OLS, using observations 1-29
Dependent variable: Price

              coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const       137791        61465.0       2.242    0.0354  **
  Size            68.3694      15.3895    4.443    0.0002  ***
  Lotsize       2303.22      7226.54      0.3187   0.7529
  Bedrooms      2685.32      9192.53      0.2921   0.7729
  Bathrooms     6832.88     15721.2       0.4346   0.6681
  Age           -833.039      719.335    -1.158    0.2593
  Monthsold    -2088.50      3520.90     -0.5932   0.5591

Mean dependent var   253910.3   S.D. dependent var   37390.71
Sum squared resid    1.37e+10   S.E. of regression   24935.73
R-squared            0.650553   Adjusted R-squared   0.555249
F(6, 22)             6.826098   P-value(F)           0.000342
Log-likelihood      -330.7412   Akaike criterion     675.4824
Schwarz criterion    685.0535   Hannan-Quinn         678.4799

Excluding the constant, p-value was highest for variable 4 (Bedrooms)
```

Figure A.3: Gretl: Sample output.

Some statistical analysis requires use of commands not included in the base installation of R. In that case the **package** that includes the command needs to be installed in the library subdirectory of R, and any R script that uses the commands must first call the **library** for necessary package. A statistical procedure may be included in several different packages, with different command names, so there can be more than one way to implement a given procedure in R.

R stores both data and output as objects in a file with file-name extension **.Rda**. Additionally output can be directed to a log file using command `sink`.

It is easiest to begin by using **R Commander**, a free Windows interface front-end to R with drop-down menus that needs to be installed separately from installation of R. R Commander provides two windows. An output window lists the output from the commands. A script window provides the associated commands in the R programming language. This provides a convenient way to learn the commands, as is the case for Stata and Gretl. And from the script window one can directly run (or submit) commands. So it is easy to modify an existing command, such as adding R options not made available through the drop-down menus, and execute the modified command. Note that, like many packages, R is case sensitive.

R Commander provides only a subset of R through the Windows interface, so ultimately you need to program in R.

For direct programming in R a good free Windows interface to use is **R Studio**. The `sink` command in R creates a log file, but this does not list all commands and results. An alternative is to use no log file. Then a complete text listing of both commands and results is given in the Console Window of R Studio, after running the script. This output can be cut and pasted into a Word processor.

The following R commands install packages that are used in the R code below. This code should be executed if these packages are not already included in the R libraries for your installation of R.

```
# Once only - Install any needed packages in your version of R
install.packages("foreign")
install.packages("psych")
install.packages("MASS")
install.packages("sandwich")
install.packages("TeachingDemos")
```

An example R script, with many of the key commands needed for this book, is the following:

```
# aed_example.R   R example script written 4/1/2014
# Clear the workspace
rm(list=ls())
```

```
# Create a log file with commands and results - requires package TeachingDemos
library(TeachingDemos)
txtStart("Rresults.txt")
# Read in the Stata data set - requires package foreign
library(foreign)
data.appxA <- read.dta(file = "HOUSE.DTA")
# Allow variables in database to be accessed simply by giving names
attach(data.appxA)
# Example of help command
help(load)
# Summary statistics - very basic
summary(data.appxA)
# Detailed descriptive statistics for one variable - uses package pscyh
library(psych)
describe(data.appxA)
# List the first ten observations
head(data.appxA)
# Two-way scatterplot with regression line and save as PNG file
png("graphr.png")
reg1 <- lm(Price~Size)
plot(Size,Price)
abline(reg1)
dev.off()
# Multivariate OLS regression - same as chapter 13
model1.ols <- lm(Price ~Size+Lotsize+Bedrooms+Bathrooms+Age+Monthsold)
summary(model1.ols)
# Confidence interval for model parameters - uses package MASS
library(MASS)
confint(model1.ols)
# Same regression with het-robust standard errors - uses package sandwich
# Need to manually construct t statistics, p values and confidence interval
library(sandwich)
cov.robust <- vcovHC (model.ols, type="HC1")
se.robust <- sqrt(diag(cov.robust))
coeffs <- coef(model.ols)
t.robust <- coeffs / se.robust
df <- model.ols$df.residual
crit <- qt(.975,df)
pvalue <- 2*pt(abs(coeffs/se.robust),22,0,0),
summary.olsrobust <- cbind(coeffs, se.robust, t.robust, pvalue,
```

