ECN 102: Analysis of Economics Data

Midterm Exam SS1 2026

Remy Beauregard, Department of Economics, UC Davis

This exam consists of 5 short-answer questions (with sub-parts) and 5 multiple choice questions. You will have a maximum of 100 minutes to complete this exam without prior accommodation.

Show all work to receive full credit. You may use only the calculators provided by the instructor. For questions requiring computation, it suffices to express final answers with 2 decimal places.

There is a formula sheet and scratch paper provided at the end of this exam. You may remove this page and discard it after the exam. If you plan to use this extra page for your final answers, please write your name and student ID at the top to ensure it is not lost.

This exam is worth 70 points.

Name:

Student ID:

Question 1: Summary Statistics [15 points]

A city transportation survey records the number of minutes each commuter spent traveling to work on a given day. Nearly all commuters report travel times between 30 and 45 minutes, but a small number of individuals report commutes longer than 2 hours. All values have been verified to be correct and should not be removed from the sample.

a) Which statistic of **central tendency** should we use for this data? Why? [3 points]

b) Which statistic of **spread** should we use for this data? Why? [3 points]

c) What can we infer about the **skewness** of this dataset from the description above? What can we infer about the **mean** of the data relative to the **median**? [3 points]

d) Can we expect this dataset to be **normally distributed** based on the information above? Please justify your answer. [3 points]

e) Suppose we **log-transform** the commute times and find that the resulting distribution appears approximately normal. How should we describe the original (untransformed) distribution of commute times? Would such a log transformation be appropriate if a small number of individuals had instead reported commute times shorter than 5 minutes? Why or why not? [3 points]

Question 2: Summation Notation [6 points]

Please calculate the following sums:

a) [3 points]

$$\frac{1}{4}\sum_{i=1}^{4}(i+3)$$

b) [3 points]

$$\sum_{i=2}^{4}(i^2-2)$$

Question 3: Correlation [10 points]

We observe the following Stata output for a dataset of automobile characteristics:

```
corr price mpg weight
```

```
(obs=74)

             |    price      mpg   weight
-------------+---------------------------
       price |   1.0000
         mpg |  -0.4686   1.0000
      weight |   0.5386  -0.8072   1.0000
```

a) Which pair of variables has the strongest **linear relationship**? Which pair has the weakest? Explain your answer. [2 points]

b) Compute R^2 for a regression of *price* on *mpg*. Interpret this value in words. [4 points]

c) If we **standardized** both *price* and *mpg* before running a regression of *price* on *mpg*, what would b_2 be equal to? Explain your answer. [4 points]

Question 4: Univariate Inference [11 points]

We are given the following Stata output:

```
invttail(59,0.025)= 2.0009954
invttail(59,0.05)= 1.671093
invttail(58,0.025)= 2.0017175
invttail(58,0.05)= 1.6715528
invttail(57,0.025)= 2.0024655
invttail(57,0.05)= 1.6720289
```

Suppose we then use the `summarize` command for the three variables below. (Note: The variables *price* and *weight* have been transformed to represent units of thousands)

```
(1978 automobile data)
```

Variable	Obs	Mean	Std. dev.	Min	Max
price	59	5.923729	2.597975	3.291	13.594
weight	59	2.920169	.7739409	1.76	4.84
mpg	59	21.72881	6.056716	12	41

a) Which of these variables has the greatest **dispersion**? Which has the least? [3 points]

b) Provide a **95% CI** for the mean of *mpg*. Interpret your answer. [3 points]

c) The claim is made that the average weight of a car in 1978 was above 2.5 thousand lbs. Using your sample data, test this claim at the 5% significance level. Clearly state your **null and alternate hypotheses**, your **test statistic**, and your **conclusion**. [5 points]

Question 5: Bivariate Regression [18 points]

a) Write down the **population model** for a linear regression of y on x. [2 points]

b) What are the **four key assumptions** of our population bivariate regression model? Name and explain each briefly. [4 points]

c) Assuming our four population assumptions hold and our sample is large enough, how would we expect b_2 to be distributed? Your answer should take the form $b_2 \sim$ ___(___, ___). Use Greek letters when appropriate. [3 points]

