ECN 102: Analysis of Economics Data

Final Exam SQ 2025 – 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 2 hours to complete this exam. No additional time may be taken 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 your final answer 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 any of these extra pages for your final answers, please write your name and student ID at the top of each scratch page to ensure they are not lost.

This exam is worth 100 points.

For this exam, we will use a dataset on doctor visits that includes age, sex, income, and health statistics for a group of individuals. The variables present in the dataset are listed here.

```
Contains data from AED Doctor Visits Data.dta
 Observations:         5,190                  Data for A. Colin Cameron
                                                (2022), ANALYSIS OF ECONOMIC
                                                DATA, Amazon
    Variables:             8                  22 Jan 2022 15:50
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
female          float   %9.0g                 Equals 1 if female and 0
                                                otherwise
freepoor        float   %9.0g                 Equals 1 if free government
                                                insurance due to low income
freerepa        float   %9.0g                 Equals 1 if free government
                                                insurance due to old-age,
                                                disability or veteran
levyplus        float   %9.0g                 Equals 1 if private insurance
age             float   %9.0g                 Age in years (midpoint of 10 year
                                                age groups)
income          float   %9.0g                 Annual income in tens of
                                                thousands of dollars
illness         float   %9.0g                 Number of illnesses in past 2
                                                weeks
visits          float   %9.0g                 Number of doctor (or specialist)
                                                visits in past 2 weeks
-------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

Question 1: Summary Statistics [16 points]

a) How could we classify the data for variable *income* in this dataset? Explain three ways. [6 points]

- **Observational (no randomization)**
- **Cross-sectional (no time variation)**
- **[continuous] Numerical (money)**

b) What is the meaning of an observation with `female==0,freepoor==1` in this dataset? [2 points]

A non-female on free low-income government insurance.

c) What type of variable is *levyplus* if we include it in a regression? [2 points]

Dummy variable.

Variable	Obs	Mean	Std. dev.	Min	Max
income	5,190	.5831599	.3689067	0	1.5
illness	5,190	1.431985	1.384152	0	5
visits	5,190	.3017341	.7981338	0	9

d) Which variable above has the greatest dispersion? Which has the least? [3 points]

$$CV_{income} = \frac{0.37}{0.58} = 0.64 \quad CV_{illness} = \frac{1.38}{1.43} = 0.965 \quad CV_{visits} = \frac{0.80}{0.30} = 2.67$$

***Visits* has the greatest dispersion based on the coefficient of variation (CV); *income* has the least.**

e) What is the standard error of *illness*? (You can leave your answer as a formula) [3 points]

$$s_{illness} = \frac{1.38}{\sqrt{5190}}$$

Question 2: Bivariate Regression [16 points]

Suppose we want to run a regression to predict *visits* using *income* with OLS:

a) Write down the population regression model we are trying to estimate. [2 points]

$$visits = \beta_1 + \beta_2 income + u$$

b) What will the **independence** assumption imply for this regression? [2 points]

Error terms are uncorrelated across observations: u_i is independent from u_j.

c) What will the **linearity** assumption imply for this regression? [2 points]

The true population model is linear: $visits = \beta_1 + \beta_2 income + u$.

```
      Source |       SS           df       MS      Number of obs   =     5,190
-------------+----------------------------------   F(1, 5188)      =     81.73
       Model |  51.2660961         1  51.2660961   Prob > F        =    0.0000
    Residual |   3254.2183     5,188  .627258731   R-squared       =    0.0155
-------------+----------------------------------   Adj R-squared   =    0.0153
       Total |  3305.48439     5,189  .637017613   Root MSE        =      .792

