DiD Regression

Danny Moncada monca016

March 17, 2020

```
suppressWarnings(suppressPackageStartupMessages({
library(dplyr)
library(ggplot2)
library(stargazer)
library(readr)
library(plm)
}))
```

```
#### Load the data ####
MyData = read.csv("TSTV-Obs-Dataset.csv")


#how long is the period of observation?
max(MyData$week)-min(MyData$week)
```

```
## [1] 13
```

```
#How many subjects got TSTV? (Treated)
length(unique(MyData$id[MyData$premium==TRUE]))
```

```
## [1] 8348
```

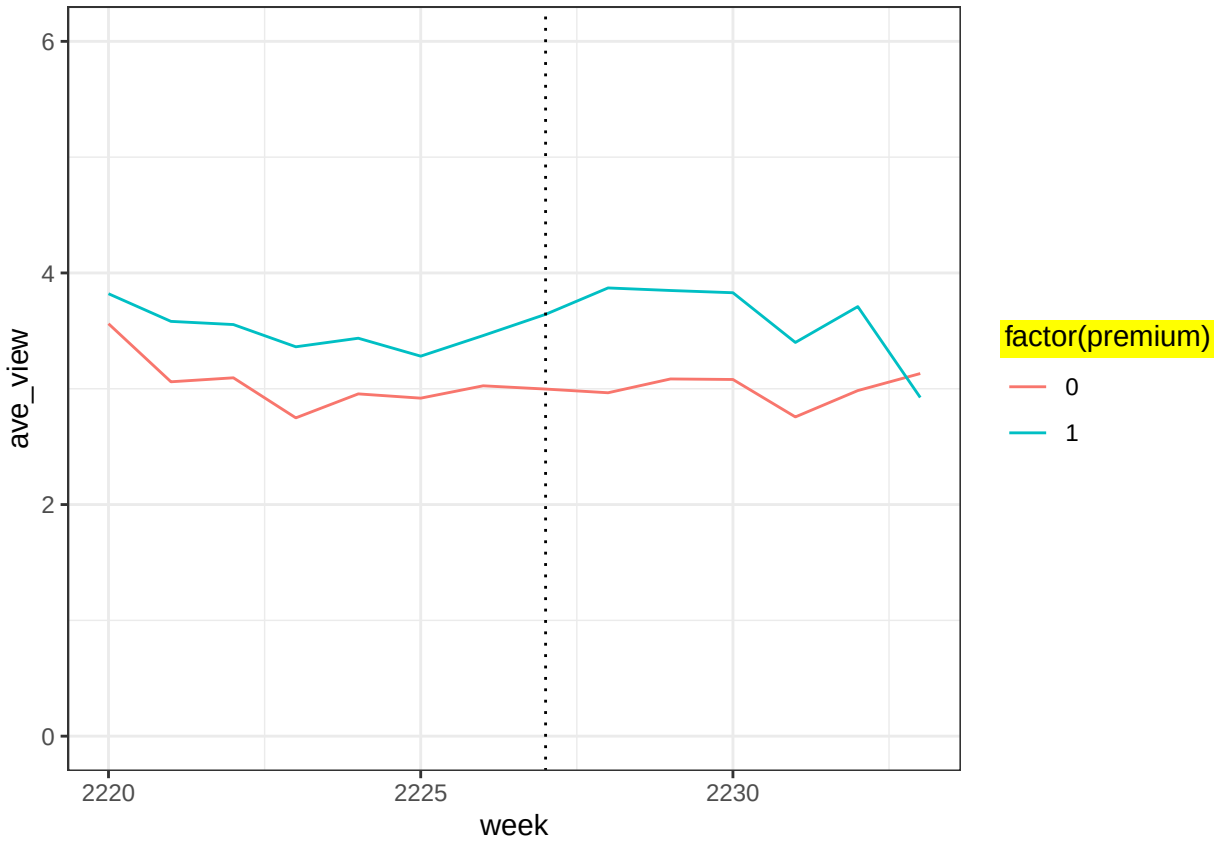
```
#How many subjects did not get TSTV? (Control)
length(unique(MyData$id[MyData$premium==FALSE]))
```

```
## [1] 41686
```

```
#In what 'week' does the "treatment" begin?
