Case Study 1: A Statistical Analysis of the Anomaly in the U.S. 2000 Presidential Election

Miya Dang, Mia Tran

2025-03-05

Introduction

The U.S. presidential election in 2000 between George W. Bush and Al Gore was one of the closest in history, with its final outcome dependent on the results in the state of Florida. Ultimately, Bush secured victory by a margin of fewer than 400 votes. However, Democratic voters in Palm Beach County argued that a confusing "butterfly" ballot layout led them to mistakenly vote for Reform Party candidate Pat Buchanan instead of Gore. This claim was supported by evidence showing that Buchanan received an unusually high percentage of votes in the county, along with a significant number of discarded ballots where voters had marked two choices.

In this case study, we will model the relationship between Bush and Buchanan votes in 2000 across Florida (excluding Palm Beach County to assess its anomaly) to answer two questions:

- *Was the number of votes for Buchanan in Palm Beach County statistically unusual compared to other Florida counties?*
- *What is the estimated number of votes intended for Gore but cast for Buchanan in Palm Beach County?*

Data Description

The dataset contains the vote counts for Buchanan and Bush in 2000 across all 67 counties in Florida.

Summary Statistics:

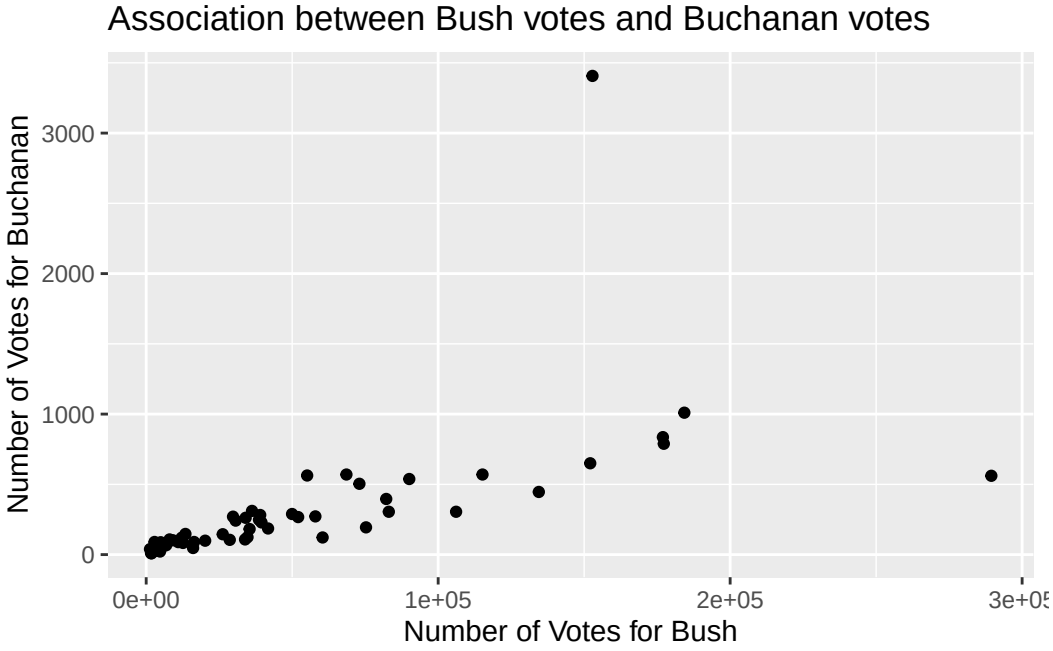
```
   Bush2000       Buchanan2000
 Min.   :  1316   Min.   :   9.0
 1st Qu.:  4746   1st Qu.:  46.5
 Median : 20196   Median : 114.0
 Mean   : 43356   Mean   : 258.5
 3rd Qu.: 56542   3rd Qu.: 285.5
 Max.   :289456   Max.   :3407.0
```

- Summary statistics indicate that, overall, Bush received significantly higher votes than Buchanan.

- Both variables have means much higher than their medians, indicating a strong right skew. This suggests that most counties had relatively low vote counts, while a small number had exceptionally high vote counts.
- The maximum number of Buchanan votes was 3,407, far exceeding the third quantile of 285.5. This vote count came from Palm Beach County, suggesting that it is an outlier.

Accompanying Graph:

The scatterplot below visualizes the relationship between Bush votes (x-axis) and Buchanan votes (y-axis) for all 67 counties:



There is a general positive trend, as counties with more Bush votes also tended to have more Buchanan votes. Palm Beach County (the mark on the top of the graph) is a notable outlier with unexpectedly high Buchanan votes (3,407). This graph shows that the data is highly left-skewed, raising concerns about potential violations of model assumptions, which we will examine in the next section.

