Analysis of Economics Data
Chapter 12: Further Topics in Multiple Regression

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 12: Further Topics in Multiple Regression

- In most applications assumptions 3-4 on the regression model errors are too restrictive
  - then default standard errors for the OLS coefficients are wrong
    - so subsequent confidence intervals and tests are wrong
  - instead we should **use appropriate robust standard errors**
    - which ones vary with the particular data application
    - this can require experience.
- For **prediction** it is important to distinguish between
  - predicting an average outcome
  - predicting an individual outcome (more difficult to do precisely).

Outline

1. Inference with Robust Standard Errors
2. Prediction
3. Nonrepresentative Samples
4. Best Estimation
5. Best Confidence Intervals
6. Best Hypothesis Tests
7. Data Science and Big Data: An Overview
8. Bayesian Methods: An Overview
9. A Brief History of Statistics and Regression

Datasets: HOUSE, REALGDPPC

12.1 Inference with Robust Standard Errors

- Continue with assumptions 1-2 so OLS estimates are still unbiased.
- Relax error assumptions 3-4 as then assumptions are more realistic
  - this leads to **different standard errors** for b_j denoted $se_{rob}(b_j)$.
- Three common complications give different $se_{rob}(b_j)$.

Complication	Robust Standard Error Type	Data Type
1. Heteroskedasticity: Error variance varies over i	Heteroskedasticity robust	Cross Section (if errors independent)
2. Clustered: Errors in same cluster are correlated	Cluster robust	Some Cross section Most Short Panel
3. Autocorrelation: Errors correlated over time	Heteroskedasticity and autocorrelation (HAC) robust	Most Time Series Some Long Panel

Inference with Robust Standard Errors (continued)

- For **implementation**, use the appropriate command in a statistical package
  - in Stata use `regress` command with the `vce( )` option
  - in R use the `sandwich` package
  - chapter 12.1.9 provides details.
- Once the appropriate standard errors $se_{rob}(b_j)$ are obtained the rest follows as usual
  - for a single parameter test use $t = (b_j - \beta_j)/se_{rob}(b_j) \sim T_v$
  - for a confidence interval on β_j use $b_j \pm t_{v;\alpha/2} \times se_{rob}(b_j)$.
- The degrees of freedom are usually $v = n - k$
  - except for cluster-robust use $v = G - 1$ where G is the number of clusters.
- **The key is to know which type of robust standard error to use.**

Heteroskedastic-Robust Standard Errors

- In many cross-section data applications
  - it may be reasonable to assume error independence across observations
  - but errors are **heteroskedastic** (**the error variance varies across observations**).
  - OLS is still unbiased under assumptions 1-2
  - but default standard errors are invalid.
- Make the following change to assumptions 1-4
  - change 3 to 3' that $\text{Var}[u_i] = \sigma_i^2$ (which depends on $x's$) and $n \to \infty$
- **The formula for** $se(b_j)$ **changes** to, say, $se_{het}(b_j)$.
- Computer output is qualitatively similar
  - $b_1, ..., b_k$ are unchanged
  - now get $se_{het}(b_1), ..., se_{het}(b_k)$
  - leading to different t-statistics and confidence intervals.

House Price Example: Heteroskedastic-Robust Standard Errors

Variable	Coefficient	Robust se	t statistic	p value	95% conf. int.	
Size	68.37	15.36	4.44	0.000	36.52	100.22
Lot Size	23020	5329	0.43	0.670	-8748	13355
Bedrooms	2685	8286	0.32	0.749	-14498	19868
Bathrooms	6833	19284	0.35	0.726	-33159	46825
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
n	29					
F(6,22)	6.41					
p-value for F	0.0005					
R^2	0.651					
St. error	24936					

House Price Example (continued)

- Same intercept and slope coefficient estimates (as still OLS).
- For individual standard errors the biggest change is 30%
  - again only *Size* is statistically significant at 5%.
- Again regressors are jointly statistically significant at 5%
  - $F = 6.41$ (compared to 6.83).
- For test of joint statistical significance of *lotsize* .... *monthsold*
  - $F = 0.46 \sim F(5, 22)$ compared to $F = 0.42$ with defaults se's
  - reject H_0 at level 0.05 as $p = .8038 > 0.05$.
- The heteroskedastic-robust standard errors can be larger or smaller than default standard errors
  - the two are generally within 30% of each other.

