Multivariate Analysis of the 2020 US Elections

Università Commerciale Luigi Bocconi

Flavio Caroli, Kristian Gjika

1. Introduction

The present study was conducted to examine the factors that may impact voting patterns in the United States 2020 presidential elections, with a specific focus on the percentage of votes received by the Democratic Party's candidates for each state.
In the interest of understanding the potential influences on voting behavior, we utilized a dataset containing information on the number of votes received by each political party as well as various socio-economic variables for each state. We chose to use a sample data set from a single country, the United States, in order to minimize potential biases that may result from disparate conditions such as wars or different government systems.
Before conducting our analysis, it was important to ensure that the dataset was in a usable format; to this end, we performed a thorough cleaning and parsing of the data to eliminate any inconsistencies or errors. Once the dataset was prepared, we applied a multilinear regression model in order to identify the variables that had a significant effect on the percentage of votes received by the Democratic Party, and finally, use some selection models to exclude useless variables and validate our final model.

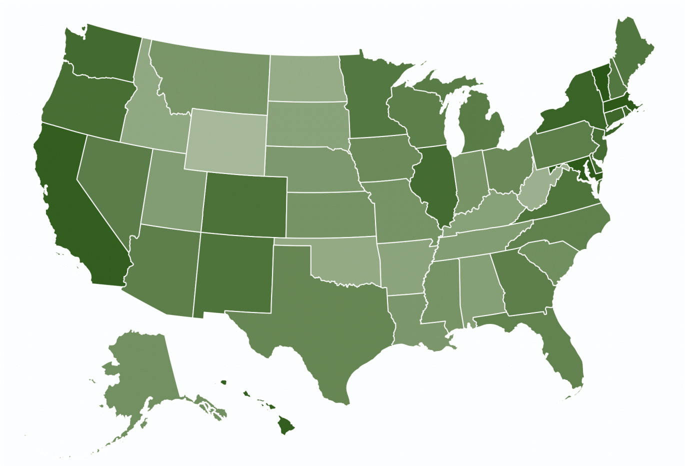


Figure 1: Votes distribution US

2. Dataset

Our dataset includes 50 data points for each variable, representing each state in the US. In order to ensure that our dataset was representative of the United States as a whole, and to avoid skewing the results of our analysis, we made the decision to exclude data from the District of Columbia due to its unique status as the seat of the U.S. federal government and home to several international organizations, in fact, we found that the data for District of Columbia was significantly different from the other states in many of the variables we analyzed.

To gain a comprehensive understanding of the factors that may impact voting patterns, we included general factors such as unemployment and median age, as well as variables specific to the cultural and historical context of the US, such as the percentage of black people and the percentage of war veterans. Additionally, we included data on COVID-19, as this pandemic significantly impacted election campaigns and polarized political opinions globally.
After collecting and cleaning the data, we normalized some variables based on the total population, such as veterans and positive cases on the day of the elections, and compiled all of the information into a single table for analysis. It follows the complete list of all variables and their abbreviation used in the R script:

- **dem**: Percentage of votes for the Democratic Party
- **General factors**:
  - **boh**: Percentage of the population with a bachelor's degree or higher
  - **wg**: Gender wage gap (USD)
  - **unr**: Unemployment rate
  - **rmhi**: Real median household income (USD)
  - **ma**: Median age
  - **cr**: Crime rate (per 100,000 people)
  - **cli**: Cost of living index
- **US-specific factors**:
  - **vet**: Percentage of the population that is a veteran
  - **fh**: Percentage of households with firearms
  - **pbp**: Percentage of the population that is Black
- **COVID-19 Factors**:
  - **fv**: Percentage of the population that is fully vaccinated against SARS-CoV-2
  - **pos**: Percentage of positive SARS-CoV-2 cases on the day of the elections

3. First Visualization

To gain insight into the distribution of the data and identify potential relationships between the variables, we plotted each of the independent variables against the percentage of votes received by the Democratic Party in R, including a representation of a linear regression in each plot. Through this preliminary analysis, two variables, **boh** and **fv** emerged as highly correlated with the percentage of votes received by the Democratic Party. On the other hand, **wg** had a negative correlation with the percentage of votes received by the Democratic Party. Nevertheless, we will further analyze our model using multilinear regression and model selection algorithms to examine these relationships more thoroughly.

