Your Informative Title Here

Case Study 1

Radiah Khan Catherine Weeks Mercer Mercer

2025-10-06

Introduction

We will introduce our research questions, background of the problem and the motive behind anayzing these questions.

Background

During the US presidential election of 2000, in Palm Beach County, Florida there was a confusing ballot that could have potentially influenced the election results for this county. During the tight race between Gore and Bush, Palm Beach County had an unexpectedly high number of votes for Buchanan, from the Reform party. This may be due to a confusing ballot in this county, where if not read closely, someone attempting to cast a vote for Gore could have mistakenly filled in the circle for Buchanan.

Research Questions

In this case study, we aim to answer three questions:

- Is there a significant relationship between Bush and Buchanan votes within Palm Beach County?
- How does the relationship between Bush and Buchanan votes differ across other Florida counties?
- How many miscast votes were there for Buchanan in Palm Beach County?

Motivation behind the research questions

We want to determine if the ballot in the Palm Beach county lead to people mistakenly casting votes for Buchanan when they meant to cast the vote for Gore.

Data Description

In this section, we will introduce the variables we are going to use for this case study. Later, we are going to perform an exploratory data analysis to see the distribution of votes for both Buchanan and Bush.

Variables Used

- `County` : Names of the county in Florida during the 2000s election
- `Buchanan2000` : Number of votes for Buchanan in the specific county in Florida during that election
- `Bush2000` : Number of votes for Bush in the specific county in Florida during that election

Exploratory Data Analysis

```
#creating a new variable that only contains Buchanan's election data
election_summary_buchanan <- election|>
  select(Buchanan2000)

#printing out the first 5 rows of the Buchanan table
head(election_summary_buchanan, 5)
```

```
  Buchanan2000
1          262
2           73
3          248
4           65
5          570
```

```
#creating a new variable that only contains Bush's election data
election_summary_bush <- election|>
  select(Bush2000)

#printing out the first 5 rows of the Bush table
head(election_summary_bush, 5)
```

```
  Bush2000
1    34062
2     5610
3    38637
4     5413
5   115185
```

```r
#Saving summary table of Buchanan election data
election_count_summary_buchanan <-summary(election_summary_buchanan)

#printing the summary table created above with captions.
kable(as.data.frame(election_count_summary_buchanan,
      caption = "Summary of Vote Counts (Buchanan) in the 2000s Election"))
```

Var1	Var2	Freq
	Buchanan2000	Min. : 9.0
	Buchanan2000	1st Qu.: 46.5
	Buchanan2000	Median : 114.0
	Buchanan2000	Mean : 258.5
	Buchanan2000	3rd Qu.: 285.5
	Buchanan2000	Max. :3407.0

```r
#Saving summary table of Bush election data
election_count_summary_bush <-summary(election_summary_bush)

#printing the summary table created above with captions.
kable(as.data.frame(election_count_summary_bush,
      caption = "Summary of Vote Counts (Bush) in the 2000s Election"))
```

Var1	Var2	Freq
	Bush2000	Min. : 1316
	Bush2000	1st Qu.: 4746
	Bush2000	Median : 20196
	Bush2000	Mean : 43356
	Bush2000	3rd Qu.: 56542
	Bush2000	Max. :289456

```r
#Creating a new data set that is better for EDA
election_new <- election |>

  #pivoting longer to make separate box plots later
  pivot_longer(cols = c("Buchanan2000", "Bush2000"),
               names_to = "Candidate",
               values_to = "Count") |>
  mutate(Candidate = recode(Candidate,
                            "Bush2000" = "Bush",
                            "Buchanan2000" = "Buchanan"))

#printing out first 5 lines of new data frame
head(election_new, 5)
```

```
# A tibble: 5 x 3
  County  Candidate Count
  <chr>   <chr>     <dbl>
1 Alachua Buchanan    262
2 Alachua Bush      34062
3 Baker   Buchanan     73
4 Baker   Bush       5610
5 Bay     Buchanan    248
```

```r
#creating a box plot with the pivoted data
ggplot(data = election_new,
       aes(x = Count, fill = Candidate))+
  geom_boxplot() +

  #adding a title and x label
  labs(title ="Vote Counts for Bush and Buchanan in Florida",
       x = "Vote Counts (In log scale)") +

  #creating multiple plots based on the candidate
  facet_wrap(~Candidate)+

  #scaling the x axis to log10 to make plot more legible
  scale_x_log10()+

  #making the plot pretty :)
  theme_minimal()+
  theme(plot.margin = margin(10,10,10,10))
```