Cluster-Robust Standard Errors

- In many cross-section data and panel data applications
  - **errors may be independent across clusters but correlated within cluster**
  - and additionally errors are **heteroskedastic**.
- **Cross-section data example**
  - independent errors for individuals in different villages but correlated for individuals in the same village.
- **Panel data example**
  - errors may be independent across individuals but correlated over time for a given individual.
- Then must use **cluster-robust standard errors**
  - these can be several times default or het-robust standard errors!
  - with correlation within cluster, adding an observation to a cluster gives less than a completely new independent piece of information
  - cluster-robust correct s for this reduced estimator precision!

Cluster-Robust Standard Errors

- OLS is still unbiased but default standard errors are too small.
- Make the following changes to assumptions 1-4
  - change 3 to 3': $\text{Var}[u_i|x_i's] = \sigma_i^2$ (so heteroskedastic)
  - change 4 to 4': correlated errors for observations in same cluster
    - and need $G \to \infty$ where G is the number of clusters.
- The formula for $se(b_j)$ changes to, say, $se_{Clu}(b_j)$
  - inference uses $T(G-1)$
    - note the much smaller degrees of freedom.
- Implementation requires specifying a variable for the clusters.

Cluster-Robust Standard Errors in Practice

- Cluster-robust standard errors **can be several times** the default or heteroskedastic-robust standard errors.
- The difference with default or heteroskedastic-robust se's gets greater
  - the more observations there are per cluster
  - the more highly correlated the regressors are within cluster
  - the more highly correlated the errors are within cluster.
- **It is essential to use cluster-robust standard errors if needed.**
- It can sometimes be difficult to know how to form the clusters.
  - data examples are given in chapters 13.4.4, 13.6.4 and 17.3.1.

HAC-Robust Standard Errors for Time Series

- Time series models often have **autocorrelated errors**
  - an autocorrelated error is one that is correlated with errors in previous periods (e.g. $u_t = 0.8u_{t-1}$).
- If errors are autocorrelated then default standard errors are invalid
  - instead use heteroskedastic- and autocorrelation-robust **(HAC) standard errors**.
- Make the following changes to assumptions 1-4
  - change 2 to 2': error has mean zero conditional on current and past values of the regressors.
  - change 3 to 3': $\text{Var}[u_t|x_t's$ and past $x_t's] = \sigma_t^2$
  - change 4 to 4': errors are correlated up to m periods apart and $T \rightarrow \infty$
- The formula for $se(b_j)$ changes to, say, $se_{HAC}(b_j)$
- The lag length m needs to be specified or be data determined
  -
    - a rule of thumb is $m = 0.75 \times T^{1/3}$ where $T = \#$ of observations.
- Data examples are given in chapters 13.2, 13.3 and 17.8.

12.2 Prediction

- **Predicting a value** is straightforward.
- Predict for a given value of regressors, say $x_2 = x_2^*, ..., x_k = x_k^*$ using

$$\widehat{y}|x_2^*, ..., x_k^* = b_1 + b_2 x_2^* + ... + b_k x_k^*.$$

- Example: regress *Price* on just *Size*
    - Predict a 2000 square foot 4-bedroom house will sell for \$262,559
    - since, using estimates reported in Section 10.4, $\widehat{y} = 115017 + 73.771 \times 2000 = 262559$.
- But **estimating the standard error of the prediction is subtle**
    - it depends on whether we are predicting an average outcome or an individual outcome.