Modeling Process and Results

Originally, we fitted a simple linear regression model for predicting Buchanan votes from Bush votes. Let *Buchanan* denote the votes for Buchanan in Palm Beach County and *Bush* denote the votes for Bush in Palm Beach County. Our original model is in the form:

$$Buchanan = \beta_0 + \beta_1 \left(Bush\right).$$

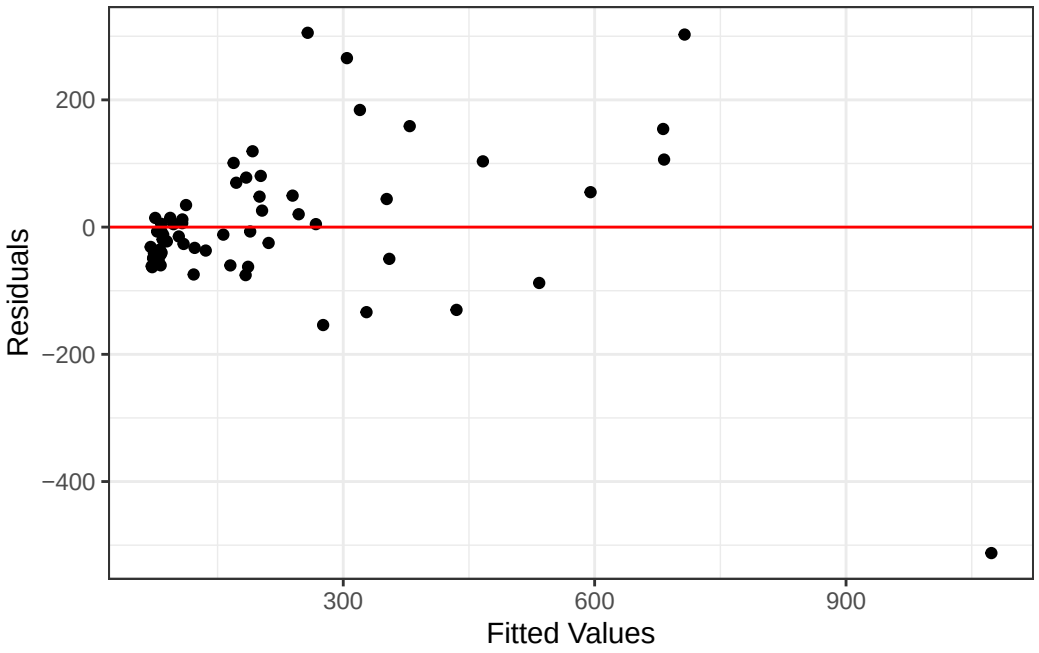
where:

- β_0 is the intercept of the model, representing the expected value of the number of votes for Buchanan when the number of votes for Bush is 0.

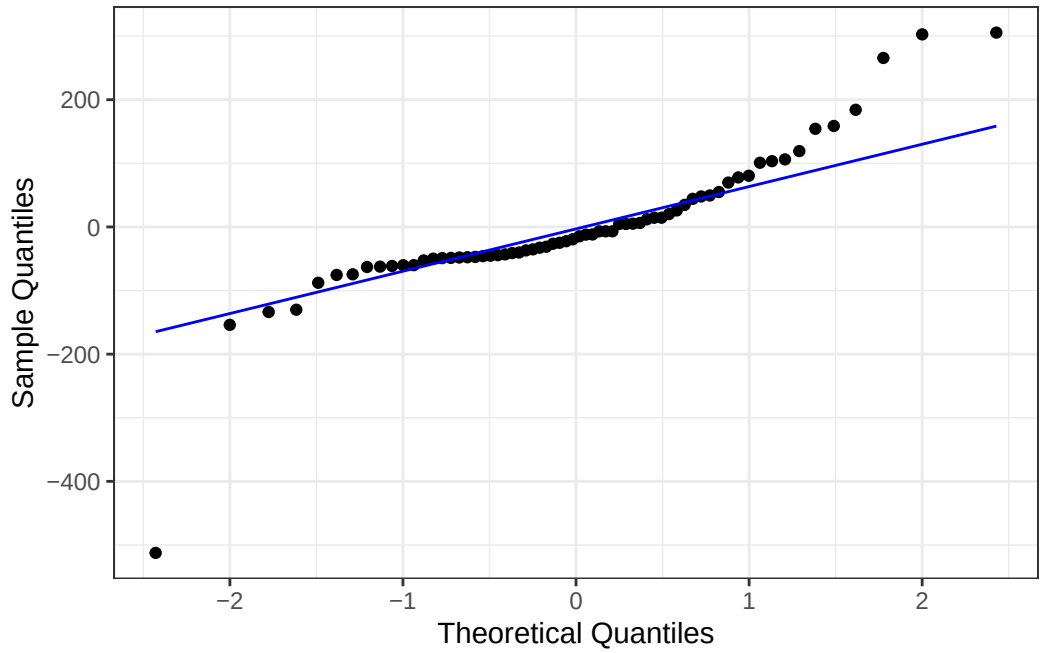
- β_1 represents the percentage change in Buchanan's votes for a 1% increase in Bush's votes. The table below shows the summary statistics of the simple linear regression model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	65.5735	17.3304	3.7837	0.00034
Bush2000	0.0035	0.0003	13.9226	0.00000

