36-402 DA Exam One

Sarah Li (sarahli)

3/24/2023

Introduction

(1) In the 2000 U.S. Presidential Election, large speculation had risen after George Bush (Republican) defeated Al Gore (Democrat) by a measly margin of 537 votes in Florida. If Bush hadn't won over Florida, Al Gore would have became president. Many in-person voters from Palm Beach county claimed to have voted for Buchanan instead of Gore using the butterfly ballot due to misalignment. This shortage would have taken away the votes Gore needed to defeat Bush. In this data report, we will be analyzing whether the difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan in PBC was larger than what we would expect, if it difference is statistically significant.

(2) We will use the available datasets of countyFL, and ballotPBC to do our analyses. The former contains the election-day vote counts for the 67 counties in Florida for Bush, Gore, and Buchanan, and the latter contains partial and anonymized individual level ballots for presidential and senatorial votes in Palm Beach County.

(3) After, doing diagnostics on three models, we decided after transforming all the variables in countyFL, a simple linear model using the log (Gore votes) was the best predictor for our response absbuchananDiff. It had a relatively low training error but also the best prediction error. To go along with that, we measured a 95% confidence interval for the the difference in Buchanan votes given that day in Palm Beach county and observed that the original observed difference lied outside of the interval, deeming it surprising and statistically significant. All in all, we conclude Buchanan should have received around 2430 less votes on election day, which could have allowed Gore to win.

Exploratory Data Analysis

(1) We begin by creating four new variables in our county-level data: totalVotes, buchananVotesProp, absBuchananVotesProp, and absBuchananDiff. absbuchananDiff (the difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan) is our target response variable. A snippet of the dataset is shown below.

```
##     county goreVotes bushVotes buchananVotes absVotes absBuchanan totalVotes
## 1  Alachua     47365     34124           263    10694          40      81752
## 2    Baker      2392      5610            73     1111           4       8075
## 3      Bay     18850     38637           248    12587          37      57735
## 4 Bradford      3075      5414            65     1126           4       8554
## 5  Brevard     97318    115185           570    31811          83     213073
## 6  Broward    387703    177902           795    48525          83     566400
##   absBuchananVotesProp absBuchananDiff
## 1          0.003740415   -5.233685e-04
## 2          0.003600360    5.439888e-03
## 3          0.002939541    1.355947e-03
## 4          0.003552398    4.046386e-03
## 5          0.002609160    6.597915e-05
## 6          0.001710459   -3.068568e-04
```

(2) Before modeling, we must conduct some visual diagnositcs on our key variables. Since the original distributions of all the potential predictors (not absBuchananDiff) are heavily skewed to the right, we transform them using log and do any necessary shifting to avoid NaN values. The transformed distributions are shown in Figure 1 below.

