A Close Election: A Closer Look

Sirohi Kumar and Chaira Harder

2024-02-22

INTRODUCTION

The choice of American President – and to a lesser extent Congress – can determine the course of public policy and politics both domestically and internationally for the next 4 years. The election of 2000 was a particularly controversial one in American history, especially regarding Florida, whose close vote counts resulted in several false calls of the election for candidate Al Gore, despite the final reports calling the election for George Bush.

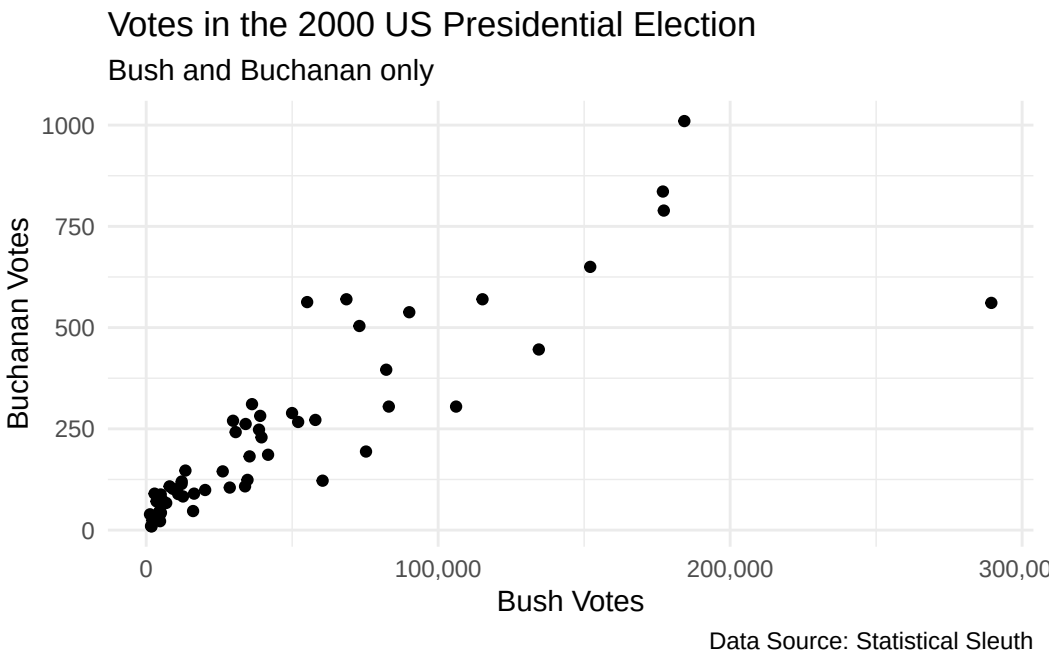
The cause of this upset can be attributed to many factors, key among which has been identified by analysts was the unique design of the ballot in Palm Beach County. Many Democrat voters complained the layout of the ballot led them to accidentally choose the Reform Party Candidate Pat Buchanan, who was projected to receive many fewer votes than he did in Palm Beach County. In this case study, we attempt to determine if there's evidence that Buchanan received an uncommonly high number of votes in the 2000 election. We do so by analyzing the relationship between the votes received by Bush (as the explanatory variable) and Buchanan (as the response variable) across several Florida counties.

RESULTS

Data wrangling

The data used in this case study comes from the Sleuth2 library in R. The dataset is of the number of votes for candidates George Bush and Pat Buchanan in the 2000 election from each county in Florida. For our data-wrangling process we removed one observation from the dataset. In order to generate a model without the Palm Beach county data, we removed the "Palm Beach'' row. The dataset originally has 67 observations, so the dataset we used for this analysis has a total of 66 observations.

Explore Data



This distribution is vaguely linear, although heavily grouped towards the left side of the x-axis and the bottom of the y-axis. The wide range of values on both axes, however, makes it difficult to determine the exact relationship between these variables.

Find Most Appropriate Model

Generate various models

We are generating four linear regression models to examine which would be best suited for the prediction of Buchanan votes. The four models are as follows:

- BVB: Buchanan versus Bush votes, both in original numeric value.
- LVB: The natural Log of Buchanan votes versus Bush votes in original numeric value.
- BVL: Buchanan votes in original numeric value versus the natural log of Bush votes.
- LVL: The natural log of Buchanan versus the natural log of Bush votes.

