BUSN 41201 HW1

Group 11

1. Intro to Datasets

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

```
##    ProductId             UserId Score       Time                  Summary Nrev
## 1 B005HG9ESG #oc-R163CP16SRRI50     5 2012-08-02 Light, refreshing water!    3
## 2 B005HG9ERW #oc-R163CP16SRRI50     5 2012-08-02 Light, refreshing water!    3
## 3 B005HG9ET0 #oc-R163CP16SRRI50     5 2012-08-02 Light, refreshing water!    3
## 4 B006Q820X0  #oc-R19QDOY2PXS15     3 2012-05-19   Good Coffee, Bad K-cup    1
## 5 B006Q820X0 #oc-R1B9W981WGB5D0     3 2012-06-15           Weak and bland    1
## 6 B007TGDXMK #oc-R1I879FCTH83GM     2 2012-04-16   Nice but a little weak    2
##   Length                Prod_Category Prod_Group
## 1      3                        Water    Grocery
## 2      3                        Water    Grocery
## 3      3                        Water    Grocery
## 4      4 Single-Serve Capsules & Pods    Grocery
## 5      3 Single-Serve Capsules & Pods    Grocery
## 6      4 Single-Serve Capsules & Pods    Grocery
```

```
dim(data)
```

```
## [1] 13319     9
```

```
words<-read.table("words.csv")
head(words)
```

```
##           V1
## 1      about
## 2   absolute
## 3 absolutely
## 4 absorbable
## 5     action
## 6   actually
```

```
words<-words[,1]
length(words)
```

```
## [1] 1125
```

```
doc_word<-read.table("word_freq.csv")
names(doc_word)<-c("Review ID","Word ID","Times Word" )
```

```
library(gamlr)
```

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

```
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
```

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

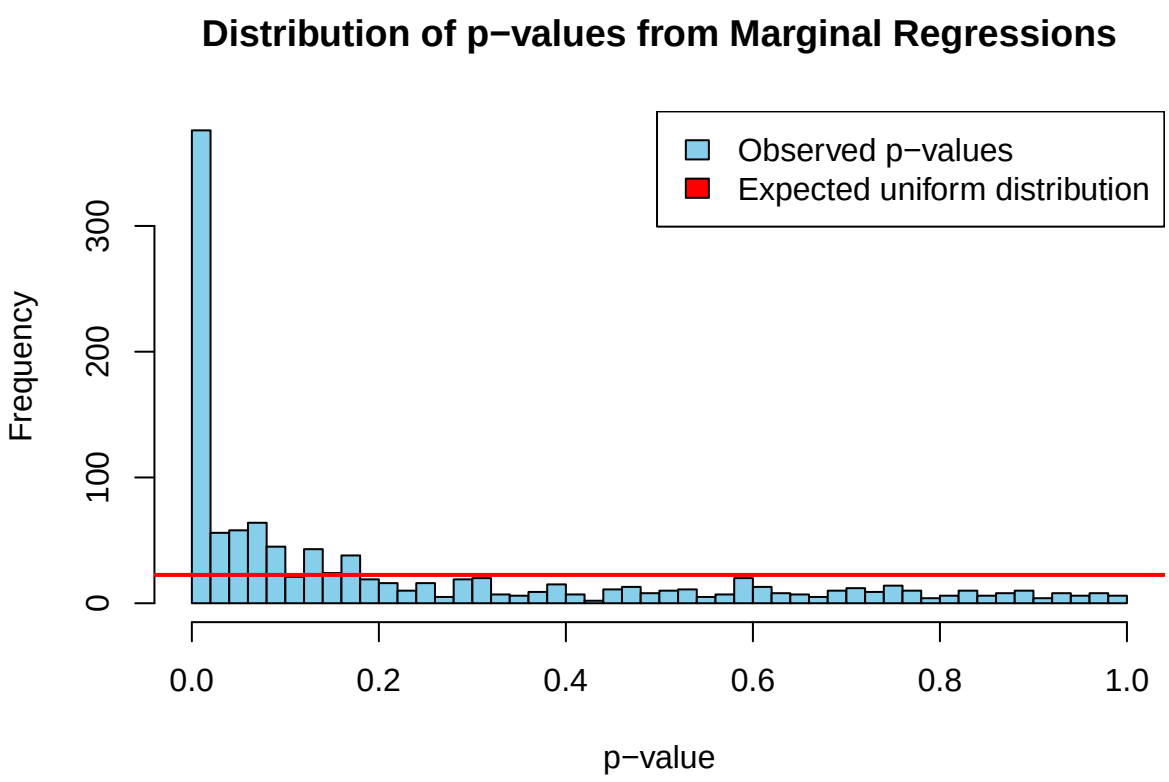
```
library(parallel)

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