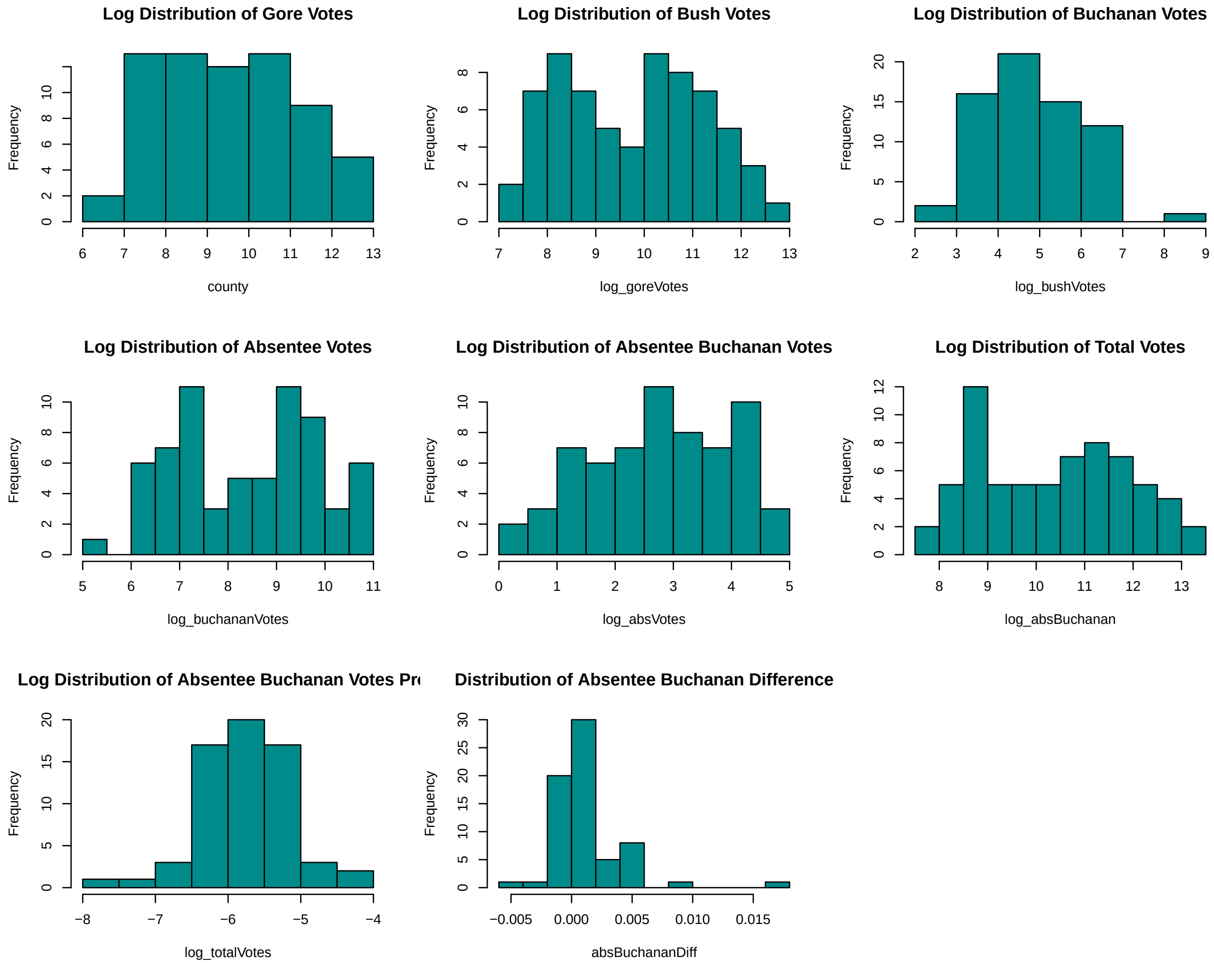


Figure 1: EDA on the Variables

(3) From the histograms, we can see that absBuchananDiff is already nearly symmetric and unimodal. Transforming it using log did not change the distribution. The other variables, after transformation, became more symmetric, though log_bushVotes and log_buchananVotes seem to show signs of possible multi-modality. We will proceed with this caution in mind.

(4) After assessing the quality of our potential predictors and transforming them as necessary, we want to ensure that there isn't high correlation among the predictors given the original data. To do that, we construct a pairs plot and visually analyze both that and the correlation matrix.

