Problem Set 4

ECON 480 - Fall 2019

Due by Thursday, October 10, 2019

Instructions

For this problem set, you may submit handwritten answers on a plain sheet of paper, or download and type/handwrite on the PDF.

Alternatively, you may download the `.Rmd` file, do the homework in markdown, and email to me a single `knit`ted `html` or `pdf` file (and be sure that it shows all of your code).

You may work together (and I highly encourage that) but you must turn in your own answers. I grade homeworks 70% for completion, and for the remaining 30%, pick one question to grade for accuracy - so it is best that you try every problem, even if you are unsure how to complete it accurately.

Theory and Concepts

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. Explain how we use Directed Acyclic Graphs (DAGs) to depict a causal model: what are the two criteria that must hold for identifying a causal effect of X on Y? When should we control a variable, and when should we *not* control a variable?

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.

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

- $Price$: 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		
Experience	0.1129	−0.2995	1.0000

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

$$\widehat{\text{Wage}} = -0.9049 + 0.5414\, Education$$

She runs another regression, and gets the results:

$$\widehat{\text{Experience}} = 35.4615 - 1.4681\, Education$$

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 Questions

Answer the following questions using `R`. When necessary, please write answers in the same document (knitted `Rmd` to `html` or `pdf`, typed `.doc(x)`, or handwritten) as your answers to the above questions. Be sure to include (email or print an `.R` file, or show in your knitted `markdown`) your code and the outputs of your code with the rest of your answers.

11. Download the `heightwages.csv` dataset. 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. Interpret the coefficient on `height85`.
c. How much would someone who is 5'10" be predicted to earn per hour, according to the model?
d. Would adolescent height cause an omitted variable bias if it were left out? Explain using both your intuition, and some statistical evidence with `R`.
e. Now add adolescent height to the regression, and write the new regression equation below, as before. Interpret the coefficient on `height85`.
f. How much would someone who is 5'10" in 1985 and 4'8" in 1981 be predicted to earn, according to the model?
g. What happened to the estimate on `height85` and its standard error?
h. Is there multicollinearity between `height85` and `height81`? Explore with a scatterplot.**[1]
i. Quantify how much multicollinearity affects the variance of the OLS estimates on both heights.[2]
j. Reach the same number as in part I by running an auxiliary regression.[3]
k. Make a regression table from part B and D using `huxtable`.

[1] Hint: to avoid overplotting, use `geom_jitter()` instead of `geom_point()` to get a better view of the data.
[2] Hint: You'll need the `car` package.
[3] Hint: There's some missing `wage96` data that may give you a different answer, so `filter()` your data here by `!is.na(wage96)` before running this regression - this will include only observations for `wage96` that are not `NA`'s.