Problem Set 4: Instrumental Variables

PSCI 2301, Spring 2025, Prof. Brenton Kenkel

2025-03-05

This problem set is due **Wednesday, March 19** at 11:59pm. Turn it in on Brightspace, following these instructions:

1. Download the `set4_answers_template.qmd` Quarto file, and enter your responses in that file.

2. When you are done, render your responses to PDF. I recommend using the "Render on Save" option the whole time you're working on the problem set, so you can catch errors sooner rather than later.

3. Upload **both** your Quarto source file and your PDF output to Brightspace.

There's a no-questions-asked, one-time-use, 72-hour extension: If you turn the assignment in after the deadline, but before 11:59pm on Saturday, March 22, there won't be any penalty to your grade—but then you won't have this option on subsequent problem sets.

Part A: Instrumental variables concepts

Read the article "Do Conditional Cash Transfers Affect Electoral Behavior?" (Ana De La O, *American Journal of Political Science*, 2013) before answering these questions. The PDF is posted with the assignment on Brightspace as `delao2013conditional.pdf`.

1. Identify the key components of De La O's causal analysis.

   - Unit of analysis
   - Population of interest

- Treatment variable
- Outcome variable(s)

2. In De La O's analysis, the main treatment is randomly assigned. Why, then, does she see a need to employ an instrumental variable? Briefly explain in your own words why the raw difference of means might have selection bias in this analysis.

3. Identify the instrumental variable De La O uses. Then briefly explain why it would be expected to meet each of the three conditions for a valid instrument: independence, first stage, and exclusion restriction. Do you see any potential issues with any of these assumptions?

4. As we discussed in class, an instrumental variables analysis estimates a *local* average treatment effect, namely the average among compliers. What is a "complier" in the context of De La O's instrument and treatment? To what extent do you think it is reasonable to generalize from the compliers to the population at large?

5. In your own words, summarize the main conclusions of De La O's analysis, and the statistical evidence she presents for these conclusions.

Part B: Instrumental variables in practice

You will work with the replication data from the De La O 2013 article, via the file `data_set4_dlo.csv`. This file consists of the observations from the 90% of the sample used in De La O's main analysis (see page 7 of the article). It contains the following columns.

Variable name	Description
votes_1994	Raw total number of votes cast in 1994
turnout_1994	Proportion of eligible voters who voted in 1994
turnout_2000	Proportion of eligible voters who voted in 2000
pri_1994_votes	Raw number of votes for the PRI party in 1994
pri_2000_share	Vote share for the PRI party in 2000
pan_1994_votes	Raw number of votes for the PAN party in 1994
pan_2000_share	Vote share for the PAN party in 2000
prd_1994_votes	Raw number of votes for the PRD party in 1994
prd_2000_share	Vote share for the PRD party in 2000
intent_to_treat	Indicator for random assignment to early treatment
early_progresa	Share of households that enrolled early
poverty	Poverty index, 1-5 scale

Variable name	Description
population	Total population as of 1994
share_eligible	Share of population eligible for the program
num_villages	Number of villages in the precinct

1. For each of the variables listed in Table 2 (page 7 of the article), use a data visualization to assess how well-balanced it is as a function of treatment. Do you notice any substantial imbalances?

   Note: When deciding on a visualization technique, keep in mind that the treatment variable here is continuous.

   Note: The "Randomly Assigned Villages == 1" and "Randomly Assigned Villages == 2" variables mentioned in the table are not included in this version of the dataset, so you don't need to assess those.

2. For the same variables as in the previous question, assess how well-balanced they are across the two possible values of the instrument. Do you notice any substantial imbalances? How do the results here compare to what you found in the last question?

3. Use visualizations to assess the first-stage relationship between the instrument and the treatment, and the reduced-form relationship(s) between the instrument and the outcome(s). What relationships are suggested by these visualizations? Based on what you've found here, what would you guess about the direction and magnitude of treatment effects?

4. At first without including any controls, estimate the treatment effect(s) using instrumental variables, and assess their statistical significance against a null hypothesis of zero effect. How do your results compare to the findings that De La O reports in the bottom panel of Table 3 (page 8)?

5. Rerun your regressions from the last part, but including the controls that De La O mentions in the caption for Table 3. How close do your results come to hers?

   Note: To match her treatment of the "number of villages" variable, you will need to treat it as a categorical variable. You can treat a variable as categorical in a regression formula via the syntax `outcome ~ ... + factor(name_of_variable)`.

6. Change the set of controls in a way that you think is reasonable. Add a variable, subtract a variable, log-transform a variable, etc. First explain why you made the changes you did. Then rerun your regressions from the last part, but with the new control specification. How much do your results and conclusions change compared to the last step?