Homework 1 - Linear Regression Review

update.packages()

NAME: Bryan Ellerbrock

NETID: bje24

DUE DATE: Friday, February 12, 2016 by 5pm

Homework 1 Instructions

1. For each problem:

    a) Answer all questions

    b) Insert code chunks directly under any problems that require you to use R and type in code as needed. In particular, make sure code chunks are included for any requested plots.

    c) Answer any questions related to the problem in the .Rmd document directly under the question

    d) Note: Occasionally when you insert a code chunk it may not go where you intend it to. If this happens, you can cut and paste it into the correct spot. Make sure the code chunk is aligned to the left margin of this document. Often it may be easier to just store a code chunk on your clip board and paste it in when you need one.

    e) You may need to knit your document occasionally to answer questions related to R output.

2. Submit your homework as a pdf document compiled from R markdown. Your file should be named *LastFirst*-HW1.pdf.

Question 1 (10 points)

Make up a data set that you will find to be a relevant source of examples. The data set can be anything, but is intended to be helpful for you to translate statistical model and tests into real world statements. Suggested sources: - a data set that you already have, or something made up to be close to it - a data set you expect to collect in future research - data that you would like to collect if only it were possible - data (possibly collected from somewhere else) associated with your field of study - data about some hobby or area of personal interest to you

You do not need to make up values for your data, but you should answer

(2 points)

a) How many experimental units (subjects, observations etc) are in the data?

173 plots of 9 different cassava accessions (grown in 6 different trials over 3 locations and 2 years)

b) What measurements are being taken?

Phenotypes that include measurements of 30 different agronomic, morphological, quality and stress traits. Genotypes that include measurements of nearly 200,000 single nucleotide polymorphisms. Of which maybe 1% are polymorphic between the 9 accessions.

In particular, you should record (but not be able to control) at least one of each of the following measures

(6 points)

a) Continuous variables that could take any value (may need to be positive) Plenty of these. The most important may be fresh root weight in kg.
b) Binary variables (yes/no, true/false, 1/0 etc) The snp genotypes are either binary or trinary, depending on the format. Disease traits are measured on a 1-5 scale, but could be converted to presence/absence to make them binary.
c) Count variables (0, 1, 2, . . .) Many examples, including fresh root count. (2 points)

Additionally, you should have some variables that you have controlled. These could be due to selection (specified numbers of men and women, specified numbers of dogs in certain weight categories) or because of active treatments you have applied (applying certain pesticides to some plants and not to others, for example).

Hmm. Within trials, all the phenotypes measured are responding variables. I guess the different accessions/genotypes are the manipulated or independent variable. And between the 6 different trials, the 3 different locations and 2 different years are specifically selected or controlled.

Question 2 (Adapted from Ramsey and Schafer Ex 8.25)

The U.S. presidential election of November 7, 2000 was one of the closest in history. As returns were counted on election night it became clear that the election would be determined by the outcome in the state of Florida. At one point in the evening the state was projected to be carried by the Democratic nominee, Al Gore but that projection was retracted a few hours later. Early in the morning of November 8, the networks projected that George W. Bush had carried the state and with it the presidency. This projection, too, was later retracted. It ultimately took weeks of legal arguments about voting irregularities to decide a winner.

One of the controversies at the time (although it did not play a significant role in the legal decisions) was the "butterfly ballot" of Palm Beach County. A number of Democratic voters complained that this ballot was confusing: voters were required to fill in a circle going down the middle of the ballot with the candidate's names on either side. Filling in the second circle counted as a vote for the Reform Party candidate, Pat Buchanan. However, since Al Gore's name appeared second on the left hand side of the ballot, it was speculated that some voters who intended to vote for Gore may have mistakenly filled in the second circle instead of the third.

Two pieces of evidence were advanced in support of this. The first was a large number of ballots discarded because both the second and third circles were filled in. The second was that Buchanan had an unusually high percentage of the vote in that county.

