Workshop 3 - Solution

Treatment effect estimation

MSBX-5130: Customer Analytics

2/6/2020

1) Setup

Workshop description

- We will calculate average treatment effects (ATEs) using two methods (and datasets):
  - Using randomized control trial data (`RCT_data.csv`)
  - Using observational panel data (`DiD_data.csv`)

Workshop task workflow

1. Setup
2. Average treatment effect (ATE) estimation with Randomized Controlled Trials (RCTs)
    1. Summarize experiment
    2. Randomization checks
    3. Basic ATE estimation
    4. ROI calculation
    5. Adding pre-treatment controls
3. Average treatment effect (ATE) estimation with observational panel data
    1. Basic Differences in Differences
    2. Differences in Differences with full panel controls
    3. Prediction with plm()
        1. plm() R^2 with fixed effects

2) Average treatment effect (ATE) estimation with Randomized Controlled Trials (RCTs)

Here we will analyze the effectiveness of an online display ad campaign.

Imagine you are the marketing manager at Nordsaksingdale's, an apparel retailer. You worked with an ad exchange to run an experiment to measure the effectiveness of an online display ad campaign. The experiment was designed such that treated users saw the campaign ad, and control users saw a public service announcement (PSA). You must now analyze the resulting dataset: "RCT_data.csv" (Canvas). The data contain the following variables:

`treatment`	1 if user is in Treatment group & 0 if user is in Control group

`sales`	Sales at the retailer in 2 weeks of campaign ($)
`past_sales`	Sales at the retailer in 2 weeks prior to the campaign ($)
`female`	1 if user is female, 0 otherwise

2.1) Summarize experiment

2.1.1) Load the data from `RCT_data.csv` to a dataframe called `RCT_DF`

```
RCT_DF = read.csv("RCT_data.csv")
```

2.1.2) Use `describe()` from the `psych` package to summarize dataframe `RCT_DF`

```
library(psych)
describe(RCT_DF)
```

```
           vars      n mean   sd median trimmed mad min   max range skew
treatment     1 418172 0.50 0.50      1    0.50   0   0  1.00  1.00 0.00
sales         2 418172 0.87 2.71      0    0.09   0   0 55.35 55.35 3.93
past_sales    3 418172 0.58 1.89      0    0.03   0   0 34.03 34.03 4.06
female        4 418172 0.41 0.49      0    0.38   0   0  1.00  1.00 0.39
           kurtosis se
treatment     -2.00  0
sales         19.88  0
past_sales    20.70  0
female        -1.85  0
```

How many users (consumers) are represented in the study?

There are a total of 418,172 users in the data.

What fraction of users are treated? What fraction are female?

50% of users are treated; 41% are female.

What is the are average sales per user? What does the median value tell us?

Average sales per user are $0.87. The fact that the median is 0 tells us that most observations of sales are zero – i.e., sales (of any value) are rare events.

2.1.3) How many users are in the test and control conditions?

```
sum(RCT_DF$treatment == 0)
```

```
[1] 208866
```

```
sum(RCT_DF$treatment == 1)
```

```
[1] 209306
```

There are 208,866 users in the control condition and 209,306 users in the test condition.

2.2) Randomization checks

Next we perform a series of tests to verify that the experiment has been properly randomized. In this case, we have two pre-treatment observed variables: `past_sales` and `female`.

2.2.1) Randomization check of `past_sales`

First, compare the mean of `past_sales` across treatment groups.

```
describe(RCT_DF$past_sales[RCT_DF$treatment==0])
```

```
   vars      n mean   sd median trimmed mad min   max range skew kurtosis se
X1    1 208866 0.58 1.89      0    0.03   0   0 34.03 34.03 4.08    21.17  0
```

```
describe(RCT_DF$past_sales[RCT_DF$treatment==1])
```

