Problem Set 1: OLS Review

EC 421: Introduction to Econometrics

Due *before* midnight on **Saturday, 30 January 2021**

DUE Upload your answer on Canvas *before* midnight on Saturday, 23 January 2021.

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.

If you are using RMarkdown, you can turn a single file, but it must be a `html` or `pdf` file with **both** your R code **and** your 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. **Cheating includes copying from your classmates, from the internet, and from previous assignments.**

Setup

Q01. Load your R packages. You're 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 individuals' personal income (variable: `personal_income`). *Note:* Household income is in tens of thousands of dollars (so a value of `3` implies an income of $30,000.)

Don't forget to label your plot's axes. A title wouldn't be be, either.

Q05. Compare the distributions of personal income for (1) women *vs.* men and (2) black individuals *vs.* white individuals. Are the differences at the extremes of the distribution or at the center (*e.g.*, mean and median)?

Note: Your answer should include four histograms (women, men, black, and white).

Hints

- There is an indicator for female in the data called `i_female`. There are also indicators for *black* and *white* names `i_black` and `i_white`.
- You can take a subset of a variable using the `filter()` variable from the `tidyverse`. *E.g.*, to take find all married individuals in the `ex_df` dataset, you could use `filter(ex_df, i_married == 1)`.

Q06. Create a scatterplot (AKA: dot plot) with commute time (`time_commuting`, which the length of the individual's morning commute, in minutes) on the `y` axis and personal income on the `x` axis.

Q07. Based upon your plot in **Q06**: If we regressed commute time on income, do you think the coefficient on income would be *positive* or *negative*? **Explain** your answer.

Q08. Run a regression that helps summarize the relationship between commute length and personal income. Interpret the results of the regression—the meaning of the coefficient(s). Comment on the coefficient's statistical significance.

Q09. 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?

Regression refresher: Varying the specification

Note: In this section, when I ask you to "comment on the statistical significance," I want you to tell me whether the coefficient is significantly different from zero at the 5% level. You do not need write out the full hypothesis test.

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

Q11. Did the sign of the coefficient on personal income surprise you based upon your figure in **06**? Explain.

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

Q13. 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

Note: We're now moving to thinking about the time at which an individual leaves her home to go to work (`time_depart`). This variable is measured in minutes from midnight (so smaller values are earlier in the day).

Q14. Regress departure time (`time_depart`) on the indicator for female (`i_female`) **and** the indicator for whether the individual was married at the time of the sample (`i_married`). Interpret the intercept and **both** coefficients (commenting on their statistical significances).

Q15. What would need to be true for `age` to cause omitted-variable bias. Explain the requirements and whether you think they are likely to cause bias in this setting.

Q16. Add `age` to the reression you ran in **Q14**. Do the results of this new regression suggest that `age` was causing omitted-variable bias? Explain your answer.

Q17. Now regress departure time on `i_female`, `i_married`, **and their interaction**. (You should have an intercept and three coefficients: the two variables and their interaction.) Interpret the coefficient on the interaction and comment on its statistical significance.

Hint: In R you can get an interaction using `:`, for example, `lm(y ~ x1 + x2 + x1:x2, data = fake_df)`.

Q18. For this last regression, we are going to do something totally different. Our outcome variable is going to be an indicator for whether the individual has internet access (`i_internet`). Regress this internet-access variable on a two explanatory variables: (1) an indicator for whether the household's location is considered urban `i_urban` (vs. rural) and (2) an indicator for whether the individual is black (`i_black`).

Interpret the intercept and coefficients. Comment on their statistical significance.

The bigger picture

Write short answers to each of these questions. No math needed: Just explain your reasoning.

Q19. Define the term *standard error*.

Q20. For exogeneity, we write $E[u|x] = 0$. What does this mathematical expression mean for the relationship between u and x?

Description of variables and names

Variable	Description
`state`	State abbreviation
`i_urban`	Binary indicator for whether home county is 'urban'
`age`	The individual's age (in years)
`i_asian`	Binary indicator for whether the individual identified as Asian
`i_black`	Binary indicator for whether the individual identified as Black
`i_hispanic`	Binary indicator for whether the individual identified as Hispanic
`i_indigenous`	Binary indicator for whether the individual identified with a group indigenous to North Am.
`i_white`	Binary indicator for whether the individual identified as White
`i_female`	Binary indicator for whether the individual identified as Female
`i_male`	Binary indicator for whether the individual identified as Male
`i_grad_college`	Binary indicator for whether the individual graduated college (estimated)
`i_married`	Binary indicator for whether the individual was married at the time of the sample
`personal_income`	Total (annual) personal income (tens of thousands of dollars)
`i_foodstamps`	Binary indicator for whether the individual uses 'foodstamps' (SNAP)
`i_health_insurance`	Binary indicator for whether the individual has health insurance
`i_internet`	Binary indicator for whether the individual has access to the internet
`time_depart`	The time that the individual typically leaves for work (in minutes since midnight)
`time_arrive`	The time that the individual typically arrives at work (in minutes since midnight)
`time_commuting`	The length of time that the individual typically travels to work (in minutes)

I've tried to stick with a naming convention. Variables that begin with `i_` denote binary/indicator variables (taking on the value of `0` or `1`).