# Predictive Analysis of Voting Trump

Qiyu Huang & Yuhan Zhu

11:59PM Nov 4th

## Estimating the factors that influence voting for Trump among the residence of the US

## Model

The main objective of the study is to come up with parameter estiamtes for the linear regression model of factors that the influence the likelihood of one voting for Trump. Here we are interested in predicting the popular vote outcome of the 2020 American federal election Singh et al (2017). To do this we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

### Model Specifics

The binary logistic regression model will be used to model the proportion of voters who will vote for Donald Trump. This is a naive model, the age,foreign_born,gender,interest,registration+vote_2016,vote_2020, vote_intention, which is recorded as a numeric variable, to model the probability of voting for Donald Trump. The logistic regression model is appropriate since the study involves estimating the influence of several variables on the voting pattern which take binary outcomes. The interest will be estimating the odds of votering having trump as their preferred candidate. The general form of the model is represented as;

$$ln\left(\frac{P}{1-P}\right)$$

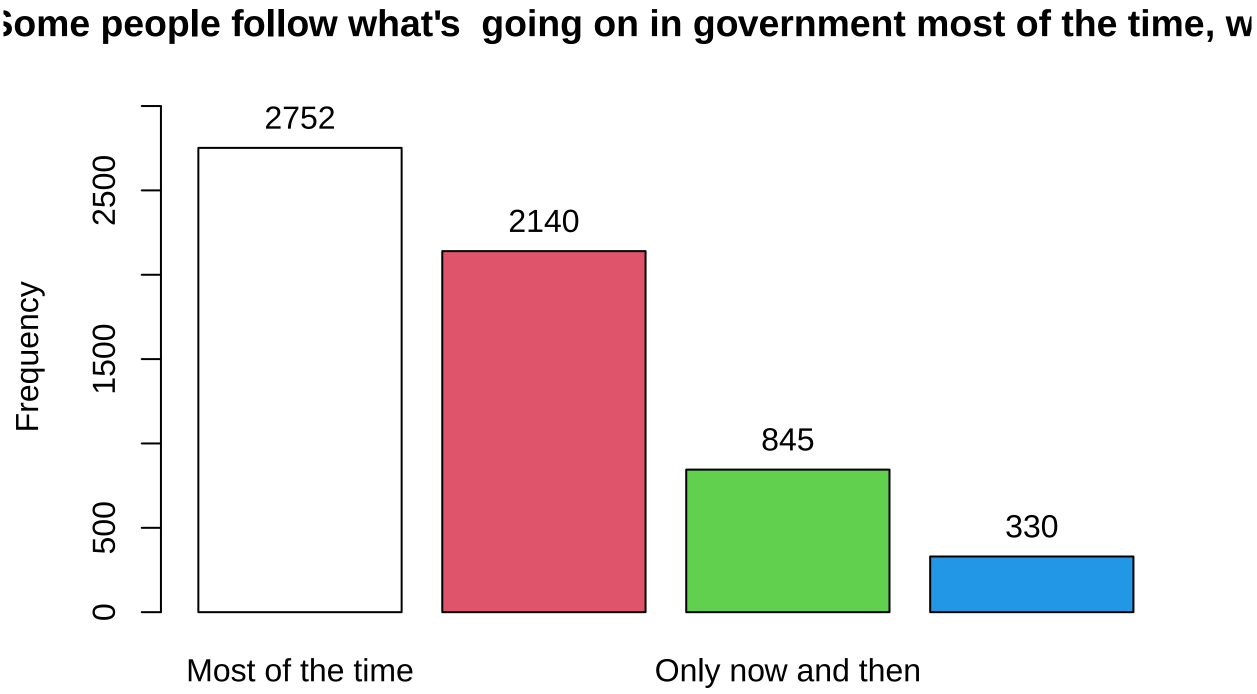
where we model the log odds of the event, wher p represents the probability of the event.

$$Z_i = ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n$$

Where $y$ represents the proportion of voters who will vote for Donald Trump. Similarly, $\beta_0$ represents the intercept of the model, and is the probability of voting for Donald Trump at age 0. Additionally, $\beta_1$ represents the slope of the model. So, for everyone one unit increase in age, we expect a $\beta_1$ increase in the probability of voting for Donald Trump. The above equation can be modeled using the glm() by setting the family argument to "binomial". But we are more interested in the probability of the event, than the log odds of the event. The ods of an events presents the raltive risk of tendncey of the desired outcome occuring given certain measures or values of the indipendent variables.The log odds of the event, can be converted to probability of event as follows:

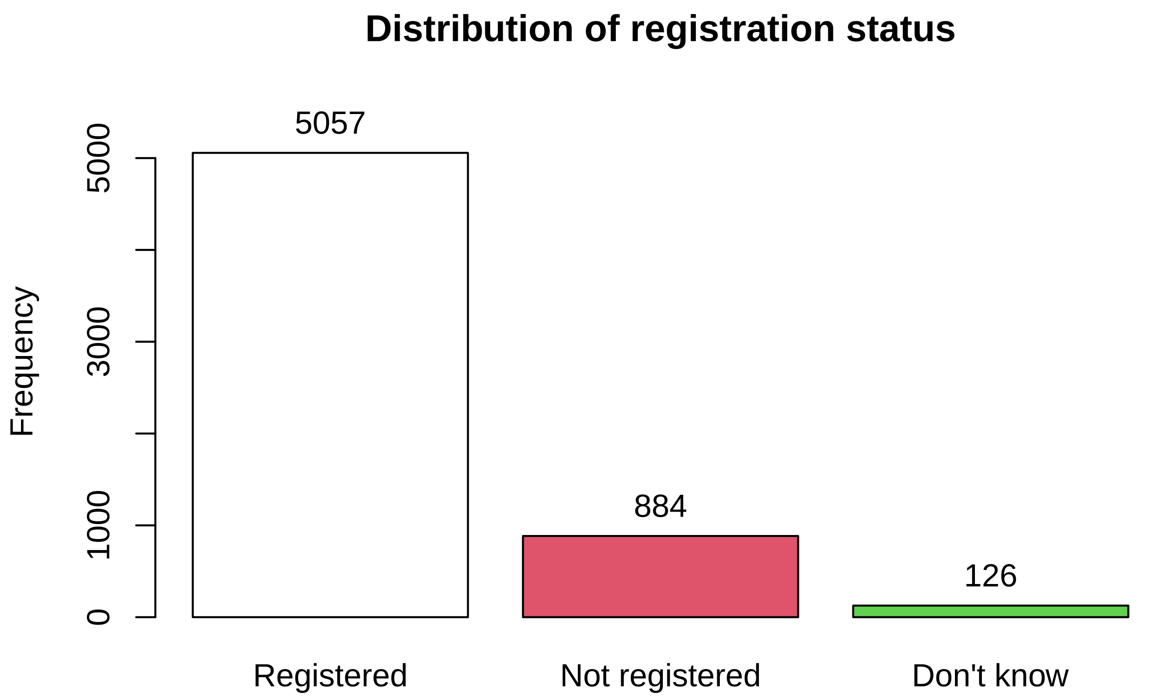
$$P_i = 1 - \left(\frac{1}{1+e_i^z}\right)$$

```
tab1(survey_data$interest, sort.group = "decreasing", cum.percent = TRUE,main = "Some people follow wha
```



```
## survey_data$interest :
##                      Frequency Percent Cum. percent
## Most of the time          2752    45.4         45.4
## Some of the time          2140    35.3         80.6
## Only now and then          845    13.9         94.6
## Hardly at all              330     5.4        100.0
##   Total                   6067   100.0        100.0
```

