Butterfly Ballot Blunder: An Investigation of the 2000 Presidential Election

Dani Justo and Zoe Khan

2025-03-05

Introduction

The 2000 U.S. presidential race was a highly contested election between Al Gore and George W. Bush, with the latter eventually winning by a hair due to Florida voters. However, something strange was afoot: people noticed that the "butterfly" ballot design for Palm Beach county in Florida was associated with a very high vote percentage for another candidate, Pat Buchanan. It was theorized that the ballot design led people to inadvertently vote for Buchanan instead of Gore. Our goal in this report is to analyze the relationship between votes for Bush and Buchanan in order to predict how many votes were potentially miscast.

Data Description

Ask about what summary statistics should be included!!!

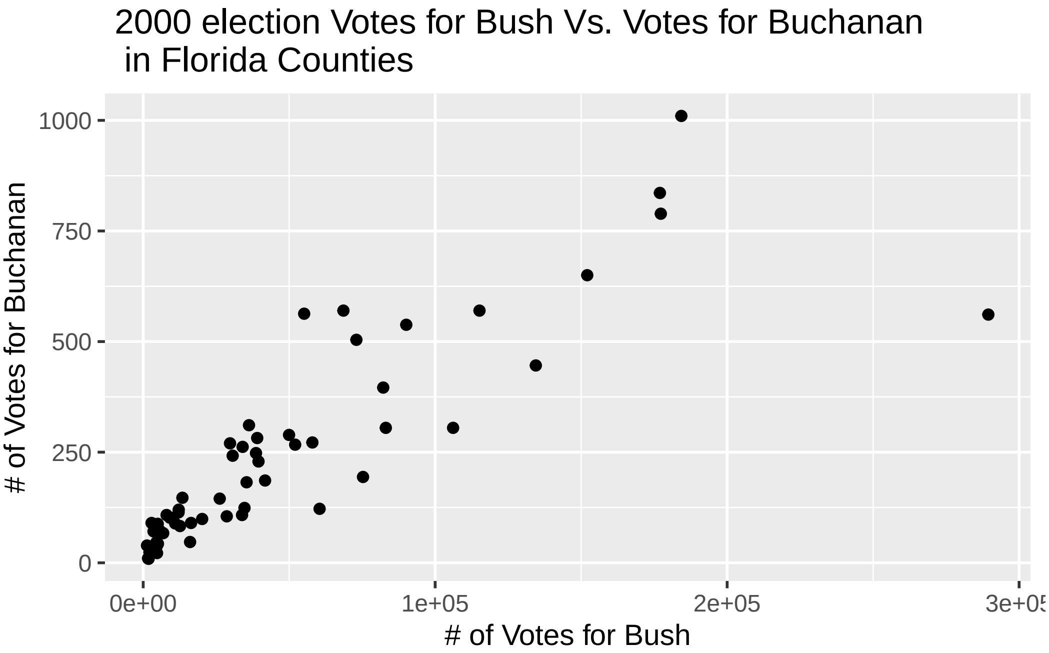
For our data analysis, we used the The "Dramatic U.S. Presidential Election of 2000" data set from the `Sleuth2` package. This data set includes the number of votes for Buchanan and Bush in all 67 counties in Florida during the 2000 U.S. presidential election. We want to see if the confusing "butterfly" layout of the ballot used in Palm Beach county caused the discrepancy in votes between Bush and Buchanan, so we've removed Palm Beach county from our data set. This allows us to observe relationship between Bush and Buchanan's votes under "normal" circumstances, namely, if the election really hadn't been affected by the ballot.

The summary statistics in the table below suggest that Bush received more votes than Buchanan in Florida on average.

Mean Bush	Mean Buchanan
41696.82	210.76

The scatter plot displayed below has a correlation coefficient of 0.867, suggesting a strong positive association between the number of votes for Bush and the number of votes for Buchanan. The

R^2 of the linear regression model is 0.862 indicates that the number of votes for Bush account for 86.2% of the variability in the number of votes for Buchanan.



