2000 Presidential Election Ballot Analysis

James "Morgan" Hawkins

3/24/2023

Introduction

In the 2000 Presidential Election, Palm Beach County had out of the ordinary election results. The proportion of votes for Buchanan that were cast on election day was far higher than what we would expect. It's hypothesized that this abnormal amount of election day votes is because the butterfly ballot caused voters to mistakenly vote for Buchanan instead of Gore. We are interested if this hypothesis is supported by relevant statistical analysis. For this analysis, we will use a data set containing vote counts from the 2000 presidential election for the 67 counties in Florida. What makes the issue of miscast votes in Palm Beach County so serious in this election is that the battle for Florida between Bush and Gore was close and a deciding factor in Bush's victory. So, we are also interested in exploring how many votes Buchanan would have received in the absence of the butterfly ballot. We will use a data set containing anonymized voter-level information from Palm Beach County to obtain an estimate for this answer.

Our analysis showed a surprisingly large difference in the amount of votes cast in person for Buchanan over the amount of absentee votes cast for Buchanan. We estimate that if the butterfly ballot had not been used, Buchanan would have received around 2500 fewer votes in Palm Beach County.

Exploratory Data Analysis

We will begin by creating variables that contain total votes, the proportion of votes cast for Buchanan in person, the Proportion of votes cast for Buchanan via absentee ballots, the proportion of votes cast for Bush, and the Proportion of votes cast for Gore. We will

also add a variable that is the difference between the proportion of in-person votes cast for Buchanan and the proportion of absentee votes cast for Buchanan.

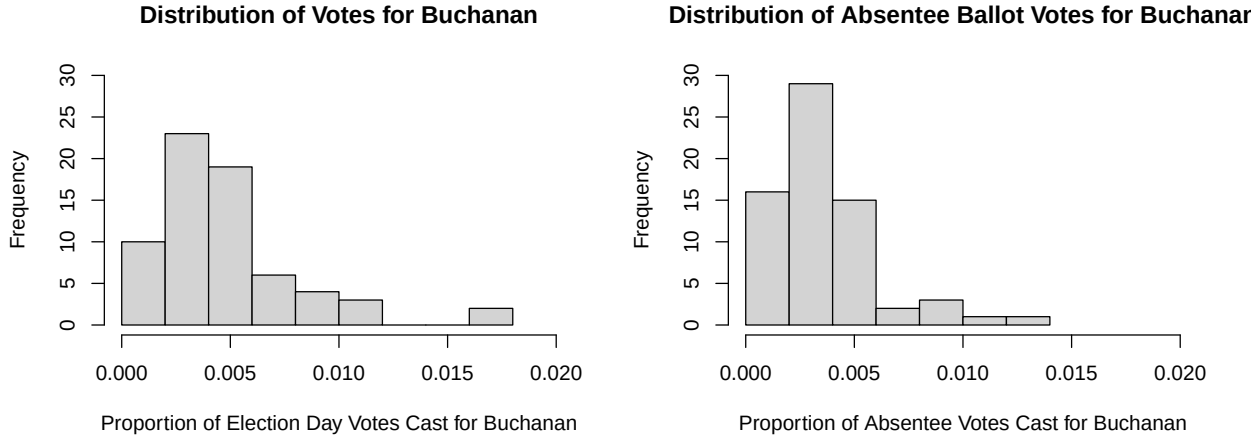

Figure 1: Distribution of Votes for Buchanan

We will begin our exploratory data analysis by looking at the distribution of votes cast for Buchanan on election day and via absentee ballots. In Figure 1 we see that both distributions appear somewhat normal, but are right skewed. We also notice that there is an outlier on the histogram showing the distribution of election day votes for Buchanan. However, this data point is not Palm Beach County it is Calhoun county. On the histogram showing the distribution of absentee ballot votes cast for Buchanan, there aren't any clear outliers, but we do notice a tail on the right side of the distribution. These two counties on the right side of the distribution are Dixie and Gulf county. We note that Palm Beach falls in the middle of the distribution with a proportion of absentee ballot votes going to Buchanan of around 0.0022. Palm Beach County had the fourth highest difference between its proportion of election day voters for Buchanan and proportion of absentee voters for Buchanan. However, it also had the 8th highest proportion of voters who voted on election day.

Next, looking at the distributions of the proportions of election day votes for Bush and Gore in Figure 2 we see that both appear somewhat normal with no outliers. Both distributions have similar standard deviations of .092 and .093 for Bush and Gore respectively.

