Data Analytics in Business

Zhaohu (Jonathan) Fan

Information Technology Management

Scheller College of Business

Georgia Institute of Technology

September 19, 2024

Week 5: Upcoming deadlines and updates

- **Week 5 (Module 5)** is now available on Canvas.
- **(Graded) Self-Assessment 4** has been released and is due by this Sunday, September 22, at 11:59 PM EST.
- **(Graded) Homework #1:** has been released and is due by this Sunday, September 22, at 11:59 PM EST. This assignment includes:
    - **Homework #1, Part 1 (Theoretical): One attempt allowed.**
    - **Homework #1, Part 2 (Computation): One attempt allowed.**
    - You can work on both parts as much as you want within the due period, but remember to click "submit" only when you're completely ready.
- **Piazza Forum:** Always open for questions! It's the perfect place to interact with our teaching team and your classmates.
    - Simply click on "Piazza" in the left panel of our Canvas course page.

Vote for Your favorite TA of September

Scan me

- Or click on the link provided below.
  - Survey link

Main topics

- **Analytics & Modeling (weeks 1-5)**
    - Week 5 (Module 5): Treatment Effect, Randomized Controlled Experiments, and Natural Experiments
        - Difference-in-differences (DiD)
        - How DiD is integrated into regression models
        - Manual Calculation of DiD

What is Difference-in-differences (DiD) approach?

- Difference in Difference method (DiD) compares not the outcomes Y but **the *change* in the outcomes pre- and post-treatment**. This is a quasi-experiment approach.
  - **DiD is a statistical method used to evaluate the effect of a policy or treatment.**
  - It compares changes in outcomes over time between a treatment group (exposed to the intervention) and a control group (not exposed).
    - Formula: $\left(\text{After}_{\text{Treatment}} - \text{Before}_{\text{Treatment}}\right) - \left(\text{After}_{\text{Control}} - \text{Before}_{\text{Control}}\right)$
    - **Estimating the DID estimator (method 1: generate the interaction). We will call this interaction ‘did’.**

The Difference-in-Differences (DiD) approach can be presented either in a mathematical formulation or in a table.