Modeling Process

When we explored an initial linear model, we found violations of linearity, equal variance, and normality, so we decided to use a transformed model to rectify this issue. After some experimentation, we found that a double-log transformation satisfied conditions and would allow us to continue with our inquiry. Let $Bush_i$ denote the votes for Bush, and $Buchanan_i$ denote the votes for Buchanan. Our final transformed linear model is

$$E[log(Buchanan_i)|log(Bush_i)] = \beta_0 + \beta_1 log(Bush_i)$$

The null hypothesis is that there is no relationship between votes for Bush and Buchanan. The alternative hypothesis is there there is a relationship:

$$H_0 = \beta_1 = 0$$

$$H_A = \beta_1 \neq 0$$

The estimates and standard errors of the model parameters are below:

	Estimate	Std. Error	t value	P-value
(Intercept)	-2.34	0.35	-6.61	0
log(Bush2000)	0.73	0.04	20.32	0

We get a very small p-value ($\alpha = 0.05$), so we reject the null hypothesis and conclude that there is a relationship between votes for Bush and Buchanan. In our prediction interval, we are 95%

confident that an individual county with 152,846 votes for Bush would have between 250 and 1,394 votes for Buchanan. However, the true number of votes for Buchanan was 3,407, which is well out of range of the prediction interval. Taking the difference between the true estimate and the interval values, we might predict that between 1,763 and 3,157 votes were miscast.

Conclusions

From our analysis, we conclude that the number of votes for Buchanan in Palm Beach county was indeed discrepant, falling outside the range of values we might expect in an election where the butterfly ballot had not been used.

Our analysis is limited to the relationship between the number of votes for Bush and Buchanan, but does not account for how votes for may have been miscast for other candidates as a result of the butterfly ballot. Additionally, although we estimated that between 1,763 and 3,157 were miscast in favor Buchanan, our estimation is limited by a lack of information concerning the number of votes we could expect were intended for Buchanan, potentially allowing us to overestimate or underestimate the expected range of values for the number of miscast votes.

Because our study is observational, we can derive no causal effect.

There is a slight left skew in the distribution of the residuals in the normality plot, but it is slight so we were not concerned that this severely violates the normality condition.

R Appendix

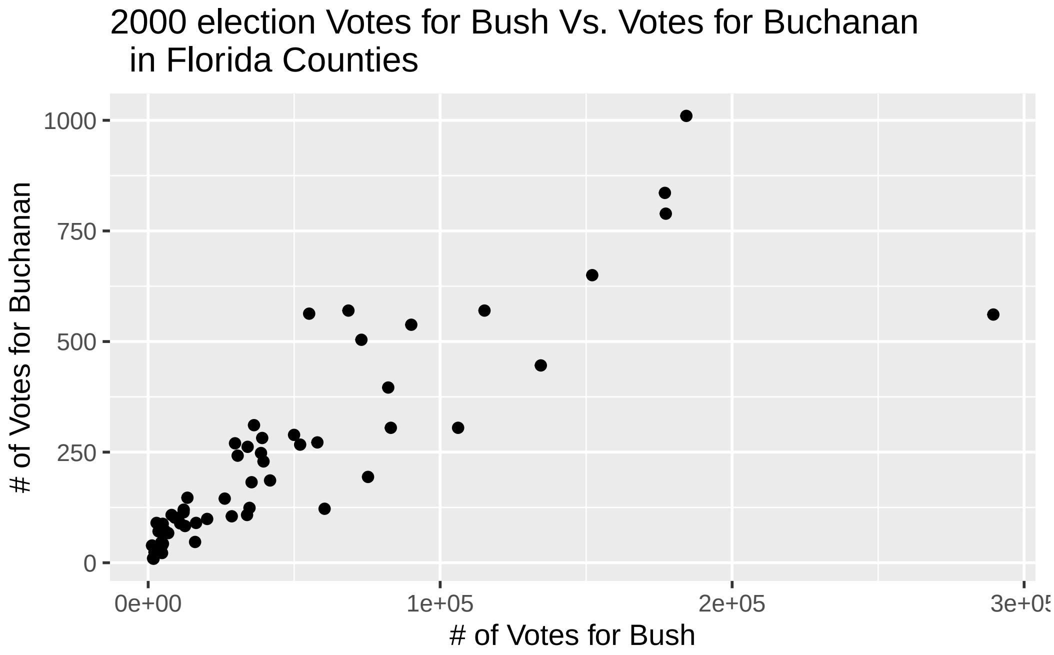
```r
# Import packages
library(tidyverse)
library(Sleuth2)
library(kableExtra)
library(performance)
library(broom)
library(equatiomatic)

# Loading the case study data
election <- Sleuth2::ex0825

# Creating a second data set with Palm Beach County excluded
election_wo_pb <- election |> filter(County != "Palm Beach")

# Display summary statistics
summary_statistics <- election_wo_pb |>
  summarize(mean_bush = mean(Bush2000),
            mean_buchanan = mean(Buchanan2000))
kable(summary_statistics)
```

