Lab 9: Difference in Difference Estimators

Sidak Yntiso sgy210@nyu.edu

April 06, 2020

Identifying assumptions

$$Y_{it} = \pi D_{it} + X_{it}'\beta_i + \alpha_i + \delta_t + \epsilon_{it}$$

- Full rank regression matrix (variation in X)
- Zero conditional mean of the errors: $\mathbb{E}[\epsilon_{it}|D_{it}, X_{it}, \alpha_i, \delta_t] = 0$ for $t = 1, .., T$

Identifying assumptions

$$Y_{it} = \pi D_{it} + X_{it}'\beta_i + \alpha_i + \delta_t + \epsilon_{it}$$

- Full rank regression matrix (variation in X)
- Zero conditional mean of the errors: $\mathbb{E}[\epsilon_{it}|D_{it}, X_{it}, \alpha_i, \delta_t] = 0$ for $t = 1, .., T$
- Conditional independence of errors: $Cov(\epsilon_{it}, \epsilon_{jt}|D_{it}, X_{it}, \alpha_i, \delta_t) = 0$
- Homoskedasticity of the errors: $Var(\epsilon_{it}|D_{it}, X_{it}, \alpha_i, \delta_t) = \sigma^2$
  - or cluster robust standard errors or block bootstrap

Empirical illustrations

Paper 1

- Multiple time periods; same treatment initiation period
- Visualizing parallel trends

Empirical illustrations

Paper 1

- Multiple time periods; same treatment initiation period
- Visualizing parallel trends

Paper 2

- Multiple time periods; different treatment initiation periods
- Effect hetereogeneity; measurement error

Paper 1

Effect of Medicaid expansion (Sommers et al 2012 (NEJM))

- What is the effect of expanded adult Medicaid eligibility
- Expansion states (New York, Maine, and Arizona) passed new laws in 2000/01
- Comparison group is neighboring states without expansions.
- Outcome is disease-related county-level mortality from the CDC

DiD Design

- first difference: expansion states and nonexpansion states
- second difference: before and after reform

First difference

```r
rm(list=ls())
library(estimatr)
medicaid_study = haven::read_dta("medicaid_study.dta")

#what's the mean death rate for expansion
#states in post-reform period (after 2001)
lm2005 <- lm_robust(cruderate~MedicaidExpansion,
                    subset(medicaid_study,year==2005))
summary(lm2005)$coefficients[,c(1:3)]
```

```
##                     Estimate Std. Error   t value
## (Intercept)         296.6068   7.412903 40.012231
## MedicaidExpansion    -9.8283   9.793611 -1.003542
```

Difference-in-means estimates by year

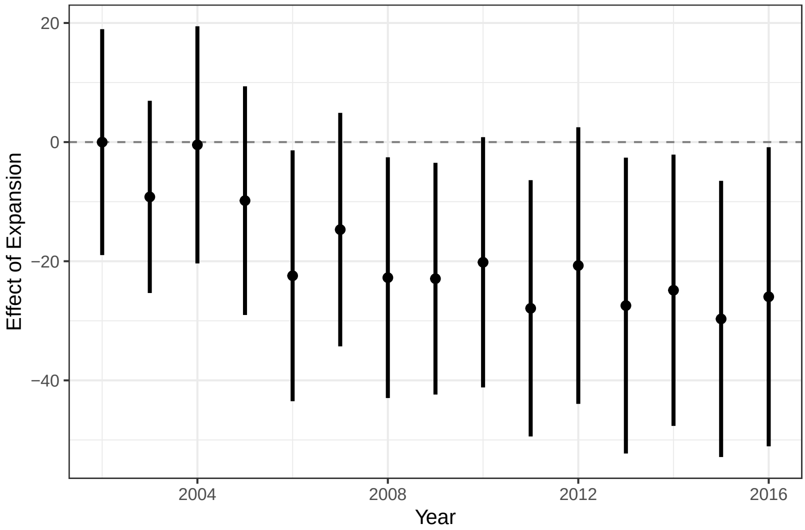
```r
dim_estimates <- c()
se_estimates <- c()
#what's the mean death rate by treatment
for (j in c(2002:2016)){
  lmj <- lm_robust(cruderate~MedicaidExpansion,
                   subset(medicaid_study,year==j))
  dim_estimates <- c(dim_estimates,
                     summary(lmj)$coefficients[2,1])
  se_estimates <- c(se_estimates,
                    summary(lmj)$coefficients[2,2])
}
#store results for years 2002-2016
dat1 <- data.frame(year=c(2002:2016),
                   dim_estimates = dim_estimates,
                   se_estimates = se_estimates)
```