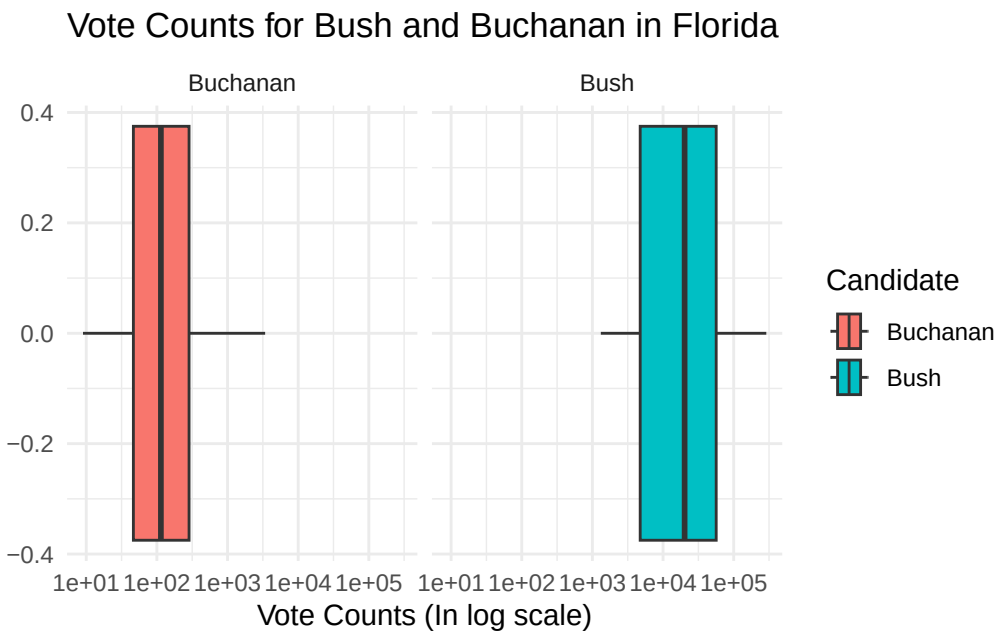


Figure 1

From the Figure 1, these box plots show the distribution of Bush and Buchanan votes in the different Florida counties. It is worth noting that the x axis has been changed to be log10, so the differences between the two plots are more significant. From this we can gather that Bush had an overall greater mean count of votes by a significant margin as well as a more consistent voter base across all counties. Meanwhile, Buchanan had less consistent voters, with some counties having much higher counts of votes than others, all while still maintaining lower vote counts than Bush overall.

Analysis Model

We will use a simple linear regression model to determine the relationship between the votes for Bush and the votes for Buchanan.

