# Short Report 1

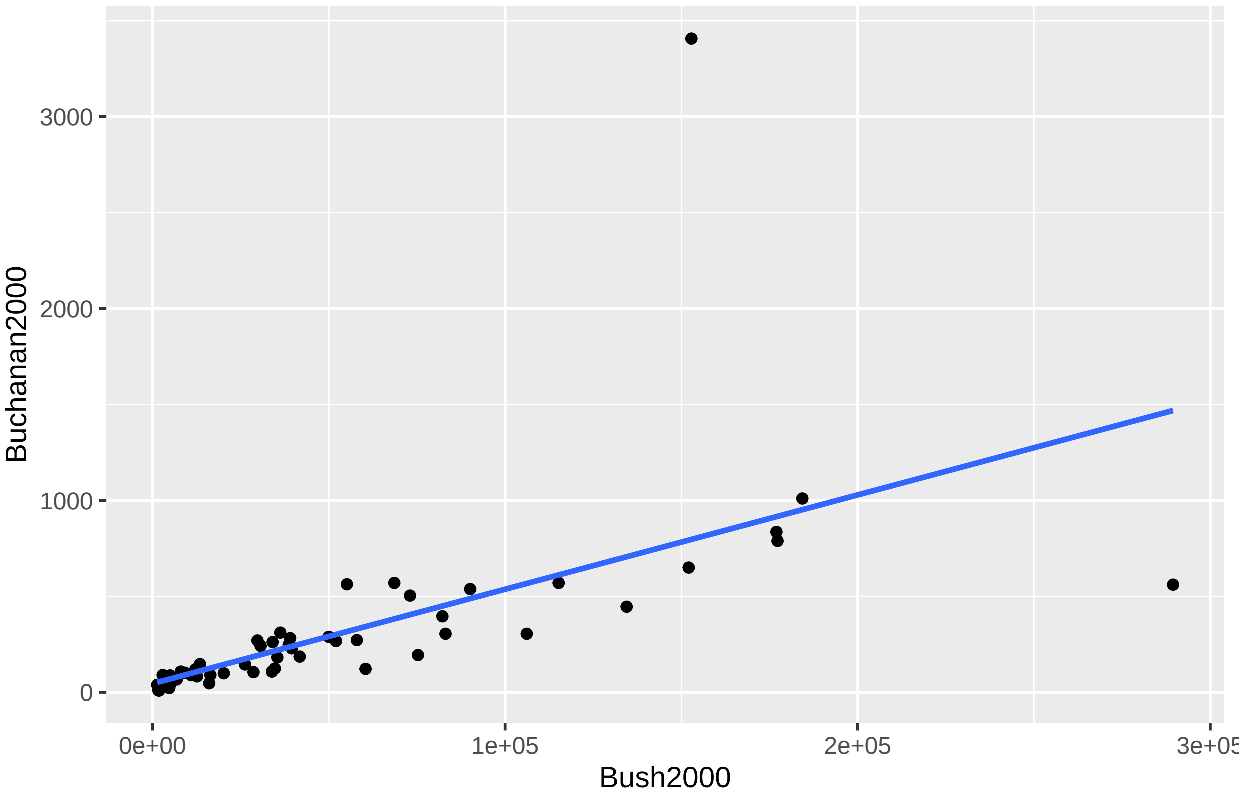
Name: Victor Huang

### Introduction

Democracy exists as the foundation which America was built on. Through voting, every citizen is able to take part of the country's democracy and make their voice heard. This ideology was put to the test during the 2000 election between Bush and Gore when democratic voters from Palm Beach County claimed that a percentage of Gore's vote was allegedly mis-allocated towards Buchanan, another presidential candidate, due to a poorly designed voting ballot. In order to make sure that everyone's voice was truthfully and accurately accounted for, we will be taking a look at the voting data between between Bush and Buchanan in Palm Beach County and seek to either dispel or prove this claim.

### Results

Votes for Bush vs. Buchanan in all Florida counties (Figure 1)



Looking at the scatterplot above, we see a representation of all Florida county votes for Bush against those for Buchanan. I also fit the regression of Bush's votes against Buchanan's votes to find a relationship between the two variables. We see that the relationship between Bush's votes and Buchanan's votes forms a mostly linear relationship. The direction also appears to be positive as Bush's votes and Buchanan's votes increase with each other. The strength of the relationship is also quite strong throughout the start of the graph. However, as we address our unusual cases, we see that there does exist a case at the top of the graph that severely deviates from the rest of the data points. And as we look locate that point from the dataset, do indeed find that it is the Palm Beach County votes data point. Based on the linear regression plotted, we get the following equation representation

