Stat 515 Final Project Report in US Presidential Elections – 2016

Merin Joy – G01118158
Volgenau School of Engineering, George Mason University, 4400 University Dr,
Fairfax, VA, 22030

Abstract

The study about the US Presidential Elections – 2016 is most of the times centered on the results obtained rather than what factors lead to those results. From the day results about the Presidential Elections where announced it has been evident that those results either created concern or relief to people around the globe. This study however does not focus on the results, but on what where the major factors that lead to those results. The factors that lead to these results can be not just one soul variable, it can be a combination of factors that lead to the results obtained. In this study, learning about those factors that influenced the results positively or negatively is done using linear regression, with the help of a scatter plot matrix it can be determined which variables are co-related and which show surprising trends. Random forest variable importance plot helps in determining which variables or factors are the most significant among the lot. Also, the relationship between the results can be best explained using linked micromaps. All these comparisons are done by applying these methods to the US Presidential Elections – 2016 dataset and generating insights on those results.

Key Words: Linear regression, scatterplot matrix, micromaps, random forest, regression trees, response variable.

1. Introduction

The US Presidential Elections are one of the very well-known elections around the world as it influences the world economy in general. The US Presidential Elections-2016 was the 58th quadrennial American presidential election, held on Tuesday, November 8, 2016. When the winning candidate got the maximum electoral votes, the defeated candidate won the maximum popular votes. This was one election in which people where not sure of what the results or consequences are going to be. It is important to see in this case that which factors lead to determine these consequences in a positive or a negative way.

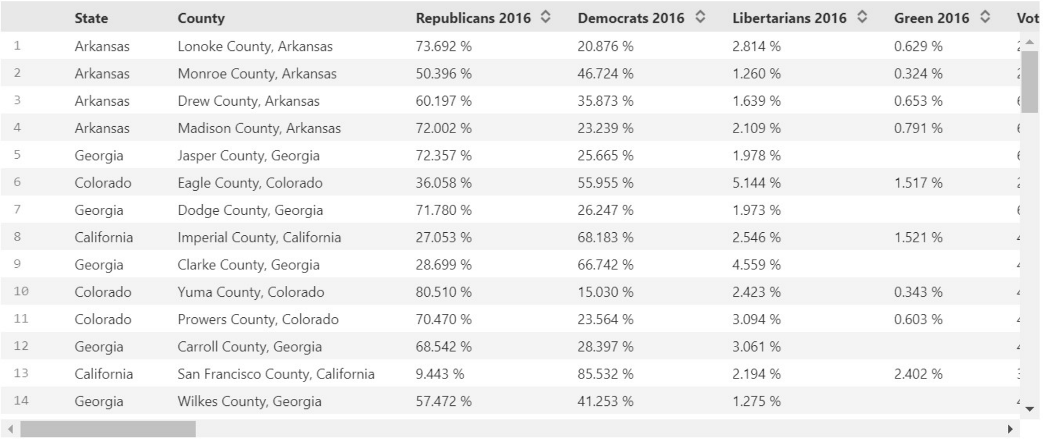
Linked micromaps appear in an ordered format and are best for comparisons between columns. By comparing the results obtained for each candidate using linked micromaps we can determine the relationship obtained between each result and determine their characteristics.

Linear regression models have been applied by finding out which response variable is the most suitable and how relationship between variables using scatterplot show interesting trends that might help in the linear regression models. By creating training and test datasets linear regression can be applied by predicting the test dataset and finding the RMSE value. Depending on the p-value we can remove the predictors who are less significant, also by applying backward elimination on the model, p-value can be decreased, have a better fstat value and the RMSE can be decreased as well.

The random forest variable importance plot specifies which predictors are the most influential and reflect most on the results obtained. The predictors are either cause a negative or a positive effect, but the variable importance plot gives us the variables that influence the results most. Using the US Presidential Elections-2016 dataset, this study confirms which factors lead in determining the results using R.

2. About the data

The dataset was obtained from the OpenDataSoft website under USA 2016 Presidential Election by County published by Deleetdk for public use. The dataset contains 159 columns and 3143 records, which later where filtered out by removing unimportant variables which do not contribute to our findings.



	State	County	Republicans 2016	Democrats 2016	Libertarians 2016	Green 2016	Vot
1	Arkansas	Lonoke County, Arkansas	73.692 %	20.876 %	2.814 %	0.629 %	2
2	Arkansas	Monroe County, Arkansas	50.396 %	46.724 %	1.260 %	0.324 %	2
3	Arkansas	Drew County, Arkansas	60.197 %	35.873 %	1.639 %	0.653 %	6
4	Arkansas	Madison County, Arkansas	72.002 %	23.239 %	2.109 %	0.791 %	6
5	Georgia	Jasper County, Georgia	72.357 %	25.665 %	1.978 %		6
6	Colorado	Eagle County, Colorado	36.058 %	55.955 %	5.144 %	1.517 %	2
7	Georgia	Dodge County, Georgia	71.780 %	26.247 %	1.973 %		6
8	California	Imperial County, California	27.053 %	68.183 %	2.546 %	1.521 %	4
9	Georgia	Clarke County, Georgia	28.699 %	66.742 %	4.559 %		4
10	Colorado	Yuma County, Colorado	80.510 %	15.030 %	2.423 %	0.343 %	4
11	Colorado	Prowers County, Colorado	70.470 %	23.564 %	3.094 %	0.603 %	4
12	Georgia	Carroll County, Georgia	68.542 %	28.397 %	3.061 %		4
13	California	San Francisco County, California	9.443 %	85.532 %	2.194 %	2.402 %	3
14	Georgia	Wilkes County, Georgia	57.472 %	41.253 %	1.275 %		4

Fig.2.1

3. Linked Micromaps

micromapST is a US linked micromap graphics which is easy and quick means of forming linked micromaps for the 51 US states and the District of Columbia. MicromapST uses standard graphics and RColorBrewer packages to create readable linked micromaps. This helps the user to explore different views of their data quickly. Each row in a micromap that's plotted represents a state. Each column can be used to define to present a different graphical representation of the user's data. For the linked maps, the main columns are MAP and state name (ID). The statistical data is presented in other columns as bars, dots, facet plots, horizontal stacked bars and scatter plots.

Fig 3.1 graph gives a clear representation of vote percentage obtained by the candidates in different states in the US. Linked micromaps helps in comparing the percentage obtained by different candidates and gives us insight into how the trends are for each person. It gives us the better understanding on what is happening for each candidate in an ordered format. Also, we can see an interesting trend between Republicans and Democrats, the vote percentages obtained by Republicans are like a mirror image of vote percentages obtained by Democrats, which further explains people's uncertainty of who was going to win the elections. The dot plot shows each state had a favorite when it came for voting for Democrats or Republicans. Whereas, trends for Greens and Libertarians are scattered and different from Democrats and Republicans but together they follow the same flow of vote percentages.