Having each of these models will allow us to look at the statistics of each model and decide, based on their statistics, which will fit our prediction best.

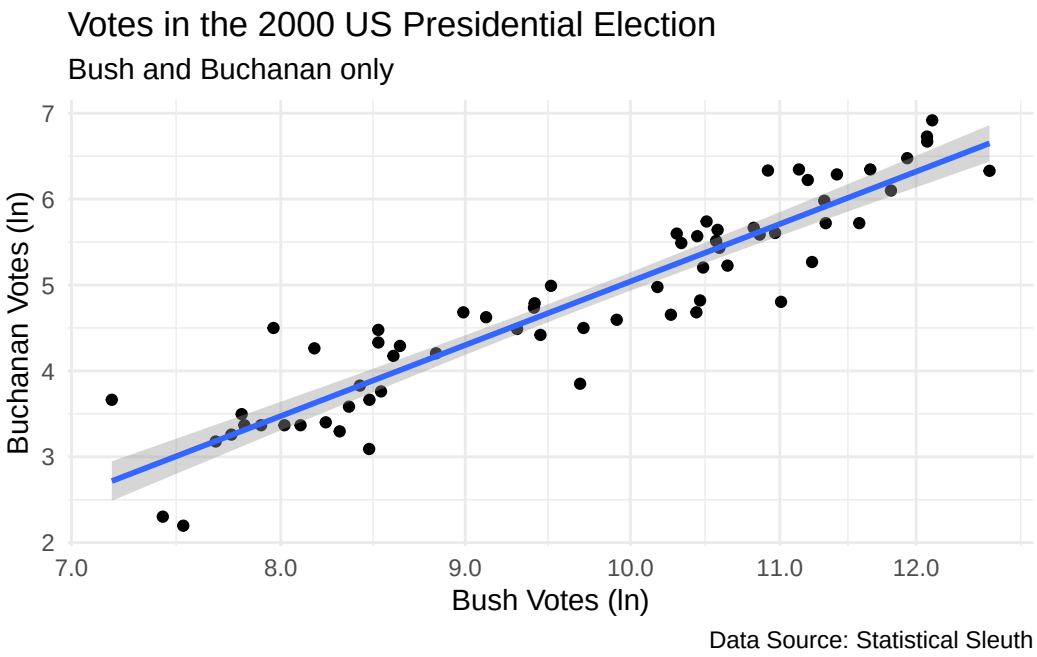
R-Squared values

We can compare the R^2 values of these various transformations using this table, which tells us that the `ln(Buchanan) vs ln(Bush)` model has the highest correlation of the four models.

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

Buchanan v Bush	ln(Buchanan) v Bush	Buchanan v ln(Bush)	ln(Buchanan) v ln(Bush)
0.7517819	0.6790001	0.571179	0.8658343

Visualize new model



This new visualization shows that the natural log of Bush and Buchanan's votes have a much more linear relationship. While there's slightly more variation towards the beginning of both axes, overall there's a high level of correlation between the logged vote counts for both candidates, which is reflected by the high R^2 value.

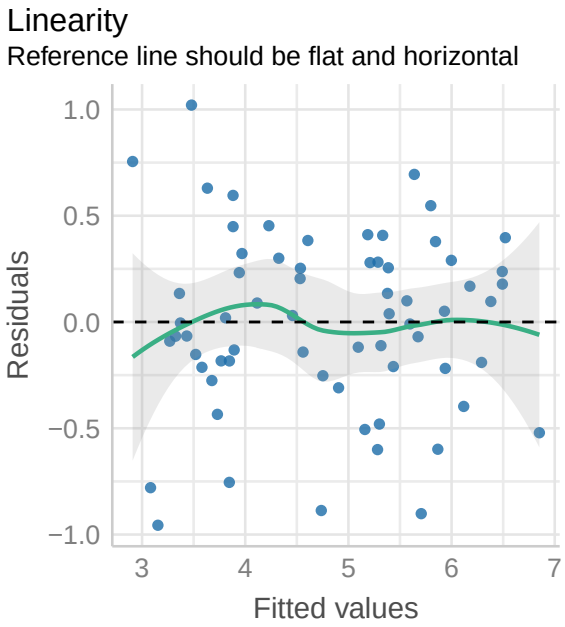
Goodness of Fit

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(Buchanan) vs Bush, Buchanan vs ln(Bush)) and also suggests a more proportional scaling between the votes, as we can see in the visual above.

