

# [Your Informative Title Here]

Sirohi Kumar and Chaira Harder

Invalid Date

*Your written short report should be the first thing your reader encounters in your case study document. While you may choose to label each of the required components/sections of the short report with its own header (see below for an example of creating a header in Quarto), you do not need to do so provided that all of the required information is included. Your report should be self-contained and written in such a way that a quantitatively inclined friend (who has taken or is taking SDS 291) could follow what you did without necessarily knowing anything about the 2000 presidential election otherwise.*

*As you write your report, you may wish to reference the guide to typesetting regression lines in Quarto using LaTeX (linked at the top of our class Moodle page), the Quarto help page for formatting documents using Markdown (), and the Quarto help page for customizing the output from executed code chunks.*

---

## INTRODUCTION

## RESULTS

### Generate Data

#### load libraries

```
library(tidyverse)
library(Sleuth2)
library(kableExtra)
library(broom)
library(performance)
```

#### data wrangling

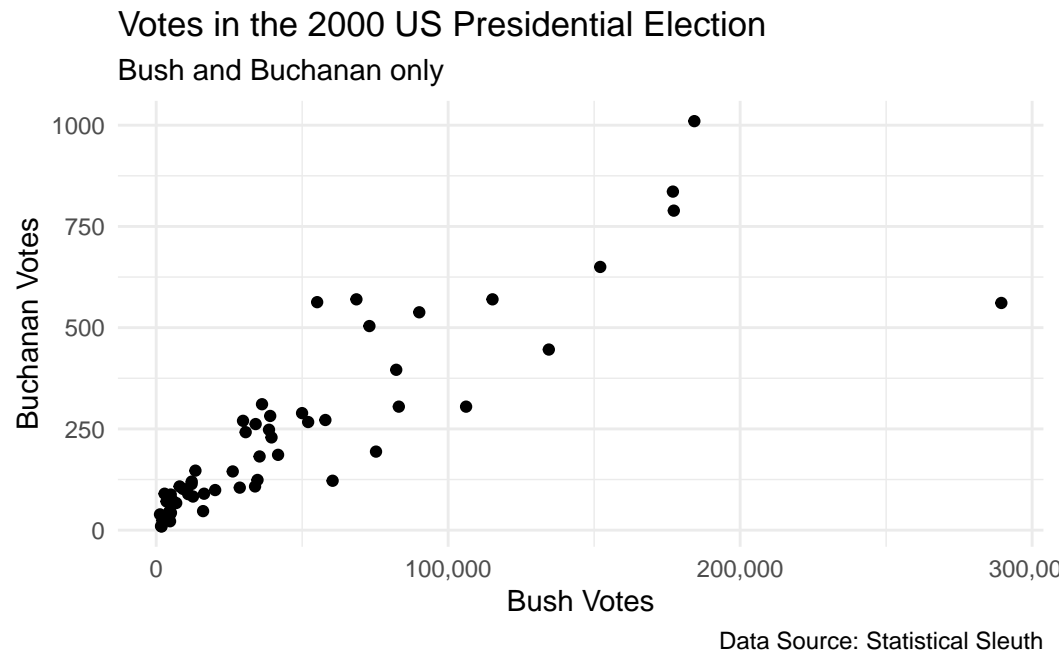
*include paragraph on data wrangling here*

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

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

## Explore Data

*describe the data layout + clearly outline the explanatory and response variables*



