Using Principal Component Analysis and Decision Trees to Determine Clinical Trials in Countries

ABSTRACT

Using principal component analysis and decision trees, the number of clinical trials that occur in a country can be predicted based upon Worldwide Governance Indicators. Two analyses were performed in R. The first was a principal component analysis that was then used to create a linear regression model. The second analysis used the Naïve Bayes classifier and decision trees that utilized boosting in order to classify like countries and determine if the number of clinical trials hosted there are similar. The results for countries diverged into two groups: those with over 1,000 clinical trials and those with less than 1,000 clinical trials. A decision tree that utilized boosting was able to create a model that could predict this outcome with 80% accuracy. The principal component analysis and linear regression showed some promise but the linear regression model had an insignificant p-value. Lastly, the model in which the clinical trial frequencies were divided into four groups proved to be disastrous with a predication accuracy of 31%. The addition of a GDP indicator raised these results to 98% and 40% respectively. The results show that indicators can successfully be used to determine whether a country will have a large number of clinical trials, but they are not able to discern different thresholds of clinical trials.

INTRODUCTION TO THE TOPIC

The world is becoming a smaller place, and, in the pharmaceutical sector that means more countries are participating in clinical trials. Until recently, most clinical trials were hosted in the United States, Canada, the United Kingdom, or Germany.[1] However, a new trend is emerging where countries from all over the world are hosting clinical trials. However, not every country is privy to this new clinical trial expansion. It appears that certain countries are favored over others and yet there is no obvious rhyme or reason to the selection of these new hosts.

This project hopes to unravel the mystery of why some countries are chosen to host clinical trials by examining different indicators such as Gross Domestic Product (GDP)[2] and Worldwide Governance Indicators[3] to find the correlation between clinical trial hosts and countries. After a review of the available literature, there does not appear to be an obvious factor, such as geographic location, that determines the likelihood that a country is a host.

I believe that by using principal component analysis (PCA) and decision trees, an algorithm can be created to determine the number of clinical trials based on the governance indicators and GDP of a country. I will use the number of clinical trials a country has hosted between 1996 and 2006 as the response variable that will be predicted by the explanatory variables from governance indicators.

Being a host of a clinical trial is important to a country for many reasons, including: bringing in foreign investments, bringing cutting edge medical treatments to a country, promoting standards in clinics and hospitals, and helping with the funding of newer and modernized clinics and hospitals. It is important not just for the economy, but also for the health of a country's citizens. Clinical trials can also help open new markets for pharmaceuticals.

WHY THIS IS INTERESTING

The pharmaceutical industry is a gigantic industry comprised of many multi-national companies. Recently, pharmaceutical companies have been expanding into countries that they did not previously have a large presence in. The number of clinical trials in a country is one means of measuring the pharmaceutical industry's presence in a country. Political science studies have looked at the presence other sectors, like telecommunications, and have been able to determine the expansion of multi-national telecommunications companies in countries based on country indices. To date, there has yet to be a similar algorithm for determining pharmaceutical company expansion into a country.

Machine learning offers the necessary tools to develop an algorithm for determining whether a country will be involved in the current expansion of clinical trials. Countries such as the Czech Republic and Hungary have seen huge expansions in the pharmaceutical industry's presences in recent years.[4] The question to be asked is why? Why these countries as opposed to others? The goal of this research is to answer this question. Using available indices, such as the governance indictors, this project will attempt to address this fundamental question.

LITERATURE REVIEW

An algorithm to determine the number of clinical trials that a country will host has not yet been created. However, this project will examine three areas of the literature in order to review what has already been achieved in other sectors using indicators to determine an occurrence of an event in a country. The three areas of the literature are: the diffusion of clinical trials in new countries, other sectors that have looked at diffusion, and the use of country indices to predict an event in a country using machine learning.

To address these three background areas of research, I have run extensive searches in both Google scholar and Northeastern University's online library. When searching for the diffusion of clinical trials, I used the following searches: diffusion of clinical trials, diffusion of pharmaceutical companies, globalization of clinical trials, trends in globalization of clinical trials, and clinical trials throughout the world. Additionally, I searched for the author Denise Dunlop. Denise Dunlop is a professor at Northeastern University who provided me with the datasets for clinical trials by country. The second area of research, diffusion in other sectors, used the following search terms: diffusion, telecom diffusion, diffusion by sector, and diffusion based on indices. Additionally, I also searched for the author Kristen Rodine-Hardy. Kristen Rodine-Hardy is another professor at Northeastern University who has done extensive research and writing on the diffusion of the telecom

sector across the world. The third area of research, indices and machine learning techniques for comparing countries, used the following search terms: machine learning and country indices, principal component analysis and countries, governance indicators, clustering and countries, principal component analysis and indices, and clustering on country indices.

The first topic is the exploration of the diffusion of clinical trials in new countries. It has already been observed that there is a shifting in the location of pharmaceutical research and development.[5] The U.S.A. is still a leader in pharmaceutical research and development, followed by other modern, westernized countries such as the UK and Germany. However, other non-"traditional" countries are beginning to host pharmaceutical research and development.[6] When looking at the locations of clinical trials, countries such as the Czech Republic, Hungary, and Estonia are beginning to conduct and host many clinical trials.[7] The current literature has established that clinical trials are becoming more globally widespread and are no longer restricted to a few choice countries.

The second area of literature review concerns other sectors that have looked at diffusion and why. Professor Rodine-Hardy has done a similar analysis using event history analysis and survival models to look at different indices and the effect they have on the liberalization of telecommunication companies.[8] Rodine-Hardy primarily examines the relationship between liberalization and membership in international organizations. The difference between this research project and her work will be the idea of working on a spectrum. In Rodine-Hardy's work, a survival analysis was necessary because membership and liberalization were one-time events. This project will instead look at indices and the number of clinical trials; both of which can vary and are not limited to a binary option.

The third area of background material concerns indices and machine learning techniques for comparing countries. Previous work using indices and principal component analysis has been used to compare the different markets of countries.[9] The goal of this type of analysis is to identify similar countries that may not usually be associated as like-countries, as well as to help identify weaknesses or deficiencies in countries that may be affecting their markets. Other principal component analyses have been conducted using indices to measure the type of political economy in countries and cluster like-countries together based on the principal component analysis.[10] The goal of the political economy based clustering is to find more optimal means by which to describe countries by using multiple indices rather than describing countries solely through a single index.

DATA SETS

This research project will primarily analyze two data sets: the ADIS list of clinical trials by country provided by Professor Dunlop and the governance indictors from the Worldwide Governance Indicators Paper by Daniel Kaufmann, et all.[11] The ADIS data set is a paid for data set managed by Springer.[12] The data set provides information about the clinical trials such as: country it was conducted in, the drug(s) tested, the phase, the company or researcher who sponsored the trial, and the phase of the trial. Professor Dunlop purchased this data set as part of her ongoing research. Additionally, the World Banks' GDP ranking[13]

will also be used in some analyses. This research project is primarily concerned with the country that the trial was conducted in. The goal of this research is to determine an algorithm through principal component analysis by which to determine the number of trials that a country will host.

ALGORITHMS

All of the analysis and data cleaning seen in this research project was conducted using R and accompanying R packages. The method and packages will briefly be discussed below. The following R Code section will walk though the step-by-step analyses conducted in this project.

To prepare the data sets for analysis, the sqldf[14] and dplyr[15] packages were used to transform the data and join the data.

In the first analysis, I used principal component analysis from the stats package in R.[16] The prcomp function uses singular value decomposition to create principal components.[17] The first and second principal components were then used to create a linear regression.

In the second analysis, I created decision trees that utilized boosting algorithms to group countries by rank and clinical trials by thresholds. The ada[18] and adabag[19] packages were used to conduct the boosting analyses. The Naïve Bayes classification analyses utilized the e1071[20] package.

R CODE

The first step was to import the ADIS dataset and Worldwide Governance Indicators,[21] then join them using SQL through the sqldf package and total up the frequency counts of clinical trials.

```
countries <- read.csv ("~/Dropbox (Personal)/School/Machine Learning/Project/ADIS_Countries_
Only.csv", header = TRUE)
head (countries)

##      country
## 1 Afghanistan
## 2 Afghanistan
## 3 Afghanistan
## 4 Afghanistan
## 5 Afghanistan
## 6 Afghanistan

dim (countries)

## [1] 395163      1
```

```
total.countries <- table (countries)
countries <- as.data.frame (total.countries, drop = FALSE)
head (countries)
```

```
##    countries Freq
## 1 Afghanistan   11
## 2    Algeria   31
## 3     Angola    6
## 4   Argentina 3111
## 5    Armenia   27
## 6   Australia 6806
```

```
write.csv (countries, file ("country_frequency"))

getwd ()
```

```
## [1] "/Users/dannywhalen/Dropbox (Personal)/School/Machine Learning/Project"
```

```
countries <- (mutate_each (countries, funs (toupper)))


indicators <- read.csv ("~/Dropbox (Personal)/School/Machine Learning/Project/CountryIndicators.csv", header = TRUE, stringsAsFactors = FALSE)


data <- sqldf ("SELECT*
          FROM countries
          JOIN indicators
          ON countries.countries=indicators.Country;")
```

```
## Loading required package: tcltk
```

```
dim (data)
```

```
## [1] 135  10
```

```
head (data)
```

```
##    countries Freq    Country VoiceAndAccountabilityRank
## 1 AFGHANISTAN   11 AFGHANISTAN                   15.76
## 2    ALGERIA   31    ALGERIA                   22.66
## 3     ANGOLA    6     ANGOLA                   16.75
## 4  ARGENTINA 3111  ARGENTINA                   58.62
## 5    ARMENIA   27    ARMENIA                   30.54
## 6  AUSTRALIA 6806  AUSTRALIA                   93.60
##   PoliticalStabilityRank GovernmentEffectivnessRank RegulatoryQualityRank
## 1                   2.91                      8.17                 11.54
## 2                  10.19                     33.65                  9.62
## 3                  34.47                     12.98                 16.83
## 4                  49.03                     45.67                 12.98
## 5                  37.86                     46.15                 60.10
## 6                  87.38                     91.83                 98.08
```

```
##   RuleOfLawRank ControlOfCorruptionRank  X
## 1         2.40                  6.25 NA
## 2        25.48                 31.73 NA
## 3        11.06                  3.37 NA
## 4        18.27                 33.17 NA
## 5        43.75                 40.38 NA
## 6        96.15                 95.19 NA
```

```
data <- data [ , -3]
data <- data [ , -9]
```

```
head (data)
```

```
##     countries Freq VoiceAndAccountabilityRank PoliticalStabilityRank
## 1 AFGHANISTAN   11                    15.76                2.91
## 2     ALGERIA   31                    22.66               10.19
## 3      ANGOLA    6                    16.75               34.47
## 4   ARGENTINA 3111                    58.62               49.03
## 5     ARMENIA   27                    30.54               37.86
## 6   AUSTRALIA 6806                    93.60               87.38
##   GovernmentEffectivnessRank RegulatoryQualityRank RuleOfLawRank
## 1                       8.17                11.54         2.40
## 2                      33.65                 9.62        25.48
## 3                      12.98                16.83        11.06
## 4                      45.67                12.98        18.27
## 5                      46.15                60.10        43.75
## 6                      91.83                98.08        96.15
##   ControlOfCorruptionRank
## 1                    6.25
## 2                   31.73
## 3                    3.37
## 4                   33.17
## 5                   40.38
## 6                   95.19
```