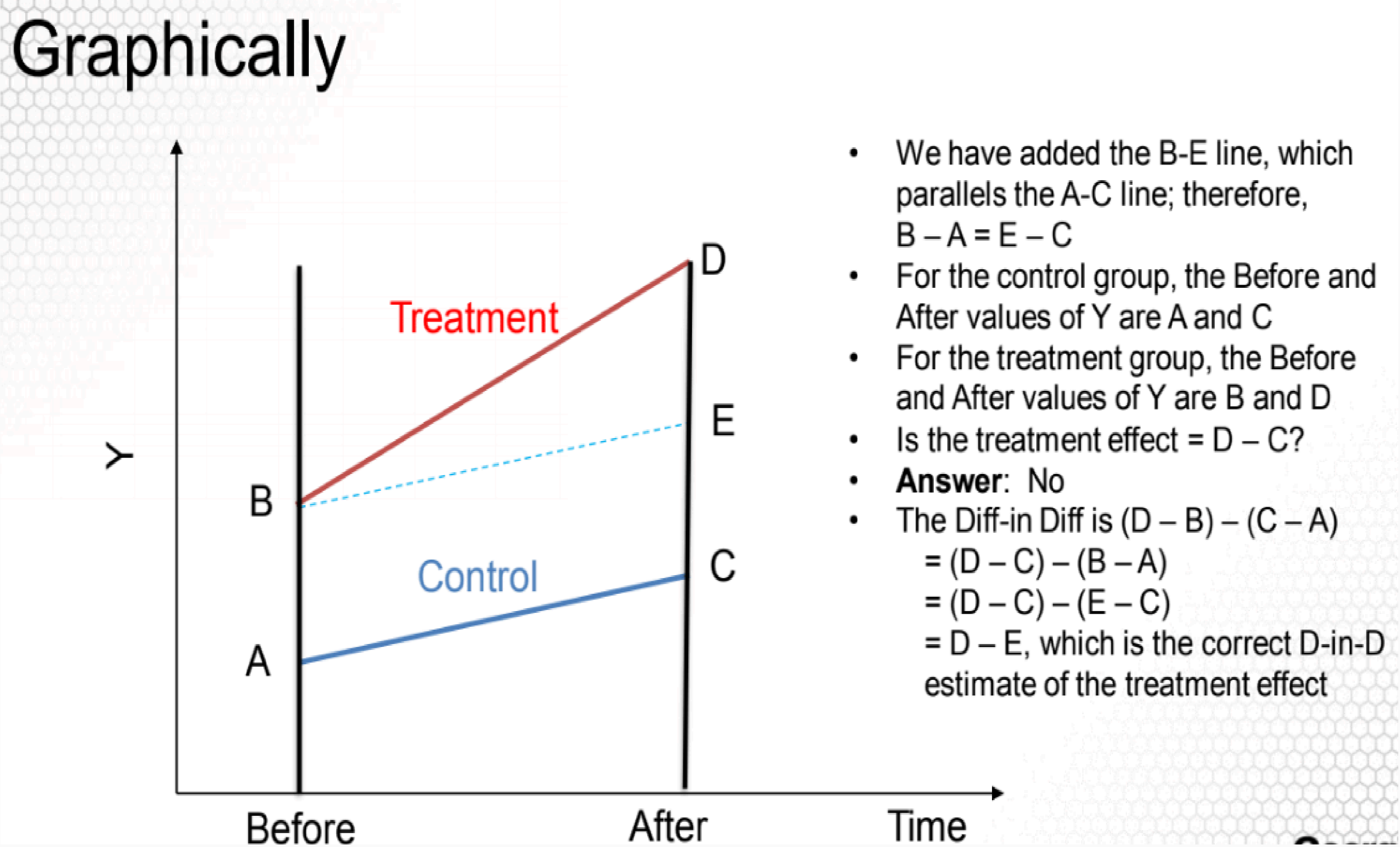
Mathematical formulation

The DiD estimator can be mathematically expressed as:

$$\text{DiD} = (Y_{\text{post,treatment}} - Y_{\text{pre,treatment}}) - (Y_{\text{post,control}} - Y_{\text{pre,control}})$$

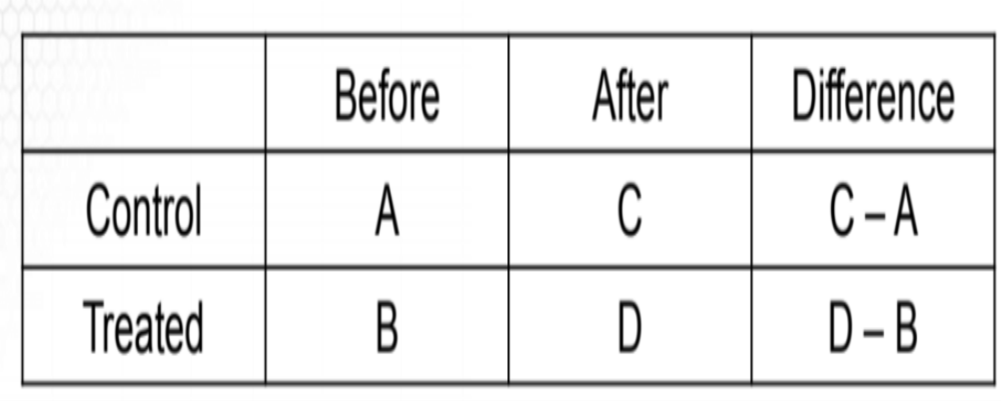
Where:

- $Y_{\text{post,treatment}}$ is the average outcome of the treatment group after the intervention.
- $Y_{\text{pre,treatment}}$ is the average outcome of the treatment group before the intervention.
- $Y_{\text{post,control}}$ is the average outcome of the control group after the intervention.
- $Y_{\text{pre,control}}$ is the average outcome of the control group before the intervention.



