Machine Learning on the 2016 Presidential Election Data Set

In this data set, my dependent variable is percent\_republican, the percentage of the vote in each county that Republican candidate Donald Trump received in the 2016 presidential election. Given that percent\_republican is a continuous variable, a natural starting point for a predictive model of percent\_republican is a linear regression model.

Initially, the 3145 observations were divided into a training set and a test set, with 80% of the observations in the training set and 20% in the test set.

As a starting point, my initial model included all of the following: independent explanatory variables in the dataset:

1. average\_age: the average age of the county’s population
2. percent\_republican\_2012: the 2012 Republican presidential candidate’s share of the vote
3. percent\_white: the percentage of the county’s population that is white
4. percent\_uninsured: the percentage of the county’s population that lacks health insurance
5. percent\_degree: percentage of county’s population that has an educational degree
6. average\_income: average annual income of the county’s population
7. percent\_unemployed: percentage of the county’s population that is unemployed

I did not include percent\_democrat\_2012 because it would be highly linearly correlated with percent\_republican 2012, since percent\_republican\_2012+percent\_democrat\_2012 is approximately 100%, excluding the 3rd party vote. In the same vain, percent\_democrat can be considered another dependent variable that is highly linearly correlated with our variable of interest, percent republican.

Here is a summary (taken from R-Studio) of this initial model:

Model 1 (all 7 explanatory variables)

lm(formula = percent\_republican ~ percent\_republican\_2012 + average\_income +

average\_age + percent\_white + percent\_uninsured + percent\_degree +

percent\_unemployed, data = na.omit(train))

Residuals:

Min 1Q Median 3Q Max

-27.9186 -1.5361 0.0924 1.8183 11.5049

Coefficients:

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

(Intercept) 8.511e+00 9.118e-01 9.333 < 2e-16 \*\*\*

percent\_republican\_2012 7.916e-01 6.854e-03 115.485 < 2e-16 \*\*\*

average\_income 3.600e-05 1.635e-05 2.202 0.0278 \*

average\_age 1.283e-01 1.532e-02 8.375 < 2e-16 \*\*\*

percent\_white 1.545e-01 5.734e-03 26.943 < 2e-16 \*\*\*

percent\_uninsured 2.839e-02 1.918e-02 1.481 0.1388

percent\_degree -4.816e-01 9.867e-03 -48.810 < 2e-16 \*\*\*

percent\_unemployed -2.247e-01 2.771e-02 -8.109 7.93e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.158 on 2482 degrees of freedom

Multiple R-squared: 0.9589, Adjusted R-squared: 0.9587

F-statistic: 8265 on 7 and 2482 DF, p-value: < 2.2e-16

In this initial linear model, model 1, all the explanatory variables are significant at an alpha =0.05 level, except for percent\_uninsured, which has a p-value of 0.139. The adjusted

R-squared for this model was 0.9587. The SSE (sum of standard errors) was 24748.3, and the root mean square error was 3.15. After obtaining of a summary of this initial model, I decided to check to see if the signs of the coefficients for each variable in the model matched the sign of the correlation coefficient of each variable with our dependent variable, percent\_republican. The only variable that had mismatched signs was average\_income, which had a positive coefficient in the initial linear regression model but a negative correlation coefficient with percent\_republican.

This mismatch in signs indicated problems with collinearity, so I created a correlation matrix among all the independent explanatory variables to check for collinearity.

Among the 7 independent explanatory variables in our intial model, the pairs of variables with the highest degree of collinearity were:

1) average\_age and percent\_white (r=0.456)

2) percent\_republican\_2012 and percent\_white(r=0.433)

# 3) percent\_white and percent\_uninsured(r= -0.473)

# 4) percent\_uninsured and average income (r=0.436)

# 5) percent\_degree and average income (r=0.51)

After this initial analysis, I decided to remove the average\_income variable from the model for the 3 following reasons:

1. Mismatched signs
2. High degree of collinearity with percent\_uninsured and percent\_degree
3. Correlation coefficient with dependent variable percent\_republican was only -0.20.

