Homework 2

Eco 231 - Undergraduate Econometrics

Prof. Carolina Caetano

1. In this question, we work with HPRICE2.dta dataset, from the textbook Introductory Econometrics by Jeffrey Wooldridge. This dataset contains various characteristics that affect median housing prices (in dollars) over 506 communities in the Boston area:

   - crime: number of reported crimes within the community, per capita
   - nox: amount of nitrogen oxide in the air, in parts per million
   - rooms: average number of rooms
   - dist: weighted distance of the community from 5 employment centers, in miles
   - radial: access index to highways
   - proptax: property tax per $1000
   - stratio: average student-teacher ratio of schools in the community
   - lowstat: percentage of people with "lower status" in the community

   The dataset is available for download on the website. Answer the following questions using STATA.

   (a) What are the numbers of observations and variables in the dataset? Are there any string variables in the dataset?

   (b) Using nox as a measure of air pollution, our goal is to identify whether housing prices are affected by air pollution. Before running regression, we might as well look at summary statistics of the variables that we are interested in. Produce a table with summary statistics of price and nox.

   (c) Make a scatter-plot to see how price (in y-axis) and nox (in x-axis) are related. More specifically, attach linear fitted line within the same scatter-plot you draw. Make sure you name the figure, and properly label the figure and its axes. Describe what you see in the graph.

   (d) Now we are ready to move on. Estimate the following simple regression:

   $$price = \alpha_0 + \alpha_1\, nox + u$$

   Interpret estimated coefficients from the regression table.

(e) Besides air pollution, there are more factors that affect housing prices as well. In particular, some of the factors may act as confounders, which prevent us from identifying casual relationship between price and nox. Suppose we want to add crime, rooms, dist, proptax, and stratio as controls. Adding these variables, re-run the OLS regression as follows:

$$price = \beta_0 + \beta_1\ nox + \beta_2\ crime + \beta_3\ rooms + \beta_4\ dist + \beta_5\ proptax + \beta_6\ stratio + u$$

Interpret estimated coefficients from the regression table. Is there any change in coefficient on nox, comparing to the previous simple regression? Why or why not?

(f) Now, we take slightly different approach to understand the regression results in (e). First, run the following regression and produce residuals, which we denote as r:

$$nox = \gamma_0 + \gamma_1\ crime + \gamma_2\ rooms + \gamma_3\ dist + \gamma_4\ proptax + \gamma_5\ stratio + e$$

Next, using residuals r, estimate the following regression:

$$price = \delta_0 + \delta_1\ r + \varepsilon$$

Now, compare β_1 and δ_1. Are these estimates the same or different? Explain why or why not. Be precise and concise.

2. We return to working with the same public-use household survey data, household.dta, from the previous assignment. Recall that our researcher is interested in the effects of wife's education attainment and labor market experience on wife's earnings in the labor market. Answer the following questions using STATA.

(a) What are the numbers of observations and variables in the dataset? Are there any string variables in the dataset?

(b) Provide summary statistics of wage, educ, and exper. In addition, show that no missing values exist within any of these variables. Your answer, shown in the output screen, should be based on STATA command(s) (you may want to use the method(s) introduced in the STATA lecture for doing this part).

(c) Before regressing wage on educ, exper, and other potential controls, we would like to make sure that all individuals in the sample have valid records of wage. To do this, our researcher has created the following three criteria: if any observation *fails* to satisfy at least one of the following conditions, it will not be considered as valid.

(#1) wage > 0 (positive earnings)

(#2) lfp $= 1$ (wives with positive earnings indeed "worked")

(#3) hours > 0 (earnings come from positive hours spent at work)

Do all observations in the dataset satisfy three conditions, (#1), (#2), and (#3)? If everything looks fine, we can move on. However, if we have any observation(s) that violate at least one of the conditions, how would you deal with such observation(s) with invalid wage before conducting regression analysis? Your answer, shown in the output screen, should be based on STATA command(s).

(d) Now, we are ready. Estimate the following regression:

$$wage = \alpha_0 + \alpha_1\ educ + \alpha_2\ exper + u$$

Interpret α_1 and α_2 from the regression table.

(e) Interpret α_0 from the regression table. How does your interpretation sound? Do you think there are any problems with α_0 that we estimated, or everything looks fine with the regression results? (Tip: to better answer this part, you might want to take a look at predicted wage of the regression model)

(f) Our researcher wants to try different regression model. In this time, she wants to regress log(wage) on educ, exper, exper^2 and other potential controls. Create new variables, named lwage and exper2, that satisfy lwage = log(wage) and exper2 = exper^2, respectively.

(g) Using your answer in (f), estimate the following regression:

$$log(wage) = \beta_0 + \beta_1\ educ + \beta_2\ exper + \beta_3\ exper^2 + u$$

Interpret estimated coefficients from the regression table. Keep in mind that the dependent variable is lwage, instead of wage. Your answer should be precise and concise.

