Differences-in-Differences Analysis

Chuyuan Zhong

2023-04-24

You have been hired as a data analyst for the Department of Education. For your first project, you have been asked to analyze a program called Aides for America (A4A). This program was funded by the Department of Education to provide teacher aides to elementary school teachers in grades 3, 4 and 5 with the goal of improving students' math test scores. Grades in a given school have multiple teachers, and the program operated at the teacher level so that within a school-grade, some teachers received additional aides and others did not. Furthermore, the program was rolled out over time so that different teachers received additional aides at different times, and some teachers did not receive any additional aides at any times.

You have been given the following teacher-year panel data:

- **mathscore** = average math test score for students in the teacher's class
- **numaides** = number of teacher aides assigned to the teacher
- **treatment** = indicator for whether teacher ever receives additional teacher aides from A4A
- **schoolid** = ID number for school.
- **grade** = categorical variable for 3rd , 4th , or 5th grade
- **yr1_treatment** = calendar year that additional A4A are assigned
- **year** = calendar year of the start of the academic year
- **teacherid** = ID number for teache r

Download dd_problem_set.csv and use this data for the following exercises:

```
df <- read.csv('dd_problem_set.csv')
head(df)
```

```
##   teacherid year mathscore numaides treatment schoolid grade yr1_treatment
## 1     30001 2001  1057.000        2         1        1     3          2008
## 2     30002 2001  1230.593        2         1        2     3          2004
## 3     30003 2001  1062.552        2         0        3     3             0
## 4     30004 2001  1090.958        2         0        4     3             0
## 5     30005 2001  1202.926        2         0        5     3             0
## 6     30006 2001  1099.034        2         0        6     3             0
```

```
cat("The dataset has", nrow(df), "observations and", ncol(df), "variables.")
```

```
## The dataset has 15508 observations and 8 variables.
```

Question 1 Confirm that you have varying dates of treatment. Explain the intuition of applying the Diff-in-Diff research design in this context to estimate impacts of the program on math test scores.

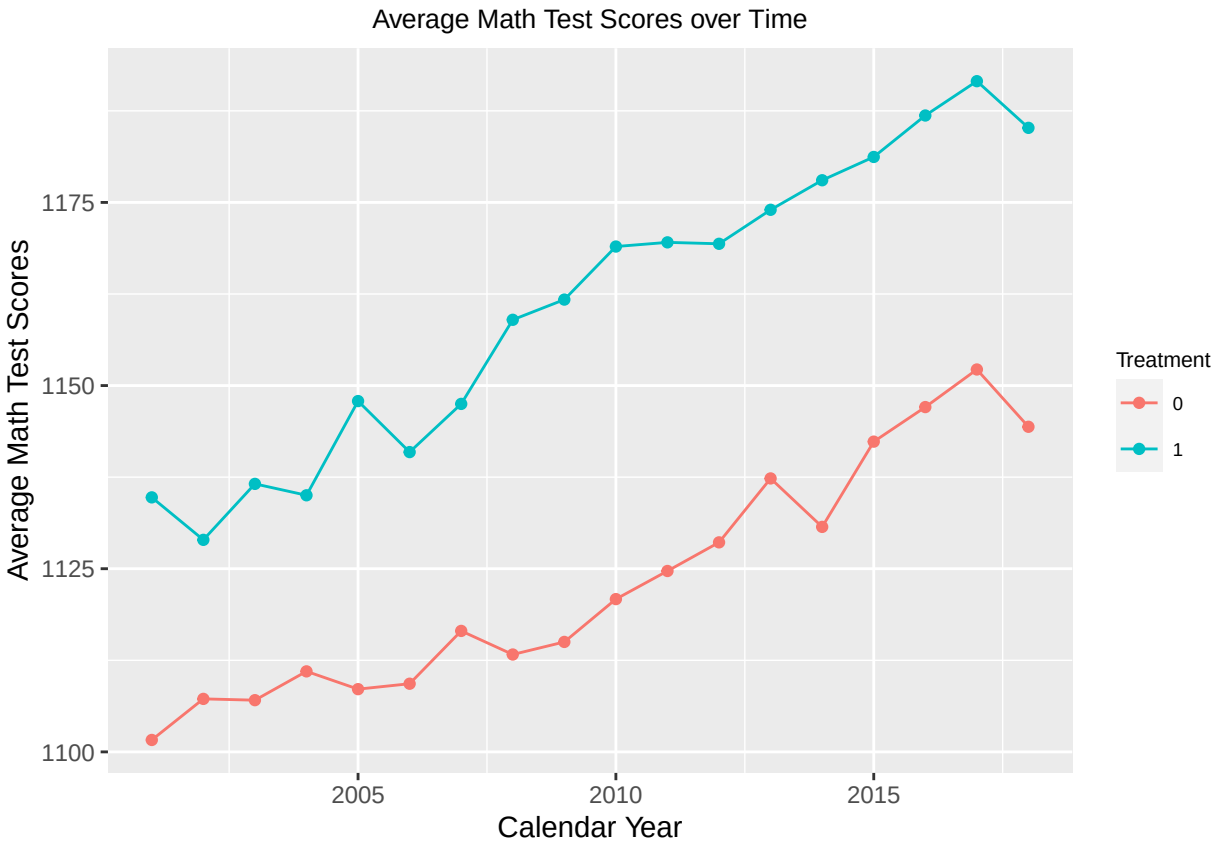
```
table(df$treatment, df$yr1_treatment)
```

```
##
##         0 2004 2005 2006 2007 2008 2009 2010 2011 2012
##   0 7872    0    0    0    0    0    0    0    0    0
##   1    0  847  862 1017  871  941  754  691  751  892
```

