Short Report 1: The Dramatic US Presidential Election of 2000

Jeanny Zhang

October 7, 2021

Introduction

This report is dedicated to investigating the possibility of a miscount in the dramatic US presidential election of 2000. After a long tug-of-war between the Democratic candidate Al Gore and the Republican candidate George Bush, Bush won the Florida state and ultiamtely the election by a margin so small that a recount was initiated. Meanwhile, voters were complaining that the structure of the ballot may lead to a confusion of the Democratic candidate Al Gore with the Reform Party candidate Pat Buchanan. These anomalies and oddities soon ballooned into a national controversy, and this report focuses on analyzing whether Buchanan received more votes than expected in Palm Beach County.

The dataset used in this report is from the Sleuth3 library with a sample size of 67. It contains data on the number of votes received by Buchanan and Bush in all counties of Florida. It has three variables: `County` (all counties in Florida), `Buchanan2000` (number of votes received by Buchanan), `Bush2000` (number of votes received by Bush). Since we are trying to check whether the observed number of votes for Buchanan match the expected value, a linear regression model was conducted fitting `Buchanan2000` on `Bush2000`, excluding the data point collected in Palm Beach County.

Results

A basic exploratory data analysis on the dataset was run first. Boxplots of values in column `Buchanan2000` (Figure 1) and column `Bush2000` (Figure 2) are shown below. The median of the number of votes received by Bush is 20196, which is higher than the median number of 114 votes received by Buchanan. The interquartile range of Bush's number of votes ranges from 4746 to 56542, which is also wider than that of Buchanan's number of votes ranging from 46.5 to 385.5. Both of the boxplots in Figure 1 and 2 appear to be right-skewed, but Buchanan's boxplot in Figure 1 seems more right-skewed than Bush's boxplot in Figure 2 with a longer right tail and an extreme outlier. The extreme outlier happens to be the data point collected in Palm Beach County with a value of 3407. This outlier is more obvious in the the scatterplot of votes for George W. Bush and Pat Buchanan in all Florida counties with the outlier having a y value over 3000 and the rest of the data points having y values within 1100.

Figure 1. Number of Votes Received by Buchanan

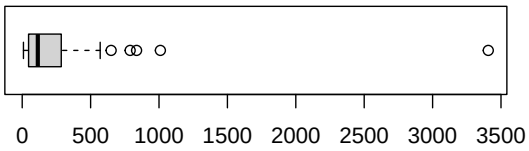
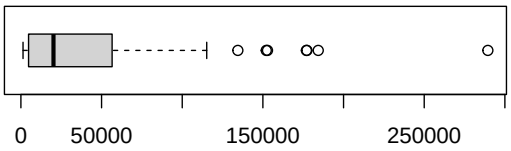


Figure 2. Number of Votes Received by Bush



To achieve linearity, log transformations were conducted on both the response variable and explanatory variable. A least squares regression line was generated fitting log10 of `Buchanan2000` on log10 of `Bush2000`, excluding the extreme outlier from Palm Beach County. The resulting regression model for the number of votes received by Buchanan is:

$$\hat{\mu}\{log_{10}(Buchanan2000) \mid log_{10}(Bush2000)\} = -1.0169 + 0.731\ log_{10}(Bush2000)$$

The model's estimated intercept is -1.0169 with a standard error of 0.154, and the model's estimated slope is 0.731 with a standard error of 0.0360. The estimated effect of Bush's votes on Buchanan's votes is statistically significant as the p-values of the intercept and the slope are both less than 0.05. The model has a r-squared

of 0.866 which suggests that about 86.6% of the variation in observed Buchanan's votes can be explained by the above regression.

The model also complies with the assumptions of a linear regression. The residual plot of the model displays no signs of violation of linearity and constant variance as the residuals are evenly distributed and there is no obvious curvature or outlier. Although the QQ plot of the model seems a bit s-shaped, aligning with the skewness shown in the boxplots, the sample size is big enough for us to overlook this violation of normality. The assumption of independence here is hard to be assessed as there is not enough information or data on the spatial, temporal, or other underlying relationship between each response and error. Overall, the linear model here is valid and it demonstrates a strong correlation between the number of votes received by Bush and that received by Buchanan.