mean_bush	mean_buchanan
41696.82	210.7576

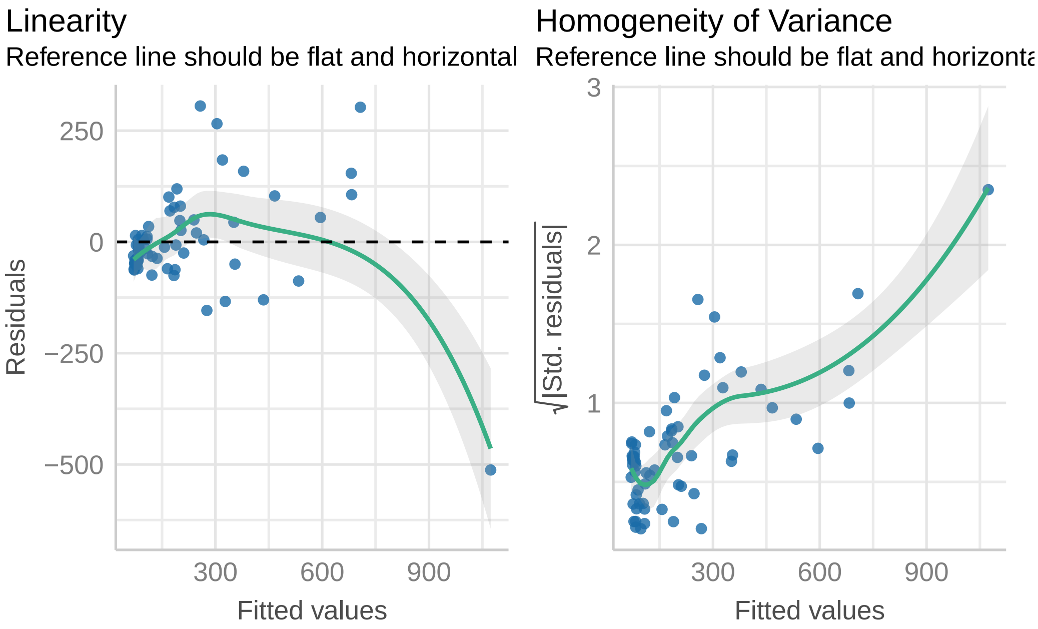
```r
# Initial model
election_model_bad <- lm(Buchanan2000 ~ Bush2000, data = election_wo_pb)
summary(election_model_bad)
```

```

Call:
lm(formula = Buchanan2000 ~ Bush2000, data = election_wo_pb)

Residuals:
    Min      1Q  Median      3Q     Max 
-512.43  -47.97  -17.09   41.78  305.45 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.557e+01  1.733e+01   3.784 0.000343 ***
Bush2000    3.482e-03  2.501e-04  13.923  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112.5 on 64 degrees of freedom
Multiple R-squared:  0.7518,	Adjusted R-squared:  0.7479 
F-statistic: 193.8 on 1 and 64 DF,  p-value: < 2.2e-16
```

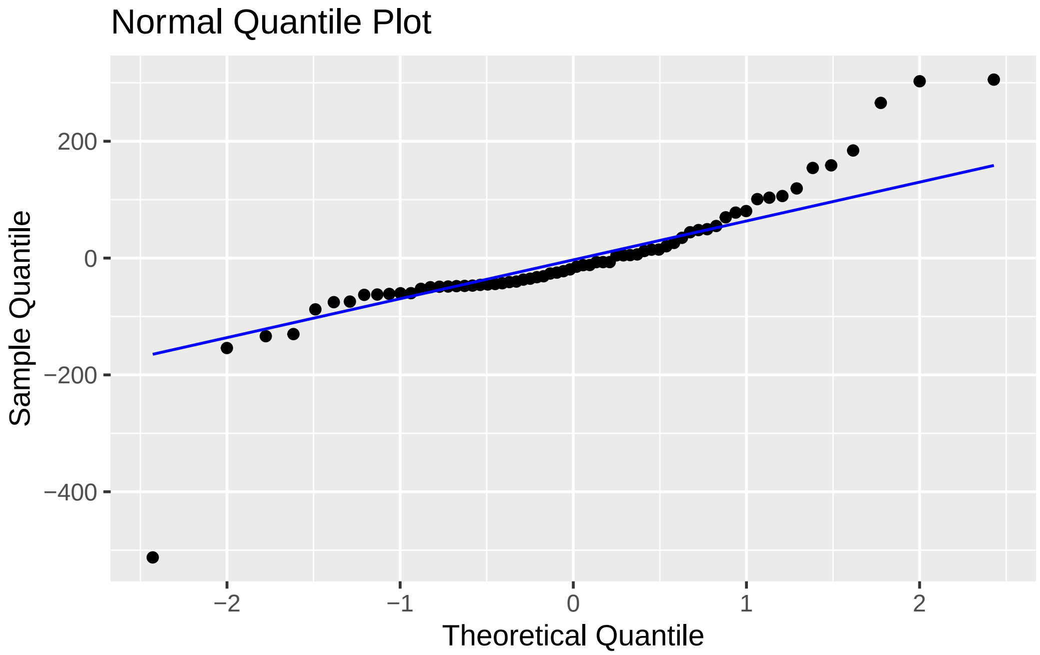
```r
# Display scatter plot of untransformed data
ggplot(data = election_wo_pb, aes(x=Bush2000,
                                  y=Buchanan2000))  +
  geom_point() +
  labs(title = "2000 election Votes for Bush Vs. Votes for Buchanan
  in Florida Counties",
       x = "# of Votes for Bush",
       y = "# of Votes for Buchanan")
```