Graphically

- We have added the B-E line, which parallels the A-C line; therefore, B – A = E – C
- For the control group, the Before and After values of Y are A and C
- For the treatment group, the Before and After values of Y are B and D
- Is the treatment effect = D – C?
- **Answer**: No
- The Diff-in Diff is (D – B) – (C – A)
  = (D – C) – (B – A)
  = (D – C) – (E – C)
  = D – E, which is the correct D-in-D estimate of the treatment effect



	Before	After	Difference
Control	A	C	C – A
Treated	B	D	D – B

- For the control group, the difference of the average Y values at time t_2 (After) and time t_1 (Before) $= \mathrm{C} - \mathrm{A}$
- For the treatment group, the difference of the average Y values at time t_2 (After) and time t_1 (Before) $= \mathrm{D} - \mathrm{B}$
- The difference between these values is called difference-in-differences (DiD)
- Diff-in-Diff $= (D - B) - (C - A)$

Variations of the table expression

- **Variations of the table expression, similar to how we used variations of the confusion matrix in the quiz questions:**
    - Perform a diagonal reflection or rotation (transposing or rotating the matrix).
    - **Swap the position of column 1 with column 2, and/or row 1 with row 2.**

Exploring DiD's core assumption

The Difference-in-Differences (DiD) approach is a statistical technique used in econometrics and social sciences.

- **This is achieved by comparing the before-and-after changes in outcomes between a treatment group and a control group.**
- **The core assumption of DiD is that both groups would have followed parallel paths in the absence of the intervention.**

Can Difference-in-Differences (DiD) estimator be negative?

- A **negative DiD estimator** indicates that the treatment effect is less than what would have been expected based on the control group's trend.
    - **In other words, the treated group's outcome has increased less (or decreased more) than what was observed in the control group after the treatment.**
    - It suggests that the treatment had a negative effect relative to what would have occurred if the treatment had not been applied, assuming all the assumptions for a valid DiD analysis hold.

Example I

Difference-in-differences (DiD)

R code

```
library(foreign)
mydata ← read.dta("https://dss.princeton.edu/training/Panel101.dta")
head(mydata, 8)
```

Output

```
##   country year           y y_bin         x1         x2          x3   opinion op
## 1       A 1990  1342787840     1 0.27790365 -1.1079559  0.28255358 Str agree  1
## 2       A 1991 -1899660544     0 0.32068470 -0.9487200  0.49253848     Disag  0
## 3       A 1992   -11234363     0 0.36346573 -0.7894840  0.70252335     Disag  0
## 4       A 1993  2645775360     1 0.24614404 -0.8855330 -0.09439092     Disag  0
## 5       A 1994  3008334848     1 0.42462304 -0.7297683  0.94613063     Disag  0
## 6       A 1995  3229574144     1 0.47721413 -0.7232460  1.02968037 Str agree  1
## 7       A 1996  2756754176     1 0.49980500 -0.7815716  1.09228814     Disag  0
## 8       A 1997  2771810560     1 0.05162839 -0.7048455  1.41590083 Str agree  1
```

For more details, see the "DiD (Manual Calculation of DiD).R" file, under 'Instructor's Session Files

Difference-in-differences (DiD)

- **Difference in Differences (DiD) method is like comparing the differences in outcomes (Y) before and after a change, between a group that experienced the change (treatment group) and a group that didn't (control group), to figure out the real impact of that change.**
    - The **'Time' column** is about when the observation was made (before or after the treatment).
    - **'Time' column** is 1 for years after 1994 and 0 for earlier years.
    - The **'treated' column** is about whether the observation was part of the group that received the treatment or not.
    - **'treated' column** is 1 for treatment group and 0 for control group.

