Project

EC 421: Introduction to Econometrics

Due *before* midnight on Saturday, 22 May 2021

Instructions

DUE: One member of your group must upload your answer on Canvas *before* midnight on Tuesday, 22 May 2021. 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 (`project.csv`). That being said - this data is data you have worked with before! It's the same as the data you worked with in PS-002.

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).

Prelude

This project is more of a choose-your-own-adventure. You will get credit for particularly insightful answers and plots. The first thing you'll need to do is pick some variables you want to include in your analysis

00. Pick a set of at least **5 different variables** (feel free to include an interaction, but that interaction does not count towards the 5.) One of these variables should be your outcome variable, and I recommend you choose a continuous variable to be your outcome. You can choose a binary outcome, but interpreting certain results will be more difficult.

Questions

01. Summarize and describe the subset you chose from your dataset. Your answer should include (at a minimum):

- Distribution of the data
- (Rough) frequency of values
- Maximum/Minimum
- At least 3 informative plots

Explain your decisions on summarizing the data. What do you learn about potential relationships? Choose summary statistics that will help you analyze this data.

02. Regress your chosen outcome variable on an intercept and the chosen variables.

03. Create a scatter plot with the residuals from **02** on the y axis and a variable of your choice on the x axis (you'll get better results if this variable is continuous.)

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, or if you think it does not, justify your answer.

07. Give an example of an omitted variable that could cause bias in the regression in **02**. If there is not one - choose one of your included variables and answer as if it were left out of your regression.

- 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. Include a new variable in your regression from **02**. Interpret the results.

09. Do any of your estimates for the effect of your independent variables on outcome change from question **02** to question **08**? Explain why this change (or lack of change) makes sense.

10. So far, 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. First, pick a subset of the data that is interesting to you (*eg. Observations in the south, observations in California, only women, etc.*) Estimate three new models. These models should not match your previous models (in **02** and **08**), but you do not need to change your outcome variable, and you are not forced to include more or fewer variables (though you should justify your specifications in each case.) Across these three new models, you should include (at least once):

- a log-transformed variable (*i.e.*, use `log`) as either an outcome or as an independent variable
- 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 college graduation on your outcome of choice. How could you use the current data to estimate this effect? Describe any regressions, estimates, figures, and/or caveats you would make.

Variable	Description
`state`	State abbreviation
`age`	The individual's age (in years)
`i_urban`	Binary indicator for whether home county is 'urban'
`i_drive_to_work`	Binary indicator for whether the individual drives to work or takes a personal car
`i_physical_disability`	Binary indicator for whether the individual has a physical disability
`socioeconomic_index`	Continuous measure of socioeconomic priviledge
`poverty_pct`	Percentage of poverty line (higher = higher income)
`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
`i_grad_highschool`	Binary indicator for whether the individual graduated high school
`i_married`	Binary indicator for whether the individual was married at the time of the sample
`i_married_mult`	Binary indicator for whether the individual has been married multiple times at the time of the sample
`personal_income`	Total (annual) personal income (tens of thousands of dollars)
`personal_nonwage_income`	Total (annual) personal non-wage income (tens of thousands of dollars)
`i_health_insurance`	Binary indicator for whether the individual has health insurance
`i_moved_in_last_year`	Binary indicator for whether the individual moved to a new location
`i_moved_out_of_state_in_last_year`	Binary indicator for whether the individual moved to a new state
`lived_abroad_in_last_year`	Binary indicator for whether the individual moved to the US in the last year
`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)
`education`	Number of years in education
`ability`	Numeric measure of 'ability' (impact of education separate from socio-economic status)
`weights`	How likely a person is to be included in a sample