Next, the data was converted into numeric objects so it could be analyzed and then split into training and test data sets.

```
data$Freq <- as.numeric (data$Freq)
```

```
data$VoiceAndAccountabilityRank <- as.numeric (data$VoiceAndAccountabilityRank)
```

```
## Warning: NAs introduced by coercion
```

```
data$PoliticalStabilityRank <- as.numeric (data$PoliticalStabilityRank)
```

```
## Warning: NAs introduced by coercion
```

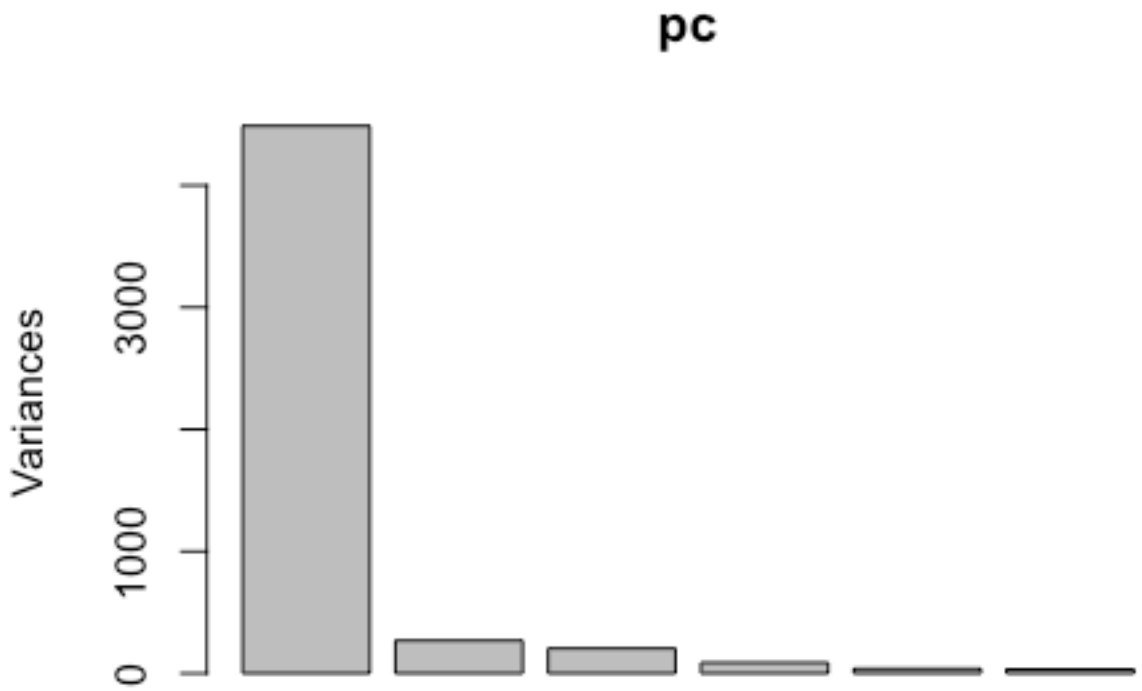
```
data$GovernmentEffectivnessRank <- as.numeric (data$GovernmentEffectivnessRank)
```

```
## Warning: NAs introduced by coercion
```

```
data$RegulatoryQualityRank <- as.numeric (data$RegulatoryQualityRank)

## Warning: NAs introduced by coercion

data$RuleOfLawRank <- as.numeric (data$RuleOfLawRank)

## Warning: NAs introduced by coercion

data$ControlOfCorruptionRank <- as.numeric (data$ControlOfCorruptionRank)

## Warning: NAs introduced by coercion

train <- data[1:67,]
test <- data[68:135,]
head (train)
```

Then, a principal component analysis was run on the governance indicators.

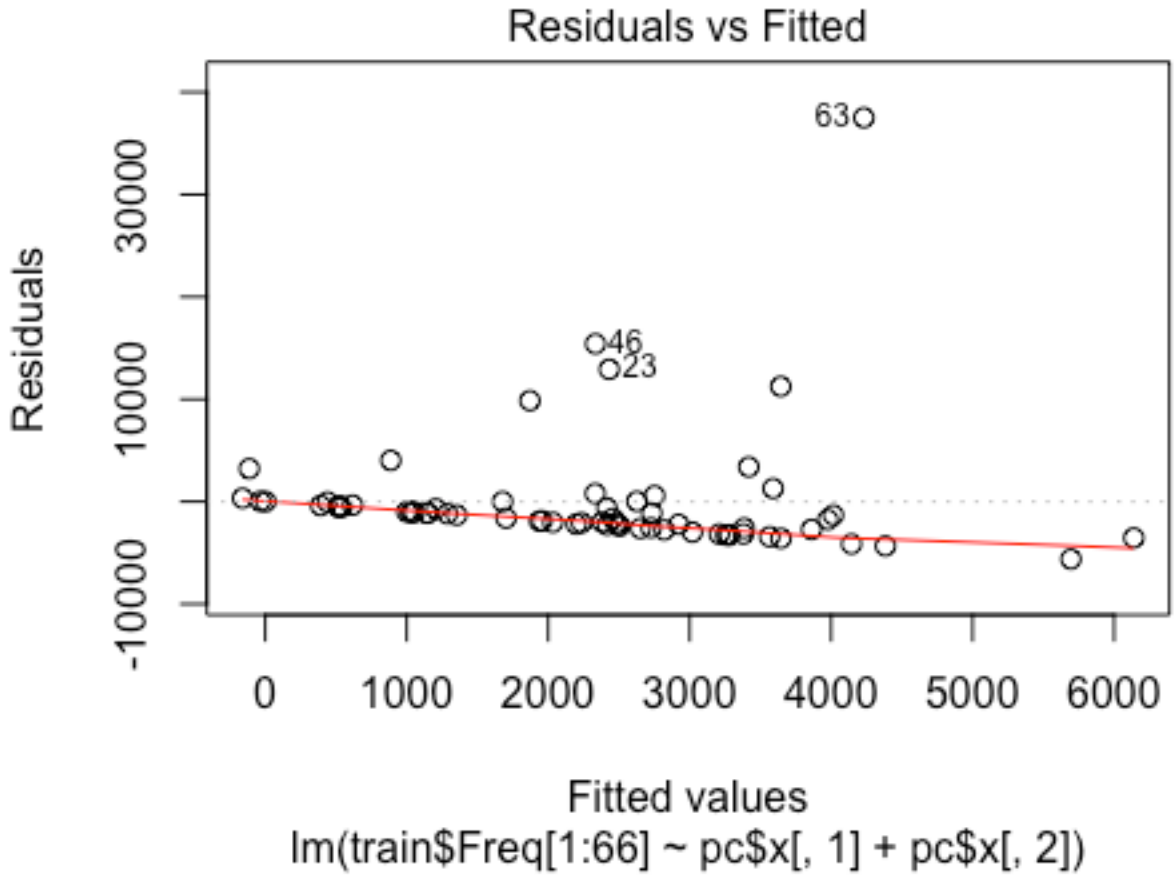
```
pc <- prcomp (na.omit(train [ , -(c(1:2))]))


summary (pc)

## Importance of components:
##                         PC1     PC2     PC3    PC4    PC5    PC6
## Standard deviation     66.9905 16.39935 14.29849 9.3857 6.20878 5.7702
## Proportion of Variance  0.8763  0.05252  0.03992 0.0172 0.00753 0.0065
## Cumulative Proportion   0.8763  0.92885  0.96877 0.9860 0.99350 1.0000

plot (pc)
```

