Home Assignment 5

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.**

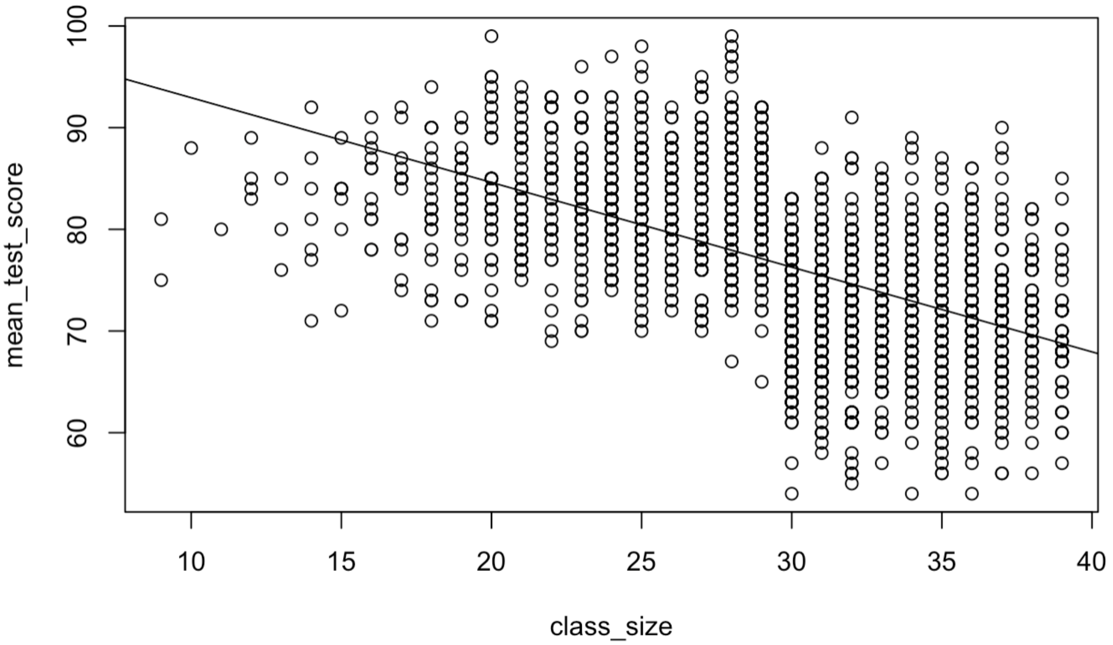


Figure: Class-size (Number of students in a class) vs Average test score

Yes, we think that grades are dependent on class size because of the following reasons -

- Firstly, by looking into the scatter plot, we see a rough negative correlation between class sizes and mean test scores. We can even see a sharp plunge at the class size of 30, where a discontinuity is detected.
- Further, we regress the mean test scores on class sizes to identify the relationship between these two variables, which is as following:

$$Mean\ test\ score = 101.3 - 0.834 * Class\ Size$$

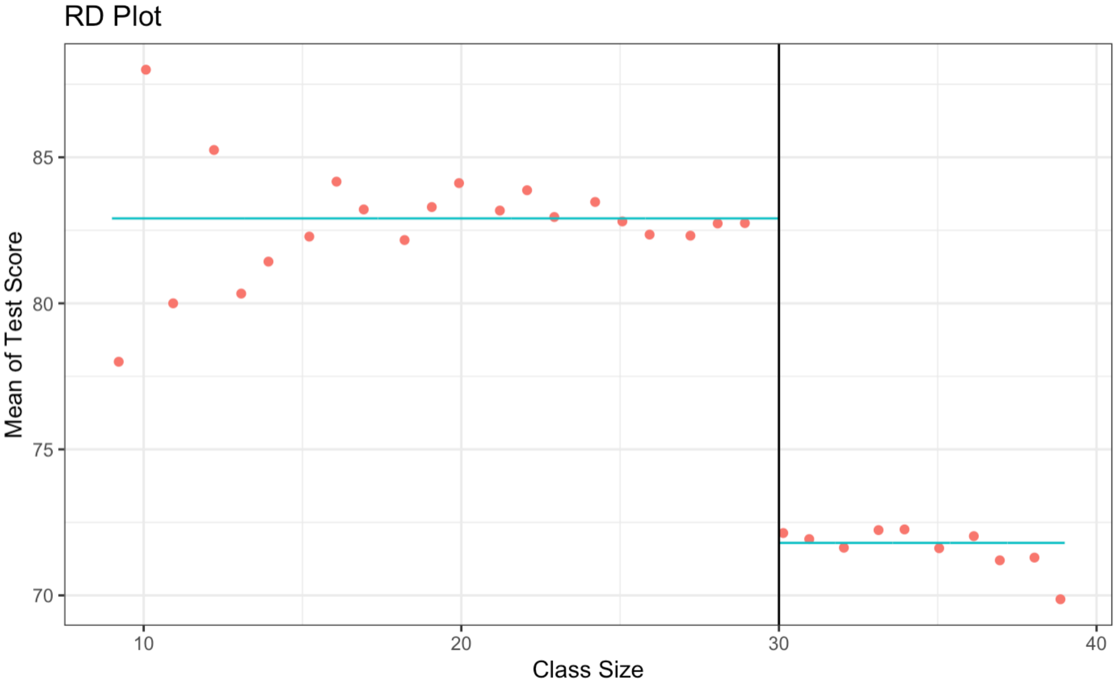
Since both intercept and class size are significant in the linear regression model, it indicates that the larger the class size is, the worse the students will perform.