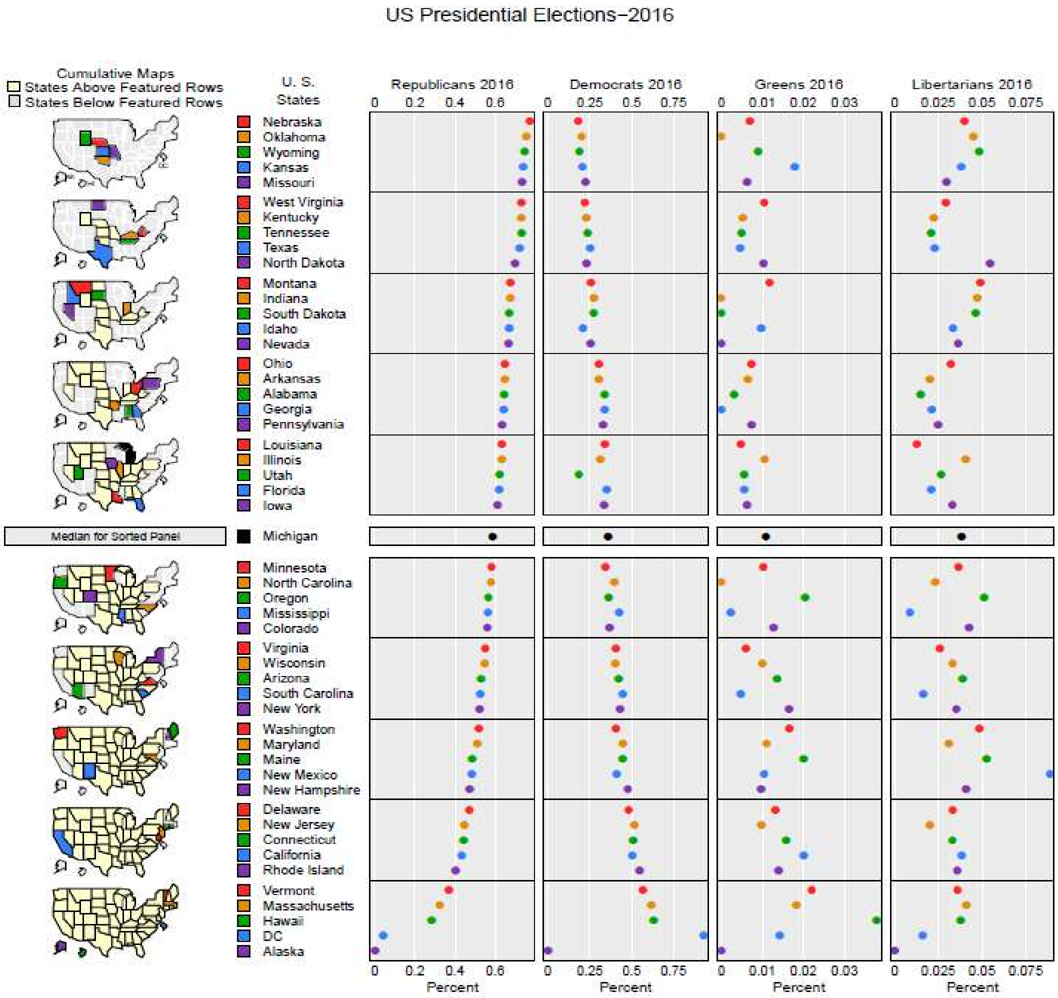


Fig. 3.1

4. Linear Regression

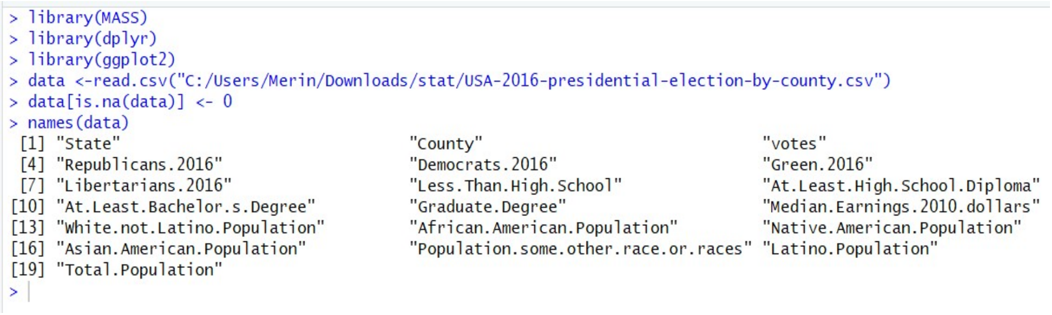
Linear regression is a model that attempts to model the relationship between two variables by fitting the linear equation to an observed data. One variable is considered to be an explanatory variable whereas the other one a dependent variable. A liner regression line has equation of a form

$$Y=a+bX,$$

Where **X** is an explanatory variable and **Y** is the dependent variable. The slope of line is **b**, and **a** is the intercept (the value of y when x=0).

4.1 Extract the data and create the training and testing data samples.

For the current model, using the Election dataset. Following are the features available in Election dataset. The problem statement is to predict “Republicans 2016” based on the set of input features.

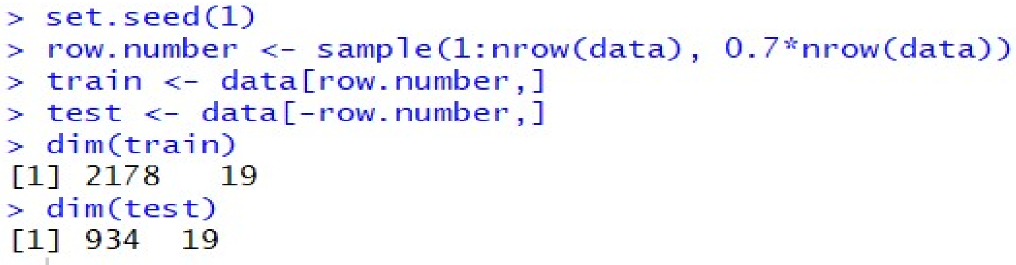


```
> library(MASS)
> library(dplyr)
> library(ggplot2)
> data <-read.csv("C:/Users/Merin/Downloads/stat/USA-2016-presidential-election-by-county.csv")
> data[is.na(data)] <- 0
> names(data)
 [1] "State"                          "County"                          "votes"
 [4] "Republicans.2016"               "Democrats.2016"                  "Green.2016"
 [7] "Libertarians.2016"              "Less.Than.High.School"           "At.Least.High.School.Diploma"
[10] "At.Least.Bachelor.s.Degree"     "Graduate.Degree"                 "Median.Earnings.2010.dollars"
[13] "White.not.Latino.Population"    "African.American.Population"     "Native.American.Population"
[16] "Asian.American.Population"      "Population.some.other.race.or.races" "Latino.Population"
[19] "Total.Population"
> |
```

Fig. 4.1.1

4.1.1 Split the sample data.

Splitting the input data into training and evaluation datset and make the model for the training dataset. It can be seen that train dataset has 2178 observations and the test dataset has 934 observations based on the 70-30 split.

