Homework 3 Template: Diff-in-diff

Application: Online Word-of-Mouth

1. Measuring the impact of online word-of-mouth

You are trying to measure the impact of online word-of-mouth on product demand in the Chinese TV market. Specifically, you are interested in finding out whether consumers' tweets about a TV show lead to higher viewership of the show. You obtain episode-level data of ratings (market-share in terms of viewership) for a large set of TV shows as well as information on the number of tweets on Sina Weibo (the Chinese version of Twitter) mentioning the name of the show on the day on which a specific episode aired. You also have data on ratings for a set of shows in Hong Kong, where Sina Weibo has almost no market penetration because Hong Kong residents mainly use Twitter (which is blocked in mainland China). For this homework use the data-set `weibo_data.csv`.

1.1 Simple regression

QUESTION: Load the data and regress (log) ratings of each show onto the (log) number of tweets per episode. Do you think this regression gives you the causal effect of tweets on show viewership? If not, do you think your estimate will be biased upwards or downwards?

```r
# load data
df = read.csv("weibo_data.csv", header = TRUE)
# regression of ratings on tweets
mod1 = lm(log_rating~log_tweet, data = df)
summary(mod1)
```

```
##
## Call:
## lm(formula = log_rating ~ log_tweet, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54639 -0.17298 -0.05115  0.11749  1.37571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.2663894  0.0032659   81.57   <2e-16 ***
## log_tweet   0.0310302  0.0009872   31.43   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2447 on 7897 degrees of freedom
##   (3528 observations deleted due to missingness)
## Multiple R-squared:  0.1112, Adjusted R-squared:  0.1111
## F-statistic: 987.9 on 1 and 7897 DF,  p-value: < 2.2e-16
```

ANSWER: No, because it causes OVB (and even possibly reverse causality). I think it will be biased upwards. Because in general the variables have a positive relationship with ratings and number of tweets (my assumption comes from the previous description; the dataset is about “information on the number of tweets on Sina Weibo”).

1.2 Geographic Diff-in-diff

(a) During the time period of your data, the Chinese government blocked the entire Sina Weibo platform due to a political scandal for three days (a dummy for those three days called `censor_dummy` is included in the data). Assume that the censorship constitutes an exogenous shock that affected the number of tweets during the three days it lasted. You want to exploit this shock in order to analyze whether ratings decreased during the censorship.

QUESTION: Run a regression of episode-level (log) ratings on show fixed effects and the censorship dummy using only data from mainland China. Interpret the coefficient on the censorship dummy. Is this result what you expected?

```
mlchina = df[df$mainland_dummy == 1,]
# mainland china regression
library(plm)
mod2 = plm(log_rating ~ censor_dummy, data = mlchina, index = c("show_id"), model ="within", effect = "
summary(mod2)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = mlchina, effect = "individual",
##     model = "within", index = c("show_id"))
##
## Unbalanced Panel: n = 193, T = 4-198, N = 7899
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.6784533 -0.0398081 -0.0032233  0.0370703  0.6555799
##
## Coefficients:
##                Estimate Std. Error t-value Pr(>|t|)
## censor_dummy -0.0121704  0.0041407 -2.9392   0.0033 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    63.588
## Residual Sum of Squares: 63.517
## R-Squared:      0.0011199
## Adj. R-Squared: -0.023901
## F-statistic: 8.63885 on 1 and 7705 DF, p-value: 0.0033005
```

ANSWER: Within the same episode, log of rating is 0.012 less if there is censorship. Yes, because censorship would affect people’s initiatives to discuss and maybe therefore give good ratings; censorship dummy is indeed statistically significant to impact the rating of the show.

(b) QUESTION: Was it necessary to control for show fixed effects in the regression above? If you ran the regression without show fixed effects, how would the interpretation of the coefficient on the censorship dummy differ?

```r
mod3 = lm(log_rating ~ censor_dummy, data = mlchina)
summary(mod3)
```

```
##
## Call:
## lm(formula = log_rating ~ censor_dummy, data = mlchina)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34193 -0.18916 -0.06868  0.12215  1.32245
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.319650   0.003025 105.685   <2e-16 ***
## censor_dummy 0.028554   0.011568   2.468   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2595 on 7897 degrees of freedom
## Multiple R-squared:  0.000771,   Adjusted R-squared:  0.0006444
## F-statistic: 6.093 on 1 and 7897 DF,  p-value: 0.01359
```

ANSWER: Yes, controlling show fixed effects ensured that the overall impact of tweets on show ratings could be assessed. If the regression was run without the fixed effects, the imposition of censorship actually shows a positive impact on ratings - which deos not make sense and is likely seen because of unbalance sample with considerably higher number of entires for time without censorship.

