ECN 102: Analysis of Economics Data

Final 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 these pages and discard them after the exam. If you plan to use these extra pages for your final answers, please write your name and student ID at the top to ensure they are not lost.

This exam is worth 100 points.

Name:

Student ID:

For this exam, we will use the 1978 automotive dataset from Stata. The variables present in the dataset are listed here.

```
Contains data from /Applications/Stata/ado/base/a/auto.dta
 Observations:            74                  1978 automobile data
    Variables:            12                  13 Apr 2022 17:45
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
make            str18   %-18s                 Make and model
price           int     %8.0gc                Price ($)
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear ratio
foreign         byte    %8.0g      origin     Car origin
-------------------------------------------------------------------------------
Sorted by: foreign
     Note: Dataset has changed since last saved.
```

Question 1: Summary Statistics [16 points]

a) Is this sample of data likely to be **observational** or **experimental**? Explain. [2 points]

b) Is this sample of data likely to be **cross-section**, **time series**, **panel**, or **repeated cross-section**? Explain. [2 points]

c) Suppose we run `summarize headroom, detail` and find $kurtosis = 2.2$. Interpret this value in words. [2 points]

d) What type of **variable** is *foreign* if we include it in a regression? Why? [2 points]

e) Which of the variables below has the greatest **dispersion**? The least? [3 points]

Variable	Obs	Mean	Std. dev.	Min	Max
price	74	6165.257	2949.496	3291	15906
weight	74	3019.459	777.1936	1760	4840
mpg	74	21.2973	5.785503	12	41

f) What is the **standard error** of *price*? What is this an estimator for? [3 points]

g) What is the meaning of an observation with `foreign==0` in this dataset? [2 points]

Question 2: Standard Errors [15 points]

Suppose we are considering different choices of standard errors for OLS regression.

a) Which OLS assumption(s) are we concerned may be violated if we are considering using **robust** standard errors? State the assumption(s) precisely in math **and** words. [2 points]

b) If we do switch from default standard errors to robust standard errors, how would our **regression coefficients** change? Explain. [2 points]

c) If we do switch from default to robust standard errors, would our **confidence intervals** generally get wider, narrower, or stay the same? Explain. [3 points]

d) In what setting should we use **clustered** standard errors? Give a specific economic example and explain which assumption(s) are being addressed. [4 points]

e) If we knew that errors were correlated **across** clusters but not **within** them, would clustered standard errors be appropriate? Explain. [2 points]

f) Under what circumstance(s) should we use **bootstrapped** standard errors? [2 points]

Question 3: Multivariate Regression [23 points]

We obtain the following results after running `gen mpgXforeign = mpg * foreign`:

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =      9.48
       Model |   183435281         3  61145093.6   Prob > F        =    0.0000
    Residual |   451630115        70  6451858.79   R-squared       =    0.2888
-------------+----------------------------------   Adj R-squared   =    0.2584
       Total |   635065396        73  8699525.97   Root MSE        =    2540.1

------------------------------------------------------------------------------
       price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         mpg |  -329.2551    74.9854    -4.39   0.000    -478.8088   -179.7013
     foreign |   -13.5874  2634.6636    -0.01   0.996    -5.27e+03   5241.0835
 mpgXforeign |    78.8883   112.4812     0.70   0.485    -145.4485    303.2250
       _cons |   1.26e+04  1527.8882     8.25   0.000    9553.2609    1.56e+04
------------------------------------------------------------------------------
```

(*Hint: review variable descriptions on page 2*)

a) Write down the **population model** with proper variable names. [2 points]

b) Interpret the coefficient b_2 on *mpg* in words. (*Hint: think about what* `foreign==0` *means.*) [2 points]

c) Interpret the coefficient a_1 on *foreign* in words. [2 points]

d) Interpret the coefficient a_2 on *mpgXforeign* in words. [2 points]

e) What conclusion can we draw from the reported F-statistic and `Prob>F` for the test of overall significance? What does this tell us about $F^*_{3,70,0.05}$? [3 points]

f) Which of our regression coefficients are statistically different from zero at $\alpha = 5\%$? Which are not? What is the name for these tests? [3 points]

g) Write down the fitted regression equation for **domestic** cars only. Write down the fitted regression equation for **foreign** cars only. [3 points]

h) Find an expression for the **marginal effect** of *mpg* in this regression. [3 points]

i) Compute $ResSS$ for a regression of *price* on just a constant term. [3 points]

Question 4: OLS and Logs [18 points]

Suppose we obtain the following results after running `gen ln_price = ln(price)`:

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     22.87
       Model |  2.70578153         1  2.70578153   Prob > F        =    0.0000
    Residual |  8.51775155        72  .118302105   R-squared       =    0.2411
-------------+----------------------------------   Adj R-squared   =    0.2305
       Total |  11.2235331        73  .153747029   Root MSE        =    .34395

------------------------------------------------------------------------------
    ln_price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         mpg |    -0.0333    0.0070    -4.78   0.000      -0.0471     -0.0194
       _cons |     9.3493    0.1535    60.91   0.000       9.0434      9.6553
------------------------------------------------------------------------------
```