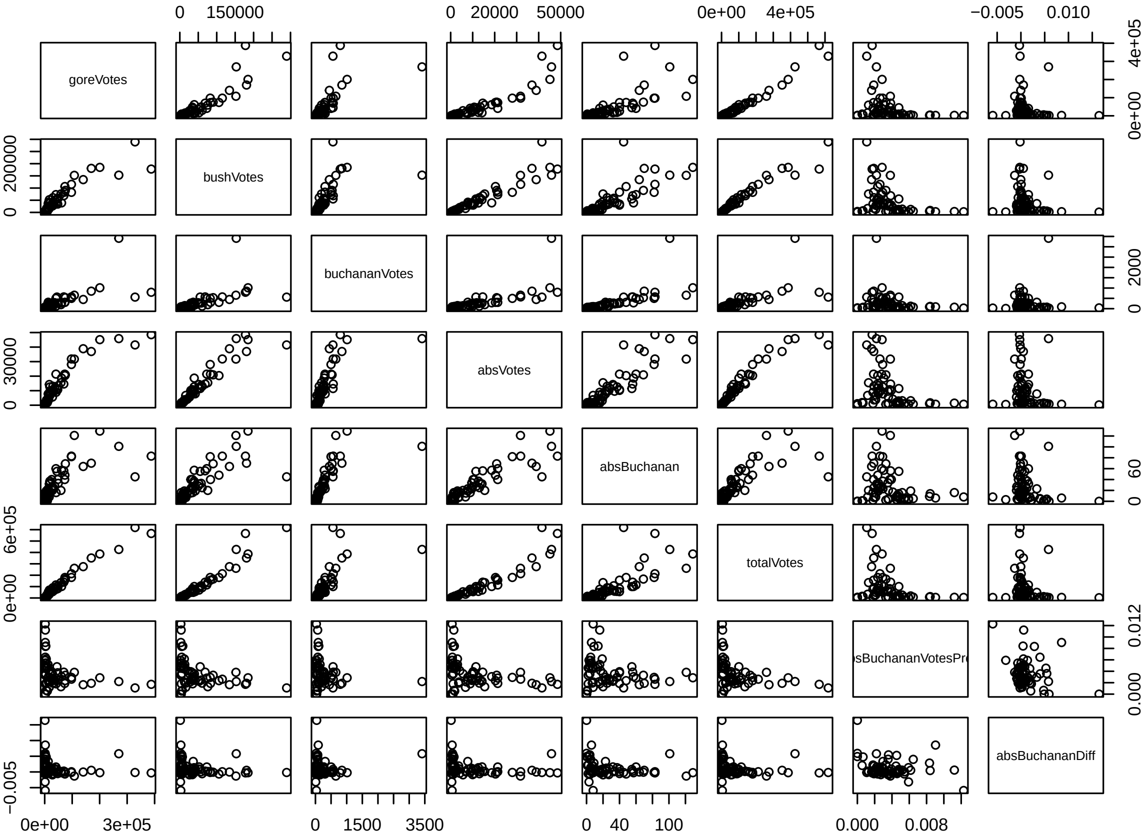


Figure 2: Pairs Plot of CountyFL Vars

(5) From Figure 2, the pairs plot demonstrated that there was heavy correlation among any pair of predictors that did not include the response variable. The correlation coefficients (matrix ommitted from report) for any two possible predictors was also relatively high. Additionally, absBuchananDiff had been calculated using buchananVotes, totalVotes, absBuchananVotesProp, absBuchanan, and absVotes. So to avoid collinearity, we decided to omit these variables as predictors and look toward just log_goreVotes and log_bushVotes for modeling. But we still keep in mind log_bushVotes' slightly bimodal distribution: there could be implication of the errors being dependent or following a non-normal distribution, which would violate our assumptions of linearity.

(6) Using the individual ballot-level data, the second dataset, we make a table showing the total number of votes for and not for Buchanan, for absentee versus non-absentee ballots and ballots with a vote for Nelson, Deckard, or neither.

##

```
##                Non-Absentee Absentee
##   Not Buchanan       378188    36331
##   For Buchanan         3261       81
```

```
##
##              Not Deckard For Deckard
##   Not Nelson      170528        1099
##   For Nelson      246234           0
```

(7) Interestingly from the tables, we can see that there are not many voters overall that were absentee compared to non-absentee. And the proportion of those who did vote for buchanan given they were non-absentee is higher than given they were absentee (.0085>.0022). This could be a possible indicator of there being more votes for Buchanan in person than expected. Additionally, the ballot data shows that Palm Beach voted significantly more for Democratic (Al Gore was also Democratic) candidate Nelson than any other candidate.

Modeling & Diagnostics

(1, 2). After choosing log_goreVotes as our predictor of interest, we begin by modeling with three types of models removing Palm Beach from the dataset: linear, kernel smoothing, and spline smoothing. For visual purposes, we decided to indicate a county as being democratic (blue) or republican (red) by by some external research before the 2000 election.

For the linear model, we have chosen to use only log_goreVotes as the predictor for absBuchananDiff. Consequently, we created three potential linear models to fit the data, one using only log_goreVotes as the predictor, one using only log_bushVotes, and one using both. From our diagnosis of the models' corresponding summary output, residuals vs fitted plot, and normal q-q plot, the model using only log_goreVotes had the lowest p-value = 0.00124 of the three models, with the coefficient of log_goreVotes also having a low p-value of 0.0012. And so we choose to use log_goreVotes as our only predictor. Contextually, this decision also makes sense given that the discrepancy stated in our research question originates from voters mistakenly choosing Buchanan instead of Al Gore. The residuals plot shows a slight fanning toward the right, and there are deviations of the