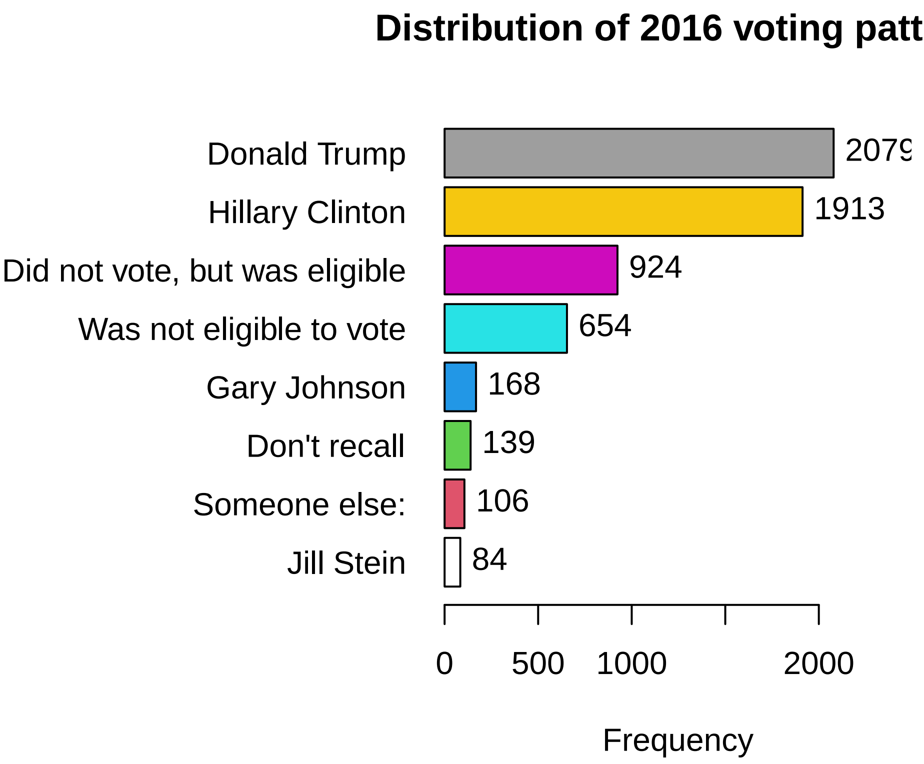
```
tab1(survey_data$registration, sort.group = "decreasing", cum.percent = TRUE,main = "Distribution of reg
```



```
## survey_data$registration :
##                 Frequency Percent Cum. percent
## Registered           5057    83.4         83.4
## Not registered        884    14.6         97.9
## Don't know            126     2.1        100.0
##   Total              6067   100.0        100.0
```

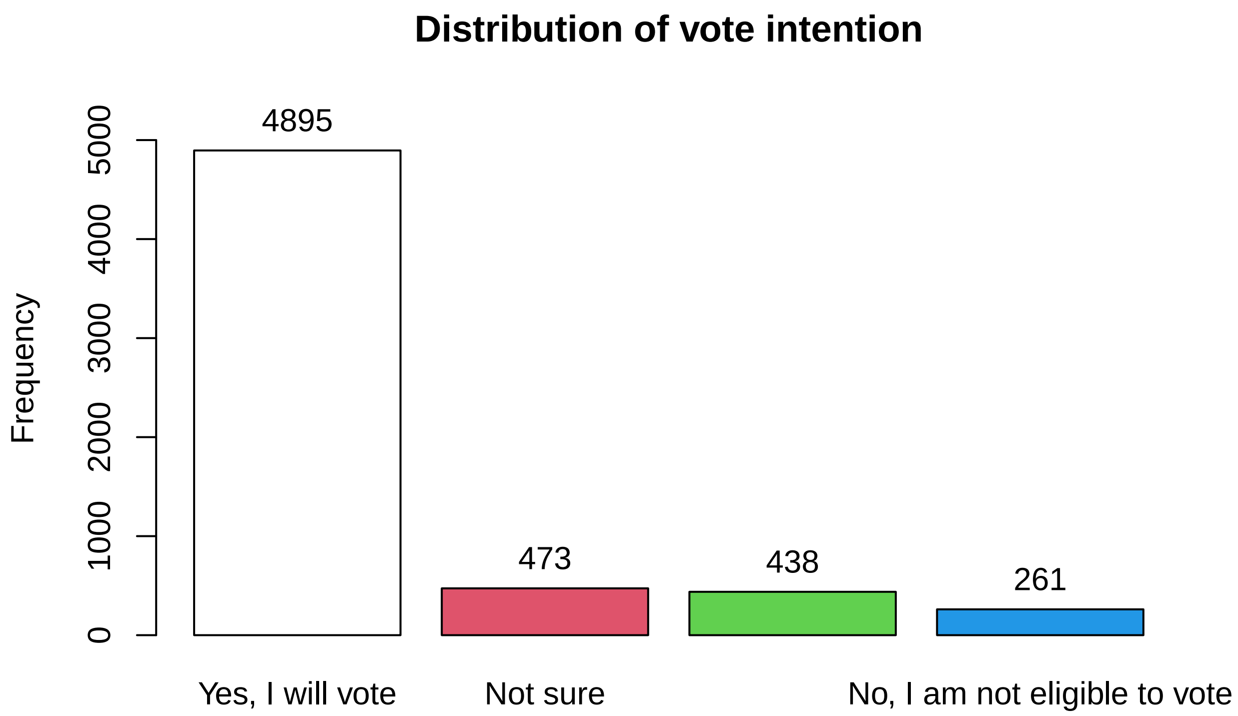
```
attach(survey_data)

tab1(survey_data$vote_2016, sort.group = "decreasing", cum.percent = TRUE,main = "Distribution of 2016 v
```



```
## survey_data$vote_2016 :
##                                 Frequency Percent Cum. percent
## Donald Trump                         2079    34.3         34.3
## Hillary Clinton                      1913    31.5         65.8
## Did not vote, but was eligible        924    15.2         81.0
## Was not eligible to vote              654    10.8         91.8
## Gary Johnson                          168     2.8         94.6
## Don't recall                          139     2.3         96.9
## Someone else:                         106     1.7         98.6
## Jill Stein                             84     1.4        100.0
##   Total                              6067   100.0        100.0
```