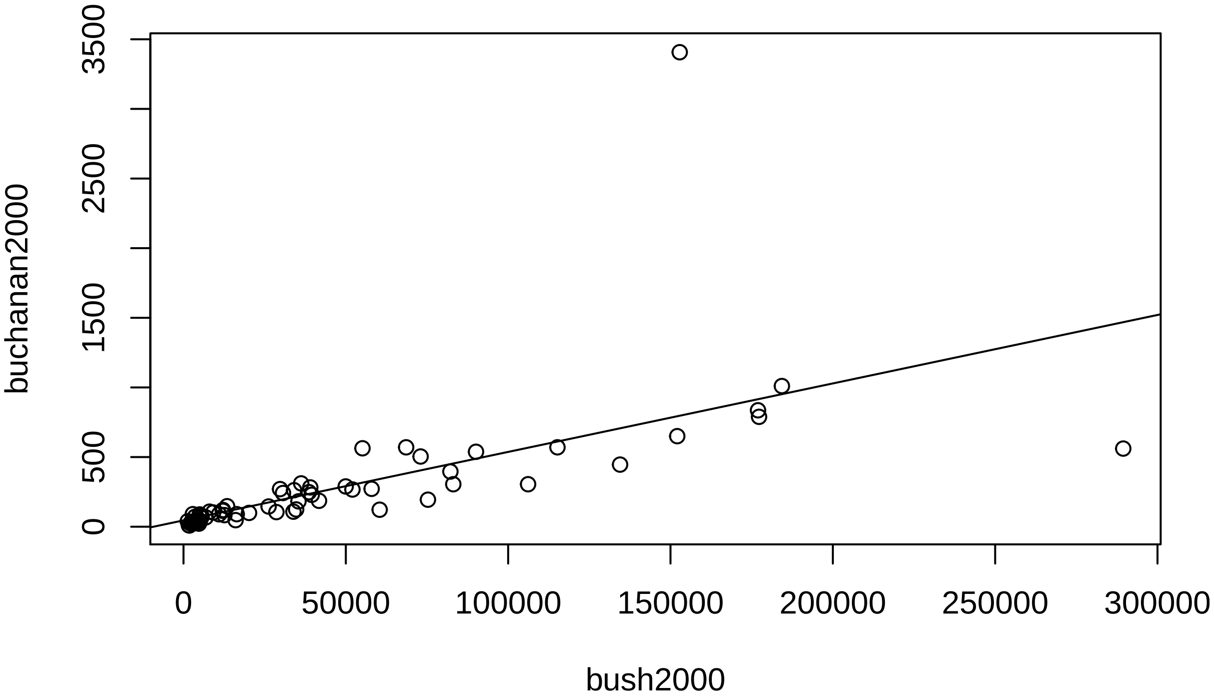
The file Election2000.csv contains the number of votes for George Bush and for Pat Buchanan in the 2000 election in each of the 67 Florida counties. The aim of this exercise is to determine if Palm Beach was anomalous among them.

1. (10 points) Plot the data along with a line of best fit. Do any patterns in the data emerge?

```r
elections200 <- read.csv("~/Personal/6020/homework/elections200.csv", header=TRUE)
summary(elections200)
```

```
##        county    buchanan2000       bush2000
##  ALACHUA : 1   Min.   :   9.0   Min.   :  1316
##  BAKER   : 1   1st Qu.:  46.5   1st Qu.:  4746
##  BAY     : 1   Median : 114.0   Median : 20196
##  BRADFORD: 1   Mean   : 258.5   Mean   : 43356
##  BREVARD : 1   3rd Qu.: 285.5   3rd Qu.: 56542
##  BROWARD : 1   Max.   :3407.0   Max.   :289456
##  (Other) :61
```

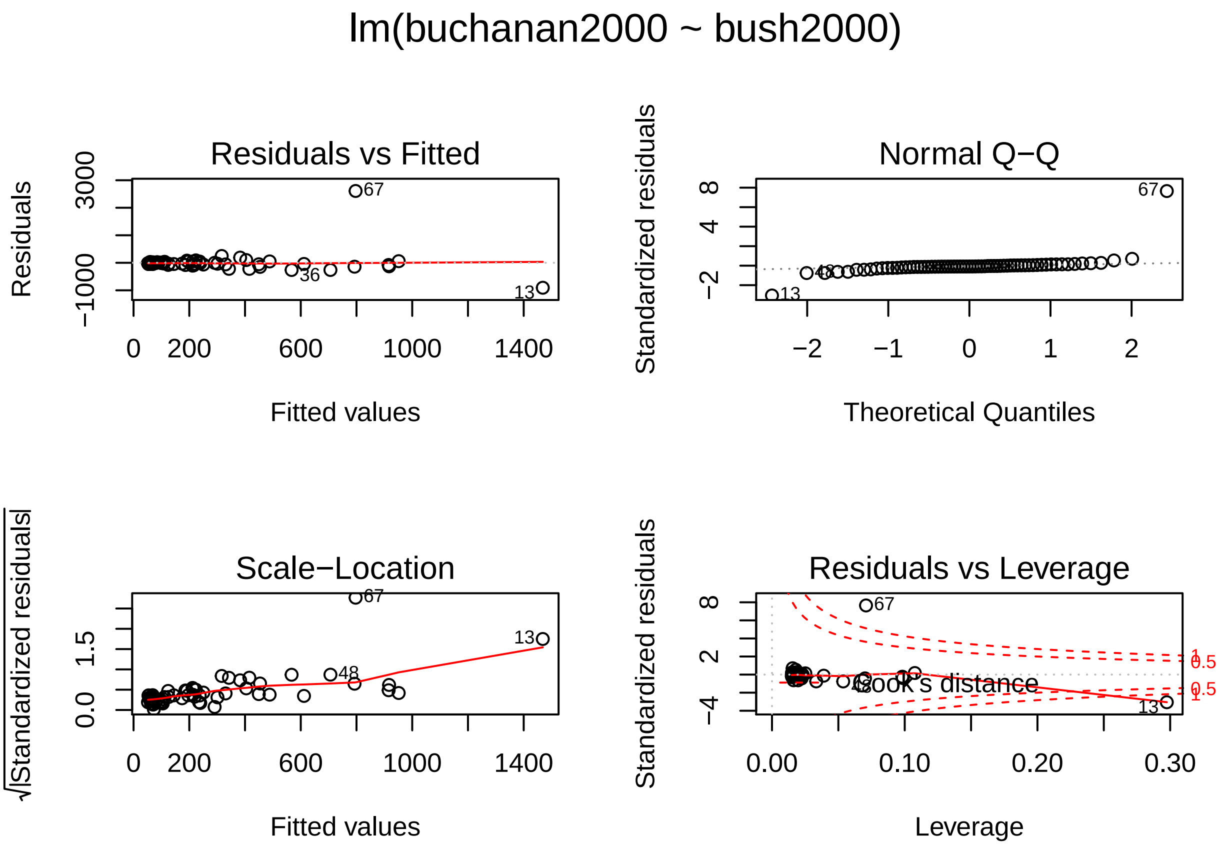
```r
attach(elections200)
bf <- lm(buchanan2000~bush2000)
plot(bush2000,buchanan2000)
abline(bf)
```