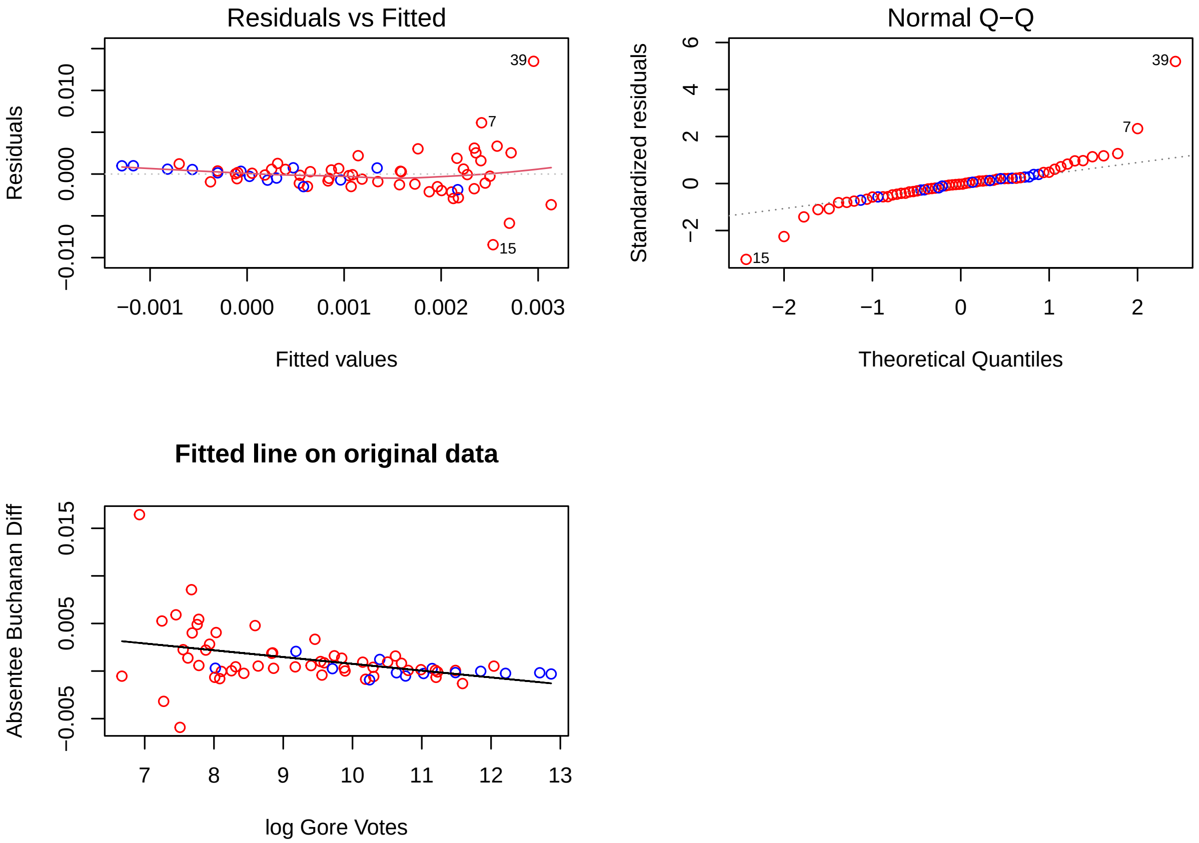
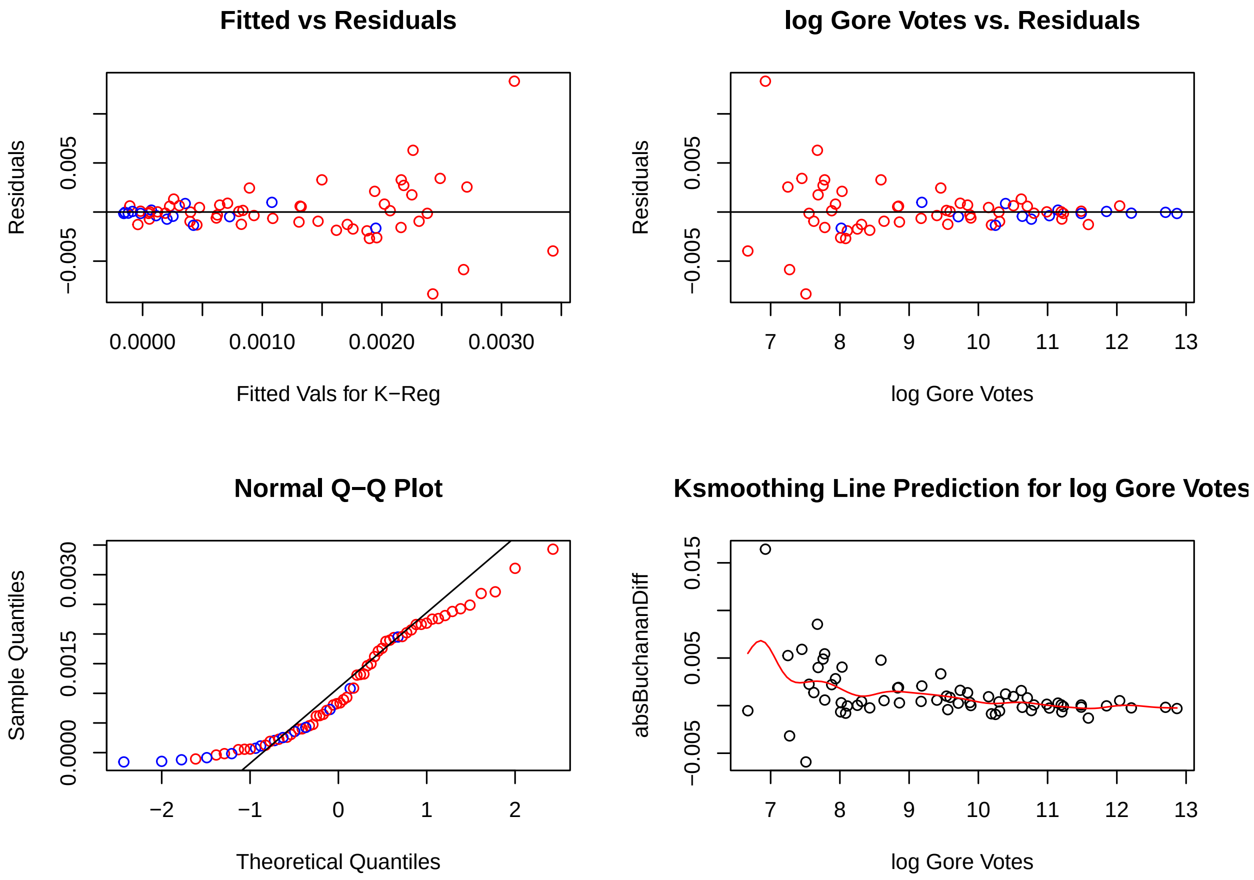


