Problem Set 1

Econometrics Review

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Sunday, 24 April 2021

DUE Your solutions to this problem set are due *before* 11:59pm on Sunday, 24 April 2021 on Canvas. **Your answers must include two files** (**1**) your responses/answers to the question (*e.g.*, a Word document) and (**2**) the R script you used to generate any answers in R. 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.**

Getting started

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 terrible, 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_male`) **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_male`, `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 = not_a_real_df)`.

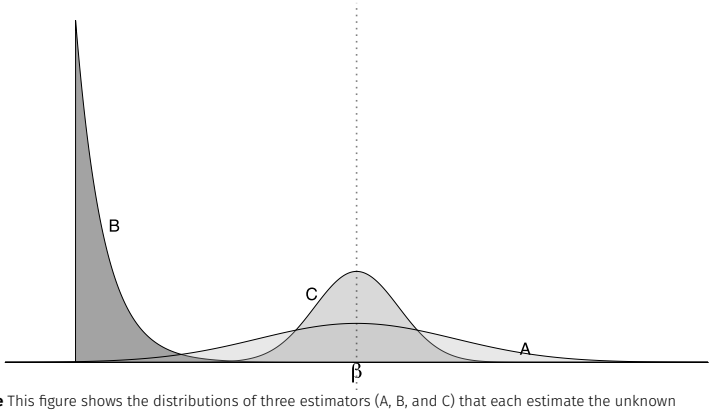
Q18. For this last regression, we are going to do something totally different. Our outcome variable is going to be an indicator for whether the individual has internet access (`i_internet`). Regress this internet-access variable on a two explanatory variables: (1) an indicator for whether the household's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

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

The bigger picture

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

Figure 1



Note This figure shows the distributions of three estimators (A, B, and C) that each estimate the unknown parameter β. E[A]= $\beta - 3$, E[B]= β, E[C]= β

Q19a. Are the estimators in Figure 1 (above) unbiased? Which ones? *Hint:* There may be more than one.

Q19b. Which of the estimators in Figure 1 (above) has the minimum variance?

Q19c. Which of the estimators in Figure 1 (above) is the best (minimum variance) unbiased estimator?

Q20. What is the definition of a *standard error*.

Q21. Exogeneity is written as $E[u|x] = 0$. What does this mathematical expression mean for the relationship between u and x?

Q22. Imagine - on a die throw where d is our die random variable and c is a coin flip with the results `heads/tails` and you know $E[d|c = heads] = 3$. If you had to pay $2 to play this game, would you expect to earn money on this game? Assume nothing about the relationship between the coin and the die.

Q23. Throughout this course, we will use the OLS estimator $\hat{\beta}$ to estimate β. Explain what it means for $\hat{\beta}$ to be biased for β.

Variable	Description
state	State abbreviation
age	The individual's age (in years)
i_urban	Binary indicator for whether home county is 'urban'
i_citizen	Binary indicator for whether the individual is a citizen (naturalized or born.)
i_noenglish	Binary indicator for whether the individual speaks English
i_only_english	Binary indicator for whether the individual speaks ONLY English
i_drive_to_work	Binary indicator for whether the individual drives to work or takes a personal car
i_asian	Binary indicator for whether the individual identified as Asian
i_black	Binary indicator for whether the individual identified as Black
i_indigenous	Binary indicator for whether the individual identified as Hispanic
i_white	Binary indicator for whether the individual identified with a group indigenous to North Am.
i_female	Binary indicator for whether the individual identified as White
i_male	Binary indicator for whether the individual identified as Female
i_grad_college	Binary indicator for whether the individual identified as Male
i_grad_highschool	Binary indicator for whether the individual graduated college
i_married	Binary indicator for whether the individual graduated high school
i_married_mult	Binary indicator for whether the individual was married at the time of the sample
personal_income	Binary indicator for whether the individual has been married multiple times at the time of the sample
i_health_insurance	Total (annual) personal income (tens of thousands of dollars)
i_internet	Binary indicator for whether the individual has health insurance
time_depart	Binary indicator for whether the individual has access to the internet
time_arrive	The time that the individual typically leaves for work (in minutes since midnight)
time_commuting	The time that the individual typically arrives at work (in minutes since midnight)
weights	The length of time that the individual typically travels to work (in minutes)