ECON/AREC 335 Introduction to Econometrics
Fall 2021
Instructor: Lackson D. Mudenda
Problem Set 3

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).

FORMART: Preferably the problem set should be typed, and your submission should be a pdf file type. However, if for whatever reason you choose not to type, please make your writing very legible.

Consider the following table for use in answering questions 1-4 below. The data set consists of information on 7,178 full-time full-year workers aged 25 to 34 years old. College is a binary variable equaling one if college. Female is a binary variable equaling one if female. Age is reported in years. Northeast, Midwest, and South are binary variables equaling one for the regions as stated.

Results of Regressions of Average Hourly Earnings on Sex and Education Binary Variables and Other Characteristics Using 2015 Data from the Current Population Survey

Dependent variable: average hourly earnings (*AHE*).

Regressor	(1)	(2)	(3)
College (X_1)	10.47 (0.29)	10.44 (0.29)	10.42 (0.29)
Female (X_2)	−4.69 (0.29)	−4.56 (0.29)	−4.57 (0.29)
Age (X_3)		0.61 (0.05)	0.61 (0.05)
Northeast (X_4)			0.74 (0.47)
Midwest (X_5)			−1.54 (0.40)
South (X_6)			−0.44 (0.37)
Intercept	18.15 (0.19)	0.11 (1.46)	0.33 (1.47)
Summary Statistics and Joint Tests			
F-statistic testing regional effects = 0			9.32
SER	12.15	12.03	12.01
R^2	0.165	0.182	0.185
n	7178	7178	7178

1. **[15 points] Using the regression results in column (2):**

(a) **Is the college-high school earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval of the difference and use it to answer the question. Show your work.**

(b) **Is the male-female earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval of the difference and use it to answer the question. Show your work.**

(c) Construct a 95% confidence interval for the expected difference between the earnings of a 29-year-old and a 32-year old female college graduate. Show your work.

2. [15 points] Using the regression results in column (3):

(a) Interpret the coefficient on *Northeast*. Make sure to indicate the size of the impact and what this is relative to (i.e. what is the base category).

(b) Do there appear to be important regional differences overall? Use an appropriate (joint) hypothesis test to explain your answer. Clearly explain what you are comparing in your test and the conclusions of your test. (Hint: There is a test statistic already calculated for you to use in the table).

(c) What would happen if a regressor for the remaining region, *West*, was included? Explain exactly why by referring to relevant concepts from the class.

3. [5 points] The coefficients on Female and Age don't appear to change as you move from column (2) to column (3) as the regional dummies are included. (Note that they probably do a little but it's out in the rounding.) What does this lack of change suggest about the relationships between these particular regressors and the regional dummies which are added in column (3)? Relate your answer specifically to the two omitted variable bias criteria.

4. [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)Adopting a "cross tabulation" approach, with finer gradations of variables within each group is one way to deal with an omitted variable problem when there are a small number of variables.

(b)The F-distribution with q degrees of freedom in the numerator and infinity in the denominator is the same as the Chi-squared distribution with q degrees of freedom when the sample is large.

(c)The joint confidence set corresponding to any null hypothesis involving more than two beta parameters is an ellipse.

EMPIRICAL /COMPUTER WORK

5. [50 points (all questions here are worth 5points each except part (a) which is 10 points)] Important: As usual, your answer should include a <u>printout</u> *(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)It is quite common in econometrics to model income variables nonlinearly. Construct a new variable called loginc where loginc=ln(PerCapitaInc). Provide summary statistics for this new variable. (Hint: Think back to how you constructed summary statistics in problem set 1.)

(b)Now run a model similar to the one that you ran as the final specification in problem set 2 except this time use loginc as the dependent variable (and we are also going to start controlling for metro status in addition to the other controls). Specifically, run a regression of loginc with unemployment rate in 2013 (UnempRate2013) as the main regressor, while also including the other regressors: percentage college educated (Ed5CollegePlusPct), percentage non-Hispanic black in 2010 (BlackNonHispanicPct2010), percentage Hispanic in 2010, and metro status in 2013 (Metro2013). Now, what is the estimated effect of UnempRate2013 in words? Also indicate if the relationship is statistically significant at the 10%, 5%, and 1% levels? Make sure that you are using heteroskedastic standard errors.

(c)What is the null hypothesis corresponding to the F-statistic as reported in the output for the regression in part (b)?

(d)What is the conclusion of the F-test as reported in the output for the regression in part (b)? Explain. (i.e. Do you reject or fail to reject the null hypothesis above and how do you know this?)

(e)What are the degrees of freedom (the specific numbers reported) for the F-test as specified in your regression output? Where are these numbers coming from for this case? (i.e., how is the computer calculating these numbers specifically given the dataset here?)

(f)Discuss what the standard error of the regression (SER), R2 and $\bar{R}^2$ in part (b) are telling you in terms of the numbers that you have found.

(g)Using what you know about the difference between the two formulas, explain specifically why the R2 and $\bar{R}^2$ statistics so similar for this case.

(h)Use an F-test to test the joint significance of the additional regressors: Ed5CollegePlus, BlackNonHispanicPct2010, HispanicPct2010, and Metro2013. Find this test statistic and clearly indicate the conclusions of the test.

(i)If you had more time to study this question and/or more/different data, what would you suggest doing next? Propose additional variables to add and/or different specifications to try and give specific reasons why you are suggesting these. Answers will vary for this part of the problem.