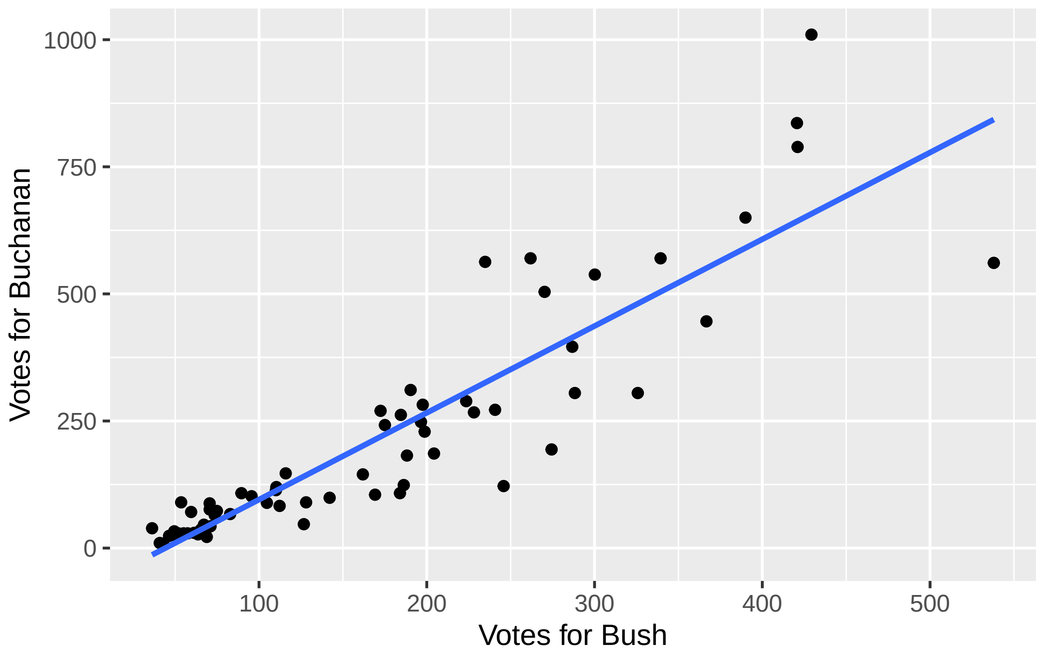
```
# Check conditions for initial model
check_model(election_model_bad, check = c("linearity", "homogeneity"))
```



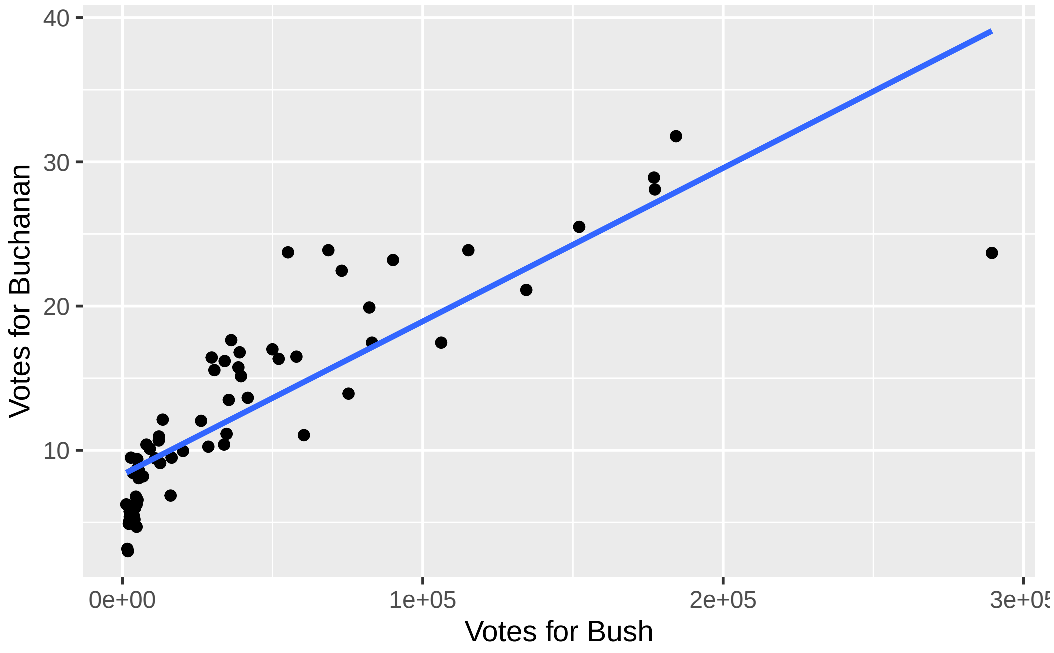
```
election_model_bad |> augment() |> ggplot(aes(sample = .resid)) +
  geom_qq()+
  geom_qq_line(col = "blue")+
  labs(title = "Normal Quantile Plot",
       x = "Theoretical Quantile",
       y = "Sample Quantile")
```



