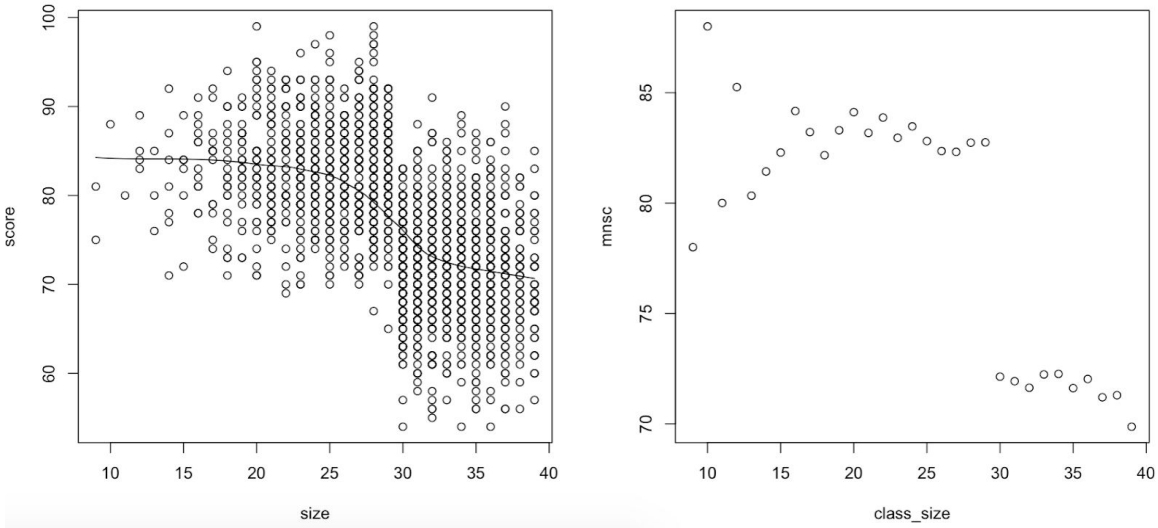
Homework 5

Team member: Nandini Basu, Linyan Dai, Maxine Li
Problem Set 1:

1. Explore the data in class.csv and see whether you think grades really do depend on the class size. Please explain the reason behind your conclusion.



From the scatter plot of class size and class average score, we could see an obvious drop with the class size increase. Then by plotting the average score of each class size, there is a clear cut-off when class size increases to 29. Therefore, we could reasonably assume that the grades depend on class size, and with a negative correlation. However, correlation does not mean causation and there could be endogeneity in the explanatory variable or a possible confounding effect.

2. Have a look at the summary statistics of the original research. Explain in detail (tell us about the steps you would follow), how you would have tackled the question of class sizes resulting in better grades using RDD.
    - Treatment
        - Class size is less or equal than 29, or class size is over 29
    - Regression model:

$$Y = a + t*D + b1*(X - c) + b2*(X - c) + e$$

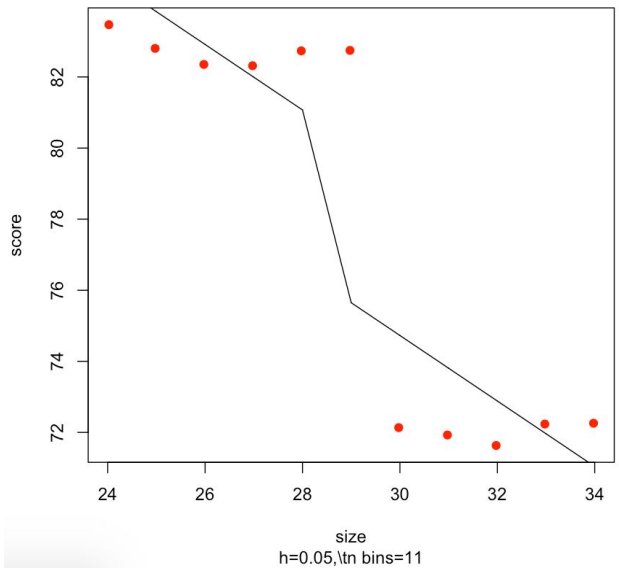
        - The c is the cut-off point, here is when class size is 29
        - D is the binary variable equal to 1 if X >= 29
    - Assumptions:

    - We assume other factors equal for both groups, including the enrollment size, the percentage of disadvantaged in each class, the reading size and math size.
    - We assume that the E[Y(0)|X=x] and E[Y(1)|X=x], are continuous in x. Also, the conditional distribution function is smooth in the covariate.
- Dataset:
    - Considering that once a class size hits 40, the school has to create a new class for the extra students. Following Maimonides's rule, there is a discontinuity in the classroom sizes as the number of students approaches 40. Hence, instead of using the full data set to check for a causal relationship between class size and test score, we use the sample which only included +5/-5 unit of class size from the treatment, class size of 29.