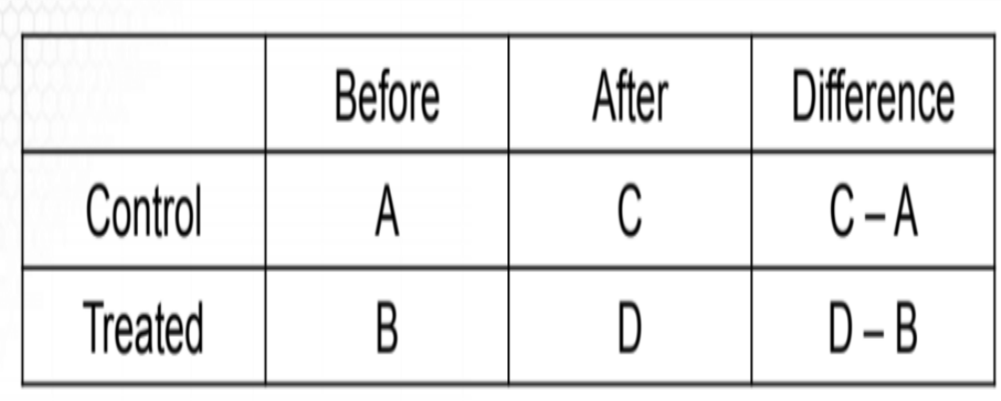
R code

```r
mydata$time=ifelse(mydata$year>=1994,1,0)
table(mydata$time)
mydata$treated = ifelse(mydata$country == "E" | mydata$country == "F" | mydata$country == "G", 1, 0)
table(mydata$treated)
```

Output

```
##
##  0  1
## 28 42
##
##  0  1
## 40 30
```

Manual Calculation of DiD



	Before	After	Difference
Control	A	C	C – A
Treated	B	D	D – B

R code

```
a = sapply(subset(mydata, treated == 0 & time == 0, select=y), mean)
b = sapply(subset(mydata, treated  == 1 & time== 0, select=y), mean)
c = sapply(subset(mydata, treated  == 0 & time == 1, select=y), mean)
d = sapply(subset(mydata, treated  == 1 & time == 1, select=y),mean)
DID = (d-b)-(c-a)
DID
```

Output

```
##           y
## -2519511630
```

For more details, see the "DiD (Manual Calculation of DiD).R" file, under 'Instructor's Session Files

Difference-in-differences (DiD)

- **Estimating the DID estimator**
    - Generate the interaction). We will call this interaction ‘did’.

R code

```
mydata$did = mydata$time * mydata$treated
head(mydata$did,50)
```

Output

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [39] 0 0 0 0 0 0 1 1 1 1 1 1
```

Difference-in-differences (DiD)

R code

```
didreg = lm(y ~ treated + time + did, data = mydata)
summary(didreg)
```

Output

```
Call:
lm(formula = y ~ treated + time + did, data = mydata)

Residuals:
       Min         1Q     Median         3Q        Max
-9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.581e+08  7.382e+08   0.485   0.6292
treated      1.776e+09  1.128e+09   1.575   0.1200
time         2.289e+09  9.530e+08   2.402   0.0191 *
did         -2.520e+09  1.456e+09  -1.731   0.0882 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Difference-in-differences (DiD)

- **Look at the Coefficient of the Interaction Term**: It represents the DiD estimator, showing the additional effect of the treatment over time, compared to the control group.