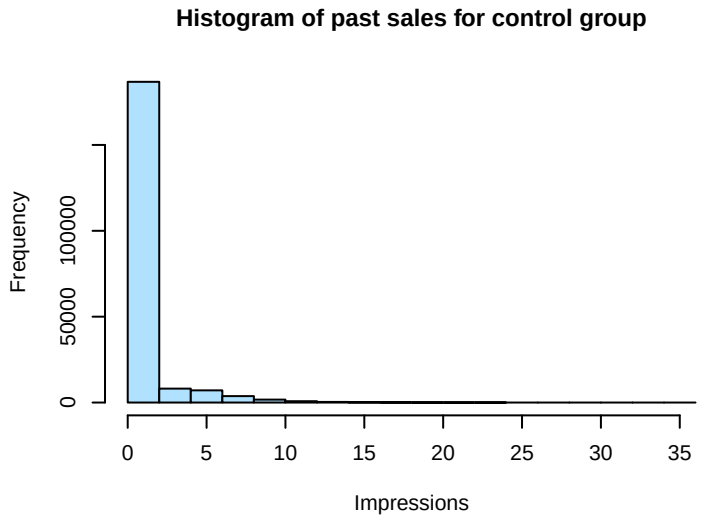
```
   vars      n mean  sd median trimmed mad min   max range skew kurtosis se
X1    1 209306 0.59 1.9      0    0.03   0   0 33.11 33.11 4.04    20.24  0
```

How similar are the means across treatment groups?

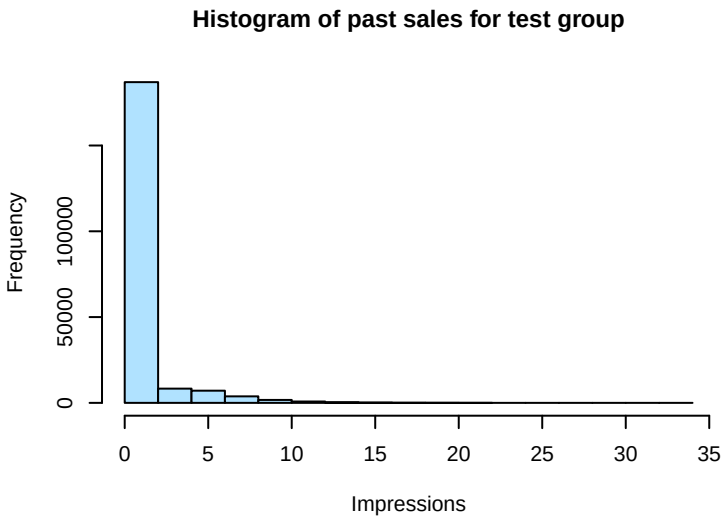
The mean of `past_sales` in the control condition is 0.58 and 0.59 in the treatment condition. These numbers look pretty similar.

Next, compare the distributions (histograms) of `past_sales` for the treament group with the control group.

```
par(cex = 0.65)
hist(RCT_DF$past_sales[RCT_DF$treatment==0],
     col  = "lightskyblue1",
     main = "Histogram of past sales for control group",
     xlab = "Impressions")
```




```
hist(RCT_DF$past_sales[RCT_DF$treatment==1],
     col  = "lightskyblue1",
     main = "Histogram of past sales for test group",
     xlab = "Impressions")
```



Do the histograms appear similar?

Yes, the histograms look quite similar.

Finally, perform a formal test of the difference in means of `past_sales` across treatment groups.

Hint: recall that `t.test(y~ind_subpop)` can be used to test if the means of `y` are different across the subsets of observations indicated by the binary variable `ind_subpop`.

```
t.test(past_sales ~ treatment, data = RCT_DF)
```

```
    Welch Two Sample t-test

data:  past_sales by treatment
t = -1.0033, df = 418168, p-value = 0.3157
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01733415  0.00559639
sample estimates:
mean in group 0 mean in group 1
      0.5811096       0.5869784
```

Are the means of `past_sales` statistically different at the 95% confidence level?

The t-test tells us that the difference in past sales is not significant across groups.

2.2.2) Randomization check of `female`

First, compare the mean of `female` across treatment groups.

```
describe(RCT_DF$female[RCT_DF$treatment==0])
```

```
   vars      n mean   sd median trimmed mad min max range skew kurtosis se
X1    1 208866 0.41 0.49      0    0.38   0   0   1     1 0.38    -1.85  0
```

```
describe(RCT_DF$female[RCT_DF$treatment==1])
```