```
                    lb=coeffs-crit*se.robust, ub=coeffs+crit*se.robust)
# Save workspace in .RData file
save.image(file="Rexample.Rdata")
# Close output file
txtStop()
# exit R
quit(save="no")
```

Lines preceded by a hash mark (#) are interpreted as comments. The package `TeachingDemos` command `txtStart` sends all commands and results to a text file and is better than the base installation command `sink`. The script requires that file HOUSE.DTA is in the working (default) directory, and not some other directory. The package `foreign` command `read.dta` reads a file as a Stata format dataset. The package `psych` command `describe` gives much more detailed descriptive statistics than the command `summary`. The command for regression automatically includes an intercept. The package `MASS` command `confint` gives confidence intervals after OLS regression – these are not included in the basic summary of OLS results. The package `Sandwich` command `vcovHC` can be used to obtain heteroskedastic-robust standard errors. The command `save.image` saves all the results (objects and functions) in an Rdata file that can subsequently read into R using command `load("Rexample.Rdata")`.

Figure A.4 gives some of the output from R.

The first set of results in Figure A.4 provides a wide range of descriptive statistics. The symmetry and excess kurtosis statistics differ a little from those obtained using the other packages, due to use of slightly different formulae.

The second set of results in Figure A.5 are regression results that are identical to those given in Table 13.4 and discussed in Chapters 13.7 and 14.5. Unlike some of the other packages, the regression output does not include an analysis of variance table as econometrics analysis does not use this table.

## A.6 Excel and Google Sheets

**Excel** is a commercial spreadsheet package that is part of Microsoft Office. It has very limited statistical capabilities compared to the preceding statistical packages, as it is a spreadsheet package and not a statistical package. If at all possible, use a statistical package rather than Excel to implement the methods in this book. **Google Sheets** can be used in place of Excel.

Excel includes single functions that compute some key statistical quantities. For example, typing in cell A31 the command =AVERAGE(A2:A30) sets cell A31 equal to the average of the values in rows 2 to 30 of column A. Excel is available in many different languages, and commands such as AVERAGE will differ with the language version.

It is much more convenient, however, to use the **Data Analysis Toolpak** in Excel as this automatically combines several commands at once, such as computing the several different

```
> library(psych)
> describe(cbind(Price, Size))

      var  n      mean       sd median trimmed      mad    min    max  range
Price   1 29 253910.34 37390.71 244000  249296 22239.00 204000 375000 171000
Size    2 29   1882.76   398.27   1800    1836   296.52   1400   3300   1900
      skew kurtosis      se
Price 1.48     2.23 6943.28
Size  1.64     3.29   73.96

> model.ols <- lm(Price ~ Size + Lotsize + Bedrooms + Bathrooms +
+ Age + Monthsold)
> summary(model.ols)

Call:
lm(formula = Price ~ Size + Lotsize + Bedrooms + Bathrooms +
    Age + Monthsold)

