Economics 102 A01-A04: Analysis of Economic Data
Cameron Fall 2022
Department of Economics, U.C.-Davis

Final Exam (A) Thursday December 8

Compulsory. Closed book. Total of 60 points and worth 45% of course grade.
Read question carefully so you answer the question.

Question scores

Question	$1a$	$1b$	$1c$	$1d$	$1e$	$2a$	$2b$	$2c$	$2d$	$2e$	$2f$	$3a$	$3b$	$3c$	$3d$	$3e$	
Points	2	1	3	1	2	1	1	2	3	1	1	1	1	4	1	1	
Question	$4a$	$4b$	$4c$	$4d$	$4e$	$4f$	$5a$	$5b$	$6a$	$6b$	$6c$	$6d$	$7a$	$7b$	$7c$	$7d$	*Mult Choice*
Points	2	1	2	2	1	1	2	2	1	2	2	2	1	1	1	1	10

Questions 1-5
Consider data on monthly family expenditure on food and monthly family income for families in South Africa in 1993.

Dependent Variable

`food` = Monthly family expenditure on food **in South African Rand**.
`lnfood` = Natural logarithm of variable `food`.

Regressors

`income` = Annual family income **in South African Rand**.
`lnincome` = Natural logarithm of variable `income`.
`metro` = 1 if live in metro area and 0 otherwise
`urban` = 1 if live in urban area and 0 otherwise
`rural` = 1 if live in rural area and 0 otherwise
`fedu` = Father's education (highest grade attended)
`medu` = Mother's education (highest grade attended)

Note: People live in exactly one out of metro, urban or rural areas.

Use the two pages of output provided at the end of this exam on:

1. Various t critical values.
2. Various descriptive statistics output and correlations for all variables.
3. Three regressions and a test.

Part of the following questions involves deciding which output to use.
You can use the output that gets the correct answer in the quickest possible way.

1. This question uses various output given in the first page of Stata output at the end of the exam.

(a) Provide an approximate graph of the distribution of family expenditure on food.
Your graph should include an appropriate scale on the horizontal axis.

(b) Give a 95% confidence interval for the population mean family expenditure on food.

(c) Perform a test at significance level .05 of the claim that the population mean family expenditure on food is less than 700 Rand.
State clearly the null and alternative hypotheses of your test, and your conclusion.

(d) Without using any of the regression output, which variable out of `metro`, `urban`, `rural`, `fedu` and `medu` will on its own best explain food expenditure? **Explain your answer.**

(e) Suppose we OLS regress food on an intercept and `metro`. Give the intercept and slope coefficient for this regression. **Explain your answer.**

2. In this question the regression studied is a linear regression of `food` on `income`.

(a) According to the regression results, by how much does family monthly food expenditure increase in response to a 1,000 Rand increase in family annual income?

(b) Give a 95 percent confidence interval for the population slope parameter.

(c) Give a 99 percent confidence interval for the population slope parameter.

(d) Test the hypothesis at significance level 5% that the coefficient of `income` equals 0.09. State clearly the null and alternative hypothesis in terms of population parameters and your conclusion.

(e) Do you think model errors are likely to be heteroskedastic in this application?
Explain your answer.

(f) Give the exact Stata command that would provide heteroskedastic-robust standard errors for this model.

3. In this question the regression studied is a linear regression of `food` on `income`.

(a) Predict the actual food expenditures of a family with income of 1,000 Rand.

(b) Suppose we want a 95 percent confidence interval for the actual value of food expenditures of a family with income of 1,000 RAND. Will the width of this confidence interval be more or less than 500 Rand?
Give a simple answer with explanation that does not require complicated calculations.

(c) What are the four key assumptions on the model for the output of this question to have OLS coefficient estimates that are unbiased and standard error estimates that are unbiased?
[One point each assumption].

(d) In addition to these four key assumptions, do you think we need to additionally assume that the errors are normally distributed in order for the p-value for variable `income` to be exactly correct?
A simple yes or no will do.

(e) Suppose the data are for people in different villages and the model error is correlated for people in the same village but uncorrelated for people in different villages. With just this complication will the OLS coefficient estimates be unbiased?
A simple yes or no will do.

4. In this question both regressions where `food` is the dependent variable are relevant.

(a) In the second model, what is the impact on family food expenditure of a one standard deviation change in mother's education?