Figure 3: Linear Model of log Gore Votes

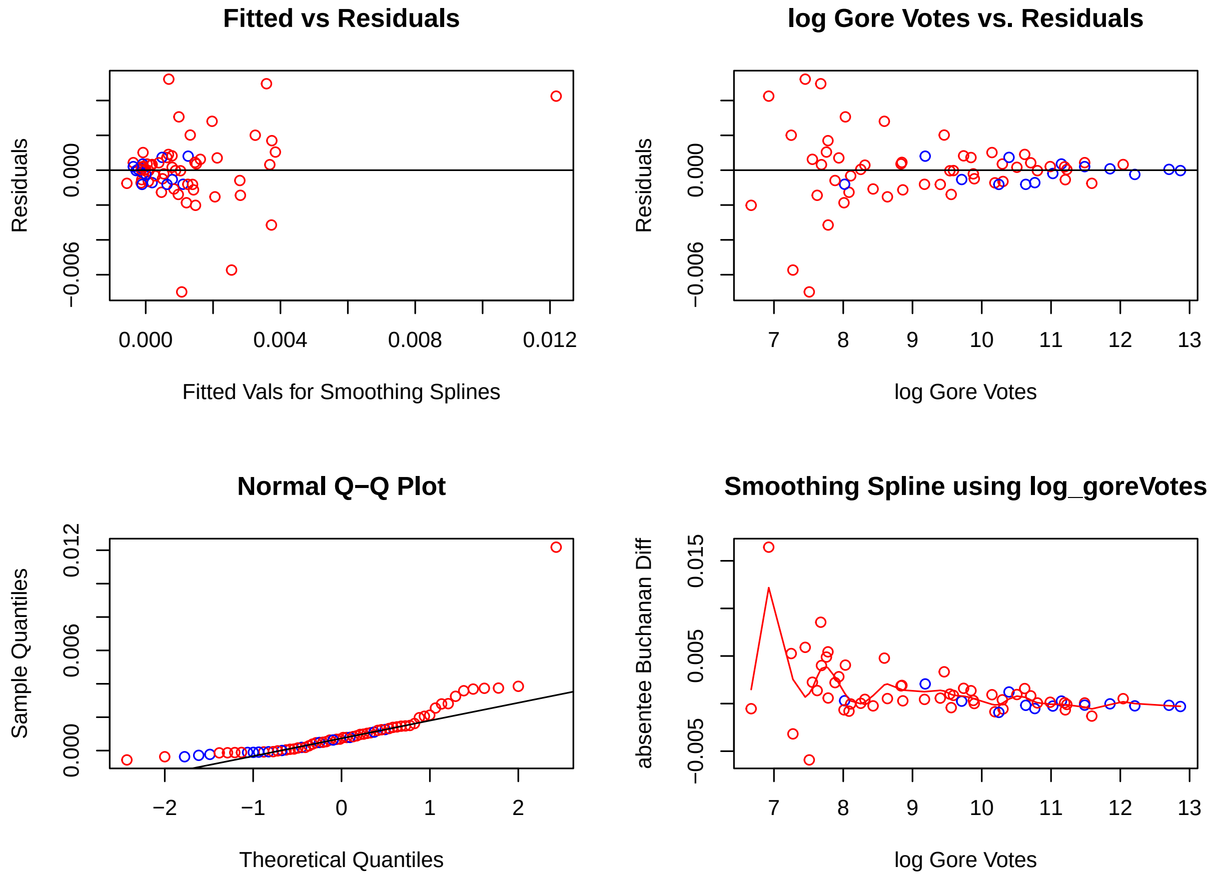
ends of the q-q plot that might indicate violations of normality; however, the plots aren't too bad so we proceed.