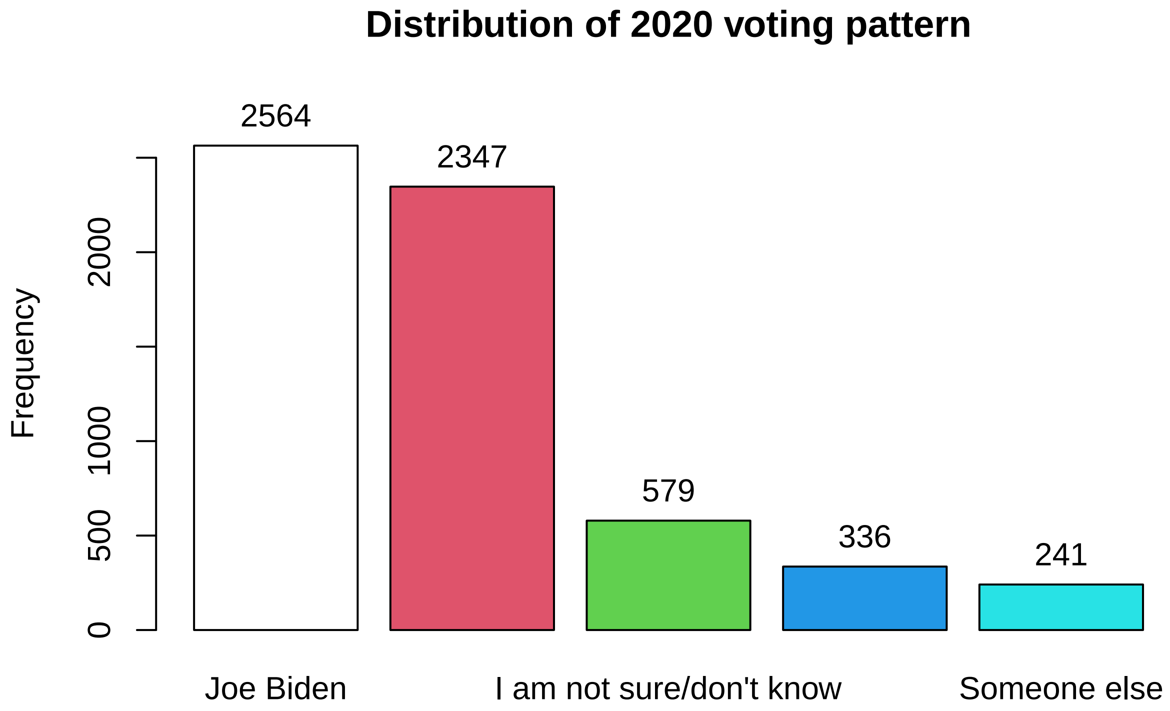
```
tab1(survey_data$vote_intention, sort.group = "decreasing", cum.percent = TRUE,main = "Distribution of
```



Distribution of vote intention

```
## survey_data$vote_intention :
##                                         Frequency Percent Cum. percent
## Yes, I will vote                             4895    80.7         80.7
## Not sure                                      473     7.8         88.5
## No, I will not vote but I am eligible         438     7.2         95.7
## No, I am not eligible to vote                 261     4.3        100.0
##   Total                                      6067   100.0        100.0
```

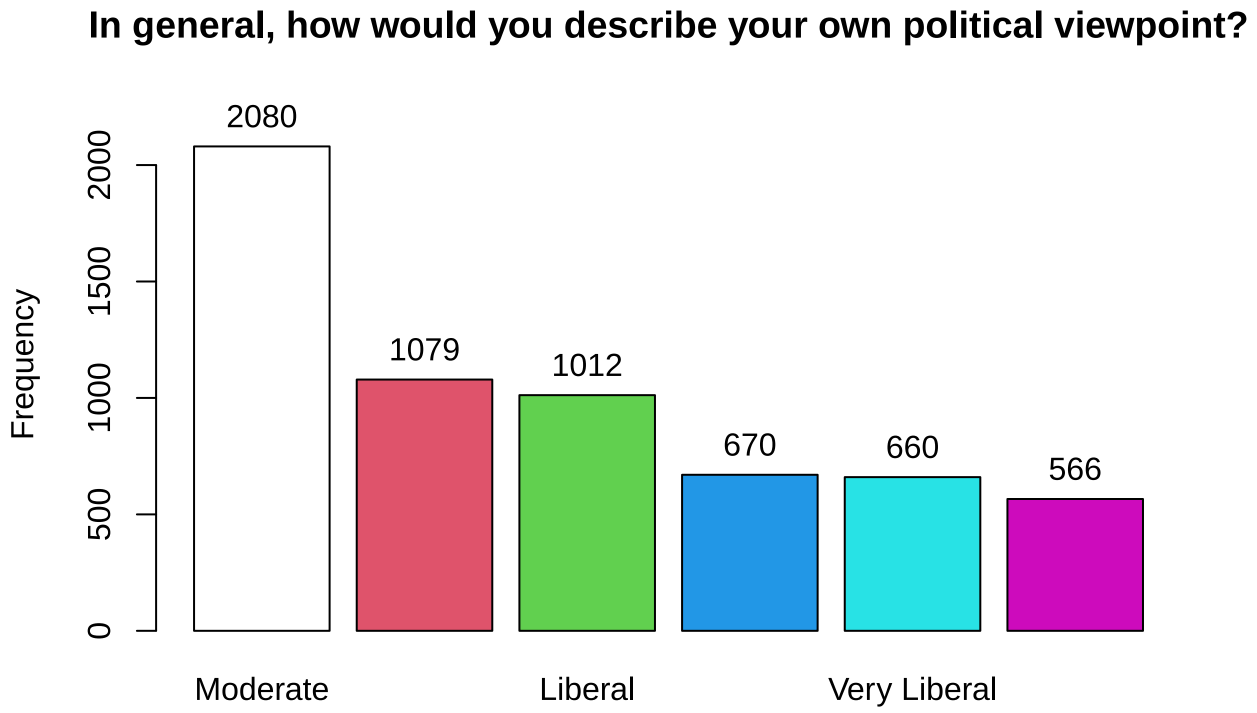
```
tab1(survey_data$vote_2020, sort.group = "decreasing", cum.percent = TRUE,main = "Distribution of 2020 
```