The data show a slightly positive linear relationship with at least one very clear outlier

2. (10 points) What is the influence of Palm Beach county on this fit as measured by cooks distance? Does it appear unusual? (The **cooks.distance()** function in R will provide cook's distance for a fitted model.)

```r
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(bf)
```



two counties look consistently problematic - they are 67 and 13

```r
elections200[c(13,67),1:3]
```

```
##         county buchanan2000 bush2000
## 13        DADE          561   289456
## 67 PALM BEACH          3407   152846
```

ok, the problem counties are dade and palm beach

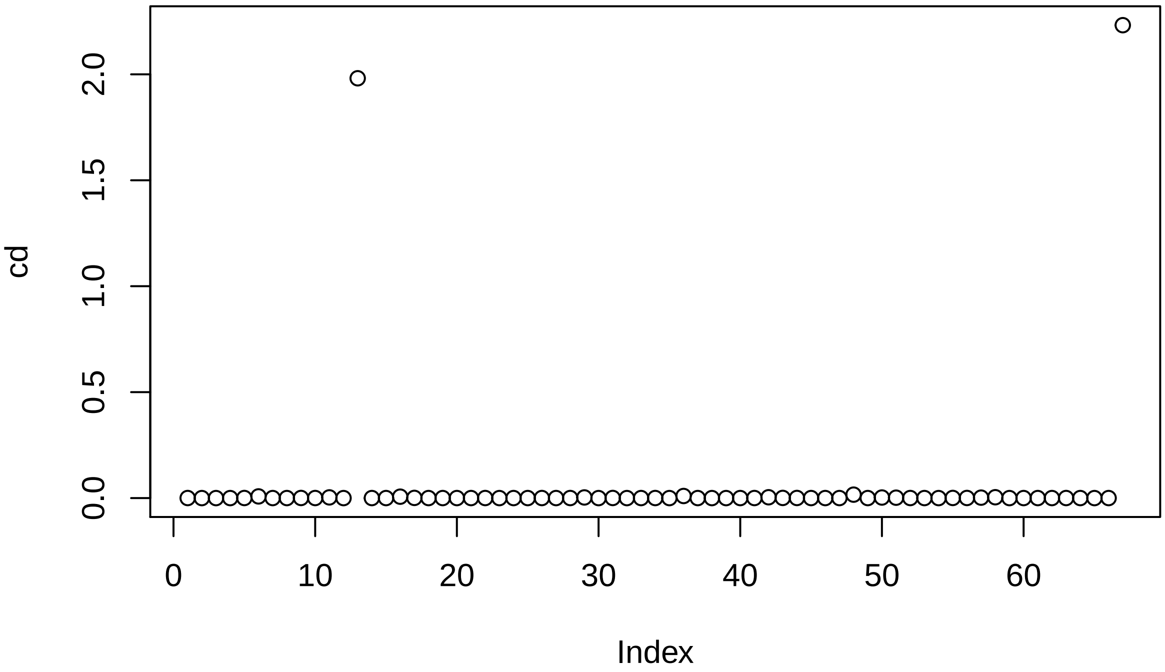
```r
cd <- cooks.distance(bf)
cd[13]
```

```
##       13
## 1.981366
```

```r
cd[67]
```

```
##       67
## 2.231935
```

```r
plot(cd)
```