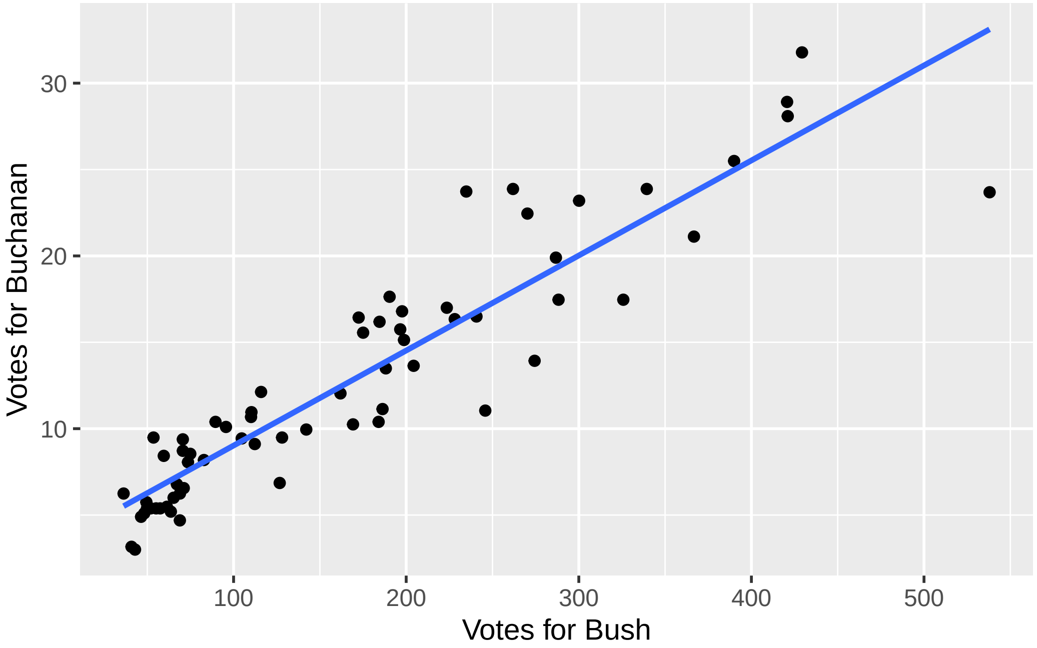
```
# Creating scatter plot sqrt transformed (x)
ggplot(data = election_wo_pb, aes(x = sqrt(Bush2000), y = Buchanan2000)) +
  geom_point() +
  labs(x = "Votes for Bush", y = "Votes for Buchanan") +
  geom_smooth(method = lm, se = FALSE)
```



