Causal Inference4

Weiheng Zhang

2022-04-17

Contents

Fixed Effects, DID and Panel data **1**
Fixed effect 1
Differences-in-Differences: pre and post treatment and control 11

```
library(FinMetric)
library(formatR)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60),
                      tidy=TRUE,
                      echo = TRUE)
```

Fixed Effects, DID and Panel data

Fixed effect

Individual Fixed Effect

- Setting

  Question: because workers in union, then they wage high; or they earn more because of they are more experienced

$$\begin{aligned}
Y_{it} &: \text{log earnings of worker i at time t;} \quad Y_{it} = Y_{0it} + (Y_{1it} - Y_{0it})D_{it} \\
D_{it} &: \text{union status;} \\
A_i &: \text{unobserved worker ability;} \\
X_{it} &: \text{other observed covariate}
\end{aligned}$$

  Suppose:

$$E(Y_{0it}|A_i, X_{it}, t, D_{it}) = E(Y_{0it}|A_i, X_{it}, t)$$

- key fixed-effects estimation assumption: A_i appears without a time subscrit in a linear model:

$$E(Y_{0it}|A_i, X_{it}, t) = \alpha + \lambda_t + A_i'\gamma + X_{it}'\beta$$

- We assume the causal effect of union membership is additive and constant:

$$E(Y_{1it}|A_i, X_{it}, t) = E(Y_{0it}|A_i, X_{it}, t) + \rho$$

- Thus, we have:

$$E(Y_{1it}|A_i, X_{it}, t) = \alpha + \lambda_t + \rho D_{it} + A_i'\gamma + X_{it}'\beta$$

$$\Rightarrow Y_{1it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}$$

  where $\alpha_i = \alpha + A_i'\gamma$.

  Individual effect: α_i

  Year effect: λ_t

  By within transformation, we can eliminate the individual effect, we can estimate ρ consistently.

Application

```
# Balanced panels
data("Grunfeld", package = "plm")
Grunfeld %>%
    select(year, firm) %>%
    table()
```

```
##       firm
## year   1 2 3 4 5 6 7 8 9 10
##   1935 1 1 1 1 1 1 1 1 1  1
##   1936 1 1 1 1 1 1 1 1 1  1
##   1937 1 1 1 1 1 1 1 1 1  1
##   1938 1 1 1 1 1 1 1 1 1  1
##   1939 1 1 1 1 1 1 1 1 1  1
##   1940 1 1 1 1 1 1 1 1 1  1
##   1941 1 1 1 1 1 1 1 1 1  1
##   1942 1 1 1 1 1 1 1 1 1  1
##   1943 1 1 1 1 1 1 1 1 1  1
##   1944 1 1 1 1 1 1 1 1 1  1
##   1945 1 1 1 1 1 1 1 1 1  1
##   1946 1 1 1 1 1 1 1 1 1  1
##   1947 1 1 1 1 1 1 1 1 1  1
##   1948 1 1 1 1 1 1 1 1 1  1
##   1949 1 1 1 1 1 1 1 1 1  1
##   1950 1 1 1 1 1 1 1 1 1  1
##   1951 1 1 1 1 1 1 1 1 1  1
##   1952 1 1 1 1 1 1 1 1 1  1
##   1953 1 1 1 1 1 1 1 1 1  1
##   1954 1 1 1 1 1 1 1 1 1  1
```

```
# Unbalanced panels

data("EmplUK", package = "plm")
EmplUK %>%
    select(year, firm) %>%
    filter(firm %in% c(1:10)) %>%
    table()
```

```
##       firm
## year   1 2 3 4 5 6 7 8 9 10
```

```
##   1976 0 0 0 0 1 1 1 1 1  1
##   1977 1 1 1 1 1 1 1 1 1  1
##   1978 1 1 1 1 1 1 1 1 1  1
##   1979 1 1 1 1 1 1 1 1 1  1
##   1980 1 1 1 1 1 1 1 1 1  1
##   1981 1 1 1 1 1 1 1 1 1  1
##   1982 1 1 1 1 1 1 1 1 1  1
##   1983 1 1 1 1 0 0 0 0 0  0
```

```r
# Balance unbalanced data

# Using 'fill' creates a new row with NAs for each missing
# time point.
EmplUK.balanced1 <- make.pbalanced(EmplUK, balance.type = "fill")
EmplUK.balanced1[1:8, ]
```