Residuals:
   Min     1Q Median     3Q    Max
-37613 -18463  -1215  16800  43130

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 137791.07   61464.95   2.242 0.035387 *
Size            68.37      15.39   4.443 0.000205 ***
Lotsize       2303.22    7226.54   0.319 0.752947
Bedrooms      2685.32    9192.53   0.292 0.772932
Bathrooms     6832.88   15721.19   0.435 0.668065
Age           -833.04     719.33  -1.158 0.259254
Monthsold    -2088.50    3520.90  -0.593 0.559114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24940 on 22 degrees of freedom
Multiple R-squared:  0.6506,  Adjusted R-squared:  0.5552
F-statistic: 6.826 on 6 and 22 DF,  p-value: 0.0003424
```

Figure A.4: R: Sample output.

quantities reported in a table of descriptive statistics. However, this toolpack provides very few statistical methods. Furthermore the data analysis toolpack is not available on Apple computers as it is an Excel add-on that is written in visual basic, a programming language specific to Microsoft Windows. Instead, an Apple user needs to purchase a commercial add-on package for the Mac that replicates the data analysis toolpack, or purchase a Windows PC version of Excel and run Windows on the Apple computer.

At the time of writing, February 2015, the **XLMiner Analysis Toolpack** is freely available for both Excel and Google Sheets and essentially replicates the Data Analysis Toolpack.

For PC users the Data Analysis Toolpak appears at the right-end of the Data menu as Data Analysis in Excel 2007 and later versions, and appears in the Tools menu of earlier versions of Excel. Because the analysis toolpak is an Excel add-in, it may not be automatically loaded when you open Excel. If this is the case, an internet search on `install excel analysis toolpak` will provide instructions on how to do so.

Figure A.5 provides some essential output from Excel analysis of file HOUSE.CSV that includes variable names in the first row. The descriptive statistics output is obtained by choosing Descriptive Statistics in the Data Analysis Toolpak. Then choose input range \$A\$1:\$B\$30, data grouped by columns, labels in first row, summary statistics, confidence level for the mean (with default 95%) and set $k^{th}$ largest and $k^{th}$ smallest at 8, as this gives the upper and lower quartiles for a sample of size $n = 29$. The output can be placed in either the current worksheet at, say, column \$A\$32, or in a new worksheet. The entry confidence level (95%) gives the half-width of a confidence interval, so the 95% confidence interval is $253910.3 \pm 14222.60 = (239687.7, 268133.0)$. The symmetry and excess kurtosis statistics differ a little from those obtained using the other packages, due to use of slightly different formulae.

The multiple regression output in Figure A.5 is obtained by choosing Regression in the Data Analysis Toolpak. Then choose input Y range \$A\$1:\$A\$30, input X range \$B\$1:\$G\$30, labels data grouped by columns, labels in first row, and confidence level (with default 95%). Again the output can be placed in either the current worksheet at, say, column \$A\$52, or in a new worksheet. The Regression Statistics output is the usual summary statistics after regression. The ANOVA output is an analysis of variance table that gives the same sum of squares decomposition as the Stata output discussed in Appendix A.2, along with the resultant F-statistic for the test of overall significance, and its associated $p$-value. The final regression output is the same as that given in Chapter 13. Note that Excel automatically includes a constant, and this is reported as the first regression coefficient. Additional regression diagnostics, included residual plots and line fit plots, are available as Regression options.

| *Price* | | *Size* | |
|---|---|---|---|
| | | | |
| Mean | 253910.34 | Mean | 1882.758621 |
| Standard Error | 6943.2807 | Standard Error | 73.95727789 |
| Median | 244000 | Median | 1800 |
| Mode | 235000 | Mode | 1600 |
| Standard Deviation | 37390.711 | Standard Deviation | 398.2721302 |
| Sample Variance | 1.398E+09 | Sample Variance | 158620.6897 |
| Kurtosis | 3.3656321 | Kurtosis | 4.717891678 |
| Skewness | 1.6472971 | Skewness | 1.823365476 |
| Range | 171000 | Range | 1900 |
| Minimum | 204000 | Minimum | 1400 |
| Maximum | 375000 | Maximum | 3300 |
| Sum | 7363400 | Sum | 54600 |
| Count | 29 | Count | 29 |
| Largest(8) | 270000 | Largest(8) | 2000 |
| Smallest(8) | 233000 | Smallest(8) | 1600 |
| Confidence Level(95.0%) | 14222.666 | Confidence Level(95 | 151.4946142 |

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.8065686 |
| R Square | 0.6505528 |
| Adjusted R Square | 0.5552491 |
| Standard Error | 24935.734 |