We proceeded to check the conditions for t-based inference with the simple linear regression model above. According to the data context, the Independence assumption is likely satisfied since the data consists of separate observations for each county in Florida (excluding Palm Beach County). Since each county reports its election results independently, the votes counted in one county do not directly influence the votes in another. Additionally, while counties may share some similarities, they operate as distinct political and administrative entities, each with its own local voting trends, demographics, and electoral processes, reducing the likelihood of systematic dependence between counties. The residual vs. fitted value plot below allows us to check the Linearity and Equal Variance conditions:



The residuals vs. fitted values plot indicates that both the Linearity and Equal Variance assumptions are violated. The clustering of residuals at the start of the line suggests a non-random pattern, violating the Linearity assumption. Additionally, the inconsistent vertical spread across fitted values indicates a lack of constant variance, suggesting a violation of the Equal Variance assumption.



The Q-Q plot suggests a violation of the Normality assumption, as the residuals deviate from the reference line not only at the extremes but also in the middle, indicating that they do not follow a normal distribution.

Since the Linearity, Equal Variance, and Normality assumptions are violated, our simple linear regression model is not appropriate for modeling the relationship between Buchanan's votes and Bush's votes. To address this, we applied a logarithmic transformation to the model.

Let $log(Buchanan)$ denote the natural logarithm of the votes for Buchanan in Palm Beach County and $log(Bush)$ denote the natural logarithm the votes for Bush in Palm Beach County. Our transformed linear regression model is:

$$E[log(Buchanan)|log(Bush)] = \beta_0 + \beta_1 log(Bush).$$

where:

- β_0 is the intercept of the model, representing the expected value of $log(Buchanan)$ when $log(Bush) = 0$. Since $log(1) = 0$, this means β_0 represents the expected value of $log(Buchanan)$ when the numbers of votes for Bush is 1.
- β_1 represents the elasticity of Buchanan's votes with respect to Bush's votes. A 1% increase in Bush's votes is associated with a β_1% change in Buchanan's votes.

The table below shows the summary statistics of the transformed linear regression model:

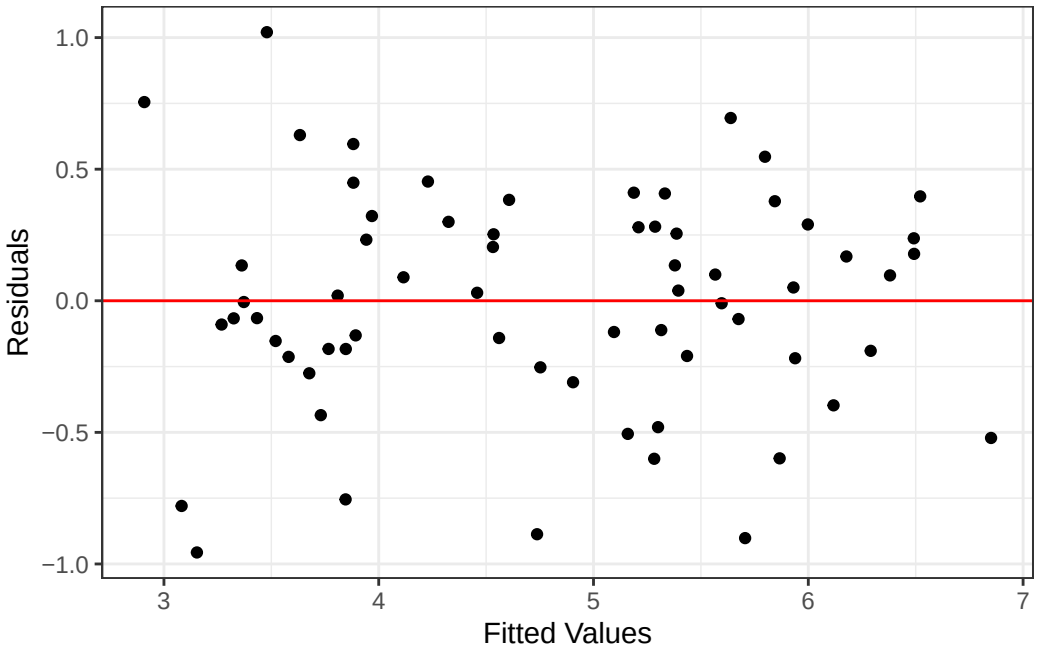
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.3415	0.3544	-6.6066	0
log(Bush2000)	0.7310	0.0360	20.3229	0