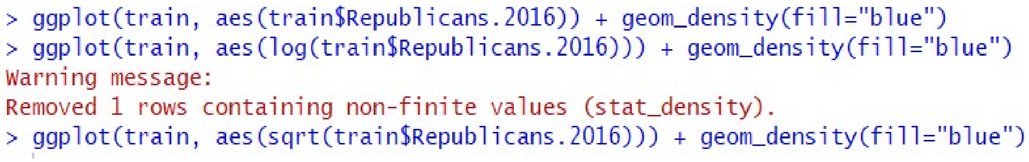


```
> set.seed(1)
> row.number <- sample(1:nrow(data), 0.7*nrow(data))
> train <- data[row.number,]
> test <- data[-row.number,]
> dim(train)
[1] 2178   19
> dim(test)
[1] 934  19
```

Fig. 4.1.1.1

4.2 Exploring the response variable

Checking the distribution of response variable “Republicans 2016”. The following figure shows the three distributions of “Republicans 2016” original, log transformation and the square root transformation. We can see that the logarithmic and square root transformations are bit skewed towards the right, whereas the original “Republicans 2016” is doing a decent job of having a distribution to normal. In the following model, we can use the original transformation.



```
> ggplot(train, aes(train$Republicans.2016)) + geom_density(fill="blue")
> ggplot(train, aes(log(train$Republicans.2016))) + geom_density(fill="blue")
Warning message:
Removed 1 rows containing non-finite values (stat_density).
> ggplot(train, aes(sqrt(train$Republicans.2016))) + geom_density(fill="blue")
```

Fig. 4.2.1

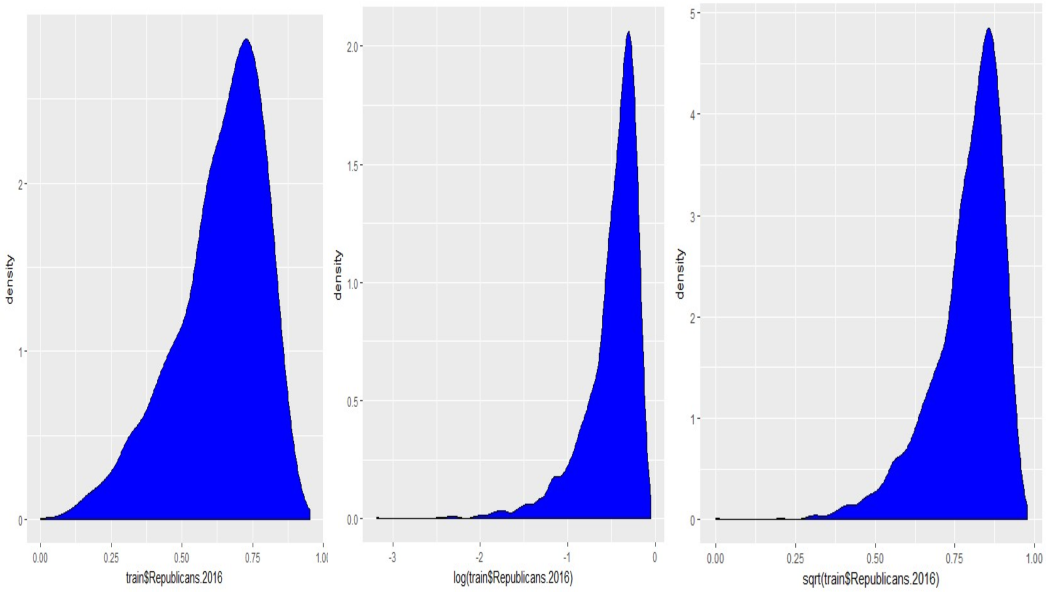


Fig 4.2.1

4.3 Scatterplot Matrix using splom function

Scatterplot matrix can be generated using the splom function that uses the lattice package. Scatterplot describes the relation between each variable. It can also be termed as a graphical representation of the co-relation matrix. Each graphical co-relation helps us in determining which variables follow unique or surprising trends.

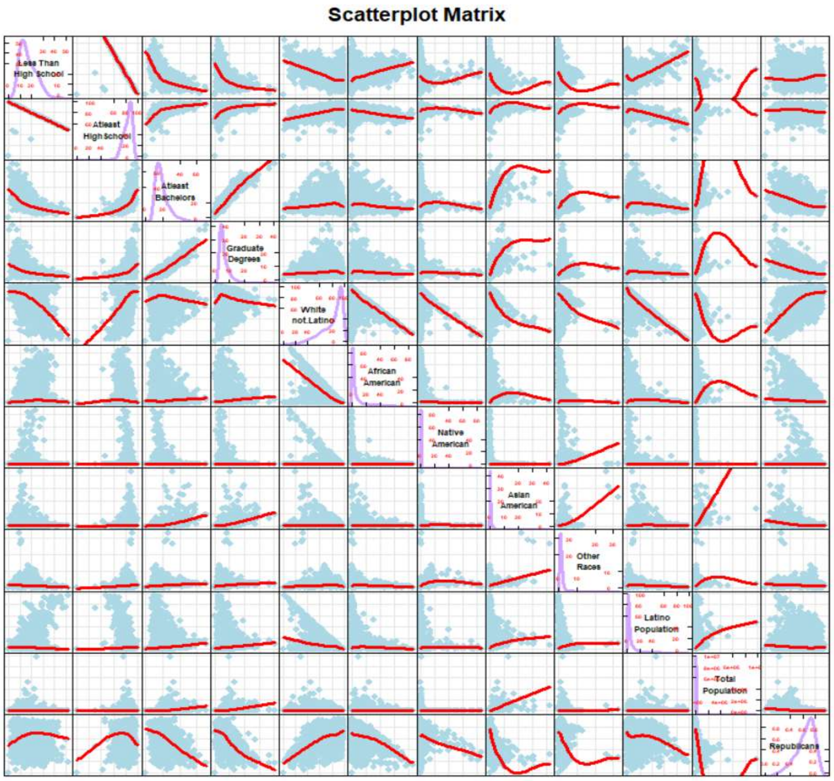


Fig. 4.3.1

Since, our response variable is “Republicans” we look at the last row of the scatterplot matrix and find the predictor variables that follow negative or positive trends. From the above plot we can say that “Almost Bachelors”, “Graduate Degrees”, “African American” and Native American” follow negative linear trends, which explains that the republican results reflect negatively due to these values. Also, an interesting observation can be made on “White not Latino” population in which the trend is showing a perfectly positive linear trend, which explains that the republican results reflect positively due to the “White not Latino” population values.