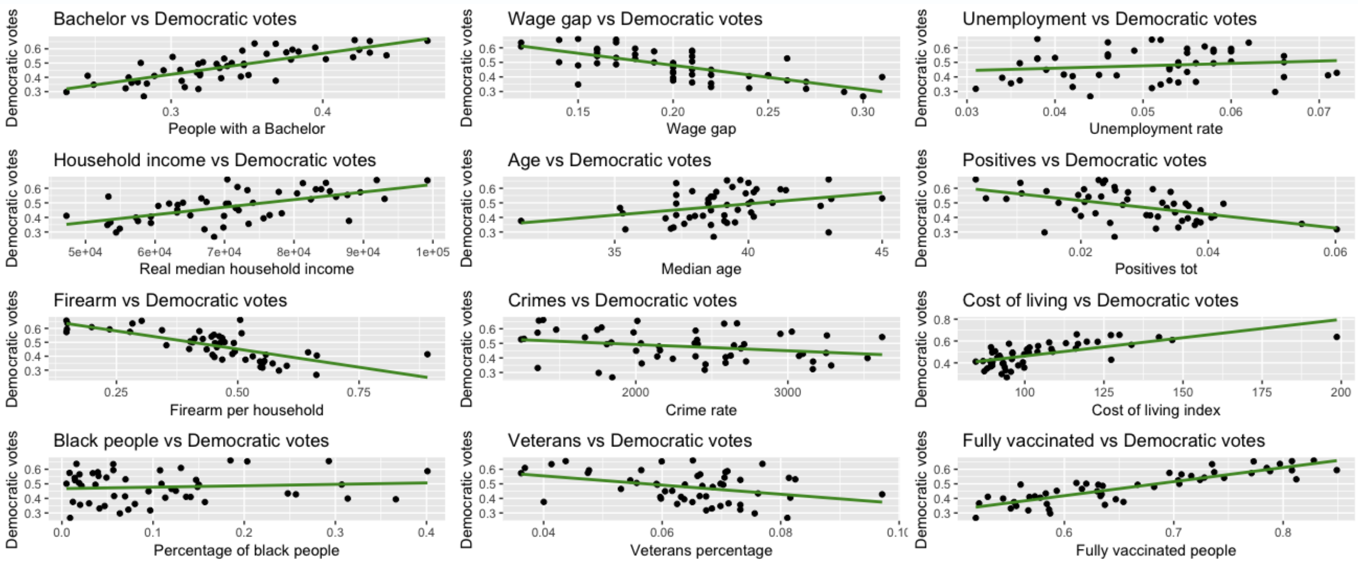


Figure 2: ggplot of all variables

4. 1^{st} multilinear regression

To identify the variables that may have had a significant effect on the percentage of votes received by the Democratic Party, we employed a multilinear regression model. The model was fitted using the "lm" function in R, with the percentage of votes received by the Democratic Party as the dependent variable and the socio-economic variables as the independent variables. Upon conducting the multivariate linear regression, it was observed that the Adjusted R-squared value indicates that our chosen variables can explain a significant portion of the variation in voting patterns. However, only four variables, **boh**, **wg**, **unr**, and **fv**, were found to be statistically significant at $\alpha = 0.05$. Given that we included a large number of variables in this initial regression, there was the risk of overfitting, and it is likely that the model could not generalize well to new data. Therefore, we attempted to simplify the model by applying model selection algorithms to identify the most relevant variables and remove those that did not significantly impact the R-squared value.

5. Model Selection

Despite the high percentage of variance explained by the initial model, which included 12 variables, many of these variables had high p-values and the overall model was prone to overfitting. In order to improve the model, we implemented two model selection methods (step-up and step-down) to identify the most relevant variables. These methods began with an empty (full) model and tested the null hypothesis ($H_0 : \beta_i = 0$) and alternative hypothesis ($H_1 : \beta_i \neq 0$) for each variable in the model, adding (removing) variables with the smallest (largest) p-values, at a significance level of $\alpha = 0.05$. This process was repeated until the optimal model was identified. The output of both model selection methods was consistent, with 8 variables being removed to simplify the model while minimizing the impact on the adjusted R-squared coefficient. The final model included the variables **boh**, **fv**, **wg**, and **unr**, which were used in a second multivariate regression analysis. The results of this process are presented in the appendix.