(c) QUESTION: Run the same regression as in part (a), but use only data from Hong Kong (and not mainland China). Make sure to control for show fixed effects. Interpret the coefficient on the censorship dummy. Is this result what you expected?

```r
hk = df[df$mainland_dummy == 0,]
# hong kong regression
mod4 = plm(log_rating ~ censor_dummy, data = hk, index = c("show_id"), model ="within", effect = "indiv
summary(mod4)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = hk, effect = "individual",
##     model = "within", index = c("show_id"))
##
## Unbalanced Panel: n = 132, T = 2-139, N = 3528
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -1.009054 -0.069506 -0.013726  0.064527  1.030956
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## censor_dummy 0.010554   0.011067  0.9536   0.3404
##
```

```
## Total Sum of Squares:    74.515
## Residual Sum of Squares: 74.495
## R-Squared:      0.00026777
## Adj. R-Squared: -0.038603
## F-statistic: 0.909322 on 1 and 3395 DF, p-value: 0.34036
```

ANSWER: Within the same episode, log of rating is 0.011 more if there is censorship. Yes, this is exactly what I expected. Censorship dummy shouldn't have an impact on show ratings in Hong Kong because the rating platforms in the two geographies are unrelated. I also thought censorship had an effect on ratings and this result proved that the trend was going up at Hong Kong, so the downward trend was indeed caused by censorship.

(d) QUESTION: Using data from both Hong Kong and mainland China, implement a difference-in-differences regression with mainland China as the treatment group and Hong Kong as the control group. In other words, you want to show that the censorship event had a differential effect in mainland China relative to Hong Kong. Make sure to control for show fixed effects. Interpret the relevant coefficients of this regression.

```
# DinD regression
mod5 = plm(log_rating ~ censor_dummy + mainland_dummy + censor_dummy * mainland_dummy, data = df, index
summary(mod5)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy + mainland_dummy + censor_dummy *
##     mainland_dummy, data = df, effect = "individual", model = "within",
##     index = c("show_id"))
##
## Unbalanced Panel: n = 325, T = 2-198, N = 11427
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.0090536 -0.0498983 -0.0050004  0.0429003  1.0309559
##
## Coefficients:
##                              Estimate Std. Error t-value Pr(>|t|)
## censor_dummy                0.0105535  0.0083309  1.2668  0.20526
## censor_dummy:mainland_dummy -0.0227239  0.0097603 -2.3282  0.01992 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Total Sum of Squares:    138.1
## Residual Sum of Squares: 138.01
## R-Squared:      0.00066014
## Adj. R-Squared: -0.02869
## F-statistic: 3.66622 on 2 and 11100 DF, p-value: 0.025604
```

ANSWER: If censorship happens in Hong Kong, log of rating increases by 0.011 compared to when there isn't censorship. If censorship happens in mainland China, log of rating decreases by -0.023 compared to when there isn't censorship.

1.3 Across-show Diff-in-diff

From this question onward, use only observations from shows in mainland China.

(a) QUESTION: The variable `av_tweets` denotes the average number of tweets associated with an episode of each show (outside of the censored time period). Therefore, this variable is show specific, but it does not vary over time. We can use this variable to capture the general level of social media interest in each show. Generate a set of three dummy variables based on the `av_tweets` variable: The first dummy is equal to one for shows with fewer than 5 tweets per episode, the second dummy is equal to one for shows with at least 5 but less than 100 tweets per episode, and the third dummy should be equal to one for shows with at least 100 tweets per episode.

```
# define dummies
mlchina$less.5 = ifelse(mlchina$av_tweets <5 ,1 ,0)
mlchina$between.5.and.100 = ifelse((mlchina$av_tweets >= 5 & mlchina$av_tweets < 100), 1, 0)
mlchina$more.100 = ifelse(mlchina$av_tweets >= 100, 1, 0)
```

