Short Report 1

Victor Huang

11/10/2021

Introduction

The presidential election between Gore and Bush was, and still is, the closest election in all of US history. From the start of the election to its tie-breaker vote in Florida, the whole election process was highly contested. While we now know in retrospect that Bush won the election, Florida became the sit for much controversy's Palm Beach County boasted a particularly high vote count for another candidate, Pat Buchanan, caused by a design flaw on the voting ballot. In this report I will explore the 2000 Florida voting election data dn see if the palm beach case and see it had a significant impact for the election result in Florida, as well as whether it is possible that palm beach county's votes might have swayed the election results.

Results

After I initially fit the regression of Bush vs. Buchanan votes, the scatter plot indicates that Palm Beach County is indeed an outlier data point as it hosts an unusually high number of votes for Buchanan for the number of Bush. This is shown below in figure 1.

As such, in order to get a model that is a more accurate representation of the true data, I will remove the Palm Beach County case before proceeding with further investigation. Looking at the dataset, the model doesn't follow the assumptions made for linear regressions. To change this I will be taking the log of both variables and creating a new regression with that. This would allow me to analyze the dataset with more accuracy and ease. Looking at the summary for our new model, we get 'log(Bush2000)' with intercept estimate of 0.73096 with standard error 0.03597 and an intercept estimate of -2.34149 and standard error 0.03597.

Afterwards, I created a prediction model (shown in figure 2) to predict a reasonable number of votes that Buchanan would expect to get for Palm Beach County as well as a qq norm plot to check the residuals (shown in figure 3). We are 95% confident that the re-poll in Palm Beach County based on our new model would have given Buchanan a vote count between 250 and 1399 votes with the predicted vote count for Buchanan being 592 votes with a standard error of 2 vote. From our qq norm plot, we see that the residual values fulfill the assumptions and no further modifications will be needed. With an upper bound of 1399 votes, we can see that the original 3407 votes for Buchanan is extremely unlikely (by the definition of the 95% confident interval). We can also conclude that out of the votes given to Buchanan, approximately 2008 to 3157 of those votes would have been intended for Al Gore instead.

```
##
## Call:
## lm(formula = log(Buchanan2000) ~ log(Bush2000), data = gb_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95631 -0.21236  0.02503  0.28102  1.02056
```