(b) Are income, living in a metro area or rural area, and parental education jointly statistically significant at 5 percent? **Explain your answer.**

(c) Is living in a metro, urban or rural area jointly statistically significant at the 5% level?
If there is insufficient information to answer this question then say so. **Explain your answer.**

(d) Propose a method to confirm that the three indicator variables `metro`, `rural`, and `urban` are mutually exclusive.

(e) Suppose instead of command `regress food income metro rural fedu medu` we give command `regress food income rural urban fedu medu`
Do you expect the coefficient of `income` to be 0.0871842? **Explain your answer.**

(f) Using an appropriate measure of goodness-of-fit, which model explains the data better - the second model (with five regressors) or the first model (with one regressor)? **Explain your answer.**

5. In this question consider the regression where `lnfood` is the dependent variable.
For each of the following regressors provide a meaningful interpretation of the various slope coefficients **in terms of impacts on food expenditures** (rather than on `lnfood`).

(a) Provide a meaningful interpretation of the estimated coefficient for `lnincome`.

(b) Provide a meaningful interpretation of the estimated coefficient for `medu`.

6.(a) Calculate $\sum_{i=1}^{n} z_i$ for $z_i = 1 + 2i^2$ and $n = 3$.

(b) Suppose Y_i is distributed with mean 40 and variance 100, though is not necessarily normally distributed. We obtain $10,000$ samples each of size 100 and for each sample compute the sample mean $\bar{y}$. What distribution do you expect the sample means to have? Provide the mean, standard deviation and, if appropriate, the distribution.

(c) Let X be the number of students who miss a midterm exam due to illness. Suppose $X = 1$ with probability 0.4, $X = 3$ with probability 0.2 and $X = 5$ with probability 0.4. What is the mean and variance of X? **Show all workings.**

(d) Consider a simple random sample of size 3 with values $15, 20, 40$.
Compute the sample standard deviation. **Show all workings.**

7. You are given the following partial Stata output

```
. regress y x z
      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  2,     18) =
       Model |      810                                 Prob > F      =
    Residual |                                          R-squared     =     (A)
-------------+------------------------------           Adj R-squared =
       Total |     1080                                 Root MSE      =     (B)
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          9       (C)       1.5
           z |          6       3.0             (D)
       _cons |        -12       3.0
------------------------------------------------------------------------------
```

(a) Calculate missing entry **(A)**.

(b) Calculate missing entry **(B)**.

(c) Calculate missing entry **(C)**.

(d) Give approximately missing entry (D).

Multiple choice questions (1 point each)

1. Suppose an investment yields a return of four percent a year compounded. Then the investment will have doubled in

a. between 0 and 10 years

b. between 10 and 20 years

c. between 20 and 30 years

d. more than 30 years.

2. Suppose X has mean μ and standard deviation σ. Then $Y = (X - \mu)/\sigma$ has

a. variance 1

b. mean 0

c. both a and b

d. neither a nor b.

3. The Stata command `use mydata` is used to

a. read in data from a text spreadsheet file `mydata.csv`

b. read in data from a Stata dataset `mydata.dta`

c. both a and b

d. neither a nor b.

4. Suppose we estimate a quadratic model and find $\widehat{y}_i = 1 + 3x_i + 2x_i^2$. Then the marginal effect of a change in x_i equals

a. $3x_i + 2x_i^2$

b. $3 + 4x_i$

c. 3

d. none of the above.

5. A linear regression of life expectancy (y) on an intercept and health spending as a percentage of GDP for various OECD countries finds that

a. for the U.S. $\widehat{y}$ is considerably higher than y

b. for the U.S. $\widehat{y}$ is considerably lower than y

c. for the U.S. $\widehat{y}$ is close to y

6. Let $\widehat{y}_i = b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$. Then the ordinary least squares estimator minimizes

a. $\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$

b. $\sum_{i=1}^{n} (y_i - \bar{y})^2$

c. $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$

d. none of the above

7. Suppose $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 4$, $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 10$ and $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 20$. Then the slope coefficient from OLS regression of y on an intercept and x equals

a. 0.5

b. 2

c. 2.5

d. 5

e. none of the above.