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.**

We assume that the summary statistics of the original research implies a prior ability of all students in terms of their reading and math score. To make a valid RDD, observations of both sides of the cutoff line should be roughly identical on pre-treatment covariates. Therefore, only if we have more related data such as reading and math scores of different classes in each school, can we make sure prior expectations for students are roughly the same by observing the distribution of all covariates.

3. **Try to use the data from question 1 and apply the procedure you described in question 2 to estimate the effect of performance on class size? What is your conclusion?**

From summary statistics, we make an assumption that students in a class of more than 30 pupils have roughly the same prior background before enrollment as those in a class of no more than 30 pupils. Then, by plotting the mean of mean test scores of each class size, we observe a dramatic drop when class size steps over 30.



Assuming that classes with a size smaller than 30 students are given treatment (like better group discussions, more efficient use of lab equipments, or can use 'smart class'), we build a regression model:

$$Mean\ of\ mean\ test\ score = 82.48 - \ 10.41 * \ Treatment - \ 0.08 * Class\ Size$$

Both the class size and treatment effect are significant, which means while holding all other variables constant, a class with more than 30 students is expected to decrease the average mean test score by 10.41.

<u>Problem Set 2:</u>

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

We use the K-means function, after scaling the matrix, to fit the 1000 phrases into k=5, 10, 15, 20, 25

Number of Clusters	Cluster sizes (Number of congressmen in each cluster)
5	3, 7, 1, 74, 444
10	1, 420, 6, 2, 5, 1, 71, 1, 16, 6
15	2, 38, 2, 1, 1, 4, 24, 5, 81, 10, 1, 1, 3, 355, 1
20	1, 2, 5, 5, 15, 1, 404, 1, 1, 42, 1, 1, 4, 7, 30, 4, 1, 2, 1, 1
25	4, 1, 2, 353, 1, 1, 1, 2, 20, 2, 1, 4, 1, 1, 14, 1, 71, 1, 12, 16, 4, 11, 2, 2, 1

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?**

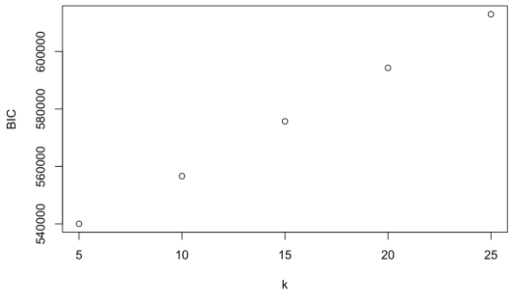


Figure: Plot of number of clusters with BIC values

Since the BIC is increasing with the clusters Since BIC is proportional to the degrees of freedom which inturn depends on K we see that BIC is increasing with K instead of decreasing,we see that it is not the most reliable way to choose K. Therefore we look at AICc.

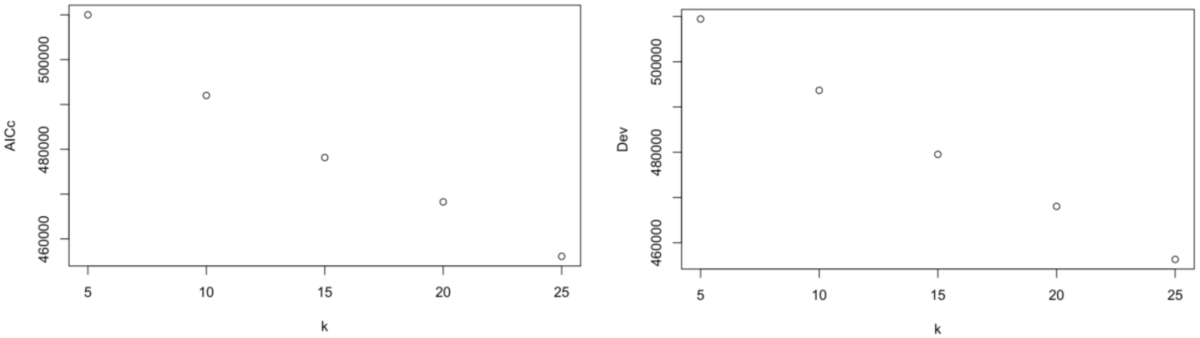


Figure: Plot of number of clusters with AICc and deviance

AICc and deviance are almost identical and we choose k = 25 since it minimizes both AICc and deviance.