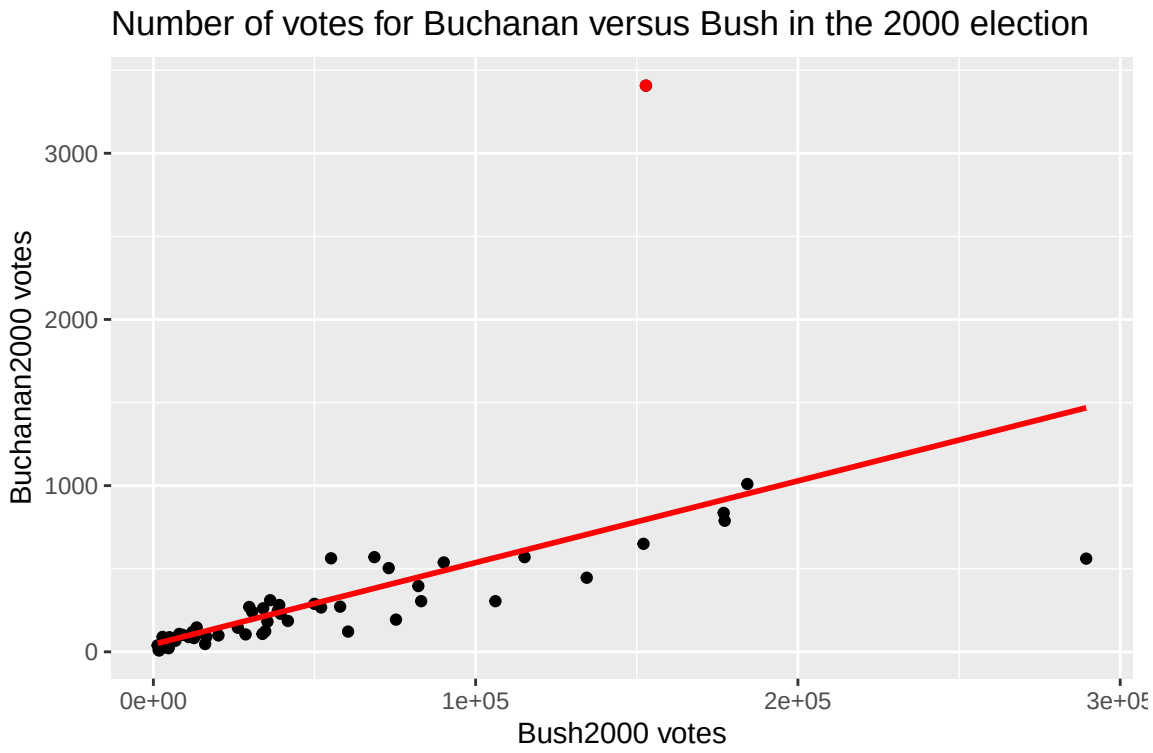


Figure 1: Scatterplot

```
##
## 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
```

Discussion

Looking at the model we created, we concluded that some of the votes given to Buchanan were actually intended for Al Gore. As such, we estimate that approximately 2008 to 3157 votes were deducted from the true vote count for Al Gore, negatively affecting his number of votes. What this means is that, assuming that this and some other counties may have some inconsistencies, it is plausible that Buchanan could have theoretically won the 2000 presidential election (this is also justified further by a Florida vote recount, going from the original 1738 vote difference to a less than 400 vote difference)

A limitation for this dataset would be its inability and failure to account for other possible variables that may have led to the misjudgment of votes in the first place. Since the voting ballot wasn't custom made for Florida's Palm Beach County, other states and their counties may have also experience similar discrepancies (perhaps to a lesser degree since they weren't reported). However, these plausible erroneous results may have favored either candidate. As such, we can not truly say that a single case like Florida's Palm Beach