min(unique(MyData$week[MyData$after==TRUE]))
```

```
## [1] 2227
```

```
# As descriptive visualization, let's look at average weekly viewership for both premium and regular vi
Week_Ave = MyData %>% group_by(week, premium) %>%
  summarise(ave_view = mean(view_time_total_hr)) %>% ungroup()
ggplot(Week_Ave, aes(x = week, y = ave_view, color = factor(premium))) +
  geom_line() +
  geom_vline(xintercept = 2227, linetype='dotted') +
  ylim(0, 6) + xlim(2220,2233) +
  theme_bw()
```




```
#### Difference in Differences Regression ####
# Interpret the treatment effect
did_basic = lm(log(view_time_total_hr+1) ~ premium*after, data=MyData)
summary(did_basic)
```

```
##
## Call:
## lm(formula = log(view_time_total_hr + 1) ~ premium * after, data = MyData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28421 -0.69919  0.07235  0.63026  2.05054
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.122544   0.001491  752.67   <2e-16 ***
## premium        0.116126   0.003613   32.14   <2e-16 ***
## after         -0.029016   0.002094  -13.86   <2e-16 ***
## premium:after  0.074558   0.005042   14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7695 on 652795 degrees of freedom
## Multiple R-squared:  0.006149,   Adjusted R-squared:  0.006145
## F-statistic:  1346 on 3 and 652795 DF,  p-value: < 2.2e-16
```

```r
# Let's try replacing the treatment dummy with subject fixed effects.
# What happened to the estimate of premium?
did_fe = plm(log(view_time_total_hr+1) ~ premium*after, data = MyData, index=c("id"), effect="individual
summary(did_fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(view_time_total_hr + 1) ~ premium * after,