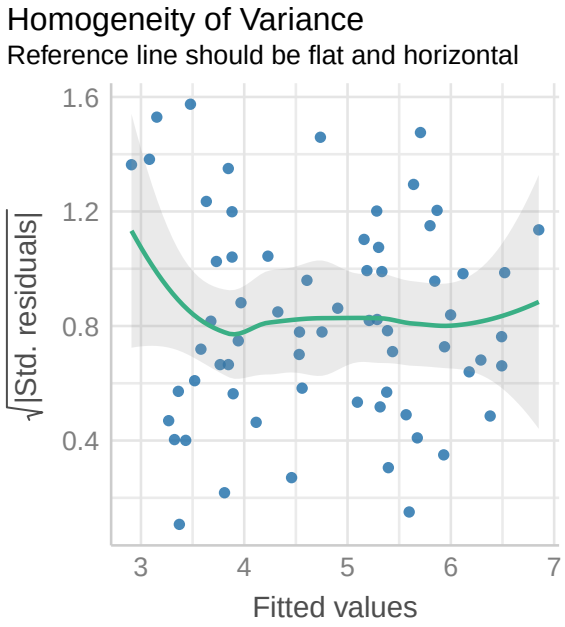
To see if our our model is truly effective and whether the votes in the dataset are truly not random – that there's a correlation – we can use the Linearity Test, Homogeneity of Variance, as well as the Normality of Residuals.

Linearity Test



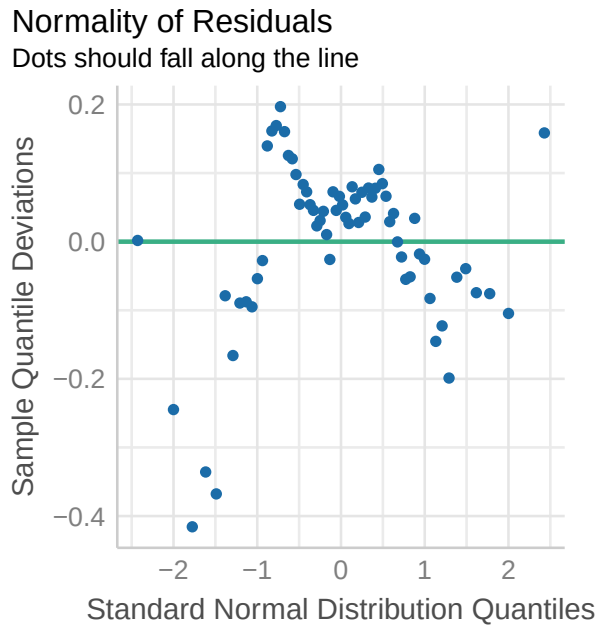
In this case, our data makes a relatively flat and horizontal reference line, thus satisfying the linearity test, meaning there is strong evidence for a linear relationship between our independent/explanatory variable *Bush2000* and dependent/outcome *Buchanan2000* variables.

Homogeneity of variance



A linear relationship requires constant variance. While our residuals are scattered, we do have constant variance across all fitted values (the green reference line is mostly horizontal and somewhat flat in the majority). Thus, we know on average the variance of the residuals does not systematically increase or decrease across our range of fitted values.

Normality of Residuals



We see in the Normality of Residuals plot above that the data points for the sample quantile deviations vs the normal distribution quantiles do not make a horizontal line shape. If they were to make a horizontal line shape, it would indicate that the residuals from our model are perfectly normally distributed. So, our LVL model does not satisfy the Normality of Residuals, but this could just be a consequence of our transformation of the data.

Predictions

Regression Line

Let $Bush_i$ denote the natural log of Bush votes in any Florida county during the 2000 election. Using the regression model from above, we can predict $Buchanan_i$, the natural log 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

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

	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. 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 indicates that Palm Beach county's votes for Buchanan, based on Bush's votes, 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.