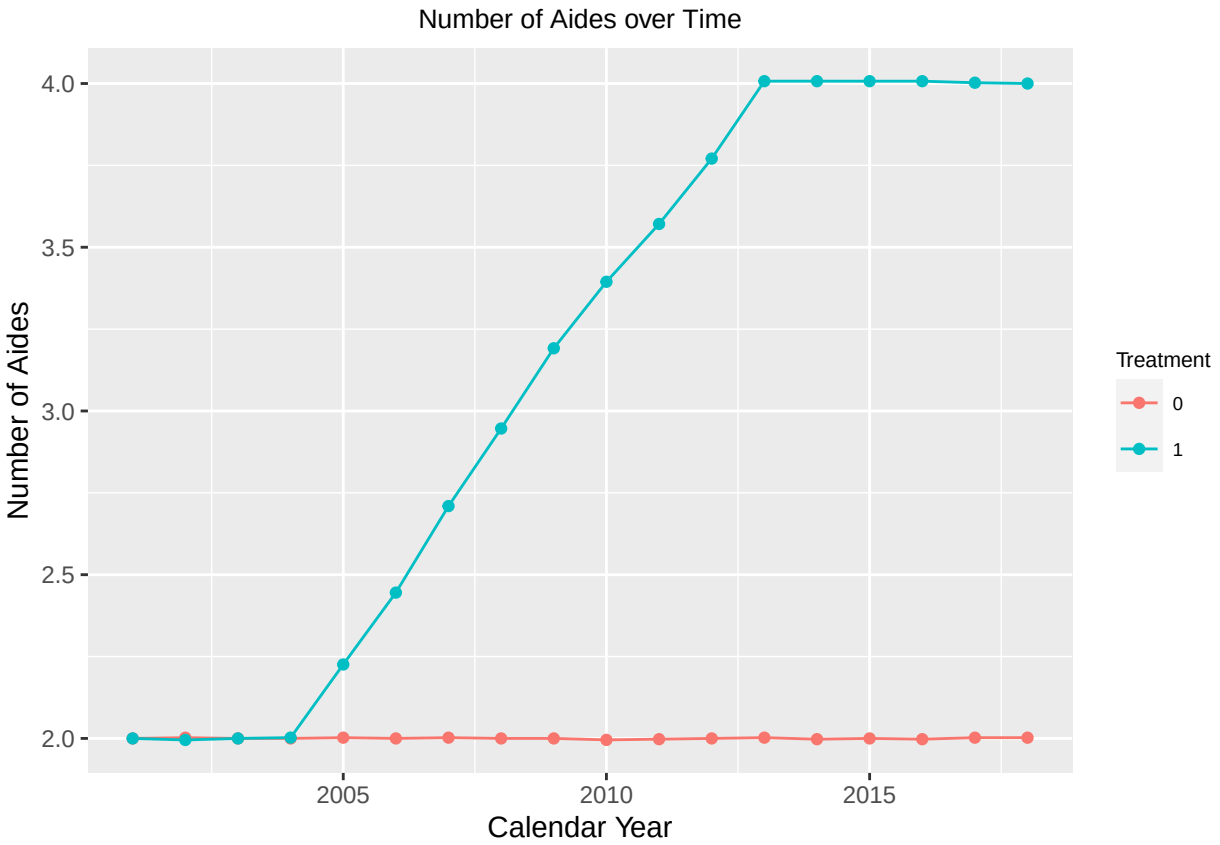
By using the `table` function we can check that there are varying dates of treatment from 2004-2012. The intuition of applying the Diff-in-Diff research design is that we can compare the math scores of teachers who received additional aids from A4A and teachers who didn't received the aids, both before and after the intervention. Since the treatment have varying dates, we assume that the timing of the treatment is random, and we assume that any differences between the treatment and control group are stable over time prior to treatment.

