Economics 102_A: Analysis of Economic Data
Cameron Winter 2019
Department of Economics, U.C.-Davis

Final Exam (A) Thursday March 21

Compulsory. Closed book. Total of 62 points and worth 45% of course grade.
Read question carefully so you answer the question.

Question scores

Question	$1a$	$1b$	$1c$	$1d$	$1e$	$2a$	$2b$	$2c$	$2d$	$2e$	$2f$	$3a$	$3b$	$3c$	$3d$	$3e$	$3f$
Points	1	3	2	2	2	1	2	3	1	1	1	1	2	1	1	2	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	2	1	1	10

Questions 1-5
Consider data on annual outpatient health expenditures for individuals in the U.S.
Outpatient spending is for health services that did not require admission to a hospital.
The data come from the Rand health insurance experiment where different individuals were assigned to one of four different levels of health insurance - see variables `coins0`, `coins25`, `coins50` and `coins95` below.

Dependent Variable

`outspend` = Annual outpatient health expenditures (paid through insurance or out-of-pocket)
`lnout` = Natural logarithm of variable `outspend`

Regressors

`age` = age in years
`lnage` = Natural logarithm of variable `age`
`education` = years of schooling
`coins0` = 1 if individual's share of health spending is 0% (free) and = 0 otherwise
`coins25` = 1 if individuals share of health spending is 25% and = 0 otherwise
`coins50` = 1 if individuals share of health spending is 50% and = 0 otherwise
`coins95` = 1 if individuals share of health spending is 95% and = 0 otherwise

Note: People have exactly one of 0%, 25%, 50% and 95% coinsurance.

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.(a) Give a 95% confidence interval for the population mean outpatient expenditures.

(b) Perform a test at significance level .05 of the claim that the population mean of outpatient expenditures is less than $1,700.
State clearly the null and alternative hypotheses of your test, and your conclusion.

(c) Without using any of the regression output, which two variables (out of `age`, `education`, `coins0`, `coins25`, `coins50` and `coins95`) do you think will best explain outpatient spending? **Explain your answer.**

(d) Suppose we OLS regress `outspend` on an intercept and `coins0`. Give the intercept and slope coefficient for this regression. **Explain your answer.**

(e) Provide an approximate graph of the distribution of the natural logarithm of outpatient spending. **Note - this is for the natural logarithm.**
Your graph should include an appropriate scale on the horizontal axis.

2. In this question the regression studied is a linear regression of `outspend` on `age`.

(a) According to the regression results, by how much does outpatient spending increase in response to ten years of aging?

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

(c) Test the hypothesis at significance level 10% that outpatient spending increases by $15 with each year of aging. State clearly the null and alternative hypothesis in terms of population parameters and your conclusion.

(d) What is the sample correlation coefficient between age and outpatient spending? **Explain your answer.**

(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 `outspend` on `age`.

(a) Predict the actual outpatient spending for someone aged 50 years.

(b) Give a 95 percent confidence interval for the conditional mean of outpatient spending for someone aged 50 years. Note that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 248{,}054$ for the variable `age`.
Give your answer as an expression involving numbers only - you do not need to complete all calculations.

(c) Using relevant output show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 248{,}054$ for `age`.

(d) Suppose we instead want a 95 percent confidence interval for the actual value of outpatient spending for someone aged 50 years. Will the width of this confidence interval be more or less than $4,000?
Give a simple answer that does not require complicated calculations.

(e) 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?
[Half point off for each assumption missing].

(f) In addition to these four key assumptions, do you think we need to additionally assume that the errors are normally distributed in order for your answer in part (b) to be exactly correct?
A simple yes or no will do.

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

(a) In the second model, based on its coefficient (and not on statistical significance) is education an important determinant of outpatient spending? **Explain your answer.**

(b) Are age, education and the health insurance indicator variables jointly statistically significant at 5 percent? **Explain your answer.**

(c) Are the four levels of health insurance jointly statistically significant at the 5% level? If there is insufficient information to answer this question then say so. **Explain your answer.**

(d) Suppose we give command `generate total = coins0+coins25+coins50+coins95`
Give the sample mean and standard deviation of variable `total`.

(e) Suppose we give command `regress outspend age education coins0 coins25 coins50`
Do you expect the coefficient of `age` to be 10.36179? **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 `lnout` is the dependent variable.
For each of the following regressors provide a meaningful interpretation of the various slope coefficients **in terms of impacts on outpatient spending** (rather than on `lnout`).

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

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

6.(a) Calculate $\sum_{i=1}^{n} z_i$ for $z_i = 3 + 2i$ and $n = 4$.

