SICSS-Istanbul

Causal inference with observational data

Dr. Serkant Adıgüzel

Sabancı University

5 July 2023

Today's agenda

- Causality and observational data
  - Confounding bias
  - Statistical control
  - Before-and-after design and cross-section comparison design
- Difference-in-differences (DiD)
  - Canonical setup
  - Staggered setup
- Regression Discontinuity Design (RDD)
- Application (DiD)

Causality, causal effects and the counterfactual

- I will use some material from the following textbook: Imai, K., & Williams, N. W. (2022). *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press
- It has an excellent R library `qss` that compiles various datasets used in the textbook. Please load it from Imai's Github: https://github.com/kosukeimai/qss

- Causality: one of the most central concepts in quantitative social science.
- We need to infer the counterfactual outcome and compare it with the factual outcome.
- Fundamental problem of causal inference: we do not observe the counterfactual!
- Need a credible way to infer unobserved counterfactual outcomes.
- Experimental research $\rightarrow$ how a treatment causally affects an outcome by assigning varying values of the treatment variable to different observations, and measuring their corresponding values of the outcome variable.
- Randomized controlled trials (RCTs): researchers randomly assign the receipt of the treatment.
    - Randomization makes sure that two groups (treatment vs. control) are similar on average except for the receipt of the treatment
    - Gold standard in establishing causality.

Causal inference: the comparison between the factual (what actually happened) and the counterfactual (what would have happened if a key condition were different).

Potential outcomes framework

Let Y and T be the outcome and treatment variable, respectively.

$Y(0)$ = the potential outcome in the absence of a treatment

$Y(1)$ = the potential outcome in the presence of a treatment

For each observation i, we can define the causal effect of a binary treatment T_i as the difference between two potential outcomes:

$$Y_i(1) - Y_i(0)$$

- But we only observe either $Y_i(1)$ or $Y_i(0)$!
- Need a credible way to infer these unobserved counterfactual outcomes!
- This requires making certain assumptions.
- "Identification assumptions"
- How are we going to estimate causal effects?
    - Randomization!

What happens when randomization is impractical, unethical, or impossible?

Observational Studies

- Unlike RCTs, researchers do not conduct an intervention.
- They just observe naturally occurring things and record the data.
- Internal validity is likely to be compromised (selection bias).
- But external validity is probably stronger than RCTs!
- Because results are more generalizable as the data comes from a real-world environment.

Let's start with an example!

Minimum wage and unemployment

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

First sentence in the paper: How do employers in a low-wage labor market respond to an increase in the minimum wage?

- They examine the impact of raising the minimum wage on employment in the fast-food industry.
- In 1992, NJ raised the minimum wage from $4.25 to $5.05 per hour.
- Answering the question within the NJ context requires: inference about the counterfactual!
- The counterfactual: NJ employment rate in the absence of such a raise in minimum wage.
    - It is not observable!
- We must use the observed data to estimate it.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

- The challenge: finding a similar state in which the minimum wage did not increase.
- They chose the neighboring state: PA.
- They claim that the fast food restaurants in these two states are comparable.
- Do you believe this claim? Why?
- **Cross-section comparison design!**
- Treatment group: Fast-food restaurants in NJ
- Control group: Fast-food restaurants in PA
- Data on the number of full-time employees, the number of part-time employees, hourly wage before and after the minimum wage increase

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

```
# Imai and Williams has an excellent R library ("qss")
# for various datasets they use in their textbook
library(qss)
library(magrittr)
library(dplyr)
data("minwage")

minwage %>% head() # same as head(minwage)
```

```
##        chain location wageBefore wageAfter fullBefore fullAfter
## 1     wendys       PA       5.00      5.25         20         0
## 2     wendys       PA       5.50      4.75          6        28
## 3 burgerking       PA       5.00      4.75         50        15
## 4 burgerking       PA       5.00      5.00         10        26
## 5        kfc       PA       5.25      5.00          2         3
## 6        kfc       PA       5.00      5.00          2         2
##   partBefore partAfter
## 1         20        36
## 2         26         3
## 3         35        18
## 4         17         9
## 5          8        12
## 6         10         9
```

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

```
# Check glimpse(minwage) as well!
dim(minwage)
```

```
## [1] 358   8
```

```
summary(minwage)
```

```
##     chain              location          wageBefore
##  Length:358         Length:358         Min.   :4.250
##  Class :character   Class :character   1st Qu.:4.250
##  Mode  :character   Mode  :character   Median :4.500
##                                        Mean   :4.618
##                                        3rd Qu.:4.987
##                                        Max.   :5.750
##    wageAfter       fullBefore       fullAfter
##  Min.   :4.250   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:5.050   1st Qu.: 2.125   1st Qu.: 2.000
##  Median :5.050   Median : 6.000   Median : 6.000
##  Mean   :4.994   Mean   : 8.475   Mean   : 8.362
##  3rd Qu.:5.050   3rd Qu.:12.000   3rd Qu.:12.000
##  Max.   :6.250   Max.   :60.000   Max.   :40.000
##    partBefore      partAfter
```

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

- Did fast-food restaurants comply with the law?

```
table(minwage$location)
```

```
##
## centralNJ   northNJ        PA   shoreNJ   southNJ
##        45       146        67        33        67
```

```
minwage$state <- ifelse(minwage$location=='PA', 'PA', 'NJ')
#minwage <- minwage %>%
#         mutate(state = if_else(location=='PA', 'PA', 'NJ'))
```

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

```
minwage <- minwage %>%
  mutate(above_min_before = ifelse(wageBefore >= 5.05, 1, 0),
         above_min_after = ifelse(wageAfter >= 5.05, 1, 0)
         )
```

```
state_props <- minwage %>%
  group_by(state) %>% summarise(prop_before = mean(above_min_before),
                                prop_after = mean(above_min_after))
```

```
state_props
```

```
## # A tibble: 2 × 3
##   state prop_before prop_after
##   <chr>       <dbl>      <dbl>
## 1 NJ         0.0893     0.997
## 2 PA         0.0597     0.0448
```

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

- Outcome: Proportion of full-time workers
- To estimate the causal effect of increasing minimum wage on employment among the NJ restaurants, we will use PA restaurants as the control group.
    - Inferring the counterfactual outcome based on the PA sample.
- How: Compare the sample mean between the NJ and PA restaurants after the NJ law went into effect (difference-in-means estimator)

```
# First, let's calculate the proportion of full-time workers
# after the min wage increase
minwage <- minwage %>%
  mutate(totalAfter = fullAfter + partAfter,
         fullPropAfter = fullAfter / totalAfter
         )
```

```
# Always check what you do!
minwage %>%
  select(chain, location, fullAfter,
          partAfter, fullPropAfter) %>% slice(1:5)
```

```
##        chain location fullAfter partAfter fullPropAfter
## 1     wendys       PA         0        36     0.0000000
## 2     wendys       PA        28         3     0.9032258
## 3 burgerking       PA        15        18     0.4545455
## 4 burgerking       PA        26         9     0.7428571
## 5        kfc       PA         3        12     0.2000000
```

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772.

```
state_averages <- minwage %>%
  group_by(state) %>%
  summarise(mean(fullPropAfter))
state_averages
```

```
## # A tibble: 2 × 2
##   state `mean(fullPropAfter)`
##   <chr>                 <dbl>
## 1 NJ                    0.320
## 2 PA                    0.272
```

```
state_averages[1, 2] - state_averages[2, 2]
```

```
##   mean(fullPropAfter)
## 1          0.04811886
```

- The results show that the increase in minimum wage in NJ did **not** increase unemployment in NJ. If anything, it caused a slight increase in employment!
- Do you agree with the conclusion? Why?

Confounding bias

Confounding bias

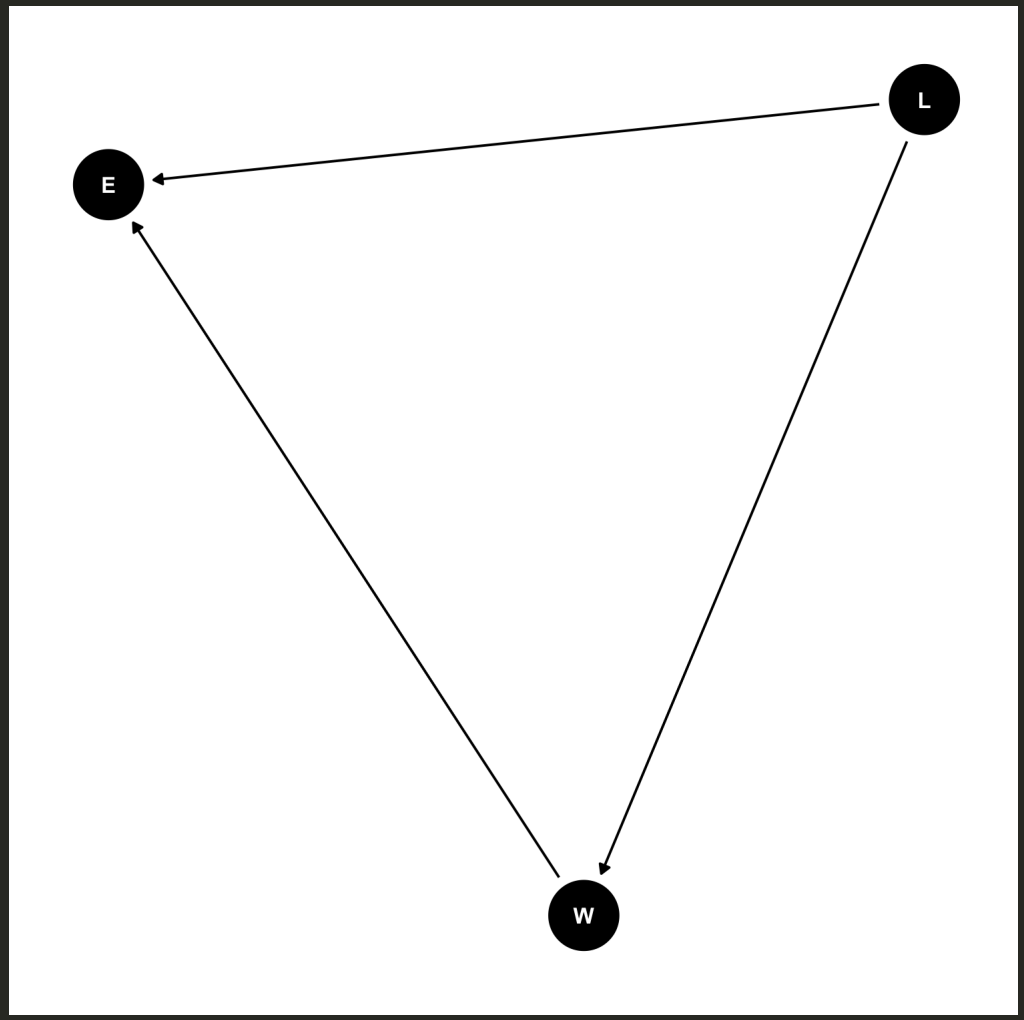
- Critical assumption: the treatment and control groups must be comparable with respect to everything related to the outcome other than the treatment!
    - This is the case in RCTs thanks to the randomization!
- Imagine a following scenario: there is a robust textile industry in NJ that attracts low-skilled workers (and no such industry in PA!)
    - Then PA restaurants cannot serve as a proper control for the NJ restaurants.
    - The fast-food restaurants in NJ could have a higher minimum wage than PA even before the wage increase to attract workers away from the textile industry.
    - Remember that we only compared mean employment rates **after** the min wage increase!
    - This means: we **cannot** infer the counterfactual outcome (NJ in the absence of minimum wage increase) from PA restaurants!

