Short Report 1

Parker Johnson

10/7/2021

Introduction

The 2000 US election between candidates Al Gore and George Bush was one of the most hotly contested elections in US history. The result of the election would end up being determined by whichever candidate won the state of Florida, with George Bush eventually winning the state by a slim recount of fewer than 400 votes. However, controversy arose in Palm Beach County, as there were an unusually high amount of votes for candidate Pat Buchanan due to the layout of the ballot being confusing for many Democratic voters. In this analysis, I will explore this case and its impact on the results of the election in Florida, and whether or not it is feasible that this case could have influenced the election results. The data for this case covers the number of votes for Buchanan versus Bush in each of the 67 counties in Florida.

Results

When I fit the regression of Bush votes on Buchanan votes, an initial scatter plot of the data indicates that Palm Beach County did contain an unusually high number of Buchanan votes for the number of votes for Bush in the county. This is indicated in Figure 1 below.

To obtain a more accurate model, I removed the point for Palm Beach County, and transformed the model into a linear model by taking the natural log of both variables. This allows me to analyze and interpret the data much easier and with more accuracy.

I then used a prediction model to predict how many votes for Buchanan we would expect for Palm Beach County given the new model. We are 95% confident that a re-poll in Palm Beach County will have a vote count for Buchanan between 250 and 1399 votes. The predicted number of votes for Palm Beach County is 592 votes with a standard error of 2 votes. The prediction model is indicated in Figure 2 below.

Since the upper bound of this prediction interval is 1399 votes, and the actual number of votes for Buchanan was 3407 votes, the likely size of the amount of votes within the 3407 votes intended for Al Gore is between 2008 and 3157 votes.

Discussion

With this model, we found that assuming some of the votes for Buchanan in the 2000 election were intended to be for Al Gore, then an estimated 2008 to 3157 of these votes for Buchanan negatively impacted the number of votes for Al Gore. This means that under this assumption, and assuming that this is the only case of such voter confusion from the 2000 election, it is possible that Al Gore should have won the state of Florida and the overall election, as the initial votes in Florida showed a difference of only 1738 votes, and the recount showed a difference of less than 400 votes.

