ECON42720 Causal Inference and Policy Evaluation

Assignment 2: IV and Regression Discontinuity

Ben Elsner

About this Assignment

This assignment teaches you how to apply a fuzzy regression discontinuity design to estimate a causal effect. As such, it is based on the material from lectures 5 (IV) and 6 (RDD).

- The assignment is due on **Friday, 3rd May 2024, 23:59**. I recommend not to wait until the last minute.
- You can work on this assignments in groups of up to 4. I will not get involved in the group formation process. If you prefer to work on your own, you can do so.
- Please submit one assignment per group on Brightspace under *Assessment > Assignments*. Make sure that all the names of the group members are on the assignment.
- The assignment should be submitted as one pdf file. It should contain a write-up with your answers to the questions (including tables and figures) and the R code you used to generate the results. Do not use code chunks; instead, provide the entire code at the end of the document.
- Formatting: I recommend using Quarto, but it's not a must. I expect nicely formatted tables and figures (use ggplot2 or similar for figures, stargazer or similar for tables). Before AI became broadly available, students could get away with ugly tables and figures because producing nices ones takes time. Not any more.
- AI Policy. I expect that you work on the solutions yourself. Feel free to use AI to help with coding and editing, but I warn against using it to come up with solutions. Please include an AI statement that explains how you used AI in your assignment.
- I will provide feedback on Brightspace to the person who submitted the assignment. Please share the feedback and the grade with the group.
- Criteria for grading: correctness of the analysis and answers, quality of the write-up, quality of the code, quality of the tables and figures, and the overall presentation of the assignment. All members of the group receive the same grade. Your grade will be on a scale from 0 to 100.

- To calculate the overall assignment grade for this module, I will take the average across both assignments and convert them to a letter grade based on the Alternative Linear Grading Scale (40% pass).

1 Tasks

The empirical application is based on a cross-sectional dataset *assign2.dta*, which contains the following variables:

- *age*: age of surveyed individual
- *logearn*: log annual earnings
- *yob*: year of birth
- *schooling*: age at which the person left school.

1.1 Instrumental Variable Strategy

A common way to obtain causal estimates is to use changes in compulsory schooling laws for identification. In this case, birth cohorts born before 1933 (yob<33) had to go to school until they were 14 years old, whereas compulsory schooling age was raised to 15 years for all cohorts born from 1933 onwards. This change in compulsory schooling laws can be used as an instrumental variable for the actual duration of schooling. The instrument is a dummy `LAW` that equals unity if a person is born 1933 or later and zero otherwise.

a) Draw a DAG that explains why a simple OLS regression of earnings on schooling yields biased estimates and why schooling reforms as an instrument can remove this bias.

b) Explain what conditions have to be fulfilled for the instrumental variable to be valid. To the extent that you can infer this from the context, discuss whether these assumptions are likely to be met in this case.

c) Explain who the compliers, always-takers, never-takers, and defiers are. If some of these groups don't exist in this context, explain why.

d) Explain why this instrumental variable strategy is akin to a fuzzy regression discontinuity design (e.g. what is the discontinuity, what is the forcing variable, why is it fuzzy, etc).

1.2 Fuzzy RDD: Graphical Inspection

Produce binscatters that illustrate the relationship between the year of birth and

- The probability that a person leaves school before age 15
- Schooling
- Log earnings

For each graph, discuss the results and their meaning for the validity of the research design.

1.3 Fuzzy RDD: Estimation

a) The goal of this analysis is to estimate the returns to education, i.e. the causal effect of one more year of education on earnings. For this purpose, estimate an OLS regression of `logearn` on `schooling`, controlling for fourth-order polynomials in age and year of birth. Interpret the coefficient of schooling.

b) Calculate the Wald estimator (without controls) "by hand' ', i.e. based on conditional averages.

c) Estimate the effect of schooling on log earnings using with 2SLS, using the indicator of whether someone was born in or after 1933 as an instrument for schooling using 2SLS. Do not include controls. Report the coefficient and standard error. Explain why the 2SLS estimate is likely to be biased without controlling for (at least) the year of birth.

d) Now estimate the returns to education with an instrumental variables estimator using the same controls as in a). Do the following:

- Estimate the first stage and reduced form and interpret the results. Compute the IV estimate from these results.
- Separately estimate the first and second stage.
- Estimate the model with an inbuilt 2SLS command such as 'ivreg'. Compare the coefficient and standard error of *schooling* to those obtained in the previous regressions in d), i.e. those based on separate estimations of the first and second stage.
- Comment on the difference between the OLS and 2SLS results. Why are they different?
- Comment on the strength of the first stage using appropriate techniques.

2 Advice

- For 2SLS estimation, I recommend the `ivreg` command from the `AER` package. There are other IV packages available, such as `ivtools` or `ivmodel`, but I recommend `AER` because it is well-documented and widely used.
- For the graphical inspection, I recommend the `binscatter` package. It is a bit tricky to install, but it is a very useful tool for RDD analysis. An alternative is the `binsreg` package, which was written by some of the godfathers of RDD. If you can't get any of these to work, you can use `ggplot2` to create the graphs.