Difference-in-Differences

Ian McCarthy | Emory University
Workshop on Causal Inference with Panel Data

Table of contents

The Idea of DD

Setup

Want to estimate $E[Y_1(1) - Y_0(1)|W = 1]$

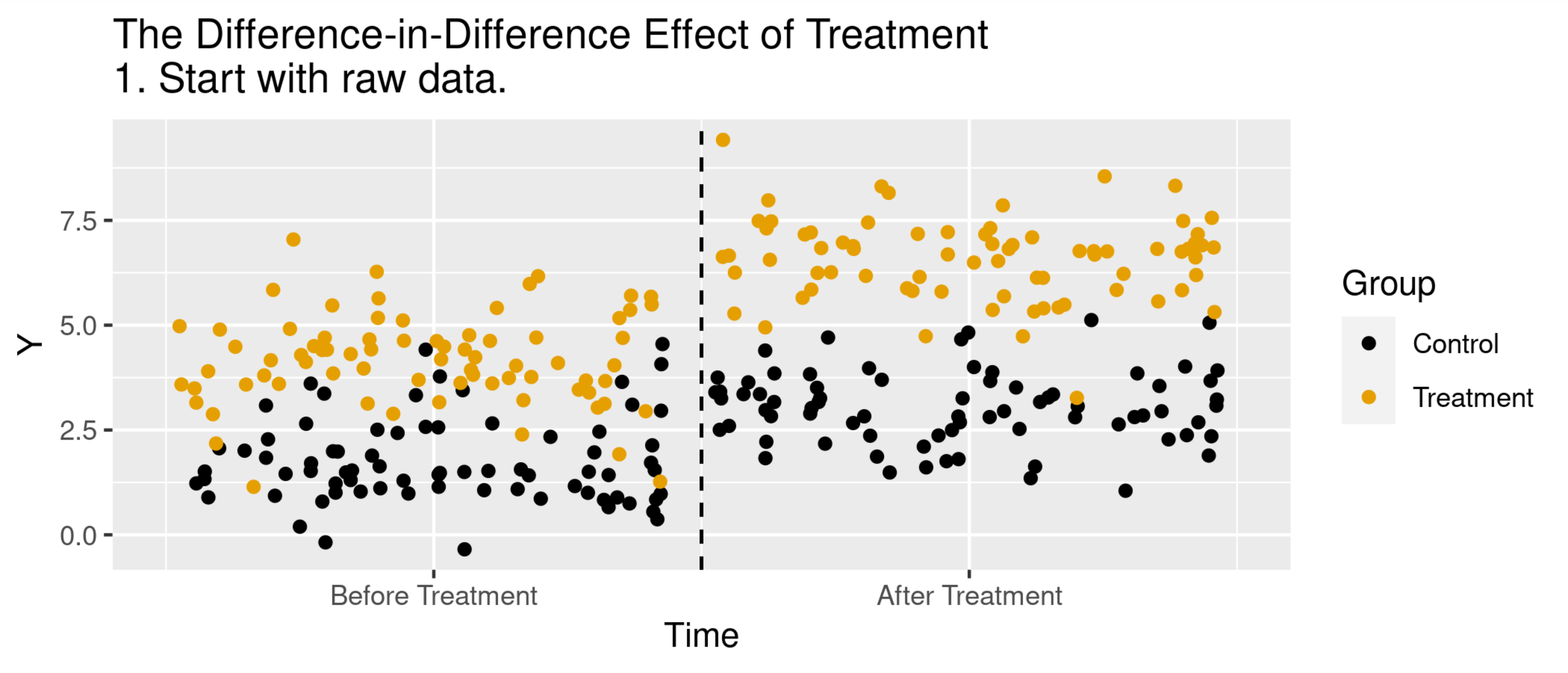
	Post-period	Pre-period
Treated	$E(Y_1(1)\|W = 1)$	$E(Y_0(0)\|W = 1)$
Control	$E(Y_0(1)\|W = 0)$	$E(Y_0(0)\|W = 0)$

Strategy 3: DD estimate...

Estimate $E[Y_1(1)|W = 1] - E[Y_0(1)|W = 1]$ using $E[Y_0(1)|W = 0] - E[Y_0(0)|W = 0]$ (pre-post difference in control group used to predict difference for treatment group)

Graphically

Animations!



