Pset7

2024-02-21

In this code block, we import the data, as well as installing any necessary packages.

```
source("./Functions.R")

install_packages_if_needed(c("utils", 'tidyr', 'plm', 'miceadds'))
library(tidyr)
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stats)
```

```
#Import the csv files
dt_psid <- data.table::as.data.table(utils::read.delim(file = "nswpsid.csv",
                                    sep = ","))

#for future use, we add an variable, id, to dt_psid:
dt_psid <- dt_psid %>% dplyr::mutate(.data = dt_psid, id = dplyr::row_number())
```

Q1

a

Here we reshape nswpsid into long format:

```
dt_long <- tidyr::gather(dt_psid, dyear2, earns, re75:re78)
```

Here we check the number of rows are indeed 5350:

```
nrow(dt_long)
```

```
## [1] 5350
```

Here we add the two variables, dyear2 and tdyear2:

```
dt_long$dyear2<-ifelse(dt_long$dyear2=="re78",1,0)

dt_long <-
  dt_long %>%
    dplyr::mutate(
        .data = dt_long,
        tdyear2 = treat*dyear2
    )
```

b

Here we create the treated data frame, and verify the number of rows is indeed 370:

```
dt_treat <- dt_long[dt_long$treat == 1,]
nrow(dt_treat)
```

```
## [1] 370
```

c

We now wish to estimate the BA comparison ρ using lm():

```
fo2 <- as.formula("earns ~ dyear2")
ba <- stats::lm(formula = fo2, data = dt_treat)
summary(ba)
```

```
##
## Call:
## stats::lm(formula = fo2, data = dt_treat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6349  -2105  -1532   1390  53959
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1532.1      441.9   3.467 0.000589 ***
## dyear2        4817.1      625.0   7.708 1.21e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6011 on 368 degrees of freedom
## Multiple R-squared:  0.139,  Adjusted R-squared:  0.1367
## F-statistic: 59.41 on 1 and 368 DF,  p-value: 1.206e-13
```

We see that the estimated ρ is 4817.1 dollars, meaning that within the treated group, the income level increased by 4817.1 dollar after the treatment (vs before the treatment).

d

We verify our result from lm “by hand”:

```
mean(dt_treat[dt_treat$dyear2==1,]$earns)-mean(dt_treat[dt_treat$dyear2==0,]$earns)
```

```
## [1] 4817.09
```

Indeed we have the same result.

e

The key assumption of BA Comparison is that each individuals (within the treated group) before the treatment is a “good control” for themselves after the treatment. That is, there is no common unobserved time trend between before and after the treatment (expressed as $\delta_{t=2} - \delta_{t=1} = 0$ in class). But if there are a common time trend across individuals (in this case, if on average the income level naturally goes up or down between 1975 and 1978), the BA Comparison would be an biased estimator of ATT.

Q2

We estimate the coefficients in eq3:

```
fo3 <- as.formula("re78 - re75 ~ treat")
fd<-stats::lm(formula = fo3, data = dt_psid)
summary(fd)
```

```
##
## Call:
## stats::lm(formula = fo3, data = dt_psid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73924  -3911   -956   3888 118683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2490.6      214.0  11.637  < 2e-16 ***
## treat         2326.5      813.9   2.859  0.00429 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10680 on 2673 degrees of freedom
## Multiple R-squared:  0.003048,   Adjusted R-squared:  0.002675
## F-statistic: 8.172 on 1 and 2673 DF,  p-value: 0.004288
```

$\widehat{ATT}^{FD} = 2326.5$

Q3

Note: We interpret $D_{i,t}$ as tdyear2, since otherwise treat does not vary across time. LSDV estimator:

```
fo4 <- as.formula("earns ~ tdyear2 + as.factor(id) + as.factor(dyear2)")
lsdv <- stats::lm(formula = fo4, data = dt_long)
```

We get the coefficient on treat here:

```
lsdv$coefficients[2]
```

```
##  tdyear2
## 2326.505
```

$\widehat{ATT}_{78}^{LSDV} = 2326.5$

Q4

We give the TWFE estimator for eq4 here:

```
pdata <- plm::pdata.frame(dt_long, index = c('id', 'dyear2'))
twfe <- plm::plm(formula = earns ~ tdyear2, data = pdata, model = 'within', effect = 'twoway
s')
```

Coefficient on Treated:

```
twfe$coefficients[1]
```

```
##  tdyear2
## 2326.505
```

We get the same number as $\widehat{ATT}_{78}^{LSDV}$ (as expected: TWFE estimator should perform similar operation as the OLS on factor(id) and factor(dyear2), or so I think).