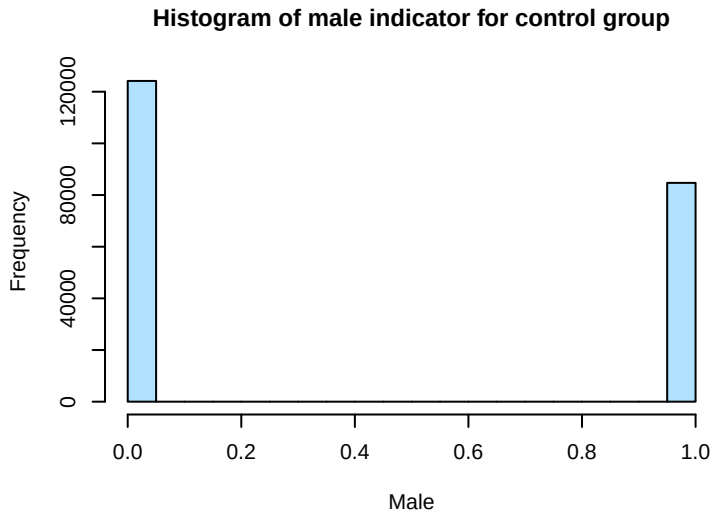
```
   vars      n mean   sd median trimmed mad min max range skew kurtosis se
X1    1 209306  0.4 0.49      0    0.38   0   0   1     1 0.39    -1.85  0
```

How similar are the means across treatment groups?

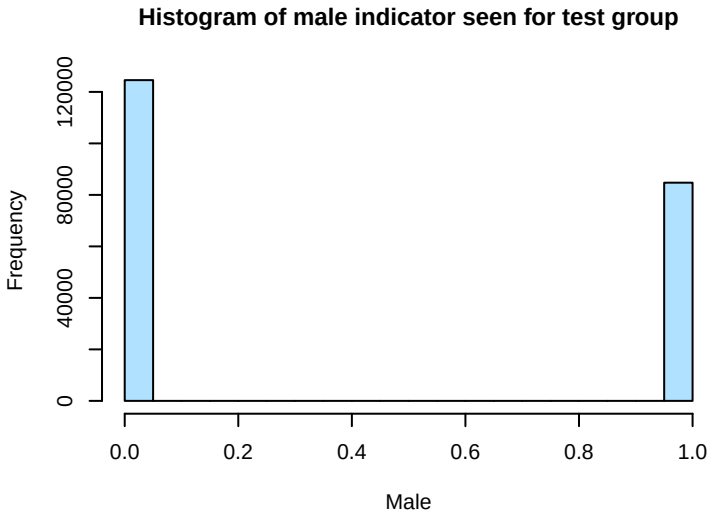
The mean of `female` in the control condition is 0.41 and 0.40 in the treatment condition. These numbers look pretty similar.

Next, compare the distributions (histograms) of `female` for the treament group with the control group.

```
par(cex = 0.65)
hist(RCT_DF$female[RCT_DF$treatment==0],
     col  = "lightskyblue1",
     main = "Histogram of male indicator for control group",
     xlab = "Male")
```




```
hist(RCT_DF$female[RCT_DF$treatment==1],
     col  = "lightskyblue1",
     main = "Histogram of male indicator seen for test group",
     xlab = "Male")
```



Do the histograms appear similar?

Yes, the histograms look quite similar.

Finally, perform a formal test of the difference in means of `female` across treatment groups.

```
t.test(female ~ treatment, data = RCT_DF)
```

```
    Welch Two Sample t-test

data:  female by treatment
t = 0.44506, df = 418168, p-value = 0.6563
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.002300184  0.003651705
sample estimates:
mean in group 0 mean in group 1
      0.4055471       0.4048713
```

Are the means of `female` statistically different at the 95% confidence level?

The t-test tells us that the difference in the proportion of female users is not significant across groups.

2.3) Average treatment effect (ATE) estimation

2.3.1) ATE estimation "by hand" (difference in means)

Calculate the ATE as: mean of the outcome (`sales`) in the treatment group, minus the mean of the outcome in the control group.

```
# ATE by hand
mean(RCT_DF$sales[RCT_DF$treatment==1])-mean(RCT_DF$sales[RCT_DF$treatment==0])
```

```
[1] 0.0328614
```

2.3.2) ATE estimation by regression

Now calculate the ATE using regression.

```
# ATE by regression
lm1 = lm(sales ~ treatment, data = RCT_DF)
summary(lm1)
```

```
Call:
lm(formula = sales ~ treatment, data = RCT_DF)

Residuals:
   Min     1Q Median     3Q    Max
-0.889 -0.889 -0.856 -0.856 54.461

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.856029   0.005931  144.34  < 2e-16 ***
treatment   0.032861   0.008383    3.92 8.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.71 on 418170 degrees of freedom
Multiple R-squared:  3.675e-05, Adjusted R-squared:  3.436e-05
F-statistic: 15.37 on 1 and 418170 DF,  p-value: 8.851e-05
```

Do your estimates of the ATE match?