Balance unbalanced data

```
##   firm year sector   emp    wage capital   output
## 1    1 1976     NA    NA      NA      NA       NA
## 2    1 1977      7 5.041 13.1516  0.5894  95.7072
## 3    1 1978      7 5.600 12.3018  0.6318  97.3569
## 4    1 1979      7 5.015 12.8395  0.6771  99.6083
## 5    1 1980      7 4.715 13.8039  0.6171 100.5501
## 6    1 1981      7 4.093 14.2897  0.5076  99.5581
## 7    1 1982      7 3.166 14.8681  0.4229  98.6151
## 8    1 1983      7 2.936 13.7784  0.3920 100.0301
```

```r
# Using 'shared.times' keeps all available firms in the
# dataset but drops all time periods where at least one
# firm has no data.
EmplUK.balanced2 <- make.pbalanced(EmplUK, balance.type = "shared.times")
EmplUK.balanced2[1:10, ]
```

```
##    firm year sector    emp    wage capital   output
## 2     1 1978      7  5.600 12.3018  0.6318  97.3569
## 3     1 1979      7  5.015 12.8395  0.6771  99.6083
## 4     1 1980      7  4.715 13.8039  0.6171 100.5501
## 5     1 1981      7  4.093 14.2897  0.5076  99.5581
## 6     1 1982      7  3.166 14.8681  0.4229  98.6151
## 9     2 1978      7 70.643 14.1036 17.2422  97.3569
## 10    2 1979      7 70.918 14.9534 17.5413  99.6083
## 11    2 1980      7 72.031 15.4910 17.6574 100.5501
## 12    2 1981      7 73.689 16.1969 16.7133  99.5581
## 13    2 1982      7 72.419 16.1314 16.2469  98.6151
```

```r
# By using 'shared.individuals' all available time periods
# are kept but only for those firms which have information
# for each of them.
EmplUK.balanced3 <- make.pbalanced(EmplUK, balance.type = "shared.individuals")

EmplUK.balanced3 %>%
    group_by(firm) %>%
    slice(1)
```

```
## # A tibble: 14 x 7
## # Groups:   firm [14]
##     firm  year sector    emp  wage capital output
##    <dbl> <dbl>  <dbl>  <dbl> <dbl>   <dbl>  <dbl>
##  1   127  1976      7  1.14  14.7    0.690   94.7
##  2   128  1976      7  1.70  14.2    0.422   94.7
##  3   129  1976      7  0.768  8.83   0.201   94.7
##  4   130  1976      3 20     29.3   23.6    102.
##  5   131  1976      9  4.10  25.6    0.504  104.
##  6   132  1976      1  0.971 20.1    0.114  127.
##  7   133  1976      4  2.17  27.6    0.284  118.
##  8   134  1976      8 13.0   15.2    1.50   117.
##  9   135  1976      8  2.37  18.6    0.398  117.
## 10   136  1976      8  0.615 24.1    0.124  117.
## 11   137  1976      2  2.60  38.4    0.664  110.
## 12   138  1976      9  1.92  27.0    0.255  102.
## 13   139  1976      9  1.05  22.2    0.147  103.
## 14   140  1976      3  1.54  29.1    0.651  105.
```

Estimation Methods

```r
# Pooled OLS via lm
pooled_ols_lm <- lm(inv ~ capital, data = Grunfeld)

summary(pooled_ols_lm)
```

Pooled Cross Sections

```
##
## Call:
## lm(formula = inv ~ capital, data = Grunfeld)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -316.92  -96.45  -14.43   17.07  481.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.23620   15.63927    0.91    0.364
## capital      0.47722    0.03834   12.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.9 on 198 degrees of freedom
## Multiple R-squared:  0.439,  Adjusted R-squared:  0.4362
## F-statistic: 154.9 on 1 and 198 DF,  p-value: < 2.2e-16
```