a) What name do we give this type of regression? [2 points]

b) How should we interpret the b_2 coefficient from this regression in words? [3 points]

c) Write an expression for the **marginal effect** of *mpg* on *price* in this regression. [3 points]

d) Find the **MEM** of *mpg* if $\overline{price} = 6165.26$. [3 points]

Suppose we obtain the following results after also running `gen ln_mpg = ln(mpg)`:

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     31.00
       Model |  3.37819527         1  3.37819527   Prob > F        =    0.0000
    Residual |  7.84533782        72  .108963025   R-squared       =    0.3010
-------------+----------------------------------   Adj R-squared   =    0.2913
       Total |  11.2235331        73  .153747029   Root MSE        =     .3301

------------------------------------------------------------------------------
    ln_price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      ln_mpg |    -0.8268     0.1485    -5.57   0.000      -1.1229     -0.5308
       _cons |    11.1415     0.4508    24.72   0.000      10.2429     12.0401
------------------------------------------------------------------------------
```

e) What name do we give this new type of regression? [2 points]

f) How should we interpret the b_2 coefficient from this regression in words? [3 points]

g) Find the **MER** of *mpg* at $mpg^* = 25$ and $price^* = \$4{,}500$. [2 points]

Question 5: F-Testing [18 points]

Refer to the regression from Question 3. Suppose we want to determine whether the **variable** country of origin has **any effect** on a car's *price*.

a) Write down the **null and alternate hypotheses** for this test. How many linear restrictions (q) would such an F-test impose? [3 points]

b) Write down the **unrestricted** model and **restricted** model for this test. [4 points]

c) Suppose we conduct the test above in Stata and find `Prob > F = 0.0387`. What should we conclude about country of origin from this result? [2 points]

d) How is this test different from the default t-test of **regressor** *foreign*? [3 points]

e) Now consider a **new, hypothetical regression** with $n = 74$. A joint F-test on this regression yields $ResSS_r = 480{,}000{,}000$ and $ResSS_u = 420{,}000{,}000$, where $k = 4$ and $q = 2$. Compute the **F-statistic**. (*Hint: you may cancel out the same number of trailing zeros from both ResSS values before computing; this does not affect the result.*) [4 points]

f) Using your F-statistic from part (e), identify the **correct critical value** at $\alpha = 5\%$ from the Stata output below, and state whether you **reject or fail to reject** H_0. [2 points]