## Find Most Appropriate Model

### Generate various models

*expand on what we're doing and why we're doing it*

### R-Squared values

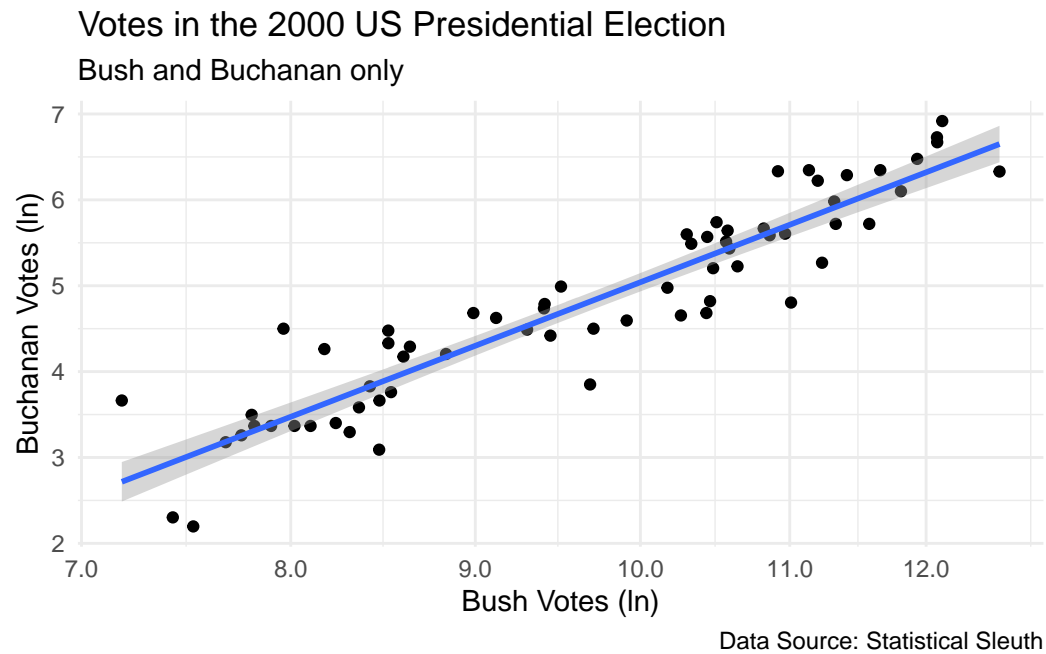
We can now compare the  $R^2$  values of these various transformations, and from this table we know that the  $\ln(\text{Buchanan})$  vs  $\ln(\text{Bush})$  model has the highest correlation.

Table 1: R-Squared (correlation) values for various models

Buchanan v Bush	$\ln(\text{Buchanan})$ v Bush	Buchanan v $\ln(\text{Bush})$	$\ln(\text{Buchanan})$ v $\ln(\text{Bush})$
0.7517819	0.6790001	0.571179	0.8658343

## Visualize new model

*describe the spread of the data in this version*



## Goodness of Fit

### Residuals

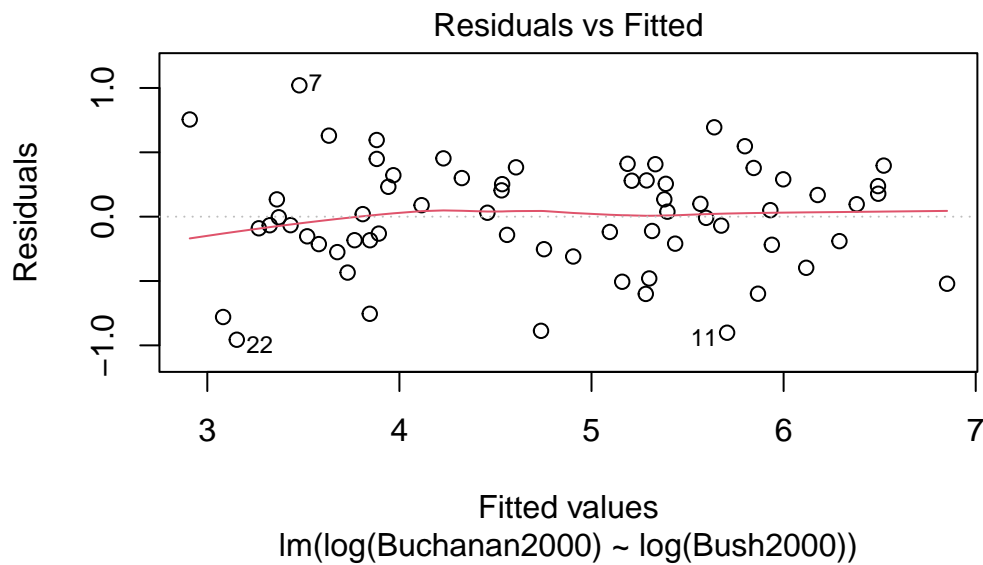
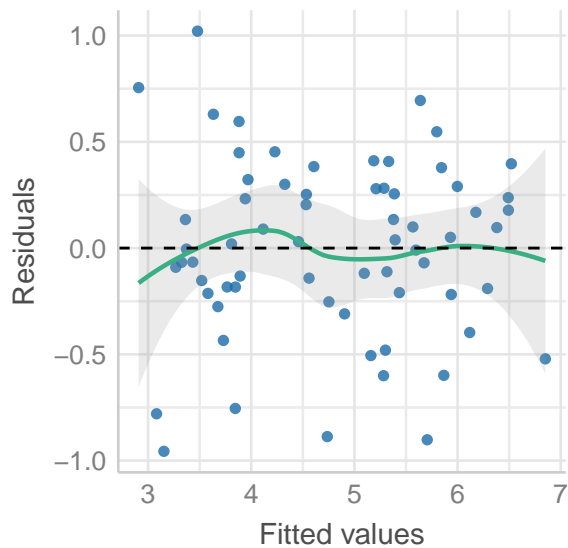
Our selected model, in which both the Bush and Buchanan votes are logged, yields the greatest  $R^2$  value at 0.8658343. With its  $R^2$  value, it captures a stronger linear relationship than its comparison models used (Buchanan vs Bush,  $\ln(\text{Buchanan})$  vs Bush, Buchanan vs  $\ln(\text{Bush})$ ) and also suggests a more proportional scaling between the votes, as we can see in the visual above.