My reasoning was that if average\_income was removed, perhaps percent\_uninsured would become a significant variable in the model.

Therefore, the second model I tested, model\_income\_removed, had all of the explanatory variables in model 1, except for average\_income. Here is a summary of this 2nd model.

Model 2 (6 explanatory variables, average\_income removed)

lm(formula = percent\_republican ~ percent\_republican\_2012 + percent\_degree +

average\_age + percent\_white + percent\_unemployed + percent\_uninsured,

data = na.omit(train))

Residuals:

Min 1Q Median 3Q Max

-28.244 -1.564 0.095 1.822 11.577

Coefficients:

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

(Intercept) 9.412875 0.815182 11.547 < 2e-16 \*\*\*

percent\_republican\_2012 0.795446 0.006632 119.936 < 2e-16 \*\*\*

percent\_degree -0.471674 0.008788 -53.673 < 2e-16 \*\*\*

average\_age 0.131899 0.015244 8.653 < 2e-16 \*\*\*

percent\_white 0.151230 0.005544 27.281 < 2e-16 \*\*\*

percent\_unemployed -0.221358 0.027687 -7.995 1.96e-15 \*\*\*

percent\_uninsured 0.010694 0.017424 0.614 0.539

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.16 on 2483 degrees of freedom

Multiple R-squared: 0.9588, Adjusted R-squared: 0.9587

F-statistic: 9627 on 6 and 2483 DF, p-value: < 2.2e-16

In this model, percent\_uninsured was still not a significant variable with p value of 0.54. All other explanatory variables were highly significant. Given that percent\_uninsured was highly correlated with percent\_white and that the correlation coefficient between percent\_uninsured and percent\_republican was 0.195, I decided to remove percent\_uninsured from the model.

Thus, I formed my final predictive model for percent republican, summarized below.

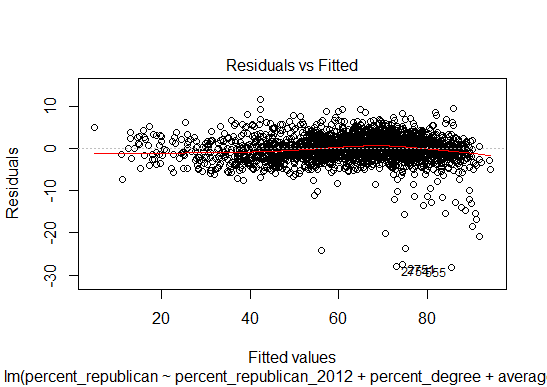
Final model(5 explanatory variables, average\_income and percent\_uninsured removed)

|  |
| --- |
| lm(formula = percent\_republican ~ percent\_republican\_2012 + percent\_degree +  average\_age + percent\_white + percent\_unemployed, data = na.omit(train))  Residuals:  Min 1Q Median 3Q Max  -28.269 -1.566 0.093 1.824 11.532  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 9.579112 0.768778 12.460 < 2e-16 \*\*\*  percent\_republican\_2012 0.797708 0.005513 144.697 < 2e-16 \*\*\*  percent\_degree -0.472347 0.008718 -54.180 < 2e-16 \*\*\*  average\_age 0.133391 0.015046 8.865 < 2e-16 \*\*\*  percent\_white 0.148985 0.004164 35.775 < 2e-16 \*\*\*  percent\_unemployed -0.218614 0.027320 -8.002 1.86e-15 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.16 on 2484 degrees of freedom  Multiple R-squared: 0.9588, Adjusted R-squared: 0.9587  F-statistic: 1.156e+04 on 5 and 2484 DF, p-value: < 2.2e-16 |

