Project Assignment for the PhD Course in Identification and Causal Inference

ETH Zurich, Spring Semester 2021
Stefan Pichler & Michael Siegenthaler

1. Decide on a causal question that you want to answer and that will guide your discussion. This causal question should not be straightforward to answer since selection is not random. From the research FAQ, answer the following questions (as short as possible):
   (a) Which experiment (randomized control trial) could ideally be used to capture the causal effect of interest?
   (b) How would the optimal data set look like and what information would it reveal?
   (c) In the real world (with real data), what would be the *two* most important threats to identification that you would have to overcome?

2. Using your favorite statistics software, generate data for the following model[1]:

$$Y_i = \mu(0) + \Delta_i D_i + U_i(0) \tag{1}$$

$$D_i^* = \alpha + \beta Z_i + V_i \tag{2}$$

$$D_i = \left\{ \begin{array}{ll} 1 & \text{if } D_i^* \geq 0 \\ 0 & \text{if } D_i^* < 0 \end{array} \right\} \tag{3}$$

$$\begin{aligned} \Delta_i &= \mu(1) - \mu(0) + U_i(1) - U_i(0) \\ &= E\{\Delta_i\} + U_i(1) - U_i(0) \end{aligned} \tag{4}$$

$$E\{U_i(1)\} = E\{U_i(0)\} = E\{V_i\} = 0 \tag{5}$$

   and where correlation between U_i and V_i is possible. Calibrate your data generating process to mimic the "real setting" you prefer.

3. Consider two different assignments into treatment.
   (a) Keeping fixed the idiosyncratic treatment effect Δ_i, distinguish between the case of a random assignment into treatment and self-selection into treatment.
   (b) After generating the data and the covariates, construct a table that shows mean values of the outcome and the covariates by treatment status and assignment mechanism (i.e. random assignment and self-selection). Are the covariates balanced across treatment status? Use an appropriate statistical test to assess this formally. Report the respective p-values of the balancing tests.

[1] To learn how to generate random variables in Stata, type "help obs" and "help uniform".

(c) Assume now that you do not know the data generating process and estimate the model for both (!) assignment types with OLS. Compare the two regressions and discuss sign and size of the bias in the self-selection case.

4. Focusing on the self-selection case, assess whether the following methods overcome the selection bias. In each case, provide a short discussion of why the method works or why it does not work:

   (a) Instrumental-Variables: *Relevant commands in Stata include -ivreg2-*

   (b) Regression-Discontinuity-Design (*-reg-, -ivreg2-, and -rd-*)

   (c) One matching approach (e.g., exact covariate matching or nearest neighbor matching based on the propensity score, *-psmatch2- and -teffects-*) *or* the post-double-selection LASSO method[2] *(-pdslasso-*).

5. In order to estimate a DiD, generate five repeated cross-sections. Assume that there are five periods and individuals belong to one of two groups (e.g. a region), of which only one is treated. Treatment takes place in period $t = 4$. Assume a common trend in the outcome between the treatment and the control group.[3]

   (a) Produce a DiD graph (i.e. a graph with t on the x-axis and Y_i on the y-axis), showing the evolution of the outcome in treatment and control group.

   (b) Estimate the DiD using an OLS regression.

   (c) Generalize the DiD to an event study DiD and produce a graph illustrating the event study coefficients. Use period $t = 3$ as the reference category.[4]

   (d) Now assume that the common trend assumption is not satisfied. Illustrate the resulting bias with the event study DiD graph.

Please hand in the programs to replicate your code!
R and Stata users may find it helpful to look at the code examples in Cunningham (2021)
Assignment Deadline: March 14, 2021

[2] In order to illustrate how the method works, you may generate a variable that is correlated to the outcome but uncorrelated to treatment status D_i. Is this variable selected?

[3] The problems with inference that this setting poses can be disregarded.

[4] One way to produce such a graph in Stata is to use the user-written *-coefplot-* command, using the options *vertical* and *connect(direct)*. The challenge is to get the omitted (baseline) coefficient into the graph. This can be done by including a variable into the regression that is dropped from it (e.g. *gen basecoef = 0*). Then use the coefplot option *omitted* and include the coefficient of this variable as the baseline effect.