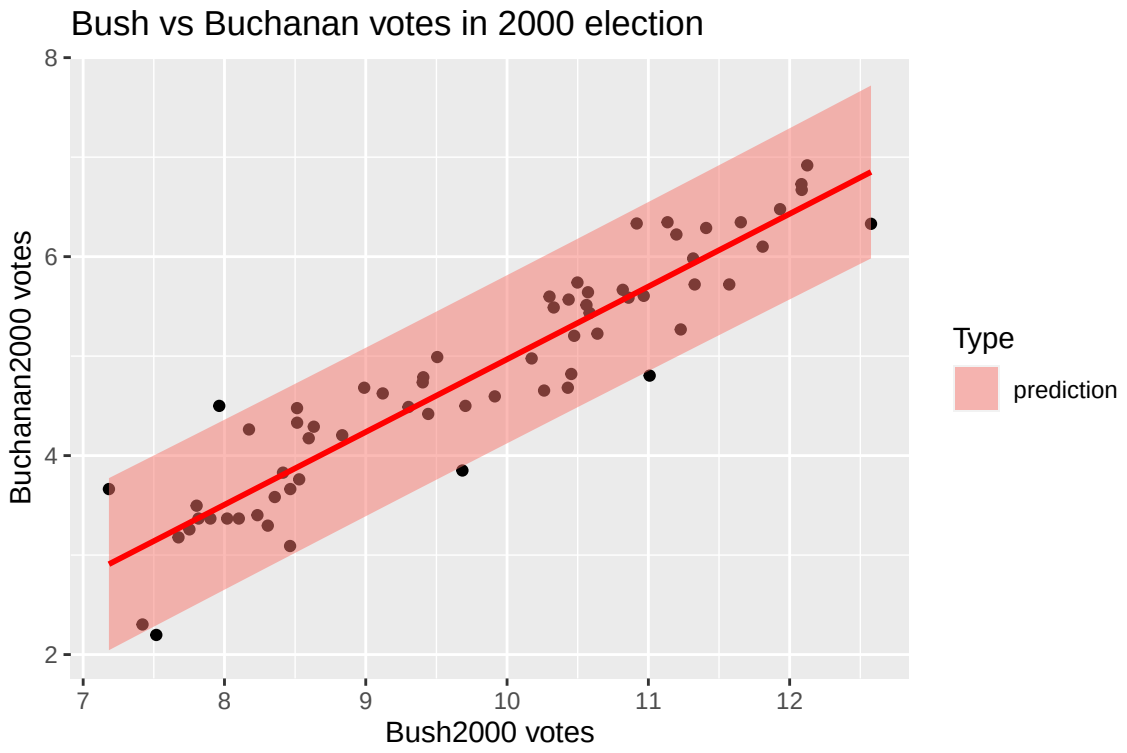


Figure 2: 95% Prediction model

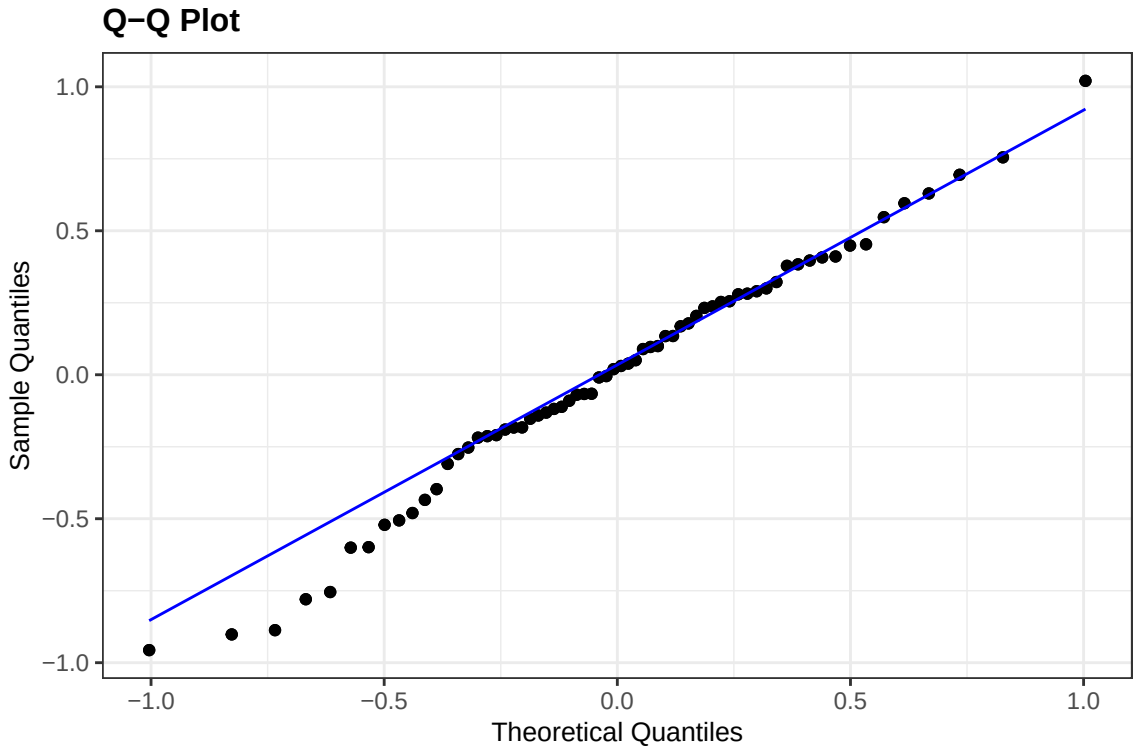


Figure 3: QQ Normal Plot

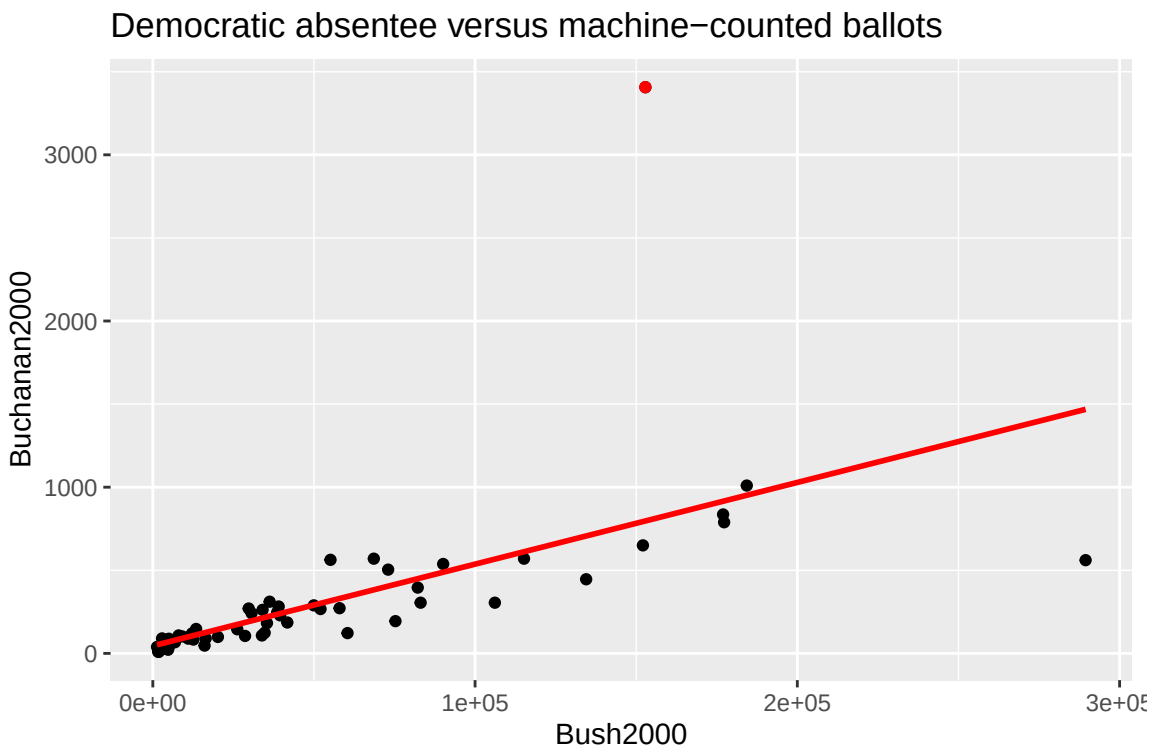
County could have significantly impacted the results for the election on a nationwide basis (think of it as if everyone was wrong to a certain degree, no one is truly extremely wrong)

Another limitation is the fact that we are simply considering two candidates. We know for a fact that there exists more than three candidates, each with their own political agenda. We are also not counting for political bias of voters (votes skewed towards the candidates home state, previous interactions between the state and the candidate, etc). As such, it would be unwise to consider a living