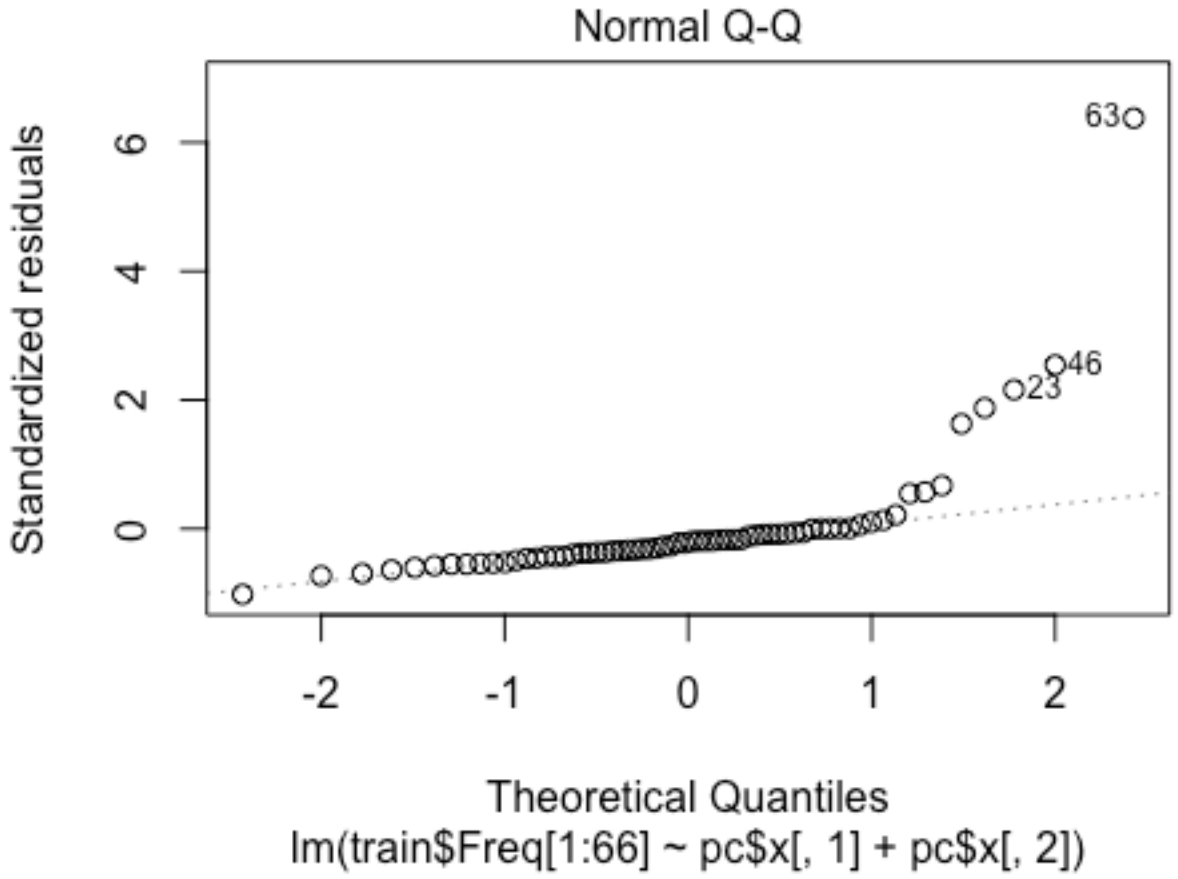


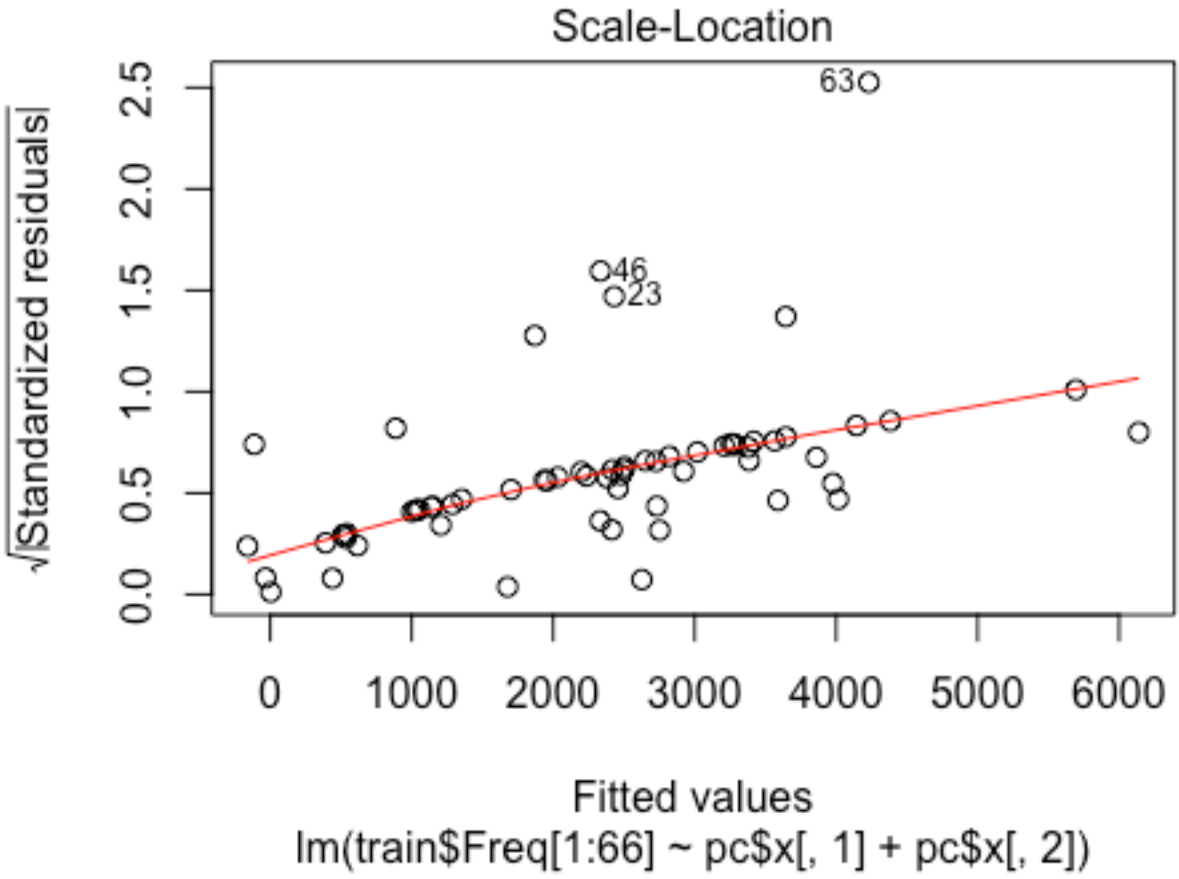

```
pc

## Standard deviations:
## [1] 66.990473 16.399345 14.298485  9.385669  6.208777  5.770177
##
## Rotation:
##                                PC1       PC2       PC3        PC4
## VoiceAndAccountabilityRank 0.3969221  0.2879511 -0.82608865  0.26783728
## PoliticalStabilityRank     0.3449369  0.8245020  0.36652715 -0.25068102
## GovernmentEffectivnessRank 0.4072261 -0.1964538  0.22700935  0.10801994
## RegulatoryQualityRank      0.4221082 -0.3292633 -0.15754533 -0.76627692
## RuleOfLawRank              0.4421866 -0.2594445  0.09293454  0.09489813
## ControlOfCorruptionRank    0.4288752 -0.1515280  0.31344026  0.50751151
##                                PC5        PC6
## VoiceAndAccountabilityRank -0.06046957  0.04148704
## PoliticalStabilityRank      0.03784343 -0.05098731
## GovernmentEffectivnessRank -0.85462253  0.04462766
## RegulatoryQualityRank       0.15271292  0.27944279
## RuleOfLawRank               0.26494206 -0.80580566
## ControlOfCorruptionRank     0.41354094  0.51601983
```

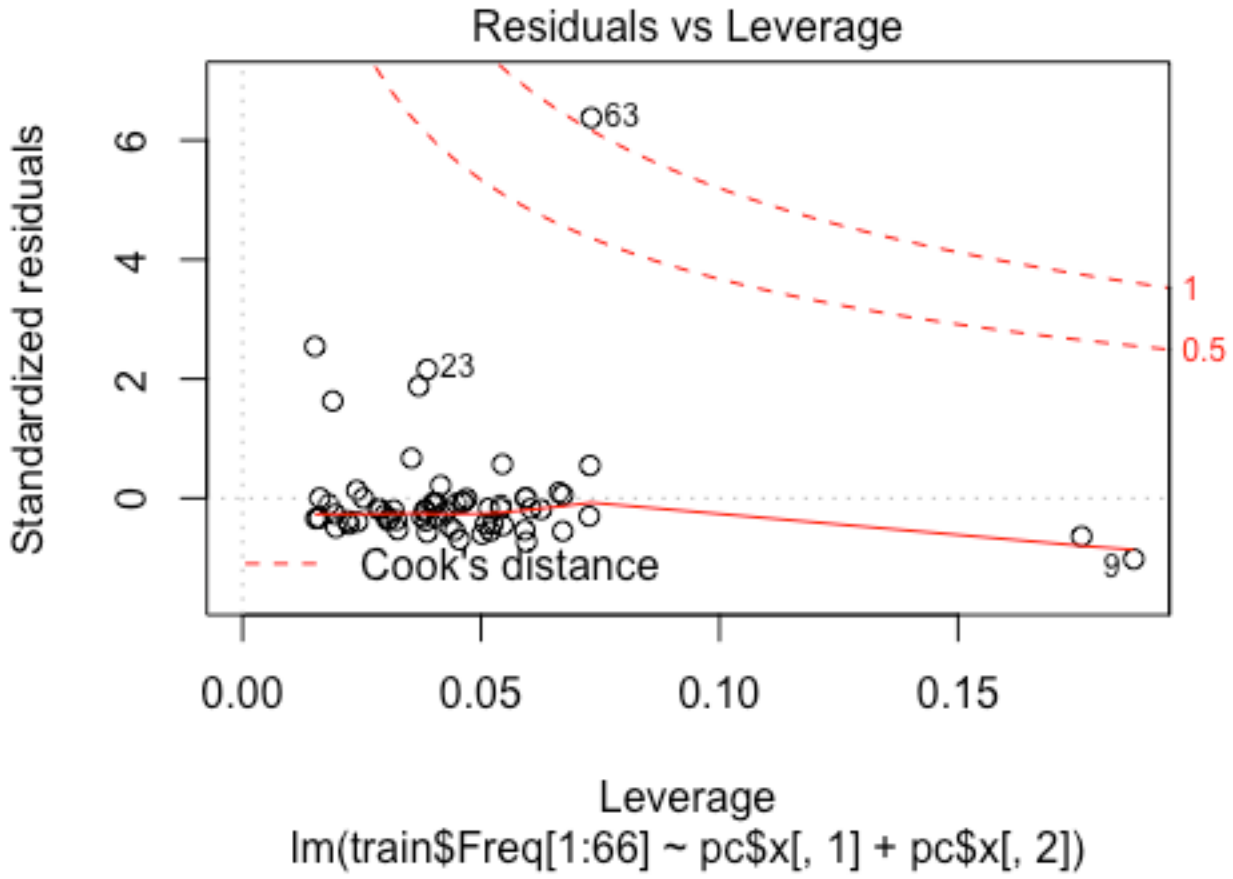
A linear regression was created from the first two principal components. The results section explains the reasoning behind using only the first two principal components.

```
reg <- lm(train$Freq [1:66] ~ pc$x [ ,1] + pc$x [ ,2])
reg

##
## Call:
## lm(formula = train$Freq[1:66] ~ pc$x[, 1] + pc$x[, 2])
##
## Coefficients:
## (Intercept)    pc$x[, 1]    pc$x[, 2]
##    2281.83       13.97       -61.39

summary (reg)

Call:

lm(formula = train$Freq[1:66] ~ pc$x[, 1] + pc$x[, 2])


Residuals:

  Min    1Q Median    3Q    Max

-4534  -2277  -1306   -373  39121


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 2281.83333  754.77016   3.023  0.00361 **

pc$x[, 1]    -0.05814    0.12333  -0.471  0.63895

pc$x[, 2]   -20.06721   12.59596  -1.593  0.11613

---

plot (reg)
```