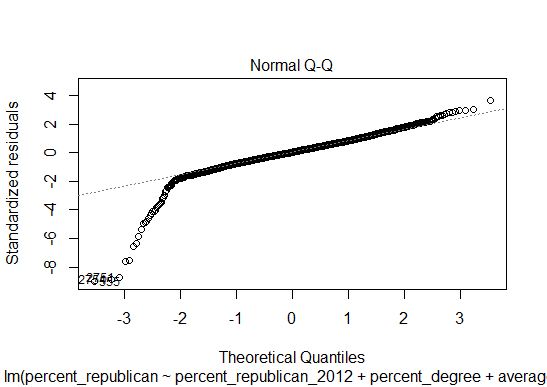
This final model was a good model. After removing two variables due to problems with

collinearity, the simplified 5 variable model has an adjusted R2 that has remained constant at a high value of 0.9587. All five variables were highly significant and had coefficients with the correct sign. Compared to our initial model with 7 explanatory variables, this simplified 5 variable model had insignificantly increased SSE and MRSE( 24748 to 24800 and 3.15 to 3.16 respectively).

I then plotted the final predictive model (model\_income\_and\_uninsued\_removed) to test for linear regression assumptions. Based on these plots, the predictive model meets the assumptions necessary for linear regression. The average of the residuals appears to be close to 0, for the most part, as seen in the plot of residual values versus the model’s predicted value of Republican vote share. The red line in the plot indicates the mean residual value; it increases to slightly above 0 at predicted values around 65% and decreased to slightly below 0 at predicted values around 80%. Overall, however, we can assume that the mean residual value is close to 0.



Another assumption of linear regressions is that the residual values are normally distributed. This plot of standardized residuals versus theoretical quantiles shows that for the most part, the residual values are distributed normally. There is a small minority of outlier values at the right side of the plot; these represent predicted values where the error was in the negative direction, meaning that for these values, the model underpredicted the Republican share of the vote. However, for a significant majority of the dataset, the residual values are normally distributed.



This final model was built using the train data set. I took this final model and used it on the test data set to create a new variable in the test data set, predict\_republican. This variable consisted of all the predicted values of the 2016 presidential election Republican percentage vote share using this model that was built using the training set. The correlation coefficient between the predicted 2016 presidential election Republican percentage vote share(predict\_republican) and the actual percentage(percent\_republican) in the test dataset was 0.974. R-squared value was 0.974^2=0.949. The high R and R-squared values indicate that this final predictive model is a good fit for the test data.

Interpreting the Model/Conclusion

Percent\_republican\_2012 is the most significant explanatory variable. With a coefficient of 0.798 in the linear regression model, it can be predicted the for every 1 percentage increase in the Republican candidate’s share of the vote in the 2012 presidential election, the Republican candidate’s share of the vote in the 2016 presidential election will increase by 0.8 percentage points. This is not surprising given that for most people party affiliation does not change over a relatively short time period of 4 years.

Another significant explanatory variable is percent\_white. With a coefficient of 0.149, it can be predicted that for every 1 percentage point increase in the portion of the population that is white, the 2016 Republican share of the presidential election vote increases by almost 0.15 percentage points. This is not surprising given that in the past 30 to 40 years, if not longer, minority voters tend to vote for Democrat candidates, while Republican candidates poll better among white voters.

Another significant explanatory variable is average\_age. With a coefficient of 0.133, the model predicts that for every 1 year increase in the average age of the county population, the 2016 Republican share of the presidential vote in that county will increase by 0.13 percentage points. This is not surprising given that Republicans historically tend to poll better with older people.

Of interest, both percent\_degree and percent\_unemployed are the two explanatory variables in the model that are negatively correlated with the Republican share of the vote in 2016. As the percentage of the population that has a college degree or that is unemployed increases, the Republican share of the vote decreases, and the Democrat share of the vote increases. Of note, of the 5 variables in this model, percent\_degree and percent\_unemployed are the two with the highest degree of correlation with the average income variable. However, they are correlated with average income in opposite directions. Percent\_degree is positively correlated with average\_income, while percent\_unemployed is negatively correlated with average income, both of which make intutitive sense. These conflicting correlations are probably contributing to the fact that average\_income is such a poor explanatory variable for the Republican share of the vote, leading it to not be included in the model. In practical terms, this means that Democrats should target both rich and poor counties. They should target counties with a high proportion of educational degree-bearing individuals; these counties are more likely to be rich. However, Democrats should also focus on turning out the vote in counties with a high percentage of unemployed individuals; these counties are more likely to be poor.