Average Outcome versus Actual Value

- **Key distinction** is between predict an average outcome and predict an individual outcome.
- **Average outcome** or **conditional mean**

$$\mathrm{E}[y|x_2^*, ..., x_k^*] = \beta_1 + \beta_2 x^* + \cdots + \beta_k x_k^*$$

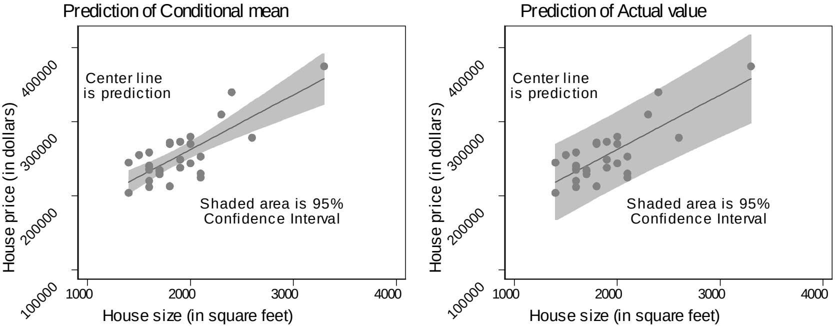
- **Individual outcome** or **the actual value**

$$y|x_2^*, ..., x_k^* = \beta_1 + \beta_2 x^* + \cdots + \beta_k x_k^* + u^*.$$

- **For both we use the same prediction** $\widehat{y} = b_1 + b_2 x_2^* + ... + b_k x_k^*$.
- **But the precision of the prediction varies with use**
  - for individual outcome we also need to predict u^* leading to noisier prediction
    - with variance necessarily at least $\mathrm{Var}[u^*]$.
- The following slide makes clear this distinction.

Example: 95% Confidence Intervals for E[y|x*] and y|x*

- Regress house *Price* on *Size*
  - predict house price at a range of house sizes
  - first panel: 95% confidence interval for the conditional mean price.
  - second panel: 95% confidence interval for actual price is much wider.



Prediction of an Average Outcome

- The **conditional mean** of y is

$$E[y|x_2^*, ..., x_k^*] = \beta_1 + \beta_2 x^* + \cdots + \beta_k x_k^*.$$

- Use $\widehat{y}_{cm} = b_1 + b_2 x_2^* + ... + b_k x_k^*$.
- $\text{Var}(\widehat{y}_{cm})$ depends on the precision of the estimates $b_1, ...b_k$.
- Define $se(\widehat{y}_{cm})$ to be the standard error of $\widehat{y}_{cm}$.
- A $100(1-\alpha)\%$ confidence interval for the conditional mean is

$$\text{E}[y|x_2^*, ..., x_k^*] \in \widehat{y}_{cm} \pm t_{n-k,\alpha/2} \times se(\widehat{y}_{CM}).$$

- $\text{Var}[\widehat{y}_{cm}] \rightarrow 0$ and $se(\widehat{y}_{cm}) \rightarrow 0$ as the estimates $b_1, ..., b_k$ become more precise.

Prediction of an Actual Value (A Forecast)

- The **actual value or forecast value** of y for $x = x^*$ is

$$y|x^* = \beta_1 + \beta_2 x^* + \cdots + \beta_k x_k^* + u^*.$$

- Use $\widehat{y}_f = b_1 + b_2 x_2^* + ... + b_k x_k^*$ as best estimate of u^* is zero.
- Then $\text{Var}(\widehat{y}_f)$ depends additionally on $\text{Var}(u^*)$
  - $\text{Var}[\widehat{y}_f] = \text{Var}[\widehat{y}_{cm}] + \text{Var}[u^*]$
- Define $se(\widehat{y}_f)$ to be the standard error of $\widehat{y}_f$
  - then $se(\widehat{y}_f) = \sqrt{se^2(\widehat{y}_{CM}) + s_{u^*}^2}$ where $s_{u^*}^2$ is estimate of $\text{Var}[u^*]$.
- A $100(1-\alpha)\%$ confidence interval for the forecast is

$$y|x_2^*, ..., x_k^* \in \widehat{y}_f \pm t_{n-k,\alpha/2} \times se(\widehat{y}_f).$$

- $\text{Var}[\widehat{y}_f] > \text{Var}[u^*]$ always, even if $b_1, ..., b_k$ are very precise.

