Econometrics HW #3

Ryan Safner

Due: Wednesday, November 7, 2018

Theory & Concepts

For the following questions, please answer the questions completely but succinctly (2-3 sentences).

1. In your own words, describe what omitted variable bias means. What are the two conditions for an omitted variable to cause a bias?

---

2. In your own words, describe what multicollinearity means. What is the cause, and what are the consequences of multicollinearity? How can we measure multicollinearity and its effects? What happens if multicollinearity is *perfect*?

---

3. In your own words, describe what a proxy variable is. When or why would we use a proxy variable, and what effects does it have on our model?

4. A recent study found that the death rate for people who sleep 6 to 7 hours per night is lower than the death rate for people who sleep 8 or more hours. The 1.1 million observations used for this study came from a random survey of Americans aged 30 to 102. Each survey respondent was tracked for 4 years. Based on the survey, would you recommend Americans who sleep 9 hours per night consider reducing their sleep to 6 or 7 hours if they wish to prolong their lives? Be specific.

Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use `R` to verify your answers, but you are expected to reach the answers in this section "manually."

5. Data were collected from a random sample of 220 home sales from a community in 2017.

$$\widehat{Price} = 119.2 + 0.485BDR + 23.4Bath + 0.156Hsize + 0.002Lsize + 0.090Age$$

- $Price$: selling price (in \$1,000s)
- BDR: number of bedrooms
- $Bath$: number of bathrooms
- $Hsize$: size of the house (in ft^2)
- $Lsize$: lot size (in ft^2)
- Age: age of the house (in years)

a. Suppose that a homeowner converts part of an existing living space in her house to a new bathroom. What is the expected increase in the value of the house?

b. Suppose a homeowner adds a new bathroom to her house, which also increases the size of the house by 100 square feet. What is the expected increase in the value of the house?

c. Suppose the R^2 of this regression is 0.727. Calculate the adjusted $\bar{R}^2$.

d. Suppose the following auxiliary regression for BDR has an R^2 of 0.841.

$$\widehat{BDR} = \delta_0 + \delta_1 Bath + \delta_2 Hsize + \delta_3 Lsize + \delta_4 Age$$

Calculate the Variance Inflation Factor for BDR and explain what it means.

6. A researcher wants to investigate the effect of education on average hourly wages. Wage, education, and experience in the dataset have the following correlations:

	Wage	Education	Experience
Wage	1.0000		
Education	0.4059	1.0000	
Experience	0.1129	-0.2995	1.0000

She runs a simple regression first, and gets the results:

$$\widehat{\text{Wage}} = \underset{(0.7255)}{-0.9049} + \underset{(0.0613)}{0.5414}\,\text{Education}$$

She then runs another regression:

$$\widehat{\text{Experience}} = \underset{(2.6678)}{35.4615} - \underset{(0.1974)}{1.4681}\,\text{Education}$$

a. If the *true marginal effect* of experience on wages (holding education constant) is 0.0701, calculate the omitted variable bias in the first regression caused by omitting experience. Does the estimate of $\hat{\beta}_1$ in the first regression overstate or understate the effect of education on wages?

b. Knowing this, what would be the *true effect* of education on wages, holding experience constant?

c. The R^2 for the second regression is 0.0897. If she were to run a better regression including both education and experience, how much would the variance of the coefficients on education and experience increase? Why?

`R` Problems

Answer the following problems using `R`. If using `R Markdown`, simply create code chunk(s) for each question and be sure all input code is displayed (i.e. `echo=TRUE`) and feel free to just turn in a single `html` or `pdf` output file for your entire homework.

If you are NOT using `R Markdown`, please follow our standard procedure: Attach/write the answers to each question on the same document as the previous problems, but also include a printed/attached (and commented!) *.R* script file of your commands to answer the questions.

7. Download the `speeding_tickets.csv` dataset from Blackboard (under Data). This data comes from a paper by Makowsky and Strattman (2009) that we will examine later. Even though state law sets a formula for tickets based on how fast a person was driving, police officers in practice often deviate from that formula. This dataset includes information on all traffic stops. An amount for the fine is given only for observations in which the police officer decided to assess a fine.