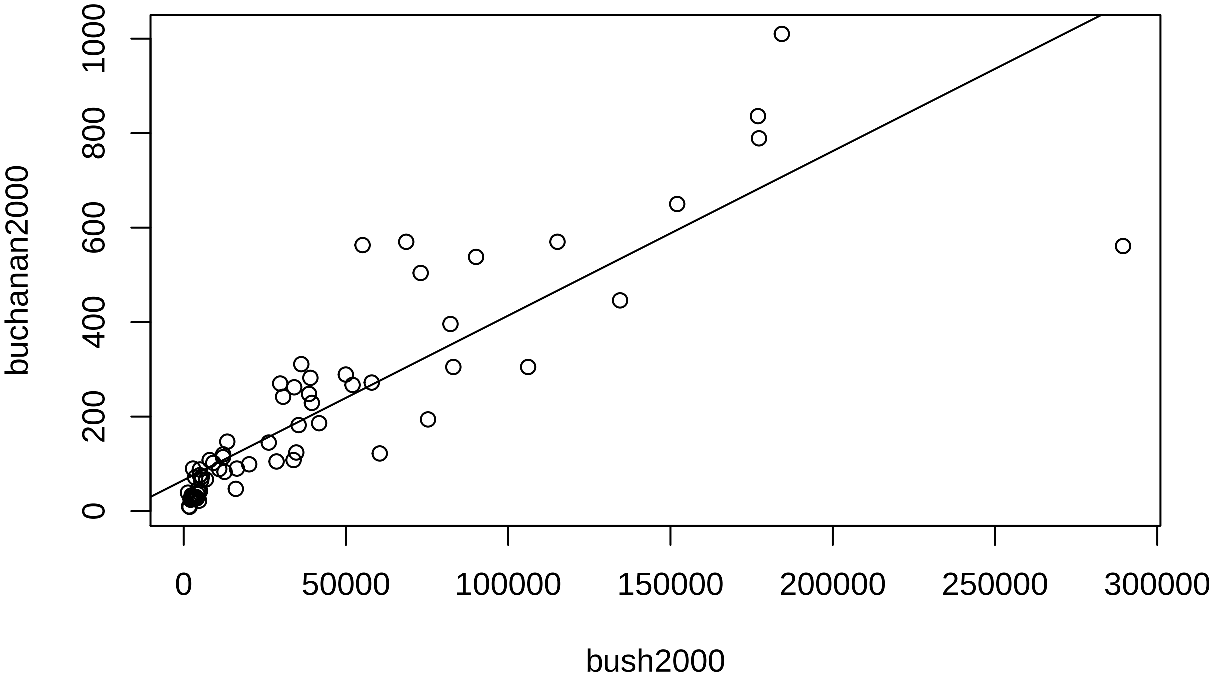
both dade and palm beach are very influential, with cooks distances of 1.98 and 2.23 respectively. Compared to they other counties, they are both unusual.

3. (10 points) Re-do the analysis without Palm Beach. Examine the residuals, does anything concern you about them? (Remember that **x**[**-j**] will remove the j'th element of **x**.)