Residuals vs Fitted
Residuals
30000
10000
-10000
63
46
23
0
1000
2000
3000
4000
5000
6000
Fitted values
lm(train$Freq[1:66] ~ pc$x[, 1] + pc$x[, 2])

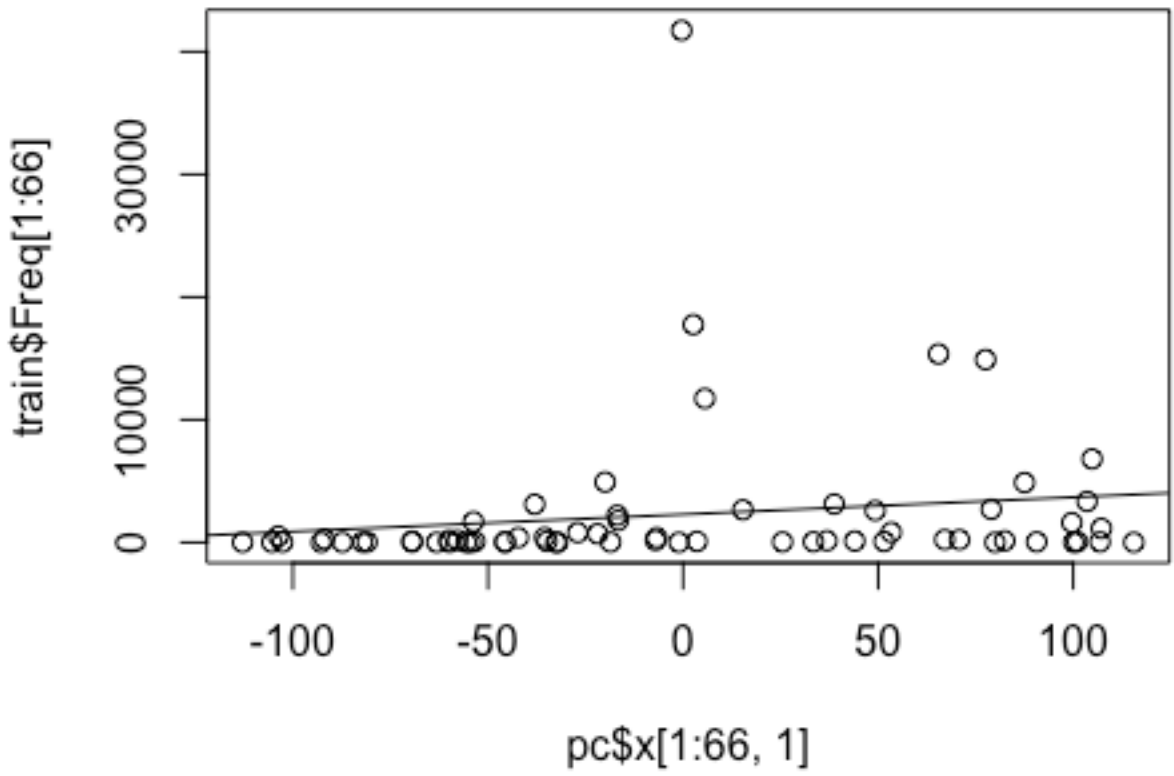

Normal Q-Q
Standardized residuals
0
2
4
6
63
46
23
-2
-1
0
1
2
Theoretical Quantiles
lm(train$Freq[1:66] ~ pc$x[, 1] + pc$x[, 2])


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
63
46
23
0
1000
2000
3000
4000
5000
6000
Fitted values
lm(train$Freq[1:66] ~ pc$x[, 1] + pc$x[, 2])



The linear regression was then plotted over graph that compared the first principal component aganist the frequency of the clinical trials.

```
plot (train$Freq [1:66] ~ pc$x[ 1:66,1])
abline (reg)

## Warning in abline(reg): only using the first two of 3 regression
## coefficients
```

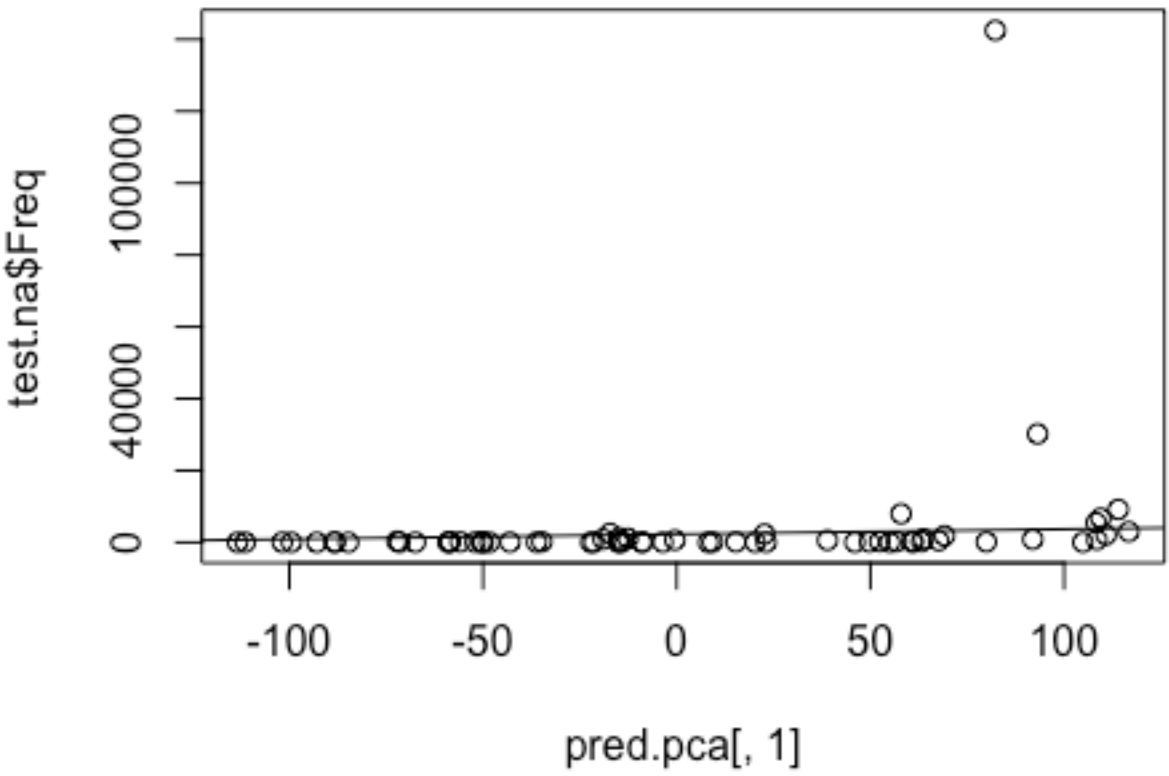


The principal components were then predicted for the test data and then predicted for the linear regression.

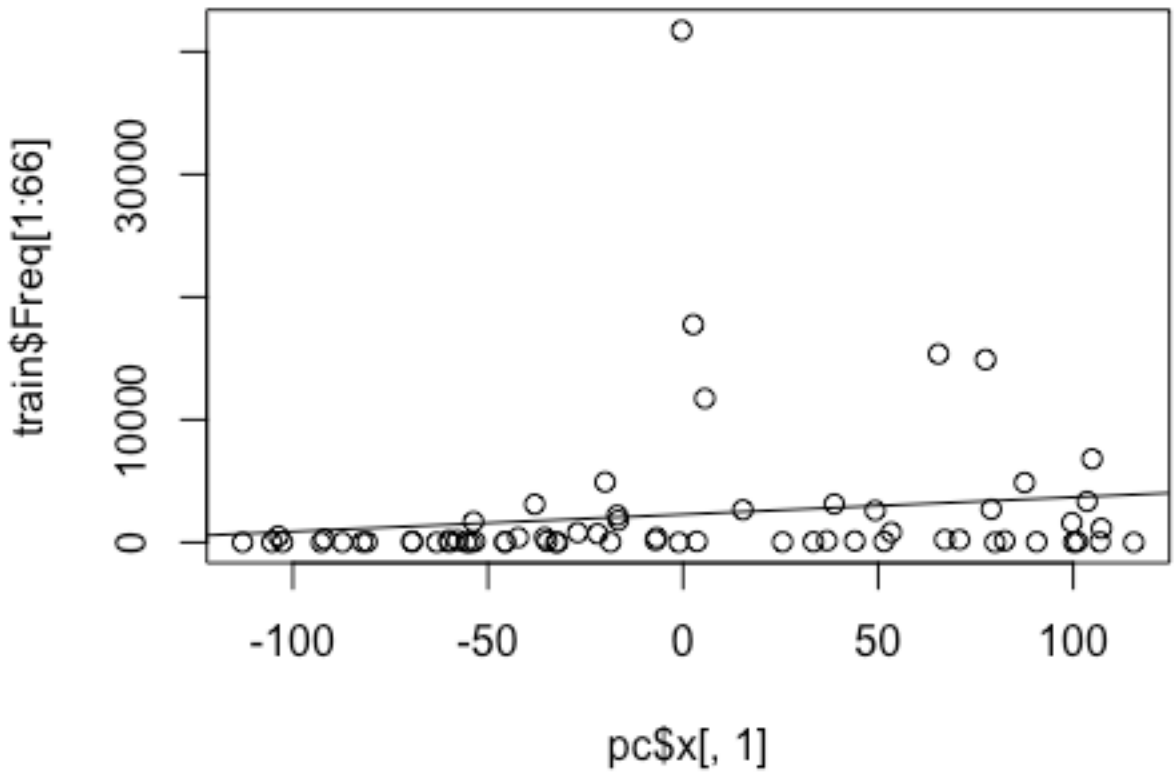
```
pred.pca <- predict (pc, test [ ,-(c(1:2))])
pred.pca

prediction <- predict(reg, data.frame (pred.pca [ , 1:2]))
prediction
```