One limitation to this data is that it fails to account for other possible cases of voter confusion in the 2000 election. It is possible that other counties around the country had similar structured ballots to the one in Palm Beach County, which could indicate that other counties contained unintended votes for the wrong candidate that negatively impacted either Bush or Gore. This means that it's quite a strong statement to say that this single case of voter confusion could have impacted the result of the entire US election.
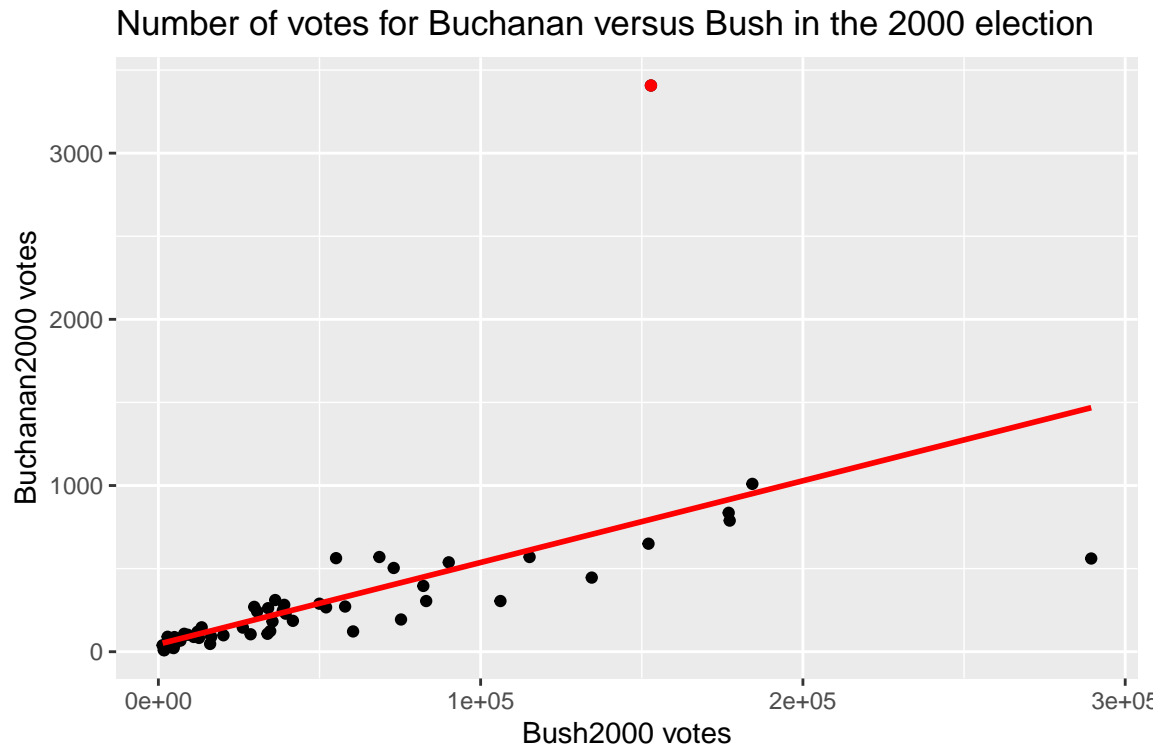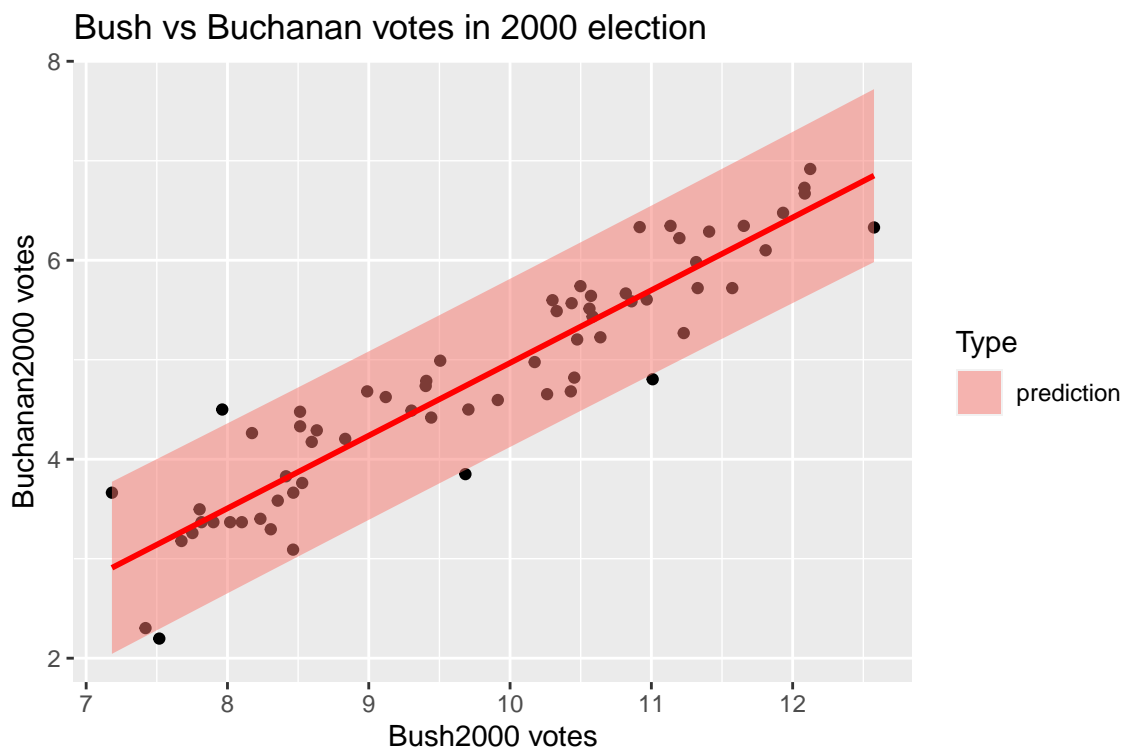
Figure 1: Scatterplot