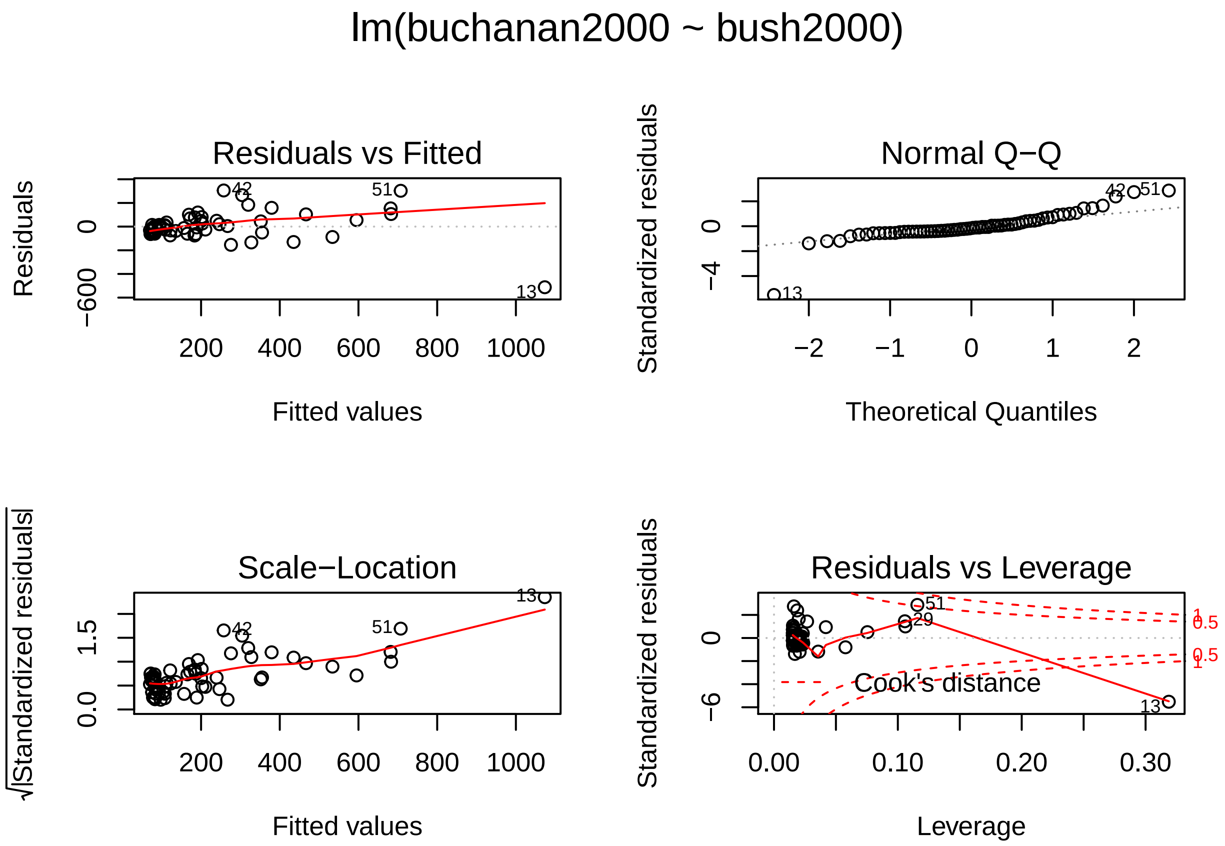
```
elect_wo_PB <- elections200[-c(67),]
summary(elect_wo_PB)
```

```
##       county    buchanan2000        bush2000
##  ALACHUA : 1   Min.   :   9.00   Min.   :  1316
##  BAKER   : 1   1st Qu.:  46.25   1st Qu.:  4745
##  BAY     : 1   Median : 111.00   Median : 18300
##  BRADFORD: 1   Mean   : 210.76   Mean   : 41697
##  BREVARD : 1   3rd Qu.: 279.50   3rd Qu.: 54362
##  BROWARD : 1   Max.   :1010.00   Max.   :289456
##  (Other) :60
```

```
detach(elections200)
attach(elect_wo_PB)
newbf <-lm(buchanan2000~bush2000)
plot(bush2000,buchanan2000)
abline(newbf)
```



```r
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(newbf)
```

