Investigating Miscast Votes in the 2000 US Presidential Election

Steffi Chern (steffic)

3/24/2023

Introduction

The 2000 presidential election in the United States was a controversy. The Democratic candidate, Al Gore, lost to George W. Bush (the Republican candidate) by only 537 votes in Florida. However, the issue was the use of the "butterfly" ballot format in Palm Beach County in Florida, which may have led to some voters mistakenly casting their vote for Pat Buchanan (the Reform party candidate) instead of Gore. **(1)** This study aims to investigate whether there is a strong evidence that votes were miscast in Florida. Specifically, we want to answer the following two questions:

- whether the difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan in Palm Beach County is larger than what would happen by chance
- how many more votes did Buchanan receive than he would have in the absence of the butterfly ballot, assuming there is sufficient evidence of votes miscast

(2) This study uses two datasets:

- County_fl.csv: It contains the election-day vote counts for Bush, Gore, and Buchanan in each of the 67 counties in Florida. In addition, there are the absentee vote counts for Buchanan and the total number of absentee votes casted in each of the 67 counties in Florida.
- BallotPBC.csv: It contains the individual level ballots for Palm Beach County, Florida, where the butterfly ballot was used. Specifically, the information regarding whether the presidential vote was for Buchanan, whether the senatorial vote was for Bill Nelson (Democratic), whether the senatorial vote was for Joel Deckard (Reform), and whether the vote was absentee or not.

(3) From our study, we obtain the conclusion that the difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan in

Palm Beach County is significant. Moreover, Buchanan received around 1745 more votes than he would have in the absence of the butterfly ballot. This further implies that the election result would've likely been different if the butterfly ballot wasn't used.

Exploratory Data Analysis

(1) To help answer the research questions, we first create four new variables based on the variables provided to us:
- *totalVotes*: total number of election-day votes for Bush, Buchanan, Gore
- *buchananVotesProp*: proportion of election-day votes for Buchanan in each county
- *absBuchananVotesProp*: proportion of absentee votes for Buchanan in each county
- *absBuchananDiff*: difference between the proportion of election-day votes for Buchanan and proportion of absentee votes for Buchanan

(2) For our univariate EDA, we first look at the histograms of the variables *buchananVotesProp*, *absBuchananVotesProp*, *bushVotes*, *buchananVotes*, and *goreVotes*. These variables are most closely related to the research questions. The histograms provide us information about the distribution of these key variables in our dataset.

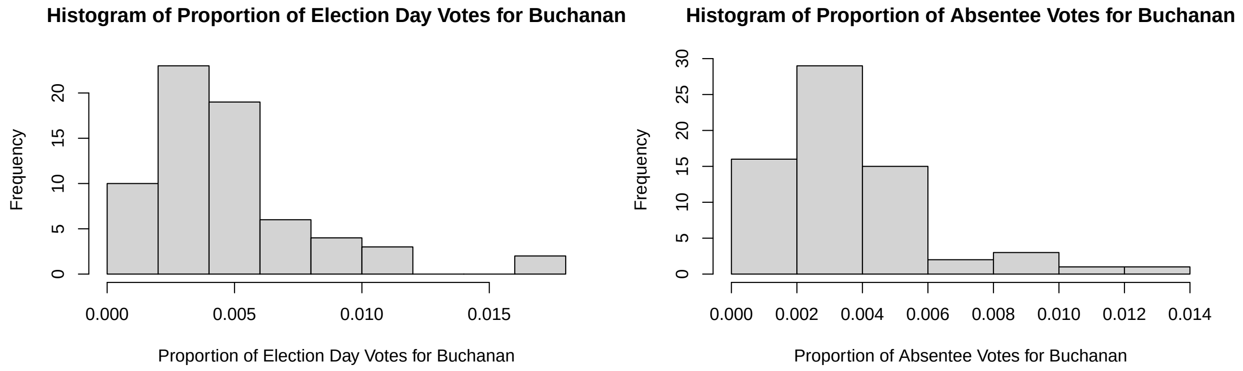

Figure 1: Histogram of Proportion of election day and absentee votes for Buchanan

(2) and **(6)** From the histograms, we see that the variables are all right skewed, unimodel, with potential outliers at the far right of the graph (see figure 1 and 2). When we compare the histograms between *buchananVotesProp* and *absBuchananVotesProp*, there seems to be a heavier tail for the distribution of *buchananVotesProp*, which may suggest that there are more people who voted for Buchanan on election-day, proportionally.