6. 2^{nd} multilinear regression

The regression analysis showed that the four variables chosen were good at explaining the variance in votes for the Democratic party. The results were reliable, as the p-values for each variable were statistically

significant and the assumptions of the regression model were satisfied after analyzing the residuals.
We checked for *Multicollinearity* by calculating the variance inflation factor (VIF) for each variable. A VIF value between 1 and 5 indicates a moderate correlation, and a value above 5 indicates a severe correlation. The variable **fv** had a somewhat high VIF value (a little more than 3), but removing it from the model significantly decreased the R-squared value, so we kept it in the model. Another assumption of the regression model is *Homoscedasticity*, which refers to the equal variance of the error terms. To assess this assumption, we plotted the residuals against the fitted values of the dependent variable. From the resulting graph, it was clear that the residuals were evenly distributed around the horizontal line, indicating the presence of homoscedasticity in the model. The *Normality* of the residuals was also checked, as this is an assumption of the regression model. We plotted a Q-Q plot of the residuals and observed that while most of the points were located on the diagonal line, but the extremes deviated from it and displayed an "S" shape. However, further testing using the K-S and S-W tests did not reject the null hypothesis of normal distribution at the alpha = 0.05 confidence level.
In addition to these analyses, we also wanted to visualize the predictive efficiency of our model.

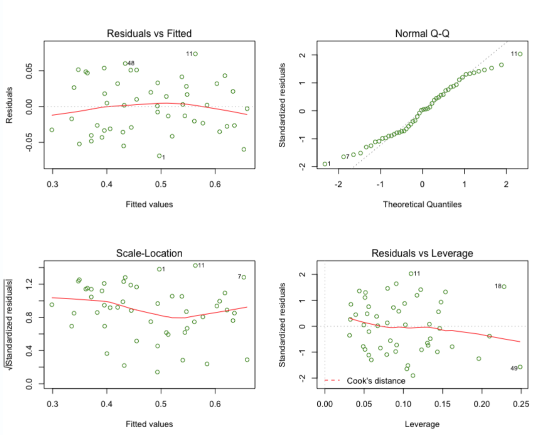


Figure 3: Plot of the residuals

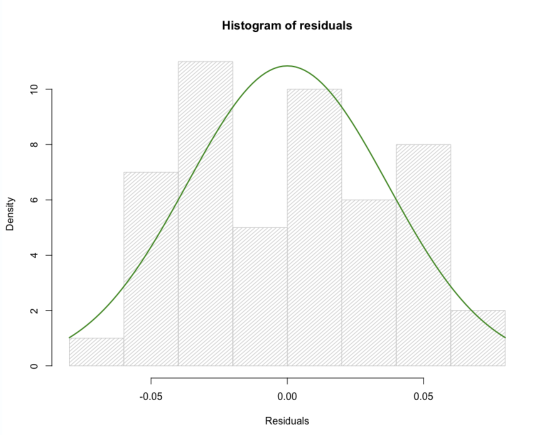
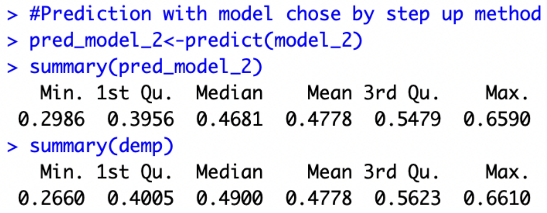


Figure 4: Histogram of the residuals



```
> #Prediction with model chose by step up method
> pred_model_2<-predict(model_2)
> summary(pred_model_2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2986  0.3956  0.4681  0.4778  0.5479  0.6590 
> summary(demp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2660  0.4005  0.4900  0.4778  0.5623  0.6610 
```

Figure 5: Prediction compared to observed votes percentages

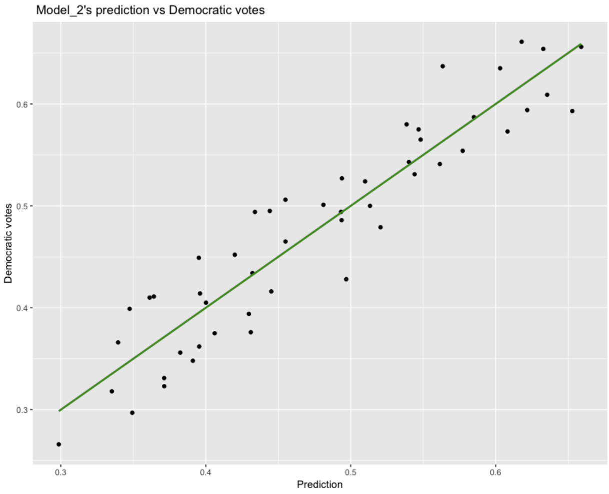


Figure 6: Predictions of the model

The resulting plot showed a strong correlation between the prediction obtained with the second model and the actually observed votes, with similar means. Overall, these results demonstrate the robustness and reliability of our model.