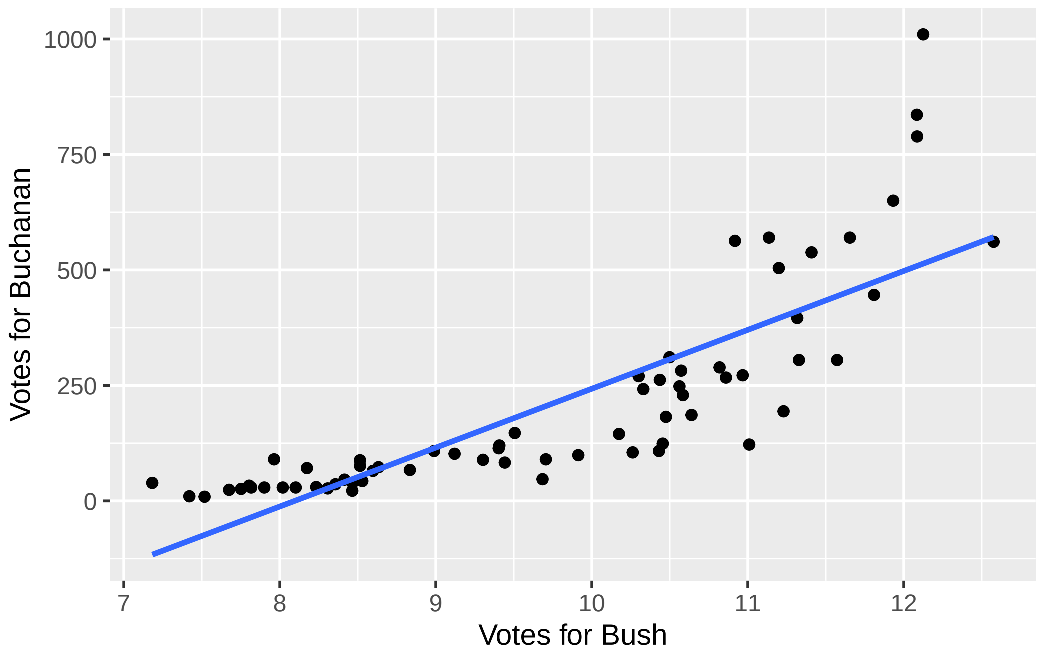
```
# Creating scatter plot sqrt transformed (y)
ggplot(data = election_wo_pb, aes(x = Bush2000, y = sqrt(Buchanan2000))) +
  geom_point() +
  labs(x = "Votes for Bush", y = "Votes for Buchanan") +
  geom_smooth(method = lm, se = FALSE)
```



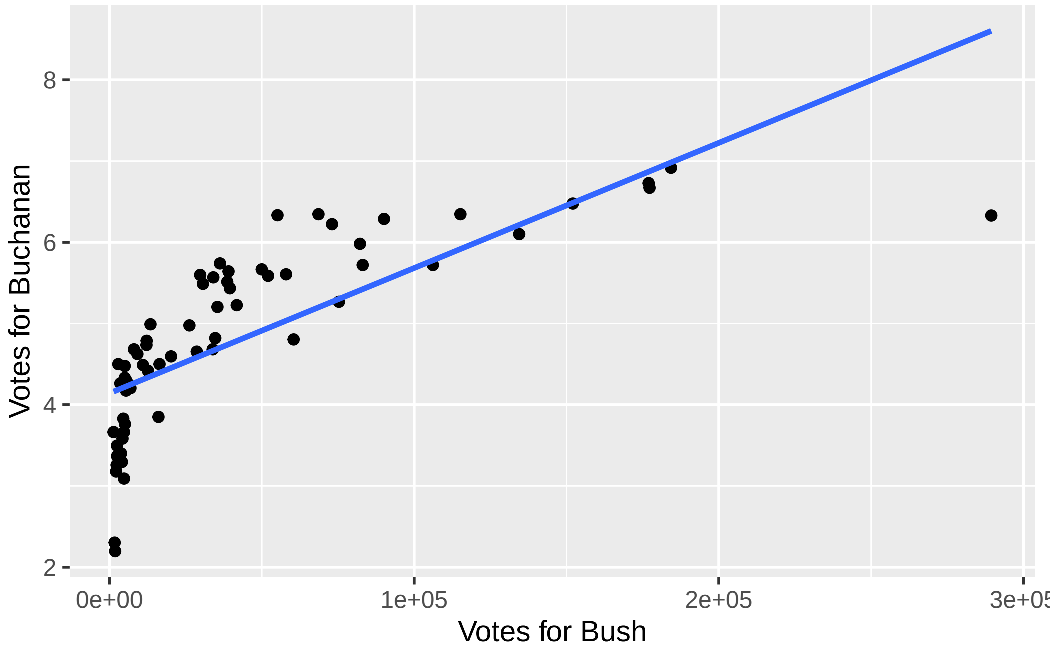
```
# Creating scatter plot sqrt transformed (both)
ggplot(data = election_wo_pb, aes(x = sqrt(Bush2000), y = sqrt(Buchanan2000))) +
  geom_point() +
  labs(x = "Votes for Bush", y = "Votes for Buchanan") +
  geom_smooth(method = lm, se = FALSE)
```