```
invFtail(1,70,0.05) = 3.9778
invFtail(2,70,0.05) = 3.1277
invFtail(2,70,0.025) = 3.8903
invFtail(3,70,0.05) = 2.7355
```

Multiple Choice [2 points each]

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

MC 1

What is true about the chance of a **Type I error** when conducting multiple hypothesis tests?

a) The total chance of a Type I error decreases with the number of tests

b) The total chance of a Type I error increases with the number of tests

c) The total chance of a Type I error is always equal to α

d) None of the above

MC 2

What is true about the **power** of a test ($1 - P$[Type II]) when conducting inference?

a) We should only run tests with a power of 1 to avoid any chance of a Type II error

b) Power will decrease as we increase the sample size n while keeping all regressors fixed

c) Power will increase when the true population parameter is closer to the null value

d) The OLS estimator already maximizes power among all linear unbiased estimators

MC 3

Suppose we have data on **vehicle body type** which takes exactly three mutually exclusive and exhaustive values: sedan, hatchback, and wagon. We want to include body type in a regression that **also has a constant term** so create a dummy variable for each type: d_{sed}, d_{hatch}, and d_{wag}. Which of the following is true?

a) We can safely include all three dummy variables along with the constant

b) We can include at most two of the three dummy variables along with the constant

c) We must include all three dummy variables and the constant to obtain unbiased estimates

d) We must encode the three body types (1, 2, 3) and include them as a single variable

e) None of the above

MC 4

What would be a valid way to increase the **precision** of our OLS estimator b_j on a regressor x_j in a multivariate regression with $j = 2, \ldots, k$?

a) Introduce a new regressor x_{k+1} that is correlated with y but uncorrelated with x_j and all other regressors (keeping sample size fixed)

b) Introduce a new regressor x_{k+1} that is correlated with x_j but uncorrelated with y and all other regressors (keeping sample size fixed)

c) Remove regressors from the model that are correlated with y but uncorrelated with x_j (keeping sample size fixed)

d) Decrease the sample size n while keeping all regressors fixed

e) None of the above

MC 5

In which of the following scenarios did we make a **mistake**?

a) We claimed that adding a new regressor to a model can never decrease R^2 but can decrease $\bar{R}^2$

b) We claimed that b_j from a multivariate regression of y on $x_2, \ldots, x_k$ will be equal to b_2 from a bivariate regression of y on x_j if x_j is uncorrelated with all other regressors

c) We claimed that heteroskedasticity in our data is problematic because it causes our OLS coefficient estimates to be biased

d) We claimed that for a single-restriction F-test of $\beta_j = 0$, the F-statistic would equal the square of the t-statistic from the test of association for x_j on y

e) None of the above

(Do not miss the BONUS on the next page!)

BONUS [up to 10 extra points]

In order to prove that $\bar{X} \sim (\mu, \sigma^2/n)$ we needed to make **three assumptions** about our random variable X. Please name each of these three assumptions and describe in detail what they mean for X (e.g. what it would mean for them to be true versus not true). You may find it helpful to use a specific example of a r.v. X in your answer, such as a coin toss or die roll.

*Note: you do **not** need to replicate the proofs for* $\mu_{\bar{X}}, \sigma^2_{\bar{X}}$

Scratch paper

Formula sheet

Univariate Data

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

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

$$ttail(df, t) = P[T > t] \quad invttail(df, p) \to t^* : P[T > t^*] = p$$

Bivariate Data

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \hat{y} = b_1 + b_2 x \quad b_1 = \bar{y} - b_2\bar{x} \quad b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_y}{s_x}$$

$$TSS = \sum(y_i - \bar{y})^2 \quad ResSS = \sum(y_i - \hat{y}_i)^2 \quad R^2 = 1 - \frac{ResSS}{TSS}$$

$$b_2 \pm t^*_{\alpha/2;n-2} \times s_{b_2} \quad t = \frac{b_2 - \beta_2}{s_{b_2}} \quad s_{b_2} = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} \quad s_e = \sqrt{\frac{ResSS}{n-2}}$$

Multivariate Data

$$\hat{y} = b_1 + b_2 x_2 + ... + b_k x_k \quad b_j = \frac{\sum_{i=1}^{n}\tilde{x}_{ji}(y_i - \bar{y})}{\sum_{i=1}^{n}\tilde{x}_{ji}^2}$$

$$b_j \pm t^*_{\alpha/2,n-k} \times s_{b_j} \quad t = \frac{b_j - \beta_j}{s_{b_j}} \quad s_{b_j} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}\tilde{x}_{ji}^2}} \quad s_e = \sqrt{\frac{ResSS}{n-k}}$$

$$\bar{R}^2 = 1 - \frac{ResSS/(n-k)}{TSS/(n-1)} \quad F_{q,n-k} = \frac{(ResSS_r - ResSS_u)/q}{ResSS_u/(n-k)}$$

$$invFtail(q, n-k, p) \to F^* : P[F > F^*] = p$$

Scratch paper