PP346 Program Evaluation
Problem set 4. Due Date: Thursday, Nov 29, 2018, by 9am, via Canvas.

For this assignment, you will use the paper "Soda Taxes and the Prices of Sodas and Other Drinks: Evidence from Mexico." Provide a write-up where you answer the questions below, selectively cutting and pasting output where needed. Be concise in your write-up; excess wordiness will be penalized. Also, submit a log file that includes commands and results for your entire analysis. Remember that your grade mostly comes from the written report—your code or log is a "replication" file that shows you did the work. Imagine that the grader won't look at your code until they're done reading the write up, and that's how you should type it up. If you put it all together in a R notebook, knit it to .pdf or .html before submission.

The assignment will implement the synthetic control method on the dataset soda_tax_long_data.dta, which you can find on Canvas. If you don't speak Spanish, you will also want to make use of prdname translations.xlsx.

Due to the advanced nature of this method, implementations are not standardized across platforms. Thus, we **strongly recommend** using Stata for this assignment, along with the package *synth runner*. Follow the instructions for installation on the author's Github (https://github.com/bquistorff/synth_runner). If you decide to do this using R, expect little to no support on using any of the implementations available on CRAN. You can also expect to write lots of loops to carry out the permutation inference.

1. How many products are in the data? Which value of product corresponds to regular sodas (prdname=regsodasndx)? How many time periods are there? Which value of date corresponds to January 2014?

2. Plot the real price (rprice) of regular sodas. Does it look like the corresponding series in Figure 2 of Grogger (2017)?

3. Using synthetic control, match on all the pre-treatment periods of the dependent variable except for the last (December 2013). Obtain period-specific estimated treatment effects for each of the 18 post-treatment time periods in the data. List them and their p-values.

4. Consider the estimated treatment effect for regular sodas, and its p-value, for the estimated treatment effect for April 2014. Given the p-value and the number of placebo treatment effects that were used to construct it, how many of those placebo treatment effects must have been larger than the treatment effect estimated for regular sodas?

5. Abadie et al (2010) recommend including the treatment product in the comparison pool for each of the placebo estimations used to construct p-values and confidence intervals. Some software, such as synth_runner, excludes it nonetheless. Under what conditions would you expect the difference to have little effect on your p-values?

6. Use the period-specific treatment effects to estimate the pooled post-tax effect reported in the first row of the first column of Table 3 (referred to as $\widehat{\Delta}$ in the text). How close is your estimate to the one reported in the Table?

7. To obtain $\widehat{\Delta}$ above, you matched on all pre-treatment values of the dependent variable, except for the very last (December 2013). Recent work suggests that this approach may lead to overfitting and suggests matching on fewer lags of the dependent variable. Compute $\widehat{\Delta}$ under the two following matching strategies and see if the results are sensitive to these choices (keep excluding December 2013 from your donor pool):

a. Match on every other lag of the dependent variable.

b. Match on every third lag of the dependent variable.

c. Now that you have the synthetic control estimate for three different lag setups (question 3/6, 7a, and 7b), is $\widehat{\Delta}$ sensitive to different matching strategies?

8. Some of the comparison products in the data are clothing items, which some might argue are rather different than foodstuffs. Re-estimate the model from 6 excluding clothing items from the comparison pool. How different are your estimates of $\widehat{\Delta}$?