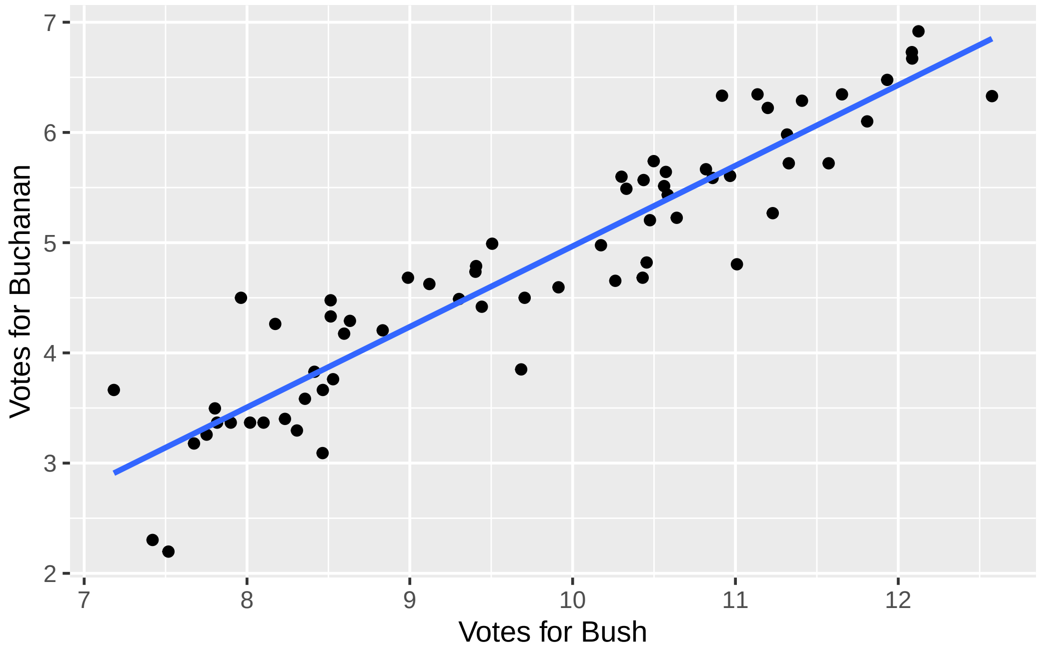
```
# Creating scatter plot log transformed (x)
ggplot(data = election_wo_pb, aes(x = log(Bush2000), y = Buchanan2000)) +
  geom_point() +
  labs(x = "Votes for Bush", y = "Votes for Buchanan") +
  geom_smooth(method = lm, se = FALSE)
```



```
# Creating scatter plot log transformed (y)
ggplot(data = election_wo_pb, aes(x = Bush2000, y = log(Buchanan2000))) +
  geom_point() +
  labs(x = "Votes for Bush", y = "Votes for Buchanan") +
  geom_smooth(method = lm, se = FALSE)
```



```
# Creating scatter plot log transformed (both) (this one wins!)
ggplot(data = election_wo_pb, aes(x = log(Bush2000), y = log(Buchanan2000))) +
  geom_point() +
  labs(x = "Votes for Bush", y = "Votes for Buchanan") +
  geom_smooth(method = lm, se = FALSE)
```

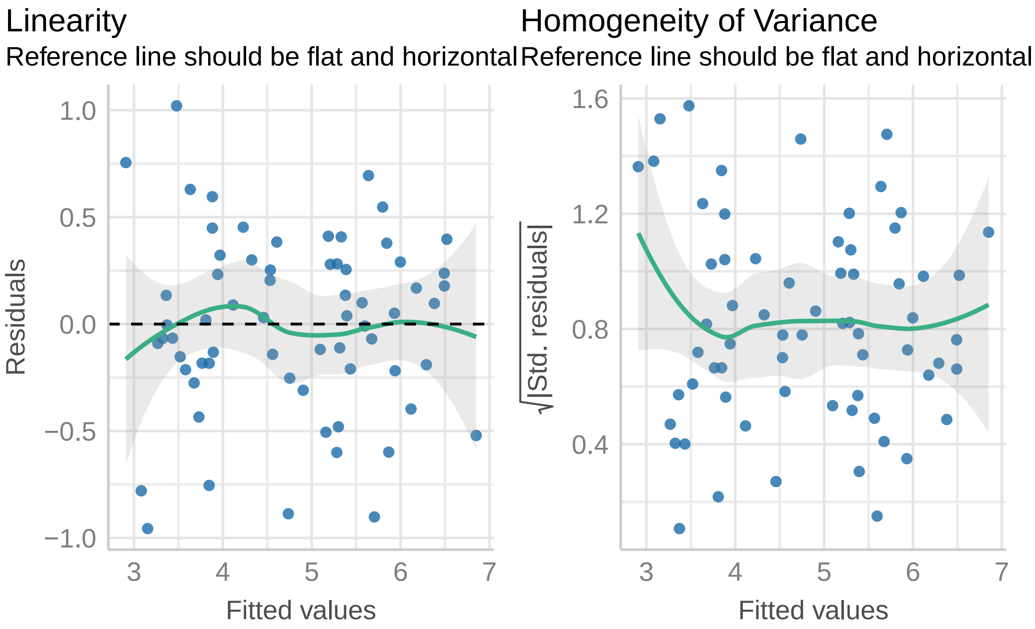


```
# Log Transformed Model
election_model <- lm(log(Buchanan2000) ~ log(Bush2000), data = election_wo_pb)
summary(election_model)
```