(b) Suppose for $X \sim (500,\ 80^2)$ we form 900 samples of size 400 and obtain 900 sample means $\bar{x}$.
What approximately do you expect the average of the $\bar{x}$ to equal? **Explain.**
What approximately do you expect the standard deviation of the $\bar{x}$ to equal? **Explain.**

(c) Let X be the number of students who miss a midterm exam due to illness. Suppose $X = 1$ with probability 0.4, $X = 2$ with probability 0.5 and $X = 6$ with probability 0.1. What is the variance of X? **Show all workings.**

(d) Consider a simple random sample of size 3 with values $5, 10, 30$.
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) =     (B)
       Model |      270                                Prob > F      =
    Residual |       90                                R-squared     =
-------------+------------------------------           Adj R-squared =
       Total |                                         Root MSE      =     (A)
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |         6        (C)       1.5
           z |         4        2.0
       _cons |        -8        2.0
------------------------------------------------------------------------------
```

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

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

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

(d) Suppose we perform an F test of $H_0 : \beta_z = 0$ against $H_a : \beta_z \neq 0$. What will the value of the F statistic be?

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 we want to transform a random variable $X \sim (\mu, \sigma)$ to a variable Y with mean 0 and standard deviation one. Then

a. $Y = (X - \mu)/\sigma$

b. $Y = (X - \mu)/\sigma^2$

c. $Y = (X - \mu)^2/\sigma$

d. $Y = (X - \mu)^2/\sigma^2$

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 an Excel formatted spreadsheet file `mydata.xlsx`

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

d. none of the above.

4. The Stata command to find the p-value for a two-tailed t-test in a regression with three regressors (including the intercept) when $t = 1.57$ and there are 40 observations is

a. 2×`invttail(1.57,37)`

b. 2×`invttail(37,1.57)`

c. 2×`ttail(37,1.57)`

d. 2×`ttail(1.57,37)`

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

a. $2x_i + x_i^2$

b. $2 + 2x_i$

c. 2

d. none of the above.

6. A linear regression of the academic performance index (y) for California high schools on various regressors finds that the most important explanators are

a. socioeconomic background measures such as fraction of students eligible for free meals

b. teacher quality measures such as whether they have teaching credentials

c. average educational attainment of parents

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 equals

a. 0.5

b. 2

c. 2.5

d. 5

e. none of the above.

8. We obtain OLS estimate $\widehat{y} = 2 + 5d$ where d is an indicator variable taking values 0 or 1. Then

a. The mean of $y = 7$ for those observations with $d = 0$

b. The mean of $y = 2$ for those observations with $d = 1$

c. both a. and b.

d. neither a. nor b.

9. 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.

10. Multicollinearity is a problem that arises when

a. the dependent variable is highly correlated with the regressors

b. the error is highly correlated with the regressors

c. the error is highly correlated with the dependent variable

d. the regressors are highly correlated with each other

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

SOME USEFUL FORMULAS

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* $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

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 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 $s_{yy} = s_y^2$].

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

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

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

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

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

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

Multiple Regression

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

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

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

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

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

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$(df1, df2, \alpha)$.

KEY CRITICAL VALUES FOR THIS EXAM

```
t_1950,.005 = 2.578  t_1949,.005 = 2.578  t_1948,.005 = 2.578  t_1944,.005 = 2.578
t_1950,.01  = 2.328  t_1949,.01  = 2.328  t_1948,.01  = 2.328  t_1944,.01  = 2.328
t_1950,.025 = 1.961  t_1949,.025 = 1.961  t_1948,.025 = 1.961  t_1944,.025 = 1.961
t_1950,.05  = 1.646  t_1949,.05  = 1.646  t_1948,.05  = 1.646  t_1944,.05  = 1.646
t_1950,.10  = 1.282  t_1949,.10  = 1.282  t_1948,.10  = 1.282  t_1944,.10  = 1.282
```

. sum outspend lnout age lnage education coins0 coins25 coins50 coins95

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	1,950	1629.187	2256.877	5.291	21107.11
lnout	1,950	6.679373	1.27082	1.666007	9.957365
age	1,950	37.77128	11.28151	21	62
lnage	1,950	3.587219	.2981845	3.044523	4.127134
education	1,950	12.28769	3.066068	0	25
coins0	1,950	.4558974	.4981789	0	1
coins25	1,950	.2512821	.4338616	0	1
coins50	1,950	.0835897	.2768426	0	1
coins95	1,950	.2092308	.4068638	0	1

. mean outspend

Mean estimation   Number of obs   =   1,950

	Mean	Std. Err.	[95% Conf.	Interval]
outspend	1629.187	51.10819	1528.954	1729.419

. sum outspend if coins0==1

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	889	1883.468	2287.072	5.291	19134.94

. sum outspend if coins0==0

Variable	Obs	Mean	Std. Dev.	Min	Max
outspend	1,061	1416.127	2209.907	9.570001	21107.11

. correlate outspend lnout age lnage education coins0 coins25 coins50 coins95
(obs=1,950)

	outspend	lnout	age	lnage	educat~n	coins0	coins25	coins50	coins95
outspend	1.0000								
lnout	0.7624	1.0000							
age	0.0402	0.0897	1.0000						
lnage	0.0437	0.0904	0.9907	1.0000					
education	0.0394	0.0590	-0.2123	-0.2002	1.0000				
coins0	0.1032	0.1786	0.0054	-0.0012	-0.0543	1.0000			
coins25	-0.0495	-0.0576	0.0059	0.0109	0.0343	-0.5303	1.0000		
coins50	-0.0521	-0.0858	0.0042	0.0049	0.0176	-0.2765	-0.1750	1.0000	
coins95	-0.0380	-0.0988	-0.0157	-0.0134	0.0179	-0.4708	-0.2980	-0.1554	1.0000

. sum lnout, detail

lnout

	Percentiles	Smallest		
1%	3.643359	1.666007		
5%	4.461069	2.258633		
10%	4.973298	2.464704	Obs	1,950
25%	5.864171	2.496917	Sum of Wgt.	1,950
50%	6.756061		Mean	6.679373
		Largest	Std. Dev.	1.27082
75%	7.572927	9.777963		
90%	8.27102	9.824061	Variance	1.614984
95%	8.692378	9.859271	Skewness	-.2593892
99%	9.372543	9.957365	Kurtosis	2.966688