4.4 Variable Importance plot using Random Forest

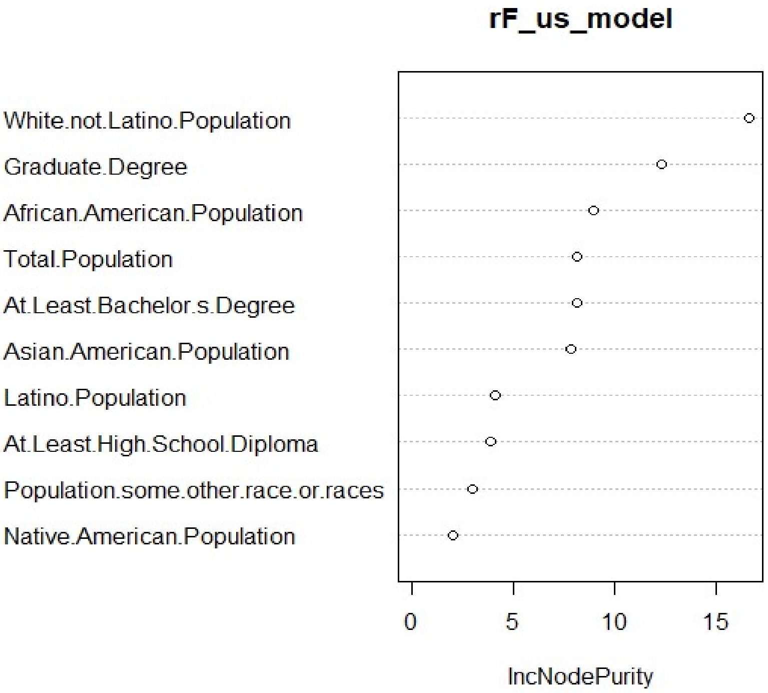
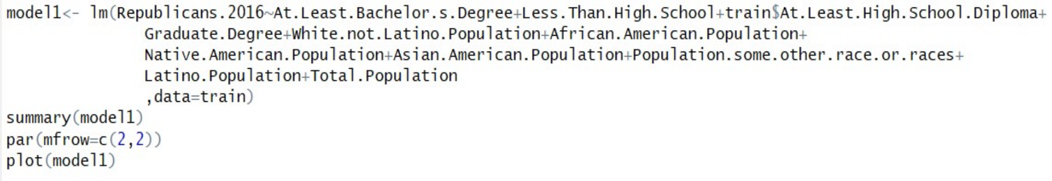


Fig. 4.4.1

According to the variable importance plot, we can find the most significant and least significant variables. In this case, White.not.Latino.Population and Graduate.Degree are the most significant variables. From the scatterplot matrix we can see that one affects the response variable positively (White.not.Latino.Population) whereas the other one affects negatively (Graduate.Degree).

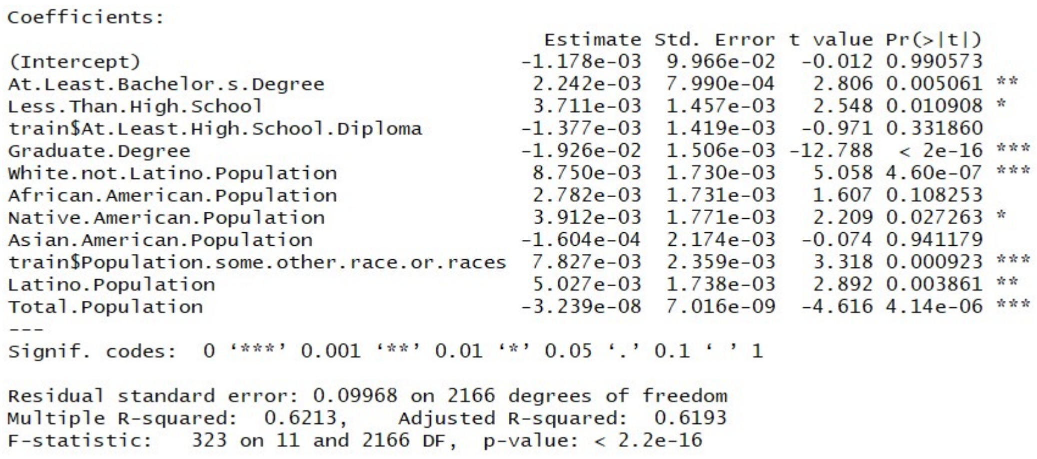
4.5 Model Building- MODEL 1

Building the first regression model by fitting the input values in the multiple regression



```
model1<- lm(Republicans.2016~At.Least.Bachelor.s.Degree+Less.Than.High.School+train$At.Least.High.School.Diploma+
              Graduate.Degree+White.not.Latino.Population+African.American.Population+
              Native.American.Population+Asian.American.Population+Population.some.other.race.or.races+
              Latino.Population+Total.Population
              ,data=train)
summary(model1)
par(mfrow=c(2,2))
plot(model1)
```

Fig. 4.5.1



```
Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                               -1.178e-03  9.966e-02  -0.012 0.990573    
At.Least.Bachelor.s.Degree                 2.242e-03  7.990e-04   2.806 0.005061 ** 
Less.Than.High.School                      3.711e-03  1.457e-03   2.548 0.010908 *  
train$At.Least.High.School.Diploma        -1.377e-03  1.419e-03  -0.971 0.331860    
Graduate.Degree                           -1.926e-02  1.506e-03 -12.788  < 2e-16 ***
White.not.Latino.Population                8.750e-03  1.730e-03   5.058 4.60e-07 ***
African.American.Population                2.782e-03  1.731e-03   1.607 0.108253    
Native.American.Population                 3.912e-03  1.771e-03   2.209 0.027263 *  
Asian.American.Population                 -1.604e-04  2.174e-03  -0.074 0.941179    
train$Population.some.other.race.or.races  7.827e-03  2.359e-03   3.318 0.000923 ***
Latino.Population                          5.027e-03  1.738e-03   2.892 0.003861 ** 
Total.Population                          -3.239e-08  7.016e-09  -4.616 4.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09968 on 2166 degrees of freedom
Multiple R-squared:  0.6213,	Adjusted R-squared:  0.6193 
F-statistic:   323 on 11 and 2166 DF,  p-value: < 2.2e-16
```