To see if our model is truly effective and the votes in the dataset are not random, we can use the Linearity test and examine the Normality of Residuals.

### Linearity Test

## Linearity

Reference line should be flat and horizontal

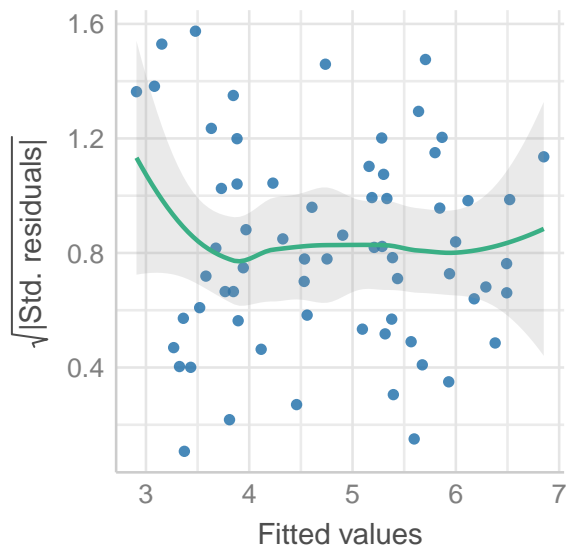


In this case, our reference line, in green, is not perfectly flat and horizontal, thus it does not pass the linearity test. But it's not bad tho...

## Homogeneity of variance

## Homogeneity of Variance

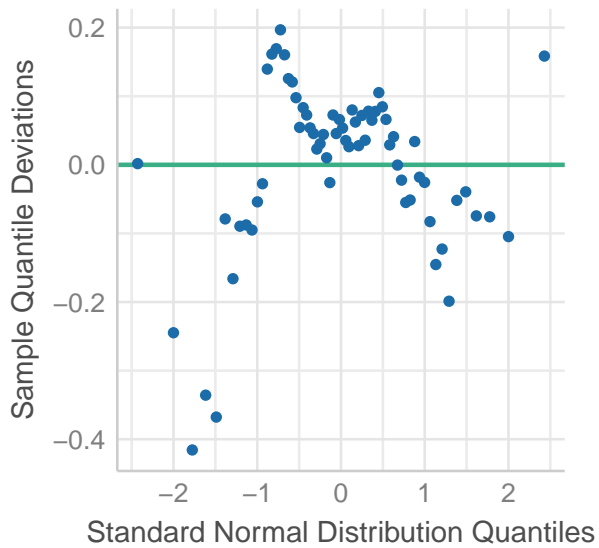
Reference line should be flat and horizontal