Confounding bias

- Any other differences that exist between the fast-food restaurants in the two states before the administration of the NJ law would bias our inference if they are also related to outcomes!
    - Can you think any other difference other than a competing industry such as textile?

```
# Let's visualize the relationship
library(ggdag) # To draw Directed Acyclic Graphs (DAG) in R
library(ggplot2)

set.seed(1990)

relationship <- dagify(W ~ L,E ~ L,E ~ W)

output <- ggdag(relationship) + theme_dag()
```

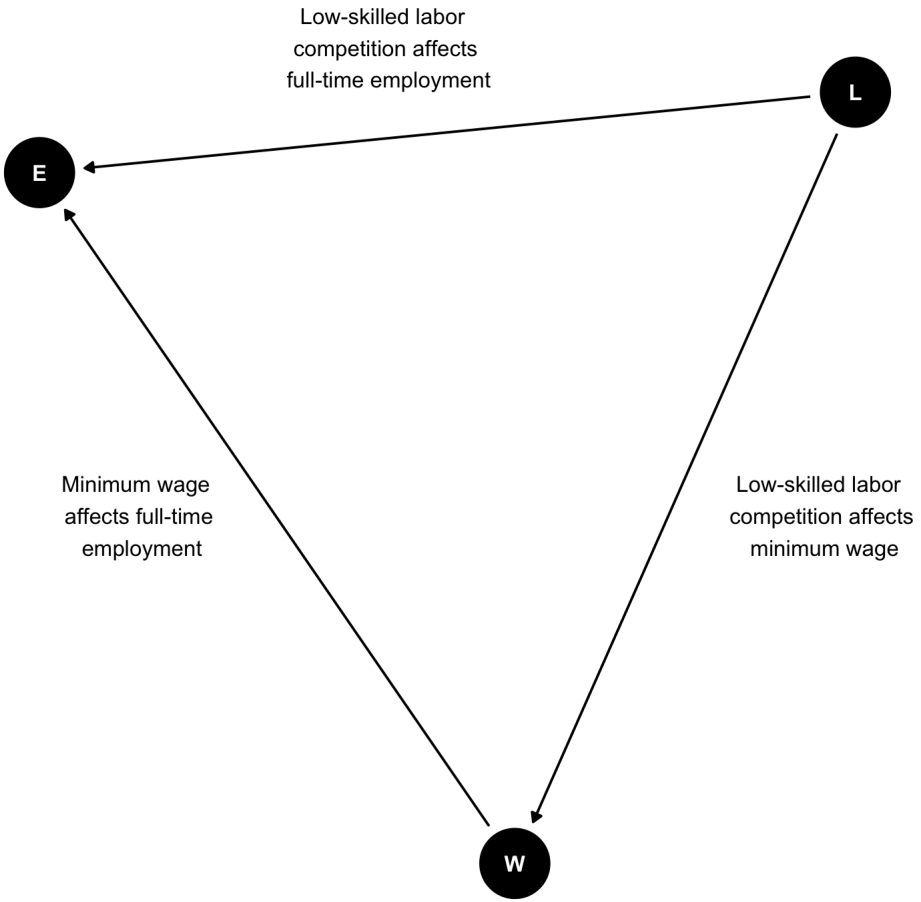
Confounding bias



```
output2 <- output + annotate(
    "text",
    x = c(-0.1, -0.4, 0.43),
    y = c(0.35, -0.2, -0.2),
    label = c(
      "Low-skilled labor \n competition affects \n full-time employme
      'Minimum wage \n affects full-time \n employment',
      'Low-skilled labor \n competition affects \n minimum wage'
    )
  )
```

The **pretreatment variables** that are associated with both the treatment and outcome variables are known as **confounders**.

- In this case, low-skilled labor competition is the confounder.



Confounding bias

- Pre-treatment variables are not affected by treatment as the name suggests.
- But, they can determine who is likely to receive treatment and influence the outcome!
- That is why, they **confound** the relationship between the treatment and the outcome $\rightarrow$ making it impossible to draw causal inferences from observational data.
- Textile industry: by increasing the demand for low-skilled labor, it might increase the demand for minimum wage increase.
    - Determining who (e.g., which state) is likely to receive treatment.
- Confounding bias due to self-selection $\rightarrow$ selection bias.
    - Why did NJ increase the minimum wage in 1992?
    - If, say, textile industry and competition for low-skilled labor was increasing the employment and demand for higher wages, then cross-sectional comparison that we did at the beginning was wrong!

Confounding bias

- Unable to control the treatment assignment: those who *self-select* themselves into the treatment will be significantly different from those who do not (both observed and unobserved).
- Observed difference between treatment and control groups cannot be attributed to the treatment!
- One solution for confounding bias: statistical control.
    - One way of doing this: sub-classification (e.g., adding controls in a regression setting).
    - Make treatment and control groups similar.

```
table(minwage$chain, minwage$state)
```

```
##
##                NJ  PA
##   burgerking  118  31
##   kfc          65  10
##   roys         73  15
##   wendys       35  11
```

Confounding bias

```
chains_state <- minwage %>% group_by(state) %>% count(chain)
chains_state
```

```
## # A tibble: 8 × 3
## # Groups:   state [2]
##   state chain          n
##   <chr> <chr>      <int>
## 1 NJ    burgerking   118
## 2 NJ    kfc           65
## 3 NJ    roys          73
## 4 NJ    wendys        35
## 5 PA    burgerking    31
## 6 PA    kfc           10
## 7 PA    roys          15
## 8 PA    wendys        11
```

Confounding bias

```
is.grouped_df(chains_state)
```

```
## [1] TRUE
```

```
chains_state <- chains_state %>% mutate(proportion = n / sum(n))
chains_state
```

```
## # A tibble: 8 × 4
## # Groups:   state [2]
##   state chain          n proportion
##   <chr> <chr>      <int>      <dbl>
## 1 NJ    burgerking   118      0.405
## 2 NJ    kfc           65      0.223
## 3 NJ    roys          73      0.251
## 4 NJ    wendys        35      0.120
## 5 PA    burgerking    31      0.463
## 6 PA    kfc           10      0.149
## 7 PA    roys          15      0.224
## 8 PA    wendys        11      0.164
```

Confounding bias

```
chains_state %>% tidyr::pivot_wider(
  names_from = state, values_from = proportion)
```

```
## # A tibble: 8 × 4
##   chain          n     NJ     PA
##   <chr>      <int>  <dbl>  <dbl>
## 1 burgerking   118  0.405 NA
## 2 kfc           65  0.223 NA
## 3 roys          73  0.251 NA
## 4 wendys        35  0.120 NA
## 5 burgerking    31 NA      0.463
## 6 kfc           10 NA      0.149
## 7 roys          15 NA      0.224
## 8 wendys        11 NA      0.164
```

Confounding bias

```
chains_state %>% tidyr::pivot_wider(-n,
            names_from = state, values_from = proportion)
```

```
## # A tibble: 4 × 3
##   chain         NJ    PA
##   <chr>      <dbl> <dbl>
## 1 burgerking 0.405 0.463
## 2 kfc        0.223 0.149
## 3 roys       0.251 0.224
## 4 wendys     0.120 0.164
```

What if, for whatever reason, Burger King has a different wage policy for its employees than other fast-food chains?

```r
library(tidyr)
minwage %>%
  group_by(state, chain) %>%
  summarise(fullPropAfterChain = mean(fullPropAfter))
```

```
## # A tibble: 8 × 3
## # Groups:   state [2]
##   state chain      fullPropAfterChain
##   <chr> <chr>                   <dbl>
## 1 NJ    burgerking              0.358
## 2 NJ    kfc                     0.328
## 3 NJ    roys                    0.283
## 4 NJ    wendys                  0.260
## 5 PA    burgerking              0.321
## 6 PA    kfc                     0.236
## 7 PA    roys                    0.213
## 8 PA    wendys                  0.248
```

```
minwage %>%
  group_by(state, chain) %>%
  summarise(fullPropAfterChain = mean(fullPropAfter)) %>%
  arrange(chain)
```

```
## # A tibble: 8 × 3
## # Groups:   state [2]
##   state chain      fullPropAfterChain
##   <chr> <chr>                   <dbl>
## 1 NJ    burgerking              0.358
## 2 PA    burgerking              0.321
## 3 NJ    kfc                     0.328
## 4 PA    kfc                     0.236
## 5 NJ    roys                    0.283
## 6 PA    roys                    0.213
## 7 NJ    wendys                  0.260
## 8 PA    wendys                  0.248
```

```
minwage %>% group_by(state, chain) %>%
  summarise(fullPropAfterChain = mean(fullPropAfter)) %>%
  pivot_wider(names_from = state,
              values_from = fullPropAfterChain)
```

```
## # A tibble: 4 × 3
##   chain         NJ    PA
##   <chr>      <dbl> <dbl>
## 1 burgerking 0.358 0.321
## 2 kfc        0.328 0.236
## 3 roys       0.283 0.213
## 4 wendys     0.260 0.248
```

```
minwage %>%
  group_by(state, chain) %>%
  summarise(fullPropAfterChain = mean(fullPropAfter)) %>%
  pivot_wider(names_from = state,
              values_from = fullPropAfterChain) %>%
  mutate(difference_full = NJ - PA)
```

```
## # A tibble: 4 × 4
##   chain         NJ    PA difference_full
##   <chr>      <dbl> <dbl>           <dbl>
## 1 burgerking 0.358 0.321          0.0364
## 2 kfc        0.328 0.236          0.0918
## 3 roys       0.283 0.213          0.0697
## 4 wendys     0.260 0.248          0.0117
```

The results show that proportion of full-time employers increased in all four fast-food chain restaurants.