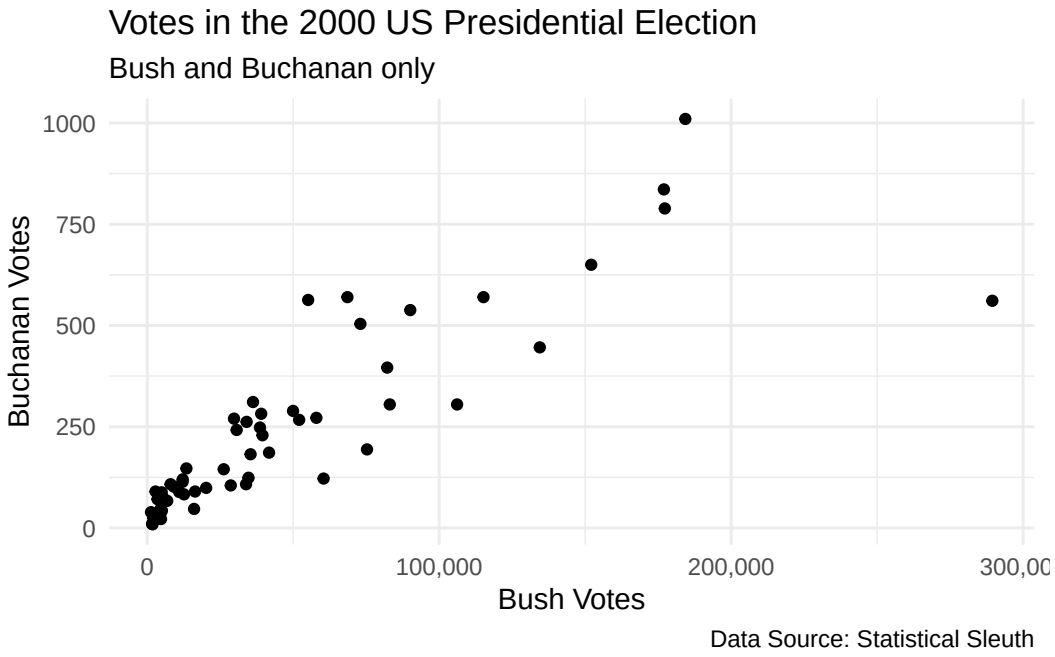
This deviation from the number of expected votes lends credence to the complaints of many Democratic voters who reported having accidentally voted for Buchanan (the Reform candidate) instead of Al Gore (the Democratic candidate), because of the confusing ballot layout. On a larger scale, this shows that 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.

R APPENDIX

```r
# importing our necessary libraries
library(tidyverse)
# Sleuth2 contains the data used in this analysis
library(Sleuth2)
# to make tables in R
library(kableExtra)
# for data wrangling
library(broom)
# for residual testing
library(performance)
# for our plot visualizations
library(ggplot2)

# --------------

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

# data wrangling:
# Creating a second dataset EXCLUDING Palm Beach County
election_wo_pb <- election |> filter(County != "Palm Beach")

# data exploration and visualization:
# visualizing our initial data using a scatterplot
ggplot(data = election_wo_pb, aes(x = Bush2000, y = Buchanan2000)) +
  geom_point() +
```

```
  scale_x_continuous(labels = scales::comma) +
  xlab("Bush Votes") +
  ylab("Buchanan Votes") +
  labs(title = "Votes in the 2000 US Presidential Election", subtitle =
   ↪  "Bush and Buchanan only", caption = "Data Source: Statistical Sleuth")
   ↪  +
  theme_minimal()
```