(3) The response variable in the county-level data is *absBuchananDiff* since we'll make

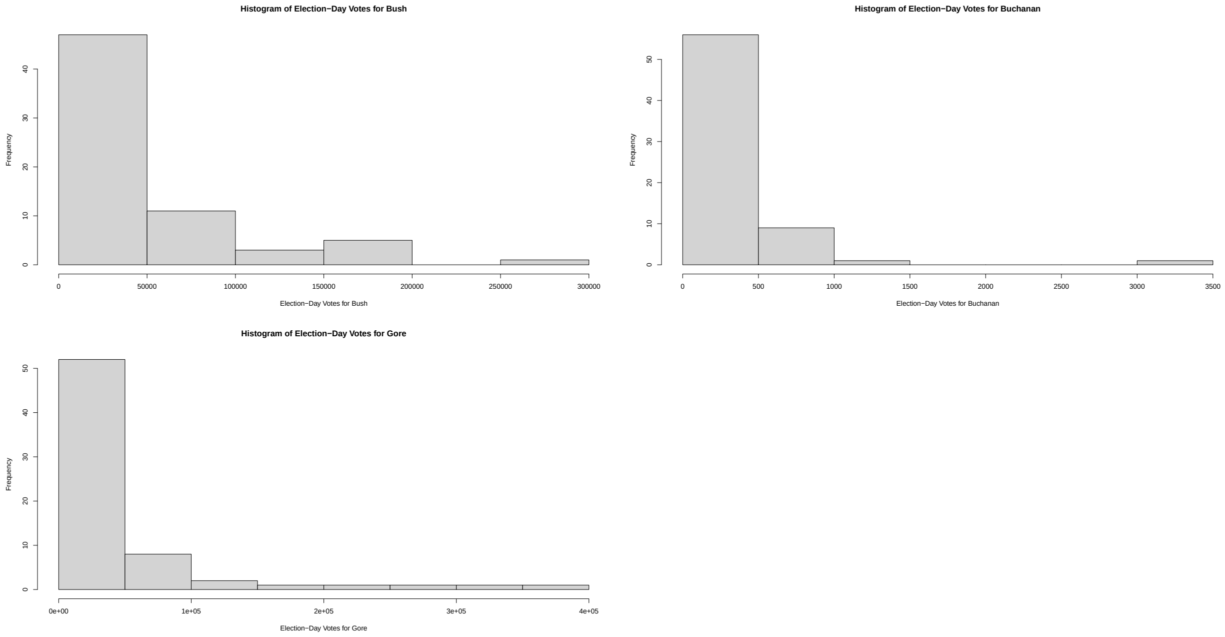

Figure 2: Histogram of Proportion of Election Day Votes for Bush, Buchanan, Gore

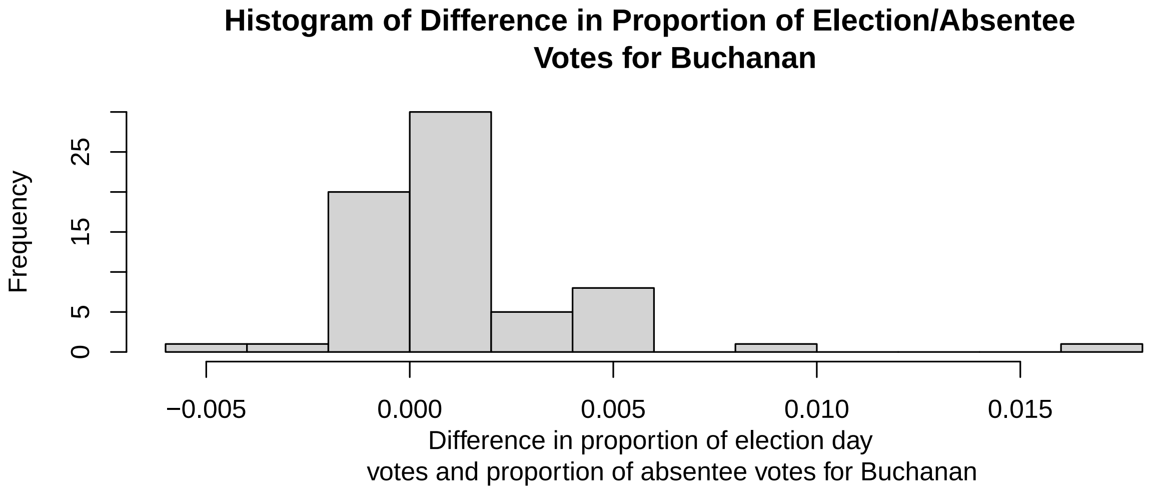

Figure 3: Histogram of Response Variable

predictions about it in later sections. We look at its distribution by plotting a histogram for this variable (see above). We observe a slightly right skewed distribution with potential outliers.

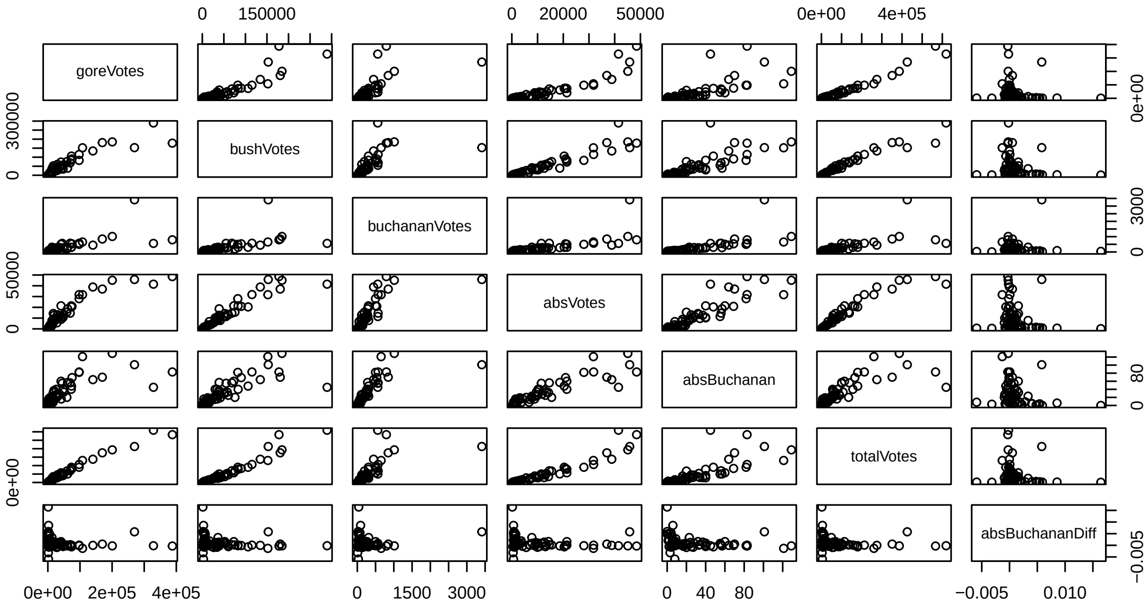

Figure 4: Scatter Plots for Response vs Predictors

(4) To explore the relationships between the predictors and the response variable, we conduct multivariate EDA through scatter plots (see figure 4), which helps us identify linear relationships. The variables *buchananVotesProp* and *absBuchananVotesProp* are not part of the predictors since they are directly used to calculate the values in the response variable. *County* is also not included since the Palm Beach County is specifically chosen to make predictions about. The other variables (excluding the 3 variables just mentioned and the response) are considered as predictors, as they could all possibly relate to the differences between the types of voting.

(5) Based on the univariate and multivariate EDA, it seems appropriate to transform the predictors. The linear relationships are not obvious for the predictors *goreVotes*, *bushVotes*, *buchananVotes*, *absVotes*, *absBuchanan*, and *totalVotes* vs the response variable (many points are clustered together at lower values of each of these predictor), thus we can try taking the log transformation of each predictor. Since some values in the predictors are 0 (log of 0 is undefined), thus we first add 1 to the original value then take the log transformation.