Our Response variable that we are interested in will be the difference between the proportion of election day votes cast for Buchanan and the proportion of absentee votes cast for Buchanan. In Figure 3 we see that our response variable is unimodal with two outliers. These two outliers are Liberty and Calhoun county. The mean of the distribution of our response is .0012 and it has a standard deviation of .0029. We are 95% confident that the true

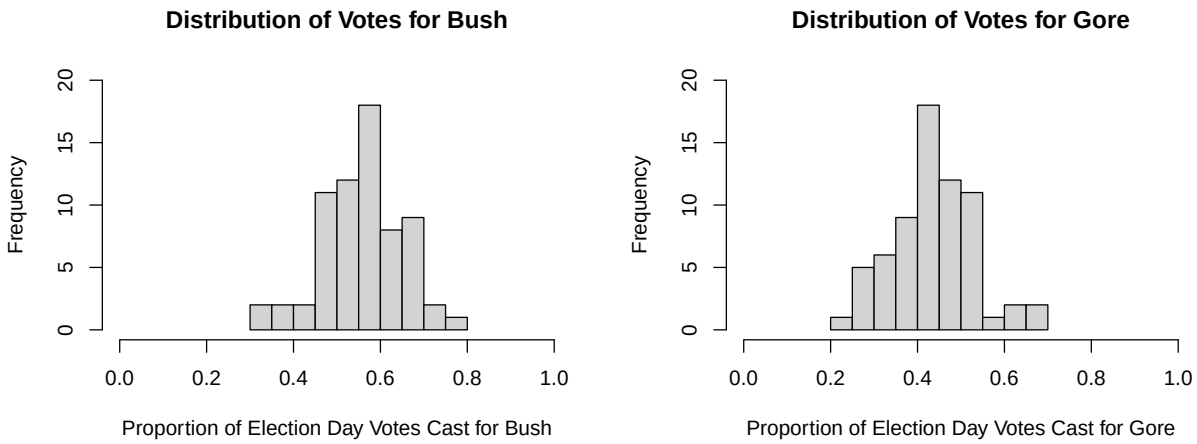

Figure 2: Distribution of Election Day Votes for Bush and Gore

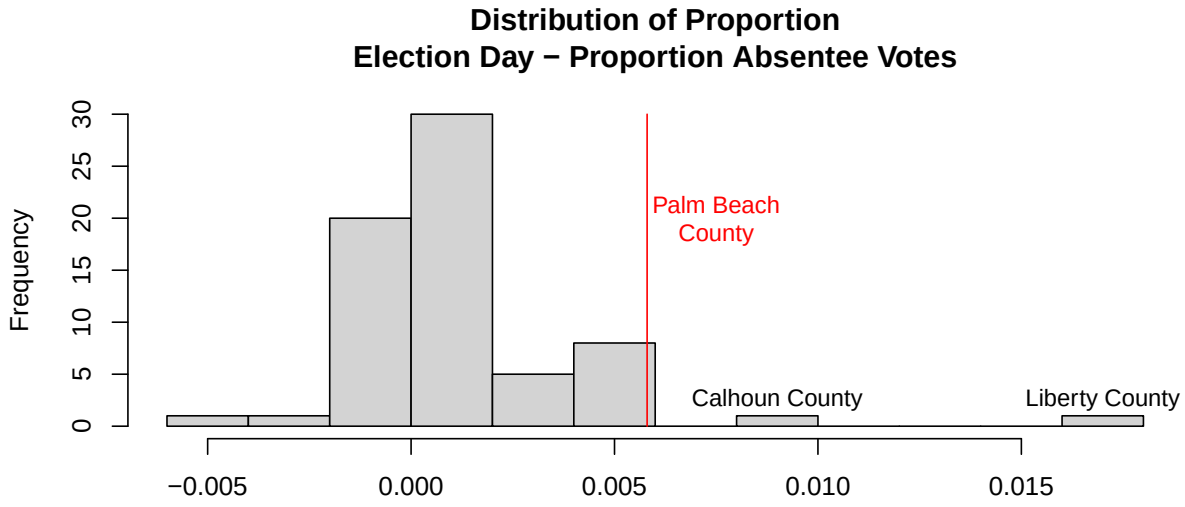

Figure 3: Difference in Voting Method Frequency for Buchanan

mean of this distribution is in the interval [-.0045, and .0069]. We note that this confidence interval includes 0.

