New York University Shanghai

Problem Set 8

BUSF-SHU 210: Business Analytics (Spring 2019)

1. Life-Saving Difference-in-Differences

Consider the London cholera epidemic in mid 1800s. In London, the water was provided by two companies: (i) the Lambeth and (ii) the Southwark and Vauxhall. Both of them obtained water from the Thames river. In early 1852, the Lambeth moved their water intake to an upriver of the Thames. In 1851, the cholera death rate of the households with water supply from the Lambeth was 61 per 10,000 homes and that of the households with water supply from the Southwark and Vauxhall was 65 per 10,000 homes. In 1852, the cholera death rate of the households with water supply from the Lambeth was 5 per 10,000 homes and that of the households with water supply from the Southwark and Vauxhall was 71 per 10,000 homes.

(a) Using the Diff-in-Diff identification strategy, what is the causal effect of moving the water intake to the upper river on the cholera death rate?

(b)) What assumption do we need for the causal effect estimated in part (a) to be valid?

2. Diff-in-Diff in Action

Google is examining how a new Ad display function could impact the click-through rate (CTR) for the advertisements of an advertiser. To do so, the company adopts the difference-in-differences method. More specifically, Google randomly selects a sample of 10,000 advertisers who start to adopt this function at the beginning of Week 0 (the treatment group), and randomly selects another sample of 10,000 advertisers who do not adopt the new function at all (the control group). It has been checked that the other relevant pre-treatment features of the treatment and control groups are balanced. Google records the CTR for the advertisements in Week -2 (2 weeks before treatment), Week -1 (1 week before treatment), Week 0 (treatment week), and Weel 1 (1 week after treatment). You job is to use Diff-in-Diff analysis to estimate the causal effect of the new function on CTR.

The data associated with the experiment described above is stored in `NewFunction.csv`. Each row in this data set represents an advertiser in a specific week. The data has 3 variables:

- $AdvertiserID$: The ID of the advertiser on Google.
- $Week$: The week associated with this row of record.
- $Adoption$: An indicator of whether the advertiser has adopted the new feature. $Adoption = 1$ if the advertiser adopted the new feature; $Adoption = 0$ if the advertiser did not adopt the new feature.
- CTR: Variable of interest, the CTR of the advertiser in the associated week.

Please briefly answer the following questions:

(a) Can you propose a method to check the parallel assumption for this study? Based on the data, is the parallel assumption satisfied?

(b) Google uses the following linear regression model to estimate the causal effect of adopting the new Ad display function:

$$CTR \approx \hat{\alpha} + \hat{\beta}_{-1} t_{-1} + \hat{\beta}_0 t_0 + \hat{\beta}_1 t_1 + \hat{\gamma} Adoption + \hat{\tau} Adoption \cdot t,$$

where

- t_{-1}: the indicator variable for Week -1. $t_{-1} = 1$ if the record is for Week -1; otherwise $t_{-1} = 0$;
- t_0: the indicator variable for Week 0. $t_0 = 1$ if the record is for Week 0; otherwise $t_0 = 0$;
- t_1: the indicator variable for Week 1. $t_1 = 1$ if the record is for Week 1; otherwise $t_1 = 0$;
- t: the indicator variable for the treatment being treated in this week. $t = 1$ if $t_0 = 1$ or $t_1 = 1$; otherwise $t = 0$.

Which estimated parameter measures the impact of adopting this new function on the CTR? Use the data to estimate this causal effect.

(c) Draw a figure to illustrate the causal effect of adopting the new function and how we use DiD to estimate it.