Introduction

Community organizers tend to have minimal resources. This scarcity limits the scope of their activities to areas that are known to have some positive relationship with whatever goal the community organizers are trying to achieve. While this allows for relatively consistent results, opportunities to increase market share are missed. Having a measure that can direct the effort of these community organizers would help in allocating their resources into new regions and potentially new supporters. In order for this measure to be effective the data used must be both relevant to the measured variable and publicly available.

The relationship between votes for the democratic senatorial candidate cast in the 2018 mid-term elections by polling locations for Hillsborough County, Florida and the demographic characteristics of the communities that surround these polling locations is the focus of this project. There is publicly available data that measures both voting behavior and demographic characteristics of a geographic region. The problem is that there is no logical relationship between this voting behavior and the geographic regions that hold the relevant demographic data. In order to create a logical relationship, the voting behavior of a precinct (the lowest grain level available) is associated with a polling location. Each polling location has an address and can be plotted in a geographic space. If demographic data can be identified with a geographic feature (here a polygon) and voting behavior can be identified with an address (here a point feature), then using a geographic information system (GIS) a logical relationship between these two features can be created according to spatial features.

This project is conducted in two parts. The first part develops an automated and scalable process that creates a logical relationship between voting behavior and the demographic characteristics of an area. The second part of this project is the statistical analysis of the produced dataset to identify relevant variables in predicting the number of votes for the democratic candidate in a senatorial race and develop a measure than can be used to evaluate communities for engagement based on demographic characteristics.

Data

The following tables and geodatabase were used to develop a dataset for statistical analysis. These tables constitute the input data for the process model. They are all publicly available.

Hillsborough County Tiger Shape files 2018

* **HilBlock (n=21850):** All block shape files for Hillsborough County
* **HilBlockGroup(n=877):** All block group shape files for Hillsborough County
* **Tlgdb\_2018\_1\_12\_fl.gdb**: Geodatabase of all shapefiles for Florida as of 2018

<https://www.census.gov/geo/maps-data/data/tiger-geodatabases.html>

**Fact**: Household income facts for block groups in Hillsborough County as of 2016. <https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>

Table 1. Fact

|  |  |  |
| --- | --- | --- |
| ID | Description | Type |
| GeoId | The unique identification code for each block group | String |
| HH | Households | Integer |
| HHMeanInc | Mean Income for Households | Integer |
| HHL50K | Households whose annual income is less than $50,000 | Integer |
| HH50100K | Households whose annual income is between $50,000 and $100,000 | Integer |
| F | Families | Integer |
| FMeanInc | Mean Income for Families | Integer |
| FL50K | Families whose annual income is less than $50,000 | Integer |
| F50100K | Families whose annual income is between $50,000 and $100,000 | Integer |
| NFMeanInc | Mean income for non-families | Integer |
| HHRent | Renters | Integer |
| HHMedRent | Median rent cost as a % of renter income | Decimal |
| OOMedP | Median home owner costs as a % of home owner income | Decimal |
| HVMean | Mean home value | Integer |

**Polls\_Geocoded (n=256)**

2018 election results by precinct in Hillsborough County.

<https://dos.myflorida.com/elections/data-statistics/elections-data/precinct-level-election-results/>

Table 2. Polls\_Geocoded

|  |  |  |
| --- | --- | --- |
| ID | Description | Type |
| Address (PK) | Polling location | String |
| City | City | String |
| Zipcode | Zip code | String |
| County | County | String |
| State | State | String |
| RegVoter | Registered voters at polling location | Integer |
| DEM (Dep Variable) | Votes for the democratic candidate for senate 2018 | Integer |
| REP | Votes for the republican candidate for senate 2018 | Integer |
| WRI | Write-in vote | Integer |
| OVER | Vote rejected due to Over vote | Integer |
| UNDER | Vote rejected due to Under vote | Integer |
| DemReg | Registered voter - democratic party | Integer |
| RepReg | Registered voter - republican party | Integer |
| NPAReg | Registered voter - no party affiliation | Integer |

**PollFact (n=256)**

PollFact is the output of the process model. Its attributes are a join of the Fact and Polls\_Geocoded tables.

Table 3. PollFact

|  |  |  |
| --- | --- | --- |
| ID | Description | Type |
| Address (PK) | Polling location | String |
| DEM (Dep Variable) | Votes for the democratic candidate for senate 2018 | Decimal |
| RegVoter | Registered voters at polling location | Integer |
| HH | Households | Integer |
| HHMeanInc | Household mean income | Integer |
| HHL50K | Households whose annual income is less than $50,000 | Decimal |
| HH50100K | Households whose annual income is between $50,000 and $100,000 | Decimal |
| F | Families | Decimal |
| FMeanInc | Family mean income | Integer |
| FL50K | Families whose annual income is less than $50,000 | Decimal |
| F50100K | Families whose annual income is between $50,000 and $100,000 | Decimal |
| NFMeanInc | Mean income for non-families | Integer |
| HHRent | Renters | Decimal |
| HHMedRent | Median rent cost as a % of renter income | Decimal |
| OOMedP | Median home owner costs as a % of home owner income | Decimal |
| HVMean | Mean home value | Integer |
| SqMiles | Square Miles | Decimal |

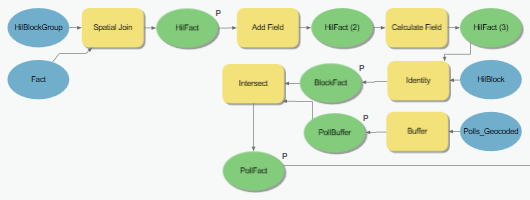
Methods and Results

**Process model for PollFact development**

The process model used to develop PollFact operates in three phases:

Phase 1: Distribute values across differing polygons (Figure 1).

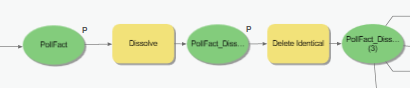
1. A spatial join is used to join the Fact table with HilBlockGroup feature; HilFact.
2. HilFact is overlaid HilBlock and an Identity function is used to attribute the HilFact values for each block group to the block features; BlockFact.
3. A 2-mile buffer is created around Polls\_Geocoded; PollBuffer.
4. An Intersect function is used to attribute PollBuffer attributes to BlockFact features that intersect with PollBuffer.



**Figure 1.** Model process used to create logical relationships between Fact and Polls\_Geocoded.

Phase 2: Aggregate observations based on poll address (Figure 2).