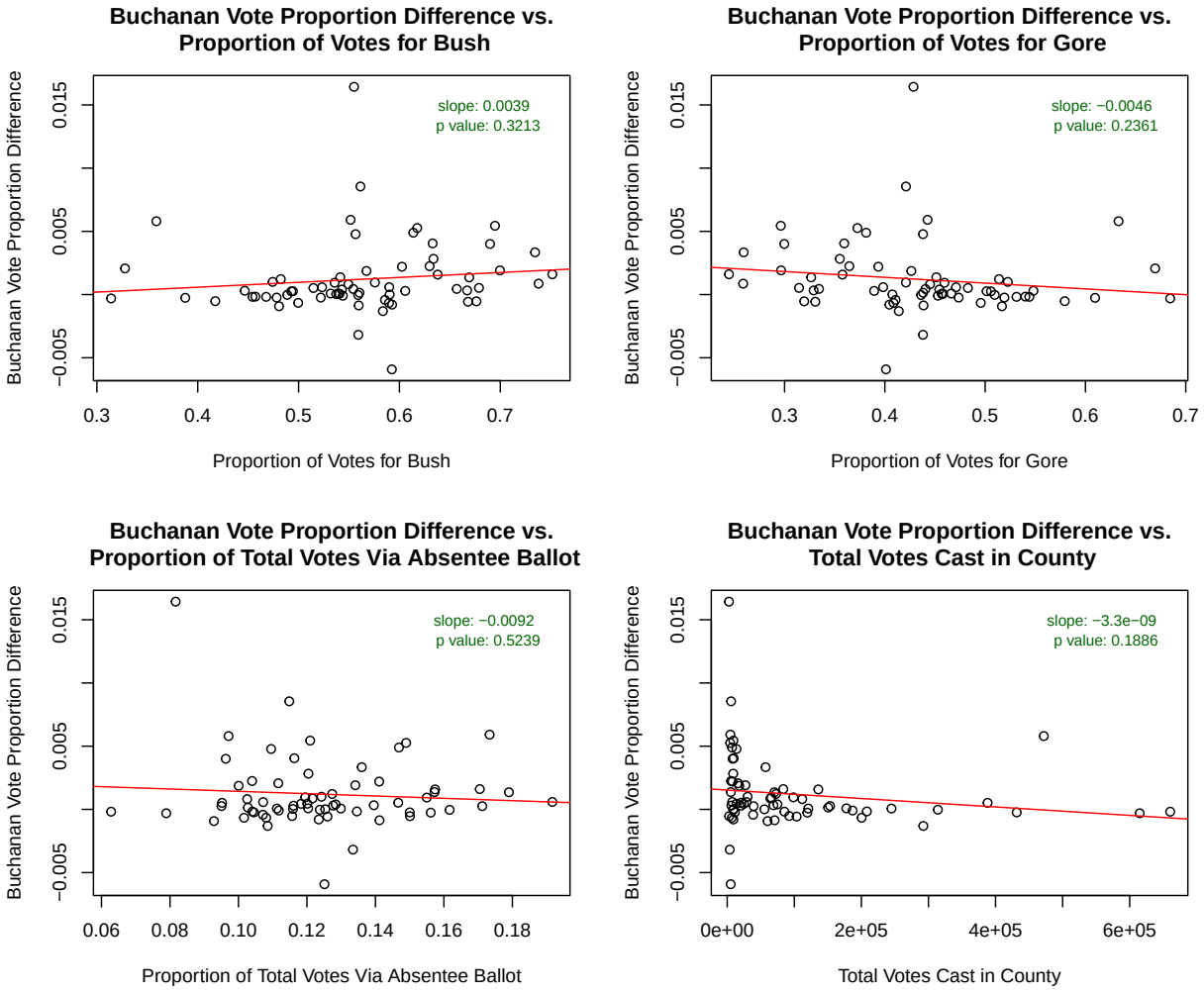

Figure 4: Relevant Pairwise Relationships with Response

For multivariate analysis, we see in Figure 4 that the proportion of votes for both Bush and Gore do not have significant relationships with the response variable. We also notice that the proportion of votes submitted via absentee ballots and the total votes cast in a county also does not have a significant relationship with our response. All four relationships appear approximately linear. Based on this multivariate analysis and our univariate analysis, we do not see any reason to apply transformations to our variables. Figure 4 appears to show linear relationships between the relevant explanatory variables and response variables. We do notice heteroskedasticity in the linear relationships that may need to be addressed through transformations, but we will investigate this during modeling.