Figure 2: 95% Prediction model

Another limitation with this data is that it only compares the number of votes for candidates Bush and Buchanan. With this set of data, we can't analyze any trends for votes for Al Gore. It is possible that Buchanan has connections to Palm Beach County, which could explain why he contained so many votes from that county, and it removes the possibility to explore whether or not other candidates had an unusually high number of votes in other counties. This limits the amount of interpretation that can be derived solely from this data set, and well as lessens the strength of our overall results if other trends have been omitted.

Appendix

The following is the R code used to reach my conclusions.

First, I created a scatterplot of the data, to visualize the outlier that is Palm Beach County.

```
GeorgeBush <- ex0825

ggplot(GeorgeBush, aes(x=Bush2000, y=Buchanan2000)) +
  geom_point() +
  geom_point(data=filter(GeorgeBush, County == "Palm Beach"), color="red") +
  geom_smooth(method="lm", se=FALSE, color="red") +
labs(x="Bush2000",
     y="Buchanan2000",
     title="Democratic absentee versus machine-counted ballots")
```

```
## `geom_smooth()` using formula 'y ~ x'
```
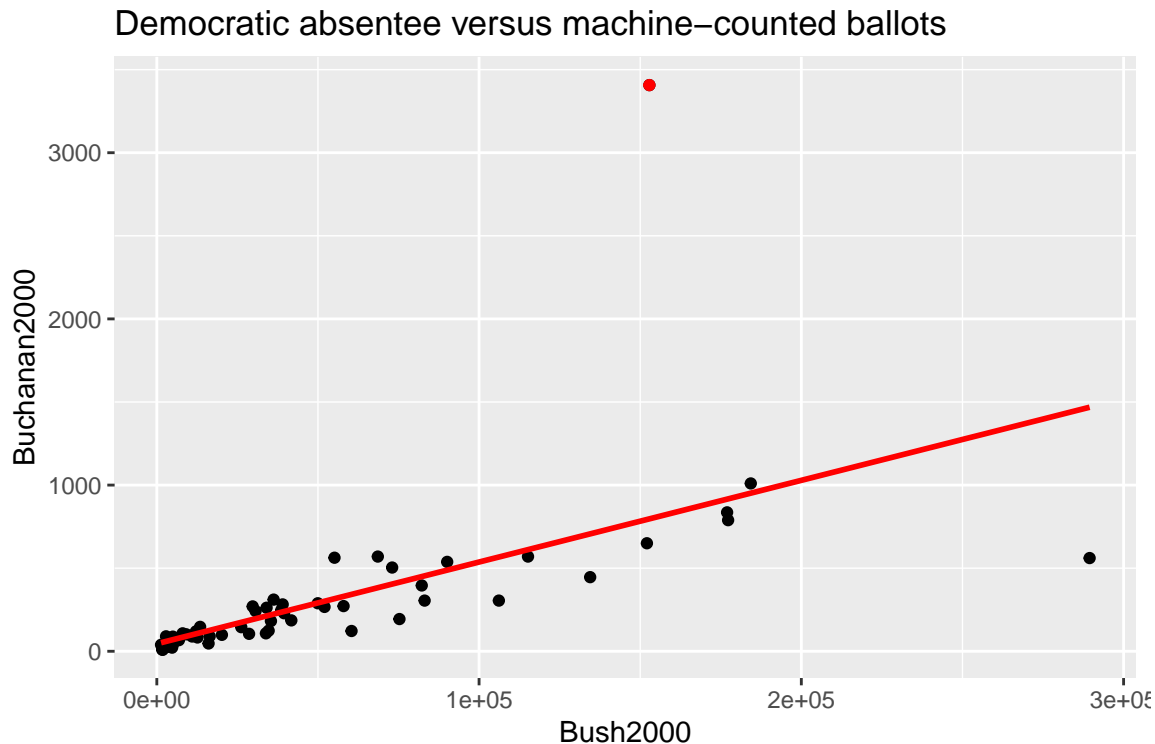