- **Positive or Negative Effect**

- **Statistical Significance**: Check if the result is statistically significant (usually indicated by p-values). A significant result means you can be more confident that the effect you're seeing is not just due to chance.

```
Call:
lm(formula = y ~ treated + time + did, data = mydata)

Residuals:
       Min         1Q     Median         3Q        Max
-9.768e+09 -1.623e+09  1.167e+08  1.393e+09  6.807e+09

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.581e+08  7.382e+08   0.485   0.6292
treated      1.776e+09  1.128e+09   1.575   0.1200
time         2.289e+09  9.530e+08   2.402   0.0191 *
did         -2.520e+09  1.456e+09  -1.731   0.0882 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Aligning DiD Calculations with Precision

- **Here's a quick clarification:**
  - **Manual Calculation of DiD:** use the exact figures (no rounding) for the DiD calculations, then round your final DiD result to two decimal places.
  - **How DiD is integrated into regression models:** the linear regression model should also be rounded to two decimal places at the end, not during the process.
- By following this approach, your results for the manual calculation of DiD will align with those from the regression models.
  - Round to two decimal places at the end of the process, not during it.

Can Difference-in-Differences (DiD) estimator be negative?(Revisit)

- In a Difference-in-Differences (DiD) analysis, you're basically comparing the change over time between two groups: one that received some kind of treatment (like a new program or policy) and one that didn't (the control group). You want to see if the treatment had any effect.

- If you're getting a negative number from your manual calculation and a positive number from your regression, here are a few points you might want to consider:

  - **Consistency in Labels**: Ensure that the labels 'After', 'Before', 'Treatment', and 'Control' are consistently used in both your manual calculation and regression model. Any discrepancy in labeling can result in opposite signs.

  - **Double-Check**: Go back and make sure that the data and labels match for both your manual and regression calculations.

Can Difference-in-Differences (DiD) estimator be negative?(Revisit)

- **When you do this manually and get a negative number, it suggests the treatment group did worse compared to the control group over time. Your intuition might tell you that the program had a negative effect on the treatment group compared to the control group.** They either didn't improve as much or possibly even did worse.
- What does "check with intuition" mean?
  - It means to think about whether your result makes sense given what you know about the situation:
  - Does it make sense that the program would make things worse for the treatment group?
  - Was there something else going on at the same time that could explain the negative number?
  - Does the negative number align with what you've observed or other information you have?

Q&A

- **Q:** which stock is most/least risky based on standard deviation. For this purpose are we to consider the standard deviation of the "market" to be a stock?`

Q&A (cont'd)

- **Q:** which stock is most/least risky based on standard deviation. For this purpose are we to consider the standard deviation of the "market" to be a stock?`
  - **A:** In simple terms, no, we usually don't treat the market's standard deviation as a stock. The market is more like a backdrop to compare how risky each individual stock is. We look at the market to get a sense of the overall risk, but we don't call it a stock itself.

Open for discussion

What is the topic for next week?

Module 2: Finance & Investments

Simple Returns

- Calculating Simple Returns in R
  - This example demonstrates how to calculate the simple return of a stock, including dividends. The formula for the simple return is:
    - $\text{Simple Return} = \frac{\text{Closing Price}_{\text{end}} + \text{Dividend} - \text{Closing Price}_{\text{start}}}{\text{Closing Price}_{\text{start}}}$

Output

```
        Date Close Dividend SimpleReturn
1 2023-01-01   100        0           NA
2 2023-01-02   105        2         0.07
```

Calculating simple returns in R

R code

```r
# Example data
data <- data.frame(
  Date = as.Date(c('2023-01-01', '2023-01-02')),
  Close = c(100, 105),  # Adjusted closing prices
  Dividend = c(0, 2)    # Dividends paid out
)

# Calculating simple return
data$SimpleReturn <- with(data, (Close + Dividend) / lag(Close)-1)

# Viewing the results
print(data)
```

Output

```
##         Date Close Dividend SimpleReturn
## 1 2023-01-01   100        0           NA
## 2 2023-01-02   105        2         0.07
```

For more details, see the "Simple-Return.R" file, under 'Instructor's Session Files

Calculating simple returns in R (cont'd)

Note:

- The `shift()` function from the `data.table` package achieves the same result as the `dplyr` package's `lag()` function.
  - Please click on the link provided below.
  - `shift` function