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

We use 1000 phrases as the number of words j. With the use of Bayes factors we see that the number of topics chosen by the model is 10. BF = exp(-BIC), therefore we choose the model with the highest BF and select the K = 10 topic model. We interpret this as for each of the 1000 phrases (token k) we have theta k topics with omega k weights. From the model we see that 10 topics (thetas) are good enough to represent the document. To bring more insights, we can further assign names to each topic.

Number of topics ->	5	**10**	15	20
log(BF)	57506.56	**77508.93**	76786.48	67996.89
Disp	3.65	**2.90**	2.47	2.23

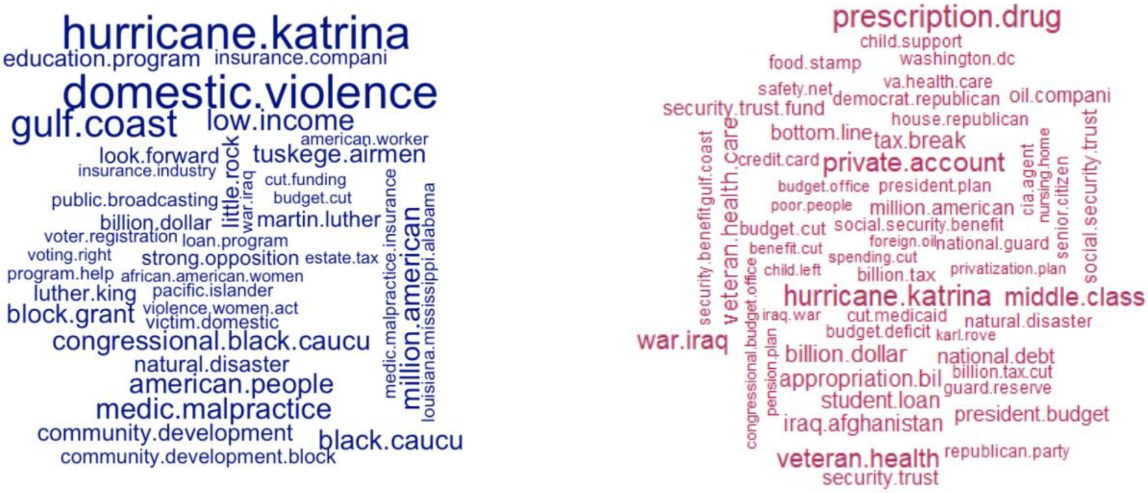


Figure: 2 of the 10 word clouds obtained

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)**

Topic	% of Republican	% of Democratic	Topic	% of Republican	% of Democratic
1	76%	24%	14	50%	50%
2	66%	34%	15	100%	0%
3	100%	0%	16	85%	15%
4	47%	53%	17	0%	100%
5	0%	100%	18	100%	0%
6	86%	14%	19	100%	0%
7	91%	9%	20	8%	92%
8	0%	100%	21	0%	100%
9	25%	75%	22	33%	67%
10	100%	0%	23	0%	100%
11	0%	100%	24	91%	9%
12	17%	83%	25	100%	0%
13	0%	100%			

Connect clustering results to partisanship by Kmeans

From the table above, topics of 4 and 14 are non-partisan, and the topic in blue is mainly about Democratic, while these in red are mainly about Republican. (When the percentage of one party is larger than 60%, we assign the doc to this party.)

Model(y = repshare)	OOS R^2 by Topic	OOS R^2 by Words
Democratic	0.01177	0
Republican	0.12421	0.12105

OOS R^2 for both models

For the topic regression model, we build regressions on two parties separately.
The first model is: repshare ~ omega, from which we could investigate the effect of estimating repshare based on topics selected by the topic modeling method. The second model is: repshare ~ percentage of phrase, from which we could investigate the effect of estimating repshare directly by words, rather than applying topic modeling method.
Using cv.glmnet, we compare the max OOS-R square between the two models for both parties, and the output is as above. From the table, we can conclude that topic modeling has a marginal advantage in estimating repshare in both parties.