(5) and **(6)** After taking the log transformation on the predictors, there seems to be a more obvious linear relationship between each of the predictors vs *absBuchananDiff* (see figure 5). Even though it's difficult to tell if the linear relationship is positive or negative, the linear assumption is met. There might be issues with multicollinearity in this case, where predictors are highly correlated with each other, which could possibly make it difficult to distinguish the individual effect of each predictor on the response variable.

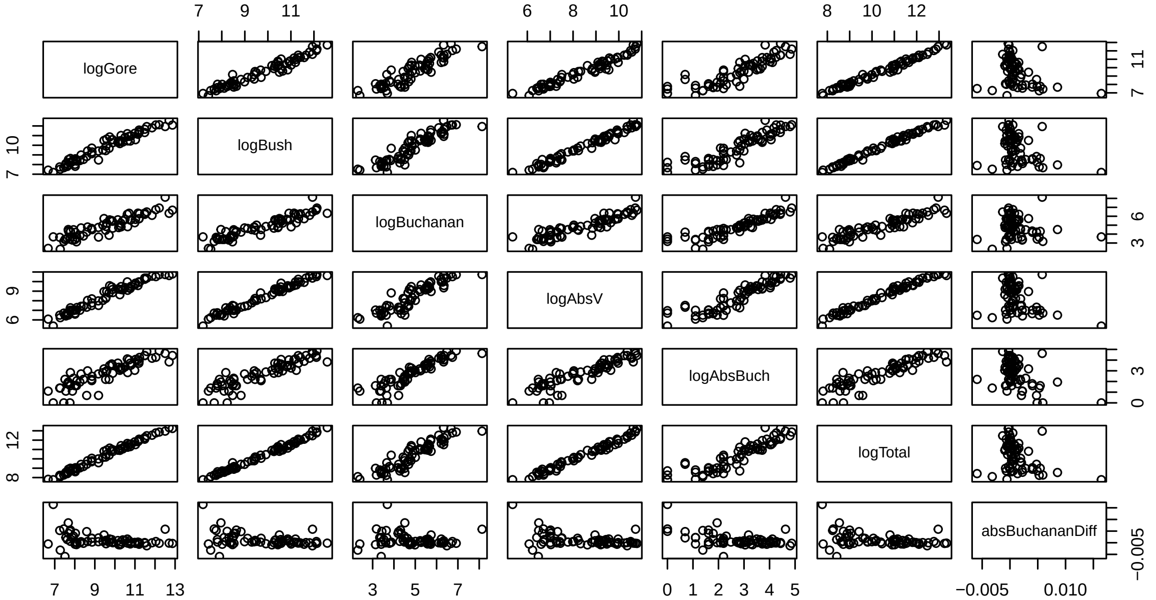

Figure 5: Scatter Plots for Response vs Predictors After Transformation

(7) To explore the individual-level ballots, we created a table showing the total number of votes for and not for Buchanan, for absentee versus non-absentee (election-day) ballots and ballots with a vote for Nelson, Deckard, or neither, as shown below.

	Absentee Voting	Election-Day Voting
Buchanan	81	3261
Not Buchanan	36331	378188

	Nelson	Deckard	Neither
Buchanan	2382	67	893
Not Buchanan	243852	1032	169635

(8) From the county-level data, we see a higher proportion of Buchanan voters voted on election day compared to that of absentee, which may imply that there were miscasts during election day.

From the individual ballot-level data, we see a higher proportion of Buchanan voters who voted for Deckard, compared to non Buchanan voters who voted for Deckard, suggesting that we could try adjusting for senatorial votes in our regressions to verify whether or not there were miscasts.

Modeling & Diagnostics

(1) We constructed three models (linear, kernel regression, smoothing spline) to predict *absBuchananDiff* from the log transformed predictors identified in the previous section. For the smoothing spline, we decided to use the log of Buchanan Election-Day votes as the predictor since it seems to be the most influential predictor of them all – the higher the election-day votes for Buchanan, the higher likelihood that the difference in election-day and absentee votes would be significant.