Forecasts can be quite imprecise

- Recall that in forecasting
  - we use $\widehat{y}_f = b_1 + b_2 x_2^* + ... + b_k x_k^*$
  - to forecast $y|x^* = \beta_1 + \beta_2 x^* + \cdots + \beta_k x_k^* + u^*$
- Even if $b_1, ..., b_k$ are very precisely estimated we still have u^*.
- So $\text{Var}(\widehat{y}_f) \geq \text{Var}(u^*)$ and $\text{St.dev.}(\widehat{y}_f) \geq \text{St.dev.}(u^*)$.
- The obvious estimate of $\text{St.dev.}(u^*)$ is the standard error of the regression s_e.
- So in large samples a 95% confidence interval for the forecast is at least as wide as

$$\begin{aligned} y|x_2^*, ..., x_k^* &\in \widehat{y}_f \pm 1.96 \times s_e \\ s_e^2 &= \tfrac{1}{n-k} \textstyle\sum_{i=1}^n (y_i - \widehat{y}_i)^2. \end{aligned}$$

Are Poor Forecasts a Problem?

- Econometric models of individual behavior can have low R^2
  - so the variance of the model error and s_e are large, so $se(\widehat{y}_f)$ is large.
  - leading to very noisy forecasts of individual outcomes
  - nonetheless the prediction of average outcomes may be quite precise, with low $se(\widehat{y}_{cm})$
  - and policy-makers often base policy on average outcomes.
- For example, many studies find that on average education has an economically and statistically significant impact on earnings
  - even though for an individual the confidence interval for forecast earnings given years of education is very wide.
- Knowing that on average greater education is predicted to lead to higher earnings encourages government to subsidize education
  - even though we cannot predict with much certainty that a given person with a high level of education will have high earnings.

Bivariate Prediction under Assumptions 1-4

- For bivariate regression under assumptions 1-4 the formula for $se(\widehat{y}_{cm})$ is

$$se(\widehat{y}_{cm}) = s_e \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- So the predicted conditional mean is more precise when
  - 1. sample y_i are closer to the regression line: then s_e is smaller.
  - 2. variation in regressors is greater: then $\sum_{i=1}^n (x_i - \bar{x})^2$ is larger.
  - 3. x^* is closer to the sample mean: then $(x^* - \bar{x})^2$ is smaller.
  - 4. sample size is larger: then $1/n$ and $(x^* - \bar{x})^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ are smaller.
- Furthermore: $se(\widehat{y}_{cm}) \longrightarrow 0$ as $n \longrightarrow \infty$ due to 4.
- When robust standard errors are used specialized software is needed to get confidence intervals.

Bivariate Forecast under Assumptions 1-4

- Again consider regression of y on x under assumptions 1-4.
- Given homoskedastic errors $\text{Var}(u^*) = \sigma_u^2$ so $s_{u^*}^2 = s_e^2$
  - then $se(\widehat{y}_f) = \sqrt{se^2(\widehat{y}_{CM}) + s_e^2}$
- For prediction of the actual value the formula for $se(\widehat{y}_f)$ is

$$se(\widehat{y}_f) = s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

- Now $se(\widehat{y}_f) \geq s_e$ does not go to zero as $n \to \infty$.

Example: House Price given Multiple Characteristics

- Predictions for a 2000 square foot house with medium lot size, four bedrooms, two bathrooms, forty-years old and sold in June.
- The predicted value is

$$\widehat{y} = b_1 + 2000b_2 + 2b_3 + 4b_4 + 2b_5 + 40b_6 + 6b_7 = 257691.$$

- Predict conditional mean assuming assumptions 1-4 hold
  - use statistical software with commands for prediction after OLS
  - $se(\widehat{y}_{cm}) = 6488$ using default standard errors
  - 95% confidence interval for the conditional mean house price
    - ★ $257691 \pm t_{22,.025} \times 6488 = (\$244{,}235, \$271{,}146)$.
- Forecast the actual value assuming assumptions 1-4 hold
  - $s_e = 24936$, $se(\widehat{y}_{cm}) = 6488$, so $se(\widehat{y}_f) = \sqrt{6488^2 + 24936^2} = 25766$.
  - 95% confidence interval for the actual house price
    - ★ $257691 \pm t_{22,.025} \times 25766 = (\$204{,}255, \$311{,}126)$.