A Dissolve function is used to aggregate all observations into unique observations of poll locations; PollFact\_Dissolve. This will have an observation count equal to the Polls\_Geocoded observation count. All duplicate values that arise during Dissolve are removed with the Delete Identical function.



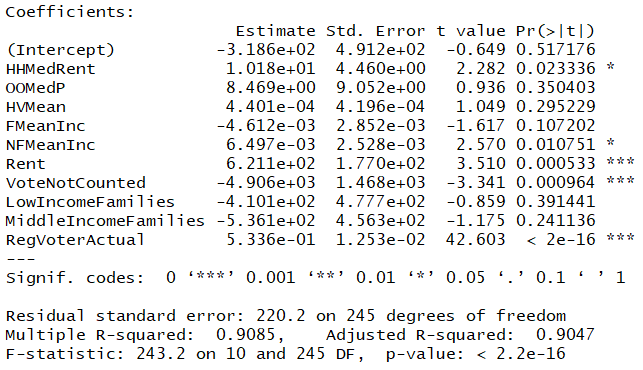
**Figure 2.** Aggregation down to unique observations.

Phase 3: Conduct initial exploratory analysis and determine spatial autocorrelation amongst observations (Figure 3).

PollFact\_Dissolve is written to a table in csv format for analysis in R. PollFact is then processed using the Spatial Autocorrelation function to determine whether PollFact\_Dissolve has to adjust for spatial autocorrelation. The results of the Spatial Autocorrelation function are then passed to the Exploratory Regression function for a general overview of relationships within the dataset.

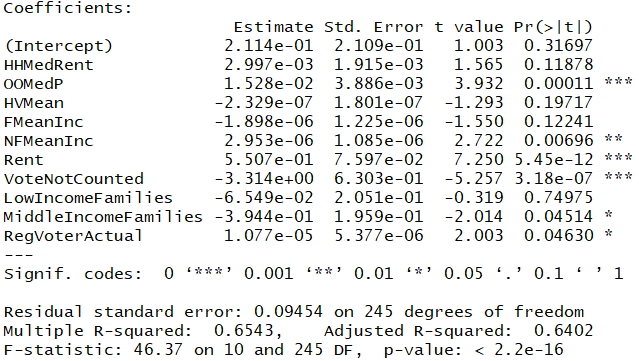
**Analysis of PollFact**

The Global Moran index suggests that spatial autocorrelation exists in PollFact. In order to mitigate this, variables are created that represent the interaction of two or more variables. These variables are of two types: discrete and bounded. The discrete variables are Median Rent Cost as % of Income, Median Home Owner Cost as % of Income, Mean Home Value, Family Mean Income, Non-Family Mean Income and Registered Voters.



**Figure 3.** Original dataset model on dependent variable DEM

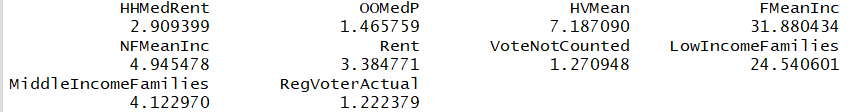
Most of the t-values are pretty small so it’s possible that some of these variables have a mean of zero and therefore not relevant in the model. The r-squared value shows a highly linear relationship between the dependent variable and all model predictors. The low p-values suggest that some of the variables have a linear relationship with the dependent variable. The high F-statistic suggests that the model beta coefficients are not zero.



**Figure 4**. Original dataset model on dependent variable DemocraticVoteRate

Some of the t-values improve with the new dependent variable. The r-squared value is significantly lower now, however, the p-values have decreased for most variables which suggests a higher likelihood of a linear relationship among variables in the model. The F-value has significantly decreased which suggests that one or more variables have a beta coefficient of zero.

The model with DemocraticVoteRate appears to have improved some with these derived variables, but not significantly. The first model had some significant multicollinearity between variables. The new model displays a significant decrease in the variance inflation factor from the original model, however, there appears to be significant correlation within model predictors. These variables are codified into dataset b.

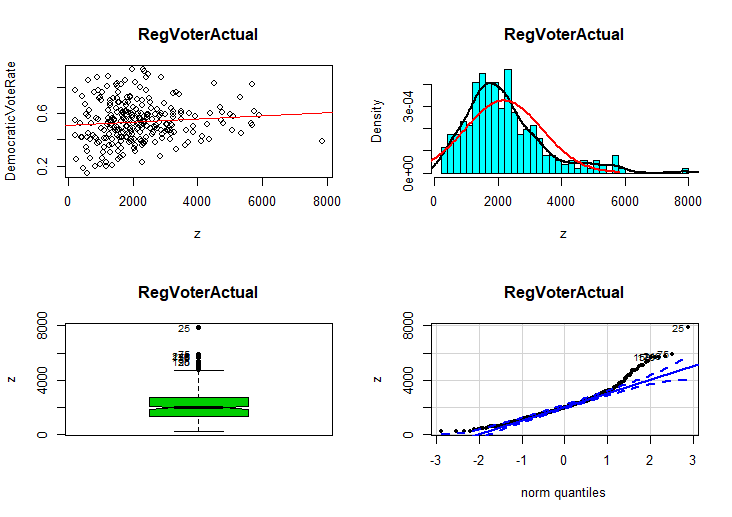


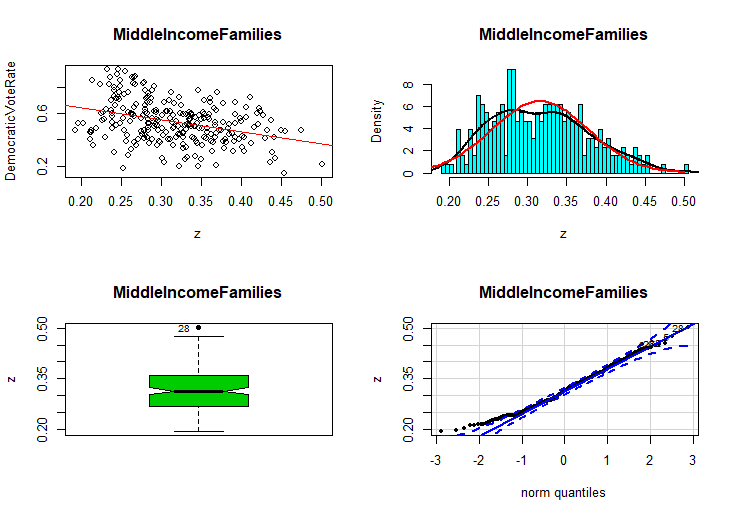
**Figure 6**. Multicollinearity amongst new model

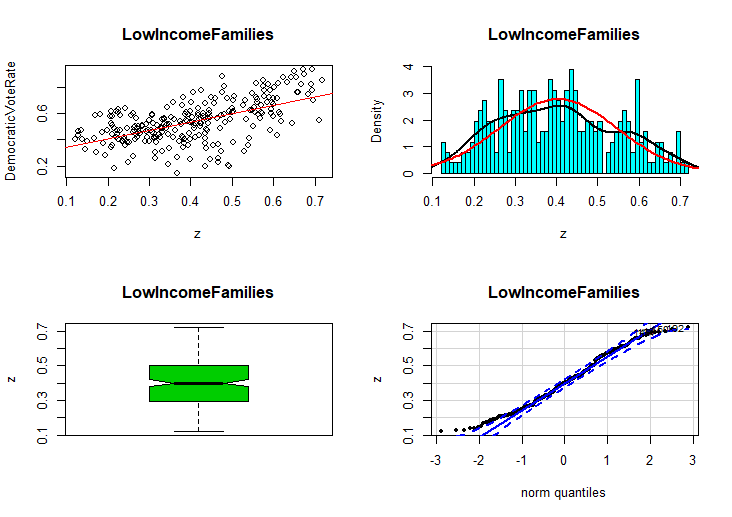
**Table 4**. Dataset b

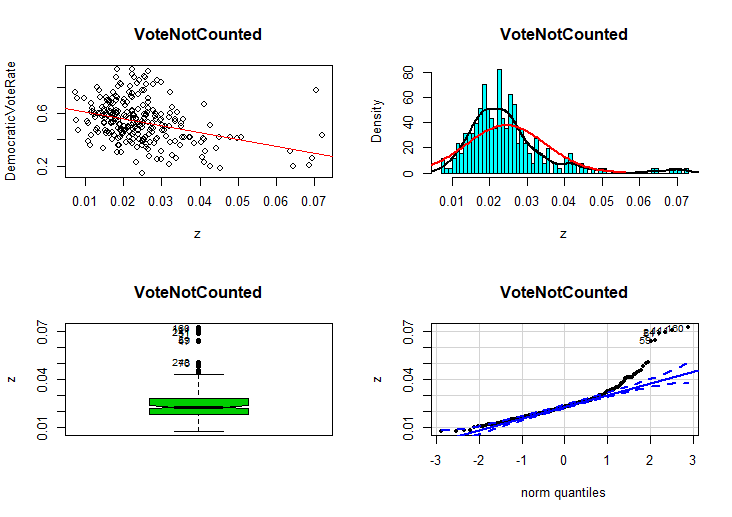
|  |  |  |
| --- | --- | --- |
| ID | Description | Type |
| HHMedRent | Median rent cost as a % of renter income | Decimal |
| OOMedP | Median home owner costs as a % of home owner income | Decimal |
| HVMean | Mean home value | Integer |
| FMeanInc | Family mean income | Integer |
| NFMeanInc | Non-family mean income | Integer |
| Rent | Renters / Households | Decimal |
| VotesNotCounted | The sum of all votes not counted divided by all votes | Decimal |
| LowIncomeFamilies | Families with annual income less than $50,000 / Households | Decimal |
| MiddleIncomeFamilies | Families with annual income between $50,000 and $100,000 / Households | Decimal |
| RegVoterActual | All registered voters at a polling location | Integer |

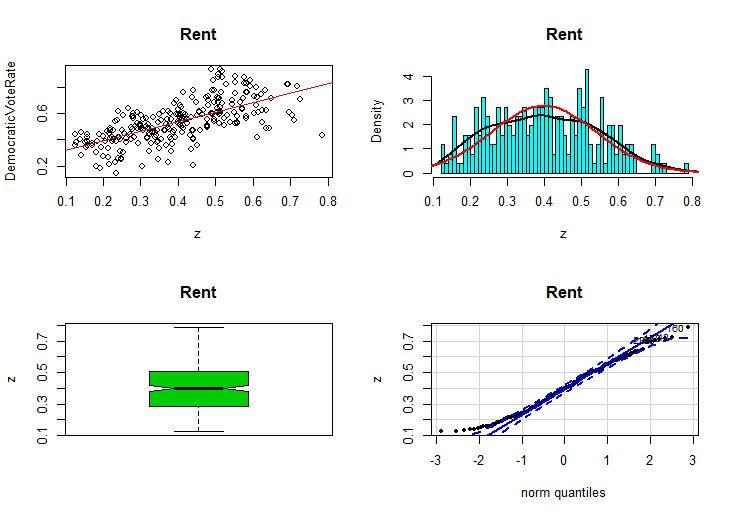
The analysis of variance (ANOVA) for model b1 suggests that the beta coefficient for six variables is not equal to zero. Instead of comparing model f-statistics across all possible combinations of models built from dataset b, the variables will be maintained and checked for normality, linearity and non-constant variance.

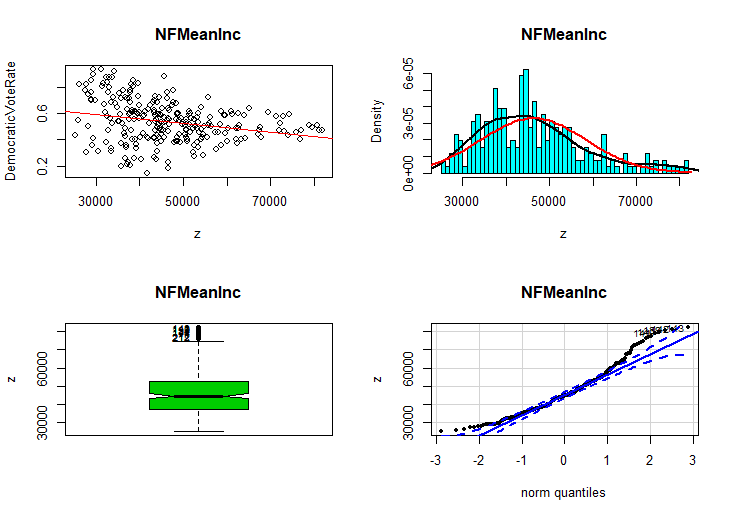


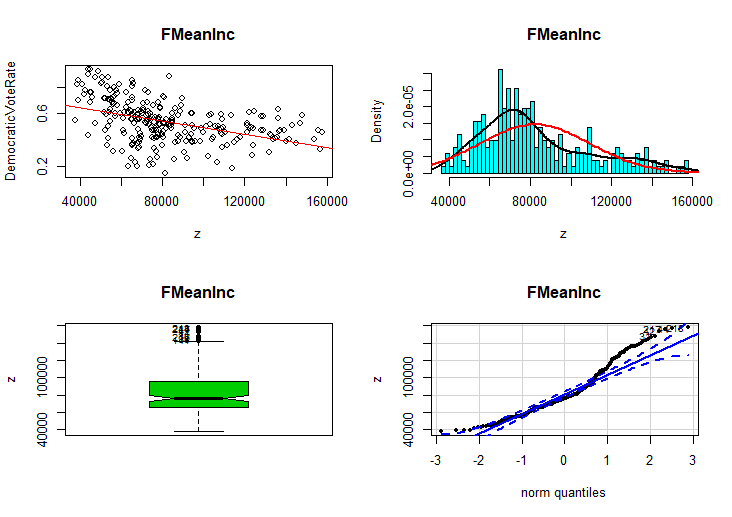


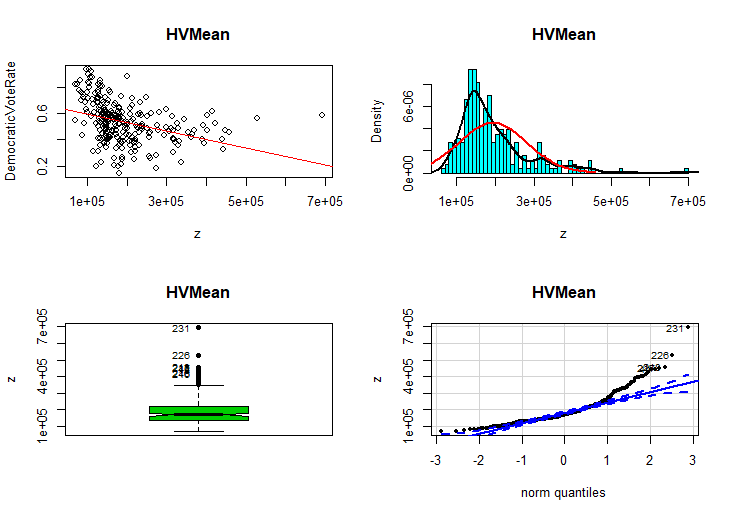


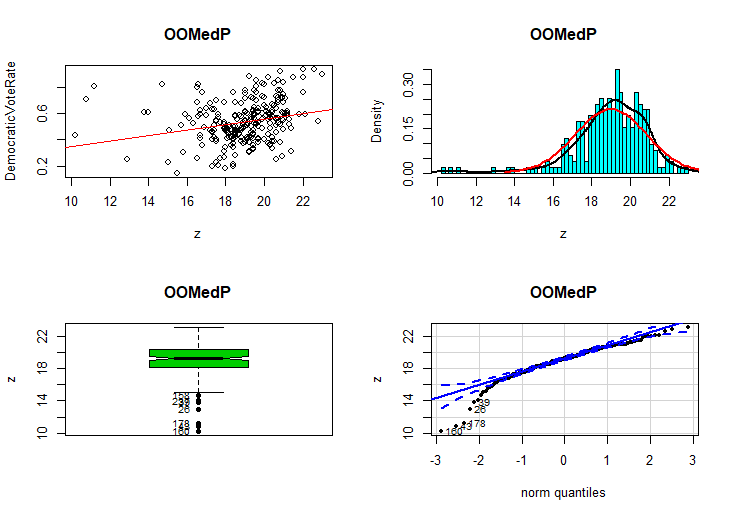


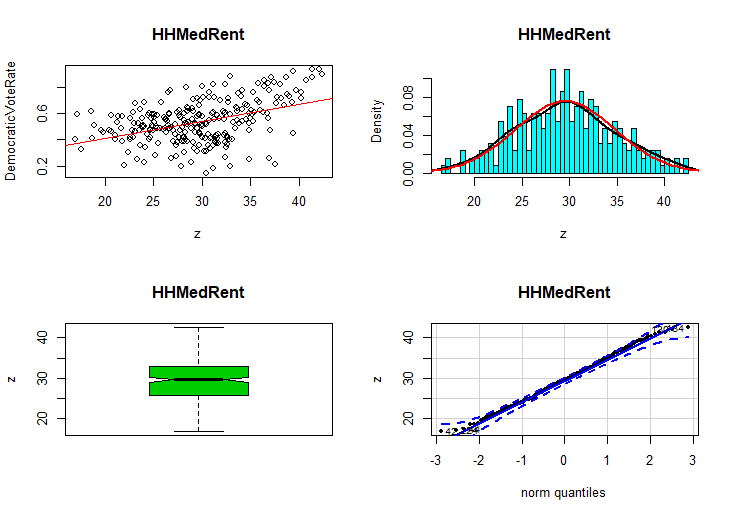










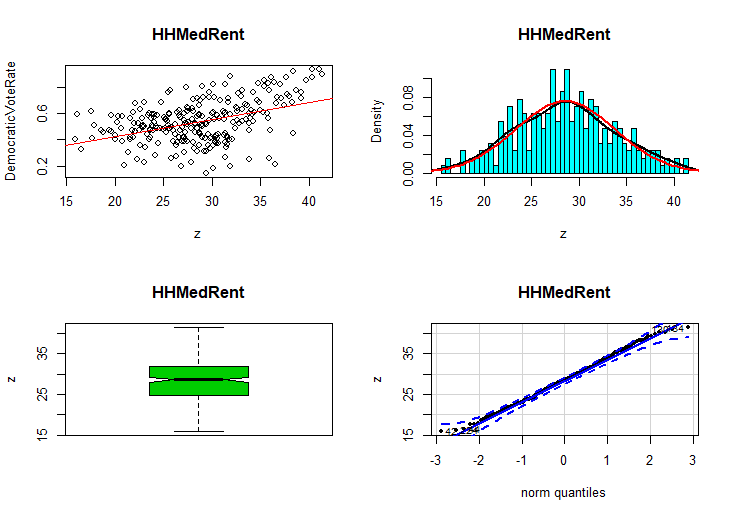


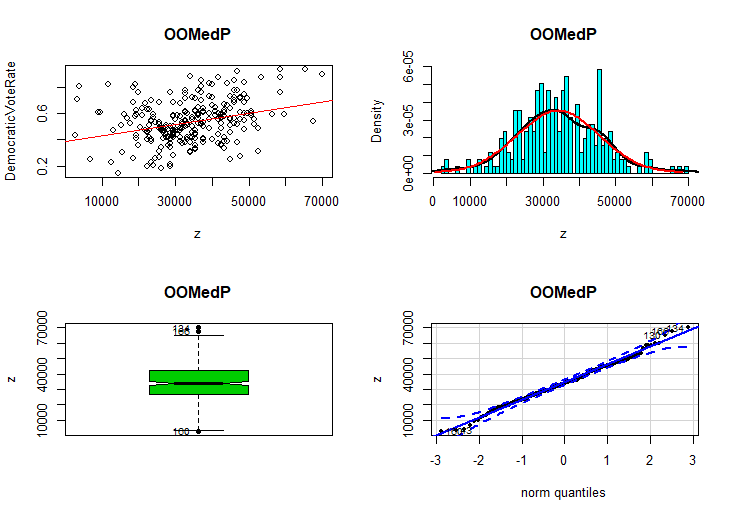
**Figure 7**. Diagnostic plots – normality

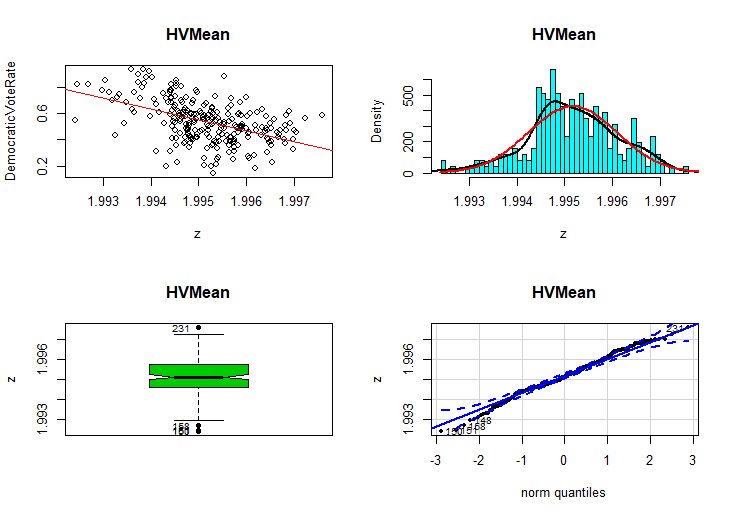
**Table 5**. Dataset c

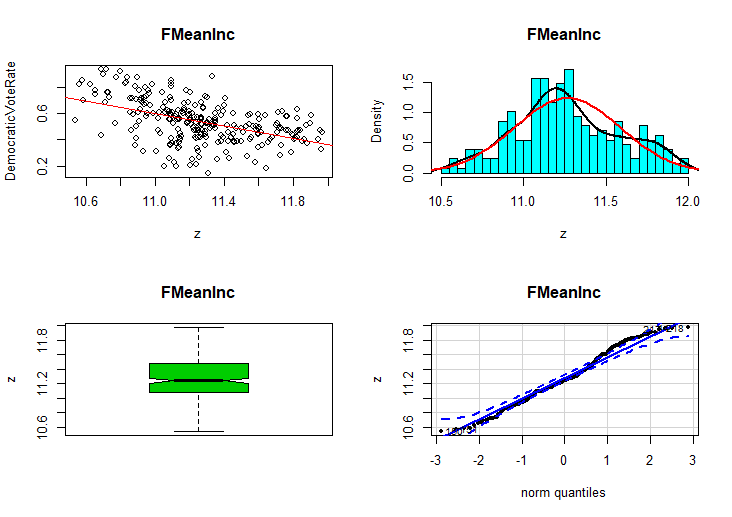


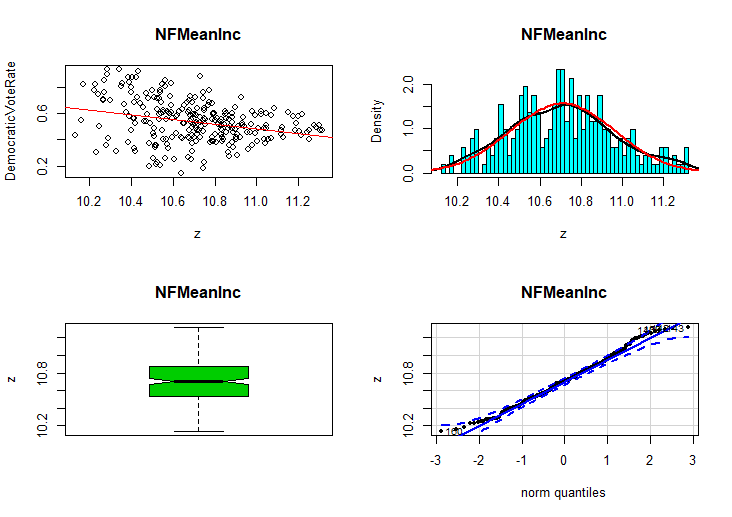
The power and logit transformations on dataset b seem to have improved the normality of the variables. Dataset c is now the working dataset. Model c1 demonstrates a slight improvement on model b1. A second ANOVA shows an improvement in most variable F-statistics.

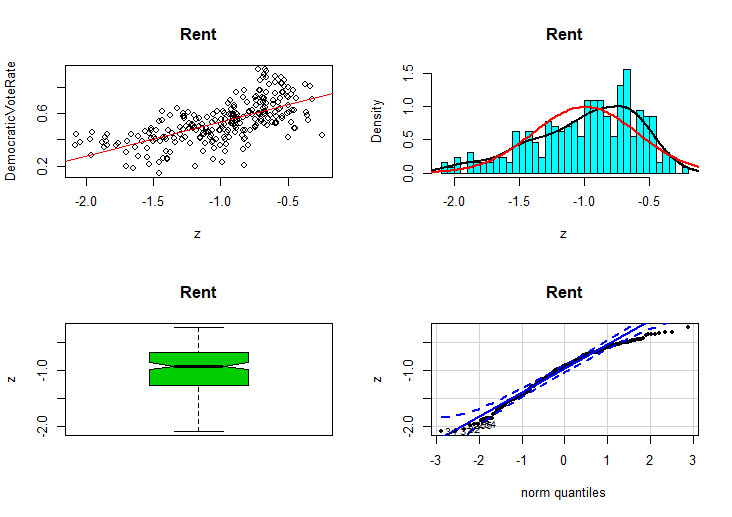


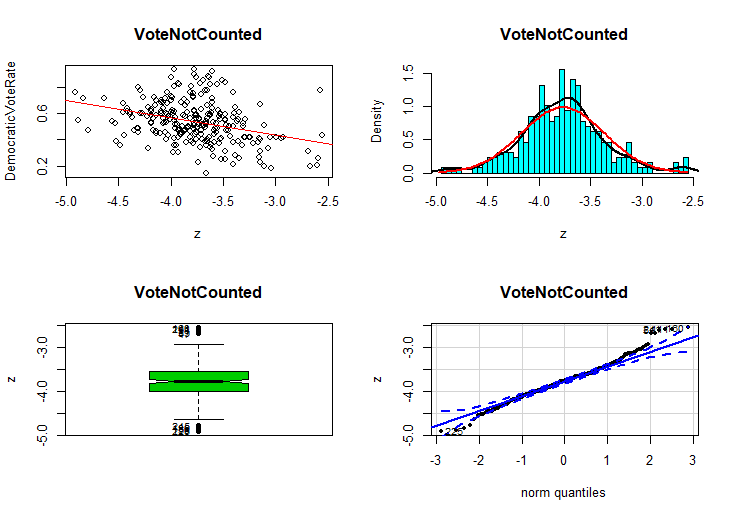


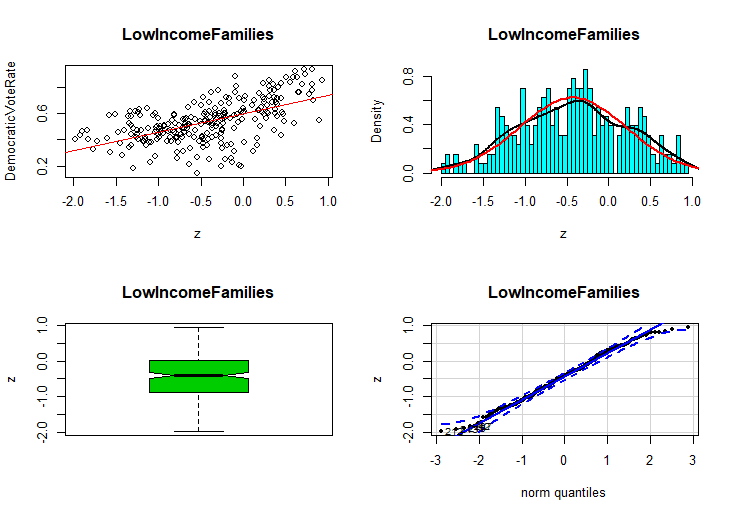


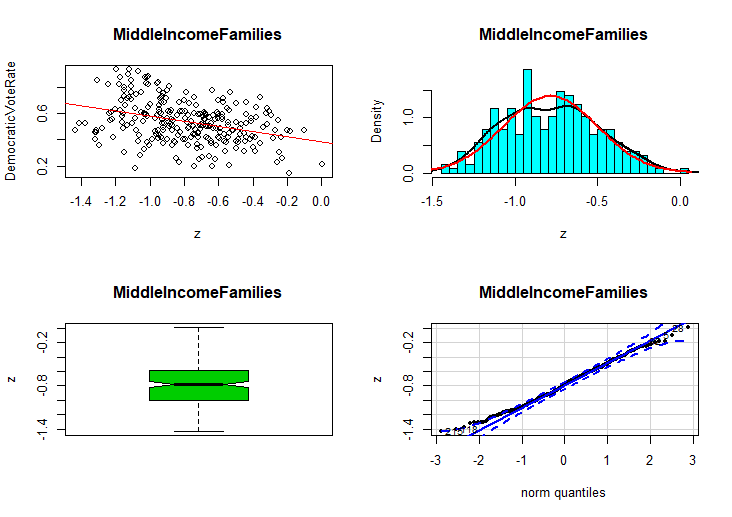


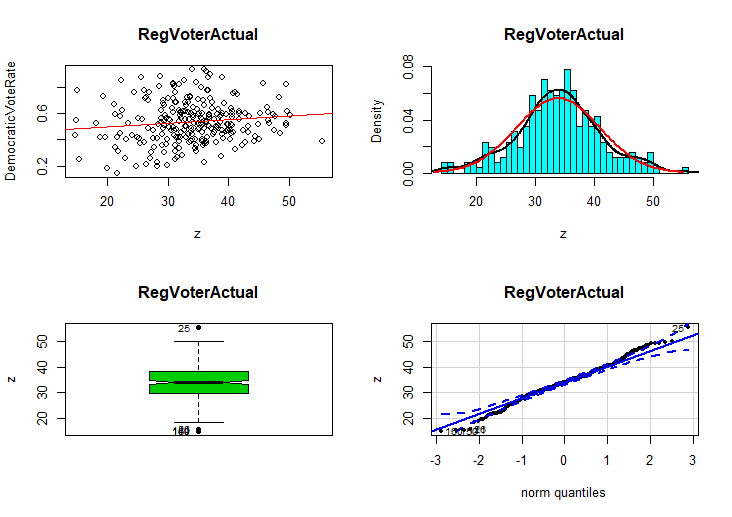






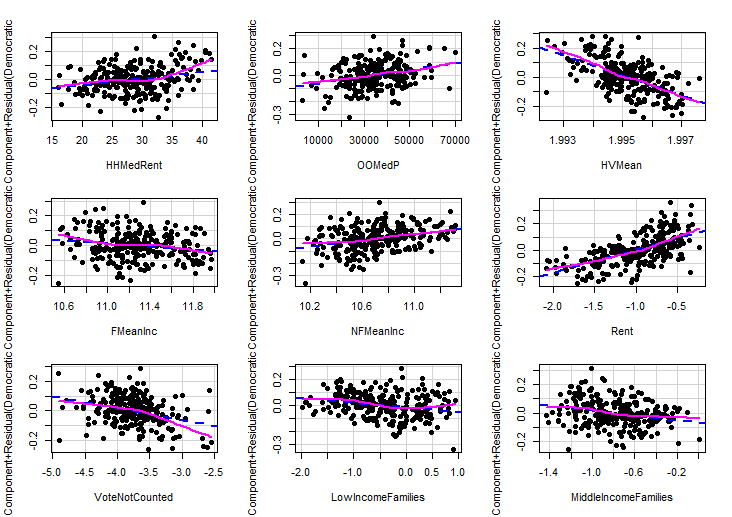






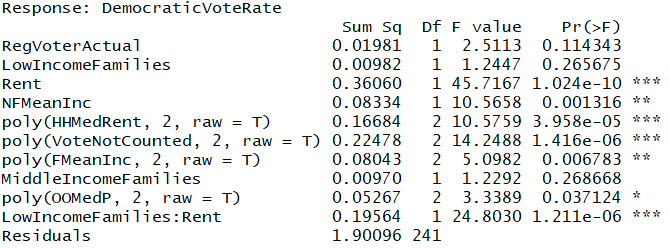
**Figure 8**. Diagnostic plots on dataset c

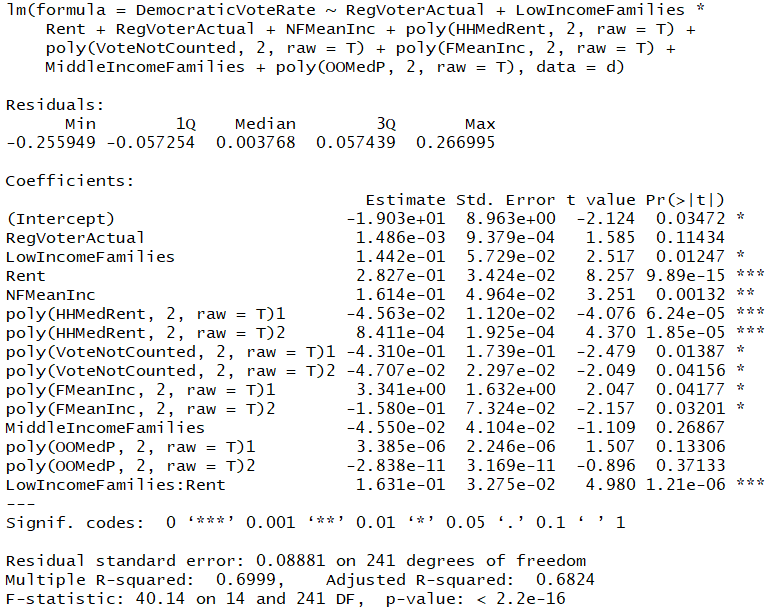
To test for non-linearity, residual plots of model c1 are created. Residual plots indicate that HHMedRent, Rent and VoteNotCounted demonstrate some non-linearity. The added-variable plots help to identify variables that may have a beta coefficient of zero. FMeanInc and LowIncomeFamilies have a slope that is visually close to zero, suggesting these variables may not be appropriate for model c1. Component + residual plots identify the following variables for a polynomial relationship: HHMedRent, FMeanInc, VoteNotCounted and LowIncomeFamilies.



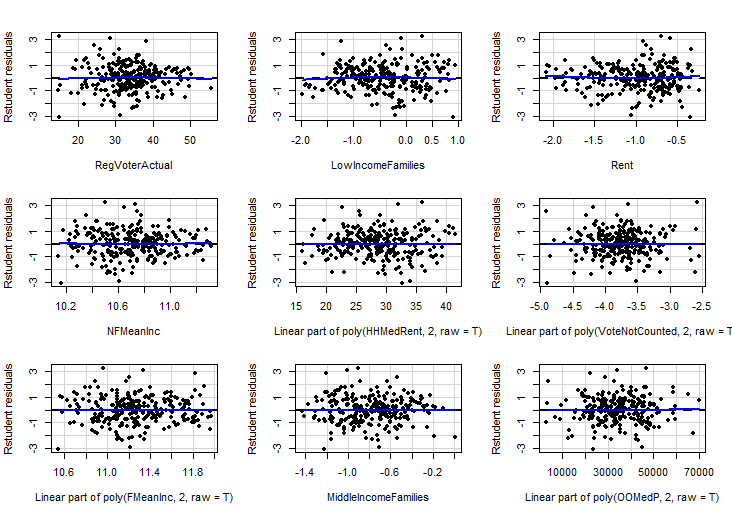
**Figure 9**. Component + Residual plots for model c1

Polynomials are introduced into our model d1. This transformation of variables mitigates non-linearity. Not much improvement between d1 and c1 in r-squared and some t-values have decreased. Model d1’s F-statistic is still pretty low so each variable’s F-statistics is compared with the dependent variable.



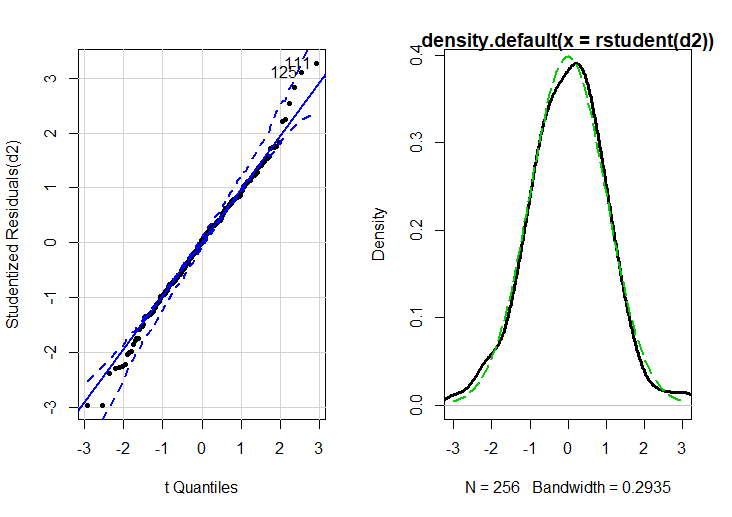


**Figure 10.** Model d1 and Anova(d1)



**Figure11**. Residual plots for model d1

Now that normality is approximately established, and non-linearity has been accounted for, non-constant variance of variable must be checked. There does not appear to be any heteroscedasticity within the overall model, however the variables RegVoterActual, NFMeanInc and the polynomial of VoteNotCounted.

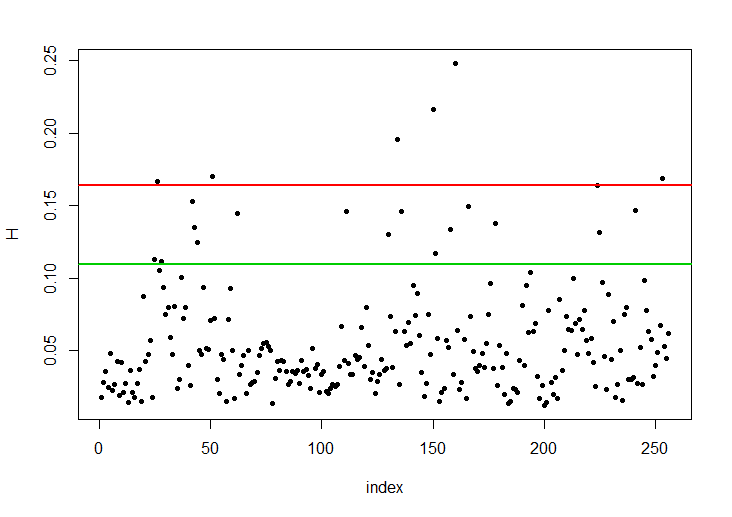


**Figure12.** Model e1 check for non-constant variance

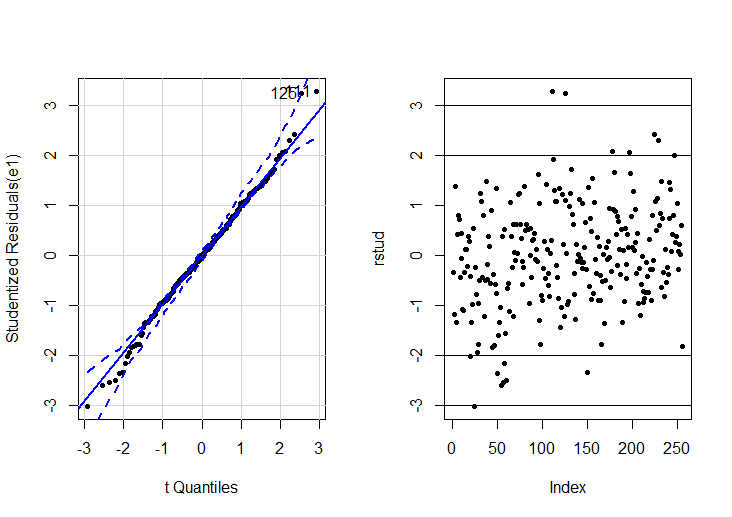
**Table 6.** Dataset e: Formalizes the polynomial transformation of selected variables



The last thing to do is to check for any observations that are having an outsized effect on model e1 regression. An index of the hatvalues for model e1 are plotted with a red line representing three times the model means and a green for two times the model means. The observations above the red line should be considered as potential outliers.



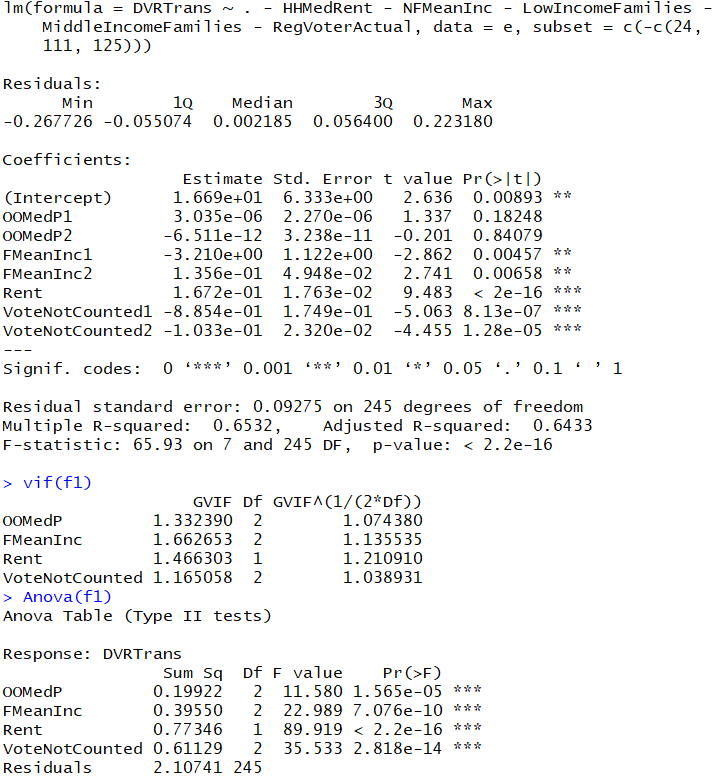
**Figure 13**. Model e1 index of hatvalues and model means



**Figure 14.** Model e1 extreme values.

Utilizing an outlier test, model adjustment and model comparison process is utilized to remove three observations whose rstudent value was not between -3 and 3. Dataset f is created that holds dataset e minus the three extreme observations. Finally, a best subset algorithm is used to identify the best variables for model f1.

**Results and Conclusion**



**Figure 15**. Final model and results

There appears to be a relationship between model f1’s variables and the percent of the vote a democratic candidate receives. At its most basic level, model f1 represents the types of neighborhoods where people are likely to vote for a democrat. Based on three financial demographic characteristics and a measure of votes not counted in the 2018 election, a prediction of the democratic vote rate can be inferred with approximately 65% accuracy and a 9% standard error. This suggests that economic characteristics are effective measures of gauging voting behavior. Consider rent. As a measure, rent is the percent of households who rent. The inverse of this is home ownership rate. As the percent of households who rent increases so does the democratic vote rate. As a family’s income decreases they become more likely to vote for the democratic candidate. As home ownership costs absorb a greater portion of income the democratic vote rate tends to increase. Collectively these variables describe a local economy. As that economy moves away from affordable home ownership, the costs associated with living in an area increase and the potential income available for local transactions decreases. The people who live in these neighborhoods are likely to vote for the democratic candidate.

There is another interesting relationship within model f1. The percent of votes not counted due to write-ins, over-counts and under-counts has a strong relationship with the democratic vote rate. As the percent of votes not counted decreases the democratic vote rate increases. There are 3 reasons that this relationship could exist:

1. The rate of the votes not counted are part of the total vote count. Inherently a relationship exists within the total vote count between its components . If true, then the following should also be true:
   1. The beta coefficients should move in the same direction whether the dependent variable is the democratic or republican vote rate.
   2. As the votes not counted rate increases, both the democratic and republican vote rates should decrease at comparable rates.
2. Democratic voters in economically depressed areas tend to make an error on their ballots (Over/Under) and/or choose to write-in candidates.
3. The rate of the votes not counted and the democratic vote rate has an artificial relationship where an increase in votes not counted has an approximately proportionate response with a decrease in the democratic vote rate.

Further research is required to confirm whether the first condition is true. If not, then why?

Registered voter data was not used in this model but was collected and analyzed concurrently with this project. Interestingly, the registered voter count published with the polling location election results are significantly higher than the official registration records at the cutoff 30 days out from the election. This could be an effect of a massive voter registration effort prior to the cutoff date and a lag time in processing these prior to publishing the official voter rolls. This could be an effect of voters requesting an exception on disclosing any identifiable information, including the precinct that they are associated with, thereby increasing the aggregate voter registration count without adjusting any variable(s) to represent that change. Initial counts suggest that this difference is not significant enough to solely account for the increase in registered voters between my count from the cutoff date and the official count for that polling location.