(2) To determine whether the three models (linear, kernel regression, and smoothing spline) fit well, we plot the residual vs fitted values and the normal QQ plots for each of them (see figure 6). The first row of the diagnostic plots corresponds to the linear model, second row corresponds to the kernel regression, and the third row corresponds to the smoothing spline.
Linear Model: mean residuals approximately 0, but residuals have slightly increasing variance and are not normally distributed
Kernel Regression: mean residuals close to 0 and approximately normally distributed, but the variance of residuals increases then decreases across the fitted values
Smoothing Spline: mean residuals around 0, but the variance of the residuals increases as the fitted values increases, and the residuals are not normally distributed
Since the biggest issue here would be the nonconstant variance of residuals due to a few outlier points, there doesn't seem to have much improvement we can do to improve our model fit. Thus, we'll proceed with our analysis with these settings.

To determine which model fits the data the best, we performed the leave-one-out cross-validation (LOOCV) and calculate the prediction error for each model. We chose LOOCV because it provides a more precise estimate of each model's predictive performance.

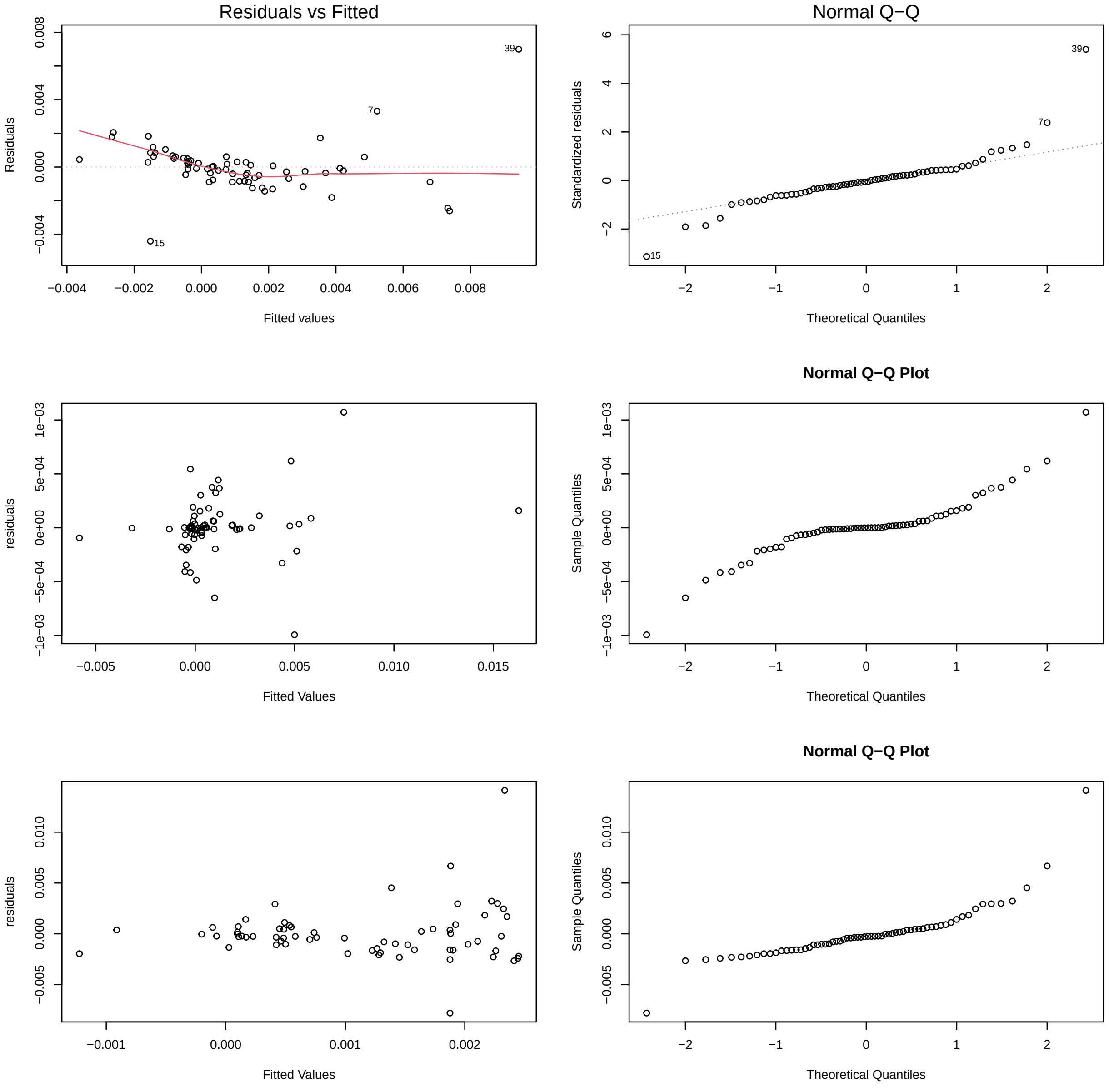

Figure 6: Diagnostics Plots (Residuals vs Fitted Values and Normal QQ Plot)

(3) After performing LOOCV, we got the following result:
Linear Model estimated prediction error: 3.066357e-06
Kernel Regression estimated prediction error: 4.31366e-06
Smoothing Spline estimated prediction error: 8.337001e-06
Linear Model estimated prediction error SE: 1.128942e-05
Kernel Regression estimated prediction error SE: 1.735514e-05
Smoothing Spline estimated prediction error SE: 2.909827e-05