Using the fitted linear regression model above, the predicted number of votes received by Buchanan in Palm Beach County is 593 when the number of votes received by Bush is 152846. The calculation procedure is shown below.

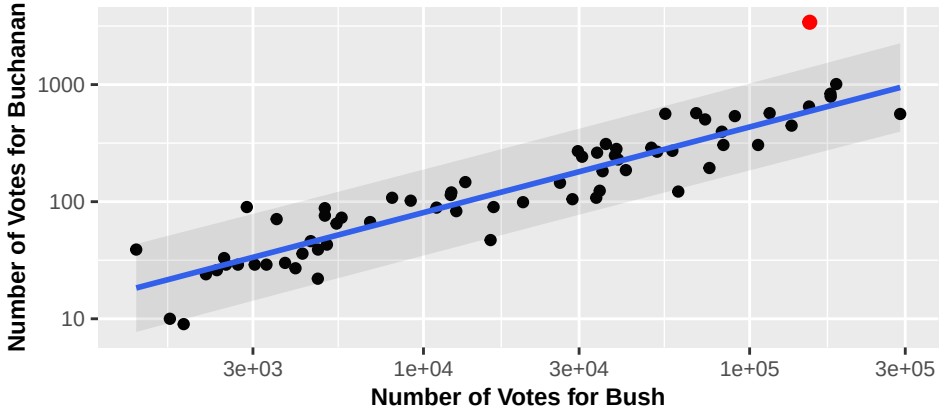
$$\hat{\mu}\{log_{10}(Buchanan2000) \mid log_{10}(Bush2000 = 152846)\} = -1.0169 + 0.731\ log_{10}(152846)$$

$$\hat{\mu}\{log_{10}(Buchanan2000) \mid log_{10}(Bush2000 = 152846)\} = 2.773$$

$$10^{2.773}$$

A scatterplot with the least squares regression line and a 95% prediction interval is shown in Figure 3. The controversial data point from the Palm Beach County is marked in red in the plot. As calculated, the prediction interval of Buchanan's vote count in Palm Beach County ranges from 251 ($10^{2.399}$) to 1400 ($10^{3.146}$) when Bush's vote count is 152846.

Figure 3. Regression of Buchanan2000 on Bush2000



Discussion

As shown in Figure 3, the observed number of votes received by Buchanan (3407) is way outside of the model's prediction interval (251 to 1400), suggesting that the observed value does not match with 95% of all future observable realizations. Therefore, Buchanan has received an unusually high number of votes in the Palm Beach County and his votes could actually contain a number of votes intended for Gore. Since the upper limit of the 95% prediction interval is 1400 and Buchanan's vote count is 3470, it is likely that around 2070 ($3470 - 1400$) votes were intended for Gore.

One of the limitations of this analysis would be the lack of assessment on the assumption of independence. There could potentially be a spatial correlation between counties and votes for Buchanan. Perhaps residents in Palm Beach County were trully supportive of Buchanan, or perhaps there are other comfounding variables we are not aware of. Unfortunately, there is not enough information to make any substantial conclusion for now.

R Code Appendix

```
# import libraries
library(Sleuth3)
library(ggplot2)
library(dplyr)
library(stargazer)
library(patchwork)

# get summary of the two quantitative columns
summary(ex0825$Buchanan2000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0    46.5   114.0   258.5   285.5  3407.0
```

```
summary(ex0825$Bush2000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1316    4746   20196   43356   56542  289456
```

```
# make boxplot for each quantitative column
par(mfcol = c(1, 2))
boxplot(ex0825$Buchanan2000, main = "Figure 1. Number of Votes Received by Buchanan", horizontal=TRUE)
boxplot(ex0825$Bush2000, main = "Figure 2. Number of Votes Received by Bush", horizontal=TRUE)
```