The results seem to be similar across different chains.

```r
m1 <- lm(fullPropAfter ~ state, data = minwage)
summary(m1)
```

```
##
## Call:
## lm(formula = fullPropAfter ~ state, data = minwage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32040 -0.20547 -0.03869  0.17960  0.67960
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.32040    0.01467  21.836   <2e-16 ***
## statePA     -0.04812    0.03392  -1.419    0.157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2503 on 356 degrees of freedom
## Multiple R-squared:  0.005622,    Adjusted R-squared:  0.002828
## F-statistic: 2.013 on 1 and 356 DF,  p-value: 0.1569
```

```
m2 <- lm(fullPropAfter ~ state + chain, data = minwage)
summary(m2)
```

```
##
## Call:
## lm(formula = fullPropAfter ~ state + chain, data = minwage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36038 -0.20303 -0.02705  0.16366  0.68342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36038    0.02154  16.733   <2e-16 ***
## statePA     -0.04915    0.03380  -1.454   0.1468
## chainkfc    -0.03856    0.03527  -1.093   0.2751
## chainroys   -0.08121    0.03343  -2.429   0.0156 *
## chainwendys -0.09143    0.04193  -2.181   0.0299 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2485 on 353 degrees of freedom
## Multiple R-squared:  0.02834,    Adjusted R-squared:  0.01733
## F-statistic: 2.574 on 4 and 353 DF,  p-value: 0.03754
```

Confounding bias

Another possible confounding bias: location

- It could be that restaurants closer to PA are more appropriate group to compare against the PA restaurants.
- Solution: further sub-classification based on location.

Confounding bias

```
m3 <- lm(fullPropAfter ~ state + chain + location, data = minwage)
summary(m3)
```

```
##
## Call:
## lm(formula = fullPropAfter ~ state + chain + location, data = minwage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40837 -0.17875 -0.02682  0.15792  0.68842
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.28507    0.03979   7.164 4.67e-12 ***
## statePA          0.02139    0.04706   0.454  0.64977
## chainkfc        -0.01851    0.03493  -0.530  0.59657
## chainroys       -0.07311    0.03276  -2.232  0.02626 *
## chainwendys     -0.09165    0.04104  -2.233  0.02618 *
## locationnorthNJ  0.12330    0.04159   2.965  0.00324 **
## locationPA            NA         NA      NA       NA
## locationshoreNJ  0.09285    0.05584   1.663  0.09728 .
## locationsouthNJ -0.01547    0.04691  -0.330  0.74183
## ---
```

Our previous design was a cross-sectional comparison. But, one could observe units (restaurants, in this case) over time, which could be a valuable source of information as well.

Longitudinal data!

Before-and-after design

Before-and-after design

- Longitudinal data: Multiple measurements taken over time on the same units (also known as panel data).
- More credible comparison of control and treatment groups than cross-section data.
    - Because panel data contains information about each unit over time.
- In the NJ-PA wage study, there is information on employment before the minimum wage increase.
- Before-and-after-design: Comparison between pre- and post-treatment
    - Focuses only on NJ!

```
# Let's create a variable for the full time
# proportion before the min wage increase

minwage <- minwage %>%
  mutate(totalBefore = fullBefore + partBefore,
         fullPropBefore = fullBefore/totalBefore)
```

Before-and-after design

Let's calculate the difference between pre- and post-treatment employment in NJ:

```
minwage %>% filter(state == 'NJ') %>%
  summarise(before_diff =
              mean(fullPropAfter) - mean(fullPropBefore)
            )
```

```
##   before_diff
## 1  0.02387474
```

The difference between NJ and PA after the min wage increase was:

```
state_averages[1, 2]- state_averages[2, 2]
```

```
##   mean(fullPropAfter)
## 1          0.04811886
```

The before-and-after analysis gives an estimate that is quite similar to the one we found before (after NJ and after PA comparison).

Before-and-after design

Pros

- Confounding factors that are specific to NJ are held constant since comparison done within NJ!
  - For instance, assume that, for whatever reason, there is a more pro-labor culture in NJ. This is held constant in this design!

Cons

- It does not solve **time-varying** confounding factors.
  - Think about the previous high-demand for low-skilled labor in NJ due to (hypothetical) textile industry. Since increase in minimum wage impacts demand for low-skilled labor, confounding bias remains!
  - Imagine another scenario: wages and employment are improving in NJ because of a **positive technological shock**.
    - We would wrongly attribute the impact of this technological shock to the minimum wage increase.

We have seen two different research designs until now:

- Cross-section design: Compare post-NJ and post-PA
- Before-and-after design: Compare pre-NJ and post-NJ

Why not combine both?

Difference-in-differences (DiD)

Difference-in-differences (DiD)

- DiD design extends the before-and-after design to address the time-varying confounding bias.
- Key assumption in DiD: the outcome variable follows a parallel trend in the absence of treatment.

Difference-in-differences (DiD)

Canonical model

- Single treatment
- Two discrete periods (pre- and post-treatment)
- 2×2 design
- Let $Y_{i,t}(1)$ be the outcome of interest for unit i at time t if the unit receives treatment and $Y_{i,t}(0)$ as the outcome for unit i at time t if it does not receive treatment.
- The average treatment effect on the treated (ATT) is the causal estimand. It is defined as the difference $Y_{i,t}(1)$ - $Y_{i,t}(0)$ averaged across the units receiving the treatment.

Canonical model

Let δ = ATT and denote D as an indicator variable (1 if treatment).

$$\delta = E\Big[Y_{i,1}(1) - Y_{i,1}(0) \mid D_i = 1\Big]$$

Assuming no anticipation of treatment, so that $Y_{i,0}(0) = Y_{i,0}(1)$ and adding and subtracting $Y_{i,0}(0)$ above, we have:

$$\delta = E\Big[Y_{i,1}(1) - Y_{i,0}(1) \mid D_i = 1\Big] - E\Big[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 1\Big]$$

The second term cannot be directly estimated in the data because $Y_{i,1}(0) - Y_{i,0}(0)$ is unobservable for a unit that receives treatment.

Because of the parallel-trends assumption, $E\big[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 1\big] = E\big[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 0\big]$, we will have:

$$\delta = E\Big[Y_{i,1}(1) - Y_{i,0}(1) \mid D_i = 1\Big] - E\Big[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 0\Big]$$

Difference-in-differences (DiD)

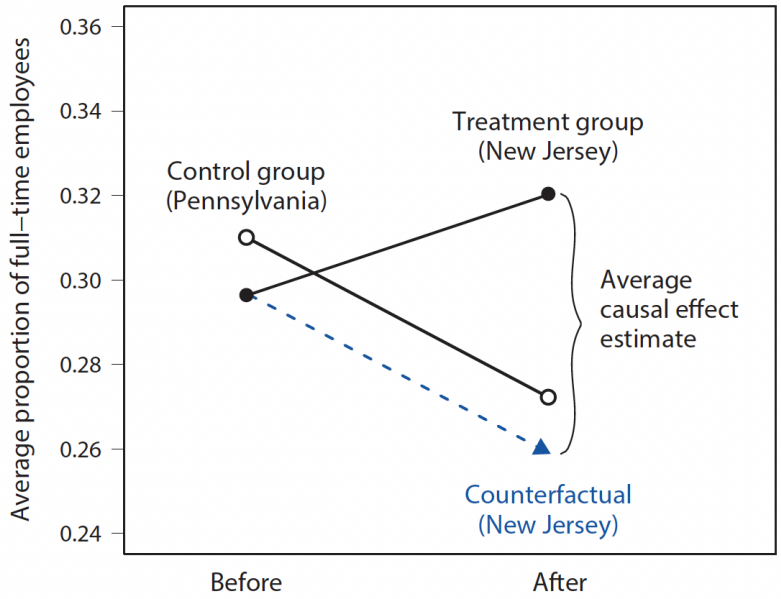


Figure 2.2. The Difference-in-Differences Design in the Minimum-Wage Study. The observed outcomes, i.e., the average proportion of full-time employees, are shown before and after the increase in the minimum wage for both the treatment group (fast-food restaurants in New Jersey; solid black circles) and the control group (restaurants in Pennsylvania; open black circles). Under the difference-in-differences design, the counterfactual outcome for the treatment group (solid blue triangle) is estimated by assuming that the time trend for the treatment group is parallel to the observed trend for the control group. The estimated average causal effect for New Jersey restaurants is indicated by the curly brace.

Parallel trends assumption

- We can estimate the counterfactual outcome for the treatment group by assuming that the time trend for the treatment group is parallel to the observed for the control group.
    - NJ would have experienced the same economic trend as PA in the absence of the treatment.
- Here, the quantity of interest under the DiD (estimand) is called: the sample average treatment effect for the treated (SATT).

$$SATT = \frac{1}{n_1} \sum_{i=1}^{n} T_i \{Y_i(1) - Y_i(0)\}$$

where $n_1 = \sum_{i=1}^{n} T_i$ is the size of the treatment group.

Let's just stop for a second.

$$SATE = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$$

$$SATT = \frac{1}{n_1}\sum_{i=1}^{n} T_i\{Y_i(1) - Y_i(0)\}$$

ATE = The average of the individual treatment effects of the population.

ATT = The average of the individual treatment effects of those treated.

In randomized experiments, ATE = ATT because the groups are comparable in expectation.

There are various different causal effects one can estimate depending on the research question: https://egap.org/resource/10-types-of-treatment-effect-you-should-know-about/

- What you want to estimate depends on your research question.
- In medical studies,for instance, you care about ATT more than ATE.

- Estimand: Our target parameter. We need to estimate this! (SATT)
- Estimator: A model/algorithm/rule that uses data to help us learn about the estimand.
- Estimate: The estimator's output.