7. k-fold cross validation

Before concluding, we wanted to verify the significance of our model, using k-fold cross-validation. This method allows us to assess the model's ability to classify new data. We began by running the train() function using the K-Nearest Neighbors method to identify the optimal value of k. The data was preprocessed by subtracting the mean and dividing it by the variance. The results showed that the optimal value of k, as determined by the smallest root mean square error (RMSE), was k = 15 (evaluated using the test data). This indicates that the training data was optimally divided into 15 samples by the second train() function used (15-fold validation) and the resulting model performed well, with low RMSE for predicting the response variables and a high R-squared coefficient. It is worth noting that the small size of the dataset, may limit the performance of the method since it is likely to improve with larger datasets for training and testing.

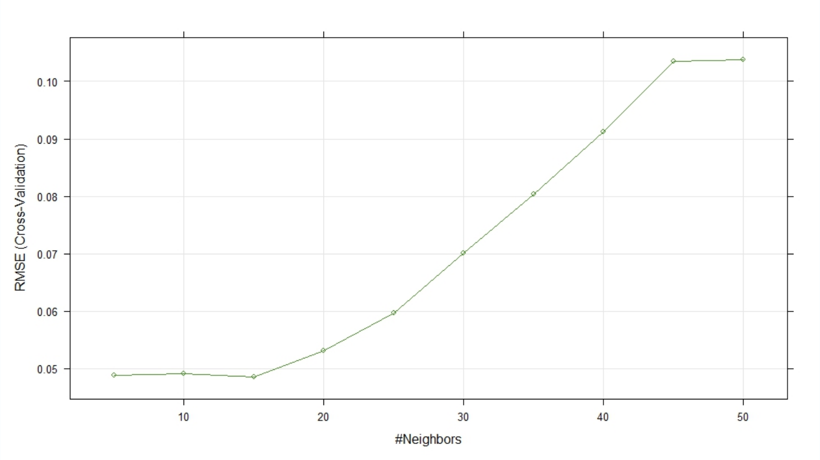


Figure 7: k-selection

8. Final Discussion

It is essential to recognize that the findings of this study should be interpreted with caution due to the limitations of the dataset. The data only includes information from the 2020 presidential election and does not take into account all potential variables that may be relevant to gain a deeper understanding of the complex and dynamic nature of political behavior in the United States; indeed, even if with our model we got a pretty high correlation with our explanatory variables and the response variable, the potential influence of other variables (such as territorial ones and social network propaganda) could improve the accuracy of the prediction, increasing the already high adjusted R-squared value we obtained. It must also be said that his study only focused on the percentage of votes received by the Democratic Party, and therefore may not provide a comprehensive understanding of voting patterns in the United States; moreover, the focus of the study was limited to the presidential elections and did not examine other types of elections such as congressional or local ones. Taking into account all the counties of every state, we could use more data points, improve the residuals analysis and increase the efficiency of the cross-validation, in order to be also able to compare a multilinear regression model, with other non-linear models, and choose the optimal one. Despite the limitations we already discussed, we are quite satisfied with the results of the analysis highlighting that higher levels of education, smaller gender wage gaps, lower unemployment rates, and higher vaccination rates against SARS-CoV-2 may be correlated with a higher percentage of votes received by the Democratic Party.

