ECN 102: Analysis of Economics Data

Final Exam FQ 2024 – 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. Make sure to read every question thoroughly and carefully. Make sure to answer all parts of all questions completely and fully explain when asked. Show all work to receive full credit.

You may use only the calculators provided by the instructor. Graphing calculators, phone calculators, and other non-standard calculators may not be used. For questions requiring computation, it suffices to express your final answer with 2 significant figures.

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, you must write your name and student ID at the top of each scratch page to ensure it is not lost.

This exam is worth 100 points.

For this exam, we will use a dataset that includes annual salary, demographic characteristics, and other metrics for a sample of individuals. The variables present in the dataset are listed here, along with their format and labels.

```
Contains data from SalaryCameron.dta
 Observations:          776                 Data for A. Colin Cameron
                                              (2022), ANALYSIS OF ECONOMIC
                                              DATA, Amazon
    Variables:           12                 22 Jan 2022 15:50
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
salary          float   %9.0g                 Annual Salary in dollars
lnsalary        float   %9.0g                 Natural Logarithm of Annual
                                                Salary
satverb         int     %32.0g                SAT test verbal score (highest
                                                self-reported score as of 2007)
satmath         int     %32.0g                SAT test mathematics score
                                                (highest self-reported score as
                                                of 2007)
highgrade       byte    %8.0g                 Highest grade ever completed
age             float   %9.0g                 Age in years
sex             byte    %14.0g                1 = female 0 = male
minority        float   %9.0g                 1 = minority 0 = not minority
height          float   %9.0g                 Height in inches
weight          int     %8.0g                 Weight in pounds
genhealth       byte    %9.0g                 Health status: 1=Excellent
                                                2=VeryGood 3=Good 4=Fair 5=Poor
actscore        float   %32.0g                ACT test score (highest
                                                self-reported score as of 2007)
-------------------------------------------------------------------------------
Sorted by:
     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 merely observing a sample of individuals.

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 of observation (people) at a single point in time; we have no *year*, *month*, *day* variable.

c) How else could we classify data for our variable *salary*? Explain. [2 points]

Numerical or continuous numerical.

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

This is a **dummy variable** because it takes only values 1 or 0 if someone is or is not a minority.

e) Are any other **variables** the same type as our variable *minority* when used in regression? [2 points]

Sex is also a dummy variable that takes value 1 for female and 0 for male.

Variable	Obs	Mean	Std. dev.	Min	Max
height	776	67.68814	4.198084	50	83
weight	776	171.3015	43.5782	92	475
salary	776	36184.23	24426.81	300	130254

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

$$CV_{height} = \frac{4.20}{67.69} = 0.06$$

$$CV_{weight} = \frac{43.58}{171.30} = 0.25$$

$$CV_{salary} = \frac{24426.81}{36184.23} = 0.68$$

Salary has the highest dispersion; *height* has the least.

g) What would we expect the standard error of *height* to be? What is this an estimator for? [3 points]

$$se_{height} = \frac{4.20}{\sqrt{776}} = 0.15$$

This is an estimator for the **population standard deviation** of the sample mean of *height.*

Question 2: Bivariate Regression [16 points]

Suppose we want to run a regression to predict *salary* using *age* with OLS.

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

$$\widehat{salary} = \beta_1 + \beta_2 age \text{ or}$$
$$salary = \beta_1 + \beta_2 age + u$$

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

$Var(u) = \sigma_u^2$ or our errors have a constant variance across values of *age*.

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

Our true population model is $salary = \beta_1 + \beta_2 age + u$ (we do not have any omitted covariates or nonlinear relationship).

```
      Source |       SS           df       MS      Number of obs   =       776
-------------+----------------------------------   F(1, 774)       =      5.78
       Model |  3.4281e+09         1  3.4281e+09   Prob > F        =    0.0164
    Residual |  4.5899e+11       774   593010915   R-squared       =    0.0074
-------------+----------------------------------   Adj R-squared   =    0.0061
       Total |  4.6242e+11       775   596669065   Root MSE        =     24352

------------------------------------------------------------------------------
      salary | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   2605.833   1083.809     2.40   0.016       478.28    4733.387
       _cons |  -34210.21   29291.25    -1.17   0.243    -91709.92     23289.5
------------------------------------------------------------------------------
```

d) Suppose we estimate this regression and find that our 95% confidence interval for b_2 is $(478.28, 4733.39)$. What conclusion about our population coefficient β_2 on *age* does this imply? [3 points]

Since 0 is not included in our 95% CI, we can reject the null for the test of association and conclude that $\beta_2 \neq 0$ or that *age* is associated with *salary*.

e) Suppose our R-squared for this regression is 0.0074. What does this mean in words? [2 points]

$0.0074 \times 100 = 0.74\%$ of the variation in *salary* is explained by *age*.

Suppose instead we estimated a regression of salary on the **log of age**.

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

Linear-log regression

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

A 1% change in *age* is associated with a $b_2/100$-unit (dollar) change in *salary*.

Question 3: Multivariate Regression [23 points]

We now estimate $\widehat{salary} = \beta_1 + \beta_2 age + \alpha_1 sex + \alpha_2 (age \times sex)$ with OLS and obtain the following:

```
      Source |       SS           df       MS      Number of obs   =       776
-------------+----------------------------------   F(3, 772)       =      4.32
       Model |  7.6433e+09         3  2.5478e+09   Prob > F        =    0.0049
    Residual |  4.5478e+11       772   589087072   R-squared       =    0.0165
-------------+----------------------------------   Adj R-squared   =    0.0127
       Total |  4.6242e+11       775   596669065   Root MSE        =     24271

