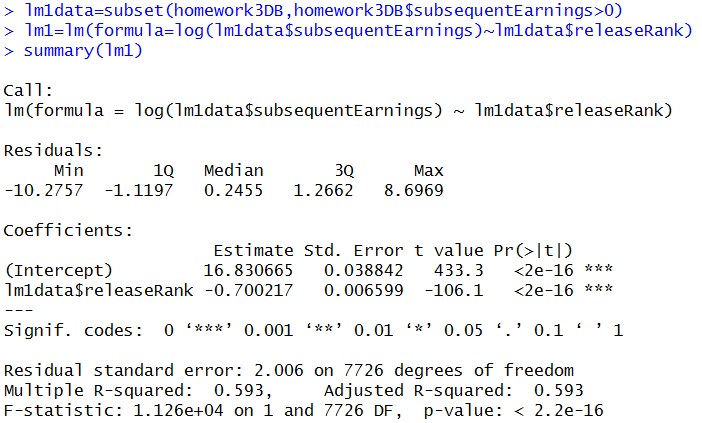
# Homework3 Suhang YAO

# Part 1: Casual Regressions (40%)

a) Run a linear regression that uses releaseRank to predict the log of subsequentEarnings. Note that log of 0 is -Infinity, and you may have to separate out those cases.

See R code

b) Summarize the regressions in paragraph form in the pdf. Display the regression output, and briefly comment the significance levels. (10%)



As the regression output shows, the coefficient value between release rank and log(subsequent earnings) is -0.700217, which means when the other factors are controlled, a change in movie rank of one unit is associated with -70% change in subsequent earnings.

P-value of T test is less than 0.05, which means the relationship between rank and subsequent earnings is significant.

c) These regressions are heavily biased. Explain, to a layperson, three different potential confounding factors that are omitted from the regression. Which do you think is the most important? What data would you need to address these factors? Answer with at least two paragraphs in the pdf. (15%)

1. Movies promotion activities will influence the subsequent earnings as well. Big movie well spend a lot of money on marketing and promotion and use famous actor, and have a better ranking and earn more money

2. Having new release movies (the competition increase)

3. The schedule of movies in cinema was different will also influence their earnings

The most important factor should be the quality between different movies

Data needed:

Cost of movies ( promotion cost, actor, director, etc.)

Schedule to play movies in cinema, in which time period, people are more willing to go see movie

d) Explain why these regressions are biased to a technical audience by describing what is in x and what is in epsilon. Write down what you think the true model might be. According to this 'true' model, what is going into the error term in the regressions in part 2a? How would that cause bias? Answer in the pdf. (15%)

Based on previous question, a better model should consider all the factors

log(subsequentEarnings)=releaseRank+advertising+ schedule of movie in cinema+ popularity of actors +error

All these factors are correlated the release rank and will change the subsequent Earnings, cause bias of coefficient higher than truth.

# Part 2: Data Cleaning (20%) ADVANCED –

See R code

# Part 3: Regression Discontinuity (40%)

a) What feature of this problem allows us to implement a regression discontinuity? Answer in a sentence or two in the pdf. (5%)

Small change in movie earnings will change movie ranking

b) Use the dataset created in part 2 to find a subset where the first and second place in the box office were close. Our threshold for 'close' is if firstWeekBoxOfficeDiff is less than .1.

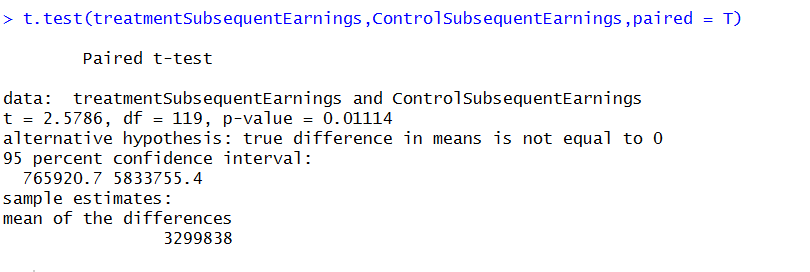
See R code

c) Consider that the 'treatment' in this case is receiving the number 1 rank. In the week of April 4th, 2008 – which movie is in the treatment group? Which movie is in the control group? Answer in the pdf. (5%)

Treatment one: Nim’s Island

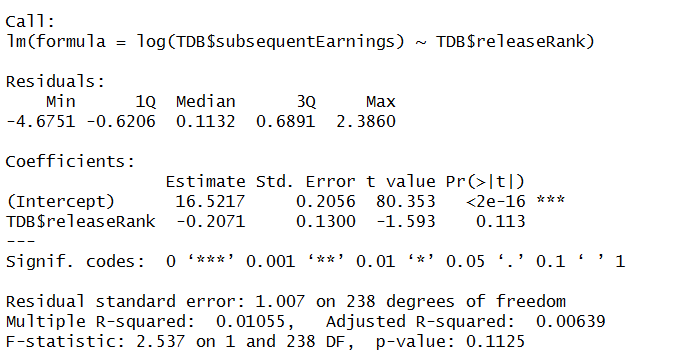
Control one: Leatherheads

d) Implement the regression discontinuity by comparing the subsequent earnings of the treatment and control in this subset. Check if the difference is significant by using the R function t.test. (10%) Hint: Look up the t.test function to find out what it does



The P-Value of paired T Test is 0.01114<0.05 , which means the difference between treatment group and control group is significant.

e) Describe this analysis in a short report to a manager (200-400 Words). In this report you should compare the result you observed here to the result in Part A – is it larger or smaller? Is this the result you expected? Why? How much additional revenue will a movie achieve by being number 1 in the box office? Describe how the discontinuity mimics random assignment since an experiment was not possible. Answer in the pdf. (20%)



The coefficient value of regression discontinuity is smaller than the part one, one umber change in rank will only cause 20%change in revenue.

Reason: In reality, majority big movies with high quality high cost are more easily to have higher release week earnings and subsequent earnings than small movies. Release week rank is not the only reason that cause subsequent earning differences. Therefore, the regression in Part A have huge bias on coefficient value. By using regression discontinuity, we remove the movies with huge release week earnings difference (the movies with huge quality difference). Only compared movie with similar quality but with different release rank to show how different release rank will influence subsequent earnings. That could give us a more accurate correlation between release week rank and subsequent earnings.

Why the discontinuity mimics random assignment since an experiment was not possible?

In this experiment, we assume that rank one movies and rank2 movies have similar qualities, and their rankings are randomly chosen. However, our sample are from real data, even though we do regression discontinuity, rank 1 movies’ release week earnings are still higher than rank2’s, So we cannot draw the conclusion that our sample data is randomly selected.