Economics 102_A01-A04: Analysis of Economic Data
Cameron Fall 2021 Department of Economics, U.C.-Davis

Second Midterm Exam (November 4) Version A

Compulsory. Closed book. Total of 34 points and worth 20% of course grade.
Read question carefully so you answer the question.
For computations final answers should be to at least four significant digits.

Question scores

Quest	1a	1b	1c	1d	1e	2a	2b	2c	2d	2e	2f	2g	2h	3a	3b	3c	3d	4a	4b	4c	Mult.cho
Points	3	1	1	2	1	1	1	3	3	1	1	1	1	2	2	1	1	1	1	1	5

1.(a) A very large random sample has mean 4.4 leading to a p-value of 0.03 for a two-sided test of whether $\mu = 5$. The claim is made that $\mu < 5$.
Specify the null hypothesis and alternative hypothesis for test of this claim.
State with explanation whether or not you reject the null hypothesis of your test at significance level 5%.

(b) Give one advantage of using the Stata command `lpoly y x` rather than `regress y x`.

(c) You are given the following information on a sample of size 25.
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 25$, $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 100$, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 50$.
What is the sample correlation between the two variables? Show calculations.

(d) Continuing with the same data as in part (c). What is the slope coefficient from regression of y on an intercept and x? Show calculations.

(e) Regression of y on an intercept and x yields total sum of squares equal to 200 and explained sum of squares equal to 30. Compute the value of R^2.

QUESTION 2 USES STATA OUTPUT GIVEN AT THE END OF THIS EXAM.
For some questions the answer is given directly in the output.
For other questions you will need to use the output plus additional computation.

The data are data on output (`q`) and capital (`k`) measured in millions of dollars for 42 U.S. states in 2014.

2.(a) When capital spending increases by one dollar what is the change in the value of output?

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

(c) Give a **90 percent** confidence interval for the population slope coefficient.

(d) The claim is made that increasing capital spending by $1 million is associated with a $13 million dollar increase in output. Test this claim using an appropriate statistical test at **significance level 1%. State clearly the null and alternative hypotheses of your test, and your conclusion.**

(e) Give the exact Stata command, including any relevant variables and numbers, that will compute the p-value of your test in part (d).

(f) Provide a prediction of `q` when `k` equals 3,000.

(g) Give the Stata command that will produce regression estimates for this model and heteroskedastic-robust standard errors.

(h) Give the standard error of the regression.

3.(a) Provide the four population assumptions used for the linear regression model.
(0.5 points per correct assumption).

(b) Which of the assumptions given in part (a) are necessary for the OLS estimates to be unbiased?

(c) You are given the following Stata code.

```
clear
 quietly set obs 80
 generate x = rnormal(5,16)
 generate u = x*rnormal(0,9)
 generate y = 4 + 8*x + u
 regress y x
```

Suppose you run this code 5,000 times leading to 5,000 estimates of the slope coefficient. What do you expect the average of these 5,000 slopes to approximately equal?

(d) Continuing with the code in part (c). For inference would it be correct to use default standard errors or do we instead need to use heteroskedastic-robust standard errors? **Explain your answer**.

4. Suppose we know that $y = 9 + 3x + u$ where $E[u|x] = 0$. A sample gives estimates $\widehat{y} = 10 + 2x$. Consider the observation $(x, y) = (3, 20)$. **Show any workings.**

(a) What is the value of the error for this observation?

(b) What is the predicted value for this observation?

(c) What is the value of the residual for this observation?

Multiple Choice Questions (1 point each)

1. For linear regression the conditional mean of y given x is

a. $b_1 + b_2 x + e$

b. $b_1 + b_2 x$

c. $\beta_1 + \beta_2 x + u$

d. $\beta_1 + \beta_2 x$

e. none of the above.

2. The total sum of squares equals

a. the sum of the squared deviations of y from the fitted value of y

b. the sum of the squared deviations of y from the sample average value of y

c. the sum of the squared deviations of the fitted value of y from the average value of y

d. none of the above.

3. A heteroskedastic-robust standard error for regression is used when

a. errors are homoskedastic and not independent

b. errors are homoskedastic and independent

c. errors are heteroskedastic and not independent

d. errors are heteroskedastic and independent

4. Regression of life expectancy on health spending as a percentage of GDP for a number of countries obtains fitted line that is

a. declining with health spending and the U.S. is below the line

b. declining with health spending and the U.S. is above the line

c. increasing with health spending and the U.S. is above the line

d. increasing with health spending and the U.S. is below the line

5. We regress y on x and find that $b_2 = 10$ with standard error 2. Given only this information

a. the regressor x is highly statistically significant

b. the regressor x is highly economically significant

c. the regressor x is highly statistically significant and highly economically significant

d. none of the above.

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

SOME USEFUL FORMULAS FOR EXAMS

Univariate Data

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

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

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

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

Bivariate Data

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

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

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

$R^2 = 1-$ ResidualSS/TSS

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

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

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

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

Multivariate Data

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

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

$b_j \pm t_{\alpha/2;n-k} \times s_{bj}$ *and* $t = \dfrac{b_j - \beta_{j0}}{s_{b_j}}$

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

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

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

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

```
. summarize q k

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
           q |         42    45094.39    45866.37   1014.207   245681.4
           k |         42     3073.95    2902.534    102.491   14057.42

. regress q k

      Source |       SS           df       MS      Number of obs   =        42
-------------+----------------------------------   F(1, 40)        =    282.90
       Model |  7.5568e+10         1  7.5568e+10   Prob > F        =    0.0000
    Residual |  1.0685e+10        40   267122708   R-squared       =    0.8761
-------------+----------------------------------   Adj R-squared   =    0.8730
       Total |  8.6253e+10        41  2.1037e+09   Root MSE        =     16344

------------------------------------------------------------------------------
           q | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           k |   14.79105   .8793993    16.82   0.000     13.01372    16.56838
       _cons |  -372.5542    3696.96    -0.10   0.920    -7844.389    7099.281
------------------------------------------------------------------------------

. di _n "Square Root of " N " = " sqrt(N) _n   ///
>       "Square Root of " N-1 " = " sqrt(N-1) _n ///
>           "Square Root of " N-2 " = " sqrt(N-2) _n

Square Root of 42 = 6.4807407
Square Root of 41 = 6.4031242
Square Root of 40 = 6.3245553


. di _n "KEY CRITICAL VALUES FOR THIS EXAM" _n _n       ///
>       "t_" dof ",.005 = " %5.3f invttail(dof,.005) _n ///
>       "t_" dof ",.01  = " %5.3f invttail(dof,.01) _n  ///
>       "t_" dof ",.025 = " %5.3f invttail(dof,.025) _n ///
>       "t_" dof ",.05  = " %5.3f invttail(dof,.05) _n  ///
>       "t_" dof ",.10  = " %5.3f invttail(dof,.10) _n

KEY CRITICAL VALUES FOR THIS EXAM

t_40,.005 = 2.704
t_40,.01  = 2.423
t_40,.025 = 2.021
t_40,.05  = 1.684
t_40,.10  = 1.303
```