------------------------------------------------------------------------------
      salary | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   1706.862   1579.951     1.08   0.280    -1394.648    4808.373
         sex |  -53271.99   58521.46    -0.91   0.363    -168152.1    61608.08
     ageXsex |   1807.953   2165.534     0.83   0.404    -2443.081    6058.986
       _cons |  -7612.026   42668.87    -0.18   0.858    -91372.78    76148.73
------------------------------------------------------------------------------
```

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

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

Dummy variable

b) Interpret the value of b_2 in words. [2 points]

Being one year older on average is *associated* with a \$1706.86 higher salary for men ($sex = 0$).

c) Interpret the value of a_1 in words. [2 points]

Women make, on average, \$53,271.99 less than men (or the difference in intercept between men and women).

d) Interpret the value of a_2 in words. [2 points]

Women make \$1807.95 more for each additional year older they are compared to men. Being one year older is associated with a $1706.86 + 1807.95 = \$3514.81$ higher salary for women ($sex = 1$) (or difference in slopes between men and women).

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

$\Pr(F > 4.32) = 0.0049 < \alpha = 0.05$: reject the null hypothesis of the default test of overall significance. At least one of our regressors/our model explains variation in our outcome variable statistically significantly.

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

All regressors have a p-value above $\alpha = 0.05$: none are statistically significantly 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]

$$b_2 \sim N\left(\beta_2, \frac{\sigma_u^2}{\sum_{i=1}^{n} \tilde{x}_{2i}^2}\right) \text{ or } b_2 \sim N(\beta_2, \sigma_{b_2}^2)$$

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

$$\frac{\partial \widehat{salary}}{\partial age} = b_2 + a_2 sex$$

i) Why are our values for R-squared and Adj R-squared different in this regression? Why might we prefer one instead of the other for multivariate regressions? [3 points]

Adjusted R-squared imposes a penalty for adding additional regressors, here $k = 4$, while regular R-squared does not. We may want a measure of model fit that penalizes endlessly adding regressors.

Question 4: Hypothesis Testing [18 points]

For a general multivariate OLS regression with sample size n, $k-1$ regressors $(x_2, ..., x_k)$ and a constant, and $j = 1, ..., k$ as our index of coefficients:

a) What are the null and alternate hypotheses for the default test of **association** for a single b_j? [2 points]

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

b) Suppose we reject this null hypothesis for the default test of association for b_j above. How would we interpret this conclusion in words? [4 points]

There is a statistically significant association between the **unique** variation in our regressor x_j and our outcome variable y.

c) 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]

$$\text{Unrestricted: } \hat{y} = \beta_1 + \beta_2 x_x + ... + \beta_k x_k \text{ ; Restricted:} \hat{y} = \beta_1 + 0 + ...0$$

d) If the p-value for the default test of association for b_j is 0.035, what would we expect the p-value to be for an F-test of the single regressor b_j? [3 points]

$p = 0.035$ (the same as for the test of association)

e) If the t-statistic for the default test of association for b_j is 2.12, what would we expect the F-statistic to be for an F-test of the single regressor b_j? [3 points]

$F = (t)^2 = 2.12^2 = 4.49$ (the square of the t-statistic for the test of association)

Question 5: Miscellaneous [17 points]

a) 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.

b) What assumption or assumptions are we concerned may be violated if we choose to use **robust** standard errors in our regression? [3 points]

Homoskedasticity assumption (no explanation needed)

c) What assumption or assumptions are we concerned may be violated if we choose to use **clustered** standard errors in our regression? [3 points]

Independence and **homoskedasticity** assumptions (no explanation needed)

d) If we suspect but **are not sure** that our data is homoskedastic, what type of standard errors should we use? Are we concerned that doing so may lead to problems? Why or why not? [4 points]

We should use **robust standard errors**. We are not concerned; using robust standard errors would not be wrong even if our data is homoskedastic; NOT using robust standard errors would be wrong if our data is NOT homoskedastic.

e) Compute the following sum [3 points]:

$$\sum_{i=2}^{3}(3i)^{i-1} - i$$

$$\sum_{i=2}^{3}(3i)^{i-1} - i = 6^{2-1} - 2 + 9^{3-1} - 3 = 6 + 81 - 5 = 82$$

Multiple Choice [2 points each]

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

MC 1 and 2

Suppose the probability of **not** rejecting H_0 is 0.95 when $\beta_j = 0$ and 0.3 when $\beta_j \neq 0$ for a test on OLS b_j.

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

a) 0.05

b) 0.95

c) 0.7

d) 0.3

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

a) 0.05

b) 0.95

c) 0.7

d) 0.3

MC 3

A trial judge says he believes he made a Type II 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 $\frac{Q-7}{3} \sim (0, 1)$, it **must** be true that:

a) $E[Q] = 0$

b) $V[Q] = 3$

c) $Q \sim N$

d) $\sigma_Q = 9$

e) None of the above

MC 5

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

a) We claimed that R^2 and $\bar{R}^2$ should always be the same for a bivariate regression.

b) We claimed that the null hypothesis for a computed statistic should be rejected at a significance level of α if the $100(1-\alpha)\%$ confidence interval for that statistic does not include the null value.

c) We claimed that the sample standard deviation of X will always be **weakly** greater ($\geq$) than the standard error of $\bar{X}$ for any sample size.

d) We claimed that the total effect of b_j from multivariate OLS of y on $x_2, ..., x_k$ will always be equal to the partial effect if x_j is independent from all other regressors.

e) None of the above