Yes

What is the advantage of using regression to calculate the ATE?

First, using regression directly gives us the standard error for the treatment effect. Second, regression methods allow us to add additional controls to our estimation of the ATE, which can increase the precision of our estimates. In some cases, adding pre-treatment control variables can also help to correct biases introduced from improper randomization of RCT's.

2.4) Return on investment (ROI) calculation

Here we demonstrate how to use the treatment effect estimate to evaluate the profiability of the ad campaign.

To perform the ROI calculation, we need additional information:

1. Average profit margin on products sold = 40%
2. Average cost for ad campaign impressions = $5 CPM ($5 per 1000 impressions)

2.4.1) Calculate incremental revenue

The first step to calculating ROI is to calculate the incremental revenue from the campaign. Since the treatment effect estimate measures the incremental effect of the campaign on sales, we can estimate the incremental revenue as:

incremental revenue = (# of treated users)*(treatment effect estimate)

Calculate the incremental revenue – store the result as `inc_rev` and print its value:

```r
# calculate inc rev as the treatment effect (on sales) times the # treated users
N_treat = sum(RCT_DF$treatment)
ate = as.numeric(lm1$coefficients["treatment"])
inc_rev = N_treat*ate
inc_rev
```

```
[1] 6878.087
```

2.4.2) Calculate incremental profit

To assess the financial return of the campaign, we need to know the incremental profit of the campaign. So, we need to translate incremental revenues to incremental profits. We can do this using the average profit margin, which tells us how much profit the firm gets (on average) for every $1 of revenue. That is:

incremental profit = (incremental revenue)*(profit margin)

Calculate the incremental profit – store the result as `inc_pft` and print its value:

```r
# calculate inc profit as the inc rev times the avg profit margin (40%)
margin = 0.4
inc_pft = inc_rev*margin
inc_pft
```

```
[1] 2751.235
```

2.4.3) Calculate campaign costs

Next, we need to calculate the cost of running the campaign. We are given the cost per (1000) impressions as $5. The campaign cost is then given by:

campaign cost = (# impressions)/1000*CPM

Calculate the campaign cost – store the result as `cost` and print its value:

```r
# calc campaign costs as # impressions/1000*CPM
cpm = 5
N_imps = dim(RCT_DF)[1]
cost = (N_imps/1000)*cpm
cost
```

```
[1] 2090.86
```

2.4.4) Calculate ROI

Finally, calculate ROI as:

100*(inc_pft-cost)/cost

and print the result

```r
# calc roi (in %) as 100*(inc_pft-cost)/cost
roi = 100*(inc_pft-cost)/cost
roi
```

```
[1] 31.58389
```

Was the campaign profitable?

Yes, 30%+ return is an excellent ROI.

How do we interpret the ROI estimate?

We interpret as: on average, $1 invested in the email camapaign returns $1.32 in profit to the firm.

2.5) Adding pre-treatment controls

For a properly designed RCT, it is unnecessary to control for pre-treatment variables (such as `past_sales` and `female`). However, including such factors in the ATE regression can improve the precision of our ATE estimate.

Estimate the ATE using a regression that include `past_sales` and `female`.

```
# w/ covariates -- efficicency
lm2 = lm(sales ~ treatment + past_sales + female, data = RCT_DF)
summary(lm2)
```

```
Call:
lm(formula = sales ~ treatment + past_sales + female, data = RCT_DF)

Residuals:
   Min     1Q Median     3Q    Max
-1.083 -1.016 -0.983 -0.667 54.651

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.982969   0.007014 140.146  < 2e-16 ***
treatment    0.032634   0.008369   3.899 9.64e-05 ***
past_sales   0.002304   0.002216   1.040    0.298
female      -0.316310   0.008537 -37.053  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.706 on 418168 degrees of freedom
Multiple R-squared:  0.003332,  Adjusted R-squared:  0.003325
F-statistic:   466 on 3 and 418168 DF,  p-value: < 2.2e-16
```

Is the ATE estimate here statistically consistent with the prior estimate (without controls)?