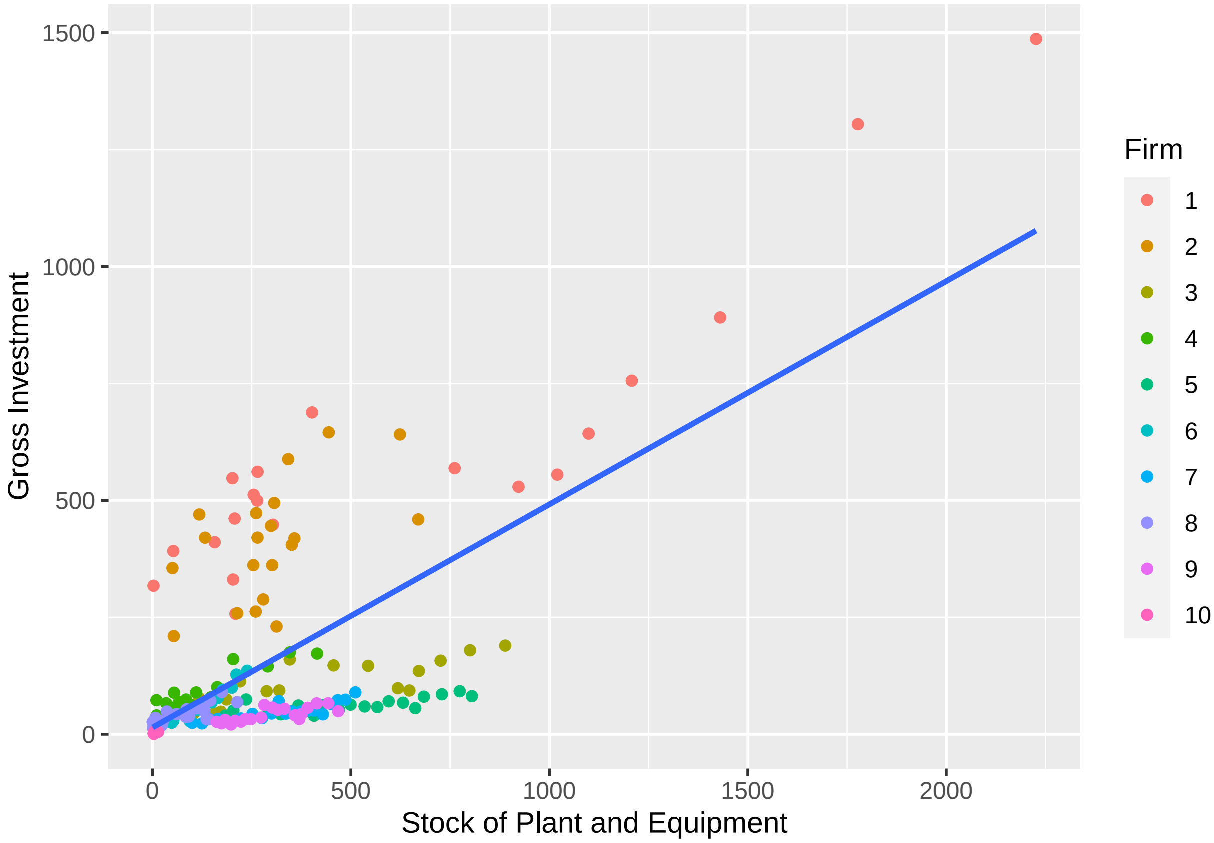
```r
# Pooled OLS via plm
pooled_ols_plm <- plm(inv ~ capital, data = Grunfeld, index = c("firm",
    "year"), effect = "individual", model = "pooling")
coeftest(pooled_ols_plm, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
```

```
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 14.23620   28.18717  0.5051 0.6140785
## capital      0.47722    0.12650  3.7724 0.0002135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(data = Grunfeld, aes(x = capital, y = inv)) + geom_point(aes(color = factor(firm,
    levels = c(1:10)))) + geom_smooth(method = "lm", se = F) +
    labs(x = "Stock of Plant and Equipment", y = "Gross Investment",
        shape = "Firm") + labs(color = "Firm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# FE using lm()
fe_model_lm <- lm(inv ~ capital + factor(firm), data = Grunfeld)

# FE using plm()
fe_model_plm <- plm(inv ~ capital + factor(firm), data = Grunfeld,
    index = c("firm", "year"), effect = "individual", model = "within")

# FE using felm
fe_model_felm <- lfe::felm(inv ~ capital | factor(firm), data = Grunfeld)

rob_se <- function(x) {
    sqrt(diag(vcovHC(x, type = "HC1", group = "firm")))
```

```r
}

se <- list(rob_se(fe_model_lm), rob_se(fe_model_plm), rob_se(fe_model_felm))

stargazer(fe_model_lm, fe_model_plm, fe_model_felm, se = se,
    header = F, keep.stat = c("n", "rsq"), omit = "factor")
```

Table 1:

