ECN 102: Analysis of Economics Data

Final Exam SS1 2026 – Answer Key

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.

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]

Observational. No randomization is taking place; we are simply recording characteristics of cars sold in 1978.

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

Cross-section. We observe different units (cars) at a single point in time (1978). There is no time dimension or repeated measurement of the same units.

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

Kurtosis measures the thickness of a distribution's tails (its propensity for extreme outliers). A normal distribution has $kurtosis = 3$. Since $kurtosis = 2.2 < 3$, the distribution of *headroom* is platykurtic: it has thinner tails and fewer extreme outliers than a normal distribution.

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

Dummy (binary indicator) variable. It takes only the values 0 (domestic) or 1 (foreign), encoding a qualitative/categorical characteristic (country of origin) as a numeric $\{0, 1\}$ indicator rather than measuring a quantity on a continuous scale.

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
```

$$CV_{price} = \frac{2949.50}{6165.26} = 0.48 \qquad CV_{weight} = \frac{777.19}{3019.46} = 0.26 \qquad CV_{mpg} = \frac{5.79}{21.30} = 0.27$$

***Price* has the greatest dispersion ($CV = 0.48$); *weight* has the least ($CV = 0.26$).**

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

$$se_{price} = \frac{s_{price}}{\sqrt{n}} = \frac{2949.50}{\sqrt{74}} = 342.87$$

This is an estimator for the population standard deviation of the sample mean of price, $\sigma_{\overline{price}}$.

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

`foreign==0` indicates a domestically produced car.

Question 2: Standard Errors [15 points]

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]

Homoskedasticity (Assumption 3). The standard assumption is $V[u_i \,|\, x_{2i}, \ldots, x_{ki}] = \sigma^2$ for all i. We use robust standard errors when we suspect the error variance is not constant across observations (heteroskedasticity): $V[u_i \,|\, x_{2i}, \ldots, x_{ki}] = \sigma^2_{u_i}$.

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

They do not change. Robust standard errors affect only how we estimate the variance of our OLS estimator; the coefficient estimates b_j themselves are computed identically by OLS regardless of which standard errors we report.

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]

Generally wider (or approximately the same). When heteroskedasticity is present, robust standard errors tend to be larger than default standard errors. Since $CI = b_j \pm t^*_{\alpha/2} \times s_{b_j}$, larger standard errors produce wider intervals. If the data is truly homoskedastic, robust and default standard errors will be approximately equal. Using robust standard errors is conservative but always valid: it cannot lead to improper inference even when homoskedasticity holds.

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]

We cluster standard errors when we believe errors are correlated within identifiable groups (clusters) but independent across groups. This addresses a violation of the independence assumption (Assumption 4). For example, if we study the effect of a farm subsidy on household crop yields, harvests within the same village may be correlated due to shared weather, infrastructure, or social networks. We would cluster at the village level, allowing errors to be correlated within villages while assuming they are independent across villages.

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

No. Clustered standard errors assume that errors are correlated *within* clusters and independent *between* clusters. The described scenario has this exactly backwards: correlation across clusters violates the between-cluster independence assumption that clustering relies on. Clustered SEs would not be appropriate here.

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

When we suspect non-linearity, i.e., our linear model may be misspecified. Bootstrapping treats the sample as the population, resamples with replacement a large number of times, and uses the variation across bootstrap samples to estimate the standard error of our statistic without assuming a correctly specified linear model.

Question 3: Multivariate Regression [23 points]

```
      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
------------------------------------------------------------------------------
```

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

$$price_i = \beta_1 + \beta_2\, mpg_i + \alpha_1\, foreign_i + \alpha_2\,(mpg_i \times foreign_i) + u_i$$

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

Among domestic cars ($foreign = 0$), a one-mpg increase in fuel efficiency is associated with a \$329.26 decrease in price, holding all else constant ($b_2 = -329.26$).

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

$a_1 = -13.59$ **is the difference in the intercept of the price–mpg relationship between foreign and domestic cars (i.e., the vertical shift for foreign cars when $mpg = 0$). Since $mpg = 0$ is outside the data range, a_1 captures the differential base price level for foreign versus domestic cars.**

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

$a_2 = 78.89$ **is the difference in the slope of the mpg–price relationship between foreign and domestic cars. For each additional mpg, foreign cars are associated with \$78.89 more in price compared to domestic cars.**

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]

Since `Prob>F` $\approx 0.000 < \alpha = 0.05$ ($F(3, 70) = 9.48$), we reject the null hypothesis of the test of overall significance ($H_0 : \beta_2 = \alpha_1 = \alpha_2 = 0$). At least one of our regressors statistically explains variation in price. Because we reject at $\alpha = 5\%$, our F-statistic must lie above the corresponding critical value: $F(3, 70) = 9.48 > F^*_{3,70,0.05}$.

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]

Only *mpg* is statistically different from zero at $\alpha = 5\%$ (`P>|t|` ≈ 0.000). The coefficients on *foreign* (`P>|t|` $= 0.996$) and *mpgXforeign* (`P>|t|` $= 0.485$) are not statistically different from zero. These are individual t-tests (tests of significance/association) of each coefficient against zero.

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

Plug in $foreign = 0$ and $foreign = 1$ respectively:

$$\text{Domestic } (foreign = 0): \quad \widehat{price} = 12{,}600.54 - 329.26 \cdot mpg$$

$$\text{Foreign } (foreign = 1): \quad \widehat{price} = (12{,}600.54 - 13.59) + (-329.26 + 78.89) \cdot mpg = 12{,}586.95 - 250.37 \cdot mpg$$

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

$$\frac{\partial \widehat{price}}{\partial mpg} = b_2 + a_2 \cdot foreign = -329.26 + 78.89 \cdot foreign$$

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

A regression of *price* on just a constant term is exactly the restricted model of the test of overall significance from part (e), so its $ResSS$ is $ResSS_r$. Rearrange the F-statistic formula and substitute the reported $F(3, 70) = 9.48$, $ResSS_u = 451{,}630{,}115$, $q = 3$, and $n - k = 70$:

$$F_{q,n-k} = \frac{(ResSS_r - ResSS_u)/q}{ResSS_u/(n-k)} \quad \Rightarrow \quad ResSS_r = ResSS_u + q \cdot F \cdot \frac{ResSS_u}{n-k}$$

$$ResSS_r = 451{,}630{,}115 + 3 \times 9.48 \times \frac{451{,}630{,}115}{70} = 451{,}630{,}115 + 28.44 \times 6{,}451{,}858.79$$

$$ResSS_r = 451{,}630{,}115 + 183{,}490{,}864 \approx 635{,}120{,}979$$

Equivalently: with only a constant term the fitted value is $\widehat{price} = \overline{price}$, so $ResSS_r = \sum (price_i - \overline{price})^2 = TSS = 635{,}065{,}396$, the Total SS already reported in the output.

Question 4: OLS and Logs [18 points]