Q5

```
dd <- stats::lm(formula = earns ~ dyear2 + treat + tdyear2, data = dt_long)
dd$coefficients[4]
```

```
##  tdyear2
## 2326.505
```

$\widehat{ATT}_{78}^{DD} = 2326.505$

Q6

We see that $\widehat{ATT}^{FD} = \widehat{ATT}_{78}^{LSDV} = \widehat{ATT}_{78}^{TWFE} = \widehat{ATT}_{78}^{DD} = 2326.505$. This is not surprising, since equations (3), (4), and (5) can be seen as different decomposition of the same equation. LSDV and TWFE estimators should also perform the same functions, as they are mechanically the same here (?).

We also see that $\widehat{ATT}^{BA} \neq 2326.505$, suggesting that the time trend is not zero.

Q7

We retrieve the four means:

```
#earns treat=1, year=78
e11 <- mean(dt_long[which(dt_long$treat==1 & dt_long$dyear2==1),]$earns)
#earns treat=0, year=78
e01 <- mean(dt_long[which(dt_long$treat==0 & dt_long$dyear2==1),]$earns)
#earns treat=1, year=75
e10 <- mean(dt_long[which(dt_long$treat==1 & dt_long$dyear2==0),]$earns)
#earns treat=0, year=75
e00 <- mean(dt_long[which(dt_long$treat==0 & dt_long$dyear2==0),]$earns)

(e11-e10)-(e01-e00)
```

```
## [1] 2326.505
```

Indeed, our estimations using lm() and other methods are confirmed “by hand”.

Q8

As mentioned earlier, BA comparison makes the assumption that the time trend, $\delta_{t=2} - \delta_{t=1}$, is equal to zero. This may not be justified in many situations, such as wages, since they have a certain trend across individuals due to macro environment and other unobservables.

On the other hand, TCC does not use the time component of the panel data, but it makes the assumption that the expected difference between treated and control groups, $E[\mu_i|D_i = 1] - E[\mu_i|D_i = 0]$ is zero. This means that differences between individuals are mean independent of treatment assignment. This assumption may also be violated in situations where treatment assignment depend on observable and unobservable characteristics, such as in our NSW studies in which individuals are randomly assigned within a certain sub-population.

DD improves over BA and TCC in two ways. First, it uses more parts of the panel data (in the quadrant table from class, DD uses all four squares rather than just two). But more importantly, DD is much more plausible when we have reasons to believe that there is a common time trend, and treatment assignment is not mean independent of all variables, observable or otherwise.

Part 2

Q9

Article read.

Q10

Loading data set

```
rm(list = ls())
dt_ff <- data.table::as.data.table(utils::read.delim(file = "fast-food-data.csv",
                                    sep = ","))
```

The authors have panel data, since they deliberately chose to interview the same individuals in different times.

Q11

```
dt_ff %>%
  group_by(state) %>%
  summarise(
    total_count = n(),
    closed_permanently = sum(status2 == 3),
    closed_for_renovations = sum(status2 == 2),
    closed_temporarily = sum(status2>3),
    refused_second_interview = sum(status2==0),
    answered_2nd_interview = sum(status2==1)
  )
```

```
## # A tibble: 2 × 7
##   state total_count closed_permanently closed_for_reno…¹ close…² refus…³ answe…⁴
##   <int>       <int>              <int>             <int>   <int>   <int>   <int>
## 1     0          79                  1                 0       0       0      78
## 2     1         331                  5                 2       2       1     321
## # … with abbreviated variable names ¹closed_for_renovations,
## #   ²closed_temporarily, ³refused_second_interview, ⁴answered_2nd_interview
```

This table shows that the distribution of responses in the 2nd interview, among the 410 individuals who are interviewed the first time. It showed that the panel data is very balanced (399 out of 410 responded to both interviews). It also showed the reason for non-response, and refusal without reason is very few (only 1 in NJ).