Figure 1. Number of Votes Received by Buchanan

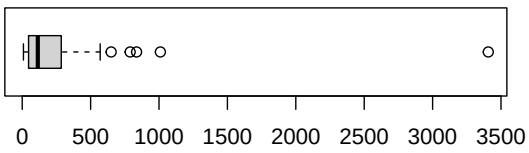
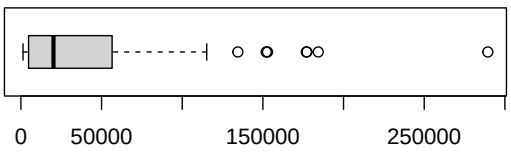


Figure 2. Number of Votes Received by Bush

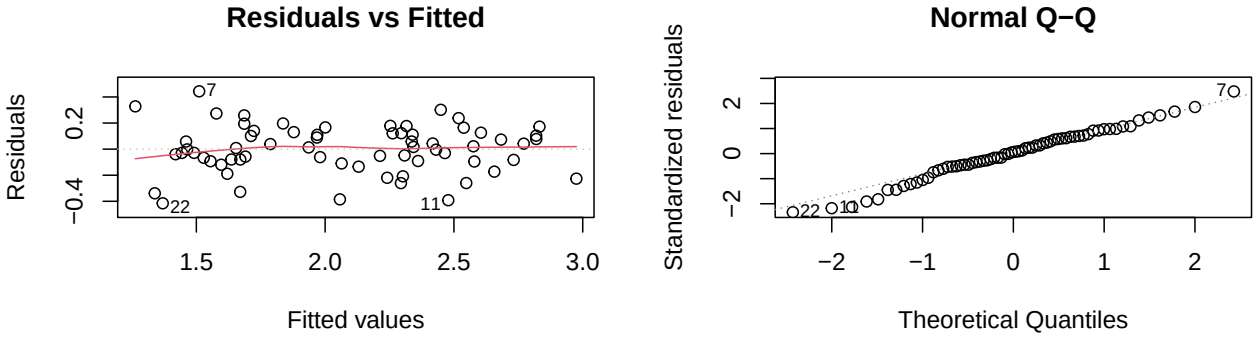



```
# filter the outlier
no_outlier <- filter(ex0825, Buchanan2000<3000)
# create SLR model
model <- lm(log10(Buchanan2000)~log10(Bush2000), data=no_outlier)
# get the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = log10(Buchanan2000) ~ log10(Bush2000), data = no_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41532 -0.09223  0.01087  0.12204  0.44323
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.01689    0.15392  -6.607 9.07e-09 ***
## log10(Bush2000)  0.73096    0.03597  20.323  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1823 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
```

```
# generate a residual plot and a QQ plot for the linear regression model
par(mfcol = c(1, 2))
plot(model, which=1, caption="", main="Residuals vs Fitted")
plot(model, which=2, caption="", main="Normal Q-Q")
```




```
# predict Buchanan2000 when Bush2000=152846
predict(model, data.frame(Bush2000=c(152846)), interval="prediction", se.fit=TRUE)
```

```
## $fit
##        fit      lwr      upr
## 1 2.772598 2.399328 3.145869
##
## $se.fit
## [1] 0.04089561
##
## $df
## [1] 64
##
## $residual.scale
## [1] 0.182317
```

```
# generate a prediction value dataframe
pred_df <- data.frame(no_outlier, predict(model, interval="prediction"))
# plot a scatterplot with the fitted least squares regression line, a 95% prediction band
# and the outlier
# the plot is not shown here because it is shown in Figure 3 above
ggplot(pred_df, aes(x=Bush2000, y=Buchanan2000)) +
  geom_point() +
  scale_y_log10() +
  scale_x_log10() +
  geom_point(data=filter(ex0825, Buchanan2000>3000), color="red", size=2) +
  geom_smooth(method="lm", se = FALSE) +
  geom_ribbon(aes(ymin=10^(lwr), ymax=10^(upr)), alpha=0.1) +
  theme(axis.text=element_text(size=8), axis.title=element_text(size=9,face="bold")) +
  labs(x="Number of Votes for Bush", y="Number of Votes for Buchanan", title="Figure 3. Regression of Bu
```