BUS 41201 Homework 1 Assignment

Veronika Rockova
Veronika.Rockova@ChicagoBooth.edu

4/8/2020

Amazon Reviews

The dataset consists of 13 319 reviews for selected products on Amazon from Jan-Oct 2012. Reviews include product information, ratings, and a plain text review. The data consists of three tables:

##Review subset.csv is a table containing, for each review, its

- ProductId: Amazon ASIN product code
- UserId: ID of the reviewer
- Score: numeric 1-5 (the number of stars)
- Time: date of the review
- Summary: review summary in words
- Nrev: number of reviews by the user
- Length: number of words in the review
- Prod Category: Amazon product category
- Prod Group: Amazon product group

Word freq.csv

is a simple triplet matrix of word counts from the review text including

- Review ID: the row index of Review subset.csv
- Word ID: the row index of words.csv
- Times Word: how many times the word occurred in the review

Words.csv

contains 1125 alphabetically ordered words that occur in the reviews.

Data exploration

The code below loads the data.

```
library(knitr) # library for nice R markdown output



# READ REVIEWS

data<-read.table("Review_subset.csv",header=TRUE)
dim(data)
```

```
[1] 13319 9
```

```
# 13319 reviews
# ProductID: Amazon ASIN product code
# UserID:  id of the reviewer
# Score: numeric from 1 to 5
# Time: date of the review
# Summary: text review
# nrev: number of reviews by this user
# Length: length of the review (number of words)

# READ WORDS

words<-read.table("words.csv")
words<-words[,1]
length(words)
```

```
[1] 1125
```

```
#1125 unique words

# READ text-word pairings file

doc_word<-read.table("word_freq.csv")
names(doc_word)<-c("Review ID","Word ID","Times Word" )
# Review ID: row of the file  Review_subset
# Word ID: index of the word
# Times Word: number of times this word occurred in the text
```

Marginal Regression Screening

We would like to pre-screen words that associate with ratings. To this end, we run a series of (independent) marginal regressions of review Score on word presence in review text for each of 1125 words.

In the starter script below, you will find a code to run these marginal regressions (both in parallel and sequentially). The code gives you a set of p-values for a marginal effect of each word. That is, we fit

$$\texttt{stars}_i = \alpha + \beta_j I[x_{ji} > 0] + \epsilon_{ji}$$

for each word term j with count x_{ji} in review i, and return the p-value associated with a test of $\beta_j \neq 0$. We'll use these 1125 independent regressions to screen words.

```
# We'll do 1125 univariate regressions of
# star rating on word presence, one for each word.
# Each regression will return a p-value, and we can
```

```r
# use this as an initial screen for useful words.

# Don't worry if you do not understand the code now.
# We will go over similar code in  the class in a few weeks.

# Create a sparse matrix of word presence


library(gamlr)
```

```
## Loading required package: Matrix
```

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

dim(spm)
```

```
[1] 13319 1125
```

```r
# 13319 reviews using 1125 words

# Create a dense matrix of word presence

P <- as.data.frame(as.matrix(spm>0))

library(parallel)

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

# The code below is an example of parallel computing
# No need to understand details now, we will discuss more later

cl <- makeCluster(detectCores())

# Pull out stars and export to cores

stars <- data$Score

clusterExport(cl,"stars")

# Run the regressions in parallel

mrgpvals <- unlist(parLapply(cl,P,margreg))

# If parallel stuff is not working,
# you can also just do (in serial):
# mrgpvals <- c()
# for(j in 1:1125){
#   print(j)
#   mrgpvals <- c(mrgpvals,margreg(P[,j]))
# }
# make sure we have names

names(mrgpvals) <- colnames(P)

# The p-values are stored in mrgpvals
```

Homework Questions:

(1) Plot the p-values and comment on their distribution. (1 point)

(2) Let's do standard statistical testing. How many tests are significant at the alpha level 0.05 and 0.01? (1 point)

(3) What is the p-value cutoff for 1% FDR? Plot and describe the rejection region. (1 point)

(4) How many discoveries do you find at q=0.01 and how many do you expect to be false? (1 point)

(5) What are the 10 most significant words? Do these results make sense to you? What are the advantages and disadvantages of our FDR anaysis? (1 point)