	Dependent variable:		
	inv		
	OLS	*panel linear*	*felm*
	(1)	(2)	(3)
capital	0.371***	0.371***	0.371***
	(0.057)	(0.062)	(0.056)
Constant	367.613***		
	(31.032)		
Observations	200	200	200
R^2	0.918	0.660	0.918
Note:	*p<0.1; **p<0.05; ***p<0.01		

```r
## Including time dimension lm()
fe_time_lm <- lm(inv ~ capital + factor(firm) + factor(year),
    data = Grunfeld)

# plm()
fe_time_plm <- plm(inv ~ capital, data = Grunfeld, index = c("firm",
    "year"), effect = "twoways", model = "within")
# felm
fe_time_felm <- lfe::felm(inv ~ capital | factor(firm) + factor(year),
    data = Grunfeld)

se <- list(rob_se(fe_time_lm), rob_se(fe_time_plm), rob_se(fe_time_felm))

stargazer(fe_time_lm, fe_time_plm, fe_time_felm, se = se, header = F,
    keep.stat = c("n", "rsq"), omit = "factor")
```

Fixed Effects Model

```r
# More than two time periods
fe_model_fd <- plm(inv ~ capital, data = Grunfeld, index = c("firm",
    "year"), effect = "individual", model = "fd")

summary(fe_model_fd)
```

First-difference Estimator

Table 2:

	Dependent variable:		
	inv		
	OLS	*panel linear*	*felm*
	(1)	(2)	(3)
capital	0.414***	0.414***	0.414***
	(0.072)	(0.057)	(0.067)
Constant	354.917***		
	(33.413)		
Observations	200	200	200
R^2	0.931	0.599	0.931
Note:	*p<0.1; **p<0.05; ***p<0.01		

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = inv ~ capital, data = Grunfeld, effect = "individual",
##     model = "fd", index = c("firm", "year"))
##
## Balanced Panel: n = 10, T = 20, N = 200
## Observations used in estimation: 190
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -240.4300  -15.2906   -4.2478    8.4604  339.4974
##
## Coefficients:
##             Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 4.456749   4.459346  0.9994 0.318877
## capital     0.199671   0.067274  2.9680 0.003387 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    584410
## Residual Sum of Squares: 558250
## R-Squared:      0.04476
## Adj. R-Squared: 0.039679
## F-statistic: 8.80927 on 1 and 188 DF, p-value: 0.003387
```

```r
# only two period Fixed effect model is same as first
# difference estimation Within estimation (two periods)
fe_model_plm_check <- plm(inv ~ capital, data = Grunfeld, subset = year %in%
    c(1935, 1936), index = c("firm", "year"), effect = "individual",
    model = "within")

coeftest(fe_model_plm_check)
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## capital  0.91353    0.85333  1.0705   0.3122
```

```r
# FD estimation (two periods)
fe_model_fd_check <- plm(inv ~ capital - 1, data = Grunfeld,
    subset = year %in% c(1935, 1936), index = c("firm", "year"),
    effect = "individual", model = "fd")


coeftest(fe_model_fd_check)
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## capital  0.91353    0.85333  1.0705   0.3122
```

```r
re_model_plm <- plm(inv ~ capital, data = Grunfeld, index = c("firm",
    "year"), effect = "individual", model = "random")

summary(re_model_plm)
```

Random Effect Model

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = inv ~ capital, data = Grunfeld, effect = "individual",
##     model = "random", index = c("firm", "year"))
##
## Balanced Panel: n = 10, T = 20, N = 200
##
## Effects:
##                    var  std.dev share
## idiosyncratic  4040.63    63.57 0.135
## individual    25949.52   161.09 0.865
## theta: 0.9121
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -164.0821  -22.2955   -3.7463   16.9121  319.9564
##
## Coefficients:
##              Estimate Std. Error z-value Pr(>|z|)
## (Intercept) 43.246697  51.411319  0.8412   0.4002
## capital      0.372120   0.019316 19.2652   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2299300
```

```
## Residual Sum of Squares: 799910
## R-Squared:      0.65211
## Adj. R-Squared: 0.65036
## Chisq: 371.149 on 1 DF, p-value: < 2.22e-16
```

```r
# FE or Pooled OLS
pFtest(fe_model_plm, pooled_ols_plm)
```