------------------------------------------------------------------------------
      visits | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0048538   .0005369     9.04   0.000     .0038013    .0059063
       _cons |   .1044828   .0244318     4.28   0.000     .0565861    .1523794
------------------------------------------------------------------------------
```

d) Suppose we estimate this regression and find that our 95% confidence interval for b_2 is $(.004, .006)$. What conclusion from the test of association for β_2 does this imply? [3 points]

Our coefficient is statistically different from zero at $\alpha = 5\%$ since 0 does not appear in the 95% confidence interval of b_2. We can reject the null hypothesis that $\beta_2 = 0$.

e) Suppose our R-squared for this regression is 0.0155. What does this mean in words? [2 points]

$0.0155 \times 100 = 1.55\%$ **of the variation in *visits* can be explained by variation in *income*.**

Suppose instead we estimated a regression of the log of visits on the log of income:

f) What would we call this type of regression? [2 points]

Log-log regression.

g) How would we interpret b_2 from this regression in words? [3 points]

A 1% change in *income* is associated with a b_2% change in *visits*.

Question 3: Multivariate Regression [23 points]

We now estimate $visits = \beta_1 + \beta_2 income + \alpha_1 freepoor + \alpha_2(income \times freepoor) + u$ with OLS:

```
      Source |       SS           df       MS      Number of obs   =     5,190
-------------+----------------------------------   F(3, 5186)      =     15.25
       Model |  28.9007111         3  9.63357037   Prob > F        =    0.0000
    Residual |  3276.58368     5,186  .631813282   R-squared       =    0.0087
-------------+----------------------------------   Adj R-squared   =    0.0082
       Total |  3305.48439     5,189  .637017613   Root MSE        =    .79487