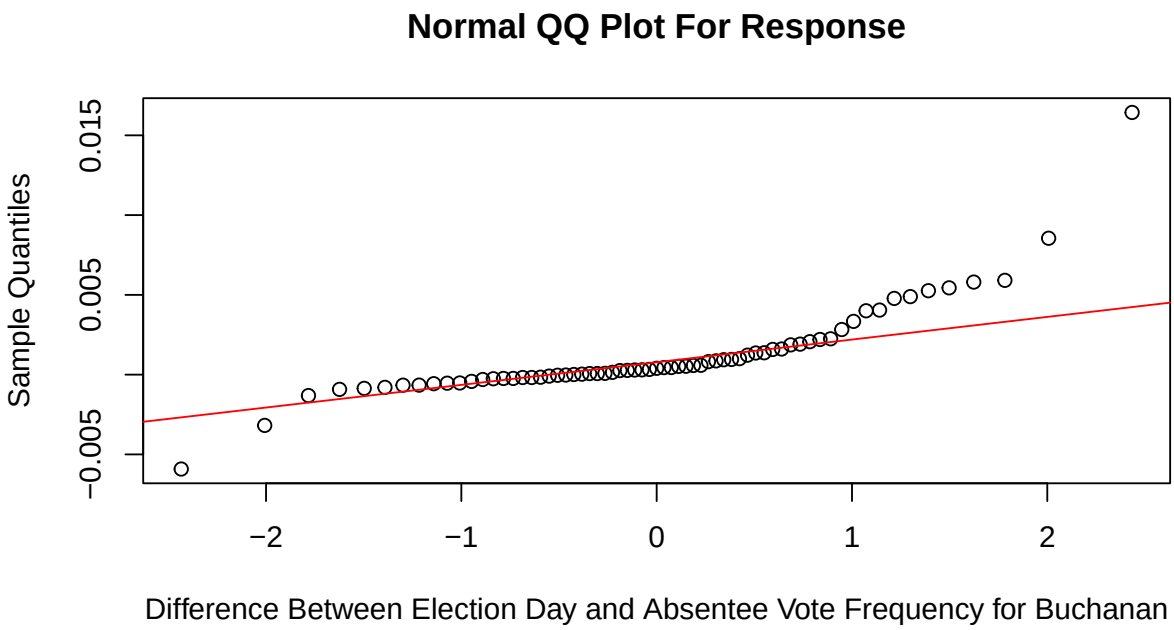

Figure 5: Checking Normality of Response

We would now like to explore to what extent our response variable is normally distributed. We can see in Figure 5 that our response appears to be somewhat normally distributed. At the end of the distributions it appears that the true distribution has wider tails than a normal distribution. However, for the majority of the data points, we see that the distribution of our response appears to be normal.

We saw earlier that `goreVotesProp` and `bushVotesProp` may be good predictors of our response. However, in Figure 6 we see that these variables are highly correlated. So, it makes sense to sum these variables during modeling.

Table 1: Ballots with Vote for Nelson (Democrat)

	other	buchanan
in person	226105	2350
absentee	17747	32

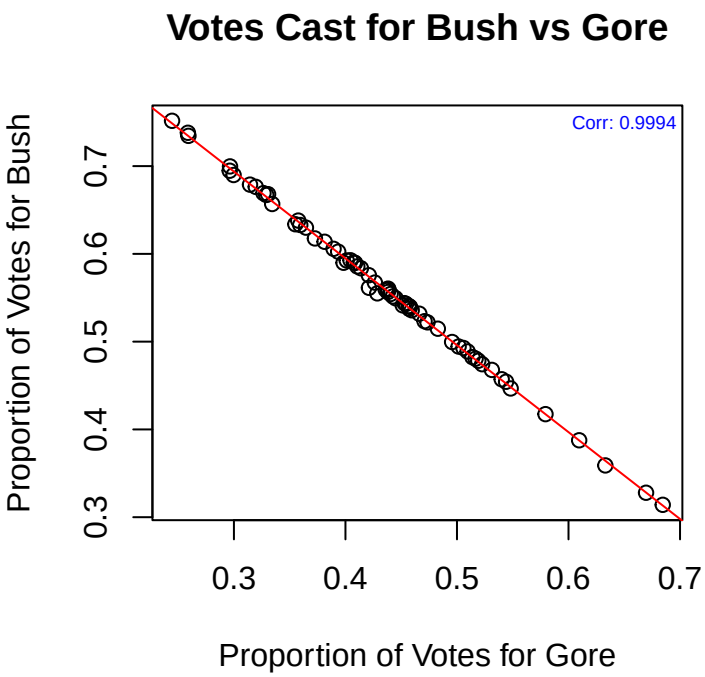

Figure 6: Correlation of Predictors

Table 2: Ballots with Vote for Deckard (Republican)

	other	buchanan
in person	941	59
absentee	91	8

Table 3: Ballots with No Senate Vote

	other	buchanan
in person	151142	852
absentee	18493	41

In the ballot level data in palm beach county, we notice in the tables above (Table 1, Table 2, and Table 3) that ballots who voted democratic for the senate had around 98.7% (95% CI [98.3%, 99.1%]) of their votes cast for Buchanan cast in person. However, ballots who voted republican for the senate only had around 88.1% (95% CI [81.5%, 95.8%%]) of their

votes cast for Buchanan cast in person.