9. Appendix A - Plots

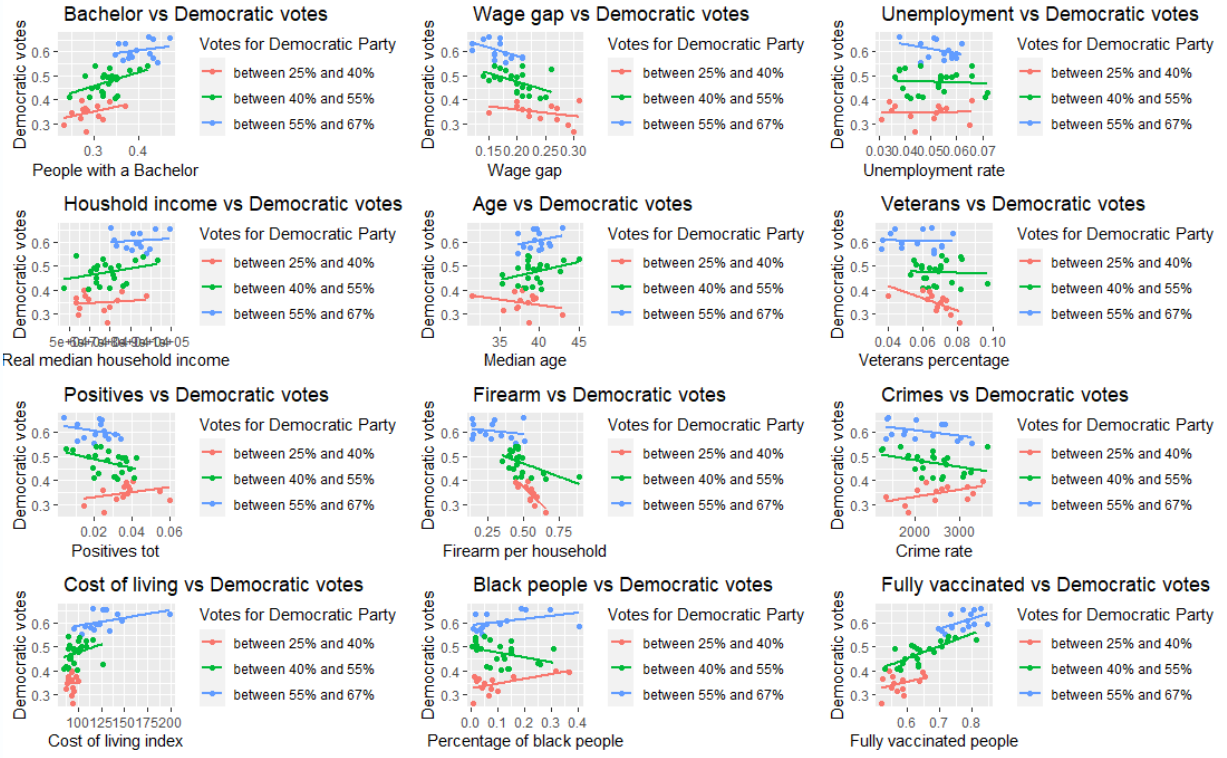
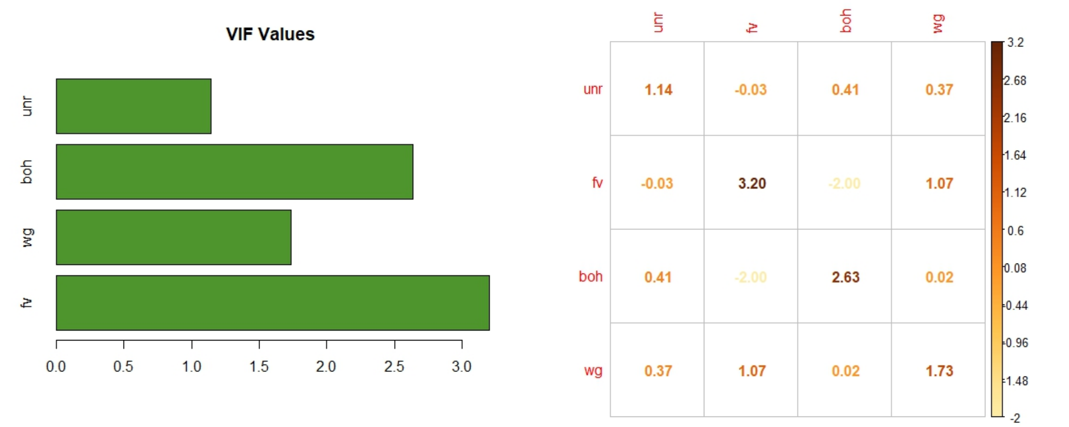


Figure 8: Plot of all variables divided by vote category



	unr	fv	boh	wg
unr	1.14	-0.03	0.41	0.37
fv	-0.03	3.20	-2.00	1.07
boh	0.41	-2.00	2.63	0.02
wg	0.37	1.07	0.02	1.73

