Causal Inference for Policy Evaluation – Spring Semester 2025

Lab Assignment 2

Due date: Friday, 18.04.2024

Instructions

Please send your solutions to mario.bernasconi@unibas.ch. The solutions should include two files. One file with your code saved in .R format. A second file with written answers, including tables and graphs, should be sent as a .pdf file and is limited to a maximum of 2 pages (12 points, single spacing). PhD student groups are allowed one more page as they have one more question. Any text or output beyond that will not be considered.

Difference-in-differences (20 points for MSc, 25 for PhD)

For the following questions use the data set "data_assignment.RData" (not the one used in the lab session). It already includes the variables connected (radius $\leq 556.6m$), submarines (date $\geq 2009Q3$) and treatment ($connected \times submarines$), as we constructed them in the lab session. It also includes a new variable: 'educ_high', a dummy for being highly educated.

Questions:

1. We want to estimate the effect of fast internet on employment (the dummy 'employed'), similarly to what we did in the tutorial:

   (a) Estimate the ATET non-parametrically and also the standard error of the estimate using 100 bootstrap replications. **(3 points)**

   (b) Estimate the ATET parametrically using a linear model (do not control for locations and time fixed effects). Do not cluster the std. errors.

      i. Is the point estimate from point (b) similar to the one in point (a)? Should we expect them to be similar? **(1 point)**

ii. Is the std. error from point (b) similar to the one in point (a)? Should we expect them to be similar? **(1 point)**

(c) Estimate again the ATET parametrically using a linear model (do not control for locations and time fixed effects), but now you should cluster the std. errors. Compare the std. errors to points (a) and (b). **(1 point)**

(d) Estimate again the ATET parametrically using a linear model, and also control for locations and time fixed effects. Cluster the standard errors at the location level. Is the point estimate for the treatment effect similar to the non-parametric estimate? Should we expect them to be similar? What could be driving the difference? **(6 points)**

2. Your classmates suggest that we should also control for the variable 'skilled' in the regression we used in point 1.d, because skilled people are more likely to be employed and also fast internet connection might facilitate skills acquisition.

   (a) Do you think that is a good idea to control for the dummy for being skilled? Explain. **(2 points)**

   (b) Run again the same regression of point 1.d but also control for 'skilled' now. Compare the result with the estimate from point 1.d. and comment. **(1 points)**

3. You are interested in the effect of fast internet on education (*educ_high*).

   (a) Replicate the common trends plot with the dummy for high education as outcome. Provide a possible explanation for the trends you observe. **(2 points)**

   (b) Conduct an event study for the outcome *educ_high* and report the results in a plot. Based on the results from this and the last question, is it reasonable to use a DiD strategy for this outcome variable? **(2 points)**

   (c) Given the common trend plot above, if you were to estimate the effect of fast internet using a DiD strategy, can we say whether our estimate would be up-ward or down-ward biased? **(1 points)**

4. **This question is for PhD students only.**

   (a) Let's go back to the non-parametric estimate of question 1.a. Adjust the function you used to compute the bootstrapped std. error in order to cluster at the location

level. Compute the clustered std. error using 100 bootstrap replications (this will take a while to run). Is the result similar to that in point 1.c? **(5 points)**

Instrumental Variable (5 points)

Read the paper for the next lab session: Angrist, J., and W. Evans (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size, *American Economic Review* 88(3): 450–477. Then answer the following questions:

Questions

5. (a) Why can't we just regress female labour supply on the number of children to estimate a causal effect? **(2 points)**

   (b) The IV strategy of the authors only identifies a very specific average treatment effect. What is that? Be precise and context-specific. **(1 point)**

   (c) Consider columns 1 and 2 in Table 7, which presents OLS and IV estimates, respectively. Provide an explanation for why the OLS estimate are lower than the IV ones (i.e. OLS over-estimates the negative effect). **(2 points)**