Distribution of 2020 voting pattern

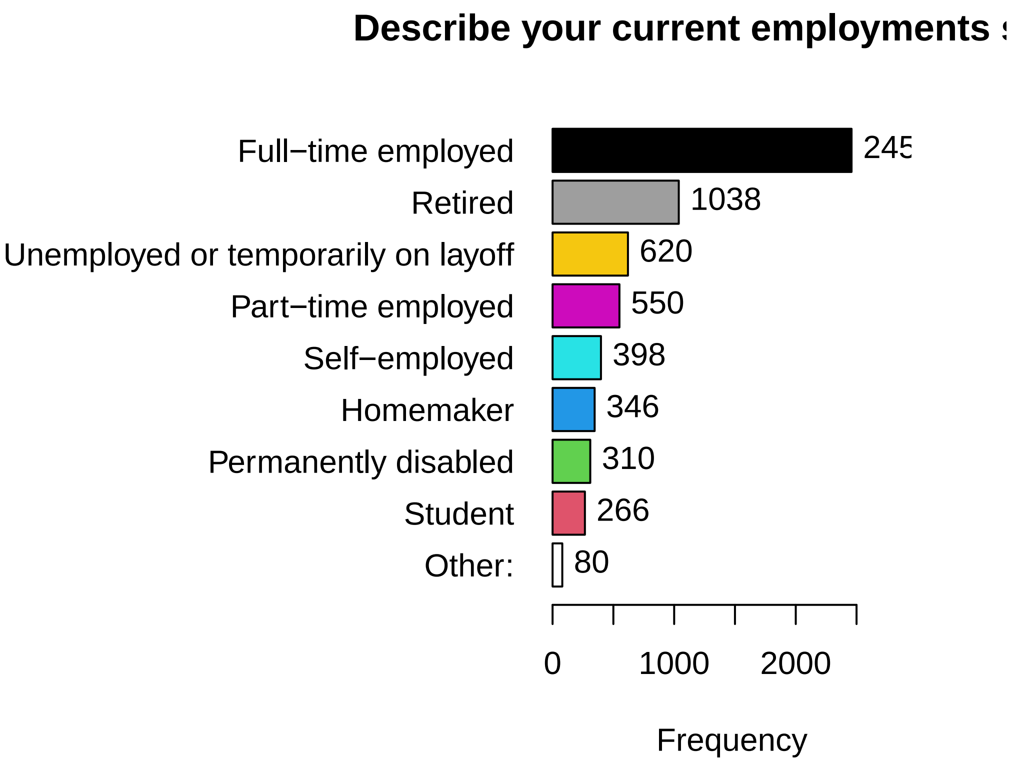
```
## survey_data$vote_2020 :
##                           Frequency Percent Cum. percent
## Joe Biden                      2564    42.3         42.3
## Donald Trump                   2347    38.7         80.9
## I am not sure/don't know        579     9.5         90.5
## I would not vote                336     5.5         96.0
## Someone else                    241     4.0        100.0
##   Total                        6067   100.0        100.0
```

```
tab1(survey_data$ideo5, sort.group = "decreasing", cum.percent = TRUE,main = "In general, how would you
```



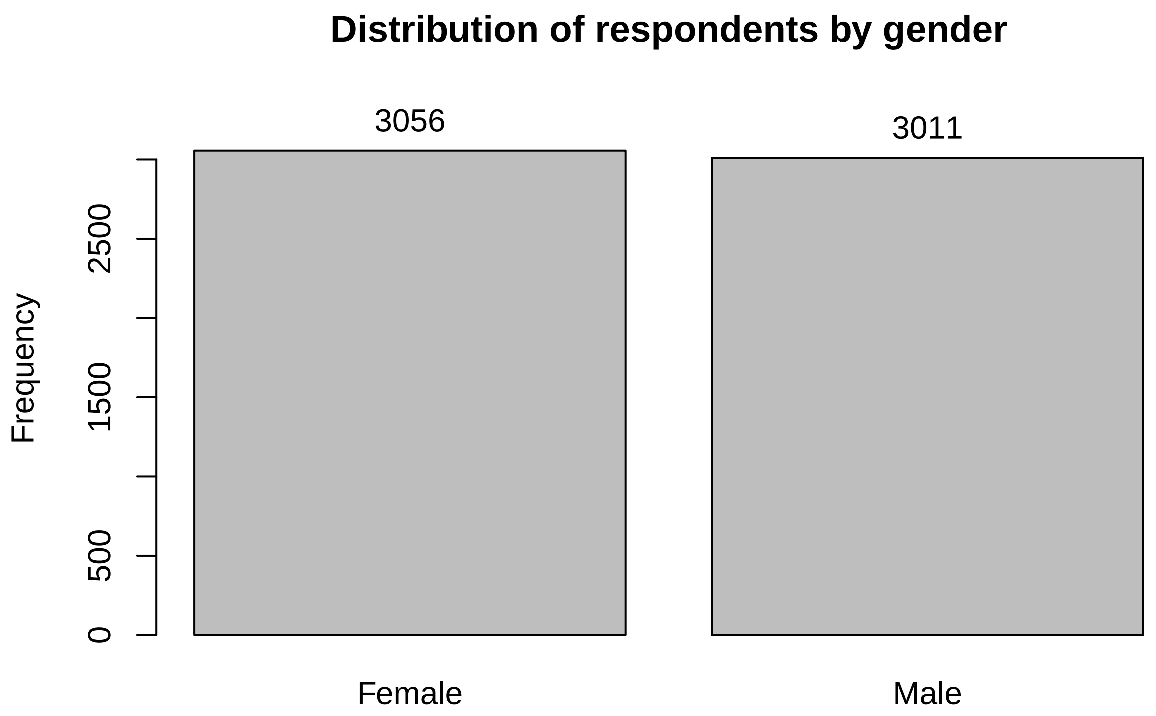
```
## survey_data$ideo5 :
##                    Frequency Percent Cum. percent
## Moderate                2080    34.3         34.3
## Conservative            1079    17.8         52.1
## Liberal                 1012    16.7         68.7
## Very Conservative        670    11.0         79.8
## Very Liberal             660    10.9         90.7
## Not Sure                 566     9.3        100.0
##   Total                 6067   100.0        100.0
```

```
tab1(survey_data$employment, sort.group = "decreasing", cum.percent = TRUE,main = "Describe your curren
```



```
## survey_data$employment :
##                                     Frequency Percent Cum. percent
## Full-time employed                       2459    40.5         40.5
## Retired                                  1038    17.1         57.6
## Unemployed or temporarily on layoff       620    10.2         67.9
## Part-time employed                        550     9.1         76.9
## Self-employed                             398     6.6         83.5
## Homemaker                                 346     5.7         89.2
## Permanently disabled                      310     5.1         94.3
## Student                                   266     4.4         98.7
## Other:                                     80     1.3        100.0
##   Total                                  6067   100.0        100.0
```

```
tab1(survey_data$gender, sort.group = "decreasing", cum.percent = TRUE,main = "Distribution of responde
```



```
## survey_data$gender :
##         Frequency Percent Cum. percent
## Female       3056    50.4         50.4
## Male         3011    49.6        100.0
##   Total      6067   100.0        100.0
```

```
# Create Training Data
input_ones <- survey_data[which(survey_data$vote_trump == 1), ]  # all 1's
input_zeros <- survey_data[which(survey_data$vote_trump == 0), ]  # all 0's
set.seed(100)  # for repeatability of samples
input_ones_training_rows <- sample(1:nrow(input_ones), 0.7*nrow(input_ones))  # 1's for training
input_zeros_training_rows <- sample(1:nrow(input_zeros), 0.7*nrow(input_ones))  # 0's for training. Pic


#training. Pick as many 0's as 1's
training_ones <- input_ones[input_ones_training_rows, ]
training_zeros <- input_zeros[input_zeros_training_rows, ]
trainingData <- rbind(training_ones, training_zeros)  # row bind the 1's and 0's

# Create Test Data
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros <- input_zeros[-input_zeros_training_rows, ]
testData<-rbind(test_ones, test_zeros)
# Creating the Model

model <- lm(vote_trump ~ age+interest+gender+vote_intention+vote_2020, data=survey_data);#summary(model


predicted <- plogis(predict(model, testData))  # predicted scores
# or
predicted <- predict(model, testData, type="response")
```

```
pROC_obj=roc(testData$vote_trump, predicted,smoothed = TRUE,
            # arguments for ci
            ci=TRUE, ci.alpha=0.9, stratified=FALSE,
            # arguments for plot
            plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
            print.auc=TRUE, show.thres=TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Warning in ci.auc.roc(roc, ...): ci.auc() of a ROC curve with AUC == 1 is always
## 1-1 and can be misleading.
```

