Jayoung Kang

Midterm

1 Marginal Significance Screening and False Discovery

To begin, we would like to find words in paper titles that associate with citations. We will look at one-at-a-time (marginal) regressions (regressing outcome on each word separately) to see if there are certain key title words that can predict citations. We will look at the following outcome $Y_i = \log(\texttt{citation}_i + 1)$ for the i^{th} research article. In the starter script, you will find a code to run the individual regressions (using parallel computing). The code gives you a set of p-values for the marginal effect of each of the words. That is, we fit

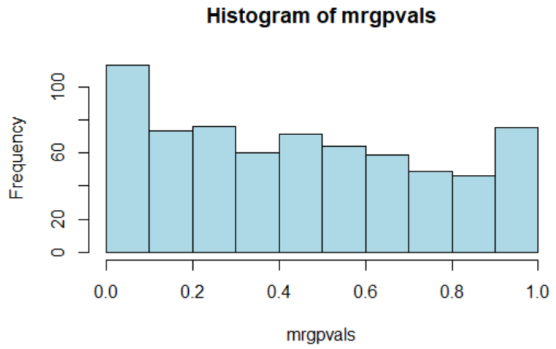
$$Y_i = \alpha + \beta_j[x_{ij} > 0] + \varepsilon_{ij}$$

for each word j with count x_{ij} in the title in the i^{th} article, and return the p-value associated with a test of $\beta_j \neq 0$.

1.1 Plot the p-values and comment on their distributions. Is there enough signal to predict citations based on the topic of the article? (12 points)

- As show in <Figure 1> there is a spike in the frequency around zero, which is indicative of a possible signal that citations can be predicted based on the title word of the article.
- If there were no signal, the distribution of the p-values would be uniform. The distribution of <Figure 1> is not a uniform distribution and is indicative of the possibility that there may be discoveries of covariates that reject the null hypothesis because there are more p-values close to zero than expected under the null.
- However, further testing would need to be conducted in order to test whether this signal is meaningful enough.

<Figure 1>



1.2 What is the alpha value (p-value cutoff) associated with 20% False Discovery Rate? How many words are significant at this level? What are the advantages and disadvantages of FDR for word selection? (14 points)

- The p-value cutoff for 20% FDR: 0.002881604
- Number of words significant at 20% FDR: 10
- Advantages of FDR for word selection:
  - This FDR analysis is effective in parsing big data and allowing us to select each word that is individually meaningful in predicting citations.
  - FDR is also advantageous in that it allows us to control for the problem of multiplicity when selecting the words.
- Disadvantages of FDR for word selection:

- FDR is valid only when p-values are approximately independent. However, for this analysis it may not be that the p-values are independent because some of the words could appear simultaneously on the title as n-gram. By extracting separate words from the title, the language structure is destroyed and the effect that a string of words has on citation is not captured.
- The p-values from the FDR do not indicate the direction of the relationship between the words and citation.

1.3 Suppose you just mark the 20 smallest *p*-values as significant. How many of these discoveries do you expect to be false? Are the *p*-values independent? Discuss. (14 points)

- The 20th smallest p-values has the p-value of 7.995731e-03 and the p-value that is the 21st smallest has a p-value of 8.255719ea-03. This would mean that the p-value cutoff is around 0.008. The FDR to get the cutoff at around 0.008 would be between 25% and 26%.
- This would mean that we would expect approximately 5 to be false discoveries (20*25% = 5)
- The p-values are likely not independent of another because the usage of some key words be correlated with another as some words like 'after' could appear simultaneously on the title as n-gram.

```
> pvals_ordered[1:20]
   selection        after  backfitting     absolute       output   covariance    rejoinder     variable
3.900475e-05 5.079644e-04 5.080861e-04 7.656960e-04 1.282430e-03 1.676580e-03 1.840808e-03 2.399703e-03
      sliced      process      dantzig    dirichlet        lasso      profile  application       stepup
2.551244e-03 2.881604e-03 3.567505e-03 3.984817e-03 4.019730e-03 5.794867e-03 6.039191e-03 6.146661e-03
  microarray    smoothing       spline        false
6.306821e-03 7.304126e-03 7.754292e-03 7.995731e-03
```

