Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due November 5, 2024
55 total points

Problem Set 7

Instructions: Answer the following questions and submit your results via email to `sean.corcoran@vanderbilt.edu`. Use your name and problem set number as the filename. Working together is encouraged, but all submitted work should be that of the individual student.

This problem will replicate analyses reported in Bifulco, Rubenstein, and Sohn (2017).[1] That study used a synthetic control design to estimate the impact of Say Yes to Education (a promise scholarship program in Syracuse, New York, which provided free college tuition to any student who graduated from a public high school in Syracuse) on total district enrollment and graduation rates. The program was implemented in 2008.

There are two separate datasets on Github containing panels of enrollment and graduation data for school districts in New York State:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/nys_data_enroll.dta, clear
use https://github.com/spcorcor18/LPO-8852/raw/main/data/nys_data_grad.dta, clear
```

Most of the variables in these datasets should be self-explanatory from their variable names and labels (although I'm not 100% sure how *target_donor* and *small_index* are defined, as they don't appear to align with the paper's selection of potential donor districts).

The authors used two potential donor pools. The comprehensive donor pool included all 275 (non-Syracuse) districts, while the restricted donor pool included 22 districts categorized as "City–Large," "City–Midsize," or "City–Small." (Note these counts are a little smaller for the graduation rate panel, which also has fewer years). NYC is excluded from the dataset.

Using the `synth2` synthetic control package in Stata, replicate the findings in this paper by reporting the elements listed below. Note you do not need to run all 6 alternative specifications of the pre-treatment years as they do in the paper (Table 1). Rather, just use their Specification (2), which uses outcomes from the first, middle, and last year of the pre-treatment period. Also include the pre-treatment average percent of district students eligible for free or reduced price lunch, percent Black, and percent Hispanic in this procedure. Run these twice, first using the comprehensive donor pool, and then again using the restricted donor pool (where *target_donor==1*).

Taken together, you will have four sets of results: two outcomes (enrollment and graduation rates) × two potential donor pools. Brownie points to those who combine results in a

[1] Thank you to Bob Bifulco and Hosung Song for providing the data used in their paper.

pleasing-to-read format.

Include these things in your results, and be sure to submit your do-file:

(a) The weights assigned to donor districts, as in Tables 1 and 5. Write a few sentences summarizing the resulting weighting used. Do they correspond to the weights reported in the paper? **(10 points)**

(b) The main synthetic control graph showing trends in Syracuse and its synthetic control, as in Figures 2 and 3. Briefly summarize what you see. **(10 points)**

(c) The treatment effect ("gap") version of the graphs in (b) showing the *difference* in mean outcomes between Syracuse and its synthetic control by year (these were not shown in the paper). **(5 points)**

(d) Point estimates of the treatment effect by year (2008, 2009, 2010, and 2011), as in Tables 3 and 6. Note the graduation rate data only include 3 post-treatment years. **(5 points)**

(e) The graph showing the gap in mean outcomes between Syracuse and its synthetic control overlaid on the placebo gaps. Briefly summarize what you see. **(10 points)**

(f) p-values from the placebo-based inference. Explain in words where these come from, and how they should be interpreted. (Note, you only need to provide a written explanation for one set of results, not every one). **(10 points)**

(g) The "leave-one-out" (loo) robustness test. Interpret the results. **(5 points)**

Notes: see the in-class exercise do-file for help, and it would (of course) help to refer to the original Bifulco et al paper. Be attentive to which district ID represents the Syracuse school district—it is not consistent across the two datasets.