Assuming conditions are met for linear regression model.

```
ggplot(election, aes(x = Bush2000, y = Buchanan2000)) +
  geom_point() +
  labs(title = "Buchanan vs Bush 2000",
       x = "Bush",
       y = "Buchanan")+
  theme_minimal()
```

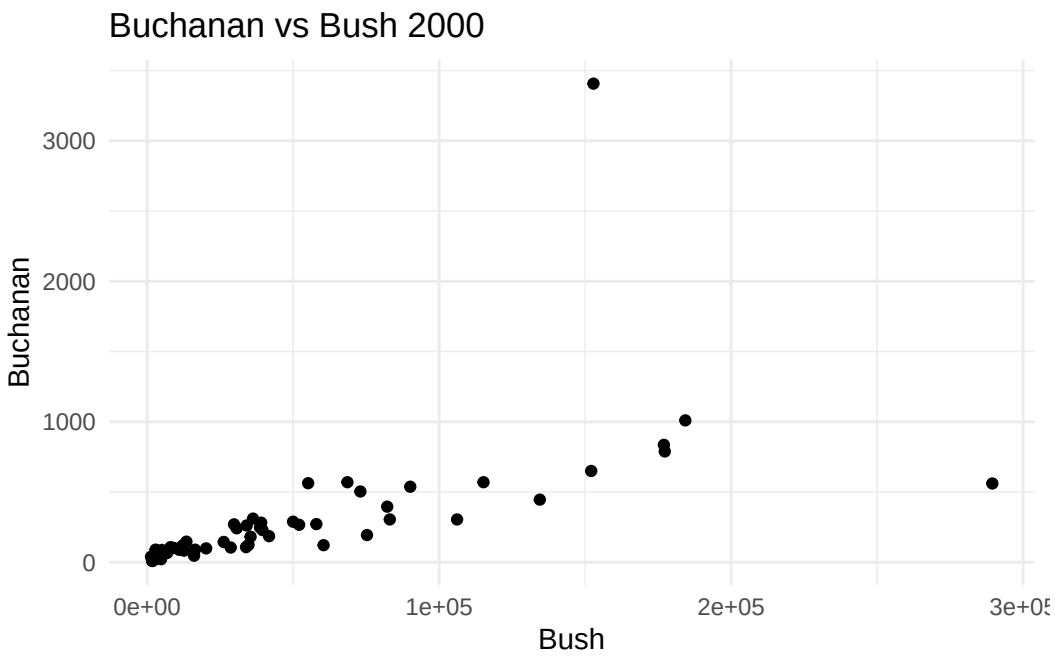


Figure 2

```
election_lm_pre <- lm((Buchanan2000) ~(Bush2000), data = election)
election_lm_pre
```

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

Coefficients:
(Intercept)     Bush2000
  45.289861     0.004917
```

```
election_summary <- tidy(election_lm_pre)
kable(election_summary, caption = "Regression summary Before Transformation")
```

Table 3: Regression summary Before Transformation

term	estimate	std.error	statistic	p.value
(Intercept)	45.2898613	54.4794230	0.8313205	0.4088361
Bush2000	0.0049168	0.0007644	6.4319610	0.0000000

```
plot(fitted(election_lm_pre), residuals(election_lm_pre))
abline(0,0)
```

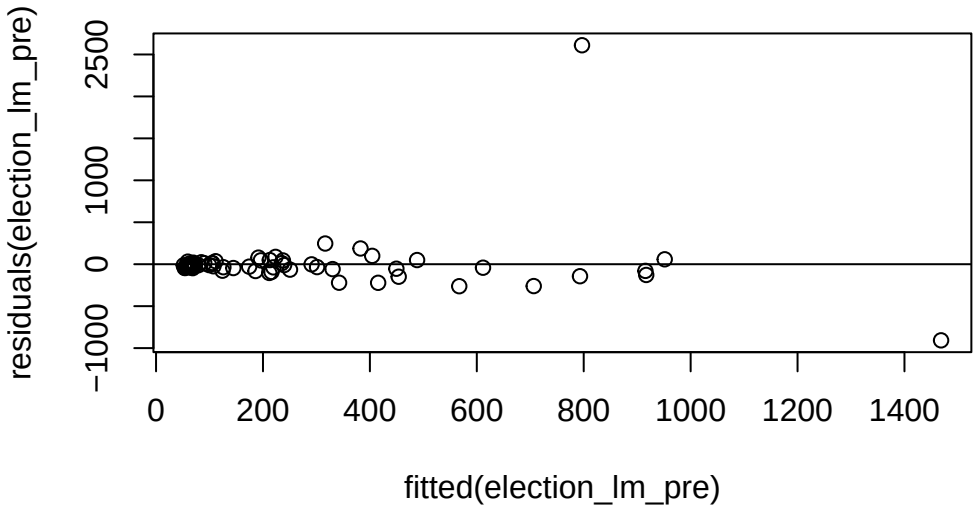


Figure 3