Difference-in-means estimates by year plot

```r
interval2 <- -qnorm((1-0.95)/2)  # 95% multiplier
#plot the effects by year
library(ggplot2)
p <- ggplot(aes(x=year,y=dim_estimates),data=dat1)+
  geom_hline(yintercept = 0, colour = gray(1/2), lty = 2)+
  geom_point(aes(x = year, y = dim_estimates),lwd = 2)+
  geom_linerange(aes(x = year,
                     ymin = dim_estimates -
                       se_estimates*interval2,
                     ymax = dim_estimates +
                       se_estimates*interval2),
                 lwd = 1)+xlab("Year")+theme_bw()+
  ylab("Effect of Expansion")
```

Visualize

```
p
```



Second difference

```r
#identify treatment states
medicaid_study$trt_state = 0
medicaid_study$trt_state[medicaid_study$state%in%
    c("New York","Maine","Arizona")] = 1

#what's the mean cruderate in post-reform period
lm2001 <- lm_robust(cruderate~I(year>=2001),medicaid_study)
summary(lm2001)$coefficients[,c(1:3)]
```

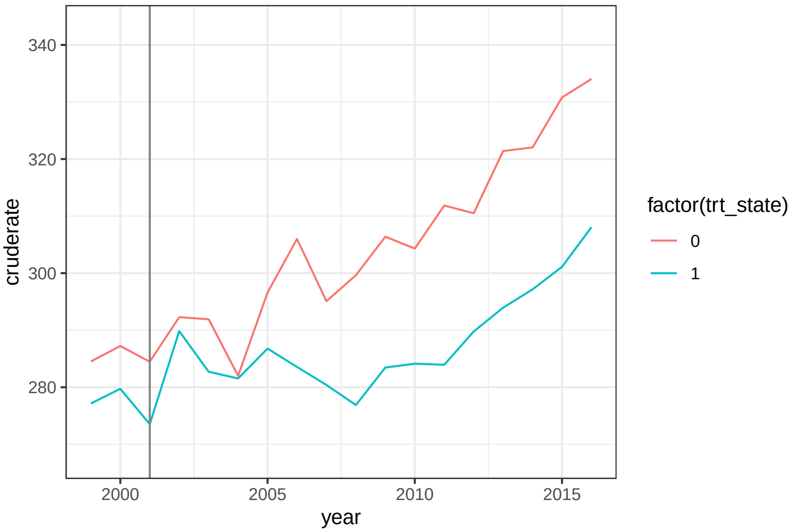
```
##                        Estimate Std. Error   t value
## (Intercept)           282.38636   3.385347 83.414316
## I(year >= 2001)TRUE    14.59126   3.650046  3.997554
```

Second difference plot

```r
#plot trends in death rate by treatment group
p2 <- ggplot(medicaid_study, aes(year, cruderate,
                color = factor(trt_state)) )  +
  stat_summary(aes(group = factor(trt_state)),
               geom="line") +
  geom_vline(xintercept = 2001, colour = gray(1/2))+
  theme_bw()
```