Tests for panel data

```
##
##  F test for individual effects
##
## data:  inv ~ capital + factor(firm)
## F = 123.39, df1 = 9, df2 = 189, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```r
# The null hypothesis is rejected in favor of the
# alternative that there are significant fixed effects.

# RE or FE Hausman Test
phtest(fe_model_plm, re_model_plm)
```

```
##
##  Hausman Test
##
## data:  inv ~ capital + factor(firm)
## chisq = 0.93423, df = 1, p-value = 0.3338
## alternative hypothesis: one model is inconsistent
```

```r
# The null hypothesis cannot be rejected here, hence we
# should use a RE model.

# Pooled OLS or RE
plmtest(pooled_ols_plm, effect = "individual", type = c("bp"))
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels
##
## data:  inv ~ capital
## chisq = 1285.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```r
# The test shows that there are significant differences
# across firms. Running a pooled OLS regression is thus not
# appropriate and the RE model is the better choice.

# one-way model or two-way model
pFtest(fe_time_plm, fe_model_plm)
```

```
##
##  F test for twoways effects
##
## data:  inv ~ capital
## F = 1.594, df1 = 19, df2 = 170, p-value = 0.06242
## alternative hypothesis: significant effects
```

```r
# Heteroskedasticity
lmtest::bptest(inv ~ capital + factor(firm), studentize = F,
    data = Grunfeld)
```

```
##
##  Breusch-Pagan test
##
## data:  inv ~ capital + factor(firm)
## BP = 386.81, df = 10, p-value < 2.2e-16
```

```r
# There is strong evidence for the presense of
# heteroskedasticity. Hence, the use of robust standard
# errors is advised.

# Serial Correlation
pbgtest(fe_model_plm)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  inv ~ capital + factor(firm)
## chisq = 73.785, df = 20, p-value = 4.338e-08
## alternative hypothesis: serial correlation in idiosyncratic errors
```

```r
# There is strong evidence that the residuals are serially
# correlated.
```

```r
# Clustered SE(OLS)
coeftest(pooled_ols_plm,
         vcov = vcovHC(pooled_ols_plm,
                       type = "sss",
                       # includes the small sample correction method as applied by Stata
                       cluster = "group"))
```

Clustered SE

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 14.23620   29.63751  0.4803 0.6315130
## capital      0.47722    0.13301  3.5878 0.0004203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Clustered SE(FE)
coeftest(fe_model_plm,
         vcov = vcovHC(fe_model_plm,
                       type = "sss",
                       cluster = "group"))
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## capital 0.370750   0.064946  5.7086 4.35e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Clustered SE(RE)
coeftest(re_model_plm,
         vcov = vcovHC(re_model_plm,
                       type = "sss",
                       cluster = "group"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 43.246697  37.815768  1.1436    0.2542
## capital      0.372120   0.065803  5.6551 5.389e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Differences-in-Differences: pre and post treatment and control

- Goal: Estimate effects of events or policy interventions that take place at an aggregate level.
- Advantages:
  - Policy interventions often take place at an aggregate level;
  - Aggregate /macro data are often available;
- Problems:
  - Selection of Control group is often ambiguous;
  - Standard errors do not reflect uncertainty about hte ability of the control group to reproduce the counterfactual of interest.
- The Synthetic Control Method
  -

Application: Card Krueger (1994)

```
# raw data
data_raw <- read_stata("/Users/waynezhang/Library/CloudStorage/OneDrive-UW-Madison/IOS-WIN/Econometric r
data_CK1994 <- data_raw %>%
    # chain value label
    mutate(chain = case_when(chain == 1 ~ "bk",
                             chain == 2 ~ "kfc",
                             chain == 3 ~ "roys",
                             chain == 4 ~ "wendys")) %>%
    # state value label
    mutate(state = case_when(state == 1 ~ "New Jersey",
                             state == 0 ~ "Pennsylvania")) %>%
    # Region dummy
    mutate(region = case_when(southj == 1 ~ "southj",
                              centralj == 1 ~ "centralj",
                              northj == 1 ~ "northj",
                              shore == 1 ~ "shorej",
                              pa1 == 1 ~ "phillypa",
                              pa2 == 1 ~ "eastonpa")) %>%
    # meals value label
```

```
    mutate(meals = case_when(meals == 0 ~ "none",
                             meals == 1 ~ "free meals",
                             meals == 2 ~ "reduced price meals",
                             meals == 3 ~ "both free and reduced price meals")) %>%
    mutate(emptot = empft + nmgrs + .5*emppt,
           # number of employee (full-time + managers + .5* part-time)
           pct_fte = empft/emptot*100
           # percentage of full-time employee
           )
```

```
# Distribution of restaurants
data_CK1994 %>%
    select(chain, state) %>%
    table() %>%
    prop.table(margin = 2) %>%
    apply(MARGIN = 2, FUN = scales::percent_format(accuracy = 0.01)) %>%
    noquote() %>%
    knitr::kable()
```