2 LASSO Variable Selection

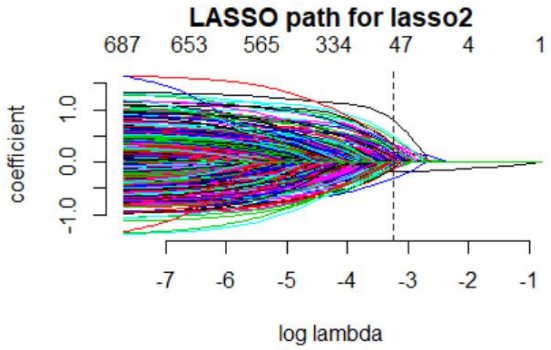
2.1 Use the LASSO method to come up with a combination of a few words that predict citations. Pick a lambda and comment on the in-sample R^2. Is there enough evidence to conclude that title predicts citations? (12 points)

- AICc Lambda: -6.022561
- In-sample R^2: 0.08006071
- The in-sample R^2 indicates that title words can explain only 8% of the variability in the citations (the continuous number of citations).
- The standardization here is set to standardize = FALSE because if it were set as TRUE it would put more penalty on words that are used more often and this might be unfair on commonly used words.
- Due to the low R^2 it is difficult to conclude that title words predict citations. Also, since we didn't account for other covariates that could have confounding effects, we cannot conclude that it is the pure effect of title words that account for the variability in citations.

2.2 Repeat the analysis from (2.1) but add extra covariates. What is the in-sample R^2 now? Describe the LASSO path and pick the top 10 strongest coefficients. What is the interpretation of the coefficient of the word `lasso`? (12 points)

- In-sample R^2: 0.2135955
- Description of LASSO path for <Figure 2>:
  - The LASSO path graph shows all covariates when log lambda is very small. As lambda increases, penalization increases and some of the weaker $\hat{\beta}$s are pulled towards 0.
  - The overlapping of lines indicates that there is some multicollinearity among the covariates.
  - The log lambda chosen for AICc is -3.240777, which is indicated as the dashed line in <Figure 2>, and shows that 66 coefficients are chosen
- Top 10 strongest coefficients: intercept, output, after, absolute, Dantzig, stepup, lasso, profile, singleindex, improving
- Interpretation of the coefficient of the word lasso: If the word lasso is included in the title, the number of citations will increase by approximately 33.9%.
-

<Figure 2>



2.3 Note that 1693 out of the 3248 papers got no citation whatsoever. We now try to predict whether a paper will get at least one citation by running a LASSO with a binary outcome $\tilde{Y}_i = \mathbf{I}(\texttt{citation}_i > 0)$. Repeat the analysis from 2.2. What is the in-sample R^2 now? What is the interpretation of the coefficient `seniority` and `female-coauthors`. Is this causal? (16 points)

- In-sample R^2: 0.1375619
- Interpretation of seniority: A unit increase in the seniority of the paper increases the odds of receiving having at least one citation by 1.016483 times.

- Interpretation of female-coauthors: A unit increase in the number of female-coauthors changes the odds of having at least one citation by 0.9586821 times, or - 4% lower odds.
- Causality in relationship: It is difficult to conclude a causal relationship between seniority and citations, as well as female-coauthors and citations. For causal inference to occur we would have to be able to determine the change in outcome as treatment ‘d’ moves independently from all other confounding covariates. The words in the title may not suffice as all of the relevant covariates to allow for causal inference to occur. There could be other unobservables that causes confounding.

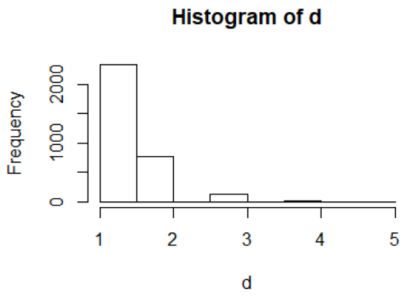
3 High-dimensional Controls, Double LASSO and Bootstrap

We want to isolate the effect of gender (the number of female coauthors) on the number of citations, controlling for all relevant words (paper topics) and other characteristics such as seniority (experience) of the authors. Our *treatment variable* will be d_i = `number of female coauthors` in the i^{th} article and we use the binary outcome $Y_i = \mathbb{I}(\texttt{citation}_i > 0)$.