8. A type 1 error of a statistical test of H_0 against H_a occurs if we

a. reject H_0 given H_0 true

b. reject H_0 given H_a true

c. do not reject H_0 given H_0 true

d. do not reject H_0 given H_a true

9. Let d be an indicator variable for whether female. The regression model $y = \beta_1 + \beta_2 x + \beta_3 d + \beta_4 d \times x + u$ is one with:

a. different slope coefficient by gender

b. different intercept coefficient by gender

c. neither a. nor b.

d. both a. and b.

10. In linear OLS regression slope coefficients estimates become biased if

a. unnecessary (or irrelevant) regressors are included

b. important regressors are omitted

c. neither a. nor b.

d. both a. and b.

Cameron: Department of Economics, U.C.-Davis

SOME USEFUL FORMULAS

Univariate Data

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad and \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$\bar{x} \pm t_{\alpha/2;n-1} \times (s_x/\sqrt{n}) \qquad and \qquad t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$\text{ttail}(df, t) = \Pr[T > t]$ where $T \sim t(df)$

$t_{\alpha/2}$ such that $\Pr[|T| > t_{\alpha/2}] = \alpha$ is calculated using $\text{invttail}(df, \alpha/2)$.

Bivariate Data

$r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \dfrac{s_{xy}}{s_x \times s_y}$ [Here $s_{xx} = s_x^2$ and $s_{yy} = s_y^2$].

$\widehat{y} = b_1 + b_2 x_i \qquad b_2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_1 = \bar{y} - b_2\bar{x}$

$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ $\quad \text{ResidualSS} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ $\quad$ Explained SS = TSS - Residual SS

$R^2 = 1-$ ResidualSS/TSS

$b_2 \pm t_{\alpha/2;n-2} \times s_{b_2}$

$t = \dfrac{b_2 - \beta_2^*}{s_{b_2}} \qquad s_{b_2}^2 = \dfrac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$

$y|x = x^* \in b_1 + b_2 x^* \pm t_{\alpha/2;n-2} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + 1}$

$\text{E}[y|x = x^*] \in b_1 + b_2 x^* \pm t_{\alpha/2;n-2} \times s_e \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$

Multiple Regression

$\widehat{y} = b_1 + b_2 x_{2i} + \cdots + b_k x_{ki}$

$R^2 = 1-$ ResidualSS/TSS $\qquad \bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$

$b_j \pm t_{\alpha/2;n-k} \times s_{bj} \qquad and \qquad t = \dfrac{b_j - \beta_j^*}{s_{b_j}}$

$F = \dfrac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k)$

$F = \dfrac{(\text{ResSS}_r - \text{ResSS}_u)/(k-g)}{\text{ResSS}_u/(n-k)} \sim F(k-g, n-k)$

$\text{Ftail}(df1, df2, f) = \Pr[F > f]$ where F is $F(df1, df2)$ distributed.