Fig. 4.5.2

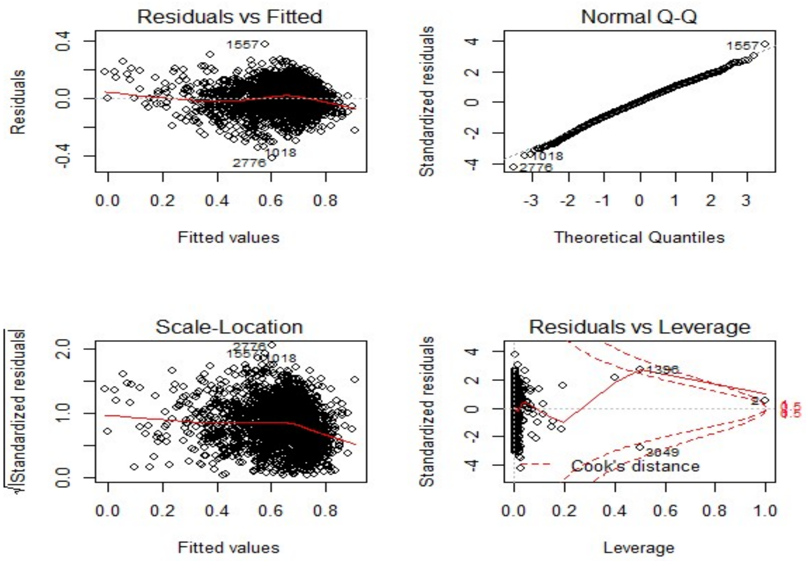


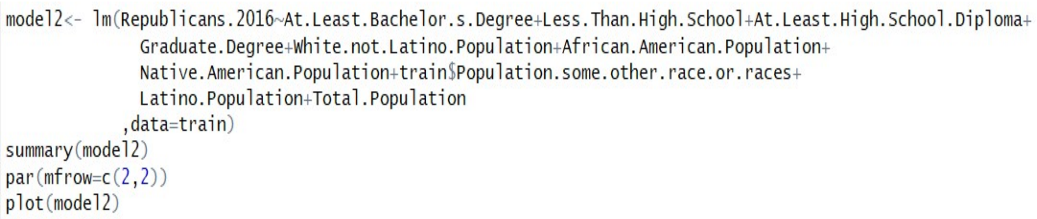
Fig. 4.5.3

4.5.1 Observations

F stat defines the collective effect of all predictor variables on response variables. In this model, Fstat 323 is far greater than 11, we can conclude that there is a relationship between predictor and response variables. Based on the p-value, we can also conclude that predictors having least p-value are the most significant and the ones having higher p-value can be eliminated in the next model. Also, the R-squared value is 0.6213 which is closer to 1, hence the model is a good fit.

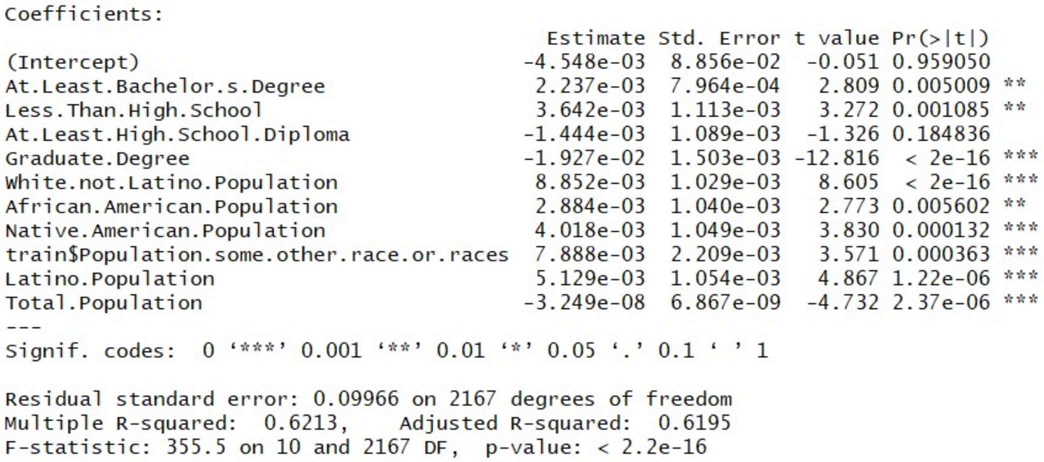
4.6. Model Building - MODEL 2

As the next step, we can remove the less significant predictors (Asian American population) depending upon the p-value and check the model again.



```
model2<- lm(Republicans.2016~At.Least.Bachelor.s.Degree+Less.Than.High.School+At.Least.High.School.Diploma+
             Graduate.Degree+White.not.Latino.Population+African.American.Population+
             Native.American.Population+train$Population.some.other.race.or.races+
             Latino.Population+Total.Population
           ,data=train)
summary(model2)
par(mfrow=c(2,2))
plot(model2)
```

Fig. 4.6.1



```
Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              -4.548e-03  8.856e-02  -0.051 0.959050    
At.Least.Bachelor.s.Degree                2.237e-03  7.964e-04   2.809 0.005009 ** 
Less.Than.High.School                     3.642e-03  1.113e-03   3.272 0.001085 ** 
At.Least.High.School.Diploma             -1.444e-03  1.089e-03  -1.326 0.184836    
Graduate.Degree                          -1.927e-02  1.503e-03 -12.816  < 2e-16 ***
White.not.Latino.Population               8.852e-03  1.029e-03   8.605  < 2e-16 ***
African.American.Population               2.884e-03  1.040e-03   2.773 0.005602 ** 
Native.American.Population                4.018e-03  1.049e-03   3.830 0.000132 ***
train$Population.some.other.race.or.races 7.888e-03  2.209e-03   3.571 0.000363 ***
Latino.Population                         5.129e-03  1.054e-03   4.867 1.22e-06 ***
Total.Population                         -3.249e-08  6.867e-09  -4.732 2.37e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09966 on 2167 degrees of freedom
Multiple R-squared:  0.6213,	Adjusted R-squared:  0.6195 
F-statistic: 355.5 on 10 and 2167 DF,  p-value: < 2.2e-16
```