Then, after removing the outlier, I checked the residual plots to see if the data was linear. The first residual plot is not linear, as the constant variance assumption is not upheld. However, after taking the natural log of both the x and and the y axis, the transformed residual plot indicates that the data is now linear.

```
GeorgeBushNoPalm <- GeorgeBush[-67, ]
```
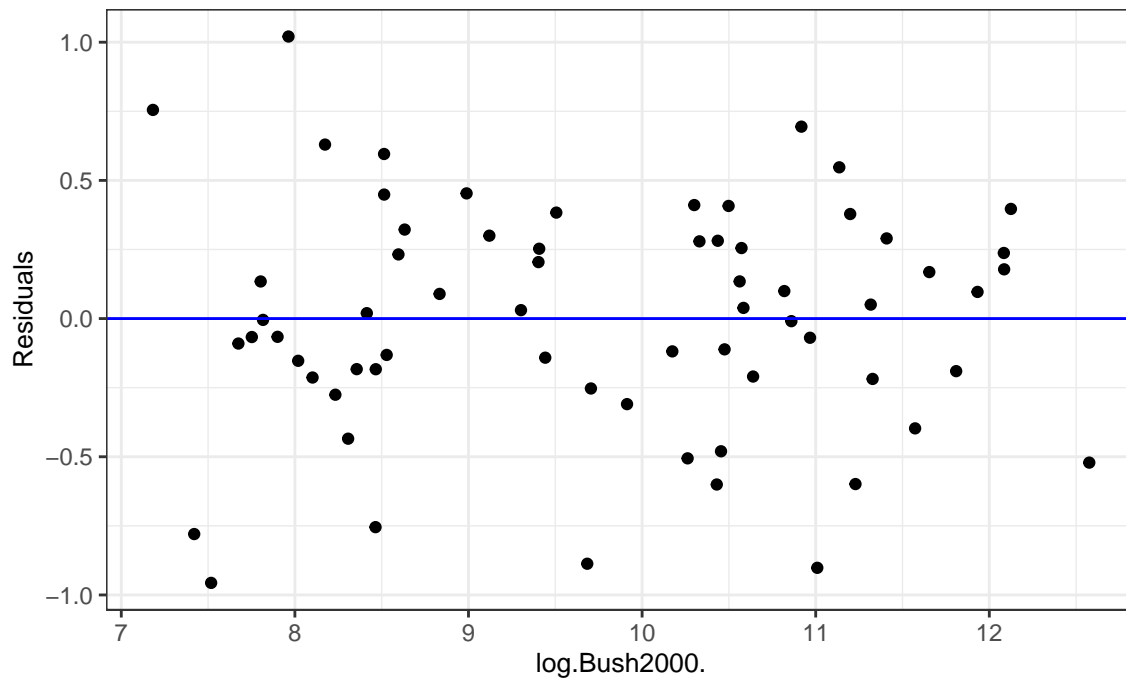
```
GeorgeBushNoPalm_lm <- lm(Buchanan2000 ~ Bush2000, data = GeorgeBushNoPalm)
resid_xpanel(GeorgeBushNoPalm_lm)
```