Image credit: https://twitter.com/WomenInStat/status/1449068312873287682?s=20&t=l2sfj5B1nQuqn4mn1rFpNA

- DiD gives us an estimate for SATT!
- The difference in the before-and-after differences between the treatment (NJ) and control (PA) groups. More formally,

$$\left(\bar{Y}^{after}_{treated} - \bar{Y}^{before}_{treated}\right) - \left(\bar{Y}^{after}_{control} - \bar{Y}^{before}_{control}\right)$$

```
d <- data.frame(pre = c(70, 50), post = c(60, 55))
rownames(d) <- c('NJ', 'PA')
```

```
d
```

```
##    pre post
## NJ  70   60
## PA  50   55
```

DiD design = (60-70) - (55-50) = (-10) - (5) = -15

Cross-section design = 60 - 55 = 5

Before-and-after design = 60 - 70 = -10

```
# Did estimate
state_differences <- minwage %>%
  group_by(state) %>%
  summarise(diff = mean(fullPropAfter) - mean(fullPropBefore))

state_differences
```

```
## # A tibble: 2 × 2
##   state    diff
##   <chr>   <dbl>
## 1 NJ     0.0239
## 2 PA    -0.0377
```

```
state_differences %>%
  pivot_wider(names_from = state, values_from = diff) %>%
  mutate(diff_in_diff = NJ - PA)
```

```
## # A tibble: 1 × 3
##       NJ      PA diff_in_diff
##    <dbl>   <dbl>        <dbl>
## 1 0.0239 -0.0377       0.0616
```

- The DiD estimate is larger than the before-and-after estimate (0.0238) because of the negative trend in PA.

- ATT can also be obtained with OLS.

$$y_{it} = \alpha + \beta_1 D_i + \beta_2 POST_t + \beta_3 (D_i \times POST_t) + \epsilon_{it}$$

- ATT is estimated with β_3 here.

```
minwage_long <- minwage %>% select(chain, state, fullPropBefore,
          fullPropAfter)
head(minwage_long)
```

```
##        chain state fullPropBefore fullPropAfter
## 1     wendys    PA      0.5000000     0.0000000
## 2     wendys    PA      0.1875000     0.9032258
## 3 burgerking    PA      0.5882353     0.4545455
## 4 burgerking    PA      0.3703704     0.7428571
## 5        kfc    PA      0.2000000     0.2000000
## 6        kfc    PA      0.1666667     0.1818182
```

```
minwage_long$ID <- seq(1:nrow(minwage_long))
```

```
minwage_long <- minwage_long %>%
  tidyr::pivot_longer(fullPropBefore:fullPropAfter)
```

```
head(minwage_long)
```

```
## # A tibble: 6 × 5
##   chain      state    ID name           value
##   <chr>      <chr> <int> <chr>          <dbl>
## 1 wendys     PA        1 fullPropBefore 0.5
## 2 wendys     PA        1 fullPropAfter  0
## 3 wendys     PA        2 fullPropBefore 0.188
## 4 wendys     PA        2 fullPropAfter  0.903
## 5 burgerking PA        3 fullPropBefore 0.588
## 6 burgerking PA        3 fullPropAfter  0.455
```

```
minwage_long$period <- ifelse(minwage_long$name == 'fullPropBefore',
        0, 1)

minwage_long <- minwage_long %>% rename(fullprop = value)
```

```
head(minwage_long, 3)
```

```
## # A tibble: 3 × 6
##   chain  state    ID name           fullprop period
##   <chr>  <chr> <int> <chr>             <dbl>  <dbl>
## 1 wendys PA        1 fullPropBefore    0.5        0
## 2 wendys PA        1 fullPropAfter     0          1
## 3 wendys PA        2 fullPropBefore    0.188      0
```

```r
minwage_long$state <- factor(minwage_long$state, levels = c('PA', 'NJ

## Now, let's see whether we find the same result
did <- lm(fullprop ~ state*period, data = minwage_long)
```

```
summary(did)
```

```
##
## Call:
## lm(formula = fullprop ~ state * period, data = minwage_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32040 -0.19568 -0.04136  0.17960  0.70347
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.30997    0.02950  10.509   <2e-16 ***
## stateNJ        -0.01344    0.03272  -0.411    0.681
## period         -0.03768    0.04171  -0.903    0.367
## stateNJ:period  0.06156    0.04627   1.331    0.184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2414 on 712 degrees of freedom
## Multiple R-squared:  0.003918,    Adjusted R-squared:  -0.0002787
## F-statistic: 0.9336 on 3 and 712 DF,  p-value: 0.424
```

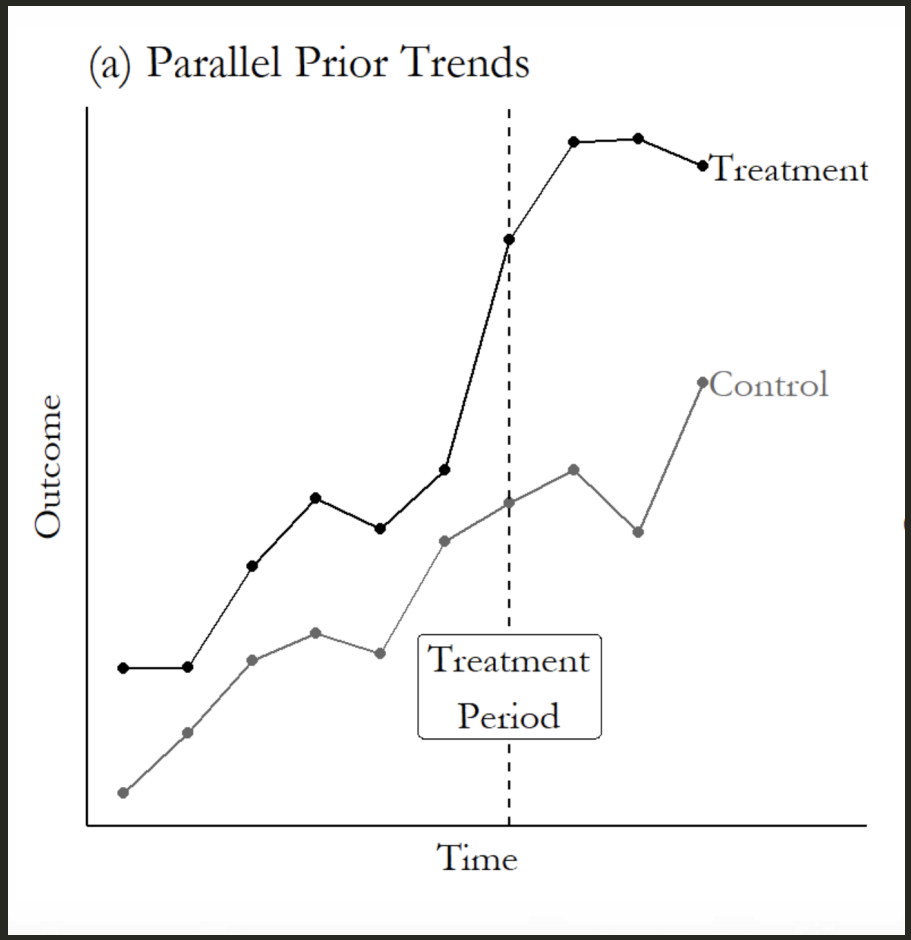
DiD design fails:

- time trend of the counterfactual outcome for the treatment group (NJ) is not parallel to the observed time trend for the control group (PA).
- How can we check this if we do not observe the counterfactual?
  - Pre-treatment trends!
- Event-study plot is useful to see whether the parallel trends assumption is satisfied. If the parallel trends assumption is satisfied, there should not be any significant difference between treatment and control before the treatment.

$$y_{it} = \alpha_i + \delta_t + \sum_{l=-K}^{-2} \mu_l D_{it}^l + \sum_{l=0}^{L} \mu_l D_{it}^l + \epsilon_{it}$$

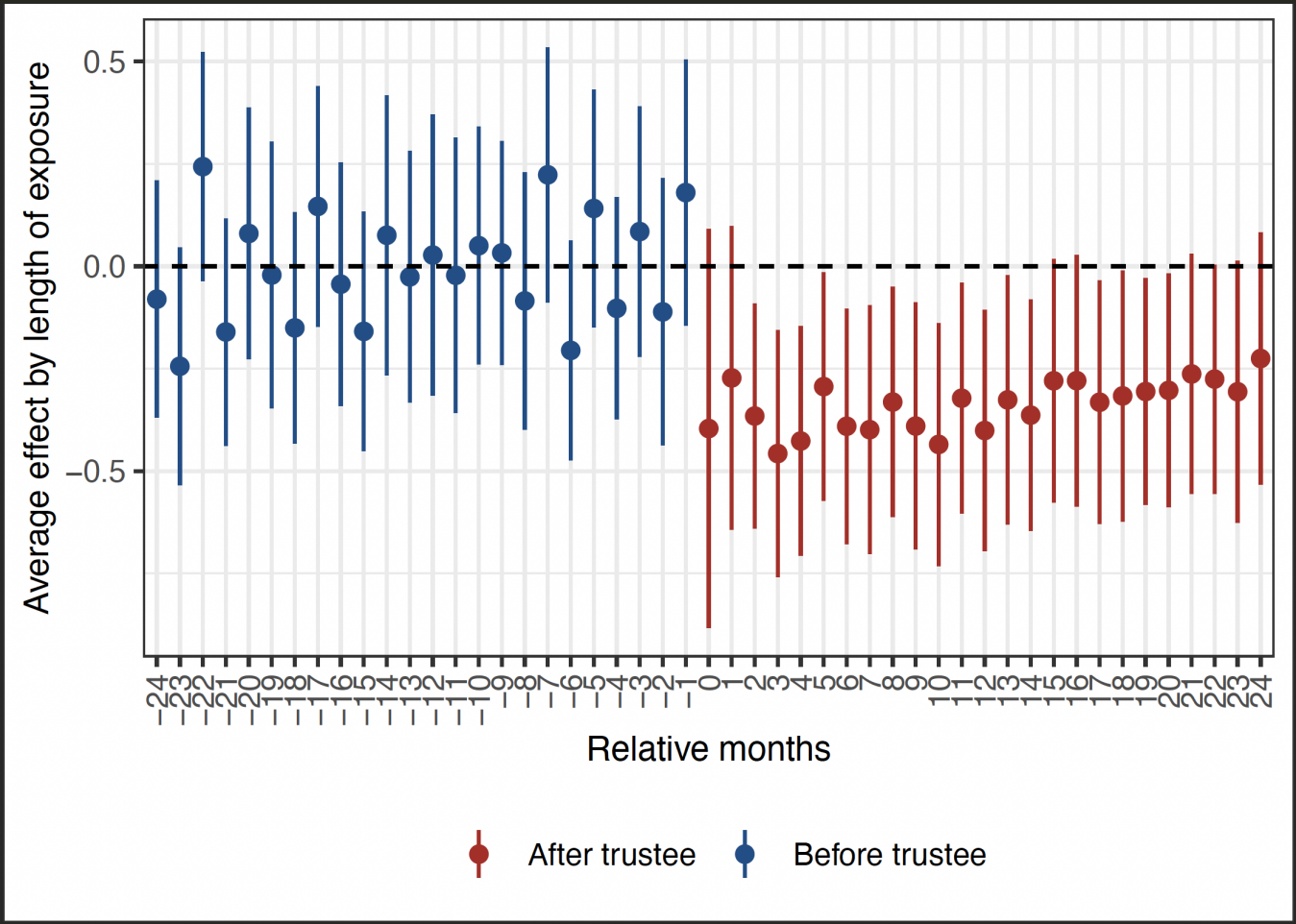
where D_{it}^l is an indicator for a treatment unit i being l periods from the start of treatment D (relative time indicators).