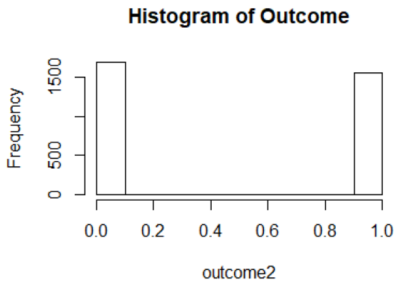
3.1 Explore the association between citations and gender (both graphically and using a marginal regression or other statistical tests) to see if there is any association. Interpret the coefficient from your marginal regression (Y on d). Predict d from x (title words and other predictors) using Poisson regression (see the code for guidance on Poisson regression), and comment on the degree of confounding we can expect. Is there any information in d independent of x? (14 points)

- Association between citations and gender:
  - Y ~ d regression
    - Coefficient interpretation: A unit increase in the number of female-coauthors changes the odds of receiving having at least one citation by 0.7666154 times, or – 23.3% lower odds
    - R^2 = 0.004202989
  - Graphical association
    - Figure 3 and 4 indicate that while the distribution of outcome between 0 and 1 is fairly even, the distribution of d is skewed towards 1 so that there are very few cases where there are more than 2 female coauthors
    - The association table on Figure 5 shows that the proportion of papers with at least one citation falls as the number of female coauthors increases. However, the proportion of papers with more than 2 female coauthors decreases rapidly as well.
    - Figure 6, 7, and 8 show that the conditional distribution tells a similar story in that as the number of female coauthors increase, there are fewer or even no citations.

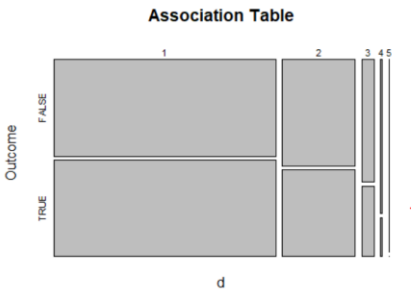
<Figure 3>



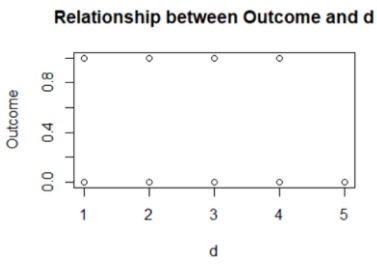
<Figure 4>



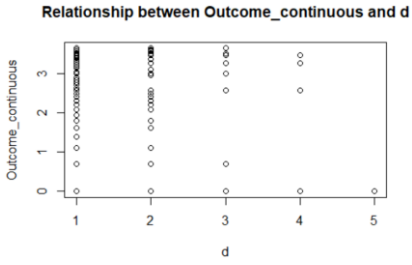
<Figure 5>



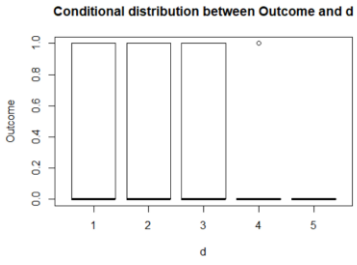
<Figure 6>



<Figure 7>



<Figure 8>



- Comment on the degree of confounding we can expect:
  - R^2 from X1~d: 0.07872685
  - Since R^2 is quite low, this indicates that the part of d that can be predicted with x's isn't very

high. This implies that the treatment effect, when estimated without the controls will not be overestimated by too much, given that the controls we have are enough to account for the confounding effect

- Information in d independent of x
  - Approximately 92% of variability in d is not explained by the x's

3.2 Isolate the effect of d by running the causal (double) LASSO. Interpret this effect and compare it to the effect obtained from the naive LASSO. (12 points)

- Double LASSO coefficient of d: 0
  - This indicates that there is no signal, so that the number of female coauthors have no effect on whether there is at least one citation, holding all other x's constant
- Naïve LASSO coefficient of d: -0.04372288
  - A unit increase in the number of female coauthors changes the odds of receiving having at least one citation by 0.9572192 times
- The difference in the two indicates that there is some confounding effect from the words in the title that is causing the effect of d to deviate from 0.

3.4 Consider the estimated treatment effect for d. We want to know how variable this estimate is (i.e. compute the standard error) and construct confidence intervals for inference. The starter script has a code to bootstrap the sampling distribution for the LASSO estimate of d selected by AICc in this regression.