Question 2 Plot average math test scores and the number of aides over calendar year for the treatment and control groups. What does this plot illustrate? For example, are the treatment and control groups similar? Explain why it's ok if there are differences between the treatment and control groups or why it's not ok. What sorts of differences between the treatment and control groups would be problematic for the Diff-in-Diff research design, and how do those factors relate to this plot?

```
dft1 <- df %>%
  group_by(year, treatment) %>%
  summarize(mean_score = mean(mathscore, na.rm = TRUE),
            mean_aid = mean(numaides, na.rm = TRUE))
dft1 <- na.omit(dft1)
ggplot(dft1, aes(x = year, y = mean_score, color = factor(treatment))) +
  geom_line() +
  geom_point() +
  labs(title = "Average Math Test Scores over Time",
       x = "Calendar Year",
       y = "Average Math Test Scores",
       color = "Treatment") +
  theme(plot.title = element_text(size = 10, hjust=0.5),
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 7))
```




```
ggplot(dft1, aes(x = year, y = mean_aid, color = factor(treatment))) +
  geom_line() +
  geom_point() +
  labs(title = "Number of Aides over Time",
       x = "Calendar Year",
       y = "Number of Aides",
       color = "Treatment") +
  theme(plot.title = element_text(size = 10, hjust=0.5),
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 7))
```