It seems that the linear model has the lowest estimated prediction error out of all the 3 models. Therefore, the linear model appears to be the best model. **(4)** When we take into account the standard errors of the estimated prediction error and construct confidence intervals of the estimated prediction errors for each model, they would overlap. This indicates that the difference between the models do not appear significant. In this case, we would choose the linear model since it is the simplest model and it has the lowest estimated prediction error.
(5) From the residuals vs fitted values plot for the linear model, we notice that the variance of the residuals are not constant across the fitted values (error assumption violated), thus we choose resampling cases with replacement as our bootstrap method.

(6) To explore our individual ballot-level data, we plotted the conditional regression function for the probability of voting Buchanan, conditioned on whether the ballot is absentee and on the senatorial vote (see figure 7).

Results

(1) Based on the linear regression model for the county-level data, we constructed a bootstrap confidence interval 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. We conducted the bootstrap procedure by resampling with cases, as explained previously. Specifically, we attained a 95% confidence interval (0.004381176 0.009776604).

(2) The observed difference in Palm Beach County is 0.005801479, which is within the 95% confidence interval. This indicates that the expected difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan in Palm Beach County is not statistically significant.

(3) From the individual ballot-level data, we computed the effect of the election-day ballot versus the absentee ballot on the proportion of Buchanan votes, adjusting for senatorial