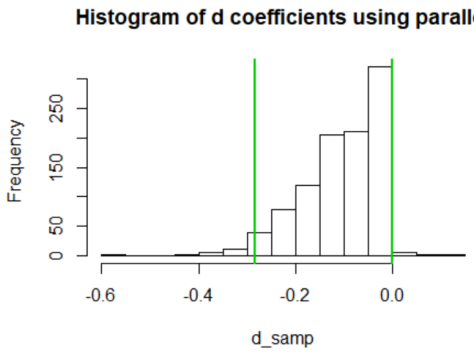
– What is the standard error for the treatment effect d?
  - Using parallel: 0.0858943
  - Without parallel: 0.0794191

– Find the 95% CI for d?
  - Using parallel:[-0.2831883, 0] (illustrated in Figure 9)
  - Without parallel: [-0.2755327, 0] (illustrated in Figure 10)

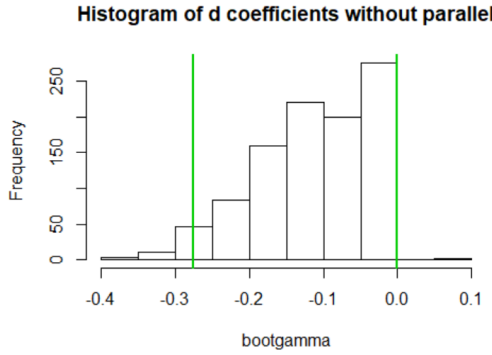
Can we safely claim that the effect is causal? (14 points)

- It is difficult to claim that the effect is causal because 0 is within the 95% CI for d. This means that we cannot reject the null hypothesis that the effect of female_coauthors is zero
- Also we cannot be sure that we accounted for all of the confounding variables for running the regression.

<Figure 9>



<Figure 10>



Codes

```
library(tidyverse)
library(gamlr)

# A list of authors
author<-read.table("authorFinal.csv",header=T)

# Covariates of papers
paper_covariates<-read.csv("PaperCovariates.txt",sep=",",header=TRUE)

# Simple triplet matrix for paper - word pairing
doc_word<-read.table("WordDocCitation.csv", header=F)

# A list of words used in the title
words<-read.table("CitationWords.csv",header=F)

# A list of papers
paper<-read.csv("paperList.txt",sep="\"",header=TRUE)



##### QUESTION 1 ##################################
# FDR: we want to pick a few words in paper titles
# that correlate with citations

spm<-sparseMatrix(
                        i=doc_word[,1],
                        j=doc_word[,2],
                        x=doc_word[,3],
                        dimnames=list(id=1:nrow(paper),words=words[,1]))

# create a dense matrix of word presence
P <- as.data.frame(as.matrix(spm>0))
library(parallel)
Outcome_continuous<-log(paper_covariates$citation+1)

margreg <- function(x){
            fit <- lm(Outcome_continuous~x)
            sf <- summary(fit)
            return(sf$coef[2,4])
}

cl <- makeCluster(detectCores())

# **** FDR Analysis ****
clusterExport(cl,"Outcome_continuous")

# run the regressions in parallel
mrgpvals <- unlist(parLapply(cl,P,margreg))

# mrgpvals<-apply(P,2,margreg) try this if parallel does not work
source("fdr.R")

## 1.1
hist(mrgpvals,col="lightblue")

## 1.2
pvals_ordered<-mrgpvals[order(mrgpvals,decreasing=F)]
p<-ncol(spm)
q<-0.2
source("fdr.R")
cutoff <- fdr_cut(mrgpvals, q)
```

```
signif <- pvals_ordered <= cutoff
table(mrgpvals<=cutoff)

## 1.3
pvals_ordered[1:20]
cutoff2 <- fdr_cut(mrgpvals, 0.25)
cutoff3 <- fdr_cut(mrgpvals, 0.26)
signif2 <- pvals_ordered <= cutoff2
table(mrgpvals<=cutoff2)
table(mrgpvals<=cutoff3)


##### QUESTION 2 ###################################
## 2.1
lasso1<- gamlr(spm, y=Outcome_continuous, standardize = FALSE, lambda.min.ratio=1e-3)

#lambda
log(lasso1$lambda[which.min(AICc(lasso1))])

# find R2 (method 1)
source("deviance.R")
pred <- predict(lasso1, newdata = spm, type="response")
R2(Outcome_continuous, pred, family = NULL)

# find R2 (method 2)
summary(lasso1)$r2[which.min(AICc(lasso1))]


## 2.2
spm2 <- sparse.model.matrix(~.-1, data=paper_covariates[,2:7])
spm3 <- cbind(spm2,spm)
lasso2<- gamlr(spm3, y=Outcome_continuous, lambda.min.ratio=1e-3)

#lambda
log(lasso2$lambda[which.min(AICc(lasso2))])

# find R2 (method 1)
source("deviance.R")
pred2 <- predict(lasso2, newdata = spm3, type="response")
R2(Outcome_continuous, pred2, family = NULL)

# find R2 (method 2)
summary(lasso2)$r2[which.min(AICc(lasso2))]

# LASSO plot
plot(lasso2, main="LASSO path for lasso2")

# Number of covariates selected
Betas <- drop(coef(lasso2))
length(Betas)
sum(Betas!=0)

# Top 10 words
o<-order(Betas,decreasing=TRUE)
Betas[o[1:10]]

# 'lasso' coef
Betas['lasso']


## 2.3
Outcome<-paper_covariates$citation>0
lasso3<- gamlr(spm3, y=Outcome, family='binomial', lambda.min.ratio=1e-3)
```

```
#lambda
log(lasso3$lambda[which.min(AICc(lasso3))])

# find R2 (method 1)
source("deviance.R")
pred3 <- predict(lasso3, newdata = spm3, type="response")
R2(Outcome, pred3, family = 'binomial')

# find R2 (method 2)
summary(lasso3)$r2[which.min(AICc(lasso3))]

# 'seniority_paper' coef
beta2['seniority_paper']
exp(beta2['seniority_paper'])

# 'female_coauthors_paper' coef
beta2['female_coauthors_paper']
exp(beta2['female_coauthors_paper'])
```

