Pol Sci 630: Problem Set 11 - Causal Inference Techniques - Differences-in-differences

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Due Date: Wednesday, November 16th, 2016, 1.25 PM (Beginning of Class)

Note 1: It is absolutely essential that you show all your work, including intermediary steps, in your (mathematical) calculations and that you comment on your R code to earn full credit (you can comment on your R code both with the use of # in the R code and in the LaTeX code). Showing all steps and commenting on code will also be required in future problem sets.

Note 2: Please submit a PDF file created through knitr containing all your answers to the problem set. knitr allows you to combine R code and LaTeX code in one document, meaning that you can include both the answers to R programming and math problems. Also submit the source code that generates the PDF file (i.e. the .Rnw file).

Note 3: Make sure that the PDF files you submit do not include any references to your identity. The grading will happen anonymously. You can submit your answer at the following website: `http://ps630-f15.herokuapp.com/`

R Programming

Problem 1: Descriptive Summary Statistics (4 points)

Do the following in R:

a) Load the *VOTE1* dataset that was used in previous problem sets.

b) Create a table of summary statistics for LaTeX that includes all interval-level variables in the dataset. Include this table in the document you submit.

Note: If you have the dataset in the same folder as your .Rnw file, you do not need to set a working directory when you compile your PDF.

Problem 2: Differences-in-differences (8 points)

Do the following in R:

The government of country Z is interested in the effect that tax reductions have on the private expenditures of the owners of small businesses. In the year 2016, a tax reduction for restaurant owners was introduced in country Z. Restaurant owners are a subset of the small business owner population.

The government would like to estimate the effect that the introduction of the tax reduction had on this particular group and, knowing your awesome statistical skills, hires you for this task. The government provides you with two datasets, one with information from the year 2015 and one with information from the year 2016. Both datasets include information on the private expenditures of small business owners. You can find those datasets on the course website under ”Meetings”.

The two datasets have the following variables:

1. Year: Year in which the private expenditures were made

2. PrivExp: The level of private expenditures in the local currency (Z-Dollars)

3. RestOwn: Equals “1” if the person is a restaurant owner, equals “0” if the person is not a restaurant owner

Assuming that there are parallel trends in the private expenditures of all small business owners, use the two datasets above to estimate the effect that the tax reductions had on the private expenditures of restaurant owners through a linear regression model. Make sure to interpret the results in your own words, including the statistical significance of the estimate. Based on your results, make a recommendation to the government regarding whether or not they should introduce tax reductions to achieve higher private expenditures of small business owners.

Note: Before you conduct this task, think carefully about how the data has to be structured to conduct a differences-in-differences analysis. When you know the structure that the data needs to have, combine the above datasets and introduce appropriate variables to achieve the structure needed.

Note: If you have the dataset in the same folder as your .Rnw file, you do not need to set a working directory when you compile your PDF.

Statistical Theory: Differences-in-differences

Problem 3 (4 points)

Answer the following questions:

a) In the tutorial it was shown that a simple differences-in-differences estimate using four groups is the following:

$$\delta = (\bar{x}_{t2} - \bar{x}_{t1}) - (\bar{y}_{t2} - \bar{y}_{t1})$$

Where x is the treatment group and y is the control group and $t1$ and $t2$ indicate time points 1 and 2 respectively, with the treatment occurring in-between.

Assuming that the parallel trends assumption is valid, how does the above calculation allow us to make a causal inference about the effect of the treatment on x? Explain carefully.

b) What happens if the parallel trends assumption is violated? Explain carefully.

c) Describe one way how the parallel trends assumption can be relaxed. What needs to be true to relax it?