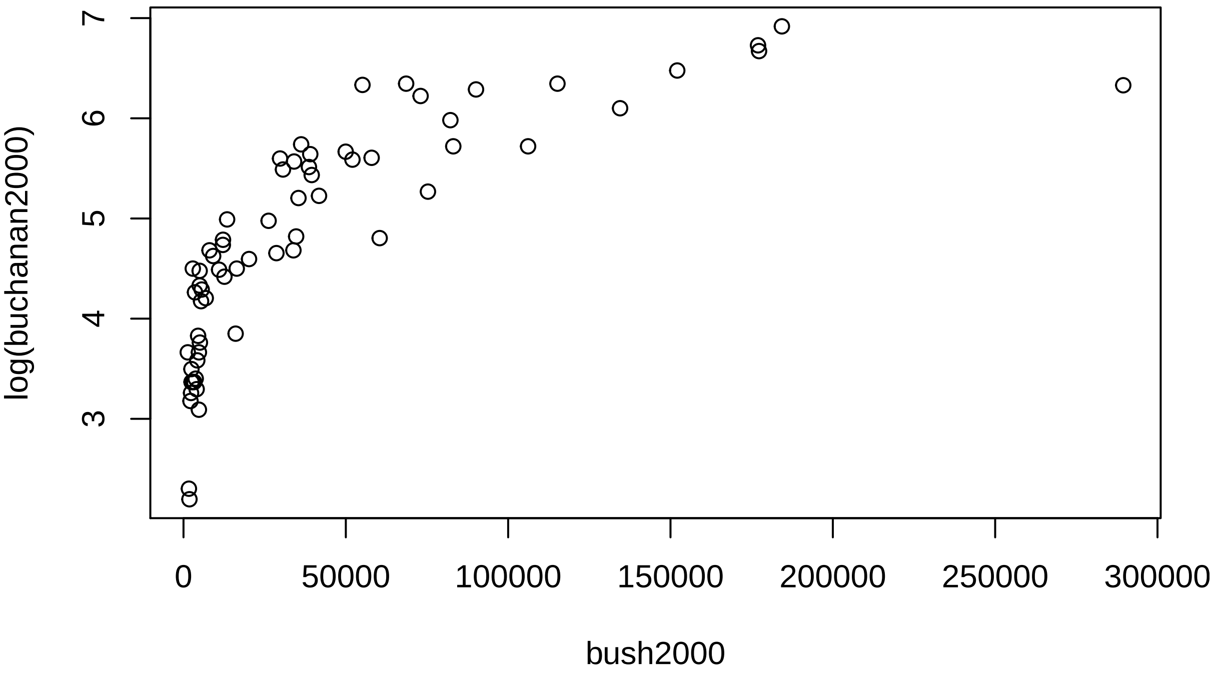


Yes. It looks like non-constant variance may be an issue, with larger residuals in larger counties. County 13, Dade county, also had a high cooks distance and stands out with a large residual and theoretical unlikeliness on the Q-Q plot.

4. (10 points) Try transforming the data (without Palm Beach) a number of ways. What model do you feel best fits the linear regression assumptions? Express the formula for the model.

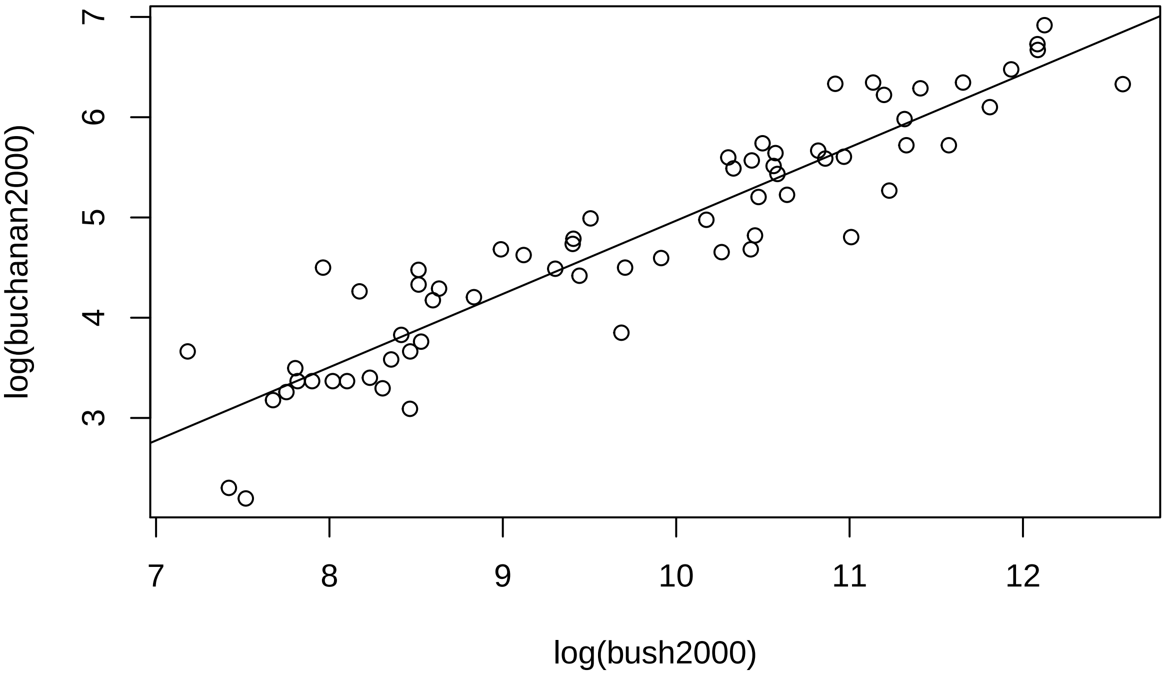
the data look mad heteroskedastic. The assumptions might be better met if it transform the Y variable with log()

```
bf3 <- lm(log(buchanan2000)~bush2000)
plot(bush2000,log(buchanan2000))
```