Finally, the predictions and training data were graphed with the first principal component verse the frequency of the clinical trials

```
plot (test.na$Freq ~ pred.pca [ ,1])
abline (reg)

## Warning in abline(reg): only using the first two of 3 regression
## coefficients
```




```
plot (train$Freq [1:66] ~ pc$x [ ,1])
abline (reg)

## Warning in abline(reg): only using the first two of 3 regression
## coefficients
```



The second analysis used Naïve Bayes and Boosting. Here, the frequency of the clinical trials was split into groups, either under 25, 25-100, 100-1,000 and over 1,000, or over 1,000 and under 1,000. The ranking of the governance indicators were split into quartiles: 0-25, 25-50,50-75 and 75-100.

Here the data is grouped and split into training and testing sets.

```
data$Freq <- as.numeric (data$Freq)
data$over1000 [data$Freq > 1000 ] <- "1"
data$over1000 [data$Freq < 1000 ] <- "0"

data$VoiceAndAccountabilityRank <- as.numeric (data$VoiceAndAccountabilityRank)

## Warning: NAs introduced by coercion

data$Voice [data$VoiceAndAccountabilityRank > 25 ] <- "1"
da
```

```
ta$Voice [data$VoiceAndAccountabilityRank < 25 ] <- "0"
data$Voice [data$VoiceAndAccountabilityRank > 50 ] <- "2"
data$Voice [data$VoiceAndAccountabilityRank > 75 ] <- "3"

data$PoliticalStabilityRank <- as.numeric (data$PoliticalStabilityRank)

## Warning: NAs introduced by coercion

data$Stability [data$PoliticalStabilityRank > 25 ] <- "1"
data$Stability [data$PoliticalStabilityRank < 25 ] <- "0"
data$Stability [data$PoliticalStabilityRank > 50 ] <- "2"
data$Stability [data$PoliticalStabilityRank > 75 ] <- "3"

data$GovernmentEffectivnessRank <- as.numeric (data$GovernmentEffectivnessRank)

## Warning: NAs introduced by coercion

data$GovEffect [data$GovernmentEffectivnessRank > 25 ] <- "1"
data$GovEffect [data$GovernmentEffectivnessRank < 25 ] <- "0"
data$GovEffect [data$GovernmentEffectivnessRank > 50 ] <- "2"
data$GovEffect [data$GovernmentEffectivnessRank > 75 ] <- "3"

data$RegulatoryQualityRank <- as.numeric (data$RegulatoryQualityRank)

## Warning: NAs introduced by coercion

data$RegQual [data$RegulatoryQualityRank > 25 ] <- "1"
data$RegQual [data$RegulatoryQualityRank < 25 ] <- "0"
data$RegQual [data$RegulatoryQualityRank > 50 ] <- "2"
data$RegQual [data$RegulatoryQualityRank > 75 ] <- "3"

data$RuleOfLawRank <- as.numeric (data$RuleOfLawRank)

## Warning: NAs introduced by coercion

data$Law [data$RuleOfLawRank > 25 ] <- "1"
data$Law [data$RuleOfLawRank < 25 ] <- "0"
data$Law [data$RuleOfLawRank > 50 ] <- "2"
data$Law [data$RuleOfLawRank > 75 ] <- "3"

data$ControlOfCorruptionRank <- as.numeric (data$ControlOfCorruptionRank)

## Warning: NAs introduced by coercion

data$Corrupt [data$ControlOfCorruptionRank > 25 ] <- "1"
data$Corrupt [data$ControlOfCorruptionRank < 25 ] <- "0"
data$Corrupt [data$ControlOfCorruptionRank > 50 ] <- "2"
data$Corrupt [data$ControlOfCorruptionRank > 75 ] <- "3"
train <- data.1000[1:70,]
test <- data.1000[70:129,]
```

Then the data is processed through Naïve Bayes and Boosting algorithms.

```
country_classifier <- naiveBayes(train [ , -1], train$over1000)
country_classifier

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train[, -1], y = train$over1000)
##
## A-priori probabilities:
## train$over1000
##       0       1
## 0.7285714 0.2714286
##
## Conditional probabilities:
##             Voice
## train$over1000        0        1        2        3
##            0 0.35294118 0.25490196 0.25490196 0.13725490
##            1 0.05263158 0.00000000 0.31578947 0.63157895
##
##             Stability
## train$over1000        0        1        2        3
##            0 0.33333333 0.33333333 0.25490196 0.07843137
##            1 0.10526316 0.21052632 0.21052632 0.47368421
##
##             GovEffect
## train$over1000       0       1       2       3
##            0 0.2549020 0.3529412 0.2352941 0.1568627
##            1 0.0000000 0.1578947 0.2105263 0.6315789
##
##             RegQual
## train$over1000        0        1        2        3
##            0 0.33333333 0.25490196 0.23529412 0.17647059
##            1 0.05263158 0.10526316 0.21052632 0.63157895
##
##             Law
## train$over1000        0        1        2        3
##            0 0.33333333 0.33333333 0.19607843 0.13725490
##            1 0.05263158 0.05263158 0.26315789 0.63157895
##
##             Corrupt
## train$over1000       0       1       2       3
##            0 0.3725490 0.3137255 0.1372549 0.1764706
##            1 0.0000000 0.2105263 0.2105263 0.5789474

country_test_pred <- predict(country_classifier, test)
CrossTable(country_test_pred, test$over1000,
       prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
       dnn = c('predicted', 'actual'))
```

```
##
##
##   Cell Contents
## |-------------------------|
## |                       N |
## |         N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  60
##
##
##              | actual
##    predicted |        0 |        1 | Row Total |
## -------------|----------|----------|----------|
##            0 |       34 |        4 |       38 |
##              |    0.723 |    0.308 |          |
## -------------|----------|----------|----------|
##            1 |       13 |        9 |       22 |
##              |    0.277 |    0.692 |          |
## -------------|----------|----------|----------|
## Column Total |       47 |       13 |       60 |
##              |    0.783 |    0.217 |          |
## -------------|----------|----------|----------|
##
##
```

```
classifier2 <- naiveBayes(train [ ,-1], train$over1000, laplace = 1)
test_pred2 <- predict(classifier2, test)
CrossTable(test_pred2, test$over1000,
        prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
        dnn = c('predicted', 'actual'))
```

```
##
##
##   Cell Contents
## |-------------------------|
## |                       N |
## |         N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  60
##
##
##              | actual
##    predicted |        0 |        1 | Row Total |
## -------------|----------|----------|----------|
##            0 |       33 |        4 |       37 |
##              |    0.702 |    0.308 |          |
```

```
## -------------|-----------|-----------|-----------|
##            1 |       14 |        9 |       23 |
##              |    0.298 |    0.692 |          |
## -------------|-----------|-----------|-----------|
## Column Total |       47 |       13 |       60 |
##              |    0.783 |    0.217 |          |
## -------------|-----------|-----------|-----------|
##
##

##Boosting Comparison

country_boosting1 <- boosting(over1000~., data=train, mfinal=20,
                     control=rpart.control(maxdepth=5))
country_predict1 <- predict.boosting(country_boosting1, newdata=test)
country_predict1$confusion

##               Observed Class
## Predicted Class  0  1
##               0 41  6
##               1  6  7

accuracy <- 1- country_predict1$error
accuracy

## [1] 0.8

# ada package
country_boosting2 <- ada(over1000~., data=train,
                  iter=50,nu=1)
country_predict2 <- predict(country_boosting2, test)
country_predict_confusion <- confusionMatrix(country_predict2, test$over1000)
country_predict_confusion$table

##          Reference
## Prediction  0  1
##          0 40  7
##          1  7  6

accuracy <- country_predict_confusion$overall[1]
accuracy

##  Accuracy
## 0.7666667
```

