Michelmore – PubPol 639 Augmented

Assignment # 2

Due Friday, February 9th, 2024 11:59pm

Please upload your assignment to Canvas. For all questions requiring you to use Stata/R, please include your Stata/R output as part of your homework.

*Content*

The purposes of this exercise is to observe and interpret the results of random assignment, conduct bivariate regression analysis and interpret regression results. This assignment requires you to use the dataset posted on Canvas titled “Earnings\_and\_Height”

*Collaboration*

*With your study group* you can discuss the problem set and work on Stata/R. Note that you’ll learn Stata/R faster if you work alone sometimes.

*By yourself*, you will write up the *content* of the problem set, in your own words. Do *not* exchange these typed answers with your study group. Do *not* work on a shared Word file.

**Part I: Data Analysis**

The dataset "Earnings\_and\_Height.dta" contains data from the 1994 National Health Interview Survey. This is a subset of the data used in the Anne Case and Christina Paxson paper “Stature and Status: Height, Ability, and Labor Market Outcomes”, *Journal of Political Economy*, 2008, 116(3): 499-532. The dataset contains information on 17,870 workers. A more detailed codebook is provided at the end of the problem set.

1. **Explore and describe the data (using Stata or R)**
   1. Describe the variation in earnings. What is the mean, median, min, max, 25th percentile, 75th percentile? (Note: In addition to pasting Stata or R output, please interpret what you find using 1-2 sentences).
   2. Describe the variation in years of schooling. What is the mean, median, min, max, 25th percentile, 75th percentile? (Note: In addition to pasting Stata or R output, please interpret what you find using 1-2 sentences).
   3. What is the correlation between earnings and years of schooling? Interpret the correlation using a few sentences (e.g. is it positive or negative? Weak or strongly correlated?)
2. **Regression analysis I**
   1. What is the unconditional relationship between earnings and years of education? As a first step, use a bivariate regression to regress earnings on years of education and paste the output below.
   2. Write out the equation for the sample regression line that corresponds to this regression. You should include your estimates for and in the equation.
   3. Find and label the following on your Stata or R output and interpret these statistics with words. (You should paste your Stata or R output here and clearly label each of the following.) It might be easiest for you to simply highlight each object below with a different color in the output.
      1. 
      2. Standard error of 
      3. t-test for null hypotheses that 
      4. 
      5. R2
   4. Based on this regression, what is the predicted earnings for someone with 12 years of schooling? What about 16 years of schooling? Show your work.
3. **Regression analysis II**
4. Regress earnings on the variable indicating the sex of the worker (check the codebook below to verify the values of this variable).
5. Write out the equation for the sample regression line that corresponds to this regression.
6. What is the “omitted” category in this regression?

1. Find and label the following on your Stata or R output and interpret these statistics. (You should paste your Stata/R output here and clearly label each of the following.)
   * 1. 
     2. Standard error of 
     3. t-test for null hypotheses that 
     4. 
     5. R2
2. Based on this regression, what is the predicted earnings for a worker who identifies as male? As female? Show your work.
3. **Model fit**

Based on the two models above, which characteristic – education or sex – explains more of the variation in earnings in this data?

1. **Introduction to multiple regression**

In assessing the relationship between earnings and education, you might want to control for several possible confounders. Create a **table (in Excel or Word or some other spreadsheet tool) that looks like Table 7.1 from S&W** (though you can simply report the R-squared and not the adjusted R-squared or SER). Each column should correspond to a separate regression and each row should correspond to a different variable, with education in the first row. Each cell of the table should report the coefficient and the standard error for the given variable and specification. The table should contain the follow specifications:

(Hint: to control for years of experience, generate a new variable, exp, where exp=age-educ-6; this is a fairly standard way of measuring potential years of experience in labor economics—we take a person’s age, subtract the number of years they were in school, and subtract 6 to account for school starting age, the balance is the expected number of years of work experience).