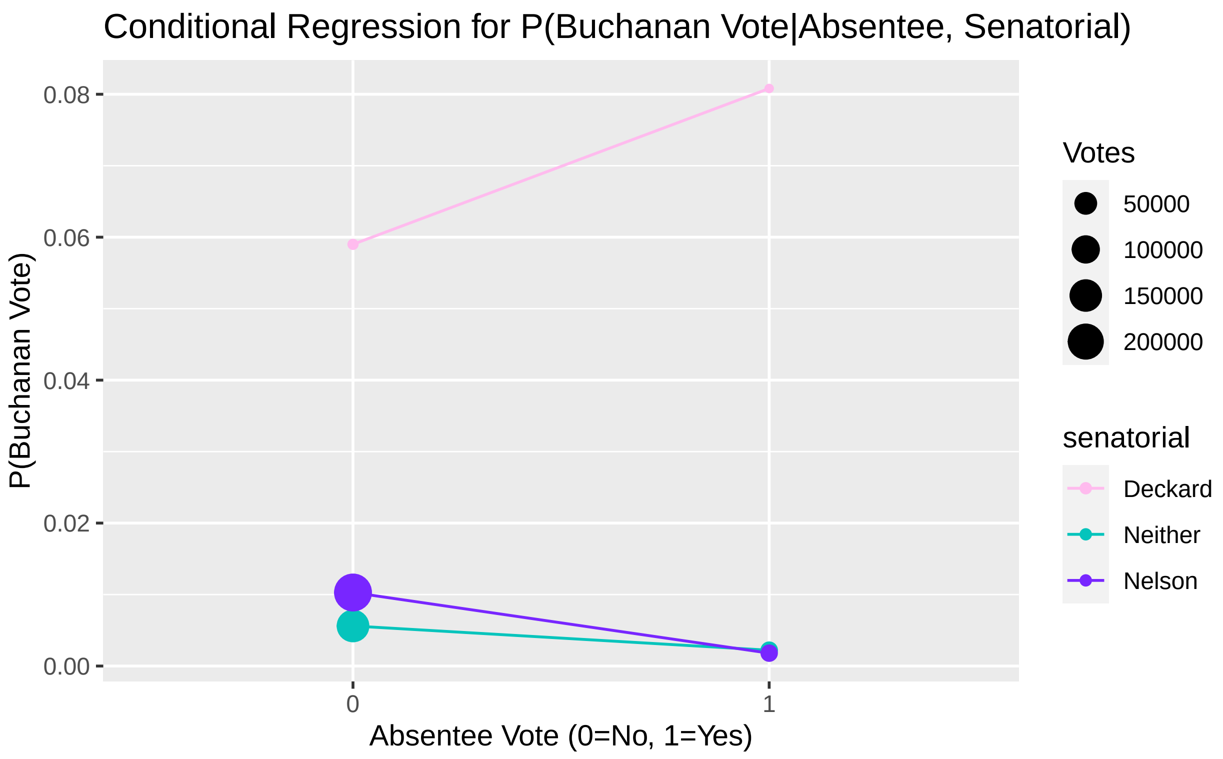

Figure 7: Conditional Regression for P(Buchanan Vote | Absentee, Senatorial

vote. There are 2 assumptions that make this a valid estimate of a causal effect – either the voters are randomly assigned to each senatorial group, or the senatorial vote variable is the only confounding variable we need to adjust for.

(4) We multiply the adjusted difference in vote proportions by the total number of non-absentee votes to estimate the expected number more or less of Buchanan votes in Palm Beach County in the absence of the butterfly ballot, which we get a result of around 1745. For this difference in vote counts to be entirely due to the butterfly ballot, we require that the senatorial vote is the only confounding variable in our case.

(5) We use the resampling cases (100 cases) bootstrap procedure to construct a 95% confidence interval of the expected number of votes more or less for Buchanan, since there's no model created in the process when calculating the adjusted effect (no residuals). We chose 100 resamples since it is a good enough number of resamples for us to get reliable estimates. After bootstrapping, we obtain a 95% confidence interval of (1684.844, 1809.546).

Conclusions

(1) To better understand the controversy related to the US election in 2000, we have constructed different models and diagnostics to determine whether the Buchanan received a surprising number of votes in Palm Beach County. Based on the results, we conclude that the difference between the proportion of election day votes for Buchanan and the proportion of absentee votes for Buchanan in Palm Beach County is not statistically significant.

(2) Based on our EDA section (specifically the residuals vs fitted values plots), we notice that the the model assumptions are not fully met. The residuals in all 3 models (linear, kernel regression, smoothing spline) have mostly mean 0, but they all suffer from heteroskedasticity to some extent. As a result, our model predictions are potentially not as accurate. However, they still provide us with reasonable insights about the dataset. There are other limitations for the county-level analysis, which includes the limited number of predictors. Other variables such as political affiliations, demographic factors, and campaign strategies could be potential predictors as well. In addition, the county-level analysis assumes that the samples collected from the counties are representative of the entire population of the counties. This may not be the case due to the fact that some people did not vote at all (may have nonresponse bias).

(3) Due to the butterfly ballot, Buchanan received an estimated of 1745 more votes, with a standard deviation of around 32. As a result, we can conclude that without the use of the butterfly ballot, it is likely that the outcome of the election would be overturned, since Gore lost to Bush by only 537 votes (i.e. if those 1745 votes belonged to Gore instead of Buchanan, Gore would've won the election).

(4) For the individual ballot-level dataset, there are certainly other confounding variables that weren't included in the dataset. This would reduce the accuracy of the estimated adjusted effect that we calculated in a previous section.