Lecture 8. DiD in Practice

Vinish Shrestha

Towson University

Canonical DiD in practice

Naive estimator

Canonical Difference in Differences Framework

Parallel trend assumption

Canonical DiD using regression

Two-way fixed effects estimator (TWFE)

Canonical DiD in practice

Evaluating the impact of Medicaid Expansion

- We are interested in evaluating the impacts of Medicaid expansion on insurance outcomes

Some context

- Part of the Patient Protection and Affordable Care Act (ACA, Obama Care)
- Original bill had Medicaid expansion: 138% of the FPL
- Supreme court decision in 2012 deemed this as unconstitutional
- Bill was reformated; expansion was then voluntary

Expansion status

- 26 states expanded Medicaid in 2014
- 19 states did not expand until 2018
- Data comes from my projects
- Uninsured data from Small Area Health Insurance Estimates (SAHIE)

Load the data (those that expanded later than 2014)

```r
# load data
library(pacman)
p_load(fixest, dplyr, ggplot2, tidyverse, patchwork, arrow, janitor)



# load in county level uninsured rate data merged with other variables
mort_allcauses <-  read_feather( file.path(datapath, "NVSS_data_county_201
                     mutate(treat = ifelse(is.na(treat) == T, "control 3", 
                     filter(yearexpand == 2014 & age == 0 & race_name == "wh
                     dplyr::select("countyfips", "year", "state.abb", "expan
                     filter(duplicated(.))   %>%
                     arrange(countyfips, year) # sort by countyfips and yea
                      # the select() function is masked
                     # by other packages, so use dplyr::select() instead
```

Head the data

```
# only keep the years 2013 and 2014 for the canonical case
dat_canonical  <- mort_allcauses  %>%
                          filter(year %in% c(2013, 2014))
head(dat_canonical)
```

```
## # A tibble: 6 x 7
##   countyfips  year state.abb expand yearexpand sahieunins138
##        <dbl> <dbl> <chr>      <dbl>      <dbl>         <dbl>
## 1       1001  2013 AL             0       2014          39.6
## 2       1001  2014 AL             0       2014          31.9
## 3       1003  2013 AL             0       2014          45.1
## 4       1003  2014 AL             0       2014          43.8
## 5       1005  2013 AL             0       2014          37.3
## 6       1005  2014 AL             0       2014          34
```

Expansion and non-expansion states

```
## The expansion states are:

##
##  AR  AZ  CA  CO  CT  DE  IA  IL  KY  MA  MD  MI  MN  ND  NH  NJ  NM  NV
## 144  30 112 106   8   6 196 196 240  28  48 164 168  79  20  39  50  26
##  OR  RI  VT  WA  WV
##  66  10  26  72 110

## The non-expansion states are:

##
##  AL  FL  GA  ID  KS  ME  MO  MS  NC  NE  OK  SC  SD  TN  TX  UT  VA  WI
## 126 134 291  84 181  32 230 149 198 144 148  88  98 190 409  54 248 136
```

Count of expansion vs non-expansion states

```
length(table(dat_canonical$state.abb[dat_canonical$expand == 1]))
```

```
## [1] 25
```

```
length(table(dat_canonical$state.abb[dat_canonical$expand == 0]))
```

```
## [1] 19
```

Expansion and non-expansion states

The two groups are as follows:

i) **Expansion states (treated)**: AR, AZ, CA, CO, CT, DE, HI, IA, IL, KY, MA, MD, MI, MN, ND, NH, NJ, NM, NV, NY, OH, OR, RI, VT, WA, WV

ii) **Non-expansion states (control)**: AL, FL, GA, ID, KS, ME, MO, MS, NC, NE, OK, SC, SD, TN, TX, UT, VA, WI, WY

Naive estimator

Naive estimator

- A naive estimate of ATT: difference in means between the treated and control groups in the period following the expansion.

```
naive  <- mean(dat_canonical$sahieunins138[dat_canonical$expand == 1 & dat_
          mean(dat_canonical$sahieunins138[dat_canonical$expand == 0 & dat_

print(naive)
```

```
## [1] -13.42305
```

Can we trust the naive estimator?

- naive estimate: uninsured rate dropped by -13.66 percentage points following the Medicaid expansion in 2014.
- But can we trust this estimate? Not really!
- Note that the estimation approach here is similar to the RTC
- But is the treatment assignment random?