Column 1: Regress earnings on education (same model estimated above)

Column 2: Regress earnings on education and years of experience

Column 3: Regress earnings on education and years of experience and sex

Column 4: Regress earnings on education and years of experience and sex and height and whether the individual is married

1. Discuss the change in magnitude and statistical significance of the coefficient on educ from column 1 to 2.
2. Interpret the coefficient on exp in column 2. Please discuss the magnitude and statistical significance.
3. From the results in column 2, would you argue that education or experience matters more for earnings? Why?
4. Discuss the difference between the R-squared across columns 1 through 4.
5. In about 2 sentences, describe the relationship between earnings and education and whether including these controls changes this relationship much.

**Codebook for Earnings\_and\_Height.dta**

sex – indicator variable, 0=female, 1=male

age – continuous variable denoting age, ranges from 25-65

mrd – marriage status:

1: married, sps in hh

2: married, sps not in hh

3: widowed

4: divorced

5: separated

6: never married

educ – individual’s years of education

cworker – class of worker

1: private company

2: federal government employee

3: state government employee

4: local government employee

5: incorporated business

6: self-employed

region – region of individual

1: Northeast

2: Midwest

3: South

4: West

Race – race/ethnicity

1: Non-Hispanic white

2. Non-Hispanic Black

3. Hispanic

4. Other

earnings – annual earnings in dollars

height – height in inches

weight – weight in pounds

Indicator variables created from categorical variables:

married – indicator variable, 0=not married, 1=married

black- indicator variable, 0=not Black, 1=Black

hisp- indicator variable, 0=not Hispanic, 1=Hispanic

white- indicator variable, 0-not white, 1=white

**Stata and R Code**

**Stata:**

Regression: reg earnings educ, robust

· Note: additional predictors can be added to run multivariate models

Generate experience proxy variable: gen exp = age-educ-6

**R:**

Regression:

m1 <- lm\_robust(earnings ~ educ, data=ps2\_df)

· Note: To run lm\_robust, install and load the estimatr library.

· Note: additional predictors can be added to run multivariate models

#generate experience proxy variable

ps2\_df<- ps2\_df %>% mutate(exp=age-educ-6)

*Creating Regression Tables*

You may just paste the output into excel and make a table that way, or use Stata or R to generate a table for you:

Stata:

The command estout will output regression results into a regression table structured in the format requested. You will need to install the estout package first:

ssc install estout, replace

Then, run the following command after you run each regression in the problem set, changing the name and title of each regression accordingly:

estimates store reg1, title(Regression 1)

Once you have run all your regressions, use the code below to compile your table. See the “Regression Tables in Stata” slides on canvas for detailed instructions on how to modify your table to include all required elements.

estout reg1 reg2 reg3 reg4, cells(b(star fmt(3)) se(par fmt(2))) starlevels(\* 0.10 \*\* 0.05 \*\*\* 0.01) legend stats(r2 N, labels("R-Squared" "Observations")) label collabels(“”)

R:

The library/command “stargazer” allows you to create regression tables in the format requested.

install.packages(“stargazer”)

install.packages(“lmtest”)

install.packages(“sandwich”)

library(“stargazer”)

library(“lmtest”)

library(“sandwich”)

m1 <- lm(y ~ x1, data=ps3\_df)

# Adjust standard errors

cov1 <- vcovHC(m1, type = "HC1")

r\_se1 <- sqrt(diag(cov1)) # Robust SEs

· Note: lm\_robust does not work with stargazer. This is why we correct SEs with Base R commands.

m2 <- lm(y ~ x1+ x2, data=ps3\_df)

# Adjust standard errors

cov2 <- vcovHC(m2, type = "HC1")

r\_se2 <- sqrt(diag(cov2)) # Robust SEs

stargazer(m1, m2, se=list(r\_se1, r\_se2), type = "text", header=FALSE, omit.stat = c("f","ser"), digits=2,column.sep.width="-5pt",title="Title",dep.var.caption="")

Note: including dep.var.lablem =”labels”, covariate.labels = “labels”, and notes = “notes” in your stargazer function as well will allow you to further customize and format the regression table.

The option for “type” can be replaced for html or latex if you are using RMarkdown.