Example: House Price with Robust Standard Errors

- Now suppose instead that model errors are heteroskedastic.
- Predict conditional mean
  - use statistical software with commands for prediction after OLS
  - $se(\widehat{y}_{cm}) = 6631$ using heteroskedastic-robust standard errors
  - 95% confidence interval for the conditional mean house price
    - ★ $257691 \pm t_{22,.025} \times 6631 = (\$243{,}939,\ \$271{,}442)$.
- Forecast the actual value
  - we additionally need an estimate of $\text{Var}[u|x^*, ..., x_k^*]$
  - it is simplest to again use $s_e^2 = 24936^2$
  - $s_e = 24936$, $se(\widehat{y}_{cm}) = 6631$, so $se(\widehat{y}_f) = \sqrt{6488^2 + 24936^2} = 25803$.
  - 95% confidence interval for the actual house price
    - ★ $257691 \pm t_{22,.025} \times 25803 = (\$204{,}178,\ \$311{,}203)$.

12.3 Nonrepresentative Samples

- Many studies use survey data that may be nonrepresentative of the population.
- If there is nonrandom sampling on variables other than the dependent variable y then OLS can estimate population parameters if we include these variables as control variables in the regression
  - e.g. include gender and race as controls.
- If there is nonrandom sampling on the dependent variable OLS does not lead to consistent estimates of population parameters
  - e.g. if high earners are omitted from survey and we want to model earnings in the population.
- Many surveys include sample weights that adjust for nonrepresentativeness
  - then population weighted least squares can be used.

12.4 Best Estimation

- An estimator b_j is **unbiased** for β_j if $E[b_j] = \beta_j$.
- An estimator b_j is **consistent** if as $n \rightarrow \infty$ any bias in $b_j \rightarrow 0$ and $\text{Var}[b_j] \rightarrow 0$.
- A best estimator has smallest variance among unbiased estimators or among consistent estimators.
- When assumptions 3-4 do not hold OLS is no longer best.
- Feasible generalized least squares (FGLS) is instead the best estimator
  - FGLS requires additionally specifying a model for the error variances and covariances and estimating this model
    - this model varies with the model for the errors
- In practice for linear regression models
  - most studies just use OLS with appropriate robust standard errors
  - this loses some precision but the loss is often not great.

12.5 Best Confidence Intervals

- Best confidence intervals are those with the shortest width at a given level of confidence.
- For standard estimators the 95% confidence interval is of form

$$\widehat{\beta}_j \pm t_{n-k,\alpha/2} \times se(\widehat{\beta}_j)$$

- So the shortest interval is that with smallest $se(\widehat{\beta}_j)$ and hence most efficient estimator.
- In practice even if assumptions 3-4 do not hold
  - most studies base confidence intervals on OLS with appropriate robust standard errors
  - this increases confidence interval width but the increase is often not great.

12.6 Best Tests: Type I and II errors

- Consider H_0 : **no disease** versus H_a : **disease is present**.
- **Two errors** can be made in hypothesis testing.
- A **type I error** (or **false positive**)
  - H_0 is rejected when H_0 is true
    - ★ so find disease even though no disease is present
  - to date we have only considered type 1 error (see Chapter 4.4).
- A **type II error** (or **false negative**)
  - H_0 is not rejected when H_0 is false
  - so find no disease when disease is present.

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 negative)
Reject H_0 : Find disease	**Type I error** (false positive)	Correct decision

Test Size and Power

- **Test size** is the **probability of a type I error**.
  - Test size is set at α, the **significance level of the test**.
- **Test power** is **one minus the probability of a type II error**
  - High power is preferred as then low Pr[type II error]
- Problem: there is a **trade-off**
  - Pr[type I error] decreases $\Rightarrow$ Pr[type II error] increases
  - e.g. Can set Pr[type I error] $= 0$ if never reject H_0.
- Solution: use **most powerful test**
  - this has highest power for given test size
  - this is a test based on most precise estimator.
- In practice while test size is set low (e.g. 5%)
  - the Pr[type II error] can be high and test power may be low.