(The table is side-ways, but should be readable)

Q12

a

Here we construct fte and fte2:

```
dt_ff <- dt_ff %>% mutate(.data = dt_ff,
                 fte = empft+nmgrs+0.5*emppt,
                 fte2 = empft2+nmgrs2+0.5*emppt2)
```

b

Here we gave the means in both variables:

```
dt_ff %>% group_by(state)%>%
          summarise(
            mean_fte = mean(fte, na.rm = T),
            mean_fte2 = mean(fte2, na.rm = T)
          )
```

```
## # A tibble: 2 × 3
##   state mean_fte mean_fte2
##   <int>    <dbl>     <dbl>
## 1     0     23.3      21.2
## 2     1     20.4      21.0
```

We see that these numbers corresponds to results from table 2.

c

The purpose of showing “Distribution of Store types” is to give a sense of the kind of stores, brands, and ownership in both states, and to highlight that fast food chain stores in both NJ and PA are similarly composed in these dimensions.

The purpose of showing “Means in Wave 1” is to highlight the similarities in other observables between stores in both NJ and PA before the minimum wage law changes. These observables include starting wages, meal prices, and number of open hours on weekdays. We see that indeed the claim made in the paper that NJ and PA have similarly behaving fast food stores, potentially justifying the mean independence assumption.

Q13

a

Since we labeled “state” variable as 1 if in NJ and 0 if in PA, we use linear regression on “state” to get our results:

```
# lm_fte <- lm(formula = fte ~ state, data = dt_ff)
# summary(lm_fte)

ttest1 <- t.test(formula = fte ~ state, data = dt_ff)
ttest2 <- t.test(formula = fte2 ~ state, data = dt_ff)
```

Here is the means in wave 1:

```
ttest1$estimate
```

```
## mean in group 0 mean in group 1
##        23.33117        20.43941
```

Here are the standard errors for PA and NJ in wave 1, respectively:

```
source("./Functions.R")
std_err(dt_ff[which(state==0 & !is.na(fte))]$fte)
```

```
## [1] 1.351149
```

```
std_err(dt_ff[which(state==1 & !is.na(fte))]$fte)
```

```
## [1] 0.5082607
```

Here are the difference between the two means and its standard error:

```
ttest1$estimate[2]-ttest1$estimate[1]
```

```
## mean in group 1
##       -2.891761
```

```
ttest1$stderr
```

```
## [1] 1.443583
```

We proceed similarly for fte2, fte2-fte, and modified fte2-fte: For the second half (fte2):

```
ttest2$estimate
```

```
## mean in group 0 mean in group 1
##        21.16558        21.02743
```

```
print(c(std_err(dt_ff[which(state==0 & !is.na(fte2))]$fte2),std_err(dt_ff[which(state==1 & !is.na(fte2))]$fte2)))
```

```
## [1] 0.9432212 0.5203094
```

```
ttest2$estimate[2]-ttest2$estimate[1]
```

```
## mean in group 1
##      -0.1381549
```

```
ttest2$stderr
```

```
## [1] 1.077213
```

b

In row 3, we have: (i) the difference between the average fte from wave 1 and wave 2 respondents in PA (wave 1 average include those who do not respond in the second time);

ii. the difference between the average fte from wave 1 and wave 2 respondents in NJ (wave 1 average include those who do not respond in the second time)

iii. the difference between (i) and (ii).

(Since this is the result from the unbalanced panel, it is somewhat offputing.) (i) can be seen as an estimation of $E[y_{i,t}|D_i=0,t=1]-E[y_{i,t}|D_i=0,t=0]$, i.e. time trend for the untreated, although attrition introduces bias in the first term.

ii. can be seen as an estimation of $E[y_{i,t}|D_i=1,t=1]-E[y_{i,t}|D_i=1,t=0]$, i.e. BA comparison estimator for the treated, but attrition also introduces bias.

iii. can be seen as an estimation of DD estimator of ATT (or the effect of NJ's MW law change on NJ stores), assuming common time trend, but again, attrition would make this estimator biased. (strictly this is multi-cross-sectional data)

Q14

a

It is balanced sample in the sense that we have data for those stores in both waves, hence “balanced” across time.