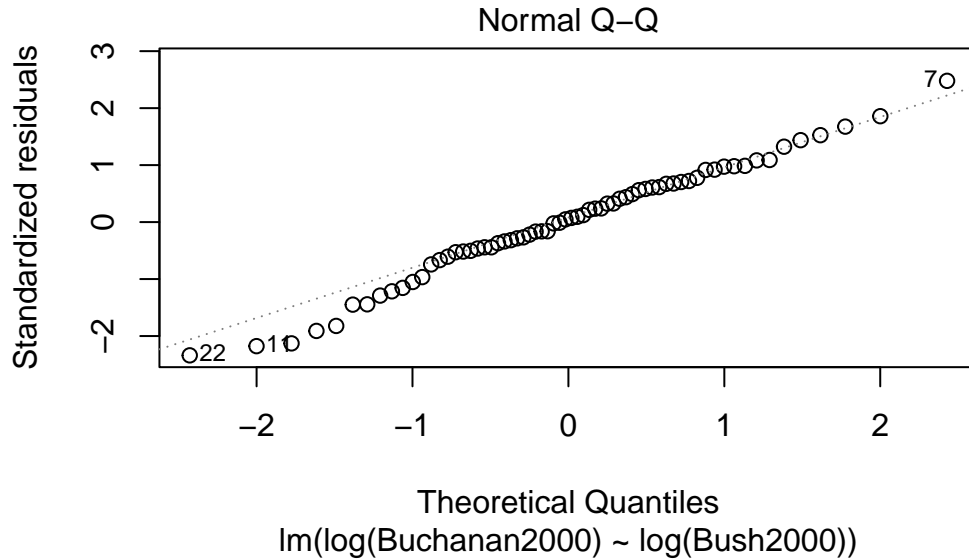


## Normality of Residuals

### Normality of Residuals

Dots should fall along the line





Sources: <http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/>

## Predictions

### Regression Line

Let  $Bush_i$  denote the  $\ln$  number of Bush votes in any Florida county during the 2000 election. Using the regression model from above, we can predict  $Buchanan_i$ , the  $\ln$  of the number of Buchanan votes in any Florida county  $i$  (during the 2000 election) based on any  $Bush_i$  value. Using this regression model, we can find

$$Buchanan_i = \beta_0 + \beta_1 (Bush_i).$$

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

This linear model predicts that for each increase in  $\ln(Bush_i)$ , that  $\ln(Buchanan_i)$  should increase by 0.7310 votes. We can now use this model to predict the number of votes Buchanan should have received in Palm Beach County, if Palm Beach county was the same as the other Florida counties.

## Prediction Interval

Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result—assuming the relationship is the same in this county as in the others

center	lower	upper
6.384143	5.524656	7.24363

We can transform these numbers to determine the non- $\ln$  vote count.

lower:  $e^{5.524656} = 250.8$  upper:  $e^{7.24363} = 1399.164$

Our prediction interval tells us that, based on this sample, 95% of the time, the number of votes for Buchanan in the 2000 election should be between 250.8 and 1399.164 votes, given that 152,846 people voted for Bush.

However, in the 2000 election, Buchanan received 3407 votes, which is over twice as large as the upper bound of our interval. This shows that Palm Beach county's votes for Bush and Buchanan are highly irregular compared to other Florida counties.

## Gore's Votes

Assuming that some of the votes cast for Buchanan were intended as votes for Gore, use the prediction interval to give an estimate for the likely number of votes intended for Gore but cast for Buchanan.

There were 3407 votes for Buchanan in the 2000 election, but our prediction interval tells us that the number of votes expected for Buchanan 95% of the time is between 250.800 and 1399.164, so the likely number of votes intended for Gore but cast for Buchanan should be between  $3407 - 250.800 = 3156.200$  and  $3407 - 1399.164 = 2007.836$ .