breathing population as simply data points on a graph without thinking and considering the possibility of other, more human, factors.

Appendix

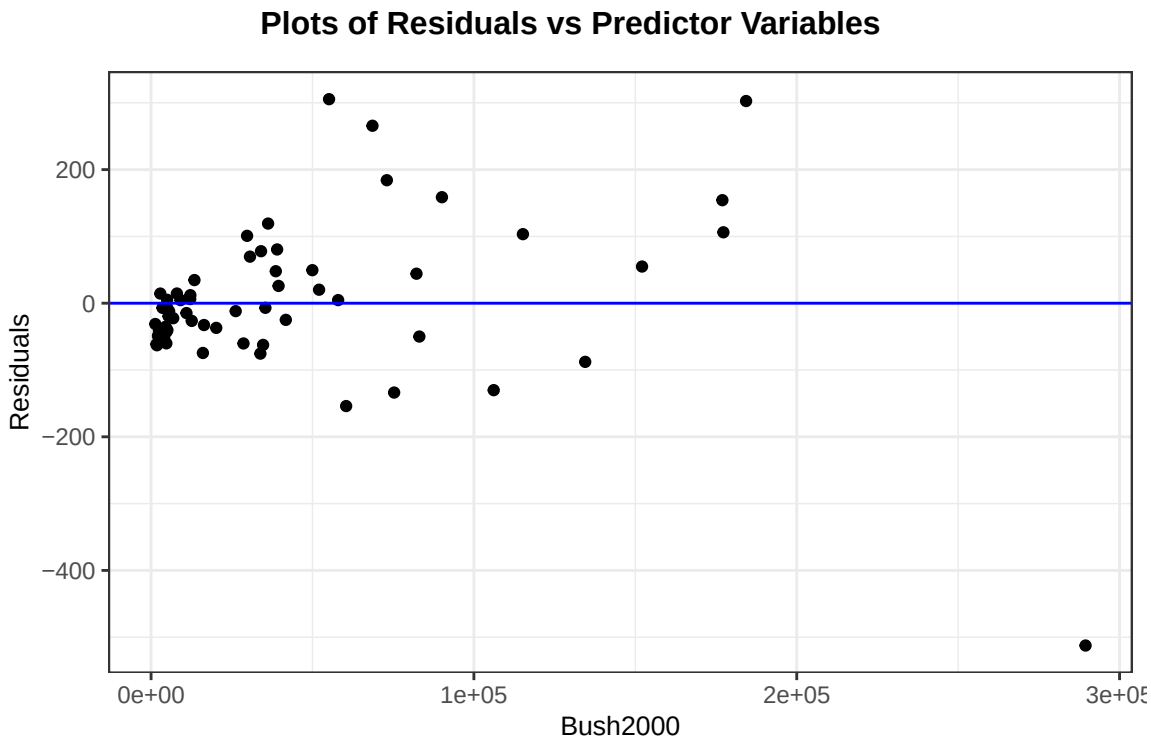
```
#take in the dataset and create a gb data table
gb <- ex0825

#create an initial plot of the data with all data points
ggplot(gb, aes(x=Bush2000, y=Buchanan2000)) +
  geom_point() +
  geom_point(data=filter(gb, County == "Palm Beach"), color="red") +
  geom_smooth(method="lm", se=FALSE, color="red") +
labs(x="Bush2000",
     y="Buchanan2000",
     title="Democratic absentee versus machine-counted ballots")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#create a new data model taking into consideration the palm beach county vote discrpency
gb_new <- gb[-67, ]
```