```
cl <- makeCluster(detectCores())

stars <- data$Score

clusterExport(cl, 'stars')

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

names(mrgpvals) <- colnames(P)
```

2. Questions

Q1. Plot the p-values and comment on their distribution.

```
hist(mrgpvals, breaks = 50, main = "Distribution of p-values from Marginal Regressions", xlab = "p-valu

# Add a line for the uniform distribution expectation
abline(h = length(mrgpvals) / 50, col = "red", lwd = 2)

legend("topright", legend = c("Observed p-values", "Expected uniform distribution"), fill = c("skyblue"
```



Based on the provided histogram of p-values from marginal regressions:

1. **Skewness towards Zero**: The distribution is heavily skewed towards zero, with a high frequency of p-values close to zero. This indicates that many words are strongly associated with the review scores.
2. **Multiple Significance**: There is a large number of p-values that are significantly lower than the expected uniform distribution would predict. This suggests that there are words that are indeed related to the review scores and that the effect is non-random.
3. **Implication for Hypotheses**: The initial impression is that many null hypotheses (the hypothesis that there is no effect) can be rejected.
4. **Potential for False Discoveries**: Despite the significant results from multiple tests, caution should be exercised about false discoveries.

Q2. How many tests are significant at the alpha level 0.05 and 0.01?

```
significant_at_05 <- sum(mrgpvals < 0.05)

# Count the number of tests significant at the alpha level of 0.01
significant_at_01 <- sum(mrgpvals < 0.01)

# Print the counts
print(paste("Number of tests significant at alpha = 0.05:", significant_at_05))
```

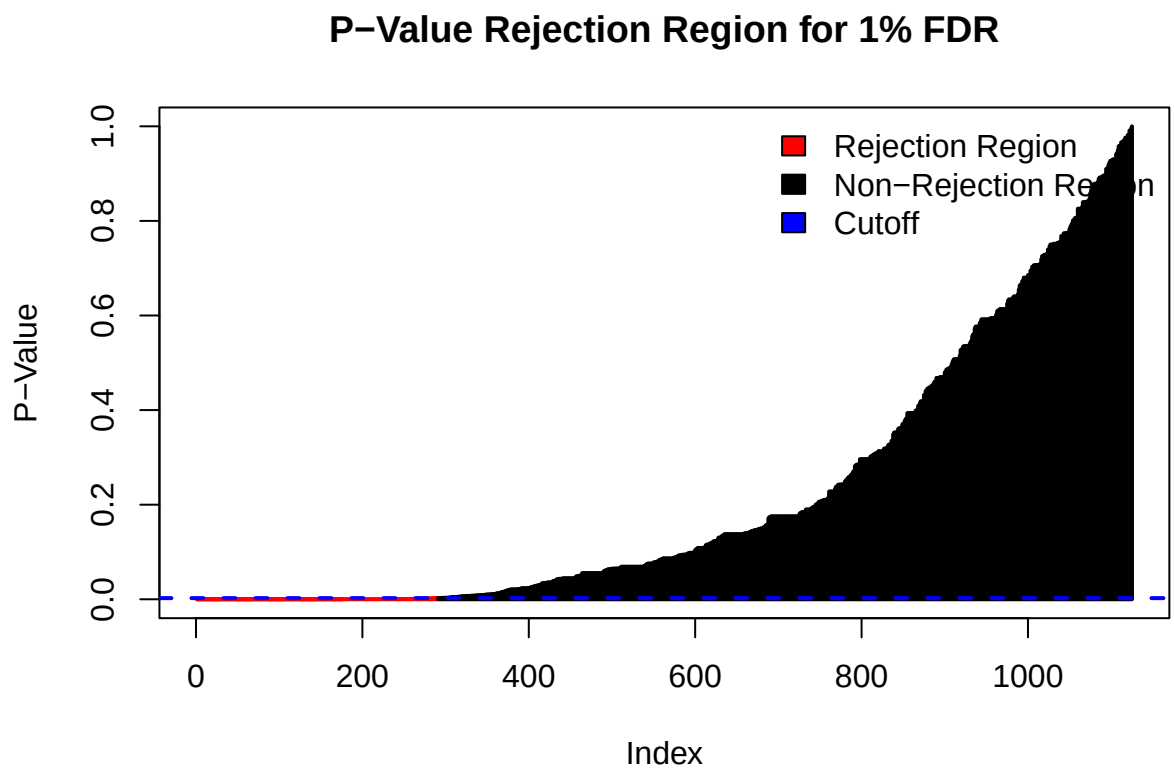
```
## [1] "Number of tests significant at alpha = 0.05: 461"
```