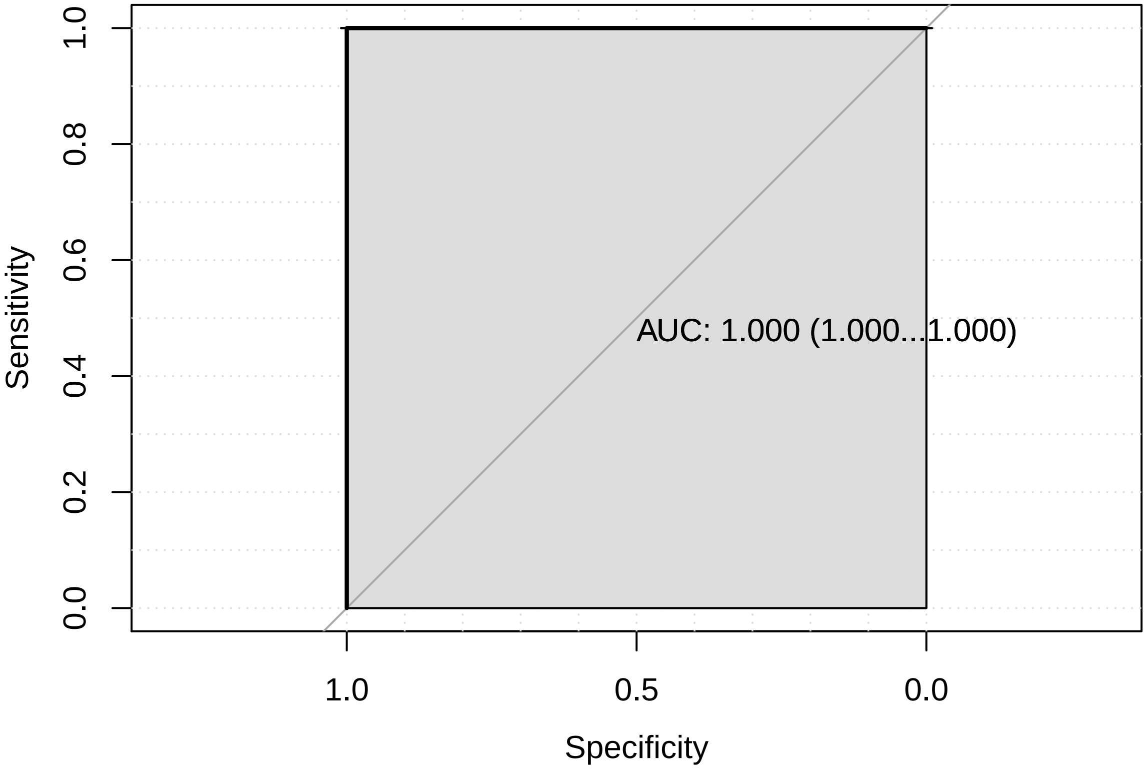
```
sens.ci <- ci.se(pROC_obj)
```

```
## Warning in ci.se.roc(pROC_obj): ci.se() of a ROC curve with AUC == 1 is always a
## null interval and can be misleading.
```

```
plot(sens.ci, type="shape", col="lightblue")
```

```
## Warning in plot.ci.se(sens.ci, type = "shape", col = "lightblue"): Low
## definition shape.
```

```
## Warning in plot.ci.se(sens.ci, type = "shape", col = "lightblue"): Low
## definition shape.
plot(sens.ci, type="bars")
```



```
# Model Results (to Report in Results section)
# summary(model)
# OR
broom::tidy(model)
```

```
## # A tibble: 13 x 5
##    term                                     estimate std.error statistic p.value
##    <chr>                                       <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)                              1.00e+ 0  1.24e-15  8.08e+14  0
##  2 age                                     -1.12e-17  1.08e-17 -1.04e+ 0  0.299
##  3 interestMost of the time                -1.30e-15  8.29e-16 -1.57e+ 0  0.117
##  4 interestOnly now and then               -3.78e-16  8.78e-16 -4.31e- 1  0.666
##  5 interestSome of the time                -1.17e-15  8.19e-16 -1.43e+ 0  0.152
##  6 genderMale                              -5.66e-17  3.49e-16 -1.62e- 1  0.871
##  7 vote_intentionNo, I will not vote but ~  8.91e-16  1.04e-15  8.54e- 1  0.393
##  8 vote_intentionNot sure                  -2.27e-15  1.04e-15 -2.17e+ 0  0.0298
##  9 vote_intentionYes, I will vote           3.67e-16  9.09e-16  4.04e- 1  0.686
## 10 vote_2020I am not sure/don't know       -1.00e+ 0  6.42e-16 -1.56e+15  0
## 11 vote_2020I would not vote               -1.00e+ 0  9.17e-16 -1.09e+15  0
## 12 vote_2020Joe Biden                      -1.00e+ 0  3.86e-16 -2.59e+15  0
## 13 vote_2020Someone else                   -1.00e+ 0  9.10e-16 -1.10e+15  0
```

```
m1=exp(model$coefficients);m1
```

```
##                                         (Intercept)
##                                           2.7182818
##                                                 age
##                                           1.0000000
##                            interestMost of the time
##                                           1.0000000
##                           interestOnly now and then
##                                           1.0000000
##                            interestSome of the time
##                                           1.0000000
##                                          genderMale
##                                           1.0000000
## vote_intentionNo, I will not vote but I am eligible
##                                           1.0000000
##                              vote_intentionNot sure
##                                           1.0000000
##                      vote_intentionYes, I will vote
##                                           1.0000000
##                   vote_2020I am not sure/don't know
##                                           0.3678794
##                           vote_2020I would not vote
##                                           0.3678794
##                                  vote_2020Joe Biden
##                                           0.3678794
##                               vote_2020Someone else
##                                           0.3678794
```

```
broom::tidy(m1)
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
## # A tibble: 13 x 2
##    names                                                x
##    <chr>                                            <dbl>
##  1 (Intercept)                                      2.72
##  2 age                                              1
##  3 interestMost of the time                         1.00
##  4 interestOnly now and then                        1.00
##  5 interestSome of the time                         1.00
##  6 genderMale                                       1.00
##  7 vote_intentionNo, I will not vote but I am eligible 1.
##  8 vote_intentionNot sure                           1.00
##  9 vote_intentionYes, I will vote                   1.
## 10 vote_2020I am not sure/don't know                0.368
## 11 vote_2020I would not vote                        0.368
## 12 vote_2020Joe Biden                               0.368
## 13 vote_2020Someone else                            0.368
```