Visualizing Parallel Trends

```
p2
```



Difference in differences

```r
#two way fixed effects estimate
m1 <- lm_robust(cruderate ~ MedicaidExpansion+
                  as.factor(year)+as.factor(countycode),
                clusters=countycode,data = medicaid_study)
summary(m1)$coefficients[c(1:2),c(1:3)]
```

```
##                     Estimate Std. Error   t value
## (Intercept)       341.909742   3.437644 99.460498
## MedicaidExpansion  -7.921419   5.020432 -1.577836
```

Testing Parallel Trends Setup

```r
#generate placebo treatments
for (j in c(1999:2001,2003:2016)){
  assign(paste("treat",j,sep=""),
         medicaid_study$trt_state*
           I(medicaid_study$year==j))
}
#DiD model with placebo treatments
m2 <- formula(paste("cruderate~",
  paste("treat",c(1999:2001,2003:2016),
          sep="",collapse="+" ),
  "+as.factor(year)+as.factor(countycode)",sep=""))
#running the model
m2 <- lm_robust(m2,data = medicaid_study,
                  clusters=countycode)
```

Placebo DiD estimates data

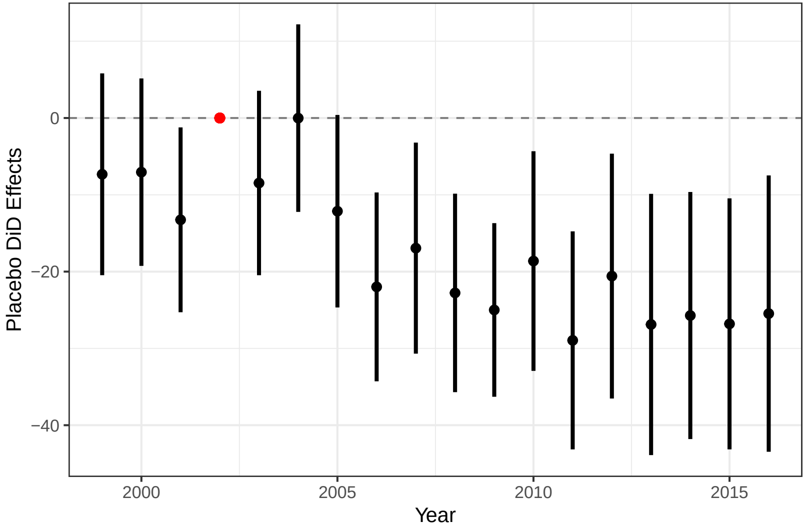
```
#storing the placebo DiD estimates
dim_estimates <- c(); se_estimates <- c()
for (j in c(2:18) ){
  dim_estimates <- c(dim_estimates,
                     summary(m2)$coefficients[j,1])
  se_estimates <- c(se_estimates,
                    summary(m2)$coefficients[j,2])
}
#saving to a dataset
dat2 <- data.frame(year=c(1999:2001,2003:2016),
                   dim_estimates = dim_estimates,
                   se_estimates = se_estimates)
```

Placebo DiD estimates plot

```r
#plotting the placebo DiD estimates by year
p3 <- ggplot(aes(x=year,y=dim_estimates),data=dat2)+
  geom_hline(yintercept = 0, colour = gray(1/2), lty = 2)+
  geom_point(aes(x = 2002, y = 0),lwd = 2,colour="red")+
  geom_point(aes(x = year, y = dim_estimates),lwd = 2)+
  geom_linerange(aes(x = year,
                     ymin = dim_estimates -
                       se_estimates*interval2,
                     ymax = dim_estimates +
                       se_estimates*interval2),
                 lwd = 1)+xlab("Year")+theme_bw()+
  ylab("Placebo DiD Effects")
```

Visualize

```
p3
```



Paper 2

Voter Identification laws: require government ID to vote (Hajnal et al (2017) (HLN) and Grimmer et al 2018)

- Strict ID states are Arizona (2004), Georgia (2005), Indiana (2005), Kansas (2011), Mississippi (2011), North Dakota (2013), Ohio (2006), Tennessee (2011), Texas (2011), Virginia (2012), and Wisconsin (2016)
- Minority voters: much less likely to hold IDs (Ansolabehere and Hersh 2016)
- What is effect of ID laws on turnout? by race?