(h) Our researcher argues that the second regression (one with log(wage) and exper^2) is preferred to the other in studying the casual effects of wife's education attainment and labor market experience on wife's earnings in the labor market. What would be her justifications of that statement? You can use regressions results as well as class materials to answer this part. Be precise and concise.

(i) Armed with the answer above, we maintain our focus on the regression model in (g). What is predicted value of log(wage) for the wife with 12 years of education (high-school graduate), and 14 years of labor market experience? How much does log(wage) change if she has one more year of labor market experience, *holding other things constant*? Do the same analysis for another wife with 12 years of education but instead, 24 years of experience. Are the effects of having 1 more year of experience identical between two cases? Why or why not?

(j) Another researcher suggests that education level of wife's father and mother may be potential confounders in our regression in (g). Run the OLS regression again, but this time add mother's and father's education of wives as *controls*. Interpret your results. How have the coefficients of educ, exper, and exper^2 changed? Do you suspect any of the controls to be redundant?

(k) Other researcher points out that the researcher should also include unemployment and largecity as additional *controls*. How would you respond to the comment? Justify your answer based on what you have done in this section so far.

3. We now return to the topic of regular exercise and mental health disorders. The file BFRSS2015.dta, which is available for download on the website, includes a subset of variables from the 2015 BFRSS (Behavioral Risk Factor Surveillance System) survey conducted by Center for Disease Control. You will need the 2015 BRFSS Annual Survey Codebook (in pdf file) and STATA to answer the following questions. For simplicity, ignore sample weights (_llcpwt) from any calculations to produce the answers.

   (a) What are the numbers of observations and variables in the data set? Are there any string variables in the dataset?

   (b) We would like to create a variable that measures the severity of mental health disorders. Generate a new variable, mhd, which satisfies the followings:

$$\begin{aligned} \text{mhd} = \text{menthlth} \quad & \text{if} \quad 1 \leq \text{menthlth} \leq 30 \\ = 0 \quad & \text{if} \quad \text{menthlth} = 88 \\ = . \text{ (missing)} \quad & \text{if} \quad \text{menthlth} = 77, 99 \end{aligned}$$

   Likewise, we would like to create a variable that measures how often the respondents are engaged in regular exercise. Generate a new variable, regexr, which satisfies the followings:[1]

$$\begin{aligned} \text{regexr} = (\text{exeroft1} - 100) \times 4.3 \quad & \text{if} \quad 101 \leq \text{exeroft1} \leq 199 \\ = (\text{exeroft1} - 200) \quad & \text{if} \quad 201 \leq \text{exeroft1} \leq 299 \\ = . \text{ (missing)} \quad & \text{if} \quad \text{exeroft1} = 777, 999, \text{ or } . \text{ (missing values)} \end{aligned}$$

   (c) Before moving on, make sure that mhd and regexr are created properly as instructed. At this point, some observations with missing regexr are from those who do not exercise at all. We would like to code such observations as "zero exercise." To do this, using replace command, impose an additional condition for regexr as follows:

$$\text{regexr} = 0 \quad \text{if} \quad \text{exerany2} = 2$$

   (d) Provide summary statistics of mhd and regexr.

   (e) As additional summary statistics, our researcher wants to know which states are under higher risks of mental health issues among residents. Using a variable _state and appropriate STATA command(s), list and name the top-5 states, excluding Guam and Puerto Rico, that exhibit

[1] 4.3 is multiplied to reports in weekly frequency, assuming that 1 month equals to 4.3 weeks over the calendar year.

higher average mhd level than the other states (you may want to use the method(s) introduced in the STATA lecture for doing this part).

(f) Generate another variable, named ageold, which satisfies the followings:

$$\begin{aligned} \text{ageold} &= 1 && \text{if} \quad \text{\_age65yr} = 2 \\ &= 0 && \text{if} \quad \text{\_age65yr} = 1 \\ &= . \text{ (missing)} && \text{if} \quad \text{\_age65yr} = 3 \end{aligned}$$

What does ageold measure? Calculate the fraction of valid observations (i.e., non-missing ageold) that have ageold = 1.

(g) Now, we are ready. Estimate the following regression:

$$mhd \;\; = \;\; \beta_0 + \beta_1 \; regexr + \beta_2 \; regexr \cdot ageold + \beta_3 \; female + u$$

where female = 1 if the respondent is female, and female = 0 otherwise. Make sure that any missing values in dependent and explanatory variables are handled properly while running regression. Interpret estimated coefficients from the regression table. Keep in mind that units of measurement for mhd and regexr are inherited from the original variables, menthlth and exeroft1, respectively.

(h) In light of the results you found above, can you conclude that more regular exercise can reduce the incidence of mental health disorders, in particular to senior people with age 65 years and more? Justify your answer. Be precise and concise.