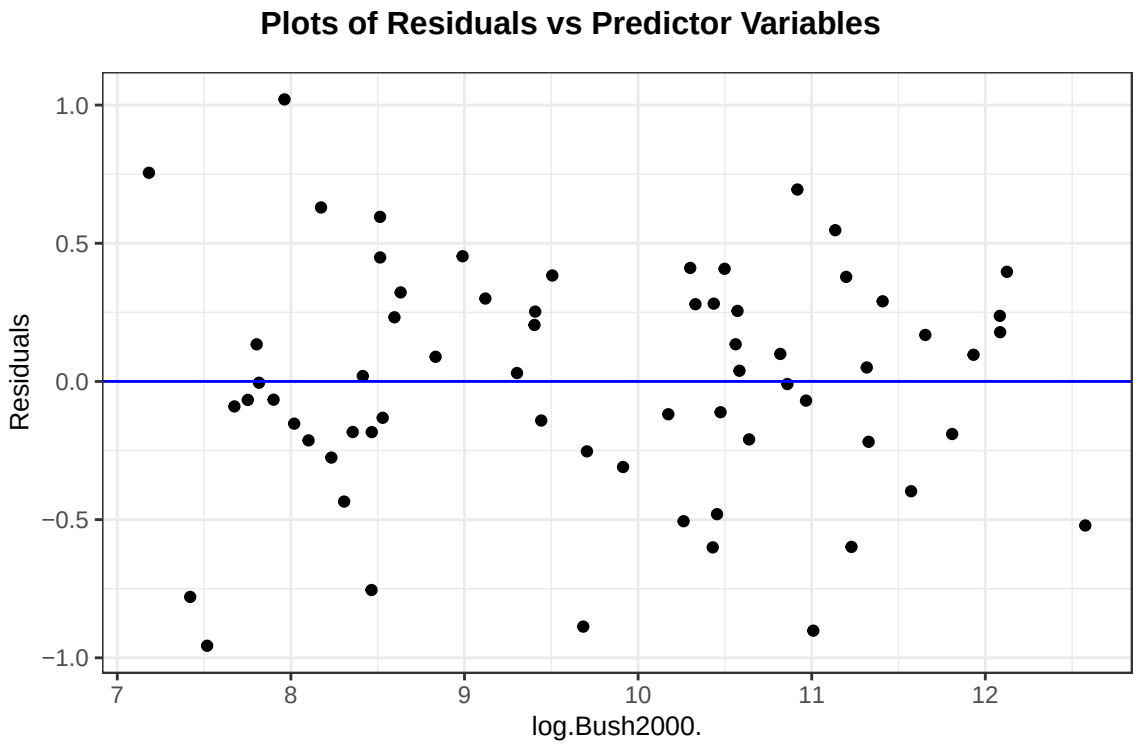
```r
#creating a new linear model after accounting for Palm Beach County
gb_new_lm <- lm(Buchanan2000 ~ Bush2000, data = gb_new)

#creating the residual panel for the new linear model
resid_xpanel(gb_new_lm)
```