The balanced sample is preferable in that it is easier to justify using this sample at t=1 as control for themselves at t=2: it is more reasonable to say that a store is similar to itself across time, than to say that a set of stores in NJ at t=1 is similar to part of this set of stores at t=2. Hence preferably, we want to have individuals giving data all of the times and compare their present selves with their past selves, rather than comparing different people at different time.

b

For the 4th row (fte2-fte):

```
dt_ff <- dt_ff %>% mutate(.data = dt_ff, d_fte = fte2-fte)

ttest3 <- t.test(formula = d_fte ~ state, data = dt_ff)
ttest3$estimate
```

```
## mean in group 0 mean in group 1
##      -2.2833333       0.4666667
```

```
print(c(std_err(dt_ff[which(state==0 & !is.na(d_fte))]$d_fte),std_err(dt_ff[which(state==1 &
!is.na(d_fte))]$d_fte)))
```

```
## [1] 1.2532690 0.4808286
```

```
ttest3$estimate[2]-ttest3$estimate[1]
```

```
## mean in group 1
##            2.75
```

```
ttest3$stderr
```

```
## [1] 1.342341
```

For the 5th row:

```
dt_ff[status2 >=4]$d_fte <- -dt_ff[status2 >=4]$fte
dt_ff[status2 ==2]$d_fte <- -dt_ff[status2 ==2]$fte

ttest4 <- t.test(formula = d_fte ~ state, data = dt_ff)
ttest4$estimate
```

```
## mean in group 0 mean in group 1
##      -2.2833333       0.2258786
```

```
print(c(std_err(dt_ff[which(state==0 & !is.na(d_fte))]$d_fte),std_err(dt_ff[which(state==1 &
!is.na(d_fte))]$d_fte)))
```

```
## [1] 1.2532690 0.4909788
```

```
ttest4$estimate[2]-ttest4$estimate[1]
```

```
## mean in group 1
##        2.509212
```

```
ttest4$stderr
```

```
## [1] 1.34601
```

c

ii. gives an estimation of BA estimator. It is comparing the stores which were affected by the law (after) with their past selves *before the law change, and see how their employment (FTE) change around the MW law change.

We see that on average, the NJ stores which were impacted by the MW law change hired 0.47 more full-time-equivalent of employees than they did before the law change. The standard error of this estimation (0.48) suggests that we do not expect this to be significantly different from “no change in hiring practice before and after MW change.”

d

iii. gives an estimation of DD estimator. We see that on average, the difference between NJ and PA fast food employment are widened by 2.75 full-time-equivalent of employees. The standard error (1.34) suggests that this is not that different from “the NJ fast food stores didn’t hire more or fewer people, relative to PA fast food stores, after the MW change.”

e

Row 5 provides an alternative interpretation of the data. Stores temporarily closing can be seen as non-reponsive to Wave 2, thus ruled out in balanced panel (as in row 4). But it can also be seen as employing 0 people in the short term. This potentailly expands our balanced panel, but also tests if our conclusion from row 4 can withstand different interpretations of the data.

Q15

Reshape to long format:

```
dt_ff_long <- tidyr::gather(dt_ff, wave, fte, fte:fte2)
dt_ff_long <- data.table::as.data.table(dt_ff_long)
dt_ff_long$wave<-ifelse(dt_ff_long$wave=="fte",0,1)
```

Q16

We now give cluster-robust regresion estimation of the formula:

```
dt_ff_long <- dt_ff_long %>% mutate (.data = dt_ff_long,
                                     statewave = state*wave)
fo <- as.formula('fte ~ wave + state + statewave')

lm1 <- miceadds::lm.cluster(data = dt_ff_long, formula = fo, cluster = "store_id")
summary(lm1)
```

```
## R^2= 0.0074
##
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 23.331169   1.346536 17.326807 2.952727e-67
## wave        -2.165584   1.218025 -1.777948 7.541243e-02
## state       -2.891761   1.439546 -2.008800 4.455832e-02
## statewave    2.753606   1.306607  2.107448 3.507880e-02
```

They are not quite the same (!). The coefficient on statewave, $\hat{\rho}$, should correspond to row 4 (iii), while $\hat{\hat{\delta}}$ should correspond to row 4 (i). What went wrong?

#Q17 We now test the alternative interpretation (row5):

```
dt_ff_long[which(status2==2 & wave ==1)]$fte <- 0
dt_ff_long[which(status2>=4 & wave ==1)]$fte <- 0

lm2 <- miceadds::lm.cluster(data = dt_ff_long, formula = fo, cluster = "store_id")
summary(lm2)
```

```
## R^2= 0.0074
##
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 23.331169   1.346523 17.326972 2.944263e-67
## wave        -2.165584   1.218013 -1.777965 7.540964e-02
## state       -2.891761   1.439532 -2.008819 4.455629e-02
## statewave    2.493204   1.310664  1.902245 5.713908e-02
```

Again, not quite right. I have no idea what went wrong.