Modeling & Diagnostics

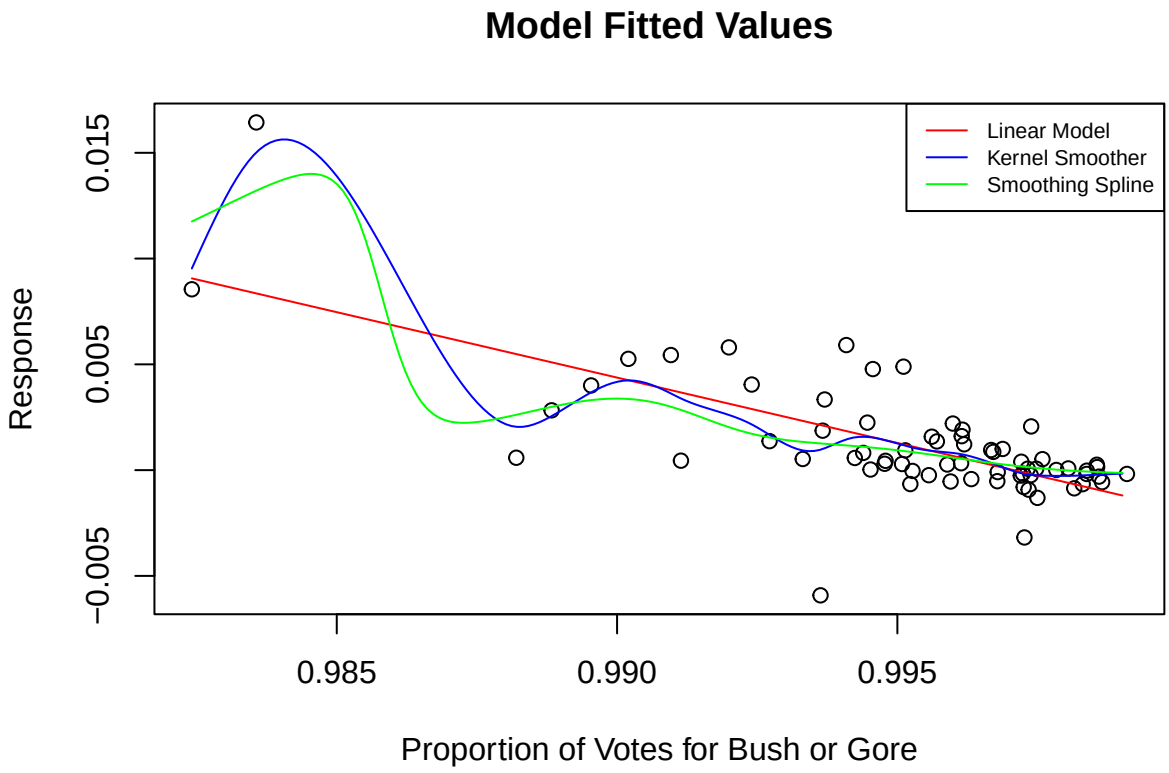

Figure 7: Fitted Values from Models

We will begin our initial modelling by fitting a linear model, a kernel smoother, and a smoothing spline our data. We saw that the proportion of vote for Gore and the proportion of votes for bush may be predictive of our response variables, but they are highly correlated. So, we will sum these variables and only include 1 explanatory variable in all three models. The name of this variables will be `goreBushVotesProp` and it is the sum of `goreVotesProp` and `bushVotesProp`. In Figure 7, we see the predicted values created by all 3 models.

Looking at the plots in Figure 8 we see that variance in the residuals appears to have a positive relationship with the fitted values in all three models. This means the residuals are not identically distributed in all three models. We also notice that the residuals do not appear to have a constant mean of 0. Using log transformation on the explanatory variables seems