(b) QUESTION: Run three separate regressions for shows with less than 5 tweets per episode, shows with 5 to 100 tweets per episode and shows with at least 100 tweets. What do you find in terms of impact of the censorship event across the three regressions?

```
# tweet activity regressions
mod6 = plm(log_rating ~ censor_dummy, data = subset(mlchina, mlchina$less.5==1),index = c("show_id"), mo
mod7 = plm(log_rating ~ censor_dummy, data = subset(mlchina, mlchina$between.5.and.100==1),index = c("sh
mod8 = plm(log_rating ~ censor_dummy, data = subset(mlchina, mlchina$more.100==1),index = c("show_id"),
summary(mod6)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = subset(mlchina,
##     mlchina$less.5 == 1), effect = "individual", model = "within",
##     index = c("show_id"))
##
## Unbalanced Panel: n = 88, T = 4-145, N = 3405
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.2956731 -0.0310456 -0.0027421  0.0295382  0.4380734
##
## Coefficients:
##                Estimate Std. Error t-value Pr(>|t|)
## censor_dummy -0.0068928  0.0051997 -1.3256   0.1851
##
## Total Sum of Squares:    17.586
## Residual Sum of Squares: 17.577
## R-Squared:      0.00052965
## Adj. R-Squared: -0.025994
## F-statistic: 1.75724 on 1 and 3316 DF, p-value: 0.18506
```

```
summary(mod7)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = subset(mlchina,
##     mlchina$between.5.and.100 == 1), effect = "individual", model = "within",
##     index = c("show_id"))
##
## Unbalanced Panel: n = 63, T = 4-198, N = 2945
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.418531 -0.045512 -0.003777  0.039892  0.655905
##
## Coefficients:
##                Estimate Std. Error t-value Pr(>|t|)
## censor_dummy -0.0042159  0.0069139 -0.6098   0.5421
##
## Total Sum of Squares:    23.695
## Residual Sum of Squares: 23.692
## R-Squared:      0.00012905
## Adj. R-Squared: -0.021736
## F-statistic: 0.371829 on 1 and 2881 DF, p-value: 0.54206
```

```
summary(mod8)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = subset(mlchina,
##     mlchina$more.100 == 1), effect = "individual", model = "within",
##     index = c("show_id"))
##
## Unbalanced Panel: n = 42, T = 9-168, N = 1549
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.6809617 -0.0526040 -0.0016863  0.0568705  0.5604936
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## censor_dummy -0.033491   0.011431 -2.9298 0.003442 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Total Sum of Squares:    22.307
## Residual Sum of Squares: 22.18
## R-Squared:      0.0056676
## Adj. R-Squared: -0.022063
## F-statistic: 8.58402 on 1 and 1506 DF, p-value: 0.0034423
```

ANSWER: The log rating decreases by 0.0068928 for shows with less than 5 tweets per episode, decreases by 0.0042159 for those between 5 to 100 tweets, and decreases by 0.033491 for shows more than 100 tweets. Only for the last model, the coefficient is statistically significant; this means that censorship has the largest effect on shows with more than 100 tweets.

(c) QUESTION: Run a difference-in-difference regression that allows for the censorship event to have a different effect for three sets of shows with the three different activity levels defined above. Interpret the relevant coefficients.

```
# across-show diff-in-diff
mod9 = plm(log_rating ~ censor_dummy + between.5.and.100 + more.100 +  censor_dummy * between.5.and.100
summary(mod9)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy + between.5.and.100 +
##     more.100 + censor_dummy * between.5.and.100 + censor_dummy *
##     more.100, data = mlchina, effect = "individual", model = "within",
##     index = c("show_id"))
##
## Unbalanced Panel: n = 193, T = 4-198, N = 7899
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.6809617 -0.0398301 -0.0029808  0.0371398  0.6559046
##
## Coefficients:
##                                  Estimate Std. Error t-value Pr(>|t|)
## censor_dummy                   -0.0068928  0.0064818 -1.0634  0.28763
## censor_dummy:between.5.and.100  0.0026769  0.0094812  0.2823  0.77770
## censor_dummy:more.100          -0.0265985  0.0107282 -2.4793  0.01318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    63.588
## Residual Sum of Squares: 63.449
## R-Squared:      0.0021828
## Adj. R-Squared: -0.023077
## F-statistic: 5.61688 on 3 and 7703 DF, p-value: 0.00076477
```

ANSWER: The change in log rating for shows with less than 5 tweets is -0.0068928 when there is censorship. The change in log rating for shows with between 5 to 100 tweets when there is censorship is -0.0068928+0.0026769 = -0.0042159. The change in log rating for shows with more than 100 is -0.0068928-0.0265985 = -0.0334913.

(d) QUESTION: Relate your findings across shows with different activity levels to the geographic difference-in-difference approach. Which regression is more informative regarding the impact of the censorship on ratings?

ANSWER: The difference-in-difference approach is more informative regarding the impact of the censorship on ratings because it not only shows the impact of censorship but also shows the relative impact of it on different categories of shows - in this case, in terms of average number of tweets.