(1, 2) cont Next, we do a kernel regression on the same dataset omitting Palm Beach.



Similarly, the residuals vs fitted and log_goreVotes vs residuals plot have a slight fanning shape. The ends of the q-q plot once again deviate from the diagonal line. The estimated line on the original data is displayed in Fig 4 above, with a bandwidth of .0679.

(1, 2) cont Lastly, we fit a Smoothing Splines model using only log_goreVotes using the default parameters.



The residuals indicate the possibility of there being an outlier toward the far right of the data; there is also an indication of fanning from the log_goreVotes vs residuals plot. The outlier seems to be very apparent in the q-q plot and spline graph as well. Overall, the residuals seem to possibly violate assumptions of independence, constant variance, and noramality like the previous plots, though more apparent in this one. The spline is graphed on top of the original data for visual purposes. As an improvement for all models, I might try to remove the outlier than might have skewed or leveraged the data (Liberty County) or tried natural cubic splining with different knot sizes to fit the third model. Otherwise, given the small sample size and possibly inherent dependency, other transformations might be too excessive.

(3) Next, we use cross-validation to determine which of the models fit best to the data in terms of prediction or generalization error, omitting PBC.

```
##     model training_error  loocv_score
## 1 Model 1   6.924147e-06 7.584911e-06
## 2 Model 2   6.805895e-06 7.861496e-06
## 3 Model 3   3.365572e-06 1.903664e-05
```

We choose LOOCV given the small size of the dataset. **(4)** Looking at our table of training errors and loocv scores, we see that Model 3 using smoothing splines performed best on training data, while Model 1 using a simple linear regression on log_goreVotes performed best on testing data. Models 1 and 2 had similar training and validation errors, though all errors are really low to begin with, with the difference between each minimal. However, we want to choose the model that would generalize to unseen or other testing data, so we pick Model 1 as the best one. Revisiting the residuals of each model, we take note that given the 66 data points (a relatively small sample size), the residuals vs fitted values don't completely follow a heteroskedastic shape about 0. The is a slight fanning toward the right and a possible outlier for Liberty county; the residuals don't seem to be highly correlated, though we cannot visually conclude that there is constant variance among the errors in any of the models. **(5)** Thus, given the uncertainty of our residuals in relation to our assumptions of normality, independence, and constant variance, it is best to perform a nonparametric bootstrap in the form of sampling n cases in order to measure uncertainty. This method does not assume accurate linear model fit and is more lenient on residual assumptions.

(6) Now we continue with plotting the conditional regression functions for the individual ballot-level data in Palm Beach. To get a sense of the probability a voter might have voted for Buchanan given that vote was absentee or not. Additionally, we would like to make take into consideration the voter's previous senatorial vote and see if the party alignment makes sense, given that Buchanan is democratic, while senators Nelson and Deckard are Democratic and Republican respectively.