Q18

a

Sample selection on employment and starting wage not na:

```
rm(list = ls())
dt_ff <- data.table::as.data.table(utils::read.delim(file = "fast-food-data.csv",
                                    sep = ","))
dt_ff <- dt_ff %>% mutate(.data = dt_ff,
                 fte = empft+nmgrs+0.5*emppt,
                 fte2 = empft2+nmgrs2+0.5*emppt2)

#set all closed, permanently and temporarily, store fte2 as 0
dt_ff[status2 == 3]$wage_st2 <- 0

dt_sub <- dt_ff[which(!is.na(wage_st)
                      &!is.na(wage_st2)
                      & !is.na(fte)
                      & !is.na(fte2)
                      )]

nrow(dt_sub)
```

```
## [1] 357
```

Indeed we have 357 observations.

b

The mean and standard deviation, respectively:

```
dt_sub <- dt_sub %>% mutate(.data = dt_sub,
                 diff_fte = fte2-fte)
mean(dt_sub$diff_fte, na.rm = T)
```

```
## [1] -0.237535
```

```
sd(dt_sub$diff_fte)
```

```
## [1] 8.825485
```

which corresponds to the paper.

c

For (i):

```
lm1 <- stats::lm(data = dt_sub, formula = diff_fte ~ state)
summary(lm1)
```

```
##
## Call:
## stats::lm(formula = diff_fte ~ state, data = dt_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.373  -3.873   0.551   4.301  27.801
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.127      1.074  -1.980   0.0484 *
## state          2.326      1.192   1.952   0.0517 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.791 on 355 degrees of freedom
## Multiple R-squared:  0.01062,    Adjusted R-squared:  0.007831
## F-statistic:  3.81 on 1 and 355 DF,  p-value: 0.05174
```

we have coefficient 2.33 and std.error 1.19.

For (ii):

```
lm2 <- stats::lm(data = dt_sub, formula = diff_fte ~ state+ as.factor(chain) + co_owned)
summary(lm2)
```

```
##
## Call:
## stats::lm(formula = diff_fte ~ state + as.factor(chain) + co_owned,
##     data = dt_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.803  -3.903   0.606   4.106  27.393
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.6972     1.2104  -1.402   0.1618
## state               2.3039     1.1955   1.927   0.0548 .
## as.factor(chain)2   0.4922     1.2992   0.379   0.7050
## as.factor(chain)3  -2.2170     1.3099  -1.693   0.0914 .
## as.factor(chain)4  -0.5120     1.4984  -0.342   0.7328
## co_owned            0.3080     1.0939   0.282   0.7785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.785 on 351 degrees of freedom
## Multiple R-squared:  0.02315,    Adjusted R-squared:  0.009231
## F-statistic: 1.663 on 5 and 351 DF,  p-value: 0.1427
```

we have coefficient 2.30 and std.error 1.20;

d

In (i), the coefficient on NJ dummy estimates the differences in employment changes over time between the two states, i.e. DD estimator, unconditioned on any other variables.

In (ii), the coefficients on NJ dummy estimates: the difference in employment changes over time between the two states, given same brand and company ownership. It is a DD estimator conditional on brand and ownership categorical variables.

Q 19

a and b

For (iii):

```
#create new variable gap
dt_sub <- dt_sub%>% mutate(data = dt_sub,
                            gap = case_when(
                              state == 0 ~ 0,
                              state == 1 & wage_st>=5.05 ~ 0,
                              state == 1 & wage_st<5.05 ~ (5.05-wage_st)/wage_st
                            )
                          )

lm3 <- stats::lm(data = dt_sub, formula = diff_fte ~ gap)
summary(lm3)
```

```
##
## Call:
## stats::lm(formula = diff_fte ~ gap, data = dt_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.924  -3.870   0.380   4.588  26.630
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5764     0.6966  -2.263   0.0242 *
## gap          15.6529     6.0802   2.574   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.757 on 355 degrees of freedom
## Multiple R-squared:  0.01833,    Adjusted R-squared:  0.01556
## F-statistic: 6.627 on 1 and 355 DF,  p-value: 0.01045
```

The coeff. (15.65) and std. error (6.08) are correct.