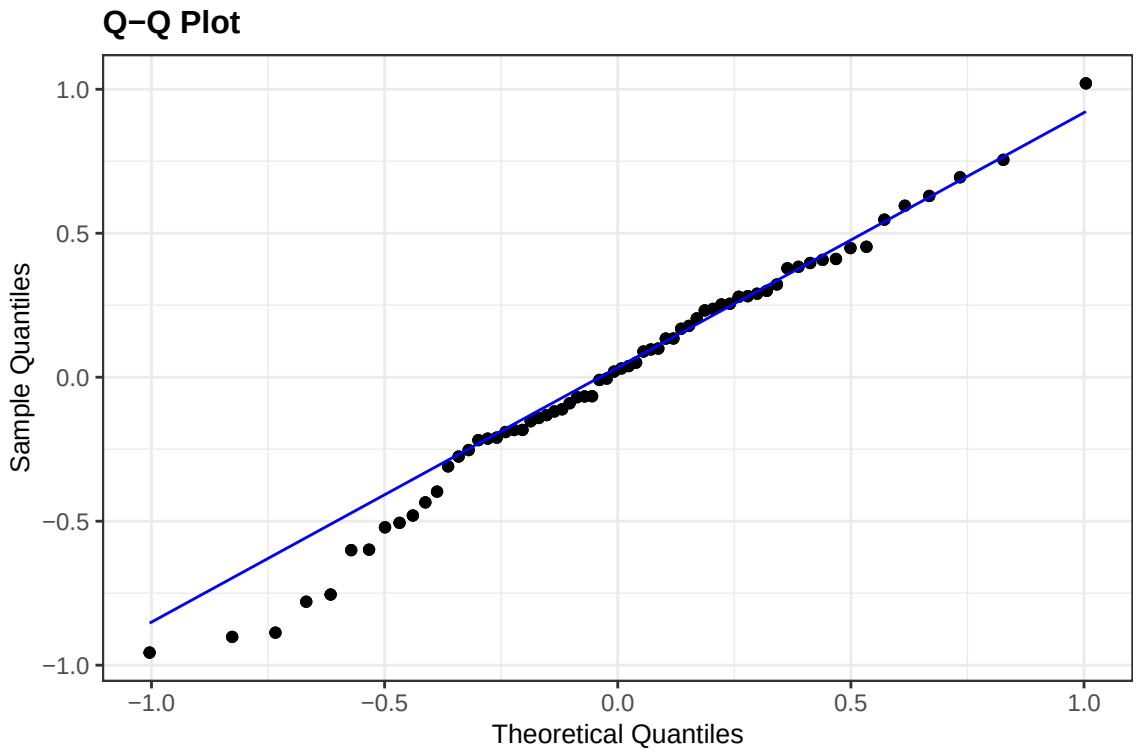

```r
#summarizing the new data model to find the parameter values and their respective standard errors
summary(gb_new_lm)
```

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

```
gb_new_lm <- lm(log(Buchanan2000) ~ log(Bush2000), data = gb_new)
resid_xpanel(gb_new_lm)
```




```
# plotting the qq norm model to check the residuals for the new linear model
resid_panel(gb_new_lm, plots = "qq")
```