Descriptive statistics

	New Jersey	Pennsylvania
bk	41.09%	44.30%
kfc	20.54%	15.19%
roys	24.77%	21.52%
wendys	13.60%	18.99%

```
# Pre-treatment means
data_CK1994 %>%
    filter(time == 0) %>%
    group_by(state) %>%
    summarise(emptot = mean(emptot, na.rm = TRUE), pct_fte = mean(pct_fte,
        na.rm = TRUE), wage_st = mean(wage_st, na.rm = TRUE),
        hoursopen = mean(hoursopen, na.rm = TRUE)) %>%
    pivot_longer(cols = -state, names_to = "variable") %>%
    pivot_wider(names_from = state, values_from = value) %>%
    knitr::kable(digits = 2, caption = "Pre-treatment means 2/15/1992 - 3/4/1992")
```

Table 4: Pre-treatment means 2/15/1992 – 3/4/1992

variable	New Jersey	Pennsylvania
emptot	20.44	23.33
pct_fte	32.85	35.04
wage_st	4.61	4.63
hoursopen	14.42	14.53

```
# Post-treatment means
data_CK1994 %>%
    filter(time == 1) %>%
    group_by(state) %>%
```

```r
    summarise(emptot = mean(emptot, na.rm = TRUE), pct_fte = mean(pct_fte,
        na.rm = TRUE), wage_st = mean(wage_st, na.rm = TRUE),
        hoursopen = mean(hoursopen, na.rm = TRUE)) %>%
    pivot_longer(cols = -state, names_to = "variable") %>%
    pivot_wider(names_from = state, values_from = value) %>%
    knitr::kable(digits = 2, caption = "Post-treatment means 11/5/1992 - 12/31/1992")
```

Table 5: Post-treatment means 11/5/1992 – 12/31/1992

variable	New Jersey	Pennsylvania
emptot	21.03	21.17
pct_fte	35.87	30.38
wage_st	5.08	4.62
hoursopen	14.42	14.65