F_α such that $\Pr[F > f_\alpha] = \alpha$ is calculated using $\text{invFtail}(df1, df2, \alpha)$.

```
KEY CRITICAL VALUES FOR THIS EXAM
t_1115,.005 = 2.580  t_1114,.005 = 2.580  t_1113,.005 = 2.580  t_1109,.005 = 2.580
t_1115,.01  = 2.330  t_1114,.01  = 2.330  t_1113,.01  = 2.330  t_1109,.01  = 2.330
t_1115,.025 = 1.962  t_1114,.025 = 1.962  t_1113,.025 = 1.962  t_1109,.025 = 1.962
t_1115,.05  = 1.646  t_1114,.05  = 1.646  t_1113,.05  = 1.646  t_1109,.05  = 1.646
t_1115,.10  = 1.282  t_1114,.10  = 1.282  t_1113,.10  = 1.282  t_1109,.10  = 1.282

. sum income lnincome food lnfood metro urban rural fedu medu

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      income |      1,115    1694.653    1649.817   13.76042     8612.5
    lnincome |      1,115    7.001659    .9995433   2.621796    9.06097
        food |      1,115    676.7994    436.4025     33.551       3775
      lnfood |      1,115    6.353889    .5743955   3.513067   8.236156
       metro |      1,115    .0618834    .2410518          0          1
-------------+---------------------------------------------------------
       urban |      1,115    .1112108     .314534          0          1
       rural |      1,115    .8269058    .3784984          0          1
        fedu |      1,115    1.403587    2.813214          0         16
        medu |      1,115    3.109417    3.598156          0         12

. mean food

Mean estimation                   Number of obs   =      1,115

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        food |   676.7994   13.06922      651.1563    702.4425
--------------------------------------------------------------

. sum food, detail

                93: HH Total Food Month Exp
-------------------------------------------------------------
      Percentiles      Smallest
 1%          137         33.551
 5%          222          65.57
10%          295         78.495       Obs               1,115
25%          415             97       Sum of Wgt.       1,115

50%          574                      Mean           676.7994
                        Largest       Std. Dev.      436.4025
75%          837           3369
90%         1147           3369       Variance       190447.2
95%         1407       3705.289       Skewness       2.671434
99%         2270           3775       Kurtosis       14.79689

. sum food if metro==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        food |         69    681.7391    327.0024    99.3875       1496

. sum food if metro==0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        food |      1,046    676.4736    442.7892     33.551       3775

. correlate income lnincome food lnfood metro urban rural fedu medu
(obs=1,115)

             |   income lnincome     food   lnfood    metro    urban    rural     fedu     medu
-------------+---------------------------------------------------------------------------------
      income |   1.0000
    lnincome |   0.8378   1.0000
        food |   0.3205   0.3084   1.0000
      lnfood |   0.3301   0.3745   0.8905   1.0000
       metro |   0.1699   0.1576   0.0029   0.0177   1.0000
       urban |   0.1587   0.1847  -0.0132  -0.0064  -0.0909   1.0000
       rural |  -0.2401  -0.2538   0.0091  -0.0060  -0.5614  -0.7731   1.0000
        fedu |   0.0981   0.1057   0.1413   0.1453  -0.0144   0.0436  -0.0271   1.0000
        medu |   0.0760   0.0872   0.1034   0.1414   0.0688   0.0931  -0.1212   0.2125   1.0000
```

```
. regress food income
```

Source	SS	df	MS		
Model	21794243.6	1	21794243.6	Number of obs	= 1,115
Residual	190363908	1,113	171036.755	F(1, 1113)	= 127.42
				Prob > F	= 0.0000
Total	212158151	1,114	190447.174	R-squared	= 0.1027
				Adj R-squared	= 0.1019
				Root MSE	= 413.57

food	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0847798	.0075105	11.29	0.000	.0700436	.0995161
_cons	533.127	17.75918	30.02	0.000	498.2817	567.9722

```
. regress food income metro rural fedu medu
```

Source	SS	df	MS		
Model	27034486.6	5	5406897.31	Number of obs	= 1,115
Residual	185123665	1,109	166928.462	F(5, 1109)	= 32.39
				Prob > F	= 0.0000
Total	212158151	1,114	190447.174	R-squared	= 0.1274
				Adj R-squared	= 0.1235
				Root MSE	= 408.57

food	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	.0871842	.0076894	11.34	0.000	.0720967	.1022717
metro	-1.965244	61.47301	-0.03	0.975	-122.5818	118.6513
rural	113.8298	39.87293	2.85	0.004	35.59489	192.0647
fedu	15.00759	4.473403	3.35	0.001	6.230304	23.78488
medu	8.470908	3.507814	2.41	0.016	1.588207	15.35361
_cons	387.6434	43.27632	8.96	0.000	302.7307	472.5561

```
. test metro rural

 ( 1)  metro = 0
 ( 2)  rural = 0

       F(  2,  1109) =    5.84
            Prob > F =    0.0030
```

```
. regress lnfood lnincome metro rural fedu medu
```

Source	SS	df	MS		
Model	62.5314393	5	12.5062879	Number of obs	= 1,115
Residual	305.01077	1,109	.275032254	F(5, 1109)	= 45.47
				Prob > F	= 0.0000
Total	367.542209	1,114	.32993017	R-squared	= 0.1701
				Adj R-squared	= 0.1664
				Root MSE	= .52444

lnfood	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnincome	.2201653	.0163539	13.46	0.000	.1880773	.2522534
metro	.0384121	.0788383	0.49	0.626	-.116277	.1931012
rural	.1745908	.0514147	3.40	0.001	.0737097	.2754719
fedu	.0176357	.0057446	3.07	0.002	.0063642	.0289072
medu	.0163607	.004504	3.63	0.000	.0075235	.025198
_cons	4.589994	.1315801	34.88	0.000	4.33182	4.848168