12.7 Data Science and Big Data: An Overview

- **Data science** or **data analytics** is the science of discerning patterns in data.
- **Machine learning** is a branch of artificial intelligence
  - algorithmically learn from data (the machine learns)
  - rather than specify a model based on expert knowledge of the particular application
  - methods include lasso, regression trees, random forests, neural networks, deep learning.
- **Big data** refers to datasets that are enormously large
  - though big data methods may also be applied to smaller datasets.

Prediction using Big Data

- Often the goal of big data is **prediction**
  - machine learning methods can predict better than earlier methods such as OLS.
- In some cases the predictions at the individual level are very precise
  - e.g. recognizing the numbers and letters on a digital image of a vehicle license plate.
- In other cases the predictions may at the individual level can be imprecise
  - but money may still be made if predict well on average
  - e.g. a better search engine than competitors
  - e.g. a better model for predicting stock prices than competitors
  - e.g. a better model for digital ad clicks than competitors.

Econometrics using Big Data

- Economists want to estimate models that are only partially specified
  - use the machine learner in part of the analysis
  - but do valid inference controlling for the machine learning.
- For example, suppose we are interested in estimating the effect of changing x on y after controlling for everything else
  - e.g. $y = \beta_1 + \beta_2 x + (\text{many control variables}) + u$
- If we included all the control variables, the estimates get very noisy (overfitting).
- Instead use a machine learner to select a subset of the control variables.

12.8 Bayesian Methods: An Overview

- An alternative to the "classical" inference approach of this book.
- Base inference on the parameter(s) of interest θ using the **posterior distribution** which combines the distribution of y given θ with a prior distribution for θ
  - the prior can be informative or uninformative.
- One advantage is that a resulting 95% Bayesian credible region can be directly interpreted as a being an interval that θ lies in with probability 0.95.
- Rarely used until recently due to intractability.
- Recent Markov chain Monte Carlo methods (MCMC) make Bayesian methods now much easier to implement.
- In very large samples or with uninformative prior get similar results to using "classical" methods.

12.8 A Brief History of Statistics and Regression

- 1733 Central limit theorem
- 1805 Least squares (without statistical inference)
- 1885 Regression
- 1888 Correlation
- 1894 The term “standard deviation”
- 1895 Histograms
- 1908 The t distribution
- 1924 The F distribution
- 1945 ENIAC (the first electronic general purpose digital computer)
- 1964 Kernel regression (a nonparametric regression method)
- 1980’s Robust standard errors
- 1984 Apple Macintosh computer (an early personal computer).

Key Stata Commands

```
* Heteroskedastic robust standard error
use AED_HOUSE.DTA, clear
regress price size bedrooms bathroom lotsize age monthsold,
vce(robust)
* HAC standard error (for the mean)
use AED_REALGDPPC, clear
pwcorr growth l.growth l2.growth l3.growth l4.growth
l5.growth
newey growth, lag(5)
* Predict conditional mean
use AED_HOUSE.DTA, clear
regress price size
display _b[_cons] + 2000*_b[size]
* 95% conf.  interval for prediction of conditional mean
lincom _cons + 2000*size
```

Some in-class Exercises

1. Suppose $y_i = \beta_1 + \beta_2 x_i + u_i$ and u_i are independent. What standard errors would you use?
2. Suppose we have $y_{ij} = \beta_1 + \beta_2 x_{ij} + u_{ij}$, with u_{ij} correlated for individuals i in the same village j but uncorrelated for individuals in different villages. What standard errors would you use?
3. Suppose $y_t = \beta_1 + \beta_2 x_t + u_t$ and the error u_t is correlated with u_{t-1}. What standard errors would you use?
4. We obtain fitted model $\widehat{y} = \underset{(0.001)}{3.0} + \underset{(0.002)}{5.0} \times x$, $n = 200$, $s_e = 2.0$, with standard errors given in parentheses. Predict y when $x = 10$.
5. For the preceding data give an approximate 95% confidence interval for $\mathrm{E}[y|x = 10]$. Hint: how precise are the OLS estimates?
6. For the preceding data give an approximate 95% confidence interval for $y|x = 10$.