From the table, our transformed linear regression model can be written as:

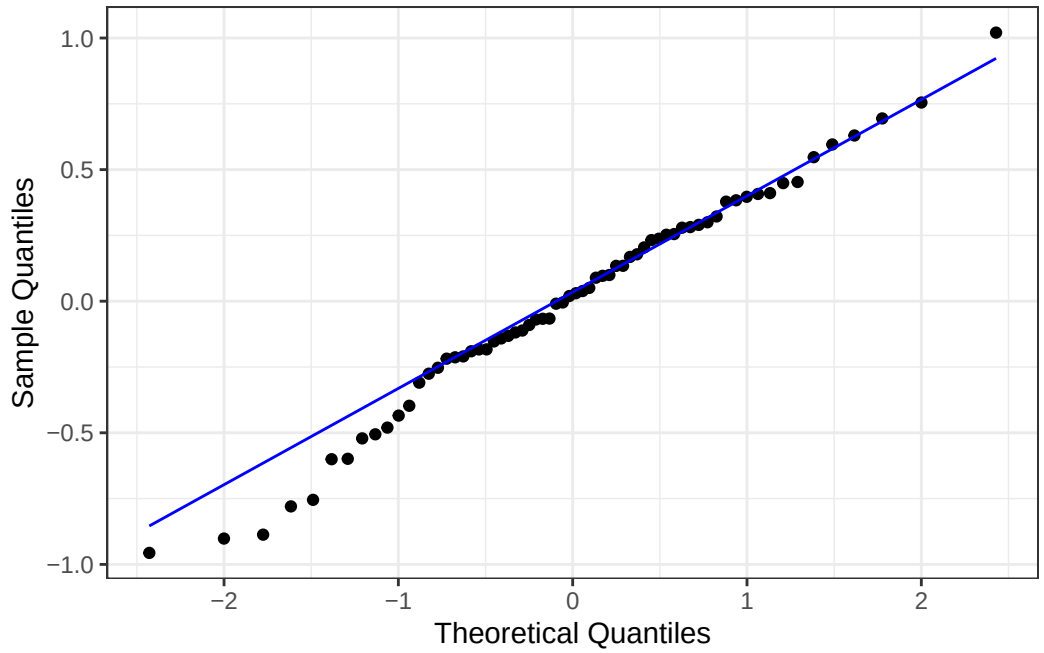
$$E[log(Buchanan)|log(Bush)] = -2.34 + 0.73log(Bush).$$

The standard error for the intercept and the slope is 0.354 and 0.036 respectively according to the table.

We proceeded to check the conditions for t-based inference with the transformed regression model above. The residual vs. fitted value plot below allows us to check the Linearity and Equal Variance conditions:



The residuals vs. fitted values plot indicates that both the Linearity and Equal Variance assumptions are now satisfied. The residuals scatter in random pattern around the $y = 0$, satisfying the Linearity assumption. Additionally, the fitted values follow a more consistent vertical spread, indicating constant variance and satisfying the Equal Variance assumption.



The Q-Q plot suggests a significant improvement in the normality assumption. While the residuals slightly deviate from the reference line at the beginning, they align more closely in the middle, indicating that they approximately follow a normal distribution.

We obtained a 95% prediction interval for the number of predicted Buchanan votes in Palm Beach County using this transformed fitted model:

Fitted	Lower	Upper
592	251	1399

The table above shows that the prediction interval is $[251, 1399]$. We are 95% confident that the votes for Buchanan when there are 152,846 votes for Bush (which is the number of votes for Bush in Palm Beach County) will fall between 251 votes and 1,399 votes.

Buchanan's real vote from Palm Beach County in 2000 election is 3,407 votes. Compared to the prediction interval above, we can calculate that an estimate for the likely number of votes intended for Gore but cast for Buchanan in Palm Beach County falls between $[2008, 3156]$.