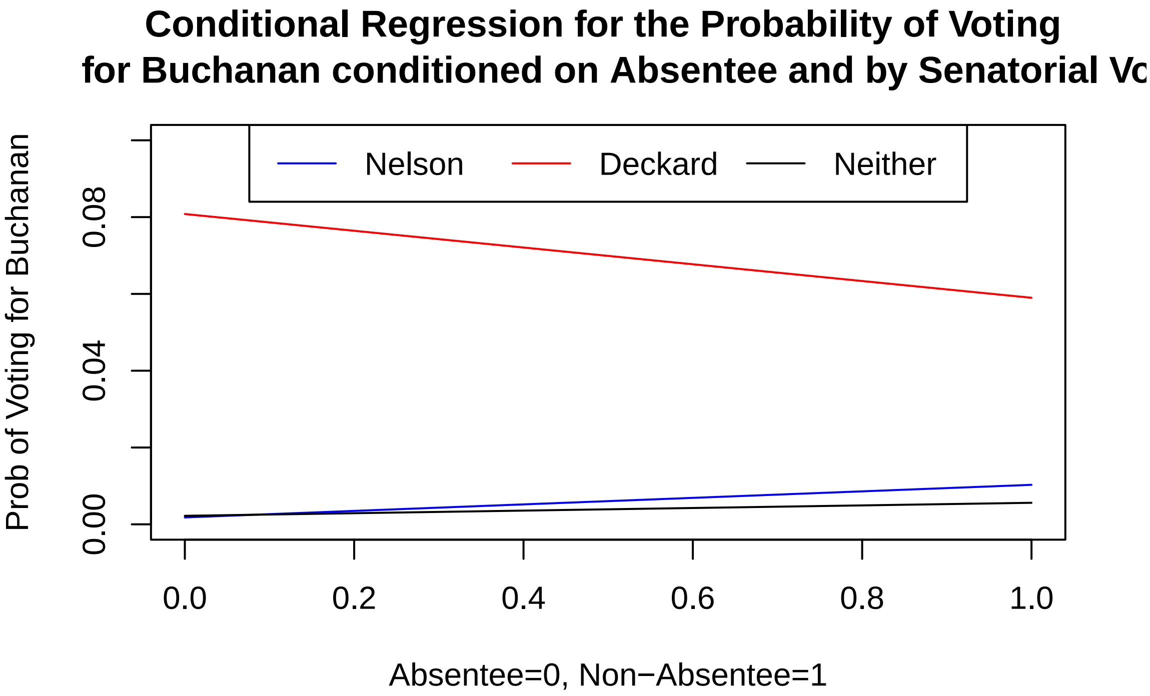


Figure 4: Conditional Regresion for Buchanan given Absentee and by Senatorial Votes

The figure demonstrates that the proportion of Buchanan voters was greater for Non-absentee (in person) voters than Absentee voters, given that they either voted for Nelson or neither party. The probability of voting for Buchanan decreased for Non-absentee voters than absentee voters given that they voted for Deckard in the senatorial elections.

Results

(1) Next, we conduct bootstrapping for the selected linear model for the county data for the expected difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan in Palm Beach County.

```
##               Method     CI.Lower    CI.Upper
## 2.5% Bootrap Cases -0.001127972 0.003116676
```

(2) We conduct a nonparametric bootstrap by resampling cases given the analyses of our residual assumptions. Since there are only 66 data points, we decide to set B to 100 in order to avoid making the confidence too narrow. We are 95% confidence that the true

