Problem Set 1: OLS Review

EC 421: Introduction to Econometrics

Due *before* midnight on Sunday, 19 April 2020

DUE Upload your answer on Canvas *before* midnight on Sunday, 19 April 2020.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
2. the R script you used to generate your answers. Each student must turn in her/his own answers.

README! The data in this problem set come from the 2018 American Community Survey (ACS), which I downloaded from IPUMS. The last page has a table that describes each variable in the dataset(s).

OBJECTIVE This problem set has three purposes: (1) reinforce the metrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics/regression.

INTEGRITY If you are suspected of cheating, then you will receive a zero. We may report you to the dean.

Setup

Q01. Load your packages. You'll probably going to need/want `tidyverse` and `here` (among others).

Q02. Now load the data. I saved the same dataset as two different formats:

- an `.rds` file: use a function that reads `.rds` files—for example, `readRDS()` or `read_rds()` (from the `readr` package in the `tidyverse`.
- a `.csv` file: use a function that reads `.csv` files—for example, `read.csv()` or `read_csv()` (from the `readr` package in the `tidyverse`.

Q03. Check your dataset. How many observations and variables do you have? *Hint:* Try `dim()`, `ncol()`, `nrow()`.

Getting to know your data

Q04. Plot a histogram of households' income (variable: `hh_income`). *Note:* Household income is in tens of thousands of dollars (so a value of `8` implies an income of $80,000.)

Q05. What are the mean and median levels of household income? Based upon this answer and the previous histogram, is household income (fairly) evenly distributed or is it more skewed? Explain your answer.

Q06. Create a histogram of household income per capita—meaning the household's income divided by the number of individuals in the household. Does dividing by the number of individuals in the household change your understanding of the income distribution? Explain your answer.

Q07. Run a regression that helps summarize the relationship between household income and household size. Interpret the results of the regression—the meaning of the coefficient(s). Comment on the coefficient's statistical significance.

Q08. Explain why you chose the specification you chose in the previous question.

- Was it linear, log-linear, log-log?
- What was the outcome variable?
- What was the explanatory variable?
- Why did you make these choices?

Q09. Plot a histogram of the time households spend commuting each day (the variable `time_commuting` is the average commuting time for a household). Is the distribution of commute time more or less equitable than income? Explain.

Regression refresher: Varying the specification

Q10. Linear specification Regress average commute time (`time_commuting`) on household income (`hh_income`). Interpret the coefficient and comment on its statistical significance.

Q11. Log-linear specification Regress the log of average commute time on household income. Interpret the coefficient and comment on its statistical significance.

Q12. Log-log specification Regress the log of average commute time on the log of household income. Interpret the coefficient and comment on its statistical significance.

Multiple linear regression and indicator variables

Q13. Regress average commute time on household income **and** the share of the individuals in the household who are non-white ethnicities (`hh_share_nonwhite`). Interpret the intercept and coefficient and comment on their statistical significance. Also compare your results to Q10. Has anything changed?

Q14. Regress average commute time on the indicator variable for whether a household moved in the last year (`i_moved`). Interpret the intercept and coefficient and comment on their statistical significance.

Q15. Add the share of the household that represents a non-white ethnicity (`hh_share_nonwhite`) to the regression in Q14. *Note:* Your outcome variable is still average household commute time, but you should now have two explanatory variables. Interpret the intercept and coefficient and comment on their statistical significance.

Q16. Did adding this second explanatory variable change the coefficient of the first variable at all? What does that tell you? Explain your answer.

Q17. Now add the interaction between your two explanatory variables in Q16 and re-run the regreation. (You should have an intercept and three coefficients—the two variables plus their interaction.) Interpret the coefficient on the interaction and comment on its statistical significance.

Q18. Did including the interaction change your understanding of the relationship between the variables? Explain.

Q19. Regress the indicator for whether the household has a smartphone (`i_smartphone`) on the household's income (`hh_income`) and the share of the household's individuals who represent non-white ethnicities (`hh_share_nonwhite`). Interpret the intercept and coefficients. Comment on their statistical significance.

The bigger picture

Q20. In the last regression (Q19), should we be concerned about omitted-variable bias? Explain your answer and provide an example of a potential omitted variable if you are concerned about omitted-variable bias.

Q21. Is R-squared a good measure of model performance? Explain your answer.

Q22. Define the term *standard error*.

Q23. What does our assumption of *exogeneity* require?

Q24. What does it mean for an estimator to be *unbiased*?

Q25. What does it mean for an estimator to be *more efficient than another estimator*?

Description of variables and names

Variable	Description
`fips`	County FIPS code
`hh_size`	Household size (number of people)
`hh_income`	Household total income in $10,000
`cost_housing`	Household's total reported cost of housing
`n_vehicles`	Household's number of vehicles
`hh_share_nonwhite`	Share of household members identifying as non-white ethnicities
`i_renter`	Binary indicator for whether any household members are renters
`i_moved`	Binary indicator for whether a household member moved in prior 1 year
`i_foodstamp`	Binary indicator for whether any household member participates in foodstamps
`i_smartphone`	Binary indicator for whether a household member owns a smartphone
`i_internet`	Binary indicator for whether the household has access to the internet
`time_commuting`	Average time spent commuting per day by each household member (minutes)

In general, I've tried to stick with a naming convention. Variables that begin with `i_` denote binary indicatory variables (taking on the value of `0` or `1`). Variables that begin with `n_` are numeric variables.