```
print(paste("Number of tests significant at alpha = 0.01:", significant_at_01))
```

```
## [1] "Number of tests significant at alpha = 0.01: 348"
```

Q3. What is the p-value cutoff for 1% FDR? Plot and describe the rejection region.

```
sorted_pvals <- sort(mrgpvals)

# Number of tests
m <- length(sorted_pvals)

# Desired FDR level
fdr_level <- 0.01

# Calculate the Benjamini-Hochberg critical values
bh_critical_values <- (1:m / m) * fdr_level

# Find where sorted p-values are less than BH critical values
cutoff_index <- max(which(sorted_pvals < bh_critical_values))

p_value_cutoff <- sorted_pvals[cutoff_index]

# Plotting
plot(sorted_pvals, type = "h", lwd = 2, col = ifelse(1:m <= cutoff_index, "red", "black"),
     xlab = "Index", ylab = "P-Value", main = "P-Value Rejection Region for 1% FDR")
abline(h = p_value_cutoff, col = "blue", lty = 2, lwd = 2)
legend("topright", legend = c("Rejection Region", "Non-Rejection Region", "Cutoff"),
       fill = c("red", "black", "blue"), bty = "n")
```




```
# Print the p-value cutoff
print(paste("P-value cutoff for 1% FDR:", p_value_cutoff))
```

```
## [1] "P-value cutoff for 1% FDR: 0.00241324881986093"
```

Q4. How many discoveries do you find at q = 0.01 and how many do you expect to be false?

```
num_discoveries <- sum(sorted_pvals <= p_value_cutoff)

# Estimate the expected number of false discoveries ### confirm
expected_false_discoveries <- num_discoveries * fdr_level

cat("Number of discoveries at q = 0.01:", num_discoveries, "\n")
```

```
## Number of discoveries at q = 0.01: 290
```

```
cat("Expected number of false discoveries among these:", expected_false_discoveries, "\n")
```

```
## Expected number of false discoveries among these: 2.9
```

Q5.

- What are the 10 most significant words?

```
sorted_mrgpvals <- sort(mrgpvals, decreasing = FALSE)

# Extract the names of the 10 most significant words
top_10_words <- names(sorted_mrgpvals)[1:10]
print(top_10_words)
```

```
##  [1] "not"          "horrible"     "great"        "bad"          "nasty"
##  [6] "disappointed" "new"          "but"          "same"         "poor"
```

- Do these results make sense to you?

It appears that the ten most significant words from the analysis of review texts are a mix of positive and negative adjectives and qualifiers, such as "not," "horrible," "great," "bad," "nasty," "disappointed," "new," "but," "same," and "poor."

These results make sense in the context of product reviews. Words that express clear sentiments or opinions are likely significant indicators of the overall score that a review might receive. For instance, words like "great" are likely associated with positive scores, while words like "horrible" and "disappointed" are likely related to negative scores. The presence of conjunctions and negations such as "but" and "not" also makes sense, as they are often used to qualify statements and can change the sentiment of a sentence.

- What are the advantages and disadvantages of our FDR analysis?

  Advantages of FDR Analysis:

  - Controls the expected proportion of false positives in multiple testing.
  - More powerful, less likely to miss true effects than methods like Bonferroni.
  - Adjustable to balance the risk of false findings against the need for discovery.
  - Suitable for large datasets with many tests.

  Disadvantages of FDR Analysis:

  - Assumes test independence; may not work well with negatively correlated tests.
  - The choice of FDR threshold can be somewhat arbitrary.
  - Controls the rate, not the number of false discoveries.
  - May not address complex multiplicity in experimental designs.