Microeconometrics II

André Portela Souza

Due: November 12th

Part 1: Instrumental variables

You have been provided with a sample of first-born children aged between 10 and 18 years old living with both of their parents. The sample was constructed from Census 2010. Each entry in the dataset corresponds to a first-born child from **both** spouses. There can be multiple first-born children in a given family/couple if the first birth from both spouses was a multiple birth. A detailed data dictionary can be found in `dicionario.xlsx`.

Your goal is to estimate the causal impact of the number of children in the family (variable `family_number_children`) on the years of education of the couple's first-born child(ren) (variable `first_child_years_of_education`).[1]

Question 1 In this item, we will explore the identifying power of the monotone treatment response (MTR) and monotone treatment selection (MTS) assumptions.

(a) State the MTR and MTS assumptions – **in the most plausible direction** – in this context. Does economic theory have any predictions on the validity of these assumptions? *Hint:* See the discussion in Ponczek and Souza (2012).

(b) Report a table with average years of schooling by number of children, as well as the frequency of each value of variable "number of children" in the sample. You may want to use individual sample weights (variable `person_weight`) in estimation in order to better account for the population of interest.

(c) Compute the upper and lower bounds on the ATE of increasing the number of children in the family from 1 to 2, 2 to 3, 3 to 4, 4 to 5, 5 to 6 etc. Compute 95% confidence intervals to these bounds using the bootstrap. You should draw samples of **households** with replacement in order to properly account for the sampling process. You may want to set the probabilty of sampling a household proportional to the household weight (variable `household_weight`) in order to better replicate the sampling process.[2].

Question 2 In this item, we will use an instrumental-variable approach in estimating the causal impact of the number of children on years of schooling of the first child. We will follow the approach in Ponczek and Souza (2012), whereby we first restrict our sample to families with two or more births from the couple (variable `family_number_births` ≥ 2). We then propose to instrument the number of children with `second_birth_ismultiplebirth`, which indicates whether the second birth of the couple was a multiple birth.[3]

(a) Make the sample restrictions previously discussed. How many second births are multiple births?

(a) Under which assumptions does an instrumental variable approach identify a treatment effect in our setting? What treatment effect? Do these assumptions seem plausible to you? Why? Looking at the dataset, do you think any controls should be included? Why?

(b) Estimate the treatment effect using 2SLS. Include any covariates you regarded as necessary in the previous item. Cluster your standard errors at the household level (you may also want to weight observations by the person weight). Is the instrument **relevant**? Why? Comment on your results.

[1] There can be older children in the family if there is an older stepchild. These do not appear on the dataset. We are thus looking at the impact on the first-born child from **both** spouses.

[2] *Hint:* In `R`, use the option `prob` in the command `sample` seen in the first course. An easy way to proceed is to adapt the code for power calculations via simulation to your setting.

[3] The authors do not consider a dummy indicating whether the first birth was a multiple birth because twins are known to be born with different traits (e.g. lower birthweight) that may affect outcomes through other channels than family size.

(c) Conduct a test for weak instruments. Are your instruments **weak**? In what sense? *Hint:* See Section 4 in Andrews et al. (2019).

(d) Report Anderson-Rubin confidence intervals. How do they compare to (b)?

(e) Compare your results with the estimates found in Question 1.

Part 2: Regression discontinuity design

For this part of the list, you have been provided with data from Amarante et al. (2016), who studies the effect of a cash-transfer program in Uruguay on health outcomes at birth. According to the authors, "the Uruguayan Plan de Atención Nacional a la Emergencia Social (PANES) was a temporary social assistance program targeted to the poorest 10 percent of households in the country, implemented between April 2005 and December 2007." Eligibility was defined via a baseline survey conducted with applicants. A probit model for the likelihood of falling below a critical per capita income level was estimated using baseline data, and households whose predicted probability exceeded some threshold were eligible to the program. **However**, due to imperfect enforcement of the rules of the program, some noneligible mothers did actually receive the cash transfer, whereas some eligible mothers failed to do so.

You have been provided with a dataset where each entry corresponds to a pair (birth,mother) during the program duration. The treatment indicator variable is `treat`. The eligibility dummy is $\texttt{eligible} = \mathbf{1}\{\texttt{running} > 0\}$, where `running` is the predicted probability of falling to poverty, already subtracted of the threshold for program eligibility. File `dic_amarante.pdf` contains the description of additional variables in the dataset.

1. State the assumptions required for the identification of a treatment effect using the discontinuity described above. What do these assumptions mean in this context? What is the interpretation of the treatment effect identified under these assumptions?

2. Report a discontinuity plot between the running variable (x-axis) and program participation (y-axis), as well as a discontinuity plot between the running variable (x-axis) and low birthweight (variable `bajo2500`). What do these plots tell you?

3. Estimate the effect of program participation on low birthweight by local linear regression. Precisely state the bandwidth selection method used, the choice of kernel, as well as whether bias correction was employed. What is the first-stage relation, at the cutoff, between program eligibility and program participation? Is it statistically significant? Comment on your second stage results.

4. In order to assess the credibility of your empirical strategy, choose a variable which you may argue is predetermined and estimate the effect of program participation at the cutoff as in the previous item. What do you find?

5. Implement a manipulation test for the running variable in your setting. What do you find?

6. Do you think there are any potential threats to identification and/or estimation in your context? Can you think of any strategies to circumvent these?

References

Amarante, V., M. Manacorda, E. Miguel, and A. Vigorito (2016, May). Do cash transfers improve birth outcomes? evidence from matched vital statistics, program, and social security data. *American Economic Journal: Economic Policy 8*(2), 1–43.

Andrews, I., J. H. Stock, and L. Sun (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics 11*(1), 727–753.

Ponczek, V. and A. P. Souza (2012). New evidence of the causal effect of family size on child quality in a developing country. *Journal of Human Resources 47*(1), 64–106.