```
residuals <- resid(election_lm_pre)
qqnorm(residuals)
qqline(residuals, col="red")
```

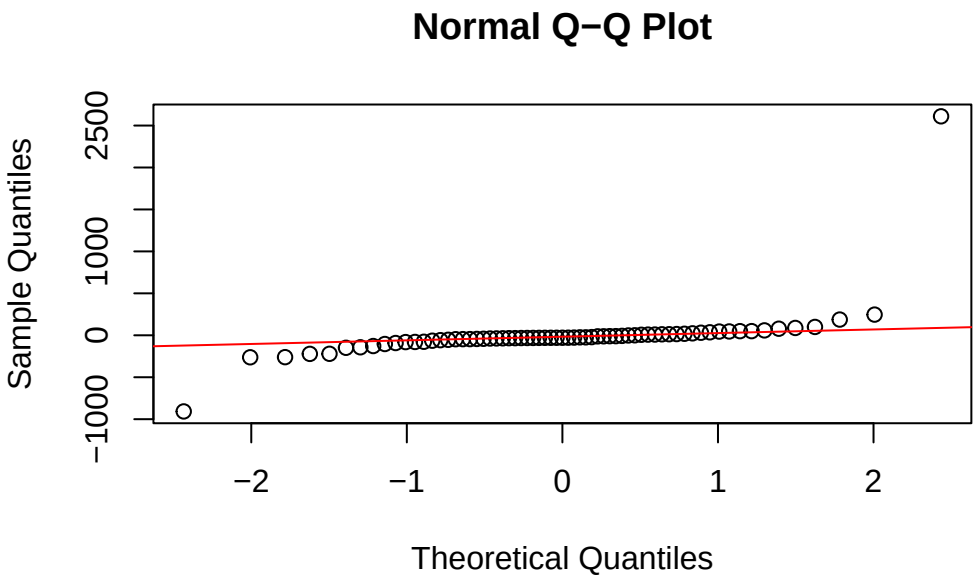
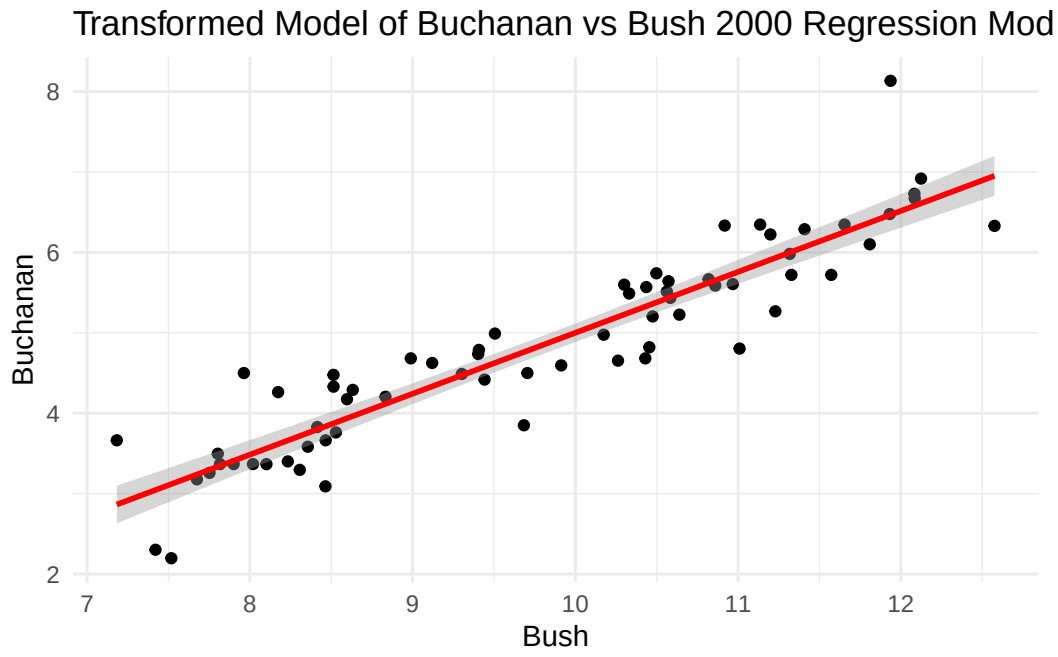


Figure 4

From the Figure 2, Figure 3 and the Figure 4, we can see that the data points are clustered. This does not meet the condition of linearity, and because of this we are going to re express the explanatory and response variables to hold the conditions.

Here is our post-transformation scatter plot.