## DISCUSSION

Our goal with this case study was to determine if, based on the vote counts for Pat Buchanan and George Bush in every Florida county, we could conclude whether Buchanan received an unusual number of votes in Palm Beach county. We used a linear model on data that had been transformed to show that for every increase in  $\ln(Bush_i)$  by one  $\ln(Buchanan_i)$  should increase by about 0.7310 votes.

We used this model to generate a prediction interval that predicted with 95% confidence that Buchanan's vote count in Palm Beach County should have been between 250.8 and 1399.164. Instead, Buchanan received 3407 votes, a highly irregular value, according to this model. Based on this, we can conclude that there is evidence Buchanan received an unusually high number of votes in Palm Beach county.

This deviation from the number of expected votes lends credence to the complaints of many Democratic voters who reported having accidentally voted for Buchanan (the Reform candidate) instead

of Al Gore (the Democratic candidate), because of the confusing ballot layout. On a larger scale, this shows that the incredibly close election of George Bush in 2000 may have been – at least in part – due to a fluke in ballot design.

However, we can't make a conclusive claim that this is the case. First of all, our model only calculates the correlation between Buchanan and Bush's votes – it cannot determine if there's a causal relationship between the two, or indeed the existence of any causal factors affecting the relationship between the two. Additionally, we cannot directly attribute the unusual number of votes for Buchanan to the ballot layout, as we haven't examined any other elections with strange ballots and, again, this is not a causal model.

---

*When you create plots for your case study report, the `echo: false` chunk option tells Quarto to include the final output of your R commands (in this case, a plot) in your rendered PDF without printing the underlying R commands that generated that plot! The message and warning flags both prevent R from printing any additional text with error messages or warnings to the PDF.*

```
# #| echo: false
#
# # Fitting the regression line for mean mortality as a function of wine
↪ consumption
# lm.wine <- lm(Mortality ~ Wine, data = wine)
#
# # Representing the regression table as a dataframe (i.e., tidying the
↪ summary() output)
# lm.wine.table <- summary(lm.wine)$coefficients
#
# # Creating a nicely formatted table from the dataframe using the kable
↪ package
# # You can find more information about this package here:
↪ https://haozhu233.github.io/kableExtra/awesome\_table\_in\_pdf.pdf
# lm.wine.table |> kbl(col.names = c("Name for Col. 1", "Name for Col. 2",
↪ "Name for Col. 3", "Name for Col. 4"), align = "c", booktabs = T,
↪ linesep="", digits = c(2, 2, 2, 4)) |> kable_classic(full_width = F,
↪ latex_options = c("HOLD_position"))
```

## R APPENDIX

*Copy and paste all code that you used for your case study into one chunk at the end of your written report. Before submitting your case study, take one final look at the R Appendix and make sure that all code is clearly visible. If you see a line running off the side of the PDF, please split the code over multiple lines using a linebreak.*



```

# #| message: FALSE
# #| warning: FALSE
#
# # Loading necessary packages
# library(tidyverse)
# library(Sleuth2)
# library(broom)
# library(kableExtra)
#
# # Loading the case study data
# election <- Sleuth2::ex0825
#
# # Creating a second dataset with Palm Beach County excluded
# election_wo_pb <- election |> filter(County != "Palm Beach")
#
# # Loading another dataset on wine consumption and heart disease mortality
# wine <- Sleuth2::ex0823
#
# # Creating a scatterplot for the relationship between mortality and wine
#   ↪ consumption
# wine |> ggplot(aes(x = Wine, y = Mortality)) + geom_point() +
#   ggtitle("Association between wine consumption and mortality rates.")
#
# # Fitting and summarizing the regression line for mean mortality
# # as a function of wine consumption
# wine.lm <- lm(Mortality ~ Wine, data = wine)
# wine.lm.table <- wine.lm |> tidy()
# wine.lm.table |> kbl(col.names = c("Name for Col. 1", "Name for Col. 2",
#                                   "Name for Col. 3", "Name for Col. 4",
#                                   "Name for Col. 5"),
#                     align = "c", booktabs = T, linesep="",
#                     digits = c(2, 2, 2, 4)) |>
#   kable_classic(full_width = F, latex_options = c("HOLD_position"))

```