$$\hat{\mu}(Bush2000|Buchanan2000) = 22911.1 + 79.1(Buchanan2000)$$

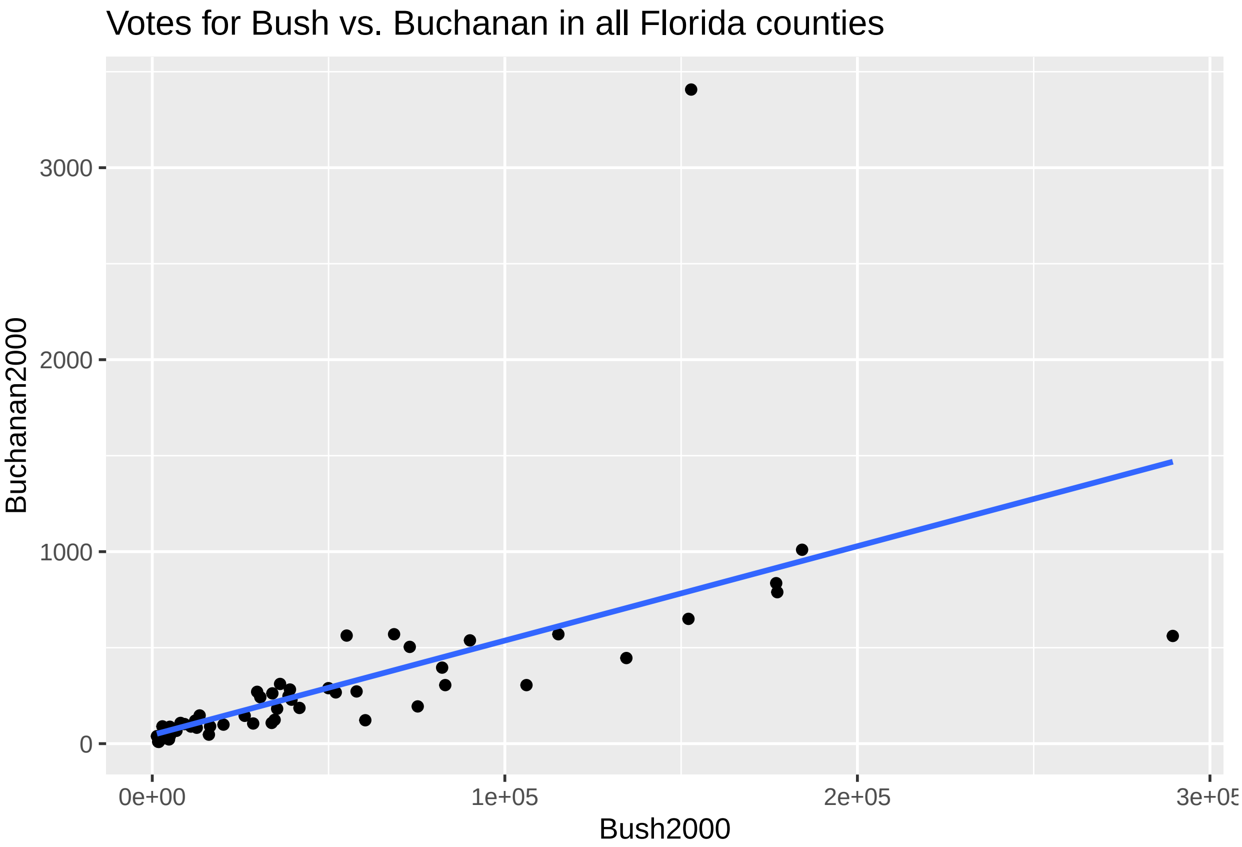
Our $\beta_0$ has a value of 22911.1 and our $\beta_1$ has a value of 79.1. From this we can see that for every one vote Buchanan gets, on average, Bush would get 79.1 votes. Checking our linear model, we get a SE of 12.3. Taking this information, we can calculate a 95% confidence interval between 54.53998, 103.6621, meaning that we are 95% confident the a one vote increase Buchanan's vote is associated with an increase in mean Bush votes of 54.54 to 103.66 points.

### Discussion

Summarizing the information above, we can see that Palm Beach County is indeed very unusual. We are initially shown this during our EDA of the scatterplot where it stands out as an outlier case among the data points. We hypothesize that Palm Beach County is indeed an outlier and we seek to prove it by using our linear regression model. Applying our linear regression model, we do in fact see that Palm Beach County does indeed fall outside of the 95% confidence interval. As such, we can conclude that there is some inaccuracy within the Palm Beach County votes. However, going through the question, I did encounter some limitations as well as some issues I would like to address. The first issue would be the lack of enough data. While 67 counties may seem like a large amount of data points, they are actually quite sparse in the scope of finding a linear regression model The second big issue would be that while we can be fairly certain that Buchanan's votes are definitely higher than they should be for Palm Beach County, we can not say for certain what the casue for this is. While it definitely could be as the democratic voters voting for the wrong candidate, it can also be caused by a variety of other reasons as it would be quite odd that only the Palm Beach County voters experienced this issue while the other county's seem quite normal.

### R Appendix

```
> ggplot(ex0825, aes(x = Bush2000, y = Buchanan2000)) + geom_point() +
+ geom_smooth(method = "lm", se = FALSE) +
+ labs(title = "Votes for Bush vs. Buchanan in all Florida counties")
```



```
> # creating the scatter plot for the dataset
> ex0825_lm <- lm(Bush2000 ~ Buchanan2000, data = ex0825) # creating the linear model
> predict(ex0825_lm, newdata = data.frame(Buchanan2000 = 3407), interval = "confidence", se.fit = TRUE)
$fit
       fit      lwr      upr
1 292408.3 214305.1 370511.5

$se.fit
[1] 39107.54

$df
[1] 65

$residual.scale
[1] 44890.7
> confint(ex0825_lm) #the confidence interval for the dataset
                   2.5 %     97.5 %
(Intercept)  10251.57966 35570.6153
Buchanan2000    54.53998   103.6621
> summary(ex0825_lm) # summary of the linear model

Call:
lm(formula = Bush2000 ~ Buchanan2000, data = ex0825)

Residuals:
    Min      1Q  Median      3Q     Max
-139562  -21969  -13626    4156  222169
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22911.1     6338.8   3.614 0.000588 ***
Buchanan2000     79.1       12.3   6.432 1.73e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44890 on 65 degrees of freedom
Multiple R-squared:  0.3889,    Adjusted R-squared:  0.3795 
F-statistic: 41.37 on 1 and 65 DF,  p-value: 1.727e-08
> 2*pt(-6.432,64) # calculating the p-value
[1] 1.824914e-08
> qt(.975 ,64) # calculating t* value
[1] 1.99773
```