##     data = MyData, effect = "individual", model = "within", index = c("id"))
##
## Unbalanced Panel: n = 50034, T = 1-14, N = 652799
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -2.583793 -0.252482  0.016201  0.296623  2.358606
##
## Coefficients:
##                 Estimate Std. Error t-value  Pr(>|t|)
## after         -0.0096263  0.0013662 -7.0462  1.84e-12 ***
## premium:after  0.0668180  0.0032670 20.4521 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    147680
## Residual Sum of Squares: 147580
## R-Squared:      0.00069803
## Adj. R-Squared: -0.082253
## F-statistic: 210.52 on 2 and 602763 DF, p-value: < 2.22e-16
```

```r
# Further add week fixed effects
did_sfe_tfe = plm(log(view_time_total_hr+1) ~ premium*after, data = MyData, index=c("id", "week"), effe
summary(did_sfe_tfe)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(view_time_total_hr + 1) ~ premium * after,
##     data = MyData, effect = "twoway", model = "within", index = c("id",
##         "week"))
##
## Unbalanced Panel: n = 50034, T = 1-14, N = 652799
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -2.594527 -0.252892  0.017542  0.295771  2.273132
##
## Coefficients:
##                Estimate Std. Error t-value  Pr(>|t|)
## premium:after 0.0682979  0.0032553   20.98 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    146610
```

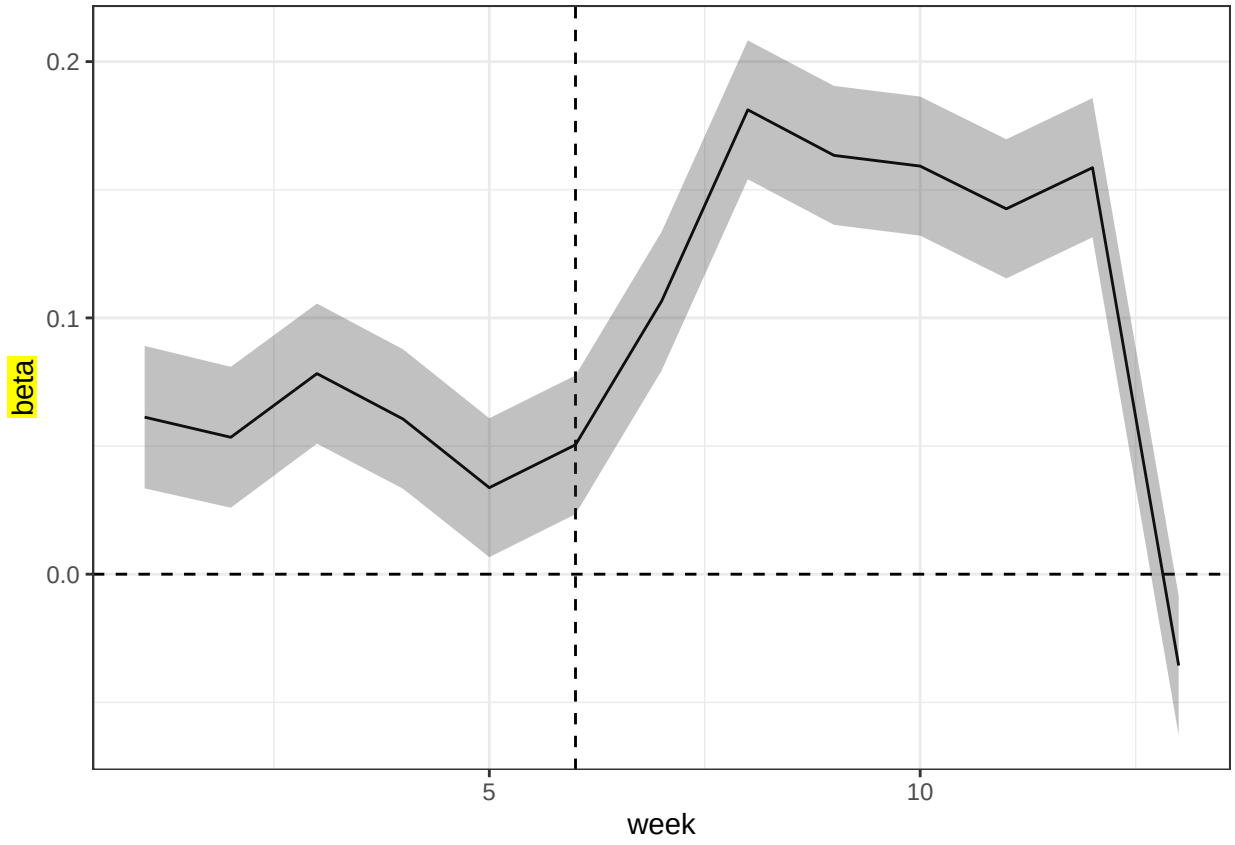
```
## Residual Sum of Squares: 146510
## R-Squared:      0.00072974
## Adj. R-Squared: -0.082241
## F-statistic: 440.172 on 1 and 602751 DF, p-value: < 2.22e-16
```