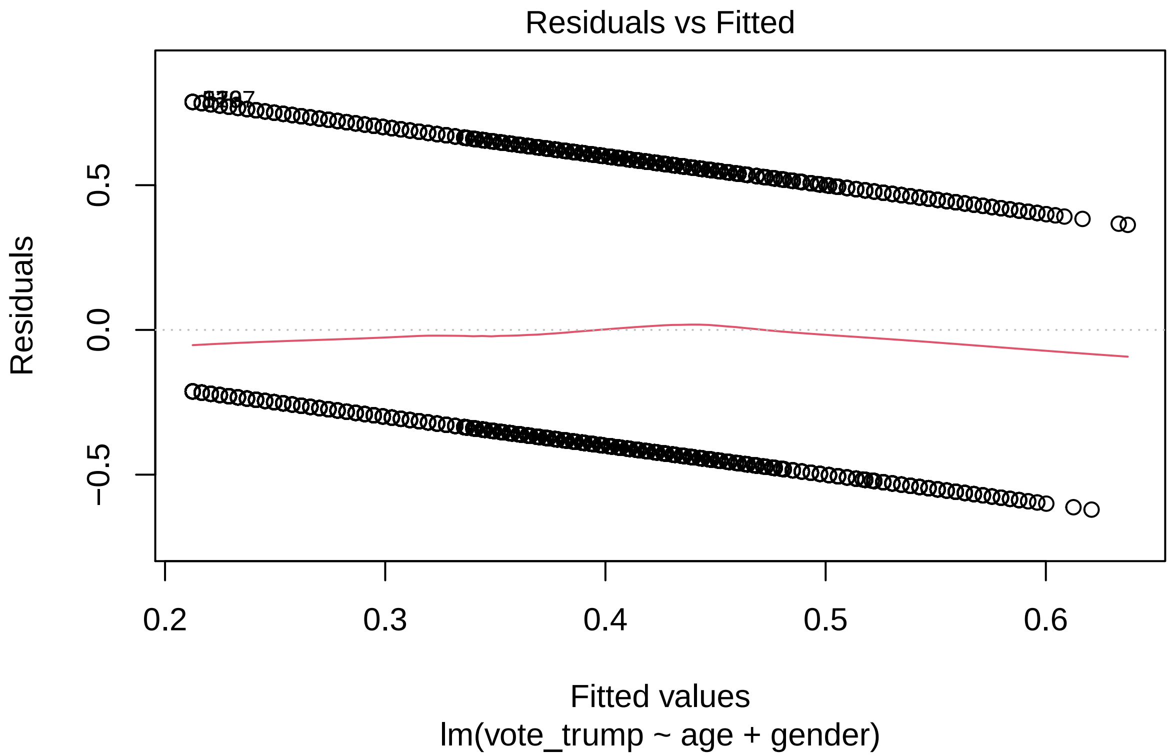
## Post-Stratification

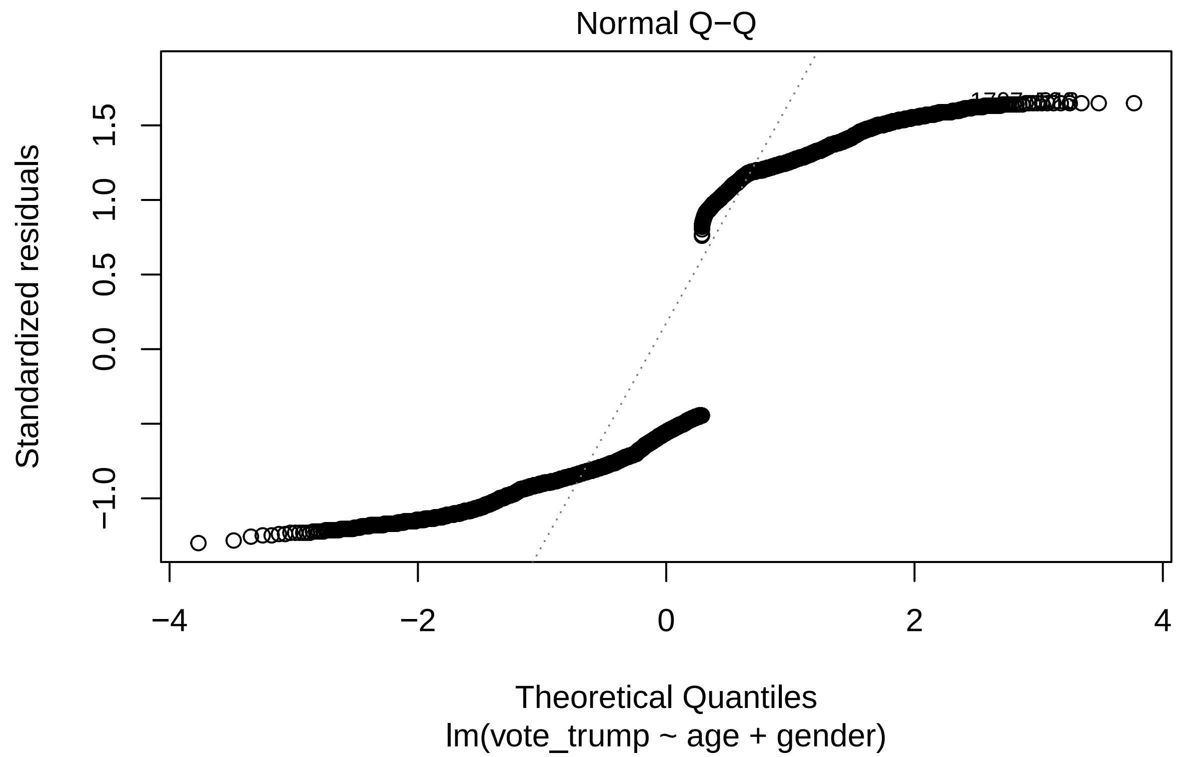
In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Here I create cells based off different ages. Using the model described in the previous sub-section I will estimate the proportion of voters in each age bin. I will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size. Survey is a good statistical tool in collection of data from people. The data collected from the survey conducted is analyzed using R-studio and findings presented as percentages in tabular forms. From the findings above most of the people of the united states are not considering to vote for Donald Trump in the 2020 general election. Only 33% of the people that participated in the survey are willing to vote for Donald Trump in 2020 general election. 84% of those who voted for Trump in 2016 are considering to vote for him again in the 2020 general election. Of the sample surveyed the white, males, those of age 65 years and above, republican and those with very conservative ideology consider voting for Donald Trump in 2020 general election. At least 30 % of the sample in each census region are willing to vote for Trump in the coming election. 8% of the democrats are also considering voting for trump while 88% of the democrats would not be voting for him. The Black race are not considering voting for trump. This is also evident in the youths who are aged 18-29 years; only 22 % of the sample showed interest in voting for Trump. 42 % of those who earn income of above 100k are willing to vote in trump in the 2020 general election whereas those of liberal ideology are not considering voting for trump, only 9% show an interest in him.

