Problem Set 3
Due Friday, December 10 at 11:59 AM (noon) via Canvas
ECON 31720, University of Chicago, Fall 2021

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. The problem in `lmw.pdf`.

2. Suppose that we observe data (Y, R, D), where $D \in \{0, 1\}$ is a binary treatment and R is a continuously distributed random variable. Let $Y(0)$ and $Y(1)$ denote the potential outcomes for Y associated with treatment D. Suppose further that

$$D\left(\mathbb{1}[R \geq c] - \mathbb{1}[R < c]\right) \geq 0.$$

   (a) Show that there exists an unobservable random variable U such that

$$D = \mathbb{1}[U \leq p(R)],$$

   where $p(r) \equiv \mathbb{P}[D = 1|R = r]$ and $U|R = r$ is distributed uniformly over $[0, 1]$ and independently of $(Y(0), Y(1))$ for all r.

   (b) Suppose that $\mathbb{E}[Y(d)|R = r, U = u]$ is a continuously differentiable function of (r, u) for both $d = 0, 1$. Suppose that $p(r)$ is continuous and differentiable at $r = c$, but that its derivative is discontinuous at $r = c$. Show that $\mathbb{E}[Y_1 - Y_0|R = c, U = p(c)]$ is point identified.

3. 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 $n = 20, 50, 200$ 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.

4. This question uses the data from "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program" by Abadie, Diamond and Hainmueller (2010). The paper and data are on Canvas. You do not need to produce standard errors for this question.

(a) For each of the 38 control states, estimate a separate difference-in-differences design using only a single state as the control while including a full set of unit and time effects. Organize these estimates into a visually appealing plot or table.

(b) Using all 38 control states together, estimate a difference-in-differences design with a full set of unit and time effects. Compare this estimate to those in the previous part.

(c) Implement the synthetic control estimator using the entire set of pre-period outcomes to determine the weights. That is, in the notation of the authors, take $\boldsymbol{X}_1 \equiv [Y_{11}, \ldots, Y_{1T_0}]'$ to be the vector that includes every pre-period outcome. Take the weighting matrix ($\boldsymbol{V}$ in the authors' notation) to be the identity matrix. How does your estimate of the treatment effect in the post-period compare to those from the previous parts?