Figure 9: Multicollinearity

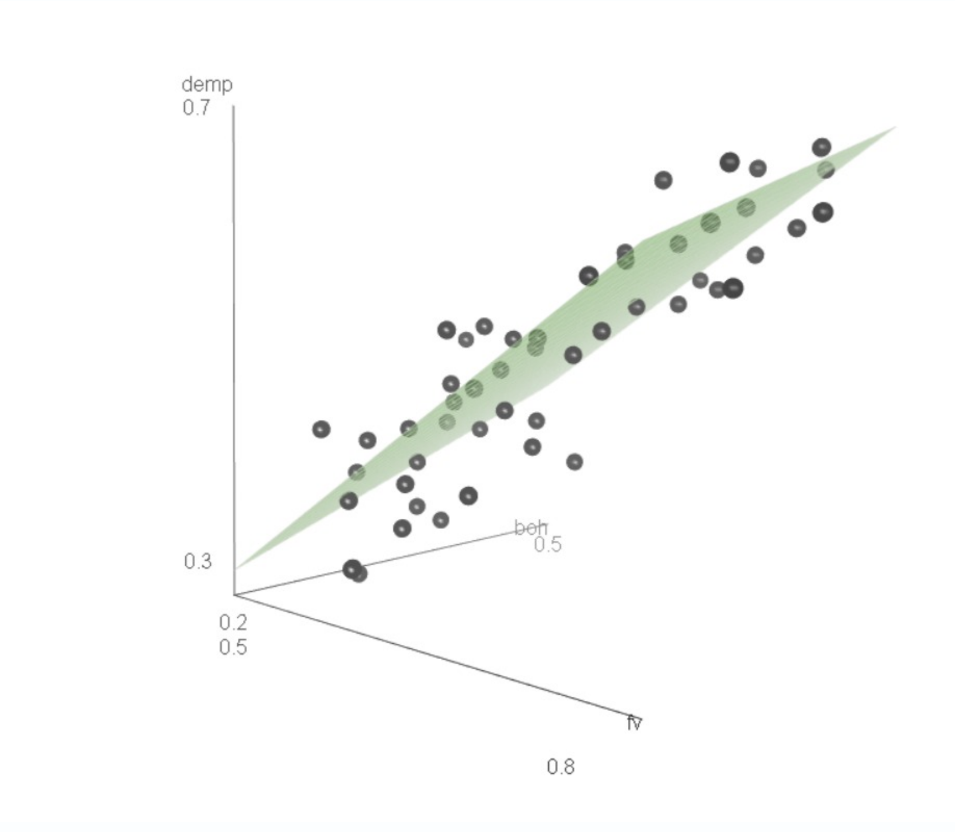
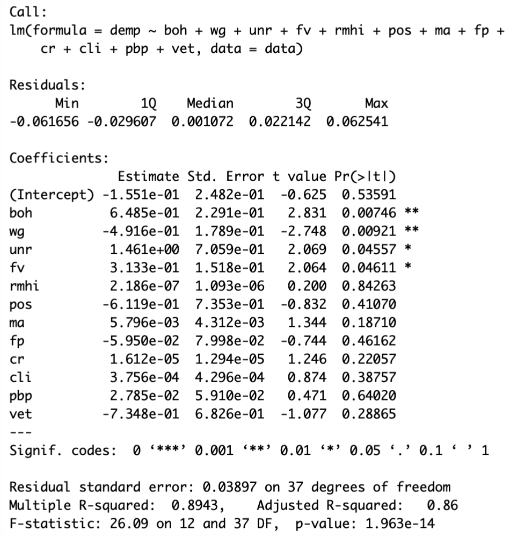


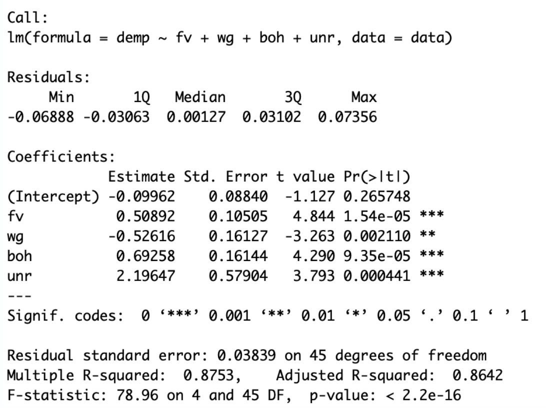
Figure 10: Plot in three dimensions with the two most influential variables

10. Appendix B - R codes' Output



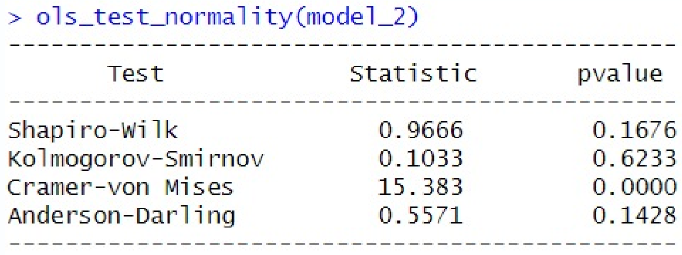
```
Call:
lm(formula = demp ~ boh + wg + unr + fv + rmhi + pos + ma + fp +
    cr + cli + pbp + vet, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.061656 -0.029607  0.001072  0.022142  0.062541

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.551e-01  2.482e-01  -0.625  0.53591
boh          6.485e-01  2.291e-01   2.831  0.00746 **
wg          -4.916e-01  1.789e-01  -2.748  0.00921 **
unr          1.461e+00  7.059e-01   2.069  0.04557 *
fv           3.133e-01  1.518e-01   2.064  0.04611 *
rmhi         2.186e-07  1.093e-06   0.200  0.84263
pos         -6.119e-01  7.353e-01  -0.832  0.41070
ma           5.796e-03  4.312e-03   1.344  0.18710
fp          -5.950e-02  7.998e-02  -0.744  0.46162
cr           1.612e-05  1.294e-05   1.246  0.22057
cli          3.756e-04  4.296e-04   0.874  0.38757
pbp          2.785e-02  5.910e-02   0.471  0.64020
vet         -7.348e-01  6.826e-01  -1.077  0.28865
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03897 on 37 degrees of freedom
Multiple R-squared:  0.8943,    Adjusted R-squared:    0.86
F-statistic: 26.09 on 12 and 37 DF,  p-value: 1.963e-14
```