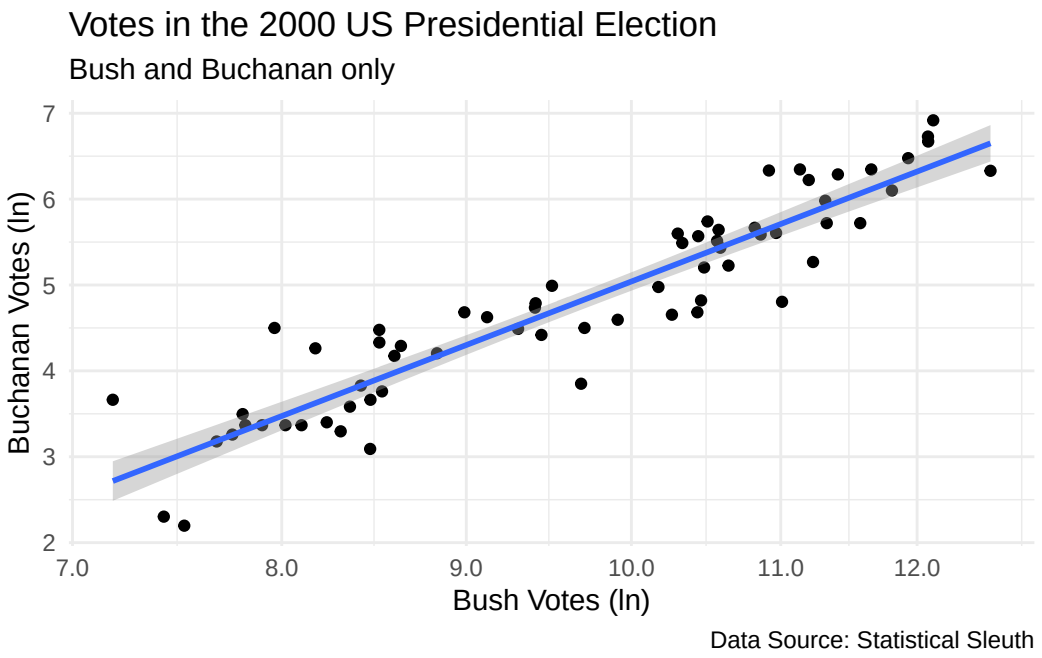

```
## Bush vs Buchanan
model_bvb <- lm(Buchanan2000 ~ Bush2000, data = election_wo_pb)
BVB <- summary(model_bvb)$r.squared

## log(Bush) vs Buchanan
model_lvb <- lm(Buchanan2000 ~ log(Bush2000), data = election_wo_pb)
LVB <- summary(model_lvb)$r.squared

## Bush vs log(Buchanan)
model_bvl <- lm(log(Buchanan2000) ~ Bush2000, data = election_wo_pb)
BVL <- summary(model_bvl)$r.squared

## log(Bush) vs log(Buchanan)
model_lvl <- lm(log(Buchanan2000) ~ log(Bush2000), data = election_wo_pb)
LVL <- summary(model_lvl)$r.squared

# creating a table of our R^2 values for our 4 models:
```

```
rsq.table = data.frame(BVB, LVB, BVL, LVL)
names = c("Buchanan v Bush", "ln(Buchanan) v Bush", "Buchanan v ln(Bush)",
 ↪  "ln(Buchanan) v ln(Bush)")
caption = "R-Squared (correlation) values for various models"

rsq.table %>% kable(caption = caption, col.names = names)
```

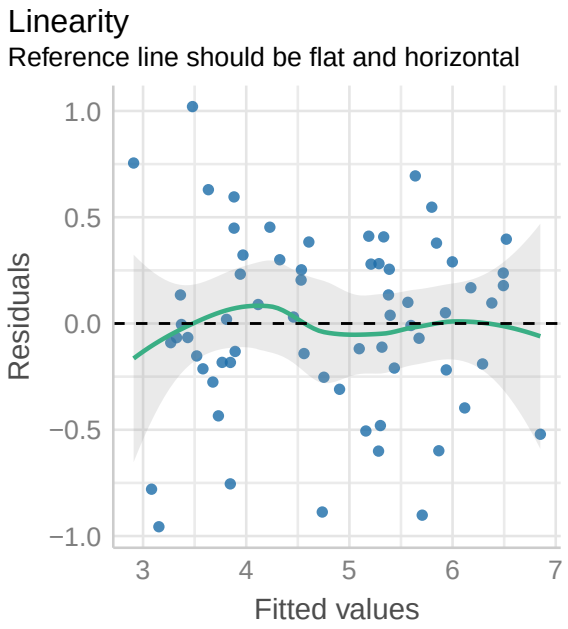
Table 2: R-Squared (correlation) values for various models

Buchanan v Bush	ln(Buchanan) v Bush	Buchanan v ln(Bush)	ln(Buchanan) v ln(Bush)
0.7517819	0.6790001	0.571179	0.8658343