```
ggplot(election, aes(x = log(Bush2000), y = log(Buchanan2000))) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Transformed Model of Buchanan vs Bush 2000 Regression Model",
       x = "Bush",
       y = "Buchanan")+
  theme_minimal()
```



Linear Model

```
election_lm_post <- lm(log(Buchanan2000) ~ log(Bush2000), data = election)
```

Residual vs Fitted

```
plot(fitted(election_lm_post), residuals(election_lm_post))
abline(0,0)
```

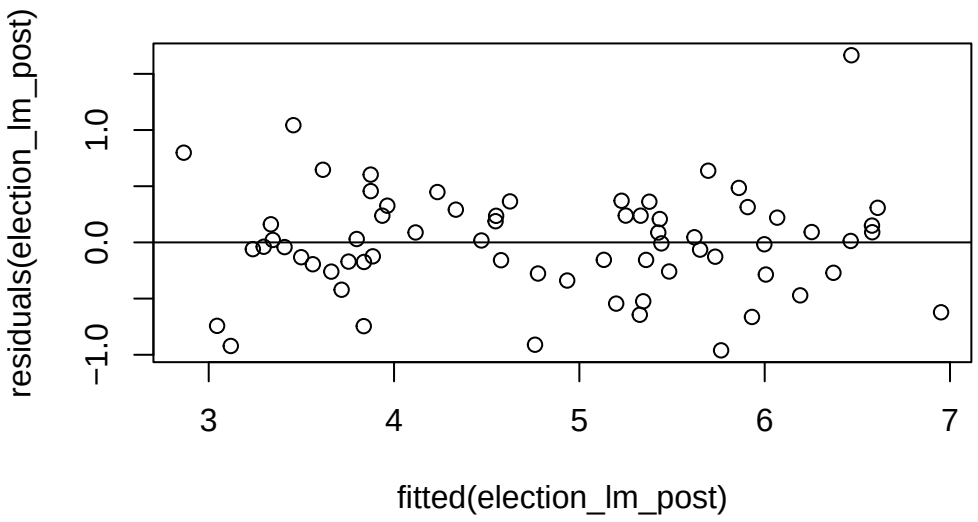


Figure 5

Normality

