Homework 3

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
library(readr)
weibo = read.csv('weibo_data.csv')
# regression of ratings on tweets
summary(lm(log_rating ~ log_tweet, data = weibo))
```

```
##
## Call:
## lm(formula = log_rating ~ log_tweet, data = weibo)
##
## 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, I do not think this gives the causal effect of tweets on ratings due to omitted variables such as the actual quality of the show in question. I think this estimate is biased upwards since it doesn't account for the rating the show would have gotten without any tweets. However, there are a variety of omitted variables and thus it is unclear whether the coefficient estimate is definitively biased upwards or downwards (i.e. other arguments could be made for a downwards bias).

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?

```
library(plm)
# mainland China regression controlling for show fixed effects
summary(plm(log_rating ~ censor_dummy,
            data = weibo[weibo$location == 'Mainland China',],
            index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = weibo[weibo$location ==
##     "Mainland China", ], 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: The coefficient on `censor_dummy` suggests that, in Mainland China, the log of the ratings during a censored period tends to be 0.0121704 lower than the log of the ratings during the non-censored periods on average. This makes intuitive sense.

(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?

ANSWER: I think it was necessary to control for show fixed effects since there are a lot of variables that coincide with `show_id` such as time of show release. Controlling for show fixed effects helps reduce these confounding factors and ultimately push the coefficient closer to what its true causal effect is. While the coefficient itself would likely change if you ran the regression without show fixed effects, the interpretation of the coefficient would not change.

(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
# hong kong regression
summary(plm(log_rating ~ censor_dummy,
            data = weibo[weibo$location == 'hongkong',],
            index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = weibo[weibo$location ==
##     "hongkong", ], 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: The coefficient on `censor_dummy` suggests that, in Hong Kong, the log of the ratings during a censored period tends to be 0.010554 greater than the log of the ratings during the non-censored periods on average. However, the coefficient is not statistically significant, which is expected since the censorship only affects Sina Weibo and thus there should be no relationship between censorship and ratings in an area like Hong Kong where most people use Twitter.

(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.

```r
# DinD regression
summary(plm(log_rating ~ mainland_dummy + censor_dummy + mainland_dummy*censor_dummy,
            data = weibo, index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ mainland_dummy + censor_dummy + mainland_dummy *
##     censor_dummy, data = weibo, 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
## mainland_dummy:censor_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: `censor_dummy`: This coefficient implies that, relative to before the censorship period, the log of the ratings in Hong Kong increased by 0.0105535 on average.

`mainland_dummy:censor_dummy`: This coefficient represents the difference-in-differences estimator. Thus, the differential effect of the censorship in Mainland China measured relative to Hong Kong is a decrease of 0.0227239 in the log of the ratings.

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.

```r
# just mainland china
weibo_main = weibo[weibo$location == 'Mainland China',]
# create categorical variables of bins
cuts = cut(weibo_main$av_tweets, breaks = c(0, 5, 100, max(weibo_main$av_tweets)+1),
           include.lowest = TRUE, right = FALSE, labels = c('<5', '5-100', '>=100'))
weibo_main$av_tweet_bins = cuts
# separate categorical values into dummies
weibo_main$av_tweet_under5 = ifelse(weibo_main$av_tweet_bins == '<5', 1, 0)
weibo_main$av_tweet_5_to_100 = ifelse(weibo_main$av_tweet_bins == '5-100', 1, 0)
weibo_main$av_tweet_greq100 = ifelse(weibo_main$av_tweet_bins == '>=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?

```r
# tweet activity regressions
summary(plm(log_rating ~ censor_dummy, data = weibo_main[weibo_main$av_tweet_under5 == 1,],
            index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
```

```
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = weibo_main[weibo_main$av_tweet_under5 ==
##     1, ], 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(plm(log_rating ~ censor_dummy, data = weibo_main[weibo_main$av_tweet_5_to_100 == 1,],
            index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = weibo_main[weibo_main$av_tweet_5_to_100 ==
##     1, ], 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(plm(log_rating ~ censor_dummy, data = weibo_main[weibo_main$av_tweet_greq100 == 1,],
            index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy, data = weibo_main[weibo_main$av_tweet_greq100 ==
##     1, ], 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: We can see that, for shows that have less than an average of 100 tweets (i.e. the first 2 regressions), the coefficient on `censor_dummy` is statistically insignificant and is very small, implying that censorship has little effect on ratings. For shows with 100 tweets or more on average, however, the coefficient on `censor_dummy` is both statistically significant and much larger in magnitude, suggesting that censorship has a larger impact on the shows that get more frequently tweeted about.

(c) QUESTION: Run a difference-in-differences 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.

```r
summary(plm(log_rating ~ censor_dummy + censor_dummy*av_tweet_5_to_100 +
            censor_dummy*av_tweet_greq100, data = weibo_main,
            index = c('show_id'), model = 'within'))
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log_rating ~ censor_dummy + censor_dummy * av_tweet_5_to_100 +
##     censor_dummy * av_tweet_greq100, data = weibo_main, 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:av_tweet_5_to_100  0.0026769  0.0094812  0.2823  0.77770
## censor_dummy:av_tweet_greq100  -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: `censor_dummy`: This coefficient represents the baseline effect of the censorship on shows with an average number of tweets that is less than 5. More specifically, it implies that censorship corresponds to a decrease of 0.0068928 in the log of the ratings for shows with an average number of tweets that is less than 5.

`censor_dummy:av_tweet_5_to_100`: This coefficient is the difference-in-differences estimator that measures the differential effect of censorship on shows with an average number of tweets that is in the range of 5-100. It implies that censorship corresponds to an increase of 0.0026769 in the log of the ratings relative to shows with an average number of tweets that is less than 5. This is counterintuitive and thus it makes sense that this coefficient is highly insignificant.

`censor_dummy:av_tweet_greq100`: This coefficient is the difference-in-differences estimator that measures the differential effect of censorship on shows with an average number of tweets that is greater than or equal to 100. It implies that censorship corresponds to a decrease of 0.0265985 in the log of the ratings relative to shows with an average number of tweets that is less than 5, suggesting that censorship has a larger negative effect on shows that are more prominent on Twitter, as expected.

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

ANSWER: The difference-in-differences regressions show similar results in that they both show that censorship negatively affects the ratings of shows. I think the regression with different activity levels is more informative, however, since it provides more information regarding which shows are most strongly affected by this censorship and it also measures these effects within the same group (that is, the whole regression uses data from Mainland China as opposed to comparing ratings in Hong Kong to ratings from Mainland China).