Parallel trends assumption



Huntington-Klein, Nick. The effect: An introduction to research design and causality. CRC Press, 2021. Available here: https://theeffectbook.net/index.html

Event-study plot: effect of trustee mayors on sealed-bid auctions



- You can generalize this simple 2×2 case to multiple time periods and units using the regression framework.

$$Y_{it} = \alpha_i + \delta_t + \beta D_{it} + \epsilon_{it}$$

where α_i and δ_t are unit and time fixed effects and β is our DiD estimate (for instance, "average effect of minimum wage increase").

- HOWEVER, when effects are heterogenous across units (for instance, states) and there are more than 2 time periods, α will be biased!

- More on this below, but bookmark this: https://asjadnaqvi.github.io/DiD/

Difference-in-differences (DiD)

Staggered model

- When there is heterogeneity in treatment timing, TWFE gives biased estimates!
- The DiD TWFE treatment effect is a "weighted average of all possible two-group/two-period DiD estimators in the data".
- An accessible paper to understand the problem behind using TWFE in such cases: Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, 144(2), 370-395.

Difference-in-differences (DiD)

Staggered model

- Suppose that we have three groups (observed from t_0 to T): a never-treated group (U), an earlier-treated group (k) that is treated at time t_k^*, and a later-treated group (l) that is treated at t_l^*. We have 4 different $2\times$ DiD comparisons:

1) Compare k or l with the control group (U) over the whole sample period (from t_0 to T).

2) The other two possible comparisons involve comparisons between the different treatment groups.

- Compare k with l over the window from t_0 to t_l^* (l = serves as control and k is the treatment group)
- Compare k with l over the window from t_k^* to T (l = treatment and k = serves as control.) Note that k is already treated too in this time window! Any problem here?

Difference-in-differences (DiD)

Staggered model

- Treated units sometimes serve as control group (an early-treated unit getting a negative weight if appears as a control for later-treated units)!
- Forbidden comparisons!
- This is problematic because the DiD estimate could reflect differences in treatment effects over time between different treatment cohorts.
- Solution: estimators that do not use already-treated units as controls!

Difference-in-differences (DiD)

The Callaway and Sant'Anna estimator (2021)

- Callaway and Sant'Anna (CS) (2021)
- It estimates cohort-time-specific treatment effects through simple 2×2s with clean controls: $ATT(g,t)$
  - $ATT(g,t) = E(Y_{i,t}(g) - Y_{i,t}(\infty) \mid G_i = g)$

We can identify $ATT(g,t)$ by comparing the expected change in outcome for cohort g between periods $g-1$ and t to that for a control group not-yet treated at period t. Under staggered versions of parallel trends and no anticipation assumption:

$$ATT(g,t) = E\Big[Y_{i,t} - Y_{i,g-1} \mid G_i = g\Big] - E\Big[Y_{i,t} - Y_{i,g-1} \mid G_i = g'\Big]$$

for any $g' > t$

Difference-in-differences (DiD)

The Callaway and Sant'Anna estimator (2021)

Since this holds for any comparison group $g' > t$, it also holds if we average over some set of comparisons G_{comp} such that $g' > t$ for all $g' \in G_{comp}$

$$ATT(g,t) = E\Big[Y_{i,t} - Y_{i,g-1} \mid G_i = g\Big] - E\Big[Y_{i,t} - Y_{i,g-1} \mid G_i \in G_{comp}\Big]$$

We can then estimate $ATT(g,t)$ by replacing expectations with their sample analogs:

$$\widehat{ATT(g,t)} = \frac{1}{N_g} \sum_{i:G_i=g} \big[Y_{i,t} - Y_{i,g-1}\big] - \frac{1}{N_{G_{comp}}} \sum_{i:G_i \in G_{comp}} \big[Y_{i,t} - Y_{i,g-1}\big]$$

- One can use never-treated units for G_{comp} or not-yet-treated units.
- The CS estimator also allows covariates.

Difference-in-differences (DiD)

The Callaway and Sant'Anna estimator (2021)

- Once you estimate cohort-specific ATTs, then it is easy to combine them to have an aggregate ATT. For instance, the average of the treatment effect l periods after adoption across different adoption cohorts:

$$ATT_l^w = \sum_g w_g ATT(g, g+l)$$

w_g chould be chosen to weight different cohorts equally, or in terms of their relative frequencies in the treated population.

- This approach has two advantages over TWFE:
  - It provides sensible estimates even under arbitrary heterogeneity of treatment effects.
  - It makes transparent exactly which units are being used as a control group to infer the unobserved potential outcomes.

Difference-in-differences (DiD)

- There are also alternative estimators to CS such as Borusyak et al. (2021)
- See the following paper for a good review: Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2022). What's trending in difference-indifferences? A synthesis of the recent econometrics literature. *arXiv preprint* arXiv:2201.01194.

The Callaway and Sant'Anna estimator: application

- R package: `did`
- The Effect of the Minimum Wage on Youth Employment
- Data: county level employment (2003 - 2007)

```
library(did)
data(mpdta)
head(mpdta, 3)
```

```
##     year countyreal     lpop     lemp first.treat treat
## 866 2003       8001 5.896761 8.461469        2007     1
## 841 2004       8001 5.896761 8.336870        2007     1
## 842 2005       8001 5.896761 8.340217        2007     1
```

The Callaway and Sant'Anna estimator: application

- You will need your data in a particular format!
    - It needs to be in long format.
    - You will need an ID (`countyreal`) and time (`year`) variable.
    - You will need a group variable (`first.treat`). This is the time period when an individual first becomes treated. For units that are never treated, this variable should be set equal to 0.

```
table(mpdta$first.treat)
```

```
##
##    0 2004 2006 2007
## 1545  100  200  655
```

```
head(mpdta, 2)
```

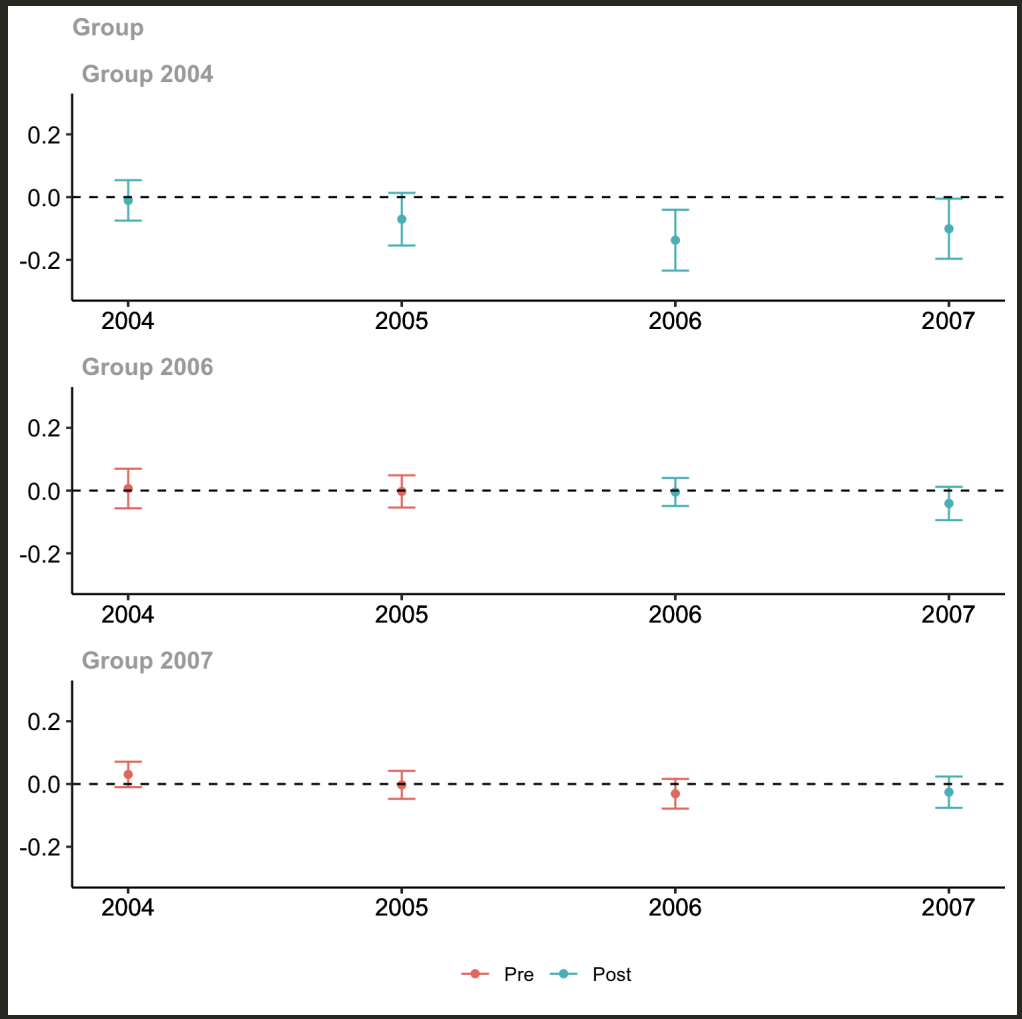
```
##     year countyreal     lpop     lemp first.treat treat
## 866 2003       8001 5.896761 8.461469        2007     1
## 841 2004       8001 5.896761 8.336870        2007     1
```

```r
# estimate group-time average treatment effects without covariates
mw.attgt <- att_gt(yname = "lemp",
                   gname = "first.treat",
                   idname = "countyreal",
                   tname = "year",
                   xformla = ~1,
                   data = mpdta,
                   )
```

```r
# summarize the results
summary(mw.attgt)
```

```
##
## Call:
## att_gt(yname = "lemp", tname = "year", idname = "countyreal",
##     gname = "first.treat", xformla = ~1, data = mpdta)
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Dif
##
## Group-Time Average Treatment Effects:
##  Group Time ATT(g,t) Std. Error [95% Simult.  Conf. Band]
##   2004 2004  -0.0105     0.0245       -0.0749      0.0539
##   2004 2005  -0.0704     0.0319       -0.1542      0.0134
##   2004 2006  -0.1373     0.0369       -0.2343     -0.0402 *
##   2004 2007  -0.1008     0.0364       -0.1966     -0.0050 *
##   2006 2004   0.0065     0.0240       -0.0566      0.0697
##   2006 2005  -0.0028     0.0196       -0.0542      0.0487
##   2006 2006  -0.0046     0.0170       -0.0492      0.0400
##   2006 2007  -0.0412     0.0202       -0.0943      0.0119
##   2007 2004   0.0305     0.0154       -0.0100      0.0711
##   2007 2005  -0.0027     0.0170       -0.0473      0.0419
##   2007 2006  -0.0311     0.0180       -0.0785      0.0163
##   2007 2007  -0.0261     0.0190       -0.0759      0.0238
## ---
## Signif. codes: `*' confidence band does not cover 0
##
```

