Project Write Up

Patrick Yu
Stat 133

August 14, 2014

1 Introduction

This project is based off of data supplied from the Harvard dialect survey ending in 2003. I will use the analyse the data using a variety of statistical tools and R to come to conlusions about the surveys.

2 First Things First

First, I wanted some simple information about the data to help me undertand it. After examining the columns I saw a great disparity in the type of answers given. I was able to use R to see which of the questions recieved the highest response on average

```
highest.avg.response.per.state <- sapply(as.matrix((levels(a$STATE))),
    function(x) which.max(colSums(a[a$STATE==x,][,5:71],na.rm = T)/nrow(a)))
states.not.q59 <- which(highest.avg.response.per.state !=10)
```

These lines of code gave me that the highets average response per state was $Q59$ for almost all of the states. This probably meant that the question was one that warranted a highly numbered response on average. On the other end of the spectrum there was $Q63$. This question had the lowest average response per state by far. However, the majority was not quite as overwhelming as $Q59$ for the highets.

```
lowest.avg.response.per.state <- sapply(as.matrix((levels(a$STATE))),
    function(x) which.min(colSums(a[a$STATE==x,][,5:71],na.rm = T)/nrow(a)))
states.not.q63 <- which(highest.avg.response.per.state != 14)
```

3 More Exciting Things

I continued and decided that I wanted to perform some principal component analysis on the data. I started off by taking the *prcomp* if the data and then looking at the standard deviations provided from the analysis.

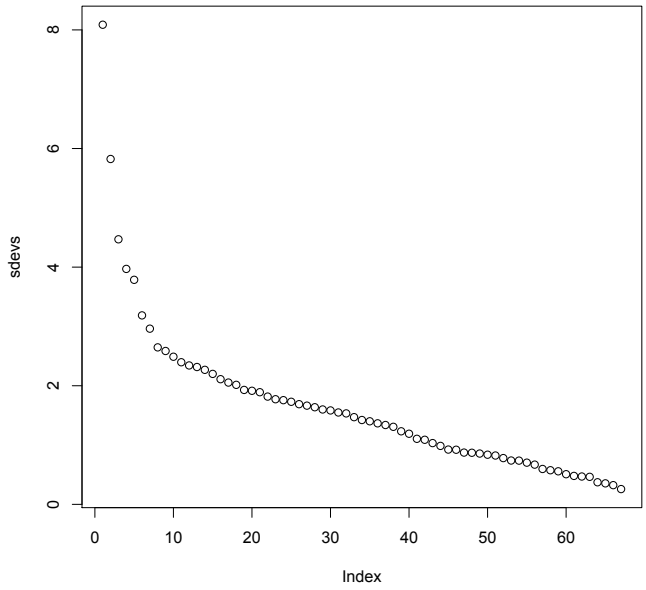
```
pc.analysis <- prcomp(a[,5:71])
sdevs <- pc.analysis$sdev
plot(sdevs)
max.sdevs <- max(sdevs)
min.sdevs <- min(sdevs)
```

The principal component analysis object I created was something I wanted to keep and plot for later, however the standard deviations now available interetsted me. The maximum standard deviation was a relatively large number. That really interested me because often times, we saw that the standard deviation for a question was less that one. I wondered what kind of question could have made such a large standard devation. It was probably something that those whom were surveyed were very divided, or not knowledgeable on.

Here is a quick plot I created in order for me to visualize the standard deviations. The plot really shows just how big the standard deviation jump was.

4 What's next?!?

Unfortunately, I hit a dead end while probing into what the standard deviations meant. I wasn't able to gather too much more information. Before I started to plot my principal component analysis, I decided I wanted to ask another question. What was the most commonly surveyed locaiton, namely



state. This interested me because I thought it might be able to shed light on where the survey took place, or where people are more willing to answer survey questions(where they were more bored). Of course, this ran the risk of the researchers doing well and spreading their surveys across the country.

The data collectors did a good job. I didn't exactly get what I wanted.

```
state.freq <- sapply(levels(a$STATE), function(x) sum(a$STATE == x,na.rm = T))
ordered.state.freq <- state.freq[order(state.freq)]
```

This piece of code told us that the most surveyed state was California, followed by New York. This didn't give me information as these are the two most populated states in the nation.

5 Back to Basics

After having the locations frequencies be relatively useless, I decided to do somthing similar to what I did earlier. I decided I wanted to take the variance of every question to see what it came up with.

```
var.by.question <- sapply(colnames(a)[5:71],function(x) var(a[,x]))
```

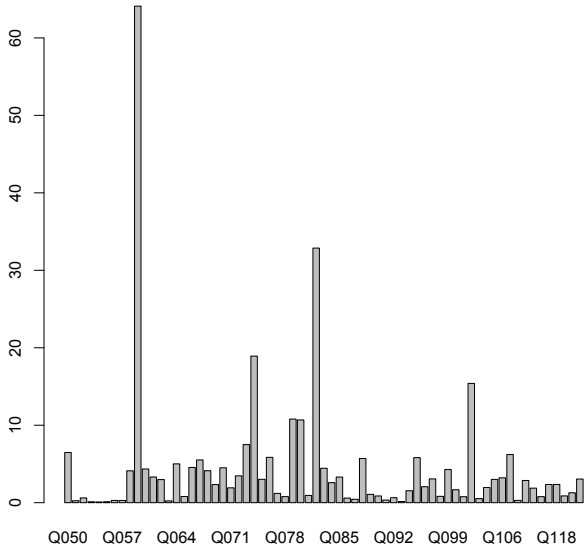
This piece of code I soon realized just gave me my standard deviations squared. I should have realized that. I was about to give this up as a failure until I realized that unlike the standard deviations I recieved from principal component analysis using *prcomp*, the *sapply* line I wrote was able to leave the variances in order. This allowed me to recognize that my question with the obscenely large standard deviation and variance(variance was 64.10198) was $Q059$.