Paper 2

Voter Identification laws: require government ID to vote (Hajnal et al (2017) (HLN) and Grimmer et al 2018)

- Strict ID states are Arizona (2004), Georgia (2005), Indiana (2005), Kansas (2011), Mississippi (2011), North Dakota (2013), Ohio (2006), Tennessee (2011), Texas (2011), Virginia (2012), and Wisconsin (2016)
- Minority voters: much less likely to hold IDs (Ansolabehere and Hersh 2016)
- What is effect of ID laws on turnout? by race?

DiD Design

- Cooperative Congressional Election Study (2006-2014)
- Dependent Variable: General/Primary Election Turnout
- Treatment: Strict Voter ID Law in state
- first difference: states with strict voter rules versus others
- second difference: year before and after reform

Data

```r
##Loading the data from HLN
dd<- read.delim('HKL_V2_data.tab', sep='\t')
#list of covariates to condition on
covs <- c("foreignb","firstgen","age","educ","inc",
         "male","married","childrenz","unionz",
         "unemp","ownhome","protestant","catholic",
         "jewish","atheist","days_before_election",
         "early_in_person","vote_by_mail",
         "no_excuse_absence_",
         "presidentialelectionyear",
         "gubernatorialelectionyear",
         "senateelectionyear",
         "marginpnew","newstrict")
```

Model

```r
#model
m1 <- formula(paste("votegenval~stricty+black+hispanic +
              asian + mixedrace + blackstricty +
              hispstricty + asianstricty +
              mixedracestricty + as.factor(year) +
              as.factor(state)+",paste(covs,collapse="+")))
###Replication Column 1, Table A9 HLN (no clustering)
model1<- lm_robust(m1,data = subset(dd,voteregpre==1))
```

Estimates

```
#store main and interaction results
results <- data.frame(
  race = c("White","Black","Hispanic","Asian","Mixed"))
 results$m1_ses <- results$m1_est <-NA
results$m1_est[1] =
  summary(model1)$coefficients['stricty',1]
results$m1_est[2] = results$m1_est[1] +
  summary(model1)$coefficients['blackstricty',1]
results$m1_est[3] = results$m1_est[1] +
  summary(model1)$coefficients['hispstricty',1]
results$m1_est[4] = results$m1_est[1] +
  summary(model1)$coefficients['asianstricty',1]
results$m1_est[5] = results$m1_est[1] +
  summary(model1)$coefficients['mixedracestricty',1]
```

Standard errors

```
#variance covariance matrix of coefficients
v_model<- vcov(model1)
#variance of main+interaction effects
#using Var(X+Y) = Var(X)+ Var(Y)+ 2*Cov(XY)
results$m1_ses[1]<- sqrt(v_model['strictly', 'strictly'])
results$m1_ses[2]<- sqrt(v_model['strictly', 'strictly']
                  +v_model['blackstrictly', 'blackstrictly']
                  +2*v_model['strictly', 'blackstrictly'])
results$m1_ses[3]<- sqrt(v_model['strictly', 'strictly']
                  +v_model['hispstrictly', 'hispstrictly']
                  +2*v_model['strictly', 'hispstrictly'])
results$m1_ses[4]<- sqrt(v_model['strictly', 'strictly']
                  +v_model['asianstrictly', 'asianstrictly']
                  +2*v_model['strictly', 'asianstrictly'])
results$m1_ses[5]<- sqrt(v_model['strictly', 'strictly']
                  +v_model['mixedracestrictly','mixedracest
                  +2*v_model['strictly', 'mixedracestrictly']
```

Summarize results

```
#t statistic
results$tval <- results$m1_est/results$m1_ses
results
```

```
##       race     m1_est      m1_ses      tval
## 1    White 0.10925251 0.008182457 13.352042
## 2    Black 0.10428520 0.011294705  9.233106
## 3 Hispanic 0.06466329 0.016683657  3.875846
## 4    Asian 0.12533809 0.039901759  3.141167
## 5    Mixed 0.08299972 0.025088427  3.308287
```

What went wrong

How could it be that "Racial and ethnic minorities . . . are especially hurt by strict voter identification laws" (HLN) if voter id laws do not suppress turnout?