Reasons why naive estimator fails

i) One way to assess the validity of naive estimate is to compare the (natural) experiment on hand with the randomized control case.

- Note that we are very far away from the randomized controlled trial in this case.
- The treatment (decision to expand Medicaid) is not random.
- Note that states voluntarily decided to expand Medicaid.
- For example, many of the southern states did not expand Medicaid.
- Also, pre-treatment uninsured rates of southern states are generally higher compared to non-southern states.

The naive comparison can simply be capturing the difference in pre-treatment characteristics correlated with the treatment assignment.

ii) The baseline characteristics among the expansion vs non-expansion states differs dramatically.

- For example, southern states have higher population of Blacks compared to non-South.

Evaluating differences in uninsured rate in 2013 (pre-treatment year)

```
naive_pre  <- mean(dat_canonical$sahieunins138[dat_canonical$expand == 1 &
                   dat_canonical$year < 2014]) -
              mean(dat_canonical$sahieunins138[dat_canonical$expand == 0 &
                   dat_canonical$year < 2014])

print(naive_pre)
```

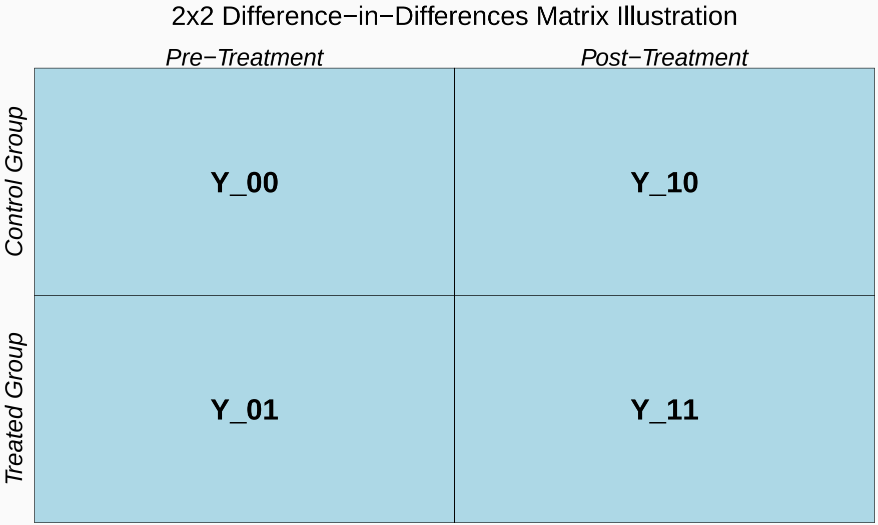
```
## [1] -7.669252
```

Canonical Difference in Differences Framework

Canonical DiD

- We would like to alleviate the aforementioned concerns.
- One way to address the second concern, i.e., outcomes in pre-treatment period may differ significantly between the treatment and control groups, is to take out the mean difference in outcome during the pre-treatment period from the mean difference in outcome post treatment. This approach uses two groups and two periods, which is termed as the canonical DiD case.

Canonical DiD 2×2 matrix.



Canonical DiD estimate (unconditional)

In the ACA-Medicaid expansion example that involves two groups and two time periods:

```
cat("did estimate: \n", naive - naive_pre)
```

```
## did estimate:
##  -5.7538
```

Canonical DiD estimate

- suggests that uninsured rate dropped by 5.81 percentage points following the Medicaid expansion in year 2014.
- let's formally visit the DiD approach to appreciate some necessary assumptions while connecting it with ATT.

Parallel trend assumption

Parallel trend assumption

$$E(Y^0(1) - Y^0(0)|D = 1) = E(Y^0(1) - Y^0(0)|D = 0)(\#eq : ptrend1) \qquad (1)$$

-*In absence of the expansion, trends in uninsured rate would evolve parallelely across the expansion and non-expansion states*

The role of parallel trend in indentification of ATT

$$\delta = E(Y^1(1)|D=1) - E(Y^0(1)|D=1) \tag{2}$$

$$
\begin{aligned}
&= E(Y^1(1)|D=1) - E(Y^0(1)|D=1) + E(Y^0(0)|D=1) - E(Y^0(0)|D=1) \\
&= \{E(Y^1(1)|D=1) - E(Y^0(0)|D=1)\} - \{E(Y^0(1)|D=1) - E(Y^0(0)|D=1)\} \\
&= \{E(Y^1(1)|D=1) - E(Y^0(0)|D=1)\} - \{E(Y^0(1)|D=0) - E(Y^0(0)|D=0)\} \\
&= \{E(Y(1)|D=1) - E(Y(0)|D=1)\} - \{E(Y(1)|D=0) - E(Y(0)|D=0)\}
\end{aligned}
$$