Plots of Residuals vs Predictor Variables
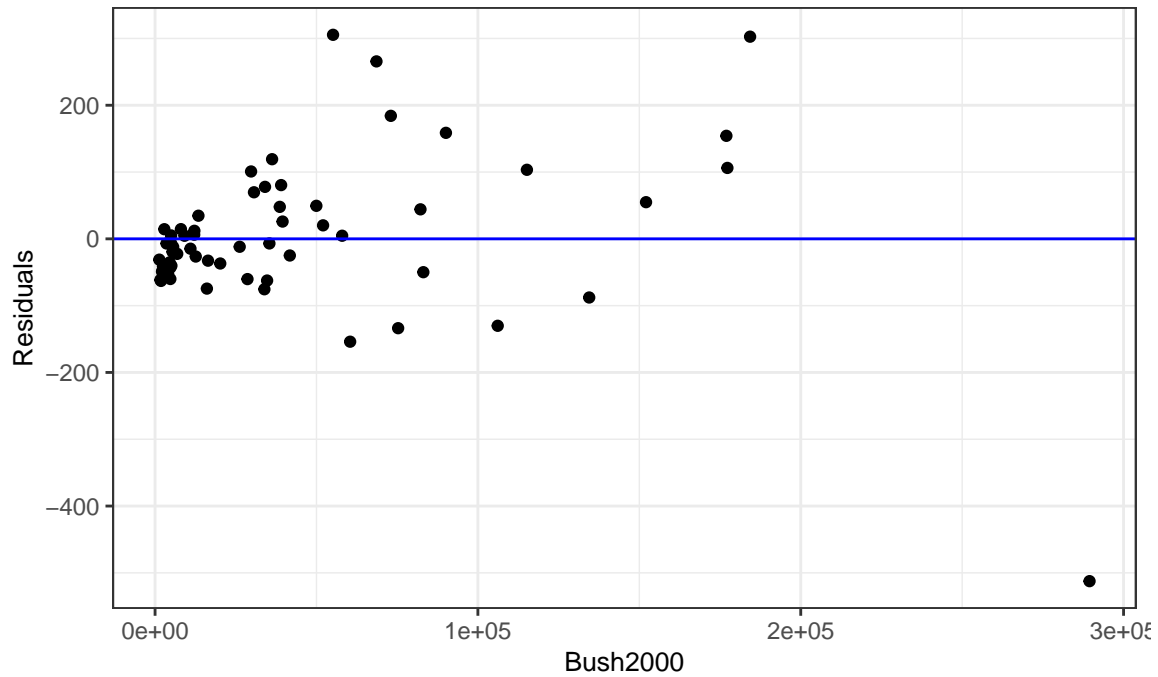

```
GeorgeBushNoPalm_lm <- lm(log(Buchanan2000) ~ log(Bush2000), data = GeorgeBushNoPalm)
resid_xpanel(GeorgeBushNoPalm_lm)
```

**Plots of Residuals vs Predictor Variables**

I then plotted the transformed model, and included the prediction bands associated with the data.

```
GeorgeBush_pred <- data.frame(GeorgeBushNoPalm,
                     predict(GeorgeBushNoPalm_lm, interval = "prediction"))
```

Warning in predict.lm(GeorgeBushNoPalm_lm, interval = "prediction"): predictions on current data ref

```
ggplot(GeorgeBush_pred, aes(x=log(Bush2000), y=log(Buchanan2000))) +
  geom_point() +
  geom_ribbon(  # add "ribbon" geom for bands
    aes(ymin = lwr,  # lower prediction bound at a given temp
        ymax = upr,  # upper prediction bound at a given temp
        fill = "prediction"),   # quick way to get a legend
    alpha = .5) +
  geom_smooth(method="lm", se=FALSE, color="red") +
labs(x="Bush2000 votes",
     y="Buchanan2000 votes",
     title="Bush vs Buchanan votes in 2000 election",
     fill = "Type")
```