With regards to overall voter turnout strategy, Republican political operatives should focus on counties where a larger percentage of the population is elderly, white, lacking an educational degree, and employed. On the other hand, Democrats should be focusing on counties where a larger percentage of the population is young, non-white, college-educated or higher, and unemployed.

Index

Model 1: (All 7 explanatory variables)

lm(formula = percent\_republican ~ percent\_republican\_2012 + average\_income +

average\_age + percent\_white + percent\_uninsured + percent\_degree +

percent\_unemployed, data = na.omit(train))

Residuals:

Min 1Q Median 3Q Max

-27.9186 -1.5361 0.0924 1.8183 11.5049

Coefficients:

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

(Intercept) 8.511e+00 9.118e-01 9.333 < 2e-16 \*\*\*

percent\_republican\_2012 7.916e-01 6.854e-03 115.485 < 2e-16 \*\*\*

average\_income 3.600e-05 1.635e-05 2.202 0.0278 \*

average\_age 1.283e-01 1.532e-02 8.375 < 2e-16 \*\*\*

percent\_white 1.545e-01 5.734e-03 26.943 < 2e-16 \*\*\*

percent\_uninsured 2.839e-02 1.918e-02 1.481 0.1388

percent\_degree -4.816e-01 9.867e-03 -48.810 < 2e-16 \*\*\*

percent\_unemployed -2.247e-01 2.771e-02 -8.109 7.93e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.158 on 2482 degrees of freedom

Multiple R-squared: 0.9589, Adjusted R-squared: 0.9587

F-statistic: 8265 on 7 and 2482 DF, p-value: < 2.2e-16

Model 2 (6 explanatory variables, average\_income removed)

lm(formula = percent\_republican ~ percent\_republican\_2012 + percent\_degree +

average\_age + percent\_white + percent\_unemployed + percent\_uninsured,

data = na.omit(train))

Residuals:

Min 1Q Median 3Q Max

-28.244 -1.564 0.095 1.822 11.577

Coefficients:

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

(Intercept) 9.412875 0.815182 11.547 < 2e-16 \*\*\*

percent\_republican\_2012 0.795446 0.006632 119.936 < 2e-16 \*\*\*

percent\_degree -0.471674 0.008788 -53.673 < 2e-16 \*\*\*

average\_age 0.131899 0.015244 8.653 < 2e-16 \*\*\*

percent\_white 0.151230 0.005544 27.281 < 2e-16 \*\*\*

percent\_unemployed -0.221358 0.027687 -7.995 1.96e-15 \*\*\*

percent\_uninsured 0.010694 0.017424 0.614 0.539

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.16 on 2483 degrees of freedom

Multiple R-squared: 0.9588, Adjusted R-squared: 0.9587

F-statistic: 9627 on 6 and 2483 DF, p-value: < 2.2e-16

Final model(5 explanatory variables, average\_income and percent\_uninsured removed)

|  |
| --- |
| lm(formula = percent\_republican ~ percent\_republican\_2012 + percent\_degree +  average\_age + percent\_white + percent\_unemployed, data = na.omit(train))  Residuals:  Min 1Q Median 3Q Max  -28.269 -1.566 0.093 1.824 11.532  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 9.579112 0.768778 12.460 < 2e-16 \*\*\*  percent\_republican\_2012 0.797708 0.005513 144.697 < 2e-16 \*\*\*  percent\_degree -0.472347 0.008718 -54.180 < 2e-16 \*\*\*  average\_age 0.133391 0.015046 8.865 < 2e-16 \*\*\*  percent\_white 0.148985 0.004164 35.775 < 2e-16 \*\*\*  percent\_unemployed -0.218614 0.027320 -8.002 1.86e-15 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.16 on 2484 degrees of freedom  Multiple R-squared: 0.9588, Adjusted R-squared: 0.9587  F-statistic: 1.156e+04 on 5 and 2484 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |