Analysis of the 1992 Presidential Election and Applications

Brianna Eskin

12/8/2020

I: Introduction

Every four years on the first Tuesday in November, the United States participates in the election of a new President. The years and months leading up to the election are filled with potential candidates working to persuade voters that they are the best choice for the role, and media outlets focus a lot of time and money into predicting how voters will cast their votes. This is usually done via polling voters to get an idea of how they plan on voting. While pre-election pollsters were generally trusted among the American people in the past, their validity was called into question after the 2016 election victory of Donald Trump, which was considered a major upset, with Hillary Clinton predicted to win in many of the polls, some as confident as 99%. Many pollsters were forced to revisit their polling practices and models to try and understand what went wrong. According to Pew Research, inaccuracies in the polling appeared to stem from **a)** the demographics of those who chose to respond to the polling surveys in the first place, with many being college educated voters who tend to vote Democrat, **b)** the quality of the surveys when compared to the sample size, **c)** inaccurate predictions of where the "Undecided" voters would end up voting, a section of voters which broke heavily in 2016 for Donald Trump (aka the "shy Trump voter"), and **d)** the true margin of error in polls being grossly understated, with the real margin of error predicted to be closer to 6 points instead of the standard 3 points[1]. Despite several improvements made to polling techniques ahead of the 2020 election, there were still many states where the polling data showed very different results from the actual election. The general pattern was pretty much the same, with the polls underestimating support for Donald Trump. Take for example the key battleground states of Wisconsin and Michigan, which despite correctly predicting that Biden would win the state, overstimated his support by 7 and 5 points repesctively. Perhaps the largest culprit, Florida was predicted to be a Joe Biden win by 3 points, but instead he lost the election by 3 points[2].

This paper will look into a potential election prediction model that does not take personal response into account. Instead, we will look at some key demographics at the county level across the United States to see if we can accurately predict how people will vote purely based on the makeup of the population. We will be able to confirm if certain stereotypes about populations hold, such as the voting tendencies of older Americans to lean towards Republican or more densely populated areas to vote Democrat. If this general model is successful, we can apply it to future elections without the unpredictable biasness and variability that comes with relying on human response, and restore the public's faith in pre-election and post-election results.

II: Datasets

For this analysis, we will focus on the 1992 Presidential election between incumbent George H.W. Bush and winner Bill Clinton. More specifically, we will focus on the percentage of people who voted for Bill Clinton on the county level across all counties in the United States. This data was compiled by Professor Larry Winner from the University of Florida, based on data gathered from the U.S. Census Bureau[3]. The response variable will be the percentage of people who voted for Bill Clinton in the election (*VotingPct*). There are nine predictors variables in our data: the median age of the population (*MedAge*), the average amount of money in the saving account of a person in the county (*Savings*), per capita income (*PerCapIncome*), percentage of people in the county living in poverty (*Poverty*), percentage of population that are veterans (*Veterans*), percentage of females in the population (*Female*), population denisty (per square mile) (*PopDensity*), the percentage of the population in a nursing home (*NursingHome*), and the crime index per capita (*Crime*). All percentages are assumed to be between 0-100. There are N=2704 county observations in the dataset. This data was originally compiled from a txt file and was converted to a csv for this analysis called Clinton-ElectionData1992.csv. A copy of the csv, as well as the code written to conduct this analysis, can be found at the following Github link: https://github.com/briannaeskin/RegressionAndTimeSeriesFall2020.

III: Data Analysis and Model Selection

We applied multiple linear regression techniques on our dataset to identify the minimum number of predictors required to explain the variability in the election results. Using all of the available data points, we first fit all of the predictors to a linear model with no polynomials, and calculated the VIF for each variable to check for potential collinearity between the predictors. Then, we checked the residuals to confirm normality $\epsilon \sim N(0, \sigma^2)$. If the assumption was violated, we checked for potential transformations using the Box-Cox plot. Next, we looked for potential outliers and leverage points and removed points with large leverage by using visual techniques, as well as calculation of the jackknife residuals and Cook's Distance. Once influential points were removed from the model, we repeated all prior analysis to check if the model was severly impacted by the influential points. Then, using BIC as the filtering criteria, we performed backwards model selection and selected the number of predictors that generated the minimum BIC. Once the smaller model was generated, we repeated all previous analysis as a way to confirm that any irregularities in the model to this point were not due to unecessary predictors.

Once the required predictors were identified and any necessary transformations were done, we further tested for model stability by randomly splitting the data into training and test sets using a 3:1 ratio. By randomly splitting the data, we eliminate any potential skewness that may possibly occur as a result of the original data being organized alphabetically by state and then county. We generated a new set of β coefficients using the training data against the original model structure. We then used these new coefficients to predict responses for the test data. Once done, we checked for model stability via comparisons of the Root Mean Square Error (RMSE), analysis of the plot of observed versus fitted values, and a Chi-Squared test of proportions, which can help confirm that the distribution of predictions is consistent and that the model can be used for new data.

IV: Results

Descriptive statistics for each variable can be found in Table 1. While visually observing the variables, we noticed some outliers in the predictors. We will wait to action on these until it is concluded that they are influential points during the model analysis. A histogram of the response variable, VotingPct, can be found in Figure 1, confirming an initial assumption that the response variable is normally distributed.

As seen in Table 2, the VIFs for each predictor are reasonable, implying little correlation between the predictors. The largest VIFs, for PerCapIncome and Poverty, were still less than 3 and the average VIF is around 1.69. As a result, we kept all of the predictors in the model for the time being. As seen in Figure 2, the residuals versus fitted values plot shows constant variance in the model, but the residuals are not normally distributed, with a long tail in the negative direction, and a calculated p-value from the Shapiro-Wilk normality test of 7.242e-05. When using the Box-Cox Method to find a potential transformation, the suggested transformation is $\lambda = 0.75$. However, after analyzing the residuals with lambdas in that range ($\lambda = 1, 0.75, 0.5$), there is little to no improvement in the distribution of the residuals, so no transformation was done. When checking for potential outliers and influential points, there was one county, Kings NY, whose jackknife residual of 6.84 was larger than the Bonferroni Correction value of 4.13, and whose Cook's Distance of 4.317 was extremely large compared to the rest of the counties (< 0.15). So we removed this county from the model and repeated the residual analysis with a new model fit.

The updated residual analyses after removing the influential county can be found in Figure 3. The residuals versus fitted values remained relatively unchanged, but the distribution of the residuals now resembles a normal distribution. Similarly, the new Box-Cox test still suggests a λ of 0.75, so there is no reason to transform the data. Backwards model selection using BIC gives that 5 predictors should be used in the model: Poverty, PopDensity, Female, Savings, Veterans (Figure 4). After a final sanity check of the residuals (Figure 5), the finalized model can be summarized in Table 3.

When comparing the RMSE of the train and test data, the values are comparable at 8.20 and 8.58, respectively. The plot of the observed versus fitted test values, with the corresponding prediction intervals, can be found in Figure 6. The model sensitivity test was insignificant, with $\chi^2 = 8.0346$ and p-value of 0.3295, suggesting the model is useful for new data (Table 4).

V: Conclusion

After analyzing the data, we were able to create a model that explained 33.27% of the variability in the percentage of the population of a given county that voted for Bill Clinton based on 5 key variables: the average savings per person in the county, the percentage of people living in poverty, the percentage of veterans, the percentage of females, and the population density. However, because the model explains only ~1/3 of the variability, it does lead to some predicting discrepancies, especially in the extreme values. In its current form, our model overestimates in counties that overwhelmingly didn't vote for Clinton and underestimates in counties that overwhelmingly did vote for Clinton. This model is very narrow in scope and makes large conclusions about the way a group of people would vote based on broad characteristics. To improve this model, we can consider other demographics, such as religious affiliation, education levels, and race/ethnicity. Another thing to keep in mind is that because this model estimates the voting percentage in a county, population is not taken into account. When trying to estimate the vote share across the state or country, county population needs to be multiplied by the predicted vote share (even though a county may vote 95% for Clinton, if it is the smallest county in a large state like California, it won't contribute much to the overall vote count). In the future, we can also further test the model stability and accuracy by applying the same model to more recent elections.

References

1. Kennedy, Courtney. *Key things to know about election polling in the United States.* https://www.pewresearch.org/fact-tank/2020/08/05/key-things-to-know-about-election-polling-in-the-united-states/ (August 5, 2020)

2. Silver, Nate. *The Polls Weren't Great, But That's Pretty Normal* https://fivethirtyeight.com/features/the-polls-werent-great-but-thats-pretty-normal/ (November 11, 2020)

3. Winner, Larry. http://users.stat.ufl.edu/~winner/data/clinton1.txt

Table 1: Descriptive Statistics for Predictors and Response

	Mean	Variance	Min	Max
VotingPct	39.61	103.67	9.55	84.64
MedAge	34.44	12.49	20.00	55.40
Savings	89939.37	1719112552.16	7472.00	631534.00
PerCapIncome	16389.29	11416454.63	6118.00	37387.00
Poverty	16.08	44.33	1.90	52.00
Veterans	11.42	5.32	2.78	27.29
Female	51.00	2.17	37.53	55.39
PopDensity	195.24	886812.73	0.40	32360.30
NursingHome	9.48	39.49	0.08	59.22
Crime	304.49	47133.09	0.00	2792.00

Table 2: Variance Inflation Factor (VIF) for Full Model

	VIF
MedAge	1.733274
Savings	1.610381
PerCapIncome	2.428866
Poverty	2.057863
Veterans	1.551142
Female	1.177141
PopDensity	1.304469
NursingHome	1.611870
Crime	1.527998

Table 3: Final Model Using BIC Selection

	Estimate	StdError	tvalue	pvalue
(Intercept)	-32.885	5.718	-5.751	< 0.001
Savings	0.000	0.000	-7.981	< 0.001
Poverty	0.714	0.026	27.377	< 0.001
Veterans	0.304	0.076	4.000	< 0.001
Female	1.177	0.114	10.311	< 0.001
PopDensity	0.003	0.000	13.522	< 0.001

Multiple R-Squared: 0.334, Adjusted R-Squared: 0.3327

Table 4: Model Sensitivity Analysis

Range	Observed	Expected
< 30	16	10
30-35	132	127
35-40	255	257
40-45	160	175
45-50	80	72
50-55	20	21
55-60	10	8
> 60	3	6

$\chi^2 = 8.0346, p = 0.3295$

Figure 1: Histogram of Voting % for Clinton by County − 1992

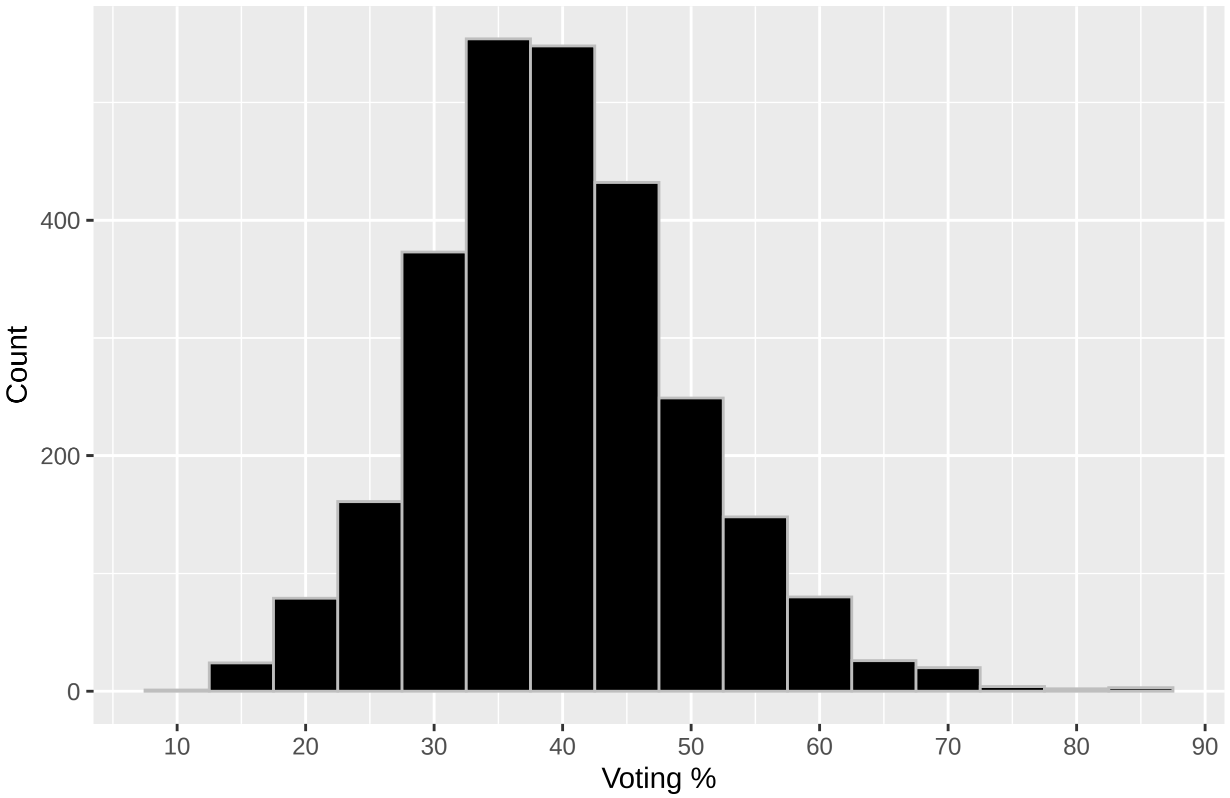


Figure 2: Residual Analysis - Full Model and Data

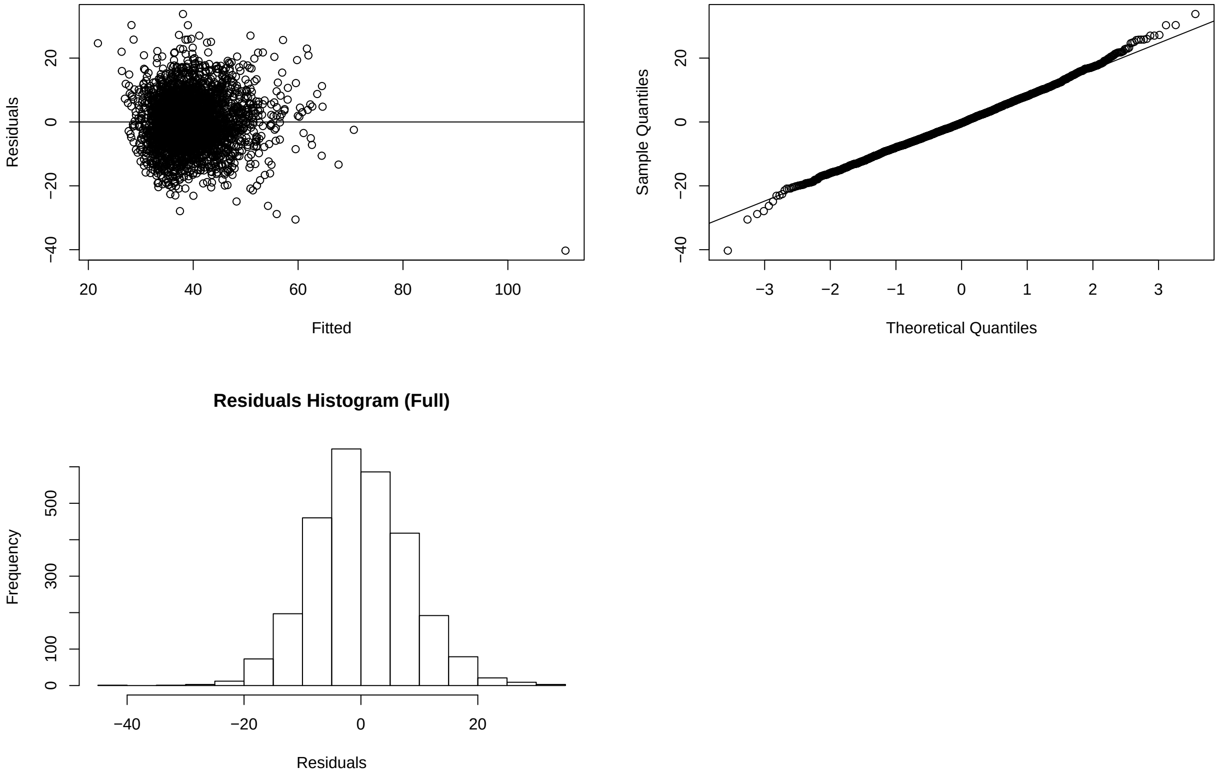

Residuals Histogram (Full)

Figure 3: Residual Analysis - Full Model and Filtered Data

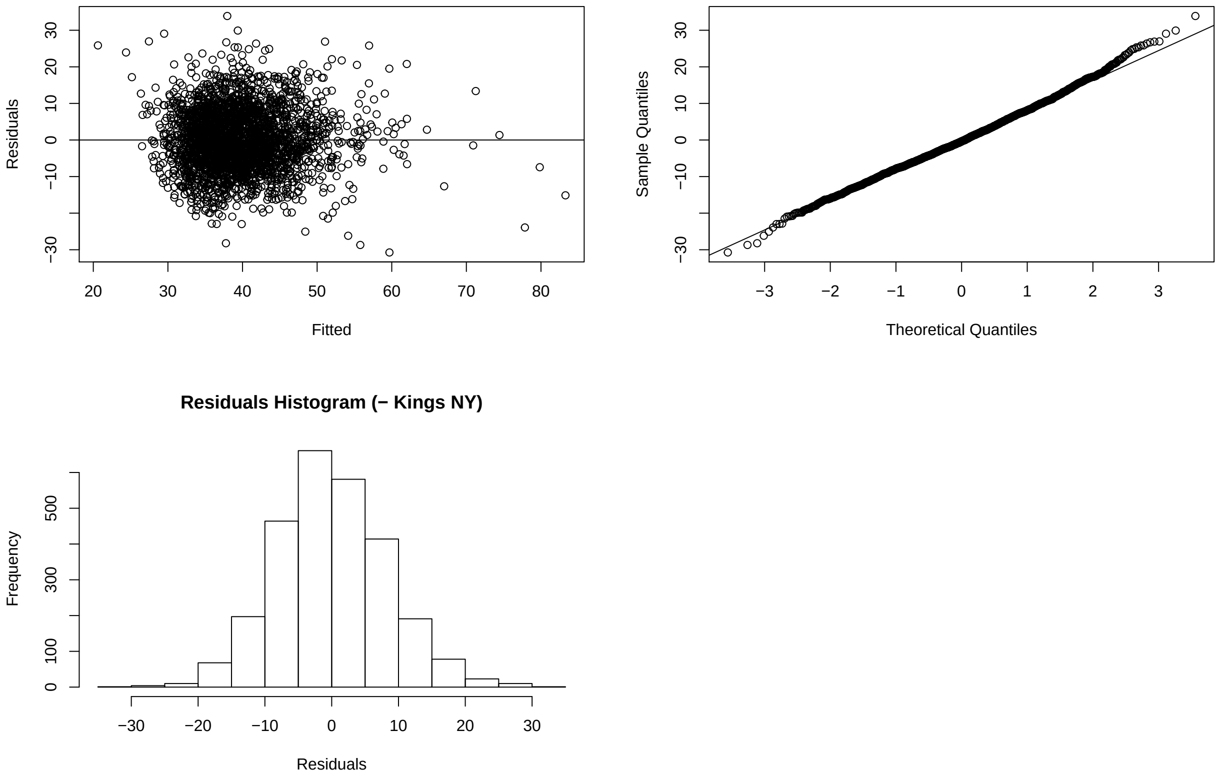


Figure 4: Backward Search− BIC

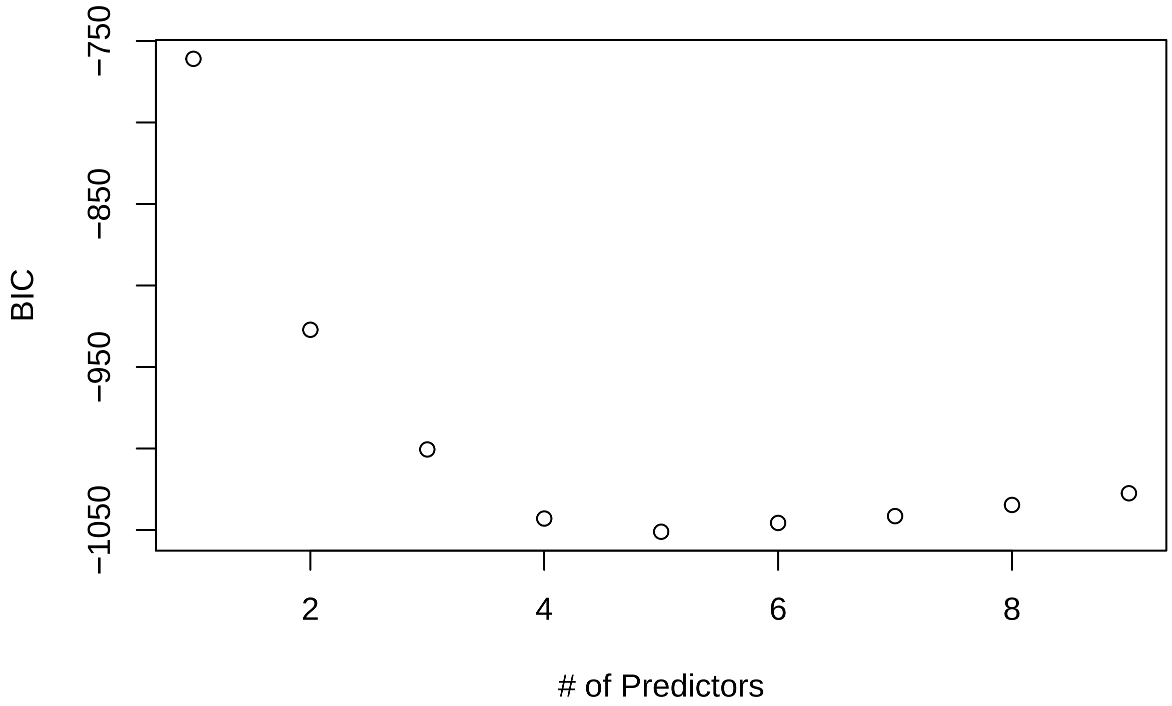
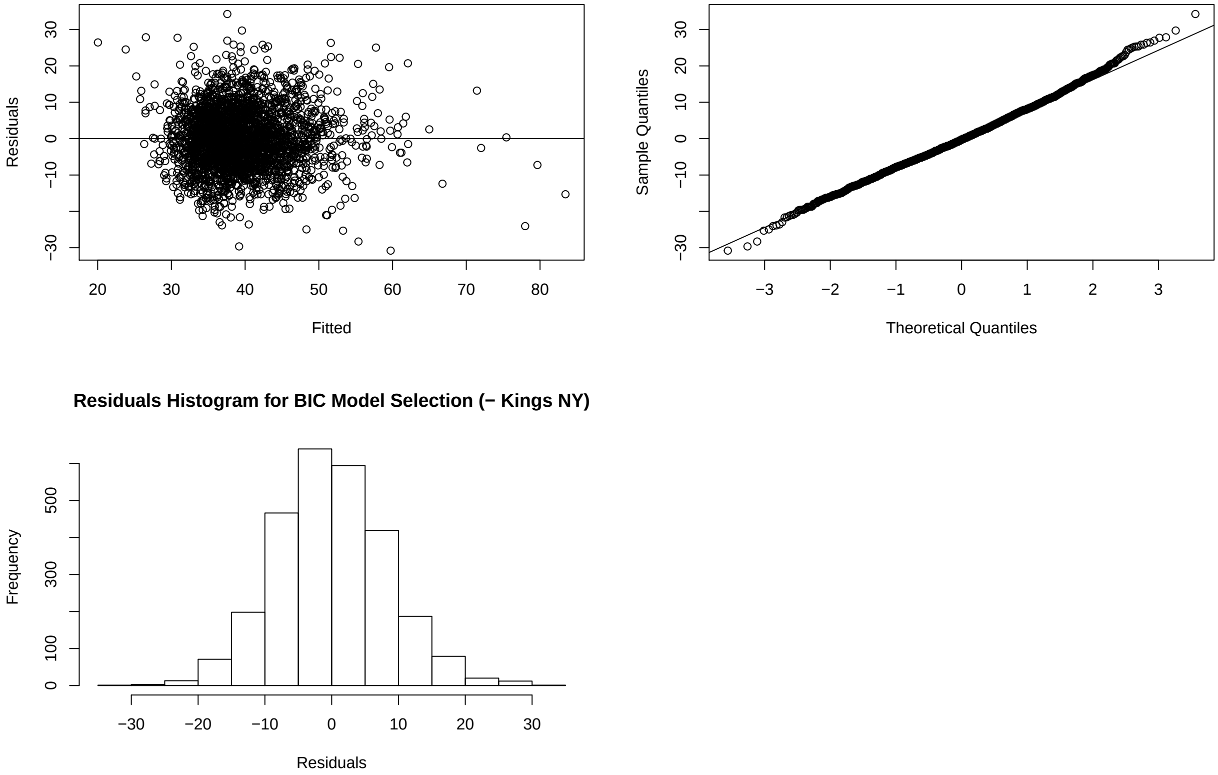
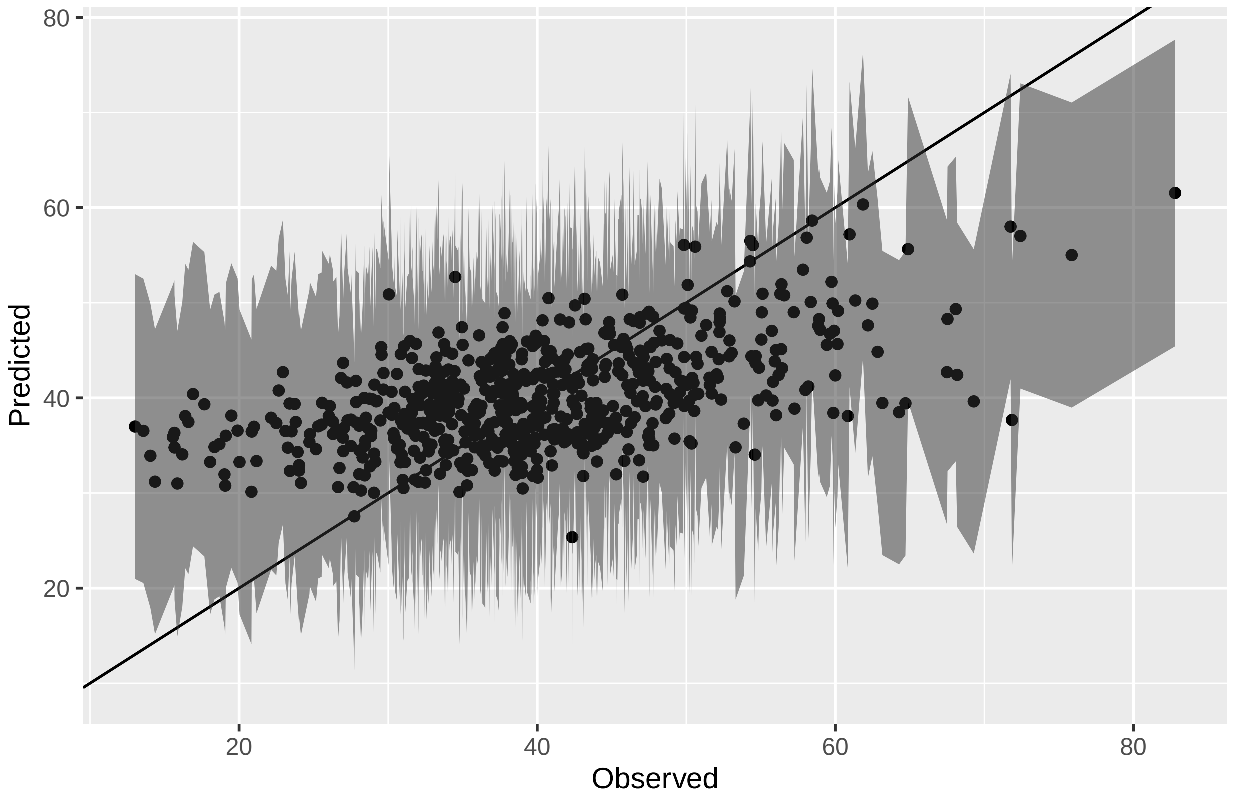


Figure 5: Residual Analysis - BIC Criteria Model and Filtered Data



Residuals Histogram for BIC Model Selection (− Kings NY)

Figure 6: Observed versus Fitted with Prediction Intervals, Test Data


Predicted
Observed