| Observations | 29 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 6 | 25466429042 | 4244404840 | 6.826098 | 0.000342425 |
| Residual | 22 | 13679397855 | 621790811.6 | | |
| Total | 28 | 39145826897 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 137791.07 | 61464.95187 | 2.241782699 | 0.035387 | 10320.55799 | 265261.573 |
| Size | 68.369419 | 15.38947183 | 4.44260984 | 0.000205 | 36.45360797 | 100.28523 |
| Lotsize | 2303.2214 | 7226.535205 | 0.318717242 | 0.752947 | -12683.69529 | 17290.138 |
| Bedrooms | 2685.3151 | 9192.525674 | 0.292119404 | 0.772932 | -16378.81621 | 21749.4465 |
| Bathrooms | 6832.88 | 15721.19154 | 0.434628635 | 0.668065 | -25770.87557 | 39436.6356 |
| Age | -833.0386 | 719.3345439 | -1.1580684 | 0.259254 | -2324.847132 | 658.769929 |
| Monthsold | -2088.504 | 3520.897859 | -0.59317359 | 0.559114 | -9390.398837 | 5213.39159 |

Figure A.5: Excel: Sample output.

Table A.2: Statistical packages: commands to calculate 95 percent confidence interval critical values.

| | Distribution and command | | | |
|---|---|---|---|---|
| | Students t | Standard normal | F distribution | Chisquared |
| | $t_{22}$ | N(0,1) | $F_{5,22}$ | $\chi^2_5$ |
| Stata | invttail(22,.025) | -invnormal(.025) | invFtail(5,.05) | invchi2(5,.05) |
| Eviews | @qtdist(.975,22) | @qnorm(.975) | @qfdist(.95,5,22) | @qchisq(.95,5) |
| Gretl | critical(t,22,.025) | critical(z,22,.025) | critical(F,5,22,.05) | critical(x,5,.05) |
| R | qt(.975,22) | qnorm(.975) | qf(.95,6,22) | qchisq(.95,6) |
| Excel | TINV(.05,22) | NORMSINV(.975) | FINV(.05,5,22) | CHIINV(.05,5) |
| Value | $t_{.025;22}$ =2.07387 | $z_{.025}$=1.95996 | $F_{.025;5,22}$ =2.66127 | $\chi^2_{.025;5}$ =11.07050 |

## A.7 Critical values and p-values

In many cases output from statistical packages includes desired confidence intervals and $p$-values for hypothesis tests.

When this is not the case one needs to use tables or commands in a statistical package to obtain the necessary critical values for confidence intervals and to obtain either $p$-values or critical values for hypothesis tests.

Unfortunately, the commands are very arcane. Different packages use different function names, the degrees of freedom may be given as the first argument or the last argument of the function, and probabilities given may be in the upper tail, lower tail or both tails depending on the package used.

The first two columns of Table A.2 present commands in various packages to obtain critical values for 95% confidence intervals and for two-sided hypothesis tests at level 0.05 based on, respectively, the $t_{22}$ distribution and the standard normal distribution. The last two columns present critical values for hypothesis tests based on, respectively, the $F_{5,22}$ distribution and the $\chi^2_5$ distribution. The degrees of freedom chosen are those for the multivariate regression of Chapter 13 with $n = 29$, $k = 7$ and, for the last two columns, a test of whether the five regressors other than *Size* are statistically significant at 5%. In all cases the null hypothesis is rejected if the test statistic exceeds the critical value.

Table A.3 presents commands in various packages to obtain $p$-values for two-sided hypothesis tests based on respectively, the $t_{22}$ distribution, the standard normal distribution, the $F_{5,22}$ distribution and the $\chi^2_5$ distribution. For the first two columns the test statistic of .292 is the $t$-statistic for test of statistical significance of the regressor *Bedrooms* in the multivariate regression of Chapter 13. The $F$-test statistic of .417 is that for test of whether the five regressors other than *Size* are statistically significant. For the chisquared distribution the test value of .417 is multiplied by the degrees of freedom, here five, to make the $\chi^2$ test

Table A.3: Statistical packages: commands to calculate p-values for a two-sided test.

| | Distribution and command | | | |
|---|---|---|---|---|
| | Students t | Standard normal | F distribution | Chisquared |
| | $\Pr[\lvert T_{22}\rvert>.292]$ | $\Pr[\lvert Z\rvert>.292]$ | $\Pr[\lvert F_{5,22}\rvert>.417]$ | $\Pr[\chi^2_5>5\times.417]$ |
| Stata | 2*ttail(22,.292) | 2*(1-normal(.292)) | Ftail(5,22,.417) | chi2tail(5,5*.417) |
| Eviews | 2*(1-@ctdist(.292,22)) | 2*(1-@cnorm(.292)) | 1-@cfdist(.417,5,22) | 1-@cchisq(5*.417,5) |
| Gretl | 2*pvalue(t,22,.292) | 2*pvalue(z,.292) | pvalue(F,5,22,.417) | pvalue(x,5,5*.417) |
| R | 2*pt(.292,22,0,0) | 2*pnorm(.292,0,1,0) | pf(.417,5,22,0,0) | pchisq(5*.417,5,0,0) |
| Excel | TDIST(.292,22,2) | 2*(1-NORMSDIST(0.292)) | FDIST(0.417,5,22) | CHIDIST(5*0.417,5) |
| Value | 0.77302 | 0.77029 | 0.83181 | 0.83726 |

comparable to the $F$ test.