```
# Let's try dynamic DiD instead.
did_dyn_sfe_tfe <- lm(log(view_time_total_hr+1) ~ premium + factor(week) + premium*factor(week), data =
summary(did_dyn_sfe_tfe)
```

```
##
## Call:
## lm(formula = log(view_time_total_hr + 1) ~ premium + factor(week) +
##     premium * factor(week), data = MyData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35401 -0.70039  0.06861  0.62780  2.03182
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.284204   0.004361 294.479  < 2e-16 ***
## premium                      0.065613   0.010272   6.388 1.68e-10 ***
## factor(week)2221            -0.116254   0.005927 -19.614  < 2e-16 ***
## factor(week)2222            -0.131373   0.005858 -22.428  < 2e-16 ***
## factor(week)2223            -0.241444   0.005830 -41.416  < 2e-16 ***
## factor(week)2224            -0.199708   0.005810 -34.372  < 2e-16 ***
## factor(week)2225            -0.216551   0.005794 -37.375  < 2e-16 ***
## factor(week)2226            -0.185174   0.005794 -31.958  < 2e-16 ***
## factor(week)2227            -0.176850   0.005803 -30.474  < 2e-16 ***
## factor(week)2228            -0.193413   0.005814 -33.266  < 2e-16 ***
## factor(week)2229            -0.159218   0.005825 -27.336  < 2e-16 ***
## factor(week)2230            -0.162324   0.005835 -27.818  < 2e-16 ***
## factor(week)2231            -0.260881   0.005847 -44.621  < 2e-16 ***
## factor(week)2232            -0.192753   0.005860 -32.896  < 2e-16 ***
## factor(week)2233            -0.190512   0.005875 -32.426  < 2e-16 ***
## premium:factor(week)2221     0.061274   0.014178   4.322 1.55e-05 ***
## premium:factor(week)2222     0.053423   0.014022   3.810 0.000139 ***
## premium:factor(week)2223     0.078268   0.013944   5.613 1.99e-08 ***
## premium:factor(week)2224     0.060519   0.013882   4.360 1.30e-05 ***
## premium:factor(week)2225     0.033752   0.013828   2.441 0.014651 *
## premium:factor(week)2226     0.050495   0.013813   3.656 0.000257 ***
## premium:factor(week)2227     0.106577   0.013818   7.713 1.23e-14 ***
## premium:factor(week)2228     0.181185   0.013827  13.104  < 2e-16 ***
## premium:factor(week)2229     0.163413   0.013834  11.813  < 2e-16 ***
## premium:factor(week)2230     0.159237   0.013841  11.505  < 2e-16 ***
## premium:factor(week)2231     0.142558   0.013850  10.293  < 2e-16 ***
## premium:factor(week)2232     0.158616   0.013857  11.446  < 2e-16 ***
## premium:factor(week)2233    -0.035631   0.013867  -2.569 0.010186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7673 on 652771 degrees of freedom
## Multiple R-squared:  0.01174,    Adjusted R-squared:  0.0117
## F-statistic: 287.2 on 27 and 652771 DF,  p-value: < 2.2e-16
```

```r
# Let's retrieve the coefficients and standard errors, and create confidence intervals
model = summary(did_dyn_sfe_tfe)
coefs_ses = as.data.frame(model$coefficients[16:28,c("Estimate", "Std. Error")])
colnames(coefs_ses) = c("beta", "se")
coefs_ses = coefs_ses %>%
  mutate(ub90 = beta + 1.96*se,
         lb90 = beta - 1.96*se,
         week = 1:nrow(coefs_ses))

# Let's connect the estimates with a line and include a ribbon for the CIs.
ggplot(coefs_ses, aes(x = week, y = beta)) +
  geom_line() +
  geom_hline(yintercept=0,linetype="dashed") +
  geom_vline(xintercept=6,linetype="dashed") +
  geom_ribbon(aes(ymin = lb90, ymax = ub90), alpha = 0.3) +
  theme_bw()
```




```r
# Time for our placebo test...
# Let's limit to pre-period data, and shift the treatment date back in time, artificially, and see if w
# Again, recall first week when treatment starts
MyDataPre <- MyData[MyData$after==0,]
max(MyDataPre$week)
```

```
## [1] 2226
```

```r
MyDataPre$after <- MyDataPre$week > 2224
did_log_basic_placebo <- lm(data=MyDataPre,log(view_time_total_hr+1)~premium+after+premium*after)
```

```r
summary(did_log_basic_placebo)
```

```
##
## Call:
## lm(formula = log(view_time_total_hr + 1) ~ premium + after +
##     premium * after, data = MyDataPre)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25939 -0.66929  0.06303  0.61911  2.03342
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.139663   0.001763 646.302   <2e-16 ***
## premium            0.119725   0.004271  28.033   <2e-16 ***
## afterTRUE         -0.056323   0.003198 -17.610   <2e-16 ***
## premium:afterTRUE -0.011916   0.007750  -1.538    0.124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.759 on 320873 degrees of freedom
## Multiple R-squared:  0.004546,   Adjusted R-squared:  0.004536
## F-statistic: 488.4 on 3 and 320873 DF,  p-value: < 2.2e-16
```