```
residuals_post <- resid(election_lm_post)
qqnorm(residuals_post)
qqline(residuals_post, col="red")
```

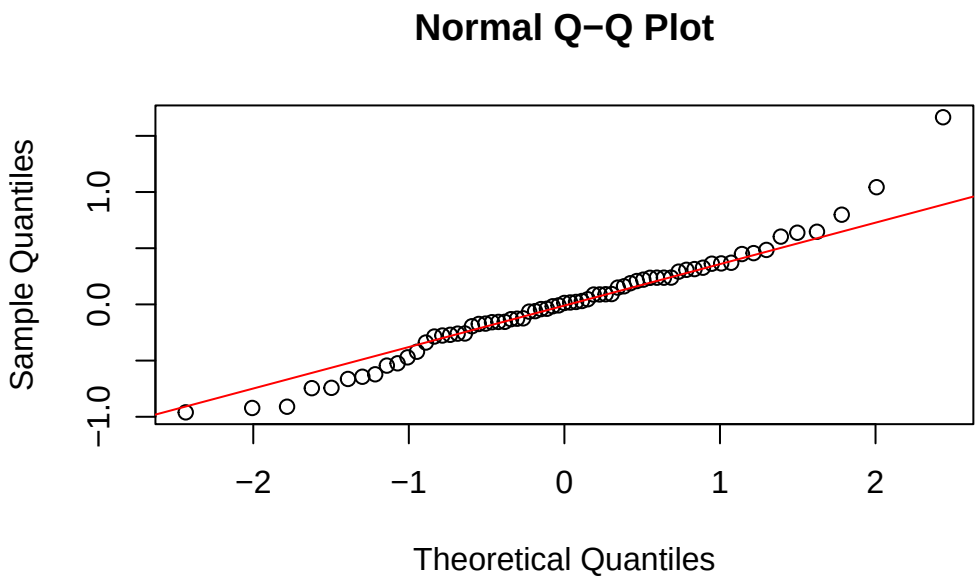


Figure 6

After it is transformed, $\log(y_i) = \beta\_0 + \beta_1 log(x_i)$

- **Null hypothesis** (H_0): $\beta = 0$
  There is no linear relationship between the votes for Bush and the votes for Buchanan in the 2000 election across Florida counties.
- **Alternative hypothesis** (H_A): $\beta \neq 0$
  There is a linear relationship between the votes for Bush and the votes for Buchanan in the 2000 election across Florida counties.

The significance level for this test is $p = 0.05$, and the confidence interval is 95.

```
#Creating a lm comparing Buchanan's and Bush's 2000 election data
election_lm <- lm(Buchanan2000 ~ Bush2000, data = election)

#saving the summary table of the election lm
election_summary <- tidy(election_lm)

#printing the summary table
kable(election_summary, caption = "Regression summary")
```

Table 4: Regression summary

term	estimate	std.error	statistic	p.value
(Intercept)	45.2898613	54.4794230	0.8313205	0.4088361
Bush2000	0.0049168	0.0007644	6.4319610	0.0000000

term	estimate	std.error	statistic	p.value

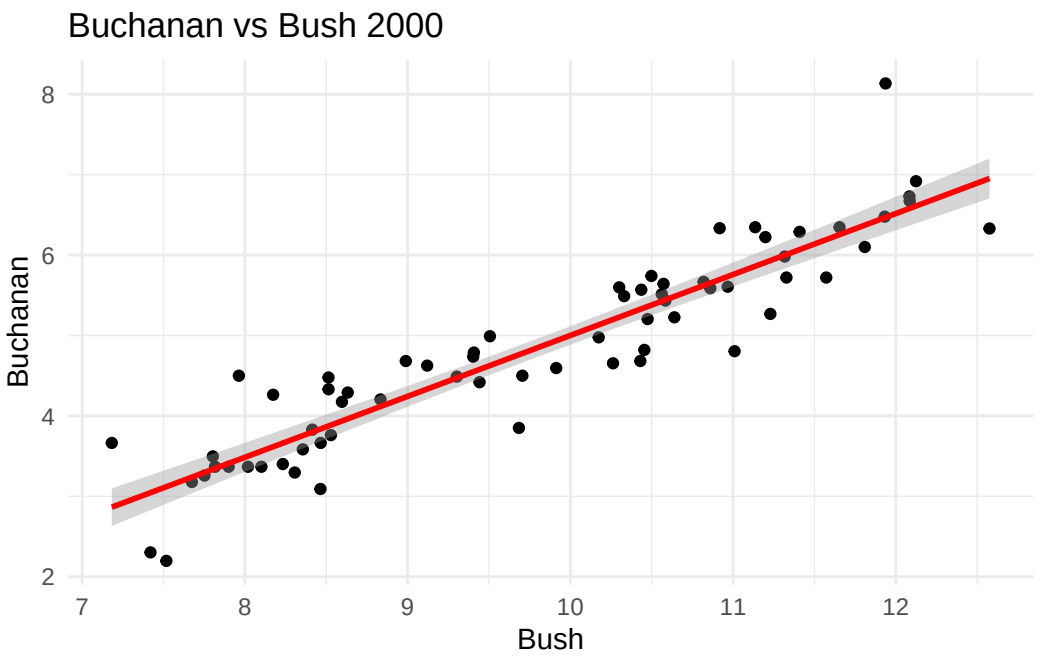
The p-value is near 0; suggesting significant relationship.