Suppose we now run a regression using our automobile dataset and obtain these results:

```
      Source |       SS           df       MS      Number of obs   =        59
-------------+----------------------------------   F(1, 57)        =     99.74
       Model |  1353.89384         1  1353.89384   Prob > F        =    0.0000
    Residual |  773.767176        57  13.5748627   R-squared       =    0.6363
-------------+----------------------------------   Adj R-squared   =    0.6299
       Total |  2127.66102        58  36.6838106   Root MSE        =    3.6844

------------------------------------------------------------------------------
         mpg | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      weight |    -6.2427     0.6251    -9.99   0.000      -7.4944     -4.9909
       _cons |    39.9585     1.8874    21.17   0.000      36.1791     43.7378
------------------------------------------------------------------------------
```

d) Write down the **sample regression model** we have estimated (with proper variable names). [2 points]

e) What is the name of the test corresponding to the default t-statistic and p-value for b_2? State its **null and alternate hypotheses**. [3 points]

f) What is the **conclusion** of this test at $\alpha = 5\%$? How many degrees of freedom does the test have? [2 points]

g) What would we **predict** the *mpg* to be for a car weighing 3.0 thousand lbs? [2 points]

Multiple Choice [2 points each]

Only one answer is correct for each question. Choose the best possible answer.

MC 1

A public health department tracks the number of flu cases reported in Sacramento County each week for two years, recording each weekly total in a spreadsheet.

This data will be:

a) Experimental, categorical, panel

b) Observational, numerical, cross-section

c) Observational, numerical, time-series

d) Experimental, numerical, repeated cross-section

e) None of the above

MC 2

If a data point has a positive **error term**, this means:

a) the point lies above the sample regression line

b) the point lies below the sample regression line

c) the point lies above the population regression line

d) the point lies below the population regression line

MC 3

Suppose we have collected data on the height of each student in a class, but realize everyone has reported **half** their true height. When we correct our data for this mistake, which of the following will be true about our variable *height*?

a) The mean and standard deviation of *height* will not change

b) The mean of *height* will double but the standard deviation will remain unchanged

c) The mean of *height* will double and the standard deviation will increase by a factor of $\sqrt{2}$

d) The mean of *height* will double and the variance will increase by a factor of 4 (quadruple)

e) None of the above

MC 4

Which is true about the **central limit theorem**?

a) The distribution of a r.v. X will become normal as $n \to \infty$

b) The distribution of sample means $\bar{X}$ will become normal as $n \to \infty$ but only if X itself is normally distributed

c) The distribution of sample means $\bar{X}$ will become normal as $n \to \infty$ even if X itself is not normally distributed

d) The distribution of sample means $\bar{X}$ will become normal as $n \to \infty$ but only if X itself is not normally distributed

e) None of the above

MC 5

For a normal distribution, which of the following is true?

a) 34% of the distribution lies between -2σ and -1σ

b) 34% of the distribution lies between μ and 2σ

c) 81.5% of the distribution lies between -2σ and 1σ

d) 81.5% of the distribution lies between -2σ and 2σ

e) None of the above

Formula sheet

Univariate Data

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

$$\bar{x} \pm t^*_{\alpha/2;n-1} \times (s_x/\sqrt{n}) \quad t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

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

$$invttail(df, p) \rightarrow t^* : P[T > t^*] = p \text{ where } T \sim T(df)$$

Bivariate Data

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x \times s_y}$$

$$\hat{y} = b_1 + b_2 x$$

$$b_1 = \bar{y} - b_2\bar{x}$$

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_y}{s_x}$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad ResSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad ExpSS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{ExpSS}{TSS} = 1 - \frac{ResSS}{TSS}$$

$$b_2 \pm t^*_{\alpha/2;n-2} \times s_{b_2} \quad t = \frac{b_2 - \beta_2}{s_{b_2}}$$

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

Scratch paper