Yes – the 95% confidence intervals have considerable overlap.

Is the standard error of the ATE estimate smaller than the prior estimate (without controls)?

Yes, but not by much in this case.

3) Average treatment effect (ATE) estimation with observational panel data

Here we will analyze the effectiveness of an end-aisle promotional display for laundry detergent. The objective is to measure the sales lift from the more prominent placement of the product in grocery stores.

A field experiment was conducted, where stores were randomly assigned to receive the treatment (promotional display) or not. Store sales (and other varibles) were observed for both treatment and control groups for 5 months (Jan-May) before the experiment began, and for the 7 months (Jun-Dec) that the end-aisle displays were placed in stores – i.e., the post-treament period spans months 6-12 (Jun-Dec).

Though stores are randomly assigned for treatment, many aspects of the environment are not controlled (local store pricing policy, other store promotional activity, etc.). In such cases, we have omitted variable concerns, and should implement strong controls (e.g. panel fixed effects) to reduce omitted variable bias when we estimate the ATE.

The field study dataset is "DiD_data.csv" (Canvas). The data contain the following variables:

`store`	Store id number
`month`	Month number
`sales`	Tide 128oz laundry detergent: unit sales
`price`	Tide 128oz laundry detergent: price ($)
`treatment`	= 1 if the store is in the treatment (promotion) condition
`post`	= 1 if the month if the treatment has been appilied

Load in the data file ("DiD_data.csv") into a dataframe called `DiD_DF`.

```
DiD_DF = read.csv("DiD_data.csv")
```

Next, create a binary numeric variable that defines the post-treatment period, called `post`, and insert it into the dataframe `DiD_DF`.

```
DiD_DF$post = as.numeric(DiD_DF$month>=6)
```

3.1) Basic Differences in Differences

Basic differences in differences tests pre/post treatment differences in outcomes across the treatment and control groups (hence the name). The simplest way to obtain the treatment effect estimate is through a regression model. In the regression setup, the treatment effect is associated with the interaction of the treatment group indicator and the post-treatment time indicator.

Estimate the basic DiD model, i.e., regress `sales` on `treatment`, `post`, and their interaction.

```
did1 = lm(sales~treatment + post + treatment:post, data=DiD_DF)
summary(did1)
```

```
Call:
lm(formula = sales ~ treatment + post + treatment:post, data = DiD_DF)

Residuals:
    Min      1Q  Median      3Q     Max
-67.934 -14.215   0.078  13.690  71.646

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     118.888      0.935 127.158  < 2e-16 ***
treatment         1.407      1.322   1.064    0.287
```

```
post              8.069      1.224   6.591 5.35e-11 ***
treatment:post    9.788      1.731   5.654 1.76e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.91 on 2396 degrees of freedom
Multiple R-squared:  0.1197,    Adjusted R-squared:  0.1186
F-statistic: 108.6 on 3 and 2396 DF,  p-value: < 2.2e-16
```

What is your ATE estimate (number and interpretation)? What is its standard error?

The ATE estimate is 9.788, meaning the promotion (on average) leads to (causes) store sales to increase by 9.788 units. The standard error is 1.731 units.

3.2) Differences in Differences with full panel controls

The basic Diff-in-Diff model controls for selection bias by allowing the treatment/control groups to have different mean outcomes (though the `treatment` parameter estimate). By replacing the `treatment` variable with individual fixed effects, we obtain much stronger (individual-specific) controls for omitted variables that could influence the ATE estimate. Similarly, replacing the `post` variable with time (month) fixed effects removes many omitted factors associated with time periods beyond the simple pre/post distinction.

