Mid-course project

EC 421: Introduction to Econometrics

Due *before* midnight on Tuesday, 12 May 2020

Instructions

DUE: One member of your group must upload your answer on Canvas *before* midnight on Tuesday, 12 May 2020. All members of the group must be listed on the submission.

IMPORTANT: As with your homework, 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 submit a single file.

README! The last page has a table that describes each variable in the dataset (`proj1.csv`).

INTEGRITY: Groups can either have **one or two members**. Only one person needs to submit your final document. If you are suspected of cheating in any way (for example, copying from someone else), then you will receive a zero. We may report you to the dean.

GRADING: Your grade for this project will be based upon the accuracy of your answers *and* how well you explain/illustrate your answers. We value short, accurate answers over long, meandering answers. Edit your answers!

EMAIL POLICY: Do not ask the GEs or the instructor for help coding or for help answering these questions. You may only ask **clarifying** questions. Use Google and the course's materials (lectures, labs, notes, assignment keys).

Questions

01. Summarize and describe the dataset. Your answer should include:

- What share of the sample is female? What share of the sample is non-white?
- How skewed are the income and education distributions?
- Create three figures (graphs) that summarize key variables.
- Create two figures (graphs) that summarize how key variables relate to each other.

Explain your decisions on summarizing the data. What do you learn about potential relationships?

02. Regress individuals' income (`income`) on an intercept and their education (`education`).

03. Create a scatter plot with the residuals from **02** on the y axis and education on the x axis.

04. Does the scatter plot from **03** suggest that **heteroskedasticity** may be present? Explain your answer.

05. More generally: Does the scatter plot from **03** suggest that there are any issues with **your specification**? Explain.

06. Explain why the regression in **02** could suffer from omitted-variable bias.

07. Give an example of an omitted variable (other than *ability*) that could cause bias in the regression in **02**. Just to be clear: Do not use the variable *ability* as your example.

- Explain how your example variable satisfies both requirements for omitted-variable bias.
- Describe the direction of the bias this variable would cause (when we estimate the effect of education on income). Explain your answer.

08. Now regress *income* on an intercept, *education*, **and** *ability*. Interpret the results.

09. Does your estimate for the effect of education on income change from question **02** to question **08**? Explain why this change (or lack of change) makes sense.

Hint: Is there a significant relationship between *education* and *ability*?

10. Up to this point, we have generally told you which regressions to run. And we've stuck with pretty simple regressions (*e.g.*, regress `y` on `x1` and `x2`). We now want you to explore the actual complexity of econometric/statistical analyses. Estimate three new models. These models should not match your previous models (in **02** and **08**). Across these three new models, you should include (at least once):

- a log-transformed variable (*i.e.*, use `log`)
- an interaction

11. How did you choose your specifications in **10**? Explain your decision making.

12. Which of your new models is "best"—if you must choose one model, which would you choose? Why?

13. For your "best" model (chosen in **12**): Interpret the coefficients and comment on their statistical significance.

14. Do you *trust* the estimates from your *best model*? Explain why/why not.

15. Suppose you want to estimate the effect of high-school graduation. How could you use the current data to estimate this effect? Describe any regressions, estimates, figures, and/or caveats you would make. *Note:* You can assume that someone with 12 years of education graduated from high school.

Variable	Description
`income`	The individual's annual income (in US dollars).
`married`	Binary indicator (1,0) for whether the individual is currently married.
`kids`	Number of children.
`nonwhite`	Binary indicator (1,0) for whether the individual belongs to a non-white ethnicity.
`female`	Binary indicator (1,0) for whether the individual is female.
`education`	Years of education.
`urban`	Binary indicator (1,0) for whether the individual's home is classified as 'urban'.
`ability`	Ability on scale of 0 (lowest) to 100 (highest).