Problem Set 4
Due Saturday, December 12 at 11:59 AM (noon) via Canvas
ECON 31720, University of Chicago, Fall 2020

Assignments must be typeset in nicely formatted LaTeX. Programming assignments must use low-level commands, be commented clearly (not excessively), formatted nicely in 80 characters per column, etc. **You will be graded on the exposition of your written answers, the clarity of your code, and the interpretability and beauty of your tables and graphics.** The problem sets are individual assignments, but you may discuss them with your classmates. Submit the problem sets through Canvas in a single zip/tar/rar file. **Late problem sets will not be accepted under any circumstances.**

1. Suppose that we observe a repeated cross-section of individuals i over two time periods. Let $T_i \in \{1, 2\}$ denote the time period in which we observe individual i. We are interested in the effect of a binary treatment $D_i \in \{0, 1\}$ on an outcome Y_i. Let $Y_i(0)$ and $Y_i(1)$ denote potential outcomes, and assume that there are constant treatment effects, so that $Y_i(1) - Y_i(0) = \alpha$ is a non-stochastic scalar. Suppose that each individual i is a member of a binary group $G_i \in \{0, 1\}$. Maintain the common trends assumption that

$$\mathbb{E}[Y_i(0)|T_i = 2, G_i = g] - \mathbb{E}[Y_i(0)|T_i = 1, G_i = g]$$

does not depend on $g = 1, 2$. In contrast to the usual difference-in-differences design, suppose that some individuals in both groups are treated in both periods, but that the increase between time periods 1 and 2 of the proportion treated in group 1 is larger than the same increase in group 0.

Is α point–identified? If so, propose an estimator that could be used to estimate α. If not, provide a counter–example.

2. Consider the following data generating process for panel data with times $t = 1, \ldots, 5$ and individuals $i = 1, \ldots, n$:

$$\begin{aligned} E_i &\sim \text{Unif}\{2, \ldots, 5\} \\ Y_{it}(0) &= -.2 + .5E_i + U_{it} \\ Y_{it}(1) &= -.2 + .5E_i + \sin(t - \theta E_i) + U_{it} + V_{it}, \end{aligned}$$

where V_{it} is serially independent standard normal, independent of E_i, and U_{it} follows the autoregressive process

$$U_{it} = \rho U_{i(t-1)} + \epsilon_{it} \text{ for } t = 2, \ldots, 5 \text{ with } U_{i1} = \epsilon_{i1},$$

where $\epsilon_{it} \sim N(0, 1)$ for each $t = 1, \ldots, 5$, independently of both E_i and V_{it}.
Note: Angles are measured in radians for the sine function.

(a) Does common trends hold in this data generating process?

(b) Let $\theta = -2$ and $\rho = .5$.

Consider a regression of Y_{it} onto a full set of cohort fixed effects, a full set of time fixed effects, and relative time dummies $D_{it}^r \equiv \mathbb{1}[t - E_i = r]$ for all $r \in \{-4, \ldots, 3\}$ except $r = -1$ and $r = -4$. Run a Monte Carlo simulation with $n = 1000$ and $n = 10000$ to evaluate the finite sample distribution for the coefficients on the relative time dummies. Report the mean together with the 2.5% and 97.5% quantiles using a single figure that combines both sample sizes and has relative time on the horizontal axis.

(c) Repeat the previous part with $\theta = 0$ and $\theta = 1$. Compare your findings across the three different values of θ.

Explain any differences, and discuss implications for empirical practice.

(d) Devise a consistent estimator of

$$\text{ATE}_3(2) \equiv \mathbb{E}[Y_{i3}(1) - Y_{i3}(0) | E_i = 2].$$

Verify that your estimator works using a Monte Carlo with $n = 10000$ for every value $\theta \in \{-2, 0, 1\}$ and $\rho = .5$.

(e) For this part let $\theta = 1$.

Consider the same regression as in part b). Run a Monte Carlo for each combination of $n = 10, 20, 50, 200$ and $\rho = 0, .5, 1$ that evaluates the empirical size (rejection probability when the null hypothesis is true) of a level .05 t–test for the coefficient on D_{it}^1 implemented using each of the following approaches:

- The classical asymptotic variance estimator under homoskedasticity.
- An Eicker–Huber–White heteroskedasticity–robust asymptotic variance estimator. (Use the HC(1) version.)
- The cluster–robust asymptotic variance estimator, clustering over individuals i. Use the finite–sample correction discussed in the supplemental notes.
- The clustered Wild bootstrap, clustering over individuals i.

Report your results using a single, well–designed table and discuss the relative performance of the different approaches.

Note: You can keep E_i as deterministic across simulations, with an equal number of observations for each value of E_i. If you don't do that, then you might have simulation draws in which the regression coefficients do not exist due to perfect collinearity.

3. Replicate Figures 10 and 11 in "Combining Matching and Synthetic Control to Trade off Biases from Extrapolation and Interpolation" by Kellogg, Mogstad, Pouliot, and Torgovitsky (September, 2020). The data is originally from "The Economic Costs of Conflict: A Case Study of the Basque Country" by Abadie and Gardeazabal (2003), and is available on Canvas.