What went wrong

How could it be that "Racial and ethnic minorities . . . are especially hurt by strict voter identification laws" (HLN) if voter id laws do not suppress turnout?

Although they control for a certain type of omitted variable, fixed-effects estimates are notoriously susceptible to measurement error.

- On one hand, outcomes tend to be persistent (e.g. union membership).
- On the other hand, measurement error often changes from year-to-year (e.g. union status may be misreported or miscoded this year but not next year).
- while union status may be misreported or miscoded for only a few workers in any single year, the observed year-to-year changes in union status may be mostly noise.

Virginia

State policy in Virginia (in effect till 2010)

- No access to vote history
- HLN code VA CCES respondents as nonvoters in 2006,2008.
- In fact their vote status are missing

```
out<- which(dd$state=='Virginia') #VA data
#VA turnout by year
vas<- table(dd$votegenval[out], dd$year[out])
vas
```

```
##
##     2006 2007 2008 2009 2010 2011 2012 2014
##   0  537    0  720    0    0    0  196  330
##   1    1    0    1    0    0    0 1064  634
```

Dropping VA

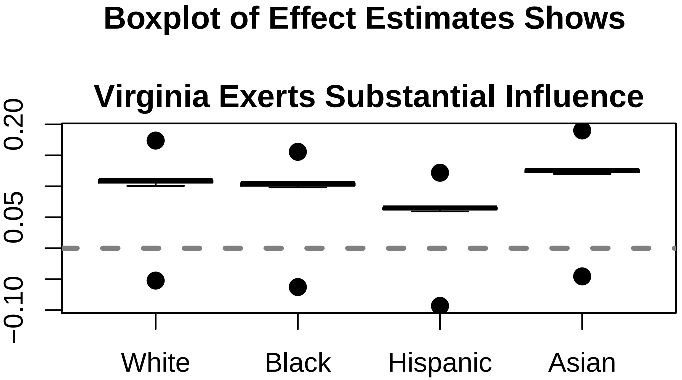
```r
##identifying states
un_state<- unique(dd$state)
un_state<- sort(un_state)
states<- as.character(rev(un_state))
##storing the models
state_early<- list()
sample_states = sample(states,size = 10)
if (!"Virginia" %in% sample_states){
  sample_states[11] = "Virginia"
}
for(z in sample_states){
  ##dropping one state at a time
  drops<- rep(0, nrow(dd))
  drops[which(dd$state==z)]<- 1
  temp<- lm_robust(m1,data =subset(dd,
          voteregpre==1&drops==0))
  state_early[[z]]<- temp
}
```

Store results

```r
#store main and interaction results
White<- Black<- Hispanic<- Asian<- c()

for(z in 1:length(sample_states)){
White[z]<-summary(state_early[[z]])$
  coefficients['stricty', 1]

Black[z]<- White[z]+summary(state_early[[z]])$
  coefficients['blackstricty',1]

Hispanic[z]<- White[z]+summary(state_early[[z]])$
  coefficients['hispstricty',1]

Asian[z]<- White[z]+summary(state_early[[z]])$
  coefficients['asianstricty',1]
}
```

Results

```
boxplot(cbind(White, Black, Hispanic, Asian), cex = 2,
        main = "Boxplot of Effect Estimates Shows
        \nVirginia Exerts Substantial Influence", pch=20)
abline(h = 0, lwd = 3, lty = 2, col=gray(0.5))
```



Final Estimates

```r
###Finally, dropping Virginia and 2006 observations
good_obs<- rep(1, nrow(dd))
good_obs[which(dd$year==2006)]<- 0
good_obs[which(dd$year==2008 & dd$state=='Virginia')]<- 0


###running the regression
model2<- lm_robust(m1,data =
                     subset(dd,voteregpre==1& good_obs==1))
```