```
#Plotting the Bush and Buchanan data with a log transformation.
ggplot(election, aes(x = log(Bush2000), y = log(Buchanan2000))) +
  geom_point() +

  #adding a line to the plot
  geom_smooth(method = "lm", col = "red") +

  #adding labels
  labs(title = "Buchanan vs Bush 2000",
       x = "Bush",
       y = "Buchanan")+

  #making it pretty :)
  theme_minimal()
```




```
#saving the summary table of the election lm (post transformation)
election_summary_post <- tidy(election_lm_post )

#printing the summary table (post transformation)

kable(election_summary_post, caption = "Regression summary Post Transformation")
```

Table 5: Regression summary Post Transformation

term	estimate	std.error	statistic	p.value
(Intercept)	-2.5771236	0.3891909	-6.621747	0
log(Bush2000)	0.7577224	0.0393594	19.251368	0

As seen in Table @ref(tab:post-summary), we set up the regression equation:

$$E[log(Buchanan2000)|log(Bush2000)] = \beta_0 + \beta_1 * Bush2000 + \epsilon$$
$$= -2.5771236 + 0.7577224 * Bush2000$$

p-value is 0; so the results are statistically significant which rejects the null hypothesis.

```
pred <- data.frame(Bush2000 = 152846)
election_lm_post |>
  augment(newdata = pred,
          interval = "prediction",
          conf.level = 0.95)|>
  select(c(".fitted", ".lower", ".upper"))|>
  exp()
```

```
# A tibble: 1 x 3
  .fitted .lower .upper
    <dbl>  <dbl>  <dbl>
1    644.   248.  1674.
```

We can say with 95% confidence that the number of Buchanan votes based on the number of Bush notes should be between 248 and 1,674 in Palm Beach County, FL. This is significantly different than the actual number of Buchanan votes in Palm Beach County, which is 3,407. Assuming some of the votes cast for Buchanan were meant to be cast for Gore, our prediction interval suggests there were likely 2,763 miscast votes.

Summary

In this case study, we set out to determine if the ballot in the Palm Beach County lead to people mistakenly casting votes for Buchanan when they meant to cast the vote for Gore. Based on our models and prediction intervals, the Buchanan votes from Palm Beach County, FL are significantly different than what we would expect to see. This brings us to the conclusion that the confusing ballot for this county likely affected the votes, and led many people to mistakenly vote for Buchanan when they meant to vote for Gore.

Limitations

Our major limitation is that we are limited to election data within Florida. We also don't know if Buchanan had an abnormally large voter base located within Palm Beach Florida. We also cannot prove that the reason Buchanan had an abnormality large vote count was because of voter fraud. If that was the case the independence condition would have been violated. However all of these limitations do not seem either likely or significant enough to affect the overall outcome to the extent we saw in the election of 2000.