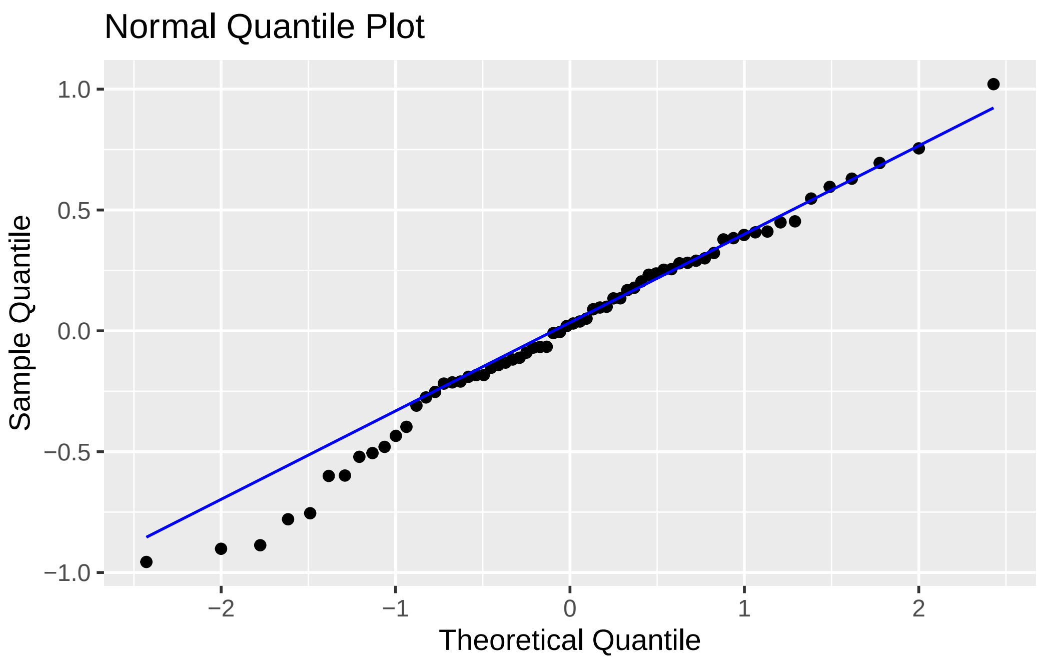
```

Call:
lm(formula = log(Buchanan2000) ~ log(Bush2000), data = election_wo_pb)

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
```

```
# Checking conditions for transformed model
check_model(election_model, check = c("linearity", "homogeneity"))
```



```
election_model |> augment() |> ggplot(aes(sample = .resid)) +
  geom_qq()+
  geom_qq_line(col = "blue")+
  labs(title = "Normal Quantile Plot",
       x = "Theoretical Quantile",
       y = "Sample Quantile")
```



```
# Making prediction interval
palm_beach <- data.frame(Bush2000 = c(152846))
augment(election_model, newdata = palm_beach, interval = "prediction")
```

```
# A tibble: 1 x 4
  Bush2000 .fitted .lower .upper
     <dbl>   <dbl>  <dbl>  <dbl>
1   152846    6.38   5.52   7.24
```

```
lower <- exp(5.52)
fitted <- exp(6.38)
upper <- exp(7.24)

equation <- extract_eq(election_model, use_coefs = TRUE)

# Extract coefficients from model
votes_table <- summary(election_model)$coefficients

# Display coefficients in (pretty) table
votes_table |> kbl(col.names = c("Estimate", "Std. Error", "t value", "P-value"),
↪   align = "c", booktabs = T, linesep="", digits = c(2, 2, 2, 4)) |>
↪   kable_classic(full_width = F, latex_options = c("HOLD_position"))
```

	Estimate	Std. Error	t value	P-value
(Intercept)	-2.34	0.35	-6.61	0
log(Bush2000)	0.73	0.04	20.32	0