3. Try to use the data from question 1 and apply the procedure you described in question 2 (maybe a simple version of what you described) to estimate the effect of performance on class size? What is your conclusion?

```
> # Test treatment effect
> rdd_mod1 <- rdd_reg_lm(rdd_object = data1, slope = 'same')
> summary(rdd_mod1)

Call:
lm(formula = y ~ ., data = dat_step1, weights = weights)

Residuals:
    Min      1Q  Median      3Q     Max
-20.727  -3.994  -0.727   4.195  18.116

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.1504     0.4243 188.890  < 2e-16 ***
D            -4.5016     0.6871  -6.552 8.17e-11 ***
x            -0.9217     0.1122  -8.213 5.14e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.286 on 1302 degrees of freedom
Multiple R-squared:  0.3606,    Adjusted R-squared:  0.3596
F-statistic: 367.1 on 2 and 1302 DF,  p-value: < 2.2e-16
```

After the regression, we get the output that the estimated effect of treatment is -4.50, which is significant at the 0.001 level.



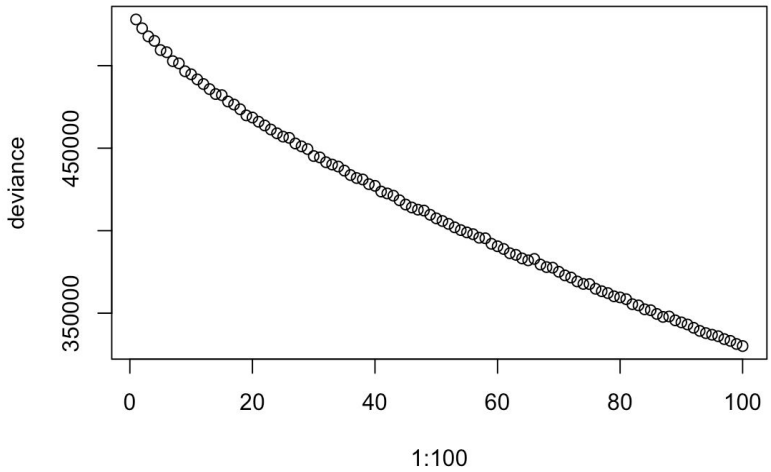
This is a visual representation of the effect as we can see the "jump" at the cutoff point.

Problem Set 2:

1. Fit K-means to the speech text of the members, comprising of the 1000 phrases, for K in 5,10,15,20,25

Please see the attached code.

2. Use BIC to choose the K and interpret the selected model. Also use the elbow curve method to identify the most optimal value of K. Compare the two values of K that you obtained. Are they equal?



The AICc is lowest for the model with 25 Ks, which means that the members can be split into 25 clusters basis the content of their speeches. The content is represented using phrase bigrams. Even when we look at the elbow model we see that the deviance decreases as the K increases hence from our options of K we will pick the highest value of K which is 25.

3. Fit a topic model for the speech counts. Use Bayes factors to choose the number of topics and interpret your chosen model.

We choose to cluster into 10 topics. Like the first topic is something about **race**, the second is probably something about **welfare**, the third is something about **illegal immigration**, fourth is about **oil**, fifth is about **retirement**, sixth is about **crime**, seventh is about **judge**, eighth is about **gun**, ninth is about **finance**, and the tenth is about **stem cell.**

4. Connect the unsupervised clusters to partisanship. Tabulate party membership by K-means cluster. Are there any non-partisan topics? Fit topic regressions for each party and repshare. Compare to regression onto phrase percentages: x <- 100 * congress109Counts / rowSums(congress109Counts)

Please see attached code for tabulations. The topics on **crime, judge** and **stem cell** have 0 coefficients, this would indicate that they are non-partisan topics. Also, looking at the graphs below we can see that for both party membership and repshare, the topic regression has lower mean-squared errors. Also for party prediction, the r-squared for the topic model is 0.5641162, which is higher than the r-squared for the bigram model. Similarly, for repshare, the r-squared for the topic model is 0.3439668 which is greater than 0.1880478, which is the r-squared for the bigram model.