BUS 41201 Big Data Midterm

Shri Lekkala

24 April 2024

Setup

```
# Reddit news data
data = read.csv("RedditNews.csv", header = FALSE, skip = 1)
data[,2:3] = data[,1:2]
data[,1] = paste0("RedditNews_",rownames(data))
date<-data[2] # this is the day of the news

subset<-date=="7/1/16" # let's take a look at news headlines on 7/1/16
# data[subset,3] # we have 24 news headlines
```

```
# Read the DJIA data
dj<-read.csv("DJIA.csv")
head(dj) # Open price, highest, lowest and close price
```

```
##         Date     Open     High      Low    Close    Volume Adj.Close
## 1 2016-07-01 17924.24 18002.38 17916.91 17949.37  82160000  17949.37
## 2 2016-06-30 17712.76 17930.61 17711.80 17929.99 133030000  17929.99
## 3 2016-06-29 17456.02 17704.51 17456.02 17694.68 106380000  17694.68
## 4 2016-06-28 17190.51 17409.72 17190.51 17409.72 112190000  17409.72
## 5 2016-06-27 17355.21 17355.21 17063.08 17140.24 138740000  17140.24
## 6 2016-06-24 17946.63 17946.63 17356.34 17400.75 239000000  17400.75
```

```
ndays<-nrow(dj) # 1989 days
```

```
# Read the words
words<-read.csv("WordsFinal.csv",header=F)
words<-words[,1]
head(words)
```

```
## [1] "ab"      "abandon" "abba"    "abbott"  "abc"     "abduct"
```

```
length(words)
```

```
## [1] 5271
```

```
# Read the word-day pairings
doc_word<-read.table("WordFreqFinal.csv",header=F)

# Create a sparse matrix
spm<-sparseMatrix(
        i=doc_word[,1],
        j=doc_word[,2],
        x=doc_word[,3],
        dimnames=list(id=1:ndays,words=words))
dim(spm)
```

```
## [1] 1989 5271
```

```
# We select only words at occur at least 5 times
cols<-apply(spm,2,sum)
index<-apply(spm,2,sum)>5
spm<-spm[,index]

# and words that do not occur every day
index<-apply(spm,2,sum)<ndays
spm<-spm[,index]

dim(spm) # we end up with 3183 words
```

```
## [1] 1989 3183
```

Question 1 Marginal Significance Screening and False Discovery

1.1 Plot the p-values and comment on their distributions (for both outcomes V and R). Is there enough signal to predict prices and volatility?

1.2 What is the alpha value (p-value cutoff) associated with 10% False Discovery Rate? How many words are significant at this level? (Again analyze both outcomes V and R.) What are the advantages and disadvantages of FDR for word selection?

1.3 Now, focus only on volatility V. 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.

Question 2 LASSO Variable Selection and Bootstrapping

2.1 Use the LASSO method to come up with a combination of a few words that predict returns R. Pick a lambda and comment on the in-sample R^2. Is there enough evidence to conclude that headlines predict returns?

2.2 Repeat the analysis from (2.1) to predict Volatility instead of Returns. Next, add one extra predictor, Volatility on a previous day. In other words, fit a LASSO model for predicting today's volatility V_t using word counts and yesterday's volatility V_{t-1}.

$V_t = a + bV_{t-1} + x_t'\beta + \epsilon_t$

What is the in-sample R2 now? Describe the LASSO path and pick the top 10 strongest coefficients. What is the interpretation of the coefficient of word "terrorist" and of V_{t-1}?

2.3 Consider the estimated of lambda selected by AICc in the model from (2.2) (using both words and V_{t-1}). We want to know how variable this estimate is. The starter script has code to bootstrap the sampling distribution for the λ selected by AICc in this regression.

– What is the standard error for the selected λ?

– Find the 95% CI for λ?

Question 3 High-dimensional Controls and Double LASSO

3.1 Explore a marginal regression (just a regression of V_t on V_{t-1}) to see if there is any correlation. Predict d from x (headlines words), and comment on the degree of confounding we can expect. Is there any information in d independent of x?

3.2 Isolate the effect of V_{t-1} by running the causal (double) LASSO. Interpret this effect and compare it to the effect obtained from the naive LASSO.

3.3 Can we safely claim that the effect is causal?

Bonus Freestyle Analysis

Provide additional analysis of the data. Points will handed out only for insightful use of data mining tools, not for scattershot application of techniques.