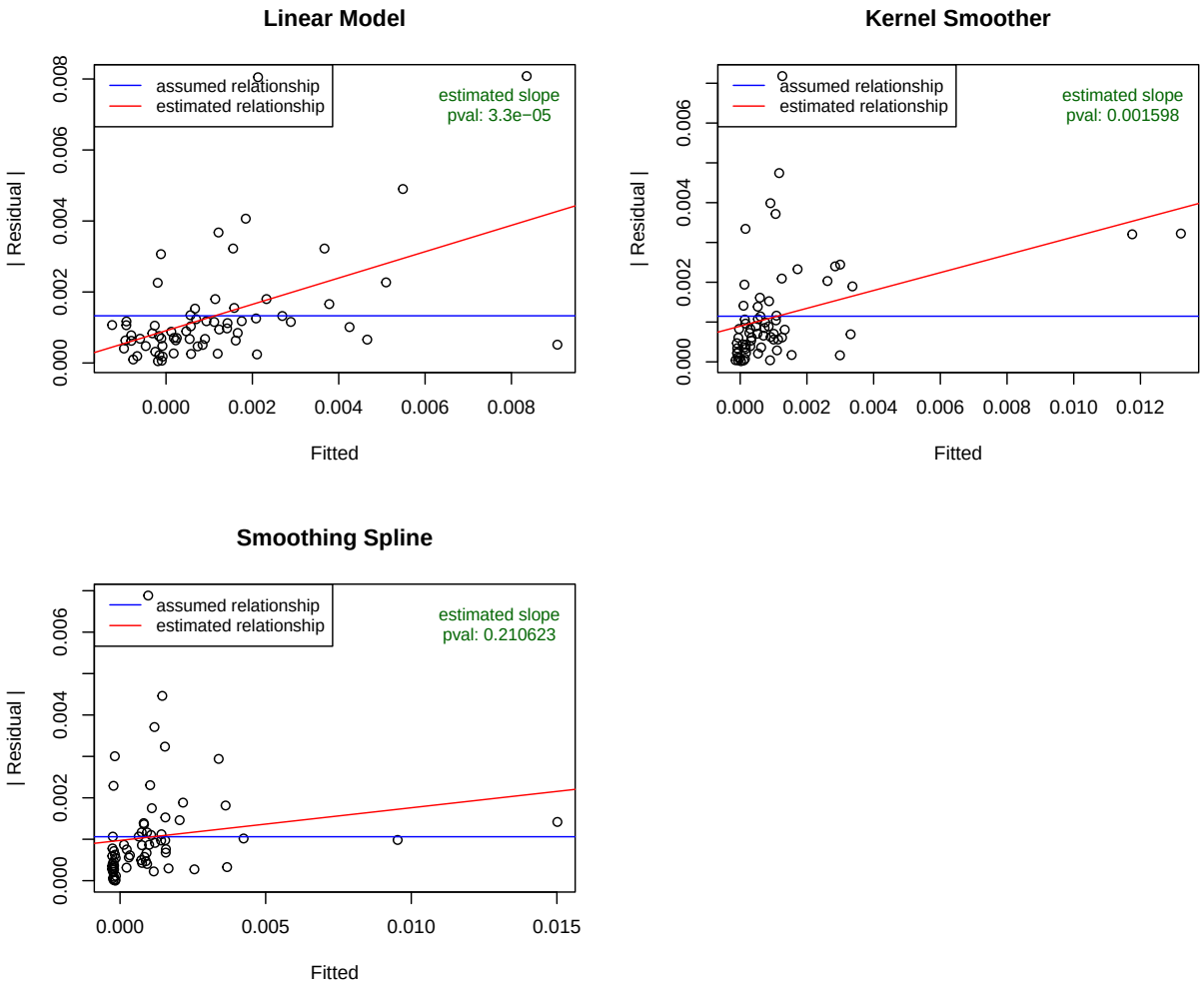

Figure 8: Residual vs. Fitted Plots

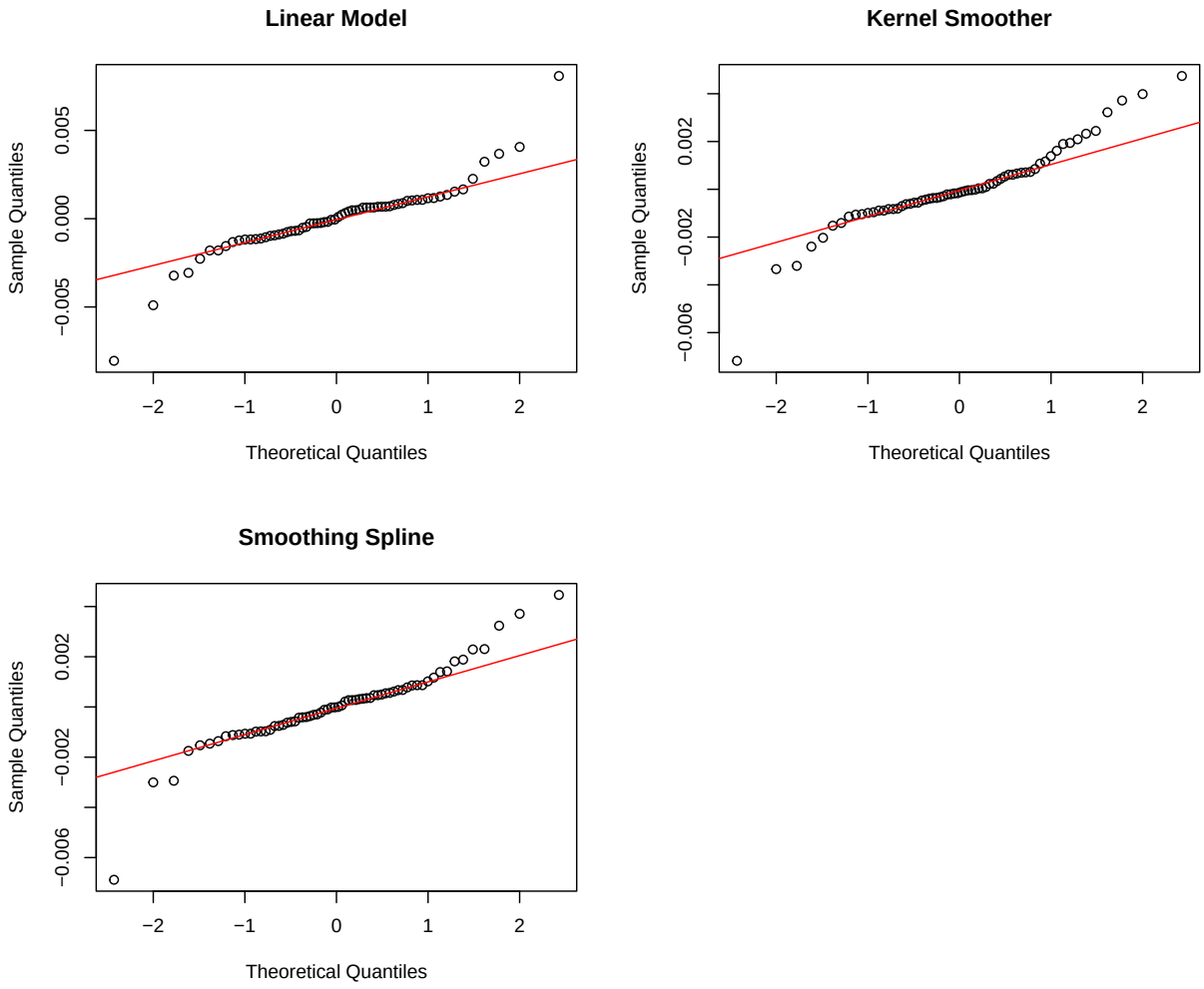

Figure 9: Normal QQ Plots for Model Residuals

to address this issue for the linear model but exacerbates the issue for the kernel smoother and smoothing spline. For the sake of consistency, we will choose to not log transform the explanatory variables. The issues that the transformation creates in the smoothing models outweighs the benefits provided to the linear model. Because we only have one predictor, our residual vs. predictor plot will show the same information as the residual vs fitted.

Looking at the distribution of the residuals for all three models in Figure 8, we notice that the smoothing spline appears to have the most normally distributed residuals. However, at the ends of the distributions, all three models appear to have wider tails than the normal distribution and are more likely to produce large residuals than .

In order to determine which of our three models fits best, we will use leave one out cross validation. Our data set only contains 67 counties so using fewer folds would have a larger impact on the bias of our estimate than if we were using a larger data set. Our prediction error estimates for the linear model, kernel smoother, and smoothing spline are .00222, .00225, and .00126 respectively. The standard error of these estimates are .00224, .00226, and .00118. From these estimates, it appears that our models perform similarly. The difference in prediction error estimate between our best model and our worst model is just .00118. This is less than one standard error from the linear and kernel model prediction error estimates. It is also just slightly more than one standard error from the spline prediction error estimate. So, we don't believe there is significant evidence that any of these models is the best. However, we still choose the spline as our best model because it has the the lowest prediction error as well as the lowest standard error on its prediction error estimate.

In Figure 7 we see that our residuals for the spline model do not appear to be identically distributed. Fitted values below 0 are far more likely to have absolute residuals that are below the average, but this is not true for fitted values above 0. So, we will choose to resample cases to perform our bootstrap.

In Figure 11 we see the conditional regression function estimated for each senatorial candidate. We notice that the probability of voting for Buchanan increases for Nelson voters and decreases for Deckard voters when voting in person rather than via absentee ballot.

Results

Now that we've modeled the our response variable `absBuchananDiff` and diagnosed possible fit issues, we can create a confidence interval for Palm Beach County's fitted value via a bootstrapping. By resampling cases from the empirical distribution, we obtain that our 95%