Estimate the differences in differences model with full panel controls – i.e., regress sales on the `treatmentXpost` interaction, price, and *include both store and time fixed effects* (I recommend using `plm()` for this, but `lm()` will work).

```
library(plm)
did2 = plm(sales~treatment:post + price + factor(month), data=DiD_DF,
              index=c("store","month"), model="within", effect="individual")
summary(did2)
```

```
Oneway (individual) effect Within Model

Call:
plm(formula = sales ~ treatment:post + price + factor(month),
    data = DiD_DF, effect = "individual", model = "within", index = c("store",
        "month"))

Balanced Panel: n = 200, T = 12, N = 2400

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-56.45040 -12.99250   0.16998  12.44866  58.53109

Coefficients:
                 Estimate Std. Error t-value  Pr(>|t|)
price           -22.88028    2.98767 -7.6582 2.811e-14 ***
factor(month)2    0.29462    2.00403  0.1470 0.8831343
factor(month)3    3.43754    2.00476  1.7147 0.0865439 .
factor(month)4    5.22084    2.00755  2.6006 0.0093687 **
factor(month)5    8.80584    2.01795  4.3637 1.338e-05 ***
factor(month)6    7.88303    2.18332  3.6106 0.0003124 ***
factor(month)7    9.68636    2.19055  4.4219 1.026e-05 ***
```

```
factor(month)8   11.46844    2.19767  5.2185 1.974e-07 ***
factor(month)9   14.56261    2.20593  6.6016 5.090e-11 ***
factor(month)10  16.36271    2.20801  7.4106 1.787e-13 ***
factor(month)11  17.26859    2.21374  7.8007 9.460e-15 ***
factor(month)12  18.94822    2.23222  8.4885 < 2.2e-16 ***
treatment:post    9.29447    1.65887  5.6029 2.374e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1033700
Residual Sum of Squares: 876350
R-Squared:      0.15221
Adj. R-Squared: 0.070028
F-statistic: 30.2038 on 13 and 2187 DF, p-value: < 2.22e-16
```

Is the ATE estimate here statisticallly consistent with the prior estimate (without panel fixed effects)? Is the ATE estimate here more precise than the prior estimate (without panel fixed effects)?

Here the estimate is 9.294 with a standard error of 1.659. This estimate is more precise and is statistically consistent with the basic diff-in-diff estimate.

3.3) Prediction with plm()

Here I demonstrate how to do prediction with `plm()`.

The code below demonstrates the equivalence of model predictions using `plm()` and `lm()` with individual fixed effects. It also shows that the R^2 values are equivalent when we include fixed effects in the `plm()` prediction.

```
# load the data
DiD_DF = read.csv("DiD_data.csv")
DiD_DF$post = as.numeric(DiD_DF$month>=6)

# estimate the DiD model with panel fixed effects, using plm()
did2 = plm(sales~treatment:post + price + factor(month), data=DiD_DF,
              index=c("store","month"), model="within", effect="individual")

# extract individual fixed effects from plm, put in dataframe
# first column contains individual id's - named store to merge with DiD_DF
fe_DF = data.frame(store = names(fixef(did2)), fe = as.numeric(fixef(did2)))

# match fixed effect (by store id)  -- afterward, fixed effects are matched to observations
merged_DF = merge(DiD_DF,fe_DF, by="store")

# get the model (X) matrix: use same formula as used to estimate the plm() model (here, did2)
X = model.matrix(sales~treatment:post + price + factor(month), data=DiD_DF)

# full prediction using matrix algebra
# note: [,-1] excludes first column (intercept) from entering prediction (replaced by fixed effects)
did2.yhat = as.numeric(merged_DF$fe + X[,-1]%*% did2$coefficients)

# compute R^2 with fixed effects included
print("plm model with fixed effects R^2 :")
```

```
[1] "plm model with fixed effects R^2 :"
```

```
cor(did2.yhat,DiD_DF$sales)^2
```

```
[1] 0.2633186
```

```
# estimate comparable lm() model
did3 = lm(sales~treatment:post + price + factor(month) + factor(store), data=DiD_DF)
# print the R^2 value
did3.sum = summary(did3)
print("lm model with fixed effects R^2:")
```

```
[1] "lm model with fixed effects R^2:"
```

```
did3.sum$r.squared[1]
```

```
[1] 0.2633186
```

```
# predict values
did3.yhat = predict(did3)

# compare plm() predictions to lm() predictions
print("test that lm() and plm() predictions are equal:")
```

```
[1] "test that lm() and plm() predictions are equal:"
```

```
all.equal(did2.yhat,did3.yhat,check.names=FALSE)
```

```
[1] TRUE
```