Storing results

```
#store main and interaction results
results <- data.frame(
  race = c("White","Black","Hispanic","Asian","Mixed"))
results$m2_ses <- results$m2_est <-  NA
results$m2_est[1]<-
  summary(model2)$coefficients['strictY', 1]

results$m2_est[2]<-results$m2_est[1]+
  summary(model2)$coefficients['blackstrictY', 1]

results$m2_est[3]<-results$m2_est[1]+
  summary(model2)$coefficients['hispstrictY', 1]

results$m2_est[4]<-results$m2_est[1] +
  summary(model2)$coefficients['asianstrictY', 1]

results$m2_est[5]<-results$m2_est[1] +
  summary(model2)$coefficients['mixedracestrictY', 1]
```

Standard Errors

```
#variance covariance matrix of coefficients
v_model<- vcov(model2)
#variance of main+interaction effects
results$m2_ses[1]<- sqrt(v_model['stricty', 'stricty'])
results$m2_ses[2]<- sqrt(v_model['stricty', 'stricty']
                  +v_model['blackstricty', 'blackstricty']
                  +2*v_model['stricty', 'blackstricty'])
results$m2_ses[3]<- sqrt(v_model['stricty', 'stricty']
                  +v_model['hispstricty', 'hispstricty']
                  +2*v_model['stricty', 'hispstricty'])
results$m2_ses[4]<- sqrt(v_model['stricty', 'stricty']
                  +v_model['asianstricty', 'asianstricty']
                  +2*v_model['stricty', 'asianstricty'])
results$m2_ses[5]<- sqrt(v_model['stricty', 'stricty']
                   +v_model['mixedracestricty','mixedracest
                   +2*v_model['stricty', 'mixedracestricty
```

Results

```r
#t statistic
results$tval <- results$m2_est/results$m2_ses
results
```

```
##       race      m2_est     m2_ses      tval
## 1    White -0.02776382 0.01067614 -2.600547
## 2    Black -0.03205925 0.01343164 -2.386845
## 3 Hispanic -0.07308328 0.01815341 -4.025872
## 4    Asian -0.05154905 0.04023828 -1.281095
## 5    Mixed -0.05296484 0.02584260 -2.049517
```

Block bootstrap

- *Block bootstrap* is when we bootstrap whole groups (states, etc) instead of *it* pairs.
- Accounts for correlations within the groups (serial correlation, etc).
- Computationally tricky because you need to keep track of all the multilevel indices.

Block bootstrap coding

```
library(dplyr)
#simplify data for bootstrap
dat <- subset(dd,voteregpre==1)  %>%
  group_by(year,state) %>%
  summarize(votegenval = mean(votegenval,na.rm = T),
            stricty = max(stricty))
#running the model
model4 <- lm_robust(votegenval ~ stricty + as.factor(year)
               as.factor(state),data =dat,clusters=as.factor

#Trick to get the indices for each group
lookup <- split(1:nrow(dat), dat$state)
names(lookup[10]); head(lookup[[10]]) #Inspect 10th state
```

```
## [1] "Florida"
```

```
## [1]  10  61 112 163 214 265
```

```
gnames <- names(lookup) #extract state names
```

Block bootstrap

```r
sims = 2000
ates <- c()
for (j in 1:sims){
  #sample from states
  star <- sample(gnames, size = length(gnames),
                 replace = TRUE)
  dat.star <- dat[unlist(lookup[star]),] #new data
  temp<- lm_robust(votegenval ~ stricty + as.factor(year) +
               as.factor(state),data =dat.star)
  ates[j] <- summary(temp)$coefficients['stricty',1]
}
sd(ates)
```

```
## [1] 0.1019129
```

```r
summary(model4)$coefficients['stricty',2]
```

```
## [1] 0.09864232
```