Figure 11: 1^{st} linear regression



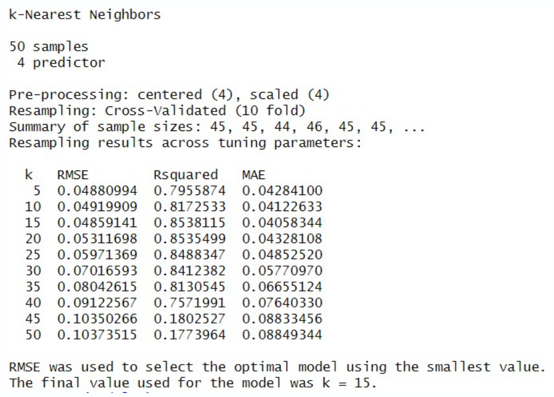
```
Call:
lm(formula = demp ~ fv + wg + boh + unr, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06888 -0.03063  0.00127  0.03102  0.07356

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09962    0.08840  -1.127 0.265748
fv           0.50892    0.10505   4.844 1.54e-05 ***
wg          -0.52616    0.16127  -3.263 0.002110 **
boh          0.69258    0.16144   4.290 9.35e-05 ***
unr          2.19647    0.57904   3.793 0.000441 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03839 on 45 degrees of freedom
Multiple R-squared:  0.8753,    Adjusted R-squared:  0.8642
F-statistic: 78.96 on 4 and 45 DF,  p-value: < 2.2e-16
```

Figure 12: 2^{nd} linear regression



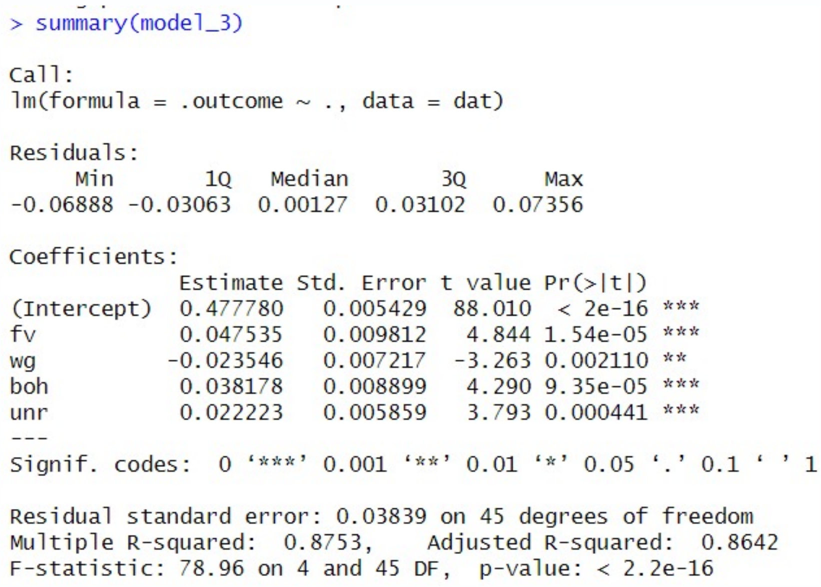
```
> ols_test_normality(model_2)
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9666         0.1676
Kolmogorov-Smirnov        0.1033         0.6233
Cramer-von Mises          15.383         0.0000
Anderson-Darling          0.5571         0.1428
-----------------------------------------------
```