```
      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]

Log-linear regression.

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

A one-unit (one-mpg) increase in fuel efficiency is associated with approximately a $-0.0333 \times 100 = -3.33\%$ **decrease in price (**$b_2 = -0.0333$**). More precisely,** $b_2 \approx (\Delta price/price)/\Delta mpg$**, so** b_2 **gives the proportional change in price per additional mpg.**

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

$$\frac{d\widehat{price}}{d\,mpg} = b_2 \cdot price = -0.0333 \cdot price$$

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

$$MEM = b_2 \cdot \overline{price} = -0.0333 \times 6165.26 = -205.30$$

```
      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]

Log-log regression.

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

A 1% increase in *mpg* is associated with a 0.8268% **decrease in *price*** **($b_2 = -0.8268$). The coefficient b_2 is the price elasticity with respect to fuel efficiency.**

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

$$MER = b_2 \cdot \frac{price^*}{mpg^*} = -0.8268 \times \frac{4500}{25} = -0.8268 \times 180 = -148.82$$

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]

$$H_0 : \alpha_1 = 0 \ \cap \ \alpha_2 = 0$$

$$H_A : \alpha_1 \neq 0 \ \cup \ \alpha_2 \neq 0$$

(Country of origin has no joint effect on price, meaning both the intercept shift and the slope shift for foreign cars are zero, versus at least one being non-zero.)

Country of origin enters the model through both *foreign* and *mpgXforeign*, so both α_1 and α_2 must equal zero under H_0. The test therefore imposes $q = 2$ linear restrictions simultaneously.

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

Unrestricted model: $price = \beta_1 + \beta_2\, mpg + \alpha_1\, foreign + \alpha_2\, (mpg \times foreign) + u$

Restricted model: $price = \beta_1 + \beta_2\, mpg + u$ (imposing $\alpha_1 = \alpha_2 = 0$)

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]

Since `Prob>F` $= 0.0387 < \alpha = 0.05$, we reject H_0 and conclude that country of origin jointly affects price at the 5% significance level. At least one of α_1 and α_2 is statistically different from zero.

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

The default t-test for *foreign* tests only $\alpha_1 = 0$ in isolation; it says nothing about the interaction coefficient α_2. Country of origin could still affect price through the slope shift (*mpgXforeign*) even if α_1 is not individually significant, which the t-test would miss entirely. The F-test (`test foreign mpgXforeign`) simultaneously tests $\alpha_1 = 0$ AND $\alpha_2 = 0$, accounting for the full joint effect of country of origin on price across both the intercept and the slope. It is the appropriate tool for a multi-coefficient joint hypothesis.

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]

$$F_{q,n-k} = \frac{(ResSS_r - ResSS_u)/q}{ResSS_u/(n-k)} = \frac{(480-420)/2}{420/(74-4)} = \frac{60/2}{420/70} = \frac{30}{6} = 5.00$$

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
```

The correct critical value is `invFtail(2, 70, 0.05)` = 3.1277, since $q = 2$ (numerator df) and $n - k = 74 - 4 = 70$ (denominator df). Since $F = 5.00 > 3.13 = F^*_{0.05}$, we **reject** H_0 at $\alpha = 5\%$.

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 $(1 - P[\text{Type II}])$ 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

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* $\mu_{\bar{X}}, \sigma^2_{\bar{X}}$

1. Common mean. We assume that all realizations of X have the same population mean; $\mathbb{E}[X_i] = \mu \; \forall \; i$. In the world of a coin flip, we are flipping a fair coin (50/50 odds) each time.

2. Common variance. We assume that all realizations of X have the same spread; $V[X_i] = \sigma^2 \; \forall \; i$. In the world of a coin flip, this again means that we are flipping a coin with the same odds each time.

3. Independence of observations. We assume that each realization of X does not depend on any previous realizations or affect any future realizations; $X_i \perp X_j \; \forall \; i \neq j$. For our coin flip, this means the odds of a heads each time are 50/50 and do not depend on what was previously flipped.