Next another analysis was performed with the frequencies split into 4 groups instead of two.

```
data$over1000 [data$Freq > 25 ] <- "1"
data$over1000 [data$Freq > 100 ] <- "2"
data$over1000 [data$Freq > 1000 ] <- "3"
```

```
data$over1000 [data$Freq < 25 ] <- "0"

country_classifier <- naiveBayes(train [ , -1], train$over1000)
country_classifier

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train[, -1], y = train$over1000)
##
## A-priori probabilities:
## train$over1000
##       0       1       2       3
## 0.2571429 0.1857143 0.2857143 0.2714286
##
## Conditional probabilities:
##            Voice
## train$over1000        0        1        2        3
##           0 0.55555556 0.11111111 0.22222222 0.11111111
##           1 0.53846154 0.23076923 0.23076923 0.00000000
##           2 0.05000000 0.40000000 0.30000000 0.25000000
##           3 0.05263158 0.00000000 0.31578947 0.63157895
##
##            Stability
## train$over1000        0        1        2        3
##           0 0.50000000 0.27777778 0.05555556 0.16666667
##           1 0.23076923 0.53846154 0.23076923 0.00000000
##           2 0.25000000 0.25000000 0.45000000 0.05000000
##           3 0.10526316 0.21052632 0.21052632 0.47368421
##
##            GovEffect
## train$over1000        0        1        2        3
##           0 0.50000000 0.27777778 0.05555556 0.16666667
##           1 0.07692308 0.53846154 0.38461538 0.00000000
##           2 0.15000000 0.30000000 0.30000000 0.25000000
##           3 0.00000000 0.15789474 0.21052632 0.63157895
##
##            RegQual
## train$over1000        0        1        2        3
##           0 0.55555556 0.22222222 0.05555556 0.16666667
##           1 0.30769231 0.23076923 0.38461538 0.07692308
##           2 0.15000000 0.30000000 0.30000000 0.25000000
##           3 0.05263158 0.10526316 0.21052632 0.63157895
##
##            Law
## train$over1000        0        1        2        3
##           0 0.50000000 0.27777778 0.11111111 0.11111111
##           1 0.30769231 0.38461538 0.30769231 0.00000000
##           2 0.20000000 0.35000000 0.20000000 0.25000000
```

```
##           3 0.05263158 0.05263158 0.26315789 0.63157895
##
##            Corrupt
## train$over1000        0        1        2        3
##           0 0.61111111 0.16666667 0.00000000 0.22222222
##           1 0.15384615 0.53846154 0.23076923 0.07692308
##           2 0.30000000 0.30000000 0.20000000 0.20000000
##           3 0.00000000 0.21052632 0.21052632 0.57894737

country_test_pred <- predict(country_classifier, test)
CrossTable(country_test_pred, test$over1000,
        prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
        dnn = c('predicted', 'actual'))

##
##
##   Cell Contents
## |-------------------------|
## |                       N |
## |         N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  60
##
##
##            | actual
##   predicted |        0 |        1 |        2 |        3 | Row Total |
## ------------|----------|----------|----------|----------|----------|
##           0 |        6 |        5 |        2 |        0 |       13 |
##             |    0.429 |    0.357 |    0.105 |    0.000 |          |
## ------------|----------|----------|----------|----------|----------|
##           1 |        2 |        3 |        7 |        4 |       16 |
##             |    0.143 |    0.214 |    0.368 |    0.308 |          |
## ------------|----------|----------|----------|----------|----------|
##           2 |        3 |        5 |        1 |        0 |        9 |
##             |    0.214 |    0.357 |    0.053 |    0.000 |          |
## ------------|----------|----------|----------|----------|----------|
##           3 |        3 |        1 |        9 |        9 |       22 |
##             |    0.214 |    0.071 |    0.474 |    0.692 |          |
## ------------|----------|----------|----------|----------|----------|
## Column Total |       14 |       14 |       19 |       13 |       60 |
##             |    0.233 |    0.233 |    0.317 |    0.217 |          |
## ------------|----------|----------|----------|----------|----------|
##
##

classifier2 <- naiveBayes(train [ ,-1], train$over1000, laplace = 1)
test_pred2 <- predict(classifier2, test)
CrossTable(test_pred2, test$over1000,
```

```
      prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
      dnn = c('predicted', 'actual'))

##
##
##   Cell Contents
## |-------------------------|
## |                       N |
## |         N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  60
##
##
##              | actual
##    predicted |         0 |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            0 |         7 |         5 |         2 |         0 |        14 |
##              |     0.500 |     0.357 |     0.105 |     0.000 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            1 |         1 |         2 |         5 |         3 |        11 |
##              |     0.071 |     0.143 |     0.263 |     0.231 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            2 |         3 |         6 |         3 |         1 |        13 |
##              |     0.214 |     0.429 |     0.158 |     0.077 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            3 |         3 |         1 |         9 |         9 |        22 |
##              |     0.214 |     0.071 |     0.474 |     0.692 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |        14 |        14 |        19 |        13 |        60 |
##              |     0.233 |     0.233 |     0.317 |     0.217 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##
##

##Boosting Comparison

country_boosting1 <- boosting(over1000~., data=train, mfinal=20,
                     control=rpart.control(maxdepth=5))
country_predict1 <- predict.boosting(country_boosting1, newdata=test)
country_predict1$confusion

##                Observed Class
## Predicted Class  0  1  2  3
##               0  4  1  0  0
##               1  2  1  5  0
##               2  6 10  8  7
##               3  2  2  6  6
```

```
accuracy <- 1- country_predict1$error
accuracy
```

```
## [1] 0.3166667
```

As a variation to the second analysis, GDP was also factored in as an indicator along side the Worldwide Governance Indicators. First the GDP data was joined to the existing data through the use of SQL.

```
GDP <- read.csv ("~/Dropbox (Personal)/School/Machine Learning/Project/GDP by
 country v2.csv", header = TRUE, stringsAsFactors = FALSE)

head (GDP)
```

```
##          Country GDP.in.Millions.USD  X X.1 X.2
## 1  United States            17419000 NA  NA  NA
## 2          China            10360105 NA  NA  NA
## 3          Japan             4601461 NA  NA  NA
## 4        Germany             3852556 NA  NA  NA
## 5 United Kingdom             2941886 NA  NA  NA
## 6         France             2829192 NA  NA  NA
```

```
GDP <- GDP [1:194 ,1:2]
GDP <- GDP [-189, ]
GDP <- GDP [-95, ]
GDP <- (mutate_each (GDP, funs (toupper)))
data <- sqldf ("SELECT*
               FROM GDP
               JOIN data
               ON GDP.Country=data.Country;")
```

```
data <- data[ , c (1, 2, 4, 6:11)]


head (data)
```

```
##          Country GDP.in.Millions.USD   Freq VoiceAndAccountabilityRank
## 1  UNITED STATES            17419000 142528                     79.80
## 2          CHINA            10360105   2212                      5.42
## 3          JAPAN             4601461  41734                     79.31
## 4        GERMANY             3852556  17751                     96.06
## 5 UNITED KINGDOM             2941886  30219                     92.12
## 6         FRANCE             2829192  14899                     89.16
##   PoliticalStabilityRank GovernmentEffectivnessRank RegulatoryQualityRank
## 1                  66.99                      89.90                 88.46
## 2                  29.61                      66.35                 45.19
## 3                  84.47                      97.12                 84.13
## 4                  79.13                      94.71                 94.23
## 5                  60.68                      92.79                 97.12
## 6                  59.22                      88.94                 82.21
##   RuleOfLawRank ControlOfCorruptionRank
```

```
## 1         89.90                    89.42
## 2         42.79                    47.12
## 3         89.42                    93.27
## 4         93.27                    94.71
## 5         94.23                    92.79
## 6         88.46                    87.98

data$over1000 [data$Freq > 1000 ] <- "1"

data$over1000 [data$Freq < 1000 ] <- "0"

train <- data.1000[c (6:75), ]

test <- data.1000[ c (1:5, 76:129),]
```

Next the data was analyzed with Naïve Bayes and Boosting algorithms for countries with over 1000 clinical trials versus countries with less than 1000 clinical trials.

```
country_classifier <- naiveBayes(train [ ,-8], train$over1000)
country_classifier

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train[, -8], y = train$over1000)
##
## A-priori probabilities:
## train$over1000
##         0         1
## 0.6142857 0.3857143
##
## Conditional probabilities:
##                GDP.in.Millions.USD
## train$over1000     [,1]     [,2]
##              0 210340.3 209926.3
##              1 837730.6 779885.3
##
##                VoiceAndAccountabilityRank
## train$over1000     [,1]     [,2]
##              0 37.43814 25.90930
##              1 79.41926 19.64118
##
##                PoliticalStabilityRank
## train$over1000     [,1]     [,2]
##              0 42.73023 27.84942
##              1 64.41889 28.18707
##
##                GovernmentEffectivnessRank
## train$over1000     [,1]     [,2]
##              0 50.63721 24.37070
```

```
##              1 80.04000 17.52163
##
##               RegulatoryQualityRank
## train$over1000     [,1]     [,2]
##              0 48.64744 28.09489
##              1 77.65333 21.58872
##
##               RuleOfLawRank
## train$over1000     [,1]     [,2]
##              0 46.90256 26.42761
##              1 78.09852 22.47007
##
##               ControlOfCorruptionRank
## train$over1000     [,1]     [,2]
##              0 44.99140 27.09546
##              1 73.21926 24.67475
```

```r
country_test_pred <- predict(country_classifier, test)
CrossTable(country_test_pred, test$over1000,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  59
##
##
##              | actual
##    predicted |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |        44 |         2 |        46 |
##              |     0.815 |     0.400 |           |
## -------------|-----------|-----------|-----------|
##            1 |        10 |         3 |        13 |
##              |     0.185 |     0.600 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        54 |         5 |        59 |
##              |     0.915 |     0.085 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```r
classifier2 <- naiveBayes(train [ ,-8], train$over1000, laplace = 1)
test_pred2 <- predict(classifier2, test)
```

```
CrossTable(test_pred2, test$over1000,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  59
##
##
##                | actual
##    predicted |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##             0 |        44 |         2 |        46 |
##               |     0.815 |     0.400 |           |
## -------------|-----------|-----------|-----------|
##             1 |        10 |         3 |        13 |
##               |     0.185 |     0.600 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        54 |         5 |        59 |
##               |     0.915 |     0.085 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
##Boosting Comparison

country_boosting1 <- boosting(over1000~., data=train, mfinal=20,
                              control=rpart.control(maxdepth=5))
country_predict1 <- predict.boosting(country_boosting1, newdata=test)
country_predict1$confusion
```

```
##                Observed Class
## Predicted Class  0  1
##               0 54  1
##               1  0  4
```

```
accuracy <- 1- country_predict1$error
accuracy
```

```
## [1] 0.9830508
```