I then realized that I had done something with that question before. It was the highest average response per state. This told me that this question was probably poorly written in relation to the other questions. It's very possible that the question just warrented high responses, but if it was also having a extremely high variance and standard deviation, something was probably wrong. Especially regarding the fact that the data we are currently using has already been cleaned. If I were to recieve this data, I would be sure to take a very close look into rewording and reworking $Q059$.

On the next page is a quick barplot of the variance. It, like the graph of the standard deviations shows the great jump. However, it also show's where the jump occured.

6 What We've All Been Waiting For

Finally I'll do those principal component analysis graphs that I talked about earlier. I decided to plot the principal component analysis in regards to two sets of questions. The first set of questions the 25% and 75% percentiles for the variance. The second set was the lowest and the highest variance(and corespondingly, standard deviation). NOTE: The red is the former, the blue is the latter in both graphs.




```
mycolors = c('red','blue')
plot(pc.analysis$x[,floor(quantile(order(var.by.question),c(.25,.75)))],col = mycolo
plot(pc.analysis$x[,order(var.by.question)[c(1,length(var.by.question))]],col = myco
```

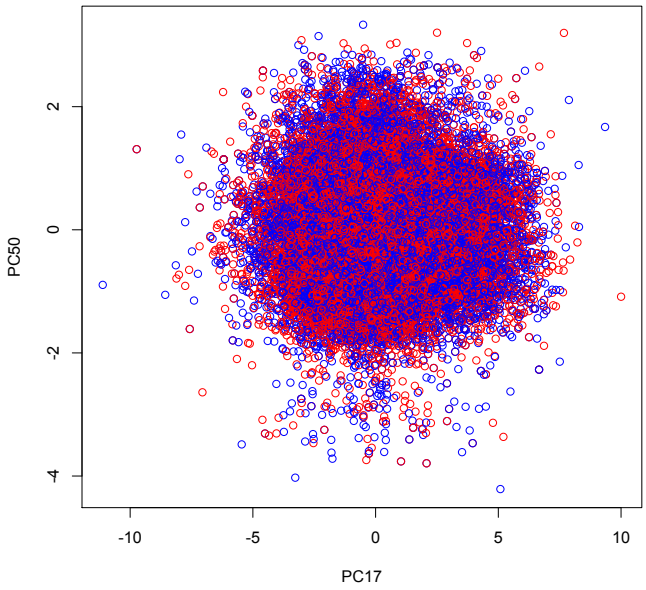
This gave me two telling graphs.

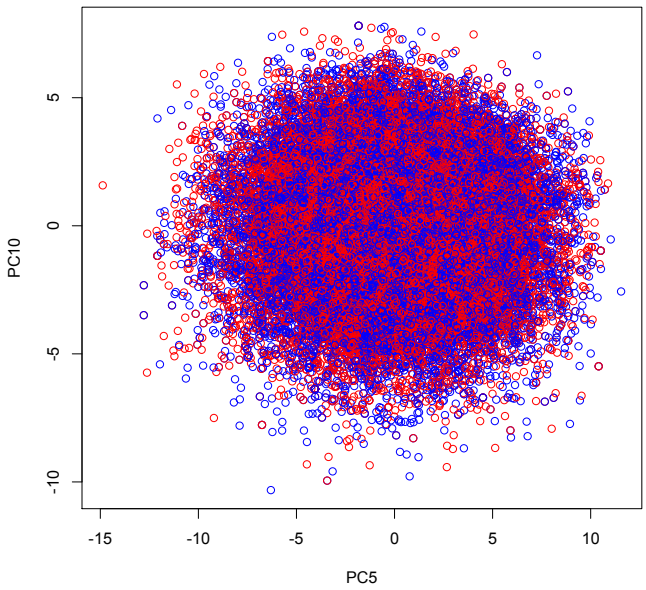
7 Parting Words

Out of all the analysis I did on this survey. The part I'm most proud of was realizing that $Q059$ had the highest standard deviation and highest average number. It really interests me as to what that question was regarding.

For my last trick, I'll plot the principal component analysis with kmeans clustering as colors.

```
k.means <- kmeans(a[,5:71],2,nstart = 10)
plot(pc.analysis$x[,floor(quantile(order(var.by.question),c(.25,.75)))],col = mycolo
```


PC50
2
0
-2
-4
-10
-5
0
5
10
PC17


PC10
5
0
-5
-10
-15
-10
-5
0
5
10
PC5

mycolors is as stated previously.