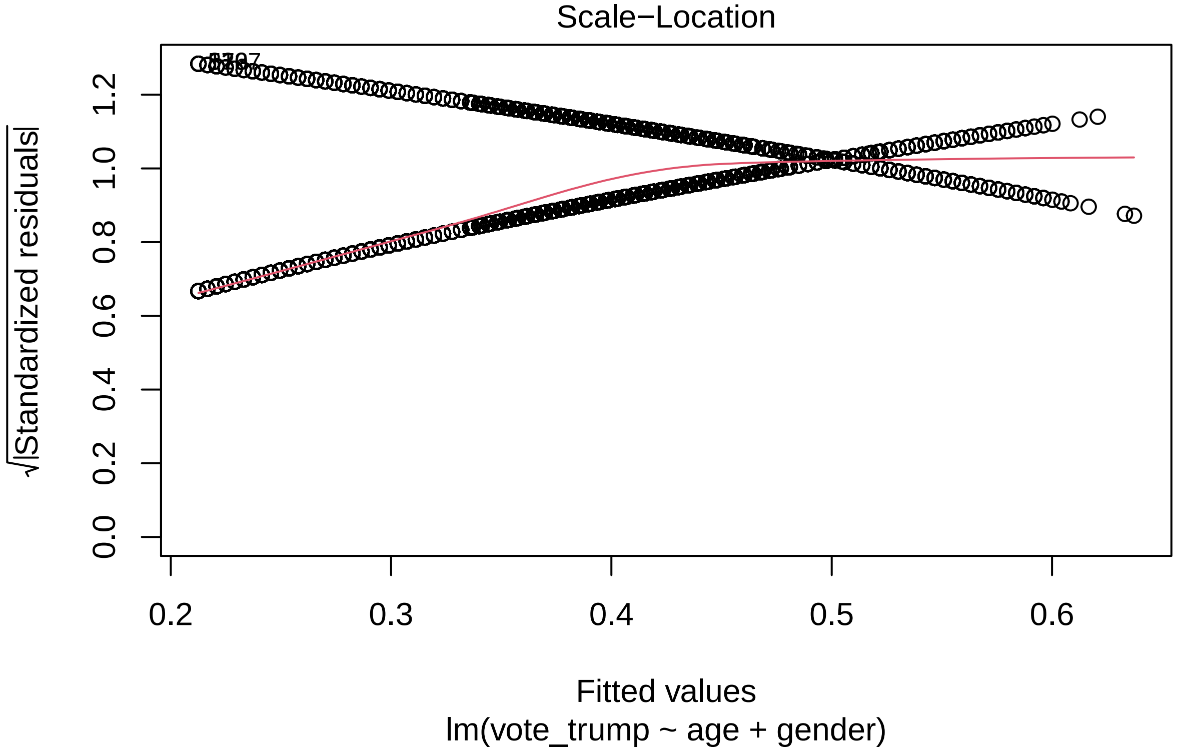
```
library(dplyr)
model2 <- lm(vote_trump ~ age+gender, data=survey_data);
summary(model2)
```

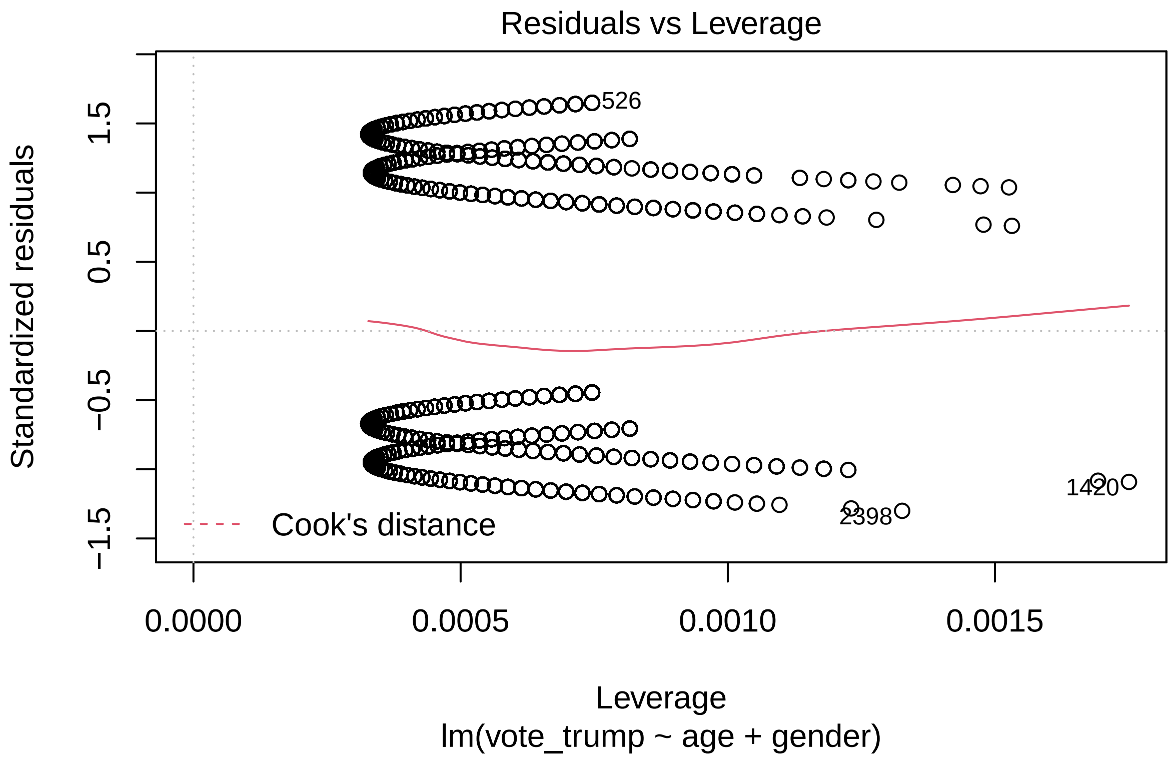
```
##
## Call:
## lm(formula = vote_trump ~ age + gender, data = survey_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6208 -0.3976 -0.2660  0.5643  0.7874
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1385438  0.0185840   7.455 1.02e-13 ***
## age         0.0041119  0.0003707  11.091  < 2e-16 ***
## genderMale  0.1244835  0.0122906  10.128  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4778 on 6064 degrees of freedom
## Multiple R-squared:  0.03806,    Adjusted R-squared:  0.03774
## F-statistic:   120 on 2 and 6064 DF,  p-value: < 2.2e-16
```

```
plot(model2)
```