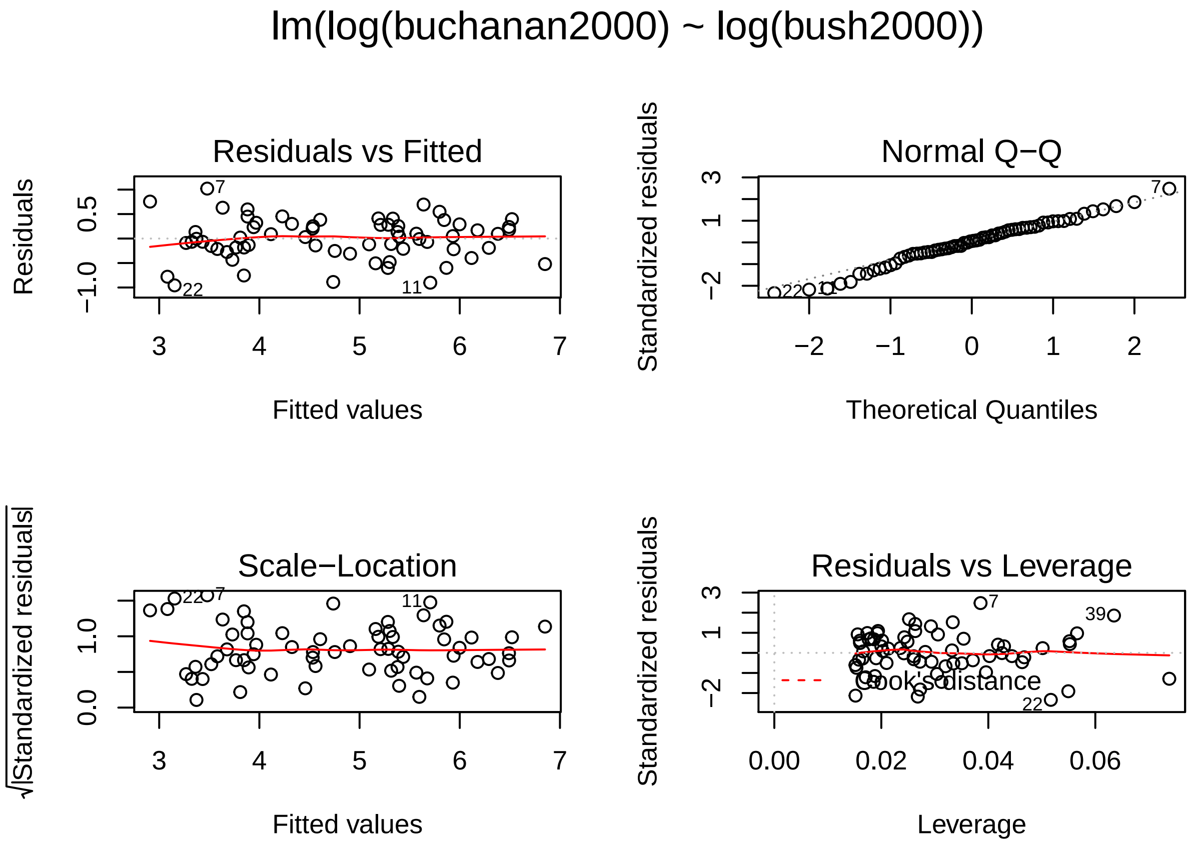


hmm, not enough, let's try transforming both.

```
bf4 <- lm(log(buchanan2000)~log(bush2000))
plot(log(bush2000),log(buchanan2000))
abline(bf4)
```



```
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(bf4)
```



wow, that looks much better.

5. (10 points) Is the model significant at the 0.05 level? How does the model imply the number of votes

for Buchanan will change as the number of votes for Bush increases?

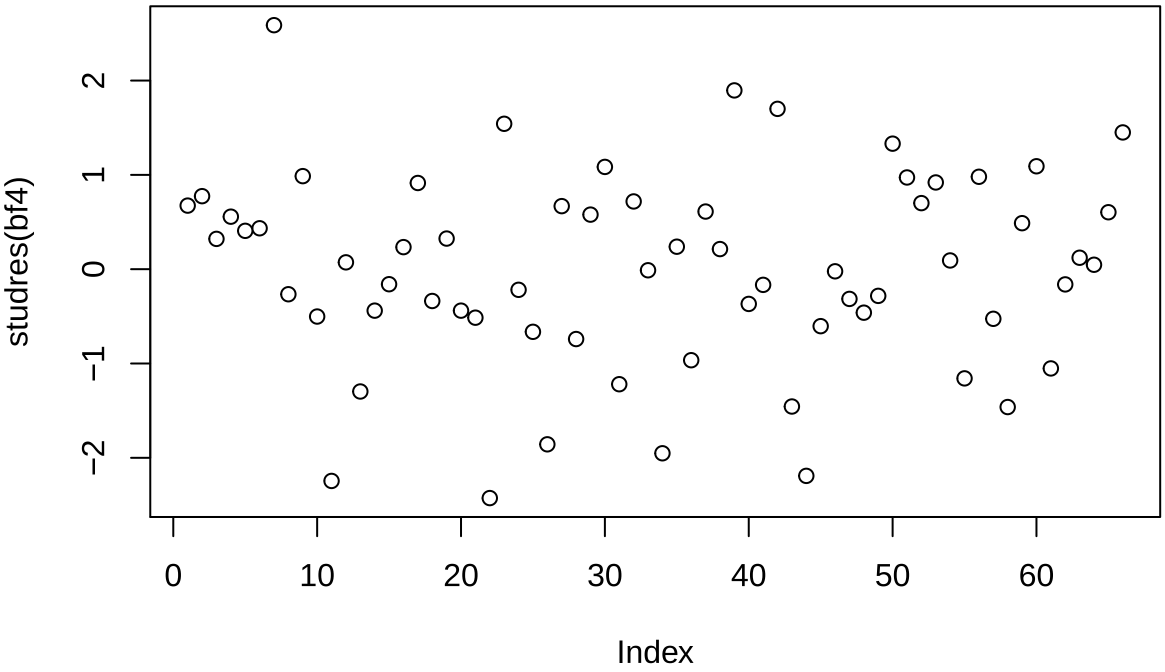
```
summary(bf4)
```

```
##
## Call:
## lm(formula = log(buchanan2000) ~ log(bush2000))
##
## 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
```