Next, the analysis using GDP is rerun using the previously discussed 4-tier breakdown of clinical trials.

```
data$over1000 [data$Freq > 25 ] <- "1"
data$over1000 [data$Freq > 100 ] <- "2"
data$over1000 [data$Freq > 1000 ] <- "3"
data$over1000 [data$Freq < 25 ] <- "0"
```

```
train <- data.1000[c (6:75), ]

test <- data.1000[ c (1:5, 76:129),]

country_classifier <- naiveBayes(train [ ,-8], train$over1000)
country_classifier

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = train[, -8], y = train$over1000)
##
## A-priori probabilities:
## train$over1000
##         0         1         2         3
## 0.1142857 0.1000000 0.4000000 0.3857143
##
## Conditional probabilities:
##                GDP.in.Millions.USD
## train$over1000     [,1]      [,2]
##              0 114207.8  66915.29
##              1 186904.7 180201.52
##              2 243665.6 236944.18
##              3 837730.6 779885.26
##
##                VoiceAndAccountabilityRank
## train$over1000     [,1]      [,2]
##              0 13.30000  6.147348
##              1 25.68714 24.142437
##              2 47.27250 24.416726
##              3 79.41926 19.641185
##
##                PoliticalStabilityRank
## train$over1000     [,1]     [,2]
##              0 30.28000 30.38303
##              1 38.76429 26.88873
##              2 47.27893 27.11112
##              3 64.41889 28.18707
##
##                GovernmentEffectivnessRank
## train$over1000     [,1]     [,2]
##              0 32.45250 27.55095
```

```
##              1 42.10143 21.45738
##              2 57.96679 21.26507
##              3 80.04000 17.52163
##
##               RegulatoryQualityRank
## train$over1000     [,1]     [,2]
##              0 30.40875 28.38544
##              1 32.00714 27.07522
##              2 58.01857 24.38474
##              3 77.65333 21.58872
##
##               RuleOfLawRank
## train$over1000     [,1]     [,2]
##              0 32.39250 30.08121
##              1 32.41714 21.34979
##              2 54.66964 23.89967
##              3 78.09852 22.47007
##
##               ControlOfCorruptionRank
## train$over1000     [,1]     [,2]
##              0 29.32875 29.92529
##              1 32.89857 19.71014
##              2 52.48964 25.58561
##              3 73.21926 24.67475
```

```
country_test_pred <- predict(country_classifier, test)
CrossTable(country_test_pred, test$over1000,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  59
##
##
##              | actual
##    predicted |         0 |         1 |         2 |         3 | Row Total |

## -------------|-----------|-----------|-----------|-----------|-----------|
##            0 |         8 |         3 |         0 |         0 |        11 |

##              |     0.381 |     0.136 |     0.000 |     0.000 |           |
```

```
## -------------|-----------|-----------|-----------|-----------|-----------|
##            1 |         7 |         9 |         2 |         0 |        18 |

##              |     0.333 |     0.409 |     0.182 |     0.000 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
##            2 |         3 |        10 |         2 |         2 |        17 |

##              |     0.143 |     0.455 |     0.182 |     0.400 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
##            3 |         3 |         0 |         7 |         3 |        13 |

##              |     0.143 |     0.000 |     0.636 |     0.600 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |        21 |        22 |        11 |         5 |        59 |

##              |     0.356 |     0.373 |     0.186 |     0.085 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
##
##
```

```
classifier2 <- naiveBayes(train [ ,-8], train$over1000, laplace = 1)
test_pred2 <- predict(classifier2, test)
CrossTable(test_pred2, test$over1000,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  59
##
##
##              | actual
##    predicted |         0 |         1 |         2 |         3 | Row Total |

## -------------|-----------|-----------|-----------|-----------|-----------|
##            0 |         8 |         3 |         0 |         0 |        11 |

##              |     0.381 |     0.136 |     0.000 |     0.000 |           |
```

```
## -------------|-----------|-----------|-----------|-----------|-----------|
##            1 |         7 |         9 |         2 |         0 |        18 |

##              |     0.333 |     0.409 |     0.182 |     0.000 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
##            2 |         3 |        10 |         2 |         2 |        17 |

##              |     0.143 |     0.455 |     0.182 |     0.400 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
##            3 |         3 |         0 |         7 |         3 |        13 |

##              |     0.143 |     0.000 |     0.636 |     0.600 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |        21 |        22 |        11 |         5 |        59 |

##              |     0.356 |     0.373 |     0.186 |     0.085 |           |

## -------------|-----------|-----------|-----------|-----------|-----------|
##
##
```

```
##Boosting Comparison

country_boosting1 <- boosting(over1000~., data=train, mfinal=20,
                              control=rpart.control(maxdepth=5))
country_predict1 <- predict.boosting(country_boosting1, newdata=test)
country_predict1$confusion
```

```
##                Observed Class
## Predicted Class  0  1  2  3
##               0  8  4  0  1
##               1  1  1  0  0
##               2 12 17 11  0
##               3  0  0  0  4
```

```
accuracy <- 1- country_predict1$error
accuracy
```

```
## [1] 0.4067797
```

RESULTS

The principal component analysis shows that the first principal component is responsible for 87.63% of the variance seen. The second principal component only explains 5.25% of the variance. The first two principal components therefore describe 92.88% of the variance. Since the first two principal components describe such a large percentage of the variance, further analysis was mainly conducted on these first two principal components. It

appears that all the indices are used to comprise the first principal component, but the Rule of Law ranking appears to carry the most weight.

The residual plots from the linear regression show that the data is not normal when regressed against the frequency of clinical trials. Furthermore, a number of outliers appear to exist in the data.

The final graphs show both the training data and test data plotted with the actual frequencies of clinical trials against the first principal component.

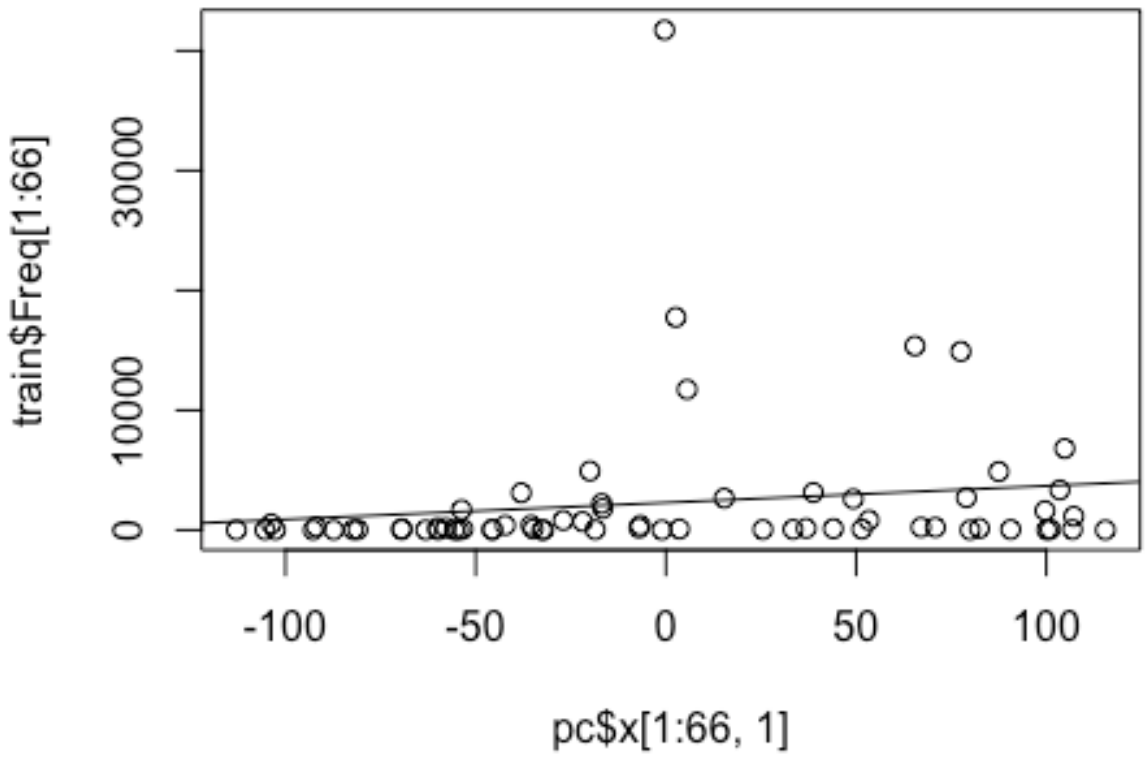
The second set of analyses used decision trees with boosting and Naïve Bayes algorithms to achieve a classifying algorithm that would group countries by the relative number of clinical trials they conducted. The addition of the GDP indicator raises the two tier system, over and under 1000 clinical trials from 76% to 98% and the 4-tier system sees a change from 31.6% to 40.6%

EVALUATION

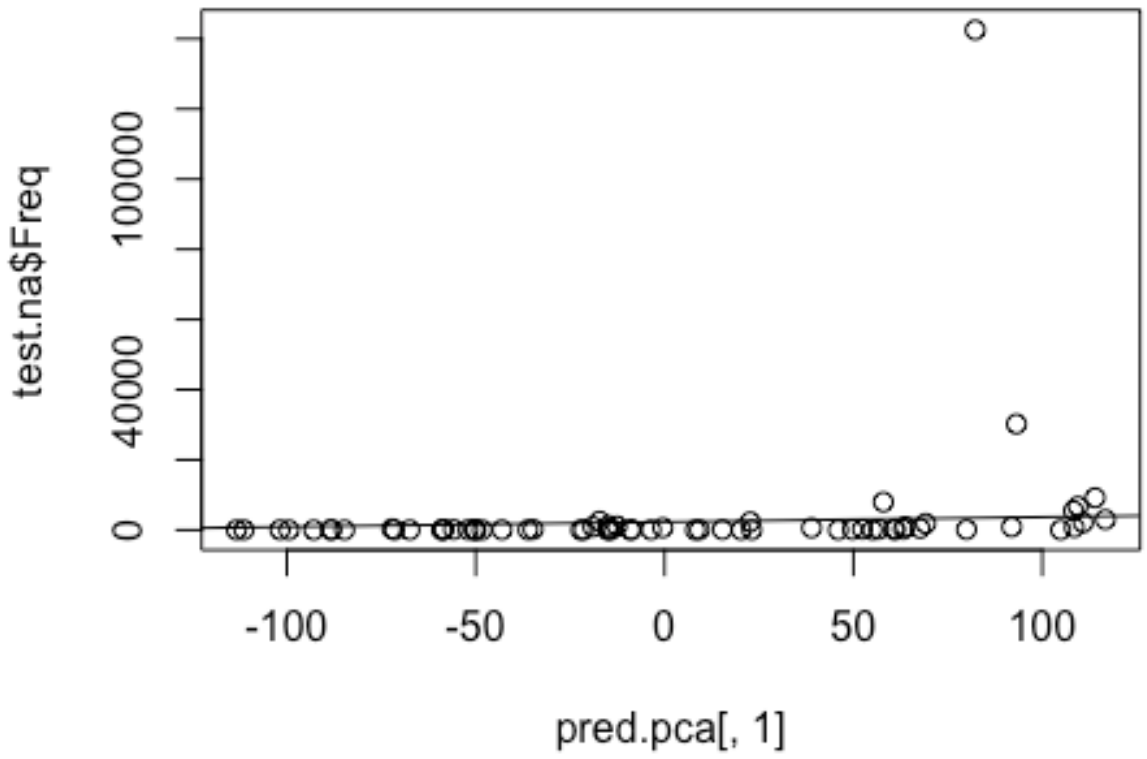
The success of this research will be evaluated based upon whether a principal component analysis or decision trees can be created that is able to predict the number of clinical trials a country will host based on country data indicators. The primary goal is to find certain indices that can generally predict the likelihood of a country being chosen as a host for clinical trials. Success can then be measured in how much variation the chosen indices can explain when examining the likelihood of a country being chosen. Success can also be partially evaluated in the number of indices needed in the principal component analysis to explain a majority of the variation; the fewer indices needed, the more likely that these indices have a real world relationship with the likelihood of being a clinical trial host and are not just the product of over-fitting.