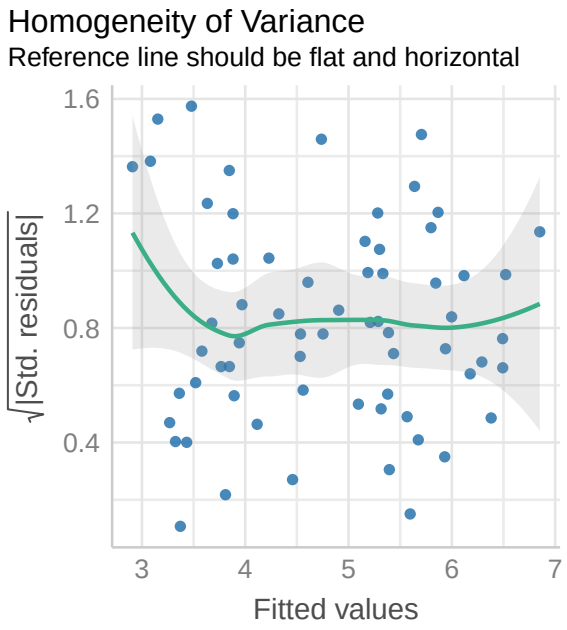
```
# visualizing our new model:
ggplot(data = election_wo_pb, aes(x = log(Bush2000), y = log(Buchanan2000)))
 ↪  +
  geom_point() +
  scale_x_continuous(trans = "log", labels = scales::comma) +
  geom_smooth(method = lm, formula = y~x) +
  xlab("Bush Votes (ln)") +
  ylab("Buchanan Votes (ln)") +
  labs(title = "Votes in the 2000 US Presidential Election", subtitle =
   ↪  "Bush and Buchanan only", caption = "Data Source: Statistical Sleuth")
   ↪  +
  theme_minimal()
```



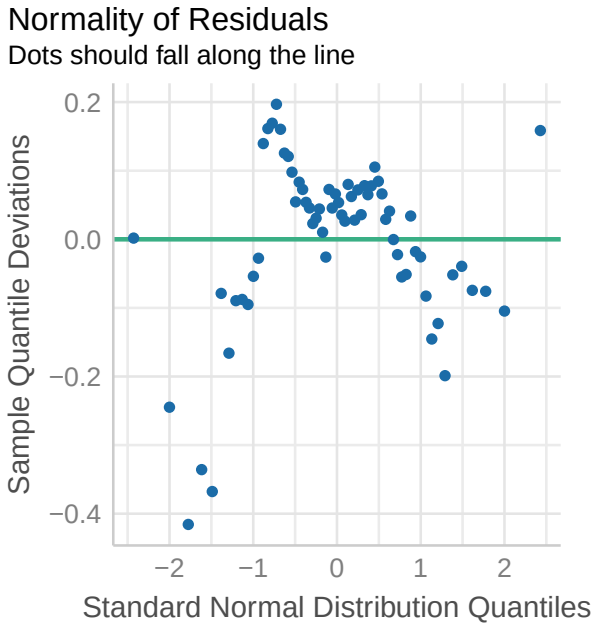
```
# Looking at our goodness of fit:
# Linearity test:
# this uses the performance library
check_model(model_lvl, check = c("linearity"))
```




```
# this is the default linearity plot which is harder to read:
# plot(model_lvl, 1)

# Homogeneity of Variance
# plot(model_lvl, 3)
check_model(model_lvl, check = c("homogeneity"))
```




```
# Normality of Residuals
check_model(model_lvl, check = c("qq"))
```




```
# plot(model_lvl, 2)
```

```r
# and some of our sources:
# Sources:
# -
 ↪  http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regress
# - https://cran.r-project.org/web/packages/performance/index.html

# our regression line summary for our chosen model
model.table <- summary(model_lvl)$coefficients

model.table |>
  kbl(col.names = c("Estimate", "Std. Error", "t value", "p value"),
      align = "c", booktabs = T, linesep="", digits = c(4, 4, 4, 4)) |>
  kable_classic(full_width = F, latex_options = c("HOLD_position"))
```

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

```r
# our 95% prediction interval for the number of Buchanan votes
predict(model_lvl, newdata = data.frame(Bush2000 = 152846), interval =
 ↪  "prediction", level = 0.95) %>%
  kbl(col.names = c("center", "lower", "upper"),
      align = "c", booktabs = T, linesep="") %>%
  kable_classic(full_width = F, latex_options = c("HOLD_position"))
```

center	lower	upper
6.384143	5.524656	7.24363