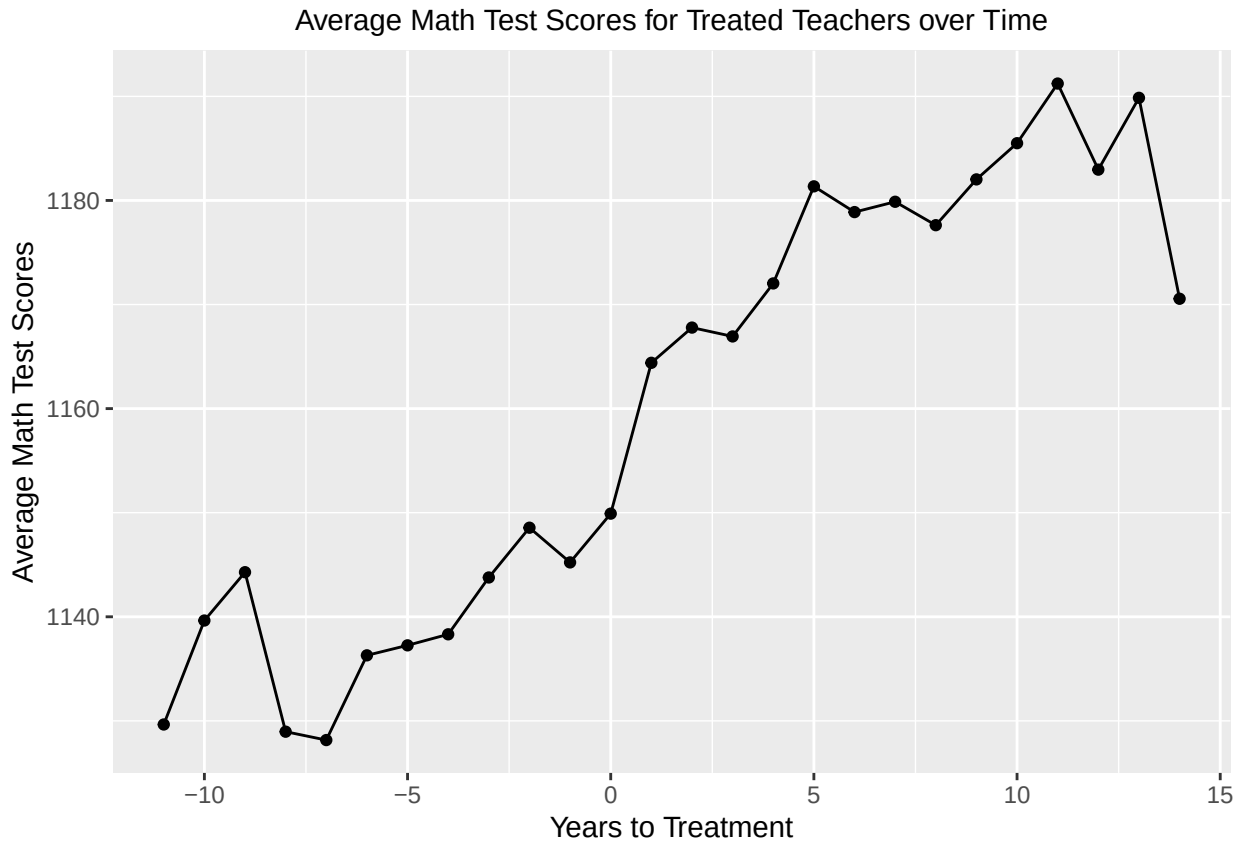


This plot illustrates the average math test scores and number of aides across years for treated and non-treated teachers. In the first panel, it is clear that the treatment and control groups do not have a similar outcome before and after the treatment. The groups tend to be parallel and stable. In the second panel, it can be seen that at the beginning, the treatment and control groups are similar. However, with the treatment intervention, there is a gap between the number of aides. It is acceptable to have a difference between the two groups as it aligns with the intuition of DD design, where initially, they show similar results, but differences emerge due to the treatment. If the treatment and control groups have a growing gap before the treatment is introduced, this would be problematic for the DD design since their trends should remain constant.

Question 3 Using observations for treated teachers only, create time since treatment and plot average math test score by time since treatment. Interpret this graph and explain how it relates to a Diff-in-Diff research design.

```
df$time_since_treatment <- ifelse(df$treatment==1, df$year-df$yr1_treatment, df$treatment)
plotdata <- aggregate(df$mathscore, list(df$time_since_treatment, df$treatment), FUN = mean)

ggplot(plotdata[plotdata$Group.2==1,], aes(x = Group.1, y = x)) +
  geom_point() +
  geom_line() +
  labs(title = "Average Math Test Scores for Treated Teachers over Time",
       x = "Years to Treatment",
       y = "Average Math Test Scores")+
  theme(plot.title = element_text(size = 11, hjust = 0.5))
```