------------------------------------------------------------------------------
      visits | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      income |  -.1804711   .0305348    -5.91   0.000    -.2403322   -.1206101
    freepoor |  -.1235766   .0947101    -1.30   0.192    -.3092484    .0620951
incomeXfre~r |  -.2656318   .2516047    -1.06   0.291    -.7588831    .2276195
       _cons |   .4156881   .0214031    19.42   0.000      .373729    .4576472
------------------------------------------------------------------------------
```

a) What type of variable is *freepoor* in this regression? [2 points]

Dummy variable.

b) Which population parameter suggests that those on low-income government insurance have 0.12 fewer doctor visits on average compared to those not on low-income government insurance? [2 points]

$$\alpha_1$$

c) Which population parameter suggests that those on low-income government insurance have 0.265 fewer doctor visits per unit of income compared to those not on low-income government insurance? [2 points]

$$\alpha_2$$

d) Which population parameter suggests that those **not** on low-income government insurance have 0.18 fewer doctor visits per unit of income? [2 points]

$$\beta_2$$

e) What conclusion should we draw from the reported F-statistic and its associated p-value? [3 points]

Since `Prob > F` $\approx 0.000 < \alpha = 0.05$, we can reject the null of our test of overall significance and say that at least one of our coefficients is non-zero for predicting our outcome variable. Our regressors jointly predict variation in our outcome variable.

f) Which of our coefficients are statistically different from zero at $\alpha = 5\%$? Which are not? [3 points]

The coefficient on *income* is statistically different from zero at 5%; those on *freepoor* and the interaction term *incomeXfreepoor* are not different from zero.

g) If our four population assumptions hold, 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]

$$b_2 \sim N(\beta_2, \sigma^2_{b_2}); \left[\sigma^2_{b_2} = \frac{\sigma^2_u}{\sum_{i=1}^{n} \tilde{x}^2_{2i}}\right]$$

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

$$\frac{\partial \widehat{visits}}{\partial income} = b_2 + a_2 freepoor$$

i) If we want to compare model fit of this regression to that of our previous bivariate model, should we use R-squared or Adj R-squared? Why? [3 points]

We should use adjusted R-squared, since our models contain different numbers of regressors. Adjusted R-squared penalizes the addition of new regressors while R-squared does not.

Question 4: Multivariate Regression [18 points]

For a multivariate OLS regression with sample size n, $k-1$ regressors $x_2, ..., x_k$, k estimated coefficients $b_1, ..., b_k$, and $j = 1, ..., k$ as our index of coefficients:

a) What is the difference between x_j and $\tilde{x}_j$? [3 points]

x_j is all the variation of our j^{th} regressor while $\tilde{x}_j$ is the *unique* variation in x_j not correlated with/independent of all other regressors in our model. We use $\tilde{x}_j$ instead of x_j in a multivariate regression.

b) Suppose we add a new regressor x_{k+1} to our model. What do we expect to happen to the R-squared of our regression? [3 points]

We expect R-squared to weakly increase (get larger or stay the same) since adding regressors to our model can never decrease our model fit.

c) Suppose we add a new regressor x_{k+1} to our model and the standard error of b_j **decreases**. What might this suggest about our model? [3 points]

This suggests that x_{k+1} improves our model fit and decreases RMSE in the numerator of s_{b_j}.

d) Suppose instead we add a new regressor x_{k+1} to our model and the standard error of b_j **increases**. What might this suggest about our model? [3 points]

This suggests that x_{k+1} is correlated with x_j, since this would decrease $\tilde{x}_j$ in the denominator of s_{b_j}.

e) Name the two models that are compared when computing the statistic for the default test of overall significance. Write down the population model for both. [6 points]

$$\text{Unrestricted: } y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

$$\text{Restricted: } y = \beta_1 + \nu$$

Question 5: Miscellaneous [17 points]

a) Compute the following sum [3 points]:

$$6 + \sum_{i=3}^{5} \left(\frac{60}{i}\right)(i-2)$$

$$= 6 + \left[\frac{60}{3}(3-2) + \frac{60}{4}(4-2) + \frac{60}{5}(5-2)\right]$$
$$= 6 + 20(1) + 15(2) + 12(3)$$
$$= 6 + 86$$
$$= 92$$

b) If the default p-values reported for all coefficients from a multivariate regression of y on $x_2, ..., x_k$ are above 0.05 but we observe `Prob > F = 0.035`, what can we conclude about our regressors? [4 points]

All of our regressors are not statistically significantly different from zero *individually* but do *jointly* explain variation in our outcome variable.

c) If we choose to use robust standard errors in our model and then find out the homoskedasticity assumption is **not** violated, should we be concerned? Why or why not? [3 points]

No. Robust standard errors are correct for both homoskedastic and heteroskedastic errors.

d) If we are told to use **clustered** standard errors in a regression model but know that population assumptions 1, 2, and 4 hold for our data, should we cluster or not? Why? [3 points]

No. We know by assumption 4 that our error terms are independent, meaning there is no need to cluster.

e) If to determine the impact a certain variable (not regressor) has on our multivariate regression model we **must use an F test/cannot use a t-test**, what does this tell us about this variable? [4 points]

This means our variable must appear in multiple regressors rather than just one (e.g. *income* above).

Multiple Choice [2 points each]

Only one answer is correct for each question. Read carefully and choose the best possible answer.

MC 1 & 2

Suppose the probability of rejecting H_0 is 0.05 when $\beta_j = 0$ and 0.95 when $\beta_j \neq 0$ for a test on OLS b_j.

The **size** of this test will be equal to:

a) 5%

b) 95%

c) 100%

d) 0%

The **power** of this test will be equal to:

a) 5%

b) 95%

c) 100%

d) 0%

MC 3

A trial judge says he believes he made a Type I error when sentencing a defendant. The precedent in court is that individuals are **innocent until proven guilty**. If the judge is correct, which of the following took place?

a) The defendant was truly guilty and was found guilty

b) The defendant was truly guilty but was found innocent

c) The defendant was truly innocent and was found innocent

d) The defendant was truly innocent but was found guilty

MC 4

If $\frac{Q-4}{4} \sim N(0,1)$, it **must** be true that:

a) $E[Q] = 0$

b) $V[Q] = 4$

c) $Q \sim N$

d) $\sigma_Q = 16$

e) None of the above

MC 5

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

a) We claimed that we could not compare the R-squared of a regression of y on x and a regression of ln(y) on x.

b) We claimed that we could not compare the R-squared of a regression of y on x and a regression of y on ln(x).

c) We claimed that the critical value approach and p-value approach for a given test would always give us the same answer of whether to reject the null hypothesis or not.

d) We claimed that the standard error of X will always be smaller than the standard deviation of X for $n > 1$.

e) None of the above