- `Amount`: Amount of fine assessed for speeding
- `Age`: Age of speeder in years
- `MPHover`: Miles per hour over speed limit

We want to know if younger people get more speeding tickets.

a. Make a scatterplot of the amount of the fine and a driver's age.

b. Run a regression of the `Amount` of the fine on `Age`. Write the equation of the estimated OLS regression, placing standard errors in parentheses below the estimated coefficients. Interpret the coefficient on `Age`.

c. How big would the difference in expected fine be between two drivers, aged 18 and 40? Is Age likely to be endogenous?

d. Now run the regression again, controlling for speed. Write the new regression equation in the same format as before. Interpret the coefficient on age, and what has happened to it (and its statistical significance)? Interpret the coefficient on speed.

e. How big is the difference for those same two drivers, who both went 10 MPH over the speed limit?

____

____

f. How about the difference between two 18 year-olds, one who went 10 MPH over, and one who went 30 MPH over?

____

____

g. Are younger people tending to drive faster? Run an auxiliary regression of MPHover on Age and interpret the coefficient on Age. How much faster or slower than the speed limit would we expect an 18 year-old and a 40 year-old to drive?

____

____

h. Suppose we had a lot less data to work with. Use only the first 1,000 observations (reassign `your.df.name2<-your.df.name[1:1000,]` to select the first 1000 rows only) and rerun the regression from part (d). Does this bias the results? What happens to the standard errors? Why? Hint: think about the formula for variance of OLS estimators.

____

____

i. Make a nice regression table of your regressions from parts b, d, and h using `stargazer`.

____

____

8. Download the `HeightWages.csv` dataset from Blackboard (under Data). This data is a part of a larger dataset from the National Longitudinal Survey of Youth (NLSY) 1979 cohort: a nationally representative sample of 12,686 men and women aged 14-22 years old when they were first surveyed in 1979. They were subsequently interviewed every year through 1994 and then every other year afterwards. There are many included variables, but for now we will just focus on:

- `wage96`: Adult hourly wages ($/hr) reported in 1996
- `height85`: Adult height (inches) reported in 1985
- `height81`: Adolescent height (inches) reported in 1981

We want to figure out what is the effect of height on wages (e.g. do taller people earn more on average than shorter people?)

a. Create a quick scatterplot between `height85` (as X) amd `wage96` (as Y).

____

____

b. Regress wages on adult height. Write the equation of the estimated OLS regression, placing standard errors in parentheses below the estimated coefficients. Interpret the coefficient on `height85`. How much would someone who is 5'10" be predicted to earn per hour, according to the model?

____

____

c. Would adolescent height cause an omitted variable bias if it were left out? Explain using both your intuition, and some statistical evidence with `R`. Hint for `R`: you will want to subset your data just to select the 3 variables we want. Then when you run your command, you will also want to add `, use="complete.obs"` since there are many missing data values.

____

____

d. Now add adolescent height to the regression, and write the new regression equation below, as before. Interpret the coefficient on `height85`. How much would someone who is 5'10" in 1985 and 4'8" in 1981 be predicted to earn, according to the model?

____

____

e. What happened to the estimate on `height85` and its standard error?

____

____

f. Is there multicollinearity between `height85` and `height81`? Explore with a scatterplot. To avoid overplotting, use `geom_jitter()` instead of `geom_point()` to get a better view of the data.

g. Quantify how much multicollinearity affects the variance of the OLS estimates on both heights. You'll need the `car` package.

h. Reach the same number by running an auxiliary regression.

Note there are some observations that have data on heights, but not the person's wage (`R` records missing data as `NA`). We only care about observations that have data for both heights and wages, so first create a new dataframe in `R` with only observations with non-missing values for `wage96` via `df.name.2<-df.name[!is.na(df.name$wage96),]` where `df.name` is the name of the dataframe you initially created. This tells `R` to create a new object `df.name.2` by subsetting `df.name` to include only observations for which `wage96` is not `NA`. [I know this is a bit complicated, but you will see every data set has its own unique challenges!] Otherwise you will get a different answer for VIF.

i. Eye color is omitted from this model. Is this likely to be a problem?

j. IQ is omitted from this model. Is this likely to be a problem?

k. Report your regression results from parts b and d with `stargazer`.