Fig. 4.6.2

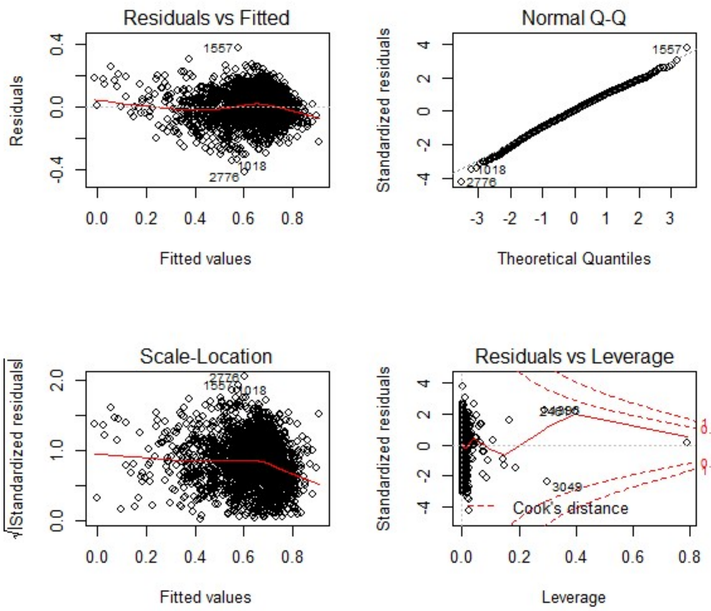


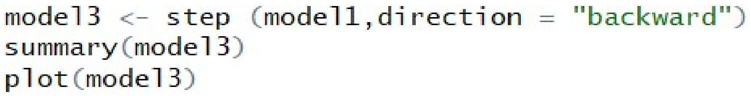
Fig. 4.6.2

4.6.1 Observations

Fstat value 355.5 is far greater than 10, we can conclude that there is a relationship between predictor and response variables. Based on the p-value on the previous model, we can conclude that most of the predictors are significant now. Also, the R-squared value is 0.6213 which is similar to the previous model.

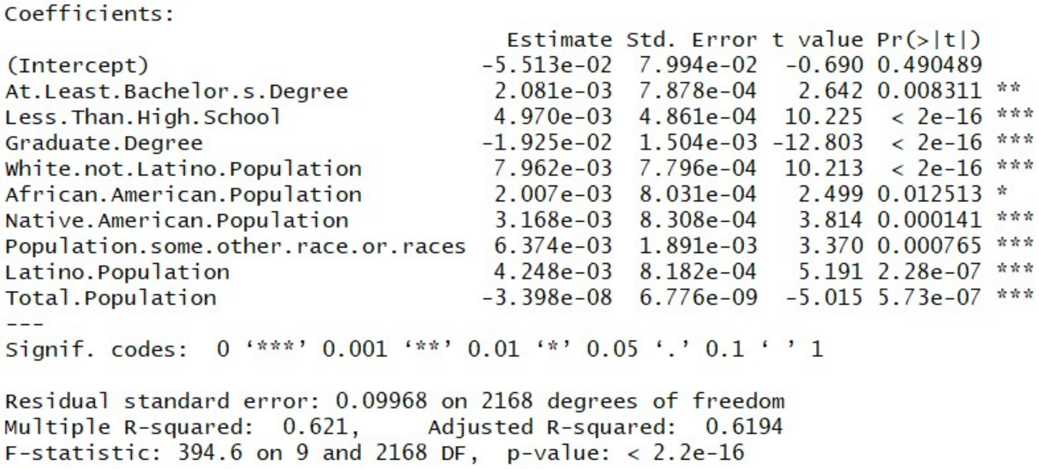
4.7 Model Building – MODEL 3 (using backward elimination on MODEL 1)

Performing backward elimination on MODEL 1



```
model3 <- step (model1,direction = "backward")
summary(model3)
plot(model3)
```

Fig. 4.7.1



```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         -5.513e-02  7.994e-02  -0.690 0.490489    
At.Least.Bachelor.s.Degree           2.081e-03  7.878e-04   2.642 0.008311 ** 
Less.Than.High.School                4.970e-03  4.861e-04  10.225  < 2e-16 ***
Graduate.Degree                     -1.925e-02  1.504e-03 -12.803  < 2e-16 ***
White.not.Latino.Population          7.962e-03  7.796e-04  10.213  < 2e-16 ***
African.American.Population          2.007e-03  8.031e-04   2.499 0.012513 *  
Native.American.Population           3.168e-03  8.308e-04   3.814 0.000141 ***
Population.some.other.race.or.races  6.374e-03  1.891e-03   3.370 0.000765 ***
Latino.Population                    4.248e-03  8.182e-04   5.191 2.28e-07 ***
Total.Population                    -3.398e-08  6.776e-09  -5.015 5.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09968 on 2168 degrees of freedom
Multiple R-squared:  0.621,	Adjusted R-squared:  0.6194 
F-statistic: 394.6 on 9 and 2168 DF,  p-value: < 2.2e-16
```

Fig. 4.7.2

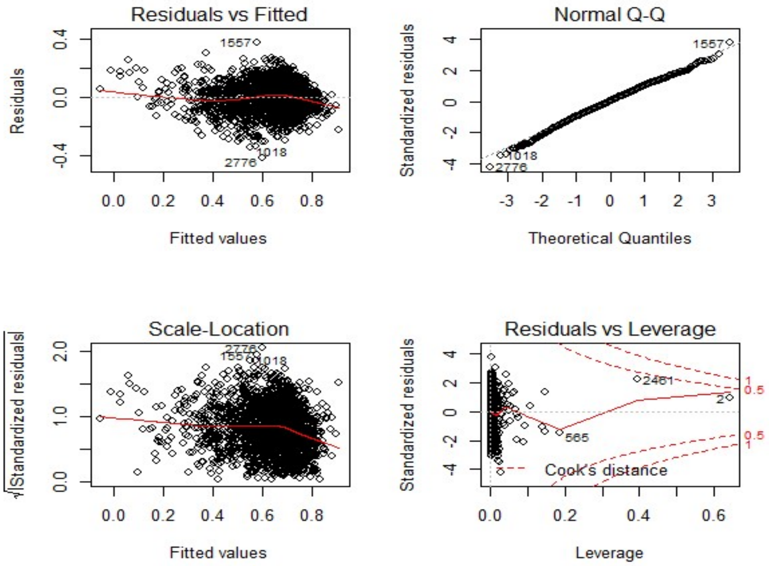


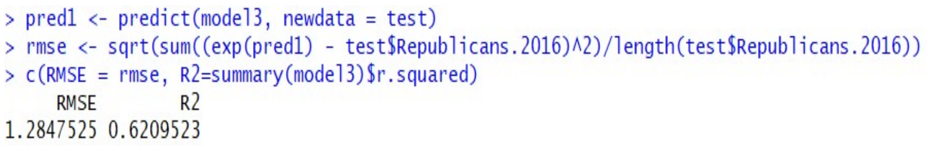
Fig. 4.7.3

4.7.1 Observations

Fstat value 394.6 is far greater than 9, we can conclude that there is a relationship between predictor and response variables. Based on the p-value on the previous model, we can conclude that all the predictors are significant now. Also, the R-squared value is 0.621 which is almost the same as the previous model MODEL 1.