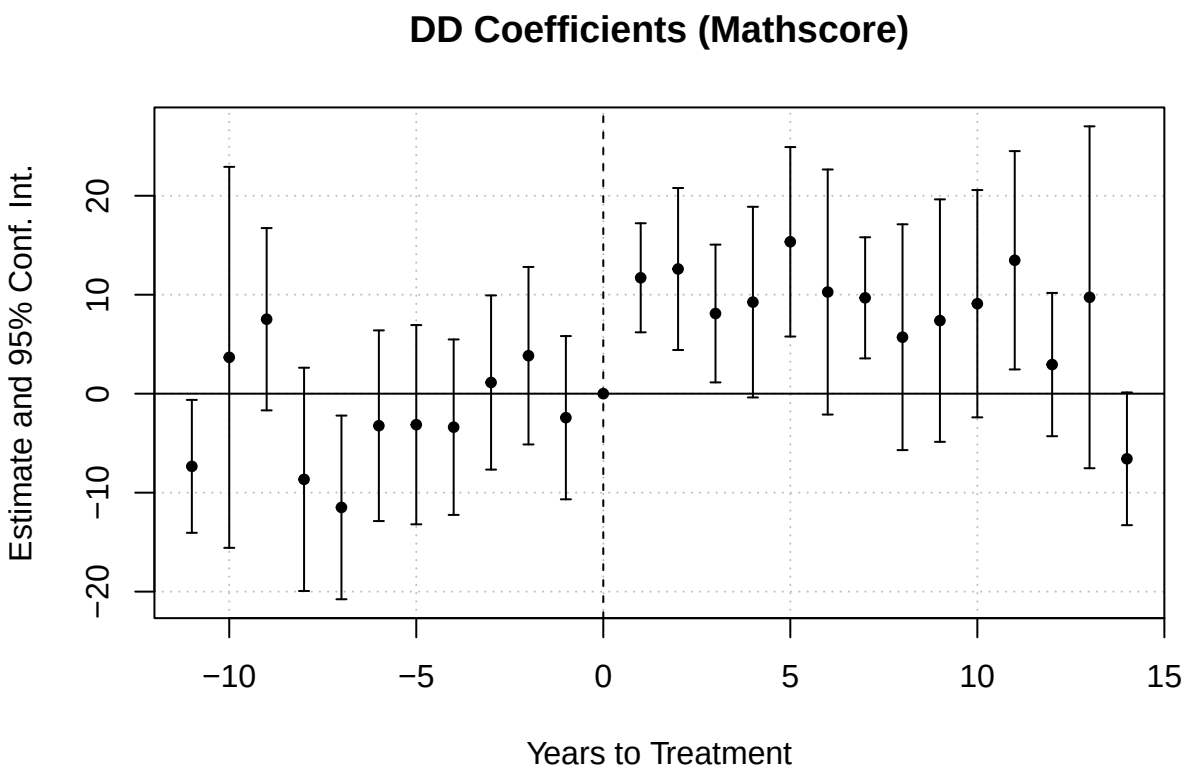
The graph illustrates a significant increase in the average math score following the treatment, this shows that the change of math scores is largely due to the treatment effect, which is related to the intuition of DD design.

Question 4 You have been asked to use the full data (treatment and control teachers), define time since treatment and estimate the following regression specification:

$$y_{it} = \alpha_0 + \alpha_1 N_i + \sum_{k=0} [\delta_k D_{ik}] + \gamma_s + \gamma_g + \gamma_t + error_{it}$$

where y_{it} denote math test scores for teacher i in year t, N_i denote an indicator for being a treated teacher, D_{ik} denotes an indicator variable for being a treated teacher and having time since treatment = k, γ_s denotes school fixed effects, γ_g denotes school fixed effects, and γ_t denotes calendar year dummies. Plot the δ_k coefficients over time since treatment. Interpret the results illustrated in the plot.