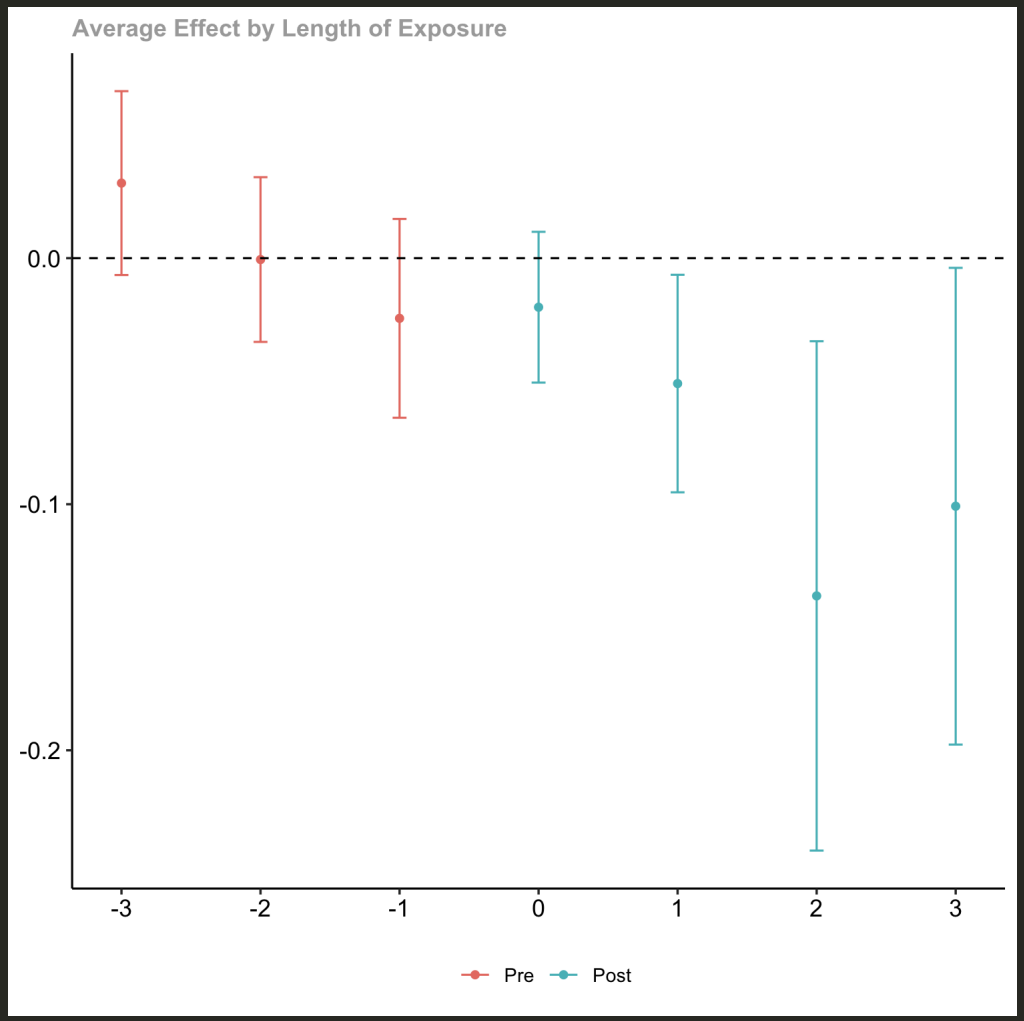
```
# set ylim so that all plots have the same scale along y-axis
ggdid(mw.attgt, ylim = c(-.3,.3))
```



```r
# aggregate the group-time average treatment effects
mw.dyn <- aggte(mw.attgt, type = "dynamic")
summary(mw.dyn)
```

```
##
## Call:
## aggte(MP = mw.attgt, type = "dynamic")
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Dif
##
##
## Overall summary of ATT's based on event-study/dynamic aggregation:
##      ATT    Std. Error     [ 95%  Conf. Int.]
##  -0.0772        0.0213    -0.1189     -0.0356 *
##
##
## Dynamic Effects:
##  Event time Estimate Std. Error [95% Simult.  Conf. Band]
##          -3   0.0305     0.0143       -0.0069      0.0679
##          -2  -0.0006     0.0128       -0.0340      0.0329
##          -1  -0.0245     0.0155       -0.0649      0.0159
##           0  -0.0199     0.0117       -0.0506      0.0107
##           1  -0.0510     0.0169       -0.0952     -0.0068 *
##           2  -0.1373     0.0397       -0.2408     -0.0338 *
##           3  -0.1008     0.0371       -0.1977     -0.0039 *
## ---
## Signif. codes: `*' confidence band does not cover 0
```

```
ggdid(mw.dyn)
```


Average Effect by Length of Exposure
0.0
-0.1
-0.2
-3
-2
-1
0
1
2
3
Pre
Post

Regression discontinuity design

- Another design that also allows us to address potential selection bias problem (due to confounders): regression discontinuity design (RDD).
- Let's see an example first.

Eggers, Andrew C., and Jens Hainmueller. "MPs for sale? Returns to office in postwar British politics." ***American Political Science Review*** **103.4 (2009): 513-533.**

- How much politicians can increase their personal wealth due to holding office?
- Data from British parliament.
- Why can't we compare MPs and non-MPs in terms of their wealth and see whether holding office increases personal wealth or not?
- Because MPs are different than non-MPs (almost everyone) in lots of observable and unobservable characteristics.
- The key idea in RDD: compare candidates who narrowly won the office with those who barely lost it.
- When one's margin of victory switches from a negative number to a positive number, we would expect a large, discontinuous, positive jump in the personal wealth of electoral candidates if serving an MP actually financially benefits them.
- Assuming that nothing else is going on at this point of discontinuity (what?), we can identify the average causal effect of being an MP on personal wealth.

Regression Discontinuity Design

- We identify the effect of being an MP on personal wealth at the threshold by comparing the candidates who barely won an election with those who barely lost it.
- We can use a regression model to predict the average personal wealth at the point of discontinuity.
- We can use a scatter plot with two regression lines to best understand the RD design.

```
# first, let's look at the data

data('MPs')

head(MPs, 2)
```

```
##     surname firstname  party ln.gross   ln.net  yob  yod
## 1 Llewellyn     David   tory 12.13591 12.13591 1916 1992
## 2    Morris     Claud labour 12.44809 12.44809 1920 2000
##   margin.pre             region      margin
## 1         NA              Wales  0.05690404
## 2         NA South West England -0.04973833
```

- We will first fit a linear regression model to the observations with a positive margin (i.e., the candidates who won the elections and became MPs)
- Then, we will fit another linear regression model to those with a negative margin (the candidates who lost).
- The difference in predicted values **at the point of discontinuity** (i.e., zero margin of victory) between the two regressions represents the average causal effect of serving as an MP on personal wealth.

```
# subset the data for separate linear regression models


labour_winners <- MPs %>% filter(party == 'labour', margin > 0)
labour_losers <- MPs %>% filter(party == 'labour', margin < 0)


tory_winners <- MPs %>% filter(party == 'tory', margin > 0)
tory_losers <- MPs %>% filter(party == 'tory', margin < 0)
```

- Now that we have two separate data frames for each party (labour and tory), we can fit our regression models.

```r
labour_fit_win <- lm(ln.net ~ margin, data = labour_winners)
labour_fit_lose <- lm(ln.net ~ margin, data = labour_losers)

tory_fit_win <- lm(ln.net ~ margin, data = tory_winners)
tory_fit_lose <- lm(ln.net ~ margin, data = tory_losers)


tory_fit_lose
```

```
##
## Call:
## lm(formula = ln.net ~ margin, data = tory_losers)
##
## Coefficients:
## (Intercept)       margin
##      12.538        1.491
```

Now that we have our regression models, we can predict wealth at death based on the margin of the victory.