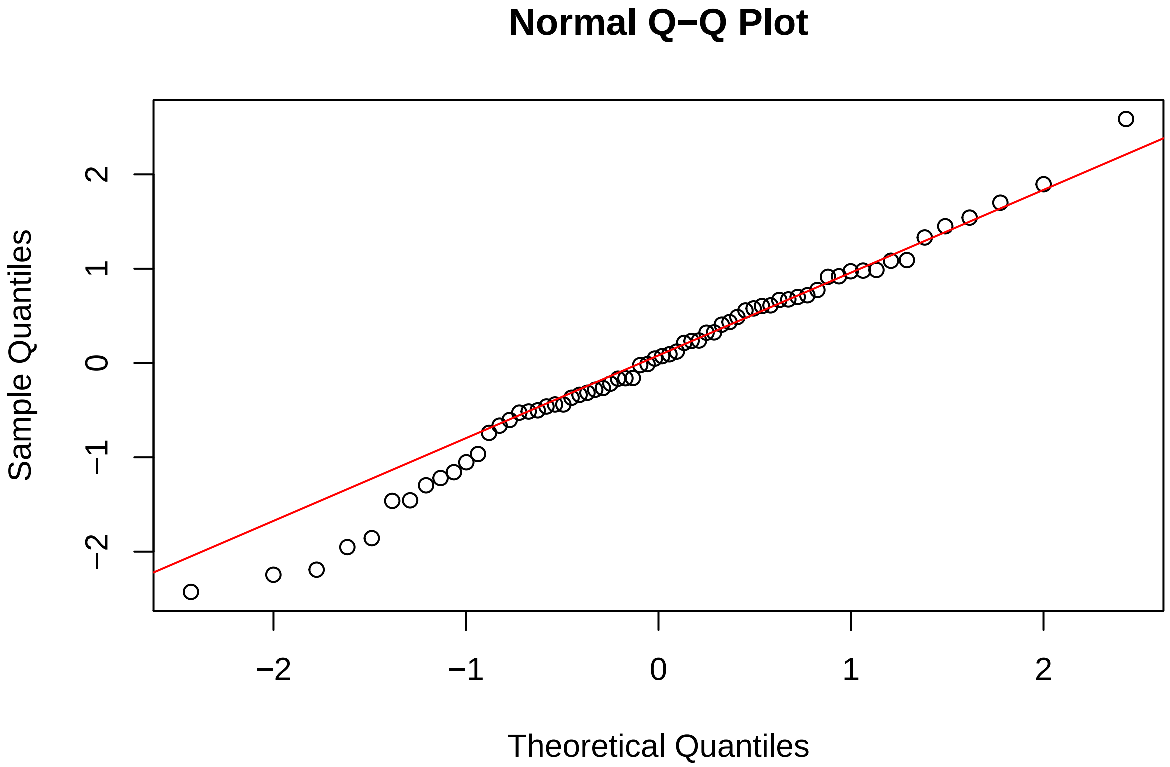
yes, the linear model using the log of both variables is highly significant, with a p-value well below 0.001. It predicts the log of the expected number of votes for Buchanan will rise by 0.73 for each increase of 1 in the log of votes for Bush.

6. (10 points) Plot the studentized residuals and produce a QQ plot of the residuals for your final model. Do the normal assumptions appear to have been met? In the MASS library, the **studres()** function will provide studentized residuals.

```
library(MASS)
plot(studres(bf4))
```



```r
qqnorm(studres(bf4))
qqline(studres(bf4),col="red")
```



Yes, after transforming both variables there are no major deviations from the theoretical line - the data appear to be normally distributed.

7. (10 points) Produce a prediction interval for the number of votes for Pat Buchanan in Palm Beach

county. Does the observed number of votes for Buchanan fall within this interval? You can obtain prediction intervals from the **predict()** function by specifying **interval="prediction"**.

```
palm_beach <- elections200[c(67),]
pb_bush <- palm_beach[3]
log(pb_bush)
```

```
##    bush2000
## 67 11.93719
```

```
pb_prediction <- predict(bf4, pb_bush, interval = "prediction")
exp(pb_prediction)
```

```
##         fit      lwr      upr
## 67 592.3769 250.