DISCUSSION

The first analysis using Principal Component analysis did not yield very promising results. The training data does show a strong increase when plotted against the first principal component.



However, when the test data was predicted using the principal components, a less notable positive correlation was seen.



The summary of the linear regression did not show significance for using either the first or second principal component as an indicator for the frequency of clinical trials.

The Naïve Bayes classifier and decision trees with boosting showed far more promise in their abilities to predict the number of clinical trials. When posed with the question whether or not a country is likely to have over 1,000 clinical trials, the boosting algorithm was able to predict this classification with 80% accuracy. However, when more groups are added, such as countries with less than 25 clinical trials, between 25 and 100, between 100 and 1,000 and over 1,000, both the Naïve Bayes classification and the decision trees with boosting lost almost all of their accuracy. Furthermore, the boosting accuracy was at 31.6%, which most likely means that the boosting actually detracted from the overall algorithm instead of improving it. The addition of the GDP indicator substantially raises the accuracy of both the 2-tier system and 4-tier system. The 2-tier system sees a change from 76% to 98% and the 4-tier system sees a change from 31.6% to 40.6%. The concern remains with the boosting algorithm that the below 50% accuracy for the 4-tier system implies that the boosting algorithm may not actually be improving the results. The positive change in the accuracy calculation by adding the additionally indicator of GDP provides reason to believe that further indicators could continue to increase the accuracy of the

prediction thereby alleviating this concern. The future research section will discuss possible solutions to raise the accuracy of the predictions.

The data is constant with two findings. First, with the addition of the GDP indicator one can reasonably predict whether a country will have a large number of clinical trials, over 1000. This can be seen from the 98% accuracy of the boosting algorithm. Overall, it appeared that boosting has yielded the most promising results for using indicators to determine the number of clinical trials in a country. The second finding demonstrates that the current indices and methods are not capable of discerning threshold numbers of clinical trials as seen in the 4-tier system. Here countries were divided into tiers based on the number of clinical trials. The accuracy in this analysis was only 40% even when the GDP indicator was utilized. Furthermore, the PCA and linear regression analysis was not able to derive a statistically significant model for determining the number of clinical trials based on the indicators.

FUTURE RESEARCH

Future research could add other indicators, such as number of hospitals or population size to help raise the accuracy further. Additionally, adding other data sets of clinical trials, such as clinicaltrials.gov, which has data from 2008 through the present could help to increase accuracy with the addition of more data points. ClinicalTrials.gov is a database of clinical trials maintained by the U.S. National Institutes of Health. [22] In comparison, ADIS is a paid-for database of clinical trials from Springer, which was provided to me by Professor Dunlap.[23] The ADIS database has clinical trial data spanning 1996 to 2006. Other indices exist such as those from the World Health Organization's Global Health Observatory,[24] and the World Bank Indicators.[25] These other indices could be utilized as a means of increasing the accuracy of the predictions and deriving a better algorithm for predicting the number clinical trials

WORKS CITED

[1] Cockburn, I. M., & Slaughter, M. J. (2010). The global location of biopharmaceutical knowledge activity: new findings, new questions. In Innovation Policy and the Economy, Volume 10 (pp. 129-157). University of Chicago Press.
[2] World Bank. (2015) GDP Ranking. http://data.worldbank.org/data-catalog/GDP-ranking-table
[3] Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters VIII: aggregate and individual governance indicators, 1996-2008. World bank policy research working paper, (4978).
[4] Cockburn, I. M., & Slaughter, M. J. (2010). The global location of biopharmaceutical knowledge activity: new findings, new questions. In Innovation Policy and the Economy, Volume 10 (pp. 129-157). University of Chicago Press.
[5] Cockburn, I. M., & Slaughter, M. J. (2010). The global location of biopharmaceutical knowledge activity: new findings, new questions. In Innovation Policy and the Economy, Volume 10 (pp. 129-157). University of Chicago Press.
[6] Thiers, F. A., Sinskey, A. J., & Berndt, E. R. (2008). Trends in the globalization of clinical trials. Nature Reviews Drug Discovery, 7(1), 13-14.
[7] Thiers, F. A., Sinskey, A. J., & Berndt, E. R. (2008). Trends in the globalization of clinical trials. Nature Reviews Drug Discovery, 7(1), 13-14.
[8] Rodine‐Hardy, K. (2015). Globalization, International Organizations, and Telecommunications. Review of Policy Research, 32(5), 517-537.
[9] Grein, A. F., Sethi, S. P., & Tatum, L. G. (2008). A Dynamic Analysis of Country Clusters, the Role of Corruption, and Implications for Global Firms. Department of Marketing and International Business. URL: idec. gr/iier/new. com.
[10] Breunig, C., & Ahlquist, J. S. (2009). Country Clustering in Comparative Political Economy.
[11] Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters VIII: aggregate and individual governance indicators, 1996-2008. World bank policy research working paper, (4978).
[12] ADIS. (2015). AdisInisght Trials. http://www.springer.com/gp/adis/products-services/adisinsight-databases/clinical-trials-insight
[13] World Bank. (2015) GDP Ranking. http://data.worldbank.org/data-catalog/GDP-ranking-table
[14] Grothendieck, G., Grothendieck, M. G., Suggests, R. H., RMySQL, R., & DBI, I. (2014). Package 'sqldf'. Perfrom SQL Selects on R Data Frames.
[15] Wickham, H., & Francois, R. (2014). dplyr: A grammar of data manipulation. R package version 0.3. 0.2.
[16] Team, R. C. (2014). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[17] Team, R. C. (2014). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[18] Culp, M., Johnson, K., & Michailidis, G. (2006). ada: An r package for stochastic boosting. Journal of Statistical Software, 17(2), 9.
[19] Alfaro, E., Gámez, M., & Garcia, N. (2013). Adabag: An R package for classification with boosting and bagging. Journal of Statistical Software, 54(2), 1-35.
[20] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. J. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

[21] Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters VIII: aggregate and individual governance indicators, 1996-2008. World bank policy research working paper, (4978).
[22] United States National Institutes of Health. (2015). Clinicaltrials.gov. https://clinicaltrials.gov/
[23] Springer. (2015). ADIS. http://www.springer.com/gp/adis
[24] World Health Organization. (2015). Global Health Observatory indicator views. http://apps.who.int/gho/data/node.imr
[25] World Bank. (2015) Data: Indicators. http://data.worldbank.org/indicator

BIBLIOGRAPHY

ADIS. (2015). AdisInisght Trials. http://www.springer.com/gp/adis/products-services/adisinsight-databases/clinical-trials-insight

Alfaro, E., Gámez, M., & Garcia, N. (2013). Adabag: An R package for classification with boosting and bagging. Journal of Statistical Software, 54(2), 1-35.

Breunig, C., & Ahlquist, J. S. (2009). Country Clustering in Comparative Political Economy.

Culp, M., Johnson, K., & Michailidis, G. (2006). ada: An r package for stochastic boosting. Journal of Statistical Software, 17(2), 9.

Cockburn, I. M., & Slaughter, M. J. (2010). The global location of biopharmaceutical knowledge activity: new findings, new questions. In Innovation Policy and the Economy, Volume 10 (pp. 129-157). University of Chicago Press.

Grein, A. F., Sethi, S. P., & Tatum, L. G. (2008). A Dynamic Analysis of Country Clusters, the Role of Corruption, and Implications for Global Firms. Department of Marketing and International Business. URL: idec. gr/iier/new. com.

Grothendieck, G., Grothendieck, M. G., Suggests, R. H., RMySQL, R., & DBI, I. (2014). Package 'sqldf'. Perfrom SQL Selects on R Data Frames.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance matters VIII: aggregate and individual governance indicators, 1996-2008. World bank policy research working paper, (4978).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., & Lin, C. J. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

Springer. (2015). ADIS. http://www.springer.com/gp/adis

Team, R. C. (2014). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

Rodine‐Hardy, K. (2015). Globalization, International Organizations, and Telecommunications. Review of Policy Research, 32(5), 517-537.

Thiers, F. A., Sinskey, A. J., & Berndt, E. R. (2008). Trends in the globalization of clinical trials. Nature Reviews Drug Discovery, 7(1), 13-14.

United States National Institutes of Health. (2015). Clinicaltrials.gov. https://clinicaltrials.gov/

Wickham, H., & Francois, R. (2014). dplyr: A grammar of data manipulation. R package version 0.3. 0.2.

World Bank. (2015) Data: Indicators. http://data.worldbank.org/indicator

World Bank. (2015) GDP Ranking. http://data.worldbank.org/data-catalog/GDP-ranking-table

World Health Organization. (2015). Global Health Observatory indicator views. http://apps.who.int/gho/data/node.imr