Normal Q−Q
Standardized residuals
−1.0
0.0
0.5
1.0
1.5
1707
−4
−2
0
2
4
Theoretical Quantiles
lm(vote_trump ~ age + gender)


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1707
0.2
0.3
0.4
0.5
0.6
Fitted values
lm(vote_trump ~ age + gender)



```
census_data1<-census_data[1:2783,]

predicted2 <- plogis(predict(model2, census_data1))
head(predicted2)
```

```
##         1         2         3         4         5         6
## 0.6140832 0.5941490 0.6023260 0.5498889 0.5794689 0.5691410
```

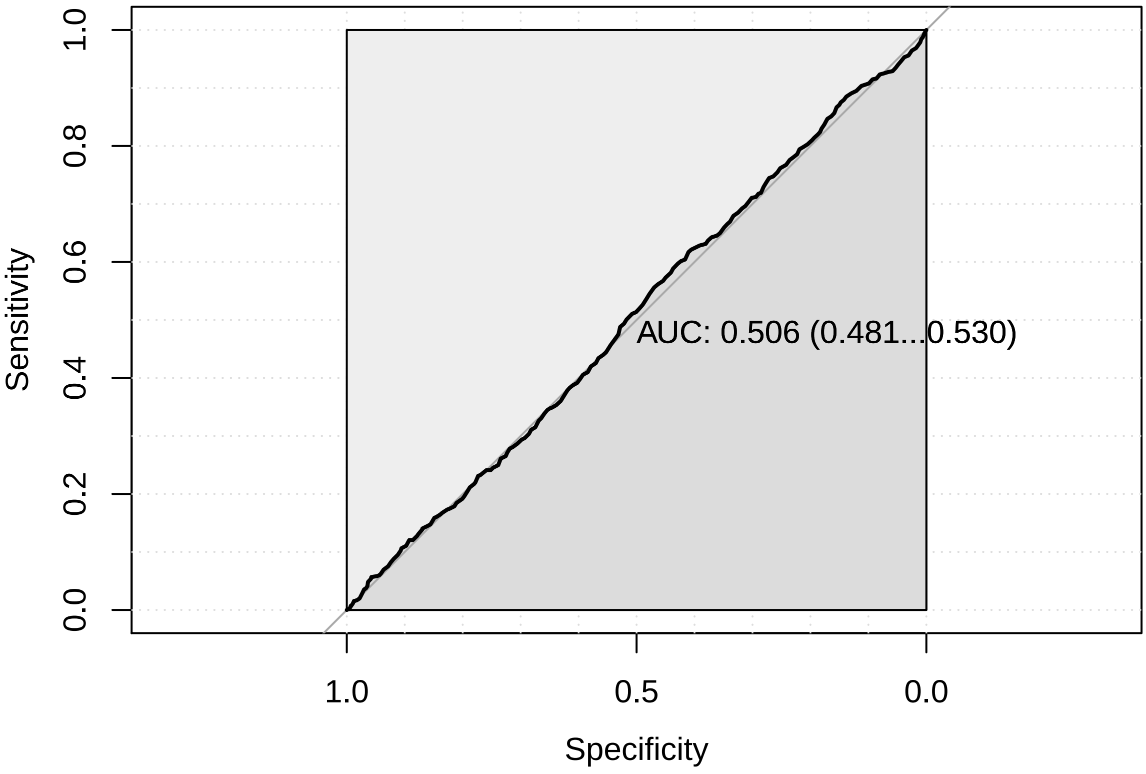
```
length(predicted2)
```

```
## [1] 2783
```

```
pROC_obj=roc(testData$vote_trump, predicted2,smoothed = TRUE,
            ci=TRUE, ci.alpha=0.9, stratified=FALSE,
            plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
            print.auc=TRUE, show.thres=TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



# Results Even before fititng the model, it was clear from the frquency tabulation that most of the individuals would not vote for trump, up to 61.3%(3720) stated that they were against Trump's bid. From the sample, only 38.7%(2347) of the indicated they would vote for Trump.

The results of the model indicated that the age of individuals, intention to vote a were significant in expalining the election outcome. As the age of an individual increases, the likelihood of that individual voting for trump decreases, this is shown by the negative age coefficent estimate.

## Discussion

The survey intended to establish how favorable is Donald Trump in the US. The survey sample findings show that 21% of the sample population consider Trump to be very favorable while 42% consider him very unfavorable, 6 % haven't heard enough about him. The 21% that consider him very favorable are those with very conservative ideology (63%), those who voted for him in the 2016 general elections, and the republicans. Those who consider Trump to be very unfavorable are those with liberal ideology, those who voted for Clinton and Jill in the 2016 general elections, the blacks and the Hispanic, the female some whites. 11% of the blacks haven't heard enough about Trump. Generally, Trump is considered unfavorable as can be inferred from the findings.

The predictive model computed the probability of voting trump. Because we have to set the length of response and predictor equal. There are 2783 rows being selected from census data, which is equal to the length of testData. The result of predicted2 tells the probability of voting trump based on gender and age factors.

### Weaknesses

In the process of conducting the analysis, it was noted that the analysis was highly impacted by presence of inconsistent observations such as missing values. A significant effort was undertaken trying to format the data in a manner would make it workable. Future procedure in data collection should be more rigorous to limit the chances of errors and inconsistencies in the data.

### Next Steps

Subsequent works related to the study should consider inclusion of more variables in the model. it would also help using other classification techniques such as the random forest model and the artificial neural network models and compare their performance with the linear regression models.

## References

Singh, P., Sawhney, R. S., & Kahlon, K. S. (2017, November). Forecasting the 2016 US presidential elections using sentiment analysis. In Conference on e-Business, e-Services and e-Society (pp. 412-423). Springer, Cham.

Survey data source; https://www.voterstudygroup.org/publication/nationscape-data-set Acs census data, IPUMS: https://usa.ipums.org/usa/index.shtml