- Let's use `data_grid()` function from the `modelr` library to create a new dataset with the values of `margin` that are relevant for a given data subset.

```
library(modelr)
y1_labour_win <- labour_winners %>% data_grid(margin) %>%
  add_predictions(labour_fit_win)

y2_labour_lose <- labour_losers %>% data_grid(margin) %>%
  add_predictions(labour_fit_lose)
```

```
head(y1_labour_win)
```

```
## # A tibble: 6 × 2
##    margin  pred
##     <dbl> <dbl>
## 1 0.00243  11.9
## 2 0.00372  11.9
## 3 0.00647  11.9
## 4 0.00938  12.0
## 5 0.0150   12.0
## 6 0.0161   12.0
```

- Note that you did not need to use the `modelr` package.
- You can create a data frame for prediction based on observed values in the sample and then use the `predict` function.

```
prediction_data <- data.frame(margin =
  seq(min(labour_winners$margin), max(labour_winners$margin),
      by = 0.001))
head(prediction_data, 2)
```

```
##        margin
## 1 0.002428998
## 2 0.003428998
```

```
nrow(prediction_data)
```

```
## [1] 366
```

```
predicted_values <- predict(labour_fit_win,
                            newdata = prediction_data)
length(predicted_values)
```

```
## [1] 366
```

```
head(predicted_values, 2)
```

```
##        1        2
## 11.94006 11.94190
```

```
summary(y1_labour_win$margin)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.002429 0.035888 0.069758 0.083104 0.103853 0.367970
```

```
summary(y2_labour_lose$margin)
```

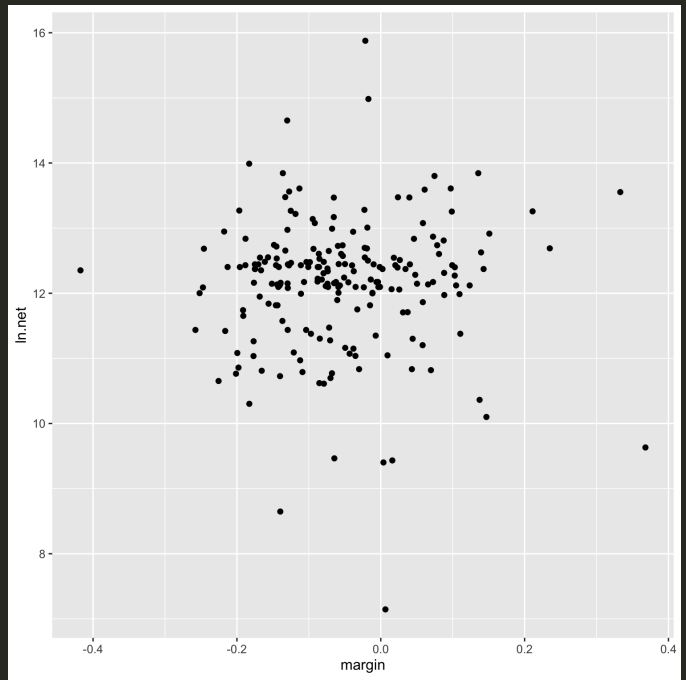
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.417494 -0.145597 -0.098552 -0.106493 -0.058526 -0.001031
```

Let's run the same models for Tories as well:

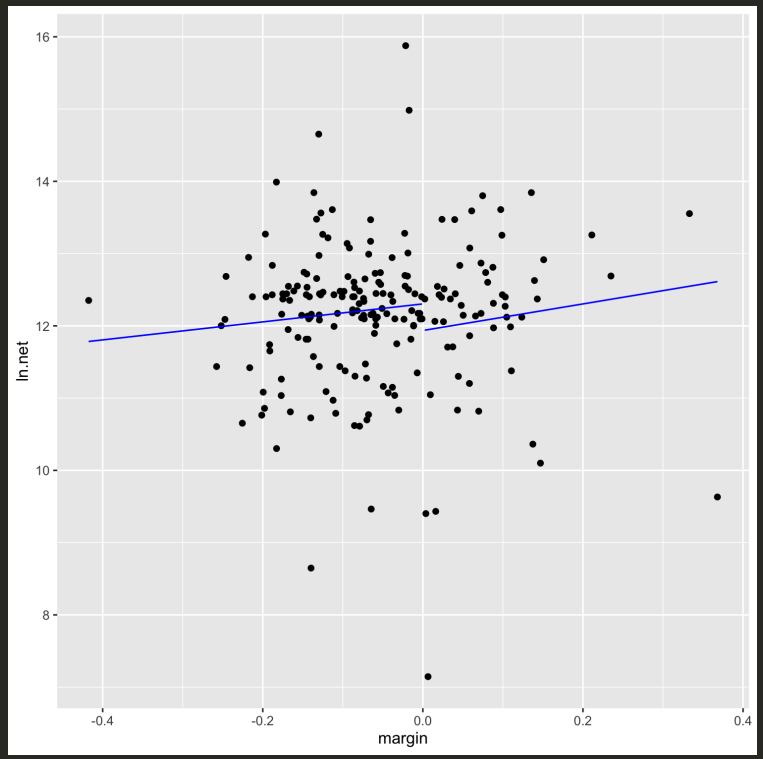
```
y3_tory_win <- tory_winners %>% data_grid(margin) %>%
  add_predictions(tory_fit_win)

y4_tory_lose <- tory_losers %>% data_grid(margin) %>%
  add_predictions(tory_fit_lose)
```

Now we can create our scatterplots in which we add our regression lines on top of observations.

```
# Code for the labor plot.
# Note that I have two data frames so I will have
# to declare datasets separately meaning that I
# cannot do it in the ggplot() part:

labor_plot <- ggplot() +
geom_point(data = labour_winners,
           aes(x = margin, y = ln.net)) +
geom_point(data = labour_losers,
           aes(x = margin, y = ln.net))

labor_plot
```

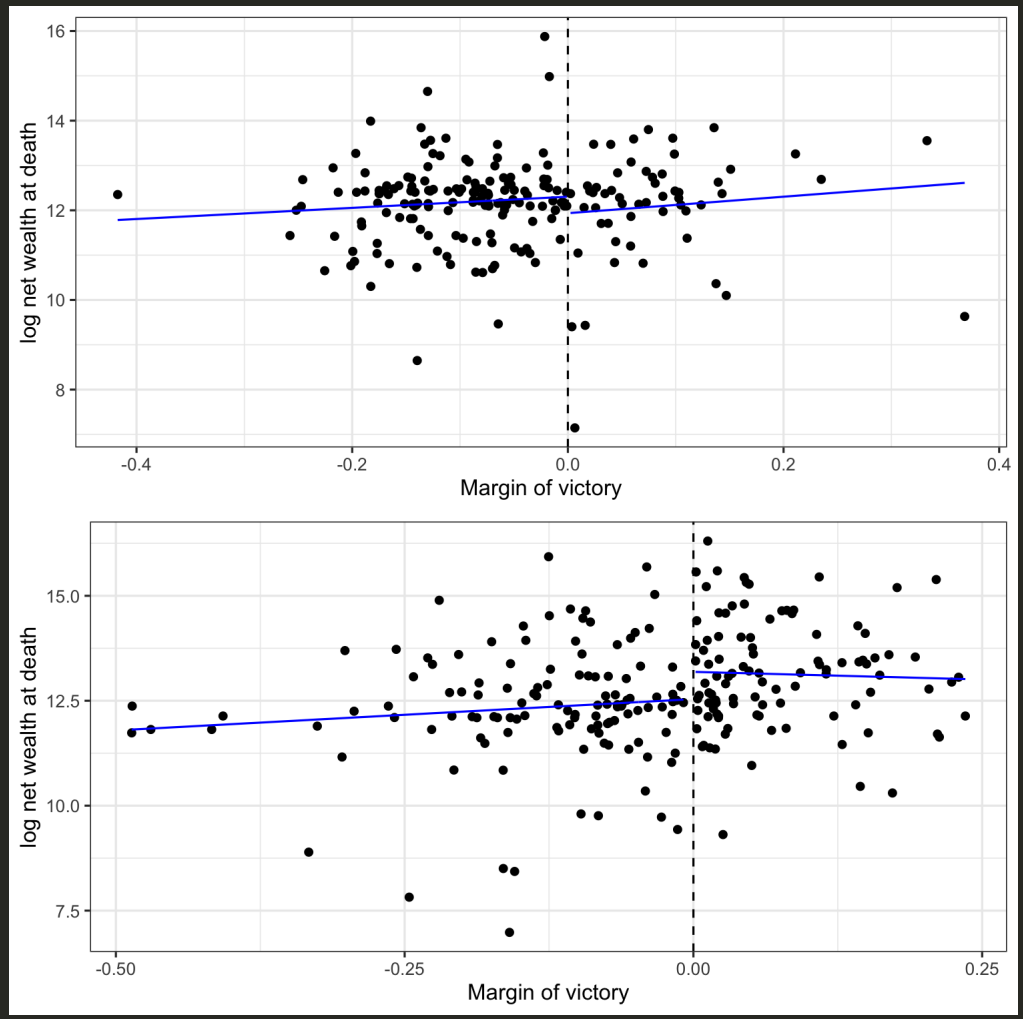


```
# Now let's add regression lines:

labor_plot <- labor_plot +
  geom_line(data = y1_labour_win, aes(x = margin, y = pred),
            color = 'blue') +
  geom_line(data = y2_labour_lose, aes(x = margin, y = pred),
            color = 'blue')

labor_plot
```



```r
# Add a vertical reference line and labels
labor_plot <- labor_plot +
  geom_vline(xintercept = 0, linetype = 'dashed') +
  labs(x = 'Margin of victory',
       y = 'log net wealth at death') +
  theme_bw()
```

```
library(gridExtra)
grid.arrange(labor_plot, tory_plot)
```



Regression Discontinuity Design

The results suggests that Tory MPs financially benefit from serving in office whereas Labour MPs do not. How large is the effect for Tory candidates?

- We can calculate predictions when the margin is 0 and then take the difference.

```
tibble(margin = 0)
```

```
## # A tibble: 1 × 1
##   margin
##    <dbl>
## 1      0
```

```
# spread_predictions add one column for each model:
tibble(margin = 0) %>%
  spread_predictions(tory_fit_win, tory_fit_lose)
```

```
## # A tibble: 1 × 3
##   margin tory_fit_win tory_fit_lose
##    <dbl>        <dbl>         <dbl>
## 1      0         13.2          12.5
```

Regression Discontinuity Design

```
tibble(margin = 0) %>%
  spread_predictions(tory_fit_win, tory_fit_lose) %>%
  mutate(rd_est = exp(tory_fit_win) - exp(tory_fit_lose)) %>%
  select(rd_est) %>% pull()
```

```
##        1
## 255050.9
```

- The estimated effect of being an MP on the personal wealth of Tory candidates is around 255,000 pounds.
- How big is this?