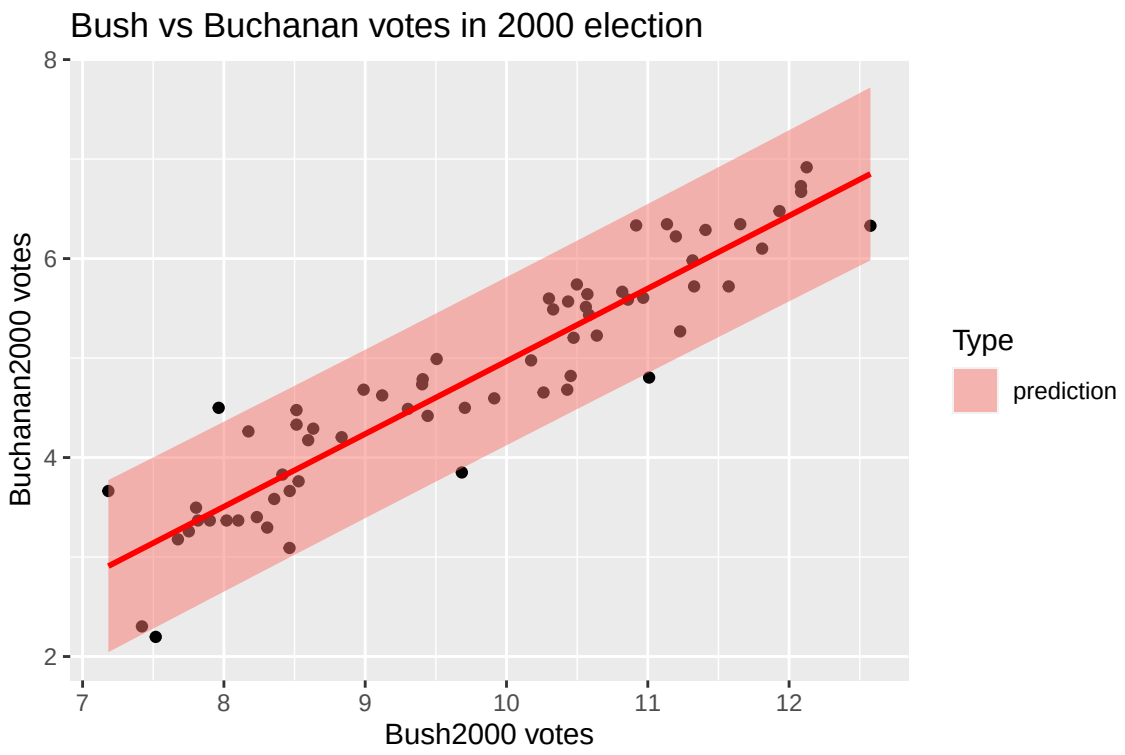


```
#creating a prediction model for the new data set
gb_new_pred <- data.frame(gb_new,
                    predict(gb_new_lm, interval = "prediction"))
```

```
## Warning in predict.lm(gb_new_lm, interval = "prediction"): predictions on current data refer to _futu
```

```
#plotting the prediction plot for the new dataset
ggplot(gb_new_pred, aes(x=log(Bush2000), y=log(Buchanan2000))) +
  geom_point() +
  geom_ribbon(
    aes(ymin = lwr,
        ymax = upr,
        fill = "prediction"),
    alpha = .5) +
  geom_smooth(method="lm", se=FALSE, color="red") +
labs(x="Bush2000 votes",
     y="Buchanan2000 votes",
     title="Bush vs Buchanan votes in 2000 election",
     fill = "Type")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#predicting the true data count
predict(gb_new_lm,   # model object
        newdata = data.frame(Bush2000 = 152846),  # new data
        interval = "prediction",   # interval type
        se.fit = T)
```

```
## $fit
##        fit      lwr     upr
## 1 6.384143 5.524656 7.24363
##
## $se.fit
## [1] 0.09416562
##
## $df
## [1] 64
##
## $residual.scale
## [1] 0.4198003
```

```
#getting the parameter value to calculate the value
summary(gb_new_lm)
```

```
##
## Call:
## lm(formula = log(Buchanan2000) ~ log(Bush2000), data = gb_new)
##
## 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
```

```
#getting rid fo the log value to get the true value
exp(6.384143)
```

```
## [1] 592.3768
```

```
exp(5.524656)
```

```
## [1] 250.8
```

```
exp(7.24363)
```

```
## [1] 1399.164
```

```
#finding the SE value using the previously found exp values
se.pred = sqrt(exp(0.09417)^2 + exp(0.4198)^2)
se.pred
```

```
## [1] 1.876882
```

```
#calculating the negative impact the bad vote count had on the candidates
3407 - 1399
```

```
## [1] 2008
```

```
3407 - 250
```

```
## [1] 3157
```