Canonical DiD using regression

The 2×2 Difference-in-Differences Estimate

- Let's begin with the canonical DiD framework using the regression format.
- I'm going to set it up as the following:

$$Y_{it} = \alpha + \tau Post_{it} \times D_i + \sigma Post_{it} + \eta D_i + \epsilon_{it} \tag{3}$$

- Let's rewrite the DiD estimator from before as:

$$\tau_{did} = \underbrace{E[Y_{11} - Y_{10}|D = 1]}_{first\ difference} - \underbrace{E[Y_{01} - Y_{00}|D = 0]}_{second\ difference}$$

What's the specification about?

- Let's look at the following conditional expectations.

1). expected outcome for treated group post treatment:
$E(Y|D = 1, Post = 1) = \alpha + \tau + \sigma + \eta$

2). expected outcome for treated group pre treatment:
$E(Y|D = 1, Post = 0) = \alpha + \eta$

3). expected outcome for control group post treatment:
$E(Y|D = 0, Post = 1) = \alpha + \sigma$

4). expected outcome for control group pre treatment: $E(Y|D = 0, Post = 0) = \alpha$

Let's apply this to our ACA-Medicaid example.

```r
# lets create the post, treat, and the interaction between the post
#  and treat (labeled as did)
dat_canonical  <- dat_canonical  %>%
                          mutate(post = ifelse(year >= 2014, 1, 0),
                                 treat = ifelse(expand == 1, 1, 0),
                                 did = post * treat)

reg_did  <- lm(sahieunins138 ~ did + post + treat, data = dat_canonical)
```

Output (extract coefficient)

```
cat("DiD estimate from regression:", "\n",
            coefficients(summary(reg_did))[2])
```

```
## DiD estimate from regression:
##  -5.7538
```

The whole of output

```r
summary(reg_did)
```

```
##
## Call:
## lm(formula = sahieunins138 ~ did + post + treat, data = dat_canonical)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.0767  -5.0184  -0.7184   4.7586  27.0540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.1460     0.1931  213.08   <2e-16 ***
## did          -5.7538     0.4172  -13.79   <2e-16 ***
```

Compare it with difference in means estimator

```r
did  <- naive - naive_pre
print(did)
```

```
## [1] -5.7538
```

Two-way fixed effects estimator (TWFE)

Two way fixed effect (TWFE) Revisited

- We have already seen the TWFE and its importance in accounting for unobserved heterogeneity.
- The TWFE is linked to the difference-in-differences setting (perhaps mistakenly).
- However, note that the TWFE estimator is not equal to the DiD estimator unless the treatment effects are homogeneous across both units and time.

TWFE estimator

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + v_{it} \; .....TWFE \tag{4}$$

- Here, Y_{it} is the outcome of individual i in period t ($t \in \{1,\ 2,\ ...,\ T\}$)
- θ_t is the time fixed effects; η_i is the unit fixed effect
- D_{it} captures whether individual i is treated in time t
- Equation above is the TWFE.

TWFE: Using Within Estimator

Data Arrange 1: Demeaning to get rid of η_i from TWFE equation (Within Estimator)

- Let's look at the concept behind the within estimator.
- In the two-period two-group case, TWFE can be written as:

$$
\begin{aligned}
Y_{i1} &= \theta_1 + \eta_i + \alpha D_{i1} + v_{i1} \\
Y_{i2} &= \theta_2 + \eta_i + \alpha D_{i2} + v_{i2}
\end{aligned}
\tag{5}
$$

- where, i is represented by 1 (treatment group) and 0 (untreated group).

Within estimator

- Adding the sub-equations and dividing by the number of time period ($T = 2$) yields:

$$\frac{Y_{i1} + Y_{i2}}{2} = \frac{\theta_1 + \theta_2}{2} + \frac{2\eta_i}{2} + \frac{\alpha(D_{i1} + D_{i2})}{2} + \frac{v_{i1} + v_{i2}}{2} \tag{6}$$

$$Y_i = \frac{\theta_1 + \theta_2}{2} + \eta_i + \alpha D_i + v_i$$

Within estimator

- Substracting the above equation from the TWFE yields the following:

$$Y_{it} - Y_i = \theta_t - \frac{\theta_1 + \theta_2}{2} + \alpha(D_{it} - D_i) + (v_{it} - v_i) \tag{7}$$

The code shows data arranging for the within estimator.