`geom_smooth()` using formula 'y ~ x'
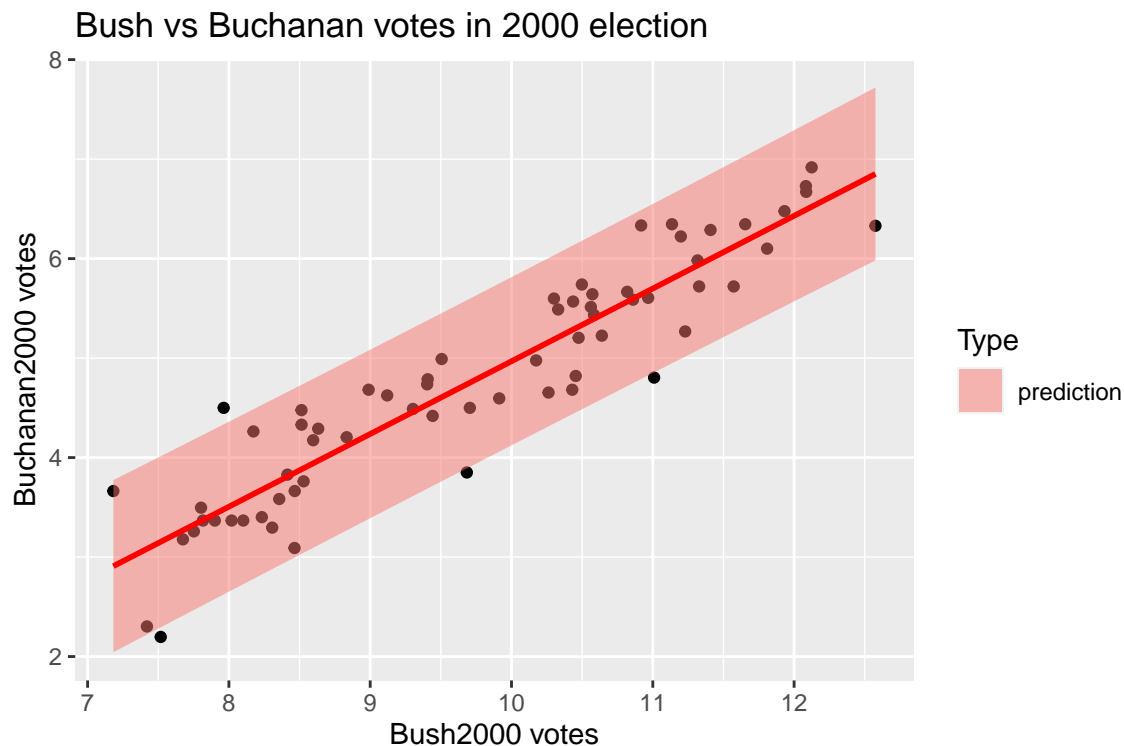
Then, I found the 95% prediction interval with the output from summary, and exponentiated the prediction interval to "unlog" the data to find the actual amount of votes we would have expected to observe for Buchanan in Palm Beach County.

```
predict(GeorgeBushNoPalm_lm,   # model object
        newdata = data.frame(Bush2000 = 152846),  # new data
        interval = "prediction",   # interval type
        se.fit = T)
```

```
## $fit
##        fit      lwr      upr
## 1 6.384143 5.524656 7.24363
```

```
## 
## $se.fit
## [1] 0.09416562
## 
## $df
## [1] 64
## 
## $residual.scale
## [1] 0.4198003
```

```
summary(GeorgeBushNoPalm_lm)
```

```
## 
## Call:
## lm(formula = log(Buchanan2000) ~ log(Bush2000), data = GeorgeBushNoPalm)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95631 -0.21236  0.02503  0.28102  1.02056
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.34149    0.35442  -6.607 9.07e-09 ***
## log(Bush2000)  0.73096    0.03597  20.323  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4198 on 64 degrees of freedom
## Multiple R-squared:  0.8658, Adjusted R-squared:  0.8637
## F-statistic:   413 on 1 and 64 DF,  p-value: < 2.2e-16
```

```
exp(6.384143)
```

```
## [1] 592.3768
```

```
exp(5.524656)
```

```
## [1] 250.8
```

```
exp(7.24363)
```

```
## [1] 1399.164
```

```
se.pred = sqrt(exp(0.09417)^2 + exp(0.4198)^2)
se.pred
```

```
## [1] 1.876882
```

Finally, I calculated how many votes were incorrectly allocated to Buchanan from Al Gore by subtracting the observed amount of votes for Buchanan from the tail values of the 95% prediction interval calculated above.

```
3407 - 1399
```

```
## [1] 2008
```

```
3407 - 250
```

```
## [1] 3157
```