QUESTION 3

3.1

```
d<-paper_covariates$female_coauthors_paper

# association between citations and gender
boxplot(Outcome~d, main='Conditional distribution between Outcome and d')
plot(x=d, y=Outcome, main = 'Relationship between Outcome and d')
plot(x=d, y=Outcome_continuous, main = 'Relationship between Outcome_continuous and d')
hist(d)
outcome2<-as.integer(as.logical(Outcome))
hist(outcome2, main='Histogram of Outcome')
t2<-table(d,Outcome)
plot(t2,main="Association Table")
reg <- glm(Outcome~d, family = 'binomial')
summary(reg)
exp(-0.26577)
1-exp(-0.26577)
1-(4477.9/4496.8)

# Stage 1 LASSO: fit a model for d on x
spm4 <- sparse.model.matrix(~.-1, data=paper_covariates[,2:6])
X1<-cbind(spm4, spm)

# Here we use Poisson regression for counts, because d is a count variable
treat <- gamlr(X1,d,family="poisson")
plot(treat)

# Extract dhat from treat
dhat<- predict(treat, X1) #type='response'?

# R2 from poisson
cor(drop(dhat),d)^2
1-cor(drop(dhat),d)^2
```

3.2

```
double_LASSO <- gamlr(cbind(d, dhat, X1), Outcome, free = 2, family='binomial', lmr=1e-4)
```

```
coef(double_LASSO)["d",]

naive <- gamlr(cbind(d,X1),Outcome, family='binomial')
coef(naive)["d",]
exp(coef(naive)["d",])


## 3.4
# With parallel
clusterExport(cl,"X1")
clusterExport(cl,"d")
clusterExport(cl,"Outcome")

boot_function <- function(ib){
        require(gamlr)
  xb <- X1[ib,]
  db <- d[ib]
  yb <- Outcome[ib]
  treatfit <- gamlr(xb,db,family="poisson")
        dhat <- predict(treatfit, xb)
        double_LASSO <- gamlr(cbind(db, dhat, xb), yb, family='binomial', free = 2)
        coef(double_LASSO)["db",]
}

boots <- 1000
n <- nrow(spm)
resamp <- as.data.frame(matrix(sample(1:n,boots*n,replace=TRUE),ncol=boots))
d_samp <- unlist(parLapply(cl,resamp,boot_function))

sd(d_samp)
hist(d_samp)
abline(v=quantile(d_samp,0.025),col=3,lwd=2)
abline(v=quantile(d_samp,0.975),col=3,lwd=2)
quantile(d_samp,0.025)
quantile(d_samp,0.975)

# Without parallel
n <- nrow(spm)
B <- 1000
bootgamma <- rep(0,B)
for(b in 1:B){
  ib <- sample(1:n, n,replace=TRUE)
  xb <- X1[ib,]
  db <- d[ib]
  yb <- Outcome[ib]
  treatb <- gamlr(xb,db,family="poisson")
  dhatb <- predict(treat, xb)
  fitb <- gamlr(cbind(db, dhat, xb), yb, family='binomial', free = 2)
  bootgamma[b] <- coef(fitb)["db",]
  print(b)
}

sd(bootgamma)
hist(bootgamma)
abline(v=quantile(bootgamma,0.025),col=3,lwd=2)
abline(v=quantile(bootgamma,0.975),col=3,lwd=2)
quantile(bootgamma,0.025)
quantile(bootgamma,0.975)
```