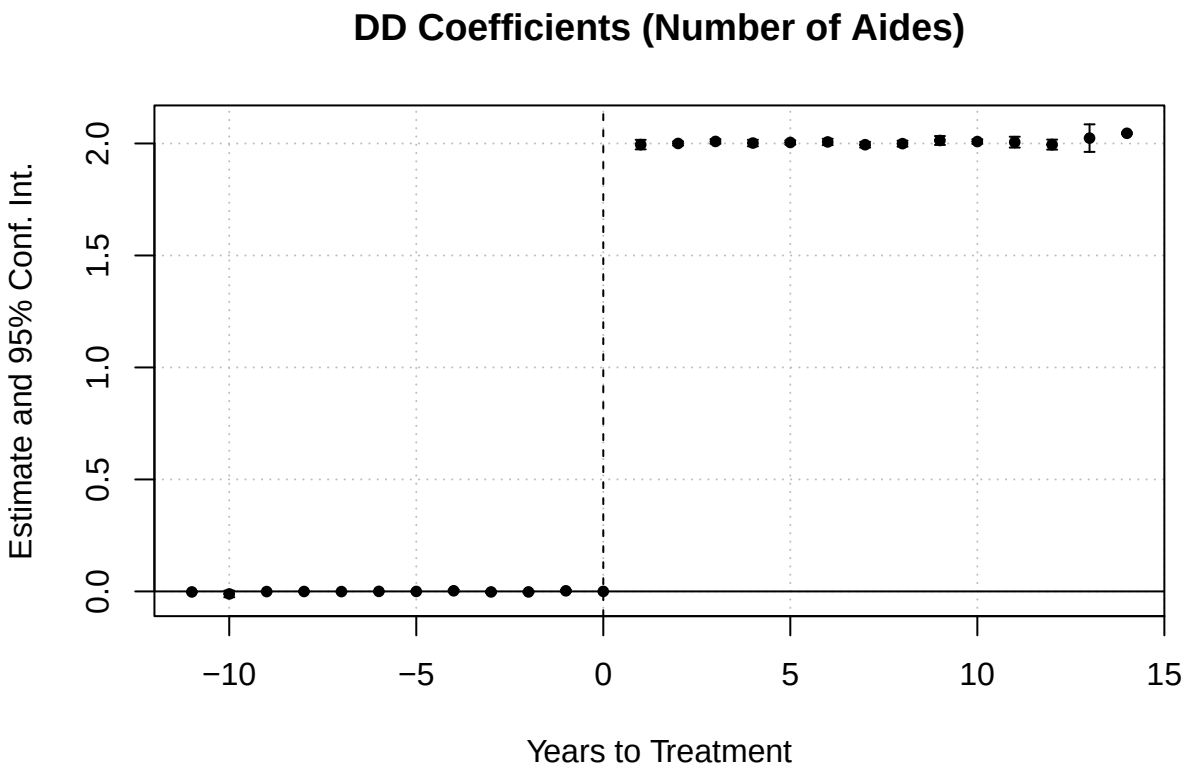
```
ddreg = feols(mathscore ~ i(time_since_treatment, treatment, ref = 0) +
              treatment | year + schoolid + grade, data = df)
iplot(ddreg, xlab = 'Years to Treatment', main = 'DD Coefficients (Mathscore)')
```



There is a positive effect since the treatment was introduced, but the effect gradually fades away as the years pass. Additionally, the coefficients are statistically insignificant before the treatment, indicating that the change in math score was stable and consistent.

Question 5 The A4A program was designed to provide additional teacher aides to teachers. Estimate the above regression specification in (4) using numaides as the outcome variable. Plot the δ_k coefficients over time since treatment. How many additional aides did treated teachers receive after the program took effect?

```
ddreg1 = feols(numaides ~ i(time_since_treatment, treatment, ref = 0) +
               treatment | year+schoolid+grade, data = df)
iplot(ddreg1, xlab = 'Years to Treatment', main = 'DD Coefficients (Number of Aides)')
```



The plot illustrates that there were no aides before the introduction of A4A. However, after the program was introduced, the number of aides started to increase. Treated teachers received two additional aides after the program took effect.

Question 6 Ultimately we want to know how much an additional teacher aid increased average math test scores. To do this, your boss suggests estimating the following regression:

$$y_{it} = \alpha_0 + \alpha_1 T_i + \delta numaides_{ik} + \gamma_s + \gamma_g + \gamma_t + error_{it}$$

Explain why there may be endogeneity concerns in this regression.

This regression examines the relationship between the number of aides and average math scores. In other words, it only estimates whether an increase in the number of aides affects the average math scores. The correlation found in this regression cannot be interpreted as a causal relationship for how much an additional teacher aid increases the average math scores. Factors such as school resources and teacher quality may also affect the number of aides.

Question 7 Use the results from (4) and (5) to answer how much an additional teacher aid increase average math test scores.

```
df$after <- ifelse(df$time_since_treatment > 0, 1, 0)
df$TAfter<- df$treatment*df$after
ta_reg <- feols(mathscore ~ TAfter + treatment | year+schoolid+grade, data = df)
ta_reg1 <- feols(numaides ~ TAfter + treatment| year+schoolid+grade, data = df)
coefficients(ta_reg)[1]/(coefficients(ta_reg1)[1])
```

```
##   TAfter
## 5.705354
```

An additional teacher aid increase associated with 5.7 points of average math scores.