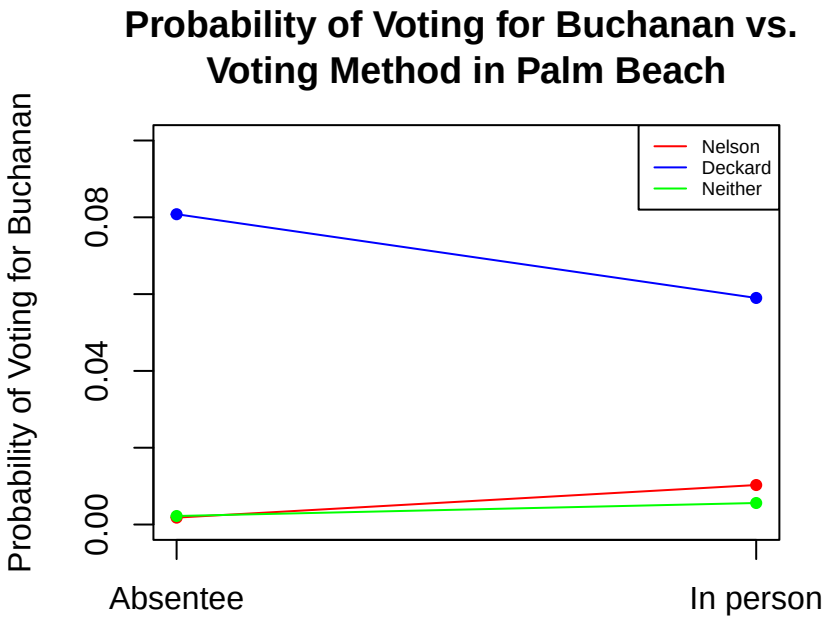

Figure 10: Conditional Regression Function for Buchanan Vote Probability

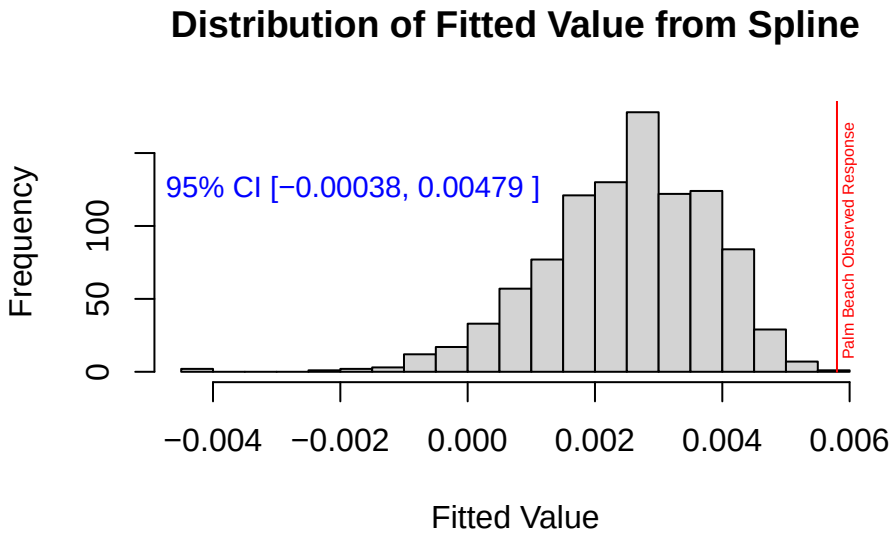

Figure 11: Palm Beach County Bootstrap Distribution

confidence interval for the fitted value of Palm Beach Bounty is [-0.00038, 0.00479]. We notice that the observed response for Palm Beach County is 0.00580 which falls outside our confidence interval of [-0.00038, 0.00479]. We can also see in Figure 10 that the observed response falls at the far right end of our estimated distribution.

Using the individual ballot level data, we can compute the treatment effect of `isabs` on `ibuchanan` adjusting for senatorial vote. Our adjusted treatment effect estimate is -.0064. This estimate can only be interpreted as a causal effect under the assumption that there are no other confounders. Multiplying this estimate by the total number of in person voters, we estimate that in the absence of the butterfly ballot, Buchanan would have received 2432.8 fewer votes. This estimate is only valid if we assume that voting in person or via absentee ballot has no impact on the candidate an individual wishes to vote for.

We are now interested in estimating a 95% confidence interval for the difference in votes we would see cast for Buchanan in the absence of the butterfly ballot. Bootstraping via resampling cases give us an estimated 95% confidence interval of [-.0065,-.0058] for our adjusted treatment effect. Multiplying these estimate by the number of in-person voters in PBC give us a 95% confidence interval for the change in votes cast for Buchanan in the absence of the butterfly ballot of [-2607.75, -2199.00]

Conclusions

Through our analysis of the county-level data we conclude that Buchanan did receive a surprising amount of votes in Palm Beach County. This conclusion was made through the finding that the amount of votes Buchanan received in Palm Beach County fell outside the bootstrapped 95% confidence interval for our predicted values from our smoothing spline. One of the limitations of this analysis is the small sample size. There are just 67 counties in Florida, so our prediction error estimates had high standard errors. If we had a larger data set, we could be more confident that we selected the appropriate model for this problem. Additionally, our small sample size means our empirical distribution may differ from the true underlying distribution by enough to make our bootstrap estimates inaccurate.

The ballot-level data allowed us to estimate the causal regression function for probability of voting for Buchanan conditioned on whether the vote was cast in person or via absentee ballot. Adjusting for senatorial vote, we estimate that Buchanan would have received 2432.8 fewer votes in the absence of the butterfly ballot. Using a case resampling bootstrap, we are 95% confidence that if we were to replicate this experiment with new data, our adjusted

treatment effect estimate would fall in the interval [-2607.75, -2199.00]. The main limitation of our analysis is again our lack of data. The only information we have on the voters is there senatorial vote, whether they voted for Buchanan, and their method fo voting. Our adjusted treatment effect estimate is only a valid causal estimate if there are no other confounders present. However, this is difficult to know with so few explanatory variables in our ballot data set.