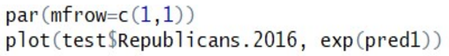
4.8 Prediction

The main objective of this study is to reduce the testing error than the training error. We will use the test data to evaluate the model that we have arrived upon, i.e, MODEL 3. By making a prediction on MODEL 3 we can evaluate the model. In the last step, by predicting the 'test' observation, we can see the comparison between predicted response and actual response value.



```
> pred1 <- predict(model3, newdata = test)
> rmse <- sqrt(sum((exp(pred1) - test$Republicans.2016)^2)/length(test$Republicans.2016))
> c(RMSE = rmse, R2=summary(model3)$r.squared)
     RMSE        R2 
1.2847525 0.6209523 
```

Fig. 4.8.1



```
par(mfrow=c(1,1))
plot(test$Republicans.2016, exp(pred1))
```

Fig. 4.8.2

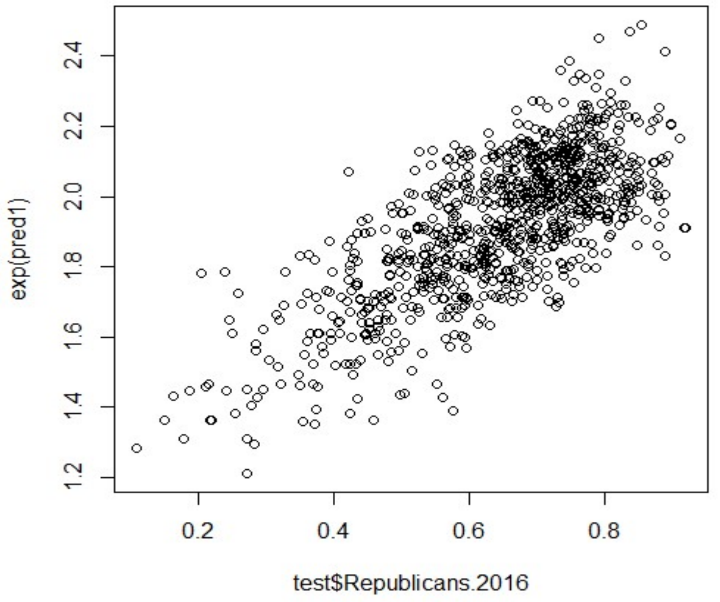


Fig. 4.8.3

RMSE explains on an average how much is the predicted value from the actual value. Based on the predicted model, RMSE is 1.2847525, we can conclude that on an average predicted value will be off by 1.2847525 from the actual value.

5. Conclusion

The model that is created can be improved using Outlier detection, Correlation detection, models such as Random Forest to improve the accuracy of the prediction. One thing we have to keep in mind is to avoid overfitting on the training dataset as its decreases the test dataset accuracy in case of overfitting.

From the Scatterplot matrix, the variable importance plot, and linear regression we can come to the conclusion that, the variables that influence the result i.e., “Republicans 2016” vote percentage are “White.not.Latino.population” and “Graduate.Degree”, showing positive and negative relations respectively. Also, the RMSE value on an average is 1.2847525, which explains that the predicted value will be away from the actual value by 1.2847525.

References

[1]. USA 2016 Presidential Election by County. (November 10,2016). Retrieved from https://public.opendatasoft.com/explore/dataset/usa-2016-presidential-election-by-county/

[2]. Pearson, Jim. (n.d). micromapST. Retrieved from https://www.rdocumentation.org/packages/micromapST/versions/1.0.5/topics/micromapST

[3]. Linear Regression. (n.d). Retrieved from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[4]. Shekhar, Prashant. (January 16, 2018). How to apply Linear Regression in R. Retrieved from https://datascienceplus.com/how-to-apply-linear-regression-in-r/

[5]. Kabakoff, Robert. (2017). Scatterplots. Retrieved from https://www.statmethods.net/graphs/scatterplot.html

[6]. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

[7]. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[8]. Linda Williams Pickle, James B. Pearson, Jr. and Daniel B. Carr (2014), micromapST: Exploring and Communicating Geospatial Patterns in U. S. State Data

[9]. Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr

[10]. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-79. https://CRAN.R-project.org/package=caret

[11]. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

[12]. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 23), 18--22.