```
exp(mean(MPs$ln.net))
```

```
## [1] 252495.8
```

- The average wealth in the dataset is 252,000 pounds!
- Being an MP doubles your wealth for Tories!

Regression Discontinuity Design

- Let $y_i(0)$ and $y_i(1)$ be the pair of potential outcomes for unit i.
- $y_i^{obs} = y_i(\omega_i)$ where ω_i is 1 if $x_i \geq 0$.
- $\tau(x) = E(y_i(1) - y_i(0) \mid x_i = x)$
- RDD focuses on estimating the average effect of the treatment at the threshold (which is 0 here):

$$\tau = \tau(0)$$

- This average can be estimated as the discontinuity in the conditional expectation of y_0^{obs} as a function of the forcing variable, at the threshold:

$$\tau = lim_{x \downarrow 0} E(y_i^{obs} \mid x_i = x) - lim_{x \uparrow 0} E(y_i^{obs} \mid x_i = x)$$

- The question is how to estimate the two limits of the regression function at the threshold:

$\mu_+ = lim_{x \downarrow 0} E(y_i^{obs} \mid x_i = x)$ and $\mu_- = lim_{x \uparrow 0} E(y_i^{obs} \mid x_i = x)$

Regression Discontinuity Design

- Tradeoff between parametric and non-parametric approaches to estimate the limits as well as global vs. local sample.
- We used a global linear regression in the example above
  - Misspecification threat and using data far away from the cutoff.
- We could have used a local linear or high-order polynomial approach or kernel regressions as well.
  - Discard the units with x_i more than some bandwidth h away from the threshold and estimate a linear or quadratic (or higher) function on the remaining units.
  - Disadvantage: kernel regressions perform poorly at the boundary.
- See the following for more: Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6), 2295-2326.

DiD Application

DiD Application

- Research question: Do elections increase biased coverage towards the opposition in competitive authoritarian regimes?
- Data: Scraped newspaper articles from Sabah (January - June 2023) about AKP, CHP, Kılıçdaroğlu and Erdoğan.
- Outcome: negative sentiment towards the opposition
- Aim: answer the research question with a DiD design!
- "Treated": CHP articles after the official announcement of elections (10 March 2023).

Collecting URLs

```
from bs4 import BeautifulSoup
import requests

hdr = {'User-Agent': 'Mozilla/5.0'} #header settings

years = [2023]
months = [1, 2, 3, 4, 5, 6]

article_links = []
for year in years:
    for month in months:
        url =
        'https://www.sabah.com.tr/sitemaparchives/post/' +
        str(year) + '-' + str(month) + '.xml'
        print('In page:', url)
        wholepage = requests.get(url, headers = hdr)
        wholesoup = BeautifulSoup(wholepage.content)
        wholelist = wholesoup.findAll('loc')

        for url in wholelist:
            url_final = url.text
            article_links.append(url_final)
```

Parsing news stories

```
import pandas as pd
import requests
from bs4 import BeautifulSoup
import dateparser

url = 'https://www.sabah.com.tr/ekonomi/son-dakika-bakan-ozhaseki-mu

hdr = {'User-Agent': 'Mozilla/5.0'} #header settings


response = requests.get(url, headers=hdr).content

soup = BeautifulSoup(response)

maintext =
soup.find('div',{'class':'newsBox'}).text.strip()
title = soup.find('h1',{'class':'pageTitle'}).text.strip()
date_extracted =
soup.find('div',{'class':'news-detail-info'}).text.strip()
```

Labeling news (based on their titles)

```
from transformers import AutoModelForSequenceClassification,
AutoTokenizer, pipeline

model =
AutoModelForSequenceClassification.from_pretrained(
  "savasy/bert-base-turkish-sentiment-cased")

tokenizer = AutoTokenizer.from_pretrained(
  "savasy/bert-base-turkish-sentiment-cased")

sentiment = pipeline("sentiment-analysis",
tokenizer=tokenizer, model=model)

title = sentiment("SON DAKİKA: Bakan Özhaseki müjdeyi verdi!
Konut teslimatları Ekim, Kasım ve Aralık ayında yapılacak")
```

DiD Application

```
library(lubridate)

# create a week variable
d$week <- week(d$date_publish)

# a dummy variable for titles with the negative sentiment
d$negative <- ifelse(d$sentiment_title == 'negative', 1, 0)

# getting rid of publish hours
d$date_publish <- as.Date(d$date_publish)

# Calculate daily averages
#day_level <- d %>% group_by(date_publish, chp) %>% summarise(average

# a dummy variable for the election period.
d$election_period <- ifelse(
  d$date_publish > '2023-03-10', 1, 0)

# 'treated' variable which is 1 for chp articles during
# election period

d$treated <- ifelse(
  d$date_publish > '2023-03-10'
  & d$chp == 1, 1, 0)
```

Remember that in a canonical 2 \times 2, we can estimate the following:

$$y_{it} = \alpha + \beta_1 D_i + \beta_2 POST_t + \beta_3 (D_i \times POST_t) + \epsilon_{it}$$

```
library(fixest)

# with the lm() function
#m1 <- lm(average_negative ~ chp * election_period +
# as.factor(date_publish), data = day_level)

# you can the faster feols() for fast fixed effect estimations
m1 <- feols(negative ~ chp*election_period
            | date_publish, data = d)
```

```r
# Since treatment is 1 for all days after
# the official announcement, election_period (beta 2 in the
# specification above) will be omitted given the day FEs
summary(m1)
```

```
## OLS estimation, Dep. Var.: negative
## Observations: 3,104
## Fixed-effects: date_publish: 181
## Standard-errors: Clustered (date_publish)
##                     Estimate Std. Error   t value  Pr(>|t|)
## chp                 0.525786   0.034666 15.167199 < 2.2e-16 ***
## chp:election_period 0.005104   0.039074  0.130618   0.89622
## ... 1 variable was removed because of collinearity (election_period)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.404939     Adj. R2: 0.284517
##                  Within R2: 0.272545
```

The interaction shows (0.005) that election period has no effect.

```
modelsummary::modelsummary(m1)
```

	Model 1
chp	0.526
	(0.035)
chp × election_period	0.005
	(0.039)
Num.Obs.	3104
R2	0.326
R2 Adj.	0.285
R2 Within	0.273
R2 Within Adj.	0.272
RMSE	0.40
Std.Errors	by: date_publish
FE: date_publish	X

Note that the canonical design above can be estimated with the more generalized version as well. You will get the same DiD estimate.

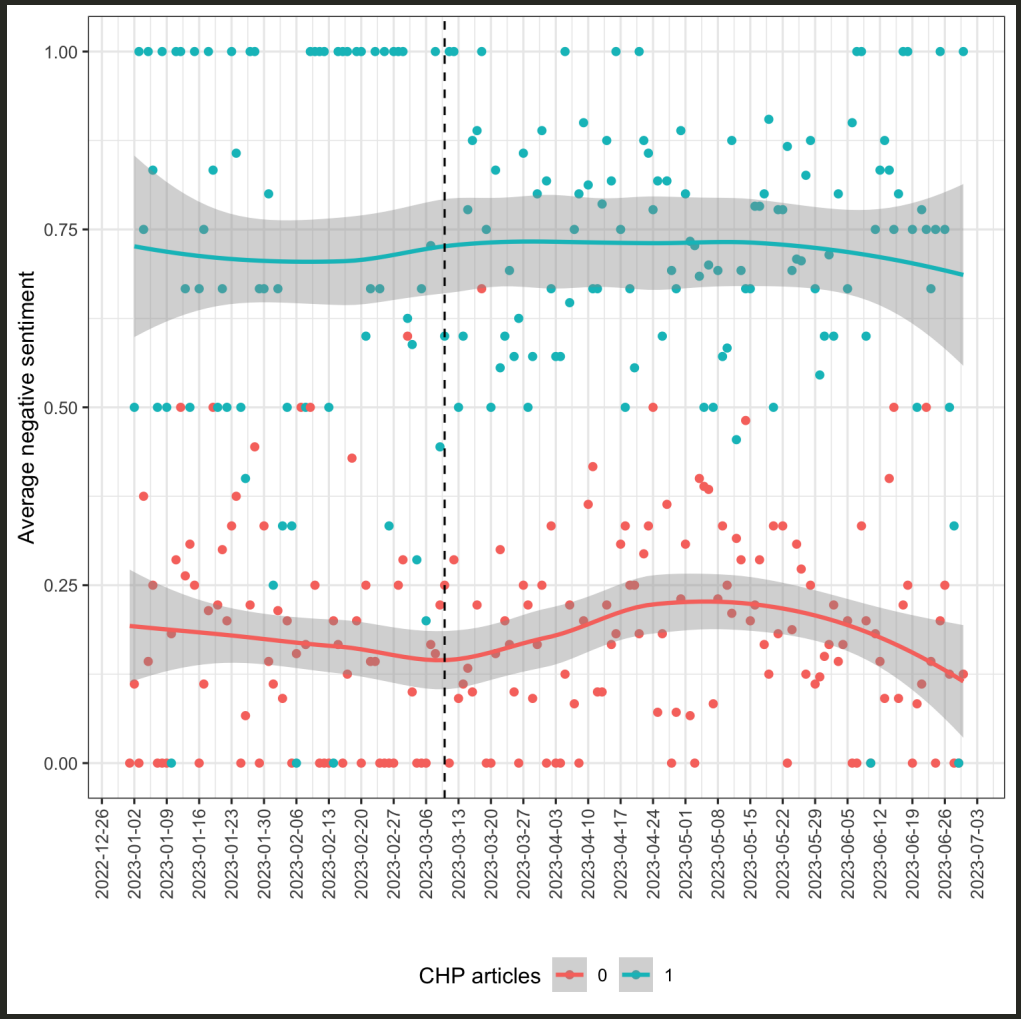
$$Y_{it} = \alpha_i + \delta_t + \beta D_{it} + \epsilon_{it}$$

```
#m2 <- lm(average_negative ~ treatment + as.factor(chp) +
# as.factor(date_publish), data = day_level)

# here we have two FEs: day and unit (CHP vs AKP)
m2 <- feols(negative ~ treated |
          date_publish + chp, data = d)


summary(m2)
```

```
## OLS estimation, Dep. Var.: negative
## Observations: 3,104
## Fixed-effects: date_publish: 181,  chp: 2
## Standard-errors: Clustered (date_publish)
##         Estimate Std. Error  t value Pr(>|t|)
## treated 0.005104   0.039074 0.130618  0.89622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.404939     Adj. R2: 0.284517
##                  Within R2: 5.989e-6
```

```r
# checking parallel trends visually
library(ggplot2)

day_level <- d %>% group_by(date_publish, chp) %>% summarise(average_

plot_sentiment <- ggplot(data = day_level, aes(x = date_publish,
            y = average_negative, color = factor(chp))) +
  geom_point() + theme_bw() + geom_smooth() +
  geom_vline(xintercept = as.Date('2023-03-10'),
             linetype = 'dashed', color = 'black')+
  labs(y ='Average negative sentiment', x = '',
       color = 'CHP articles') +
  theme(legend.position = 'bottom',
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
  scale_x_date(date_breaks  ="1 week")
```

```
plot_sentiment
```



```r
# checking parallel trends with the event-study approach

# we can have weekly estimates with week dummies by
# taking the 9th week (the week of March 10) as the baseline.
# feols makes event study easier.

event_study <- feols(negative ~ i(week, ref = 9),
                     data = d)
```

```
iplot(event_study)
```