Figure 13: Normality test



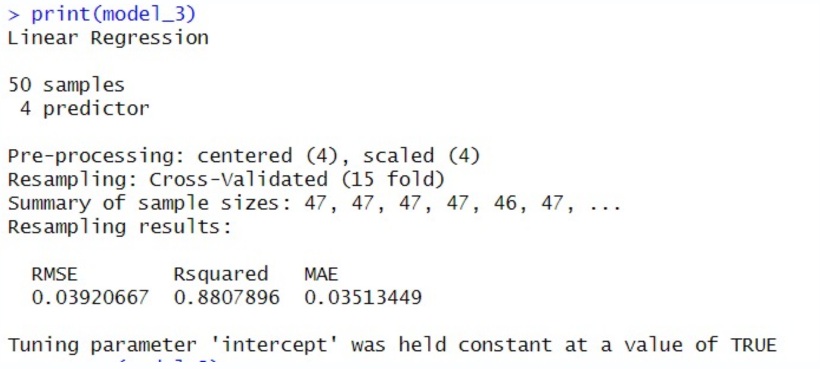
```
k-Nearest Neighbors

50 samples
 4 predictor

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 45, 45, 44, 46, 45, 45, ...
Resampling results across tuning parameters:

  k   RMSE        Rsquared   MAE
   5  0.04880994  0.7955874  0.04284100
  10  0.04919909  0.8172533  0.04122633
  15  0.04859141  0.8538115  0.04058344
  20  0.05311698  0.8535499  0.04328108
  25  0.05971369  0.8488347  0.04852520
  30  0.07016593  0.8412382  0.05770970
  35  0.08042615  0.8130545  0.06655124
  40  0.09122567  0.7571991  0.07640330
  45  0.10350266  0.1802527  0.08833456
  50  0.10373515  0.1773964  0.08849344

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 15.
```

Figure 14: kNN



```
> summary(model_3)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.06888 -0.03063  0.00127  0.03102  0.07356 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.477780   0.005429  88.010  < 2e-16 ***
fv           0.047535   0.009812   4.844 1.54e-05 ***
wg          -0.023546   0.007217  -3.263 0.002110 ** 
boh          0.038178   0.008899   4.290 9.35e-05 ***
unr          0.022223   0.005859   3.793 0.000441 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03839 on 45 degrees of freedom
Multiple R-squared:  0.8753,	Adjusted R-squared:  0.8642 
F-statistic: 78.96 on 4 and 45 DF,  p-value: < 2.2e-16
```

Figure 15: Cross-validation linear regression



```
> print(model_3)
Linear Regression 

50 samples
 4 predictor

Pre-processing: centered (4), scaled (4) 
Resampling: Cross-Validated (15 fold) 
Summary of sample sizes: 47, 47, 47, 47, 46, 47, ... 
Resampling results:

  RMSE        Rsquared   MAE       
  0.03920667  0.8807896  0.03513449

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Figure 16: Cross-validation model

11. Appendix C - Sources

- Voting results in 2020:
  `https://www.cookpolitical.com/2020-national-popular-vote-tracker`
  - Cleaned, eliminated useless data,
- Real Median Household Income by State, 2020:
  `https://fred.stlouisfed.org/release/tables?rid=249&eid=259515&od=2020-01-01#`
  - Cleaned, transposed
- Bachelor's Degree or Higher by State:
  `https://fred.stlouisfed.org/release/tables?rid=330&eid=391444&od=2020-01-01#`
  - Cleaned, transposed
- Composite cost of living index in the different states of the United States as of 2020:
  `https://www.statista.com/statistics/1240947/cost-of-living-index-usa-by-state/`
  - Hand copied
- List of U.S. states and territories by median age:
  `https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_median_age`
- Firearm percentage per household:
  `https://www.cbsnews.com/pictures/gun-ownership-rates-by-state/2/`, `https://www.gunpolicy.org/firearms/region/district-of-columbia`
- Number of Black/African-American people:
  `https://data.census.gov/table?tid=DECENNIALPL2020.P1&tp=true`
  `https://worldpopulationreview.com/states`
  - Cleaned in alphabetical order
- Number of positives to COVID - day of elections:
  `https://covidtracking.com/data`
  - Cleaned, only considering the day of elections
- Unemployment rate 16 yrs old+:
  `https://data.census.gov/table?q=employment+rate&g=0100000US$0400000&y=2020&tid=ACSST5Y2020.S301&moe=false&tp=false`
  - Cleaned/transposed from 16 years-old people or above
- Wage gap:
  `https://nwlc.org/wp-content/uploads/2019/10/Overall-Wage-Gap-State-By-State-2020.pdf`
- Crime rate by state 2020:
  `https://www.statista.com/statistics/301549/us-crimes-committed-state/`
- Number of vaccinated people 08/31/2020:
  `https://usafacts.org/visualizations/coronavirus-covid`
- Percentage of veterans:
  `https://www.va.gov/vetdata/veteran_population.asp`
  - Transformed to a percentage (divided by total pop)
- Population by state in 2020:
  `https://worldpopulationreview.com/states`
- Number of hate crimes by state 2020:
  `https://www.statista.com/statistics/737930/number-of-hate-crimes-in-the-us-by-motivation/`