The Difference-in-Difference Effect of Treatment
1. Start with raw data.
Y
7.5
5.0
2.5
0.0
Before Treatment
After Treatment
Time
Group
Control
Treatment

Average Treatment Effects with DD

Estimation

Key identifying assumption is that of *parallel trends*

$$E[Y_0(1) - Y_0(0)|W = 1] = E[Y_0(1) - Y_0(0)|W = 0]$$

Estimation

Sample means:

$$
\begin{aligned}
E[Y_1(1) - Y_0(1)|W = 1] = \quad & (E[Y(1)|W = 1] - E[Y(1)|W = 0]) \\
& - (E[Y(0)|W = 1] - E[Y(0)|W = 0])
\end{aligned}
$$

Estimation

Regression:

$$Y_i = \alpha + \beta D_i + \lambda 1(Post) + \delta D_i \times 1(Post) + \varepsilon$$

	After	Before	After - Before
Treated	$\alpha + \beta + \lambda + \delta$	$\alpha + \beta$	$\lambda + \delta$
Control	$\alpha + \lambda$	α	λ
Treated - Control	$\beta + \delta$	β	δ

Simulated data

```r
N <- 5000
dd.dat <- tibble(
  w = (runif(N, 0, 1)>0.5),
  time_pre = "pre",
  time_post = "post"
)

dd.dat <- pivot_longer(dd.dat, c("time_pre","time_post"), values_to="time") %>%
  select(w, time) %>%
  mutate(t=(time=="post"),
         y.out=1.5+3*w + 1.5*t + 6*w*t + rnorm(N*2,0,1))
```

Mean differences

```
dd.means <- dd.dat %>% group_by(w, t) %>% summarize(mean_y = mean(y.out))
knitr::kable(dd.means, col.names=c("Treated","Post","Mean"), format="html")
```

Treated	Post	Mean
FALSE	FALSE	1.522635
FALSE	TRUE	3.002374
TRUE	FALSE	4.515027
TRUE	TRUE	12.004623

Mean differences

In this example:

- $E[Y(1)|W=1] - E[Y(1)|W=0]$ is 9.0022495
- $E[Y(0)|W=1] - E[Y(0)|W=0]$ is 2.9923925

So the ATT is 6.0098571

Regression estimator

```r
dd.est <- lm(y.out ~ w + t + w*t, data=dd.dat)
summary(dd.est)
```

```
##
## Call:
## lm(formula = y.out ~ w + t + w * t, data = dd.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0038 -0.6674  0.0047  0.6609  3.6135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.52263    0.01970   77.28   <2e-16 ***
## wTRUE        2.99239    0.02795  107.07   <2e-16 ***
## tTRUE        1.47974    0.02786   53.10   <2e-16 ***
## wTRUE:tTRUE  6.00986    0.03953  152.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seeing things in action

Application

- Try out some real data on Medicaid expansion following the ACA
- Data available on GitHub (see code files for links)

Step 1: Look at the data

Stata

```
insheet using "https://raw.githubusercontent.com/imccart
gen perc_unins=uninsured/adult_pop
keep if expand_year=="2014" | expand_year=="NA"
drop if expand_ever=="NA"
collapse (mean) perc_unins, by(year expand_ever)
graph twoway (connected perc_unins year if expand_ever==
  (connected perc_unins year if expand_ever=="TRUE", col
  xline(2013.5) ///
    ytitle("Fraction Uninsured") xtitle("Year") legend(o
```

R

```
library(tidyverse)
mcaid.data <- read_tsv("https://raw.githubusercontent.co
ins.plot.dat <- mcaid.data %>% filter(expand_year==2014
  mutate(perc_unins=uninsured/adult_pop) %>%
  group_by(expand_ever, year) %>% summarize(mean=mean(pe

ins.plot <- ggplot(data=ins.plot.dat, aes(x=year,y=mean,
  geom_line() + geom_point() + theme_bw() +
  geom_vline(xintercept=2013.5, color="red") +
  geom_text(data = ins.plot.dat %>% filter(year == 2016)
            aes(label = c("Non-expansion","Expansion"),
                x = year + 1,
                y = mean)) +
  guides(linetype=FALSE) +
  labs(
```

Step 2: Estimate Effects

Interested in δ from:

$$y_{it} = \alpha + \beta \times 1(Post) + \lambda \times 1(Expand) + \delta \times 1(Post) \times 1(Expand) + \varepsilon$$

Stata

```
insheet using "https://raw.githubusercontent.com/imccart
gen perc_unins=uninsured/adult_pop
keep if expand_year=="2014" | expand_year=="NA"
drop if expand_ever=="NA"
gen post=(year>=2014)
gen treat=(expand_ever=="TRUE")
gen treat_post=(expand=="TRUE")

reg perc_unins treat post treat_post

*also try didregress
```

R

```
library(tidyverse)
library(modelsummary)
mcaid.data <- read_tsv("https://raw.githubusercontent.co
reg.dat <- mcaid.data %>% filter(expand_year==2014 | is.
  mutate(perc_unins=uninsured/adult_pop,
         post = (year>=2014),
         treat=post*expand_ever)

dd.ins.reg <- lm(perc_unins ~ post + expand_ever + post*
msummary(dd.ins.reg)
```

Final thoughts

- Key identification assumption is **parallel trends**
- We've ignored any issues with inference
- Typically want to cluster at unit-level to allow for correlation over time within units
- "Extra" things like propensity score weighting and doubly robust estimation