ECON/AREC 335 Introduction to Econometrics
Fall 2020
Instructor: Lackson D. Mudenda
Problem Set 2

Instructions (Please read!): No credit will be given *for late problem sets. Problem sets will NOT be accepted via email. Please UPLOAD TO THE PROVIDED LINK ON CANVAS. Please put your FULL name clearly at the start of your submission. Working in study groups is encouraged. However, problem sets must be written up and submitted individually. Please underline, circle, or highlight relevant numbers and final solutions. Please write all of your answers in correct numerical order. You will lose points for not following these directions. Presentation is a big part of life and will be significant to your success after you leave CSU. Start thinking about this now! Start early as most questions have several sub-parts, and note point values when allocating your time.*

It is expected that you will show your work and/or fully explain your answers for all questions in this problem set (and in this class in general).

1. [15 points] A researcher decides to run an experiment to measure the effect of study group time on homework scores. He gives each of 50 students in the same class the same problem set to complete, but some students are given 1 hour to work with a study group and others are given 2 hours to do the same. Suppose that he can monitor their study habits and somehow make sure that students are not working together in other settings beyond the hours associated with the experiment. Each student is randomly assigned into one of the two groups based on a flip of a coin. Let Y_i be the number of points achieved on the homework by the *i*th student ($0 \leq Y_i \leq 100$), let X_i denote the amount of time that the student has to study with a group in hours (X_i=1 or 2), and consider the following regression model: $Y_i = \beta_0 + \beta_1 X_i + u_i$.

(a) Explain what the term u_i represents. Why will different students have different values of u_i? Give examples.
(b) Explain why $E(u_i|X_i)$ equals 0 for this regression model.
(c) Discuss whether or not the other least squares assumptions are satisfied. Explain.

2. [20 points] Suppose that you have data on counties in the United States and have recorded mediate household income in each county in dollars (MedHHInc) and have recorded the unemployment rate in each county in percentage points (UnempRate). A regression of median household income on unemployment rates yields:

$$\widehat{MedHHInc} = 65719.6 - 3508.17 UnempRate, \qquad R^2 = 0.150845, SER = 12441.70$$

(a) Interpret the regression slope in words. Be sure to clearly indicate the economic magnitude of this result (i.e. a *what* size change in *what* variable translates into a *what* size change in *what* other variable).
(b) Does the estimated regression intercept for this model have a relevant economic interpretation? Why or why not? Make sure to give a clear explanation.
(c) Interpret the R-squared statistic in words, making sure to indicate the magnitude.
(d) Interpret the standard error of the regression statistic in words, making sure to indicate the magnitude.

3. [15 points] Is each statement True, False, or Uncertain? Explain your answers clearly and succinctly in each case for credit. (No credit will be awarded for statements of True, False, or Uncertain without accompanying explanation.)

(a) The OLS estimator minimizes average squared difference between actual values of Y_i and the actual values of X_i based on estimated line.

(b) When we run ordinary least squares regression, we don't need to worry about any significant outliers in our data because the computer's optimization problem fixes these.

(c) You are attempting to build a model that explains aggregate average wages as a function of the level of community unemployment rates. You would rather sample during a period of fluctuating unemployment rates than during a period in which unemployment rates were relatively stable.

EMPIRICAL /COMPUTER WORK

4. [50 points (all questions here are worth 5points each except part (a) which is 10 points)] Important: As usual, your answer should include a printout *(can cut and paste into a file. I will show you how to do this)* of relevant calculations on the computer (R or other software output) AND a write up of final answers following the subparts of the question. The data again are *the same as for Problem Set 1.* Use these data to answer the following questions.

(a) Run a regression to determine the impact of the 2013 unemployment rate (UnempRate2013) on the per capita income (PerCapitaInc) in a county. What is the estimated slope? Explain what this number means in words in terms of the unemployment rate and in terms of per capita income. Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels. For this first pass, use homoskedastic standard errors.

(b) Re-run the regression from part (a) but this time use heteroskedastic standard errors. Are your coefficients the same as in part (a)? Why? Are your standard errors (of your betas) the same as in part (a)? Why?

(c) Run the same regression as in part (b) but now also include the additional regressors percentage of the population that is college-educated (Ed5CollegePlusPct), percentage of the population that is black (BlackNonHispanicPct2010), and percentage of the population that is Hispanic (HispanicPct2010). Now, what is the estimated impact of unemployment rate in 2013 on per capita income? Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

(d) Provide economic/econometric intuition as to why the impact of the unemployment rate's impact on per capita income changed between parts (b) and (c). Note that I am asking you to think about the context (and hence the "story" behind these data).

(e) Construct a 95% confidence interval for the slope coefficient on UnempRate2013 in (c). Write out your calculations. Clearly indicate how this confidence interval relates to whether UnempRate2013 is statistically significant or not in this context *by relating your answer to your constructed confidence interval*.

(f) You recall from problem set 1 that both the means of per capita income and of unemployment rate in 2013 are quite different across metro and nonmetro areas. You therefore want to explore this in more detail. Run the regression from (c) using only metro areas in 2013 (Metro2013==1). [Hint: You need to restrict the data based on a criterion before running the regression.] Now, what is the estimated effect of the 2013 unemployment rate on per capita income and also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

(g) Now, run the regression from (c) using only non-metro areas in 2013 (Metro2013==0). [Hint: You need to restrict the data based on a criterion before running the regression.] Now, what is the estimated effect of the 2013 unemployment rate on per capita income and also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

(h) What did you learn from the comparison between results in parts (f) and (g)? Explain your answer. Note that I again am asking you to think about the context (and hence the "story" behind these data).

(i) Return to the full sample. Now, run a regression to determine the impact of changing the percentage of the population which is college educated (Ed5CollegePlusPct) on the per capita income (PerCapitaInc) in a county. Include controls for the unemployment rate in 2010 (UnempRate2010), percentage of the population that is black (BlackNonHispanicPct2010), and now also include a dummy variable for metro status (Metro2013). Now, what is the estimated impact of percentage with a college education on per capita income? Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.