```
. regress outspend age

      Source |       SS           df       MS      Number of obs   =     1,950
-------------+----------------------------------   F(1, 1948)      =      3.16
       Model |    16055842         1    16055842   Prob > F        =    0.0758
    Residual |  9.9112e+09     1,948  5087864.54   R-squared       =    0.0016
-------------+----------------------------------   Adj R-squared   =    0.0011
       Total |  9.9272e+09     1,949  5093492.03   Root MSE        =    2255.6

------------------------------------------------------------------------------
    outspend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   8.045322    4.52892     1.78   0.076    -.8367169    16.92736
       _cons |   1325.305   178.5266     7.42   0.000     975.1813    1675.428
------------------------------------------------------------------------------

. regress outspend age education coins25 coins50 coins95

      Source |       SS           df       MS      Number of obs   =     1,950
-------------+----------------------------------   F(5, 1944)      =      6.26
       Model |   157414231         5  31482846.2   Prob > F        =    0.0000
    Residual |  9.7698e+09     1,944  5025618.18   R-squared       =    0.0159
-------------+----------------------------------   Adj R-squared   =    0.0133
       Total |  9.9272e+09     1,949  5093492.03   Root MSE        =    2241.8

------------------------------------------------------------------------------
    outspend |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   10.36179   4.606929     2.25   0.025      1.32675    19.39683
   education |   41.37921   16.97463     2.44   0.015     8.088812     74.6696
     coins25 |  -462.7932   126.2862    -3.66   0.000    -710.4638   -215.1226
     coins50 |   -659.371    191.112    -3.45   0.001    -1034.177    -284.565
     coins95 |   -428.764   134.1436    -3.20   0.001    -691.8443   -165.6837
       _cons |   990.4725    305.762     3.24   0.001     390.8168    1590.128
------------------------------------------------------------------------------

. test coins25 coins50 coins95

 ( 1)  coins25 = 0
 ( 2)  coins50 = 0
 ( 3)  coins95 = 0

       F(  3,  1944) =    7.79
            Prob > F =    0.0000

. regress lnout lnage education coins25 coins50 coins95

      Source |       SS           df       MS      Number of obs   =     1,950
-------------+----------------------------------   F(5, 1944)      =     20.59
       Model |  158.317781         5  31.6635562   Prob > F        =    0.0000
    Residual |  2989.28621     1,944  1.53769867   R-squared       =    0.0503
-------------+----------------------------------   Adj R-squared   =    0.0479
       Total |  3147.60399     1,949  1.61498409   Root MSE        =      1.24

------------------------------------------------------------------------------
       lnout |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnage |   .4622027   .0961673     4.81   0.000     .2736008    .6508045
   education |   .0375138   .0093656     4.01   0.000     .0191461    .0558814
     coins25 |  -.3907023   .0698611    -5.59   0.000    -.5277129   -.2536918
     coins50 |  -.6245889   .1057155    -5.91   0.000    -.8319164   -.4172613
     coins95 |  -.4994225   .0741998    -6.73   0.000    -.6449421   -.3539029
       _cons |   4.815274   .3863439    12.46   0.000     4.057582    5.572966
------------------------------------------------------------------------------
```