Data arranging for Within-estimator

```
###########################
# Treatment group
###########################
treat_t <- rep(1, 1000)
period_t <- rep(c(0, 1), each = 500)
id <- rep(seq(1, 500, 1), 2) #for the panel nature of data
y_treat <- 20 * period_t + 7  + rnorm(1000, 0, 5)
treatdata <- data.frame(treat = treat_t, period = period_t, Y = y_treat, i
treatdata <- treatdata %>% mutate(Ytrans = Y - mean(Y),
                                  D = treat * period - mean(treat * period
```

Data arranging for Within-estimator (control)

```
###########################
# control group
###########################
control_t <- rep(0, 1000)
period_c <- rep(c(0, 1), each = 500)
id <- rep(seq(501, 1000, 1), 2)
y_control <- 3 +  rnorm(1000, 0, 5)
controldata = data.frame(treat = control_t, period = period_c, Y = y_contr
controldata <- controldata %>% mutate(Ytrans = Y - mean(Y),
                                       D = treat * period - mean(treat * period

data = rbind(treatdata, controldata)
```

TWFE: As the first difference

Data Arrange 2: First differencing

- Let's briefly look at the concept behind first differencing.

Write TWFE as:

$$\begin{aligned} Y_{i1} &= \theta_1 + \eta_i + \alpha D_{i1} + v_{i1} \\ Y_{i2} &= \theta_2 + \eta_i + \alpha D_{i2} + v_{i2} \end{aligned} \tag{8}$$

for $i \in \{0,\ 1\}$.

TWFE: As the first difference

- Then,

$$Y_{i2} - Y_{i1} = \theta_2 - \theta_1 + \alpha(D_{i2} - D_{i1}) + (v_{i2} - v_{i1}) \tag{9}$$

Data arranging for First diff.

```
# First the treated group
fd_treat1 <- treatdata %>% filter(period == 0) %>% dplyr::select(-c("Ytrans
colnames(fd_treat1) <- c("treat1", "period1", "Y1", "id")
fd_treat2 <- treatdata %>% filter(period == 1)%>% dplyr::select(-c("Ytrans
colnames(fd_treat2) <- c("treat2", "period2", "Y2", "id")
fd_treat <- merge(fd_treat1, fd_treat2, by = "id", all.x = T)
fd_treat <- fd_treat %>% mutate(Y_FD = Y2 - Y1,
                                D = (period2 * treat2) - (period1 * treat1
```

```r
# Then the control group
fd_control1 <- controldata %>% filter(period == 0) %>% dplyr::select(-c("Y
colnames(fd_control1) <- c("treat1", "period1", "Y1", "id")
fd_control2 <- controldata %>% filter(period == 1)%>% dplyr::select(-c("Yt
colnames(fd_control2) <- c("treat2", "period2", "Y2", "id")
fd_control <- merge(fd_control1, fd_control2, by = "id", all.x = T)
fd_control <- fd_control %>% mutate(Y_FD = Y2 - Y1,
                                    D = (period2 * treat2) - (period1 * treat1

FDdata = rbind(fd_treat, fd_control)
```

Various ways of estimating

```
##
## Call:
## lm(formula = Y ~ treat:period + treat + period, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.8122  -3.3101   0.1082   3.5773  15.8086
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.1006     0.2277  13.618   <2e-16 ***
## treat          3.7076     0.3220  11.514   <2e-16 ***
## period        -0.2390     0.3220  -0.742    0.458
## treat:period  20.4792     0.4554  44.972   <2e-16 ***
```

```
##
## Call:
## lm(formula = Ytrans ~ D + period, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.8122  -3.3101   0.1082   3.5773  15.8086
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1195     0.1971   0.606    0.544
## D            20.4792     0.4553  44.983   <2e-16 ***
## period       -0.2390     0.3219  -0.742    0.458
## ---
```

3. First Difference

```
##
## Call:
## lm(formula = Y_FD ~ D, data = FDdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.498  -4.740  -0.247   4.443  26.045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2390     0.3203  -0.746    0.456
## D            20.4792     0.4530  45.209   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion

Questions

- Does the parallel trend hold?
- We don't know for sure.
- Can we provide some suggestive evidence?

References