For (iv):

```
lm4 <- stats::lm(data = dt_sub, formula = diff_fte ~ gap + as.factor(chain) + co_owned)
summary(lm4)
```

```
## 
## Call:
## stats::lm(formula = diff_fte ~ gap + as.factor(chain) + co_owned, 
##     data = dt_sub)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -40.185  -4.152   0.268   4.373  26.507 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)  
## (Intercept)       -1.31503    0.95676  -1.374   0.1702  
## gap               14.91567    6.20533   2.404   0.0167 *
## as.factor(chain)2  0.65895    1.29231   0.510   0.6104  
## as.factor(chain)3 -1.91221    1.30835  -1.462   0.1448  
## as.factor(chain)4  0.01362    1.51518   0.009   0.9928  
## co_owned           0.40891    1.09233   0.374   0.7084  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.759 on 351 degrees of freedom
## Multiple R-squared:  0.0288, Adjusted R-squared:  0.01496 
## F-statistic: 2.082 on 5 and 351 DF,  p-value: 0.06717
```

The coeff. (14.92) and std. error (6.21) are aligned with the table 4.

For (v): (per pg 781, there are 3 nj dummies used)

```
lm5 <- stats::lm(data = dt_sub, formula = diff_fte ~ gap + as.factor(chain) + co_owned + pa1
+ pa2+ centralj+northj+southj)
summary(lm5)
```

```
##
## Call:
## stats::lm(formula = diff_fte ~ gap + as.factor(chain) + co_owned +
##     pa1 + pa2 + centralj + northj + southj, data = dt_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.274  -4.175   0.445   4.442  26.344
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.5992     1.5226  -0.394    0.694
## gap                11.9792     7.4191   1.615    0.107
## as.factor(chain)2   0.6653     1.3097   0.508    0.612
## as.factor(chain)3  -2.0895     1.3279  -1.574    0.117
## as.factor(chain)4   0.2705     1.5318   0.177    0.860
## co_owned            0.1596     1.1101   0.144    0.886
## pa1                -3.4621     2.1324  -1.624    0.105
## pa2                 0.3732     1.9130   0.195    0.845
## centralj           -1.2440     1.5556  -0.800    0.424
## northj              0.1266     1.2118   0.104    0.917
## southj                  NA         NA      NA       NA
##
## Residual standard error: 8.752 on 347 degrees of freedom
## Multiple R-squared:  0.04134,    Adjusted R-squared:  0.01647
## F-statistic: 1.663 on 9 and 347 DF,  p-value: 0.09669
```

the coeff. (11.9792) and std. error (7.4191) do not match. This may be due to lm() automatically dropping one of the regional variables (southj in this case).

C

The coefficients on the gap variable (15.65, 14.92, and 11.91 respectively) would be interpreted as the impact of NJ's MW change, while (iii) would be this impact not conditional on any additional variables, (iv) would be conditional on brands and company ownership, (v) will also be conditional on regional variations.

Q20

The stores in NJ that already paid wages higher than the new MW law before the change would provide a different kind of control group, for the stores which paid lower than new MW before the change. This is because the "high-wage" stores would not be impacted by the new MW law, but the "low-wage" ones would have to update their wage and potentially employment. Since NJ's small size, the authors argue that the within-regional differences would be stable over time (common time trend within NJ).

Estimations from this alternative "treatment-control" grouping provides a natural analogue to the NJ-PA natural experiment. For instance, if NJ-PA sees widening gap over time, but high-low wage within NJ does not, then it is possible that there are unobserable regional differences between NJ and PA stores around the MW change, and they may not be very good comparison.

Q21

One concern over the result in the paper is that the balanced panel data contains survivor bias: only those stores that remain open would be in both waves. This bias would cause us to over-estimate the overall employment, if we only use data within the balanced panel, since MW changes may outright push stores out of business.

The author accounts for this by setting the employment of closed stores after the change to 0. In this way, we expand the panel to include those stores that did not survive, and reduce the survivor bias.

The claim that this result “measure the overall effect of the minimum wage on average” is still somewhat strong, since there are still stores unresponsive in the wave 2, and we do not know their conditions. However, it is more reasonable to see the panel data that include closed stores as less biased than otherwise.