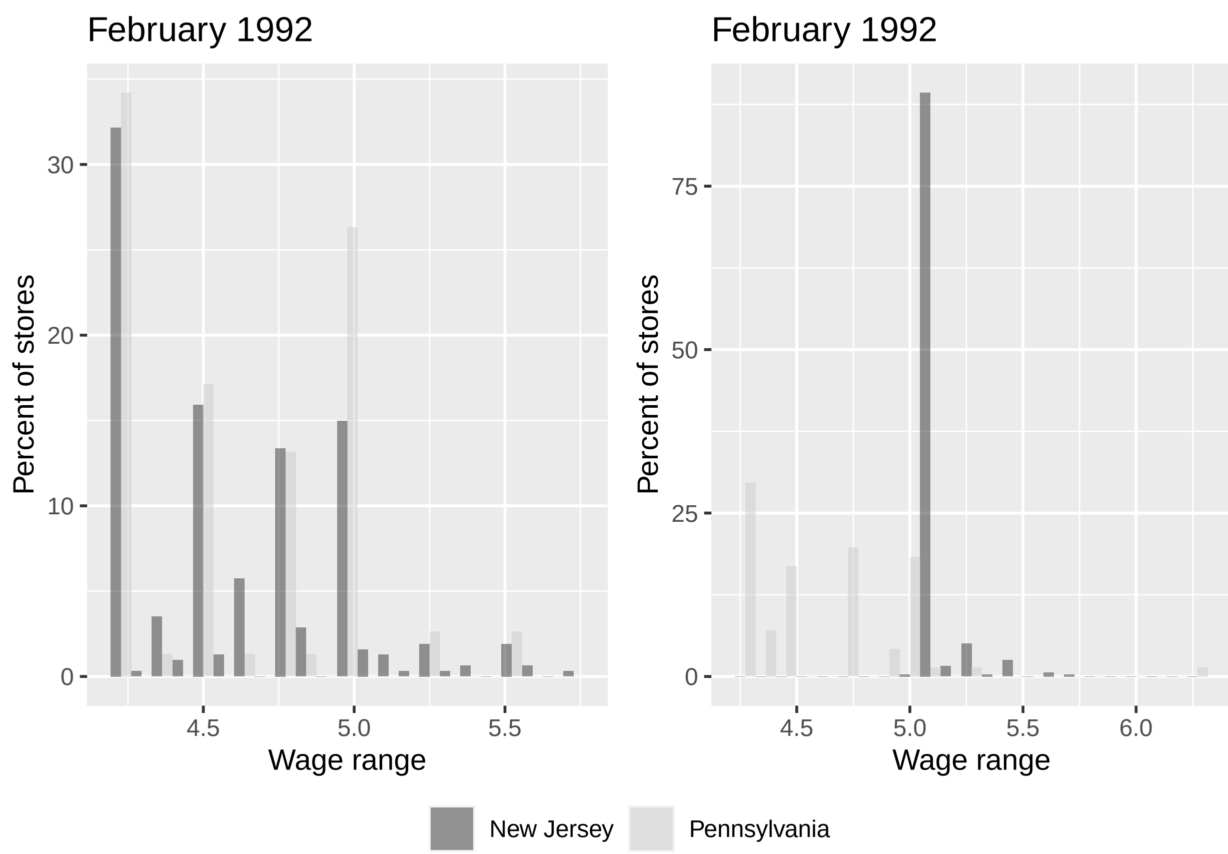
```r
# Figure 1
hist_feb <- data_CK1994 %>%
    filter(time == 0) %>%
    ggplot(aes(wage_st, fill = state)) + geom_histogram(aes(y = c(..count..[..group.. ==
    1]/sum(..count..[..group.. == 1]), ..count..[..group.. ==
    2]/sum(..count..[..group.. == 2])) * 100), alpha = 0.5, position = "dodge",
    bins = 23) + labs(title = "February 1992", x = "Wage range",
    y = "Percent of stores", fill = "") + scale_fill_grey()

hist_nov <- data_CK1994 %>%
    filter(time == 1) %>%
    ggplot(aes(wage_st, fill = state)) + geom_histogram(aes(y = c(..count..[..group.. ==
    1]/sum(..count..[..group.. == 1]), ..count..[..group.. ==
    2]/sum(..count..[..group.. == 2])) * 100), alpha = 0.5, position = "dodge",
    bins = 23) + labs(title = "February 1992", x = "Wage range",
    y = "Percent of stores", fill = "") + scale_fill_grey()
library(ggpubr)
ggarrange(hist_feb, hist_nov, ncol = 2, common.legend = TRUE,
    legend = "bottom")
```

Figure 1

```
## Warning: Removed 20 rows containing non-finite values (stat_bin).
## Removed 20 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 21 rows containing non-finite values (stat_bin).
```




```r
# First differences
diff_express <- data_CK1994 %>%
    group_by(time, state) %>%
    summarise(emptot = mean(emptot, na.rm = T)) %>%
    pivot_wider(names_from = state, values_from = emptot) %>%
    mutate(diff = `New Jersey` - Pennsylvania)
```

Calculating the treatment effect

```
## `summarise()` has grouped output by 'time'. You can override using the `.groups`
## argument.
```

```r
# The Average Treatment Effect (ATT)
diff_express$diff[2] - diff_express$diff[1]
```

```
## [1] 2.753606
```

```r
data_CK1994_mod <- data_CK1994 %>%
    mutate(treated = ifelse(state == "New Jersey", ifelse(time ==
        1, 1, 0), 0))

did_mod <- lm(emptot ~ treated + time + factor(state), data = data_CK1994_mod)
coeftest(did_mod, vcov = function(x) vcovHC(x, cluster = "group",
    type = "HC1"))
```

Calculating the DID estimator

```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                20.43941    0.50875 40.1756  < 2e-16 ***
## treated                     2.75361    1.79545  1.5337  0.12551
## time                       -2.16558    1.64121 -1.3195  0.18738
## factor(state)Pennsylvania  2.89176    1.43870  2.0100  0.04477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
panel_did <- plm(emptot ~ treated + time + factor(state), data = data_CK1994_mod,
    model = "within", index = "store")
```

```
## Warning in pdata.frame(data, index): column 'time' overwritten by time index
```

```
coeftest(panel_did, vcov = function(x) vcovHC(x, cluster = "group",
    type = "HC1"))
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## treated   2.7500     1.3359  2.0585  0.04022 *
## time2    -2.2833     1.2465 -1.8319  0.06775 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```