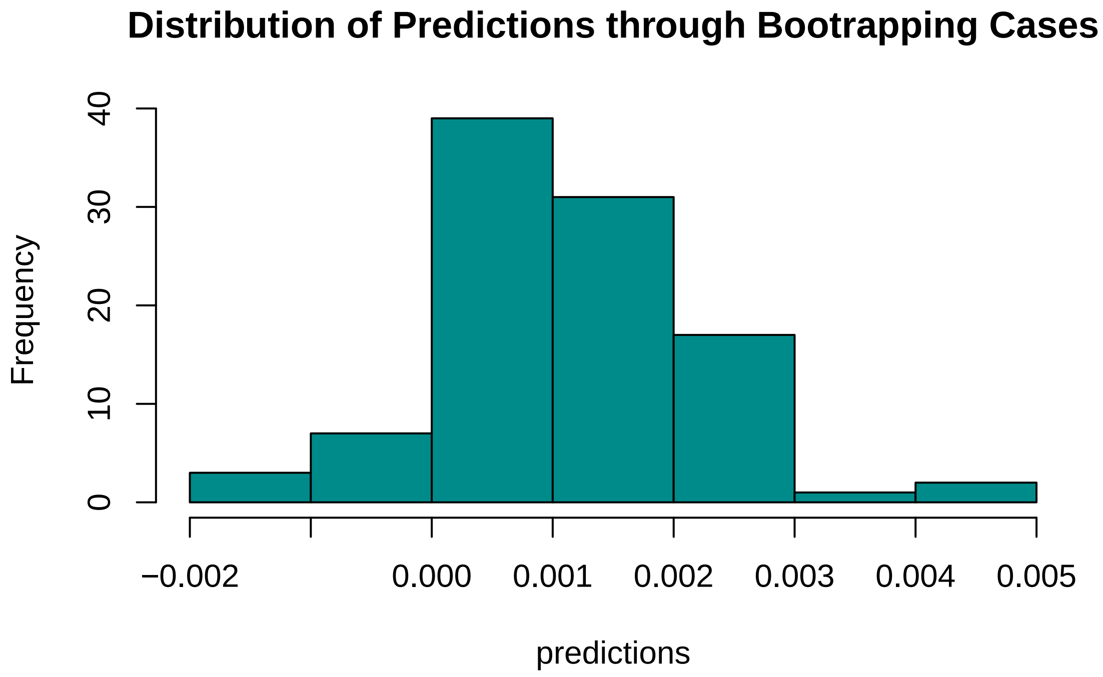


Figure 5: Histogram of the predictions of the differences of PBC

value of absBuchananDiff given Palm Beach county is within the interval (-0.0006402, 0.0035033). Interestingly, the actual value of for Palm Beach in the original dataset is 0.0058, which is outside of the confidence interval. This implies that the value fo absBuchananDiff for Palm Beach county is statistically significant and greater than what we expect to observe, as this variable is defined as the difference between the proportion of election day votes and absentee votes for Buchanan.

(3) Next, we compute the effect of the election day ballot versus absentee ballot on the proportion fo votes for Buchanan, adjusting for senatorial vote and using empirical senate vote (non-absentee). For this causal effect to be valid, we assume that the absentee and non-absentee voting was randomly assigned, or that the voting type is not dependent on the subject. This might not entirely be the case, but we proceed to calculate the relevant effect.

(4) The causal effect we calculated to be is -0.006378, and the expected number of votes we expect Buchanan to have received is calculated to be -2433, or 2433 less votes at PBC on election day. Buchanan would have gotten 979 votes in PBC without the butterfly ballot assuming no other confounding variables.

(5) To construct the 95% CI for this statistic, we do another 95% confidence interval for the expected number of votes for Buchanan with B=500 draws since the data size is much bigger The result is (-2439.332, -2419.414), meaning we are 95% confident the expected number of votes for Buchanan in PBC for the day is -2429.373. Our previous calculated value of around -2433 votes (rounded) is within the range. Since this value is in our range and our previously estimated difference seemed to fall out of our confidence interval, we conclude that the votes for Buchanan in PBC was statistically different and significant, wiht the amount of votes ranging higher than expected for him given the butterfly ballot. More specific to the research question, the expected difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan was significantly higher in our observed data compared to our bootstrap casing with uncertainty.

Conclusions

(1) Deciding to go with our linear model using log(Gore Votes) to predict the absDiffBuchananVotes, we have found through modeling and bootstrapping that Buchanan received an abnormally higher amount of votes in Palm Beach County on election day than he was expected to have. In particular, he should have received an estimated amount of 2430 less votes, the amount needed (537 votes) for Gore to have defeated Bush and won the election.

(2) This conclusions we have drawn and the modeling we have used are only valid given our assumptions about the data first-hand. To create our linear model, we assume that the data itself graphically follows a linear pattern already. More importantly, we expected the residuals/errors to be independent and normally distributed, with constant variance and no correlation. However, given what we know about the data and have seen in the residual plots, the errors seem to have a slightly fanning shape in all three model diagnostics. In reality, we also would not be surprised if the errors are correlated given a county is dependent on the affiliation of the party, and the affiliation of the party could also play a role in voting in-person or not. The limitations would be the assumptions of our model, which bootstrapping with cases does help by tackling these uncertainties within the errors. We also do not know if marginalizing on in-person voting was the sole condition that would confound the results of votes for Buchanan.

(3) But, overall, given our models and bootstrapping amidst the uncertainties, we did find statistically significant results. Buchanan would have indeed received less votes from

PBC on the day of election, specifically ~2430 less votes on average.**(4)** Some limitations of the individual ballot-level analysis include that this difference is affected only by the butterfly ballot's defect and no other confounding variables. Though, given how complex politics is, this is unlikely the case.