Appendix

micromapST

```
library(ggplot2)
library(dplyr)
library(micromapST)
source('rowTheme.R')

data <-read.csv("C:/Users/Merin/Downloads/stat/USA-2016-presidential-election-by-county.csv")
head(data)
agg_state <-data%>%group_by(State)%>%summarise(votes=sum(votes,na.rm=TRUE),
            Republicans.2016 =mean(Republicans.2016,na.rm=TRUE),
            Democrats.2016=mean(Democrats.2016,na.rm=TRUE),
            Green.2016=mean(Green.2016,na.rm=TRUE),
            Libertarians.2016=mean(Libertarians.2016,na.rm=TRUE),
            Less.Than.High.School=mean(Less.Than.High.School,na.rm=TRUE),
            Atleast.High.School=mean(At.Least.High.School.Diploma,na.rm=TRUE),
            At.Least.Bachelors=mean(At.Least.Bachelor.s.Degree,na.rm=TRUE),
            Graduate.Degree=mean(Graduate.Degree,na.rm=TRUE),
           White.not.Latino.Population=mean(White.not.Latino.Population,na.rm=TRUE),
        African.American.Population=mean(African.American.Population,na.rm=TRUE),
          Native.American.Population=mean(Native.American.Population,na.rm=TRUE),
           Asian.American.Population=mean(Asian.American.Population,na.rm=TRUE),
   Population.some.other.races=mean(Population.some.other.race.or.races,na.rm=TRUE),
   Latino.Population=mean(Latino.Population,na.rm=TRUE),
   Total.Population=sum(Total.Population,na.rm=TRUE))
head(agg_state)
agg_state[,c(1,17)]
order<- with(agg_state, order(Republicans.2016,decreasing = TRUE))
columnOrder <- c(1:6,9,8,7)
newdata<- agg_state[order,columnOrder]
blocks <-as.character(newdata$State)
newdata$State <-factor(blocks,levels=rev(blocks))
levels(newdata$State)
groups <- paste("G",1:11,sep="")
groups
size <- c(5,5,5,5,5,1,5,5,5,5,5)
newdata$Group <- factor(rep(groups, size),level=groups)
newdata$Group
order1 <- rep(1:5,5)
order1
seq <- c(order1,1,order1)
seq
label <- c('1', '2', '3', '4', '5')
temp <- labs[seq]
newdata$Record <- factor(temp,levels = label)
newdata$Record
recordColor<- rgb(
  red  = c(1.00, 1.00, 0.00, 0.10, 0.80),
  green= c(0.10, 0.50, 0.75, 0.65, 0.45),
  blue = c(0.10, 0.00, 0.00, 1.00, 1.00))

ggplot(newdata, aes(x=newdata$Republicans.2016,y=State,fill=Record,group=Group)) +
  labs(x="Votes Precentage", y="States",
      title="Votes-Republicans 2016 ")+
  geom_point(shape=21,size=3)+
  scale_fill_manual(values=recordColor)+
  guides(fill=FALSE)+
  facet_grid(Group~., scale="free_y", space="free" )+rowTheme
MicroMap <- as.data.frame(newdata)
MicroMap
rownames(MicroMap) <-MicroMap[,1]
MicroMap
MicroMap <-MicroMap[,c(1,2,3,4,5,6)]
head(MicroMap)
panel<- data.frame(
  type=c('mapcum','id','dot','dot','dot','dot'),
  lab1=c(",",'Republicans 2016','Democrats 2016','Greens 2016','Libertarians 2016'),
  lab3 = c(",",'Percent','Percent','Percent','Percent'),
  col1=c(NA,NA,3,4,5,6)
)
t(panel)
pdf(file= "Election statistics.pdf",width = 9.5, height = 10)
micromapST(MicroMap,panel,
           rowNamesCol="State",
           rowNames="full",
           sortVar = "Republicans.2016",plotNames = 'full', ascend=FALSE,
           title=c("US Presidential Elections-2016")
)
dev.off()
```

Scatterplot Matrix

```
library(tidyverse)
library(ggplot2)
library(caret)
library(dplyr)
source('hw.r')

data       <-read.csv("C:/Users/Merin/Downloads/stat/USA-2016-presidential-election-by-
county.csv")
head(data)
data[is.na(data)] <- 0
summary(data)
variables <- colnames(data)[c(8,9,10,11,13,14,15,16,17,18,19,4)]
variables
Elect<- dplyr::select(data,vars)
Elect
varnames = c("Less Than\nHigh School","Atleast\nHighSchool",
             "Atleast\nBachelors","
Graduate\nDegrees","White\nnot.Latino","African\nAmerican","Native\nAmerican",
"Asian\nAmerican","Other\nRaces","Latino\nPopulation","Total\nPopulation","Republic
ans")
offDiag <- function(x,y,...){
  panel.grid(h = -1,v = -1,...)
  panel.points(x,y,...,cex = .8, pch = 16, col = "light blue")
  panel.loess(x , y, ..., lwd = 3, col = 'red')
}
onDiag <- function(x, ...){
  yrng <- current.panel.limits()$ylim
  d <- density(x, na.rm = TRUE)
  d$y <- with(d, yrng[1] + 0.95 * diff(yrng) * y / max(y) )
  panel.lines(d,col = rgb(.83,.66,1),lwd = 3)
  diag.panel.splom(x, ...)
}
splom(Elect, as.matrix = TRUE,
      xlab = ",main = "Scatterplot Matrix",
      varnames = varnames,
      pscale = 4, varname.cex = 0.5,  varname.font = 2,
      axis.text.cex = 0.3,
      axis.text.col = "red",axis.text.font = 2,
      axis.line.tck = .5,
      panel = offDiag,
      diag.panel = onDiag)
```

Variable Importance plot

```
library(MASS)
library(randomForest)

data       <-read.csv("C:/Users/Merin/Downloads/stat/USA-2016-presidential-election-by-
county.csv")
head(data)
data[is.na(data)] <- 0
rF_us_model <- randomForest(Republicans.2016~Graduate.Degree+
                             White.not.Latino.Population+At.Least.Bachelor.s.Degree+
                             Asian.American.Population+At.Least.High.School.Diploma
                           +Total.Population+African.American.Population
                           +Native.American.Population+Asian.American.Population
                           +Population.some.other.race.or.races+Latino.Population,data=data)
rF_us_model
plot(rF_us_model)
varImpPlot(rF_us_model)
```

Linear Regression

```
library(MASS)
library(ggplot2)
library(dplyr)
source("hw.r")

data       <-read.csv("C:/Users/Merin/Downloads/stat/USA-2016-presidential-election-by-
county.csv")
```

```
head(data)
names(data)

set.seed(1)
row.number <- sample(1:nrow(data), 0.7*nrow(data))
train <- data[row.number,]
test <- data[-row.number,]
dim(train)
dim(test)

ggplot(train, aes(train$Republicans.2016)) + geom_density(fill="blue")
ggplot(train, aes(log(train$Republicans.2016))) + geom_density(fill="blue")
ggplot(train, aes(sqrt(train$Republicans.2016))) + geom_density(fill="blue")

model1<-
lm(Republicans.2016~At.Least.Bachelor.s.Degree+Less.Than.High.School+train$At.Lea
st.High.School.Diploma+
            Graduate.Degree+White.not.Latino.Population+African.American.Population+

Native.American.Population+Asian.American.Population+Population.some.other.race.or.
races+Latino.Population+Total.Population,data=train)
summary(model1)
par(mfrow=c(2,2))
plot(model1)

model2<-
lm(Republicans.2016~At.Least.Bachelor.s.Degree+Less.Than.High.School+At.Least.Hig
h.School.Diploma+Graduate.Degree+White.not.Latino.Population+African.American.Po
pulation+Native.American.Population+train$Population.some.other.race.or.races+Latino.
Population+Total.Population,data=train)
summary(model2)
par(mfrow=c(2,2))
plot(model2)

model3 <- step (model1,direction = "backward")
summary(model3)
plot(model3)

pred1 <- predict(model3, newdata = test)
rmse                    <-                    sqrt(sum((exp(pred1)                    -
test$Republicans.2016)^2)/length(test$Republicans.2016))
c(RMSE = rmse, R2=summary(model3)$r.squared)
par(mfrow=c(1,1))
plot(test$Republicans.2016, exp(pred1))
```