Summary and Conclusions

Key Findings

- The Buchanan vote count in Palm Beach County during the 2000 election was statistically unusual.
- An estimated 2,008 to 3,156 votes intended for Gore were likely miscast for Buchanan.

- These results suggest that the ballot design may have altered the election outcome. Given that Gore lost Florida by fewer than 400 votes, the potential miscast votes in Palm Beach County alone far exceed this margin.

Limitations

- The log-log transformation assumes a multiplicative relationship between Bush and Buchanan votes. While this transformation improved linearity, residual diagnostics revealed mild heteroscedasticity and a slightly left-skewed residual distribution. These imply that the prediction interval may underestimate uncertainty, particularly for counties with high Bush vote count like Palm Beach.
- The model assumes that voter behavior is similar across all counties. However, many other factors could contribute to regional differences between voter behavior, such as candidate campaigns or other demographic factors.
- The analysis demonstrate association and not causation. Further investigation would be necessary to definitively prove that the butterfly ballot caused the excess Buchanan votes.

R Appendix

```r
# Loading necessary packages
library(tidyverse)
library(Sleuth2)
library(broom)
library(kableExtra)

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

# Summary statistics
summary(election %>% select(Bush2000, Buchanan2000))

# Creating a scatterplot for the relationship between Bush and Buchanan votes
election |>
  ggplot(aes(x = Bush2000,
             y = Buchanan2000)) +
  geom_point() +
  ggtitle("Association between Bush votes and Buchanan votes") +
  xlab("Number of Votes for Bush") + ylab("Number of Votes for Buchanan")

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

# Fitting the regression line for mean Buchanan votes
# as a function of Bush votes
election_lm <- lm(Buchanan2000 ~ Bush2000, data = election_wo_pb)
```

```r
# Creating the table summarizing the estimated coefficients of the model
# and their corresponding standard errors
election_lm_table <- summary(election_lm)$coefficients
election_lm_table |> kbl(col.names = c("Estimate", "Std. Error",
                                       "t value", "Pr(>|t|)"),
                         align = "c",
                         booktabs = T,
                         linesep="",
                         digits = c(4, 4, 4, 4)) |>
  kable_classic(full_width = F, latex_options = c("HOLD_position"))

# Creating the residuals-fitted plot to check
# Linearity and Equal Variance for the original election model
election_lm |>
  augment() |>
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, col = "red") +
  xlab("Fitted Values") +
  ylab("Residuals") +
  theme_bw()

# Creating the Q-Q plot as check Normality for the original election model
election_lm |>
  augment() |>
  ggplot(aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line(col = "blue") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles") +
  theme_bw()

# Transformation
transformed_election_lm <- lm(log(Buchanan2000) ~ log(Bush2000), election_wo_pb)

# Creating the table summarizing the estimated coefficients of the model
# and their corresponding standard errors
transformed_election_lm_table_1 <- summary(transformed_election_lm)$coefficients
transformed_election_lm_table_1 |> kbl(col.names = c("Estimate", "Std. Error",
                                       "t value", "Pr(>|t|)"),
                         align = "c",
                         booktabs = T,
                         linesep="",
                         digits = c(4, 4, 4, 4)) |>
  kable_classic(full_width = F, latex_options = c("HOLD_position"))
```

```r
# Creating the residuals-fitted plot to check
# Linearity and Equal Variance for the transformed election model
transformed_election_lm |>
  augment() |>
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, col = "red") +
  xlab("Fitted Values") +
  ylab("Residuals") +
  theme_bw()

# Creating the Q-Q plot as check Normality for the transformed election model
transformed_election_lm |>
  augment() |>
  ggplot(aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line(col = "blue") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles") +
  theme_bw()

# Filter the original dataset to only Palm Beach Country
# to create prediction interval
election_pb <- election |> filter(County == "Palm Beach")

# Creating a 95% prediction interval for the number of
# predicted Buchanan votes in Palm Beach County
# after knowing the votes for Bush from Palm Beach County
# and transform back the interval
new_election <- data.frame(Bush2000 = 152846)
transformed_election_lm_table_2 <- transformed_election_lm |>
  augment(newdata = new_election,
          interval = "prediction",
          conf.level = 0.95) |>
  select(c(".fitted", ".lower", ".upper")) |> exp()

# Creating the table summarizing the converted prediction interval
# of the transformed model
transformed_election_lm_table_2 |> kbl(col.names = c("Fitted", "Lower",
                                       "Upper"),
                         align = "c",
                         booktabs = T,
                         linesep="",
                         digits = c(0, 0, 0)) |>
  kable_classic(full_width = F, latex_options = c("HOLD_position"))
```