ECN/PUB 480/580
Assignment #4
Due: Thursday March 17, 2022 by end of the day

Directions: Answer each question electronically in a MS Word or .pdf file. Compile your answers into a single computer file and then upload it to Canvas under "Assignment #4." Contact me if you have any questions.

Go to Canvas to download the data set entitled wage.dta from where this assignment is posted. This is a data set containing observations on wages and some other variables for 1,000 workers. We are going to look at what explains the wage a worker earns, whether females earn a lower wage than males, and if they do, then potentially why.

The data set includes the following variables:

wage: average hourly earnings, in dollars, for each individual
educ: number of years of education for each individual
age: age of each individual, in years
exper: number of years of work experience for each individual
marr: dummy variable taking the value of 1 if the individual is married, 0 otherwise
female: dummy variable taking the value of 1 if the individual is female, 0 otherwise

Start by opening the data set in Stata and generate 5 new variables using the `gen` command. To help you out, I give the command you need to use in parentheses.

a) the natural logarithm of wages (`gen lwage = log(wage))`
b) experience squared (`gen expersq = exper^2`)
c) a new dummy variable for `male`, which takes on the value of 1 if the individual is a male and 0 otherwise (`gen male = 1-female`)
d) a new interaction dummy variable `marrfe` taking on the value of 1 if the individual is a married female, and 0 otherwise (`gen marrfe = marr*female)`
e) a new interaction dummy variable that will be equal to 1 if the individual is a married male, and 0 otherwise (`gen marrma = marr*male`)

The point of this assignment is to conduct a regression analysis to see if the "wage gap" between male and female workers persists after we start controlling for other variables, namely education and experience.

1. Create a table in MS Word (with a title and description) that summarizes the data by gender. That is, give the average wage, average education, average experience, and average age separately for men and women. You can use the `sum` command along with an `if` statement to do this (no commas). For example, `if female == 1` means only females. Use proper units for the variables in your summary statistics and round to two places past the decimal. Do NOT use the Stata output as your table! **(3 points)**

2. Consider the following regression:

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 expersq + \beta_4 female + u_i$$

This regression is a log-linear regression. This was "case 2" for using natural logarithms in regression in Lecture #12. Estimate this regression using Stata's `reg` command. Include your output along with the command used to generate it.

a. Interpret the $\hat{\beta}_1$ and $\hat{\beta}_2$ from each equation. Based on the coefficients, how is the independent variable related to the dependent variable? In particular, how will the dependent variable change if the independent variable increases by 1 unit? Be specific and use specific units! Refer to Case 2 for the interpretation of a $\hat{\beta}$ for a log-linear regression. **(3 points)**

b. Explain the signs on the coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$. What do the signs indicate about the relationship between *exper* and *lwage*? **Hint:** refer Lecture #13, which covers using quadratic terms in regression. Is the relationship between *lwage* and *exper* a straight linear relationship or does the relationship taper over time? Why do you think this is the case? **(3 points)**

c. Interpret $\hat{\beta}_4$, which is the estimated coefficient on the *female* dummy variable. Recall that the estimated coefficients on dummy variables have an "if-then" interpretation. In other words, *if* the dummy variable equals 1, *then* the y-variable changes by the associated $\hat{\beta}$. Recall that this is a log-linear regression, so the interpretation is that if the dummy variable equals 1, then the y-variable changes by $\hat{\beta}_4$x100%. $\hat{\beta}_4$x100% would thus be the "wage gap," since it says how much more or less someone earns, in percentage terms, if that person is female. **(3 points)**

3. a. Which of the estimated coefficients (that is, the $\hat{\beta}_s$) are statistically significant and why? Recall that statistically significant means you can reject the null hypothesis that a particular $\hat{\beta}$ is equal to zero. Explain your answer referring to the t-statistics, critical values, and use a 5% level of significance. You can either use the t-statistics in the Stata output or calculate them yourself. You can ignore the constant ($\hat{\beta}_0$). Use a two-tailed test. **(3 points)**

b. Explain what the R^2 and F-statistics mean for this regression. That is, for the F-statistic, do you reject the H_0 for the F-test? **(2 points)**

4. Estimate the regression for only single people. Do this by including the statement `if marr == 0` at the end of the `reg` command (no commas). Is the wage gap (as given by the $\hat{\beta}$ on the variable *female*) statistically significant? Compare the $\hat{\beta}$ *female* in this case to that in question #2. Is the "wage gap" for single women more or less than for women in general? How much more or less? **(3 points)**

5. Estimate the regression for only people who are single and less than 30 years old. Do this by including the statement `if marr == 0 & age < 30` at the end of the `reg` command (no commas). Given how young this subsample is, **do not** include the variable *expersq*. Include the variable *female* in both equations and interpret the results as you did in question #4. Is the wage gap statistically significant in this case? Is the "wage gap" for single women more or less than for single women in general? For women in general? How much more or less? Why do you think this is? **(3 points)**

1 point extra credit: Refer to the statement "Given how young this subsample is, **do not** include the variable *expersq*" from question 11. What do I mean by this statement? That is, why would including the square of experience be inappropriate here? You can answer this in just one or two sentences.

6. Estimate the regression, including *expersq*, for everyone (e.g. don't include the "if" statement), but include both the *female* dummy variable and the *marrfe*.
a. Is the $\hat{\beta}$ on the interaction term statistically significant? How do you know **(2 points)**

b. Interpret the $\hat{\beta}$ on the interaction term. Does being married increase or decrease a female worker's wage? How much? **(3 points)**

c. Compare the $\bar{R}^2$ if you have the interaction term in the regression versus if you don't. What does this mean about the interaction term? **(2 points)**