## Data source

Kaggle.com datasets related to republican primary results from 1881 counties in the first 25 states. Using the data to train models and then predict the Republican Primary in California.

## Multiple Linear Regressions and Predictions

Regression models to predict the respective vote shares of Trump, Cruz, and Kasich, were built and then used to forecast the results for the upcoming California presidential primary. Randomized samples were initially generated for the training and validation sets comprising of 75% and 25% respectively of the overall data for each candidate. A total of twelve variables from the county related dataset were removed right at the outset because the correlation tests revealed that they had collinearity issues. Subsequently, several multiple linear regression models were evaluated for each candidate predicated on variegated mechanisms ranging from statistical significance, Stepwise algorithms and Akaike Information Criterion(AIC) methods to prediction error analysis on the validation set in R. Consequently, the final three models offering the best fit were selected and used for predicting the outcome of the forthcoming California primary. The respective final multiple linear regression models for the individual candidates are clarified below:

**Trump**

*Regression Equation: Trump Voting Share Outcome=4.81E-03 \* Persons 65 years and over Percent + (4.22E-03) \*High school graduate or higher Percent + (1.89E-03) \* Bachelor's degree or higher Percent + 4.66E-03 \* Mean travel time to work + (1.44E-06) \* Median household income + 3.92E-03 \* Black or African American Percent + 2.14E-03 \* American Indian and Alaska Native Percent + 9.86E-03 \* Asian alone Percent + 3.07E-03 \* White alone Percent + 2.75E-06 \* Retail Sales per Capita + 5.67E-04 \* Total number of firms + 1.27E-03 \* Hispanic-owned firms Percent + (3.34E-03) \* Female persons Percent*

The model for Trump indicates that higher levels of education, income, and percentage of female persons in a county have corresponding negative effects on the resultant vote share. On the other hand, it is revealed that a higher percentage of 65 years and over persons and a greater percentage of minorities in a county have a positive relationship with the vote share outcome. However, the latter may not necessarily signify that minorities are opting for Trump. In fact, one plausible hypothesis that could justify this result is as follows: A significant majority of people from minority groups are democrats and are not eligible to participate in the Republican primary process anyway; however, Trump’s message resonates with the blue collar working class white community within these generally low income minority neighborhoods and this particular group is voting for Trump in large numbers, which has vote share implications. Additionally, a higher mean travel time to work has a positive prediction linkage with Trump vote share. A suitable explanation that clarifies this connection could be that very long distances between home and work could be representative of rural voters dissatisfied with their current economic situation; Also, it turns out that a significant majority of conservative talk radio hosts are apparently unabashed champions of Trump’s candidacy and the long drives to work could serve as a useful opportunity for them to influence their regular listeners accordingly.

**Cruz**

*Regression Equation: Cruz Voting Share Outcome=(4.14E-03) \* Persons 65 years and over Percent + 2.09E-03 \* High school graduate or higher Percent + 1.35E-01 \* Persons per household + 2.45E-03 \* Homeownership rate + (4.43E-03) \* Mean travel time to work + (3.29E-03) \* Hispanic-owned firms, percent + (2.99E-06) \* Median household income + (3.83E-05) \* Population per square mile + (3.86E-03) \* Foreign born person percent + 4.27E-03 \* Language other than English spoken at home percent + (3.96E-03) \* Persons below poverty level percent*

The Cruz model highlights that higher levels of education, homeownership, and persons per household in a county have positive connections with the vote share outcome. Conversely, greater percentage of 65 years and over people, and persons below poverty level have the opposite effect on Cruz’s vote share. A negative coefficient for the ‘population per square mile’ dependent variable in the model demonstrates that Cruz is predicted to perform well in rural counties with a low population density. In addition, the percentage of people who speak in a language other than English at home variable has a positive association with vote share outcomes. Interestingly, the mean travel time to work variable has a negative correlation with Cruz’s vote share.

**Kasich**

*Regression Equation: Kasich Voting Share Outcome= (2.91E-03) \* Persons 65 years and over Percent + 4.09E-03 \* High school graduate or higher Percent + 8.89E-04 \* Bachelor's degree or higher, percent + (3.13E-02) \* Persons per household + 7.43E-04 \* Women-owned firm percent + 1.92E-03 \* Living in same house 1 year and over percent + 2.93E-05 \* Population per square mile + (3.63E-03) \* Population percent change*

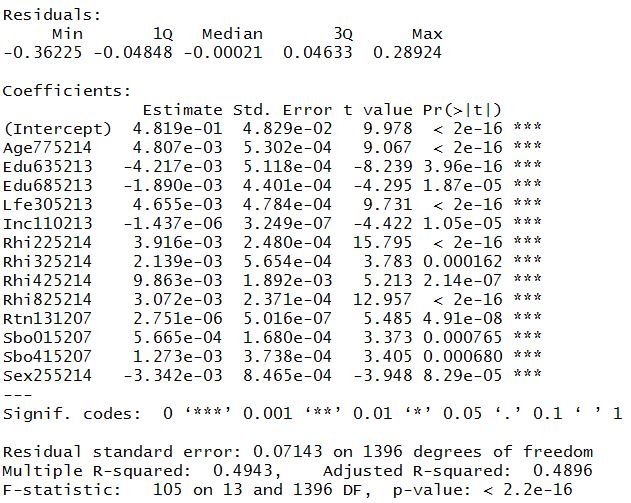
The Kasich model shows that higher education levels and percentage of women owned firms in a county have positive relationships with vote share outcomes. Also, the percentage of people living in the same house for over a year variable, which could be viewed as a proxy for stability has a positive association with Kasich’s vote share. Similarly, the model reveals that Kasich can expect better vote share outcomes within urban counties driven by the population density dependent variable. In contrast, counties with a higher percentage of 65 years and over persons are less likely to vote for Kasich.

**California Primary Predictions**

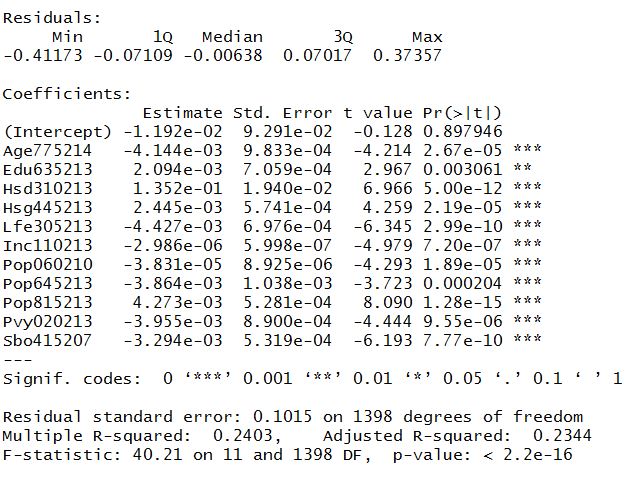
The California county data was applied to the final linear regression models to predict the vote share outcomes for Trump, Cruz, and Kasich. The results were compared with Real Clear Politics (popular professional polling data aggregator) data as of May 2nd 2016. It must be noted that 8% of voters are reported to be undecided in the Real Clear Politics aggregated poll.

|  |  |  |
| --- | --- | --- |
| **Candidate** | **Model Prediction** | **Real Clear Politics** |
| Donald Trump | 50.4% | 50.7% |
| Ted Cruz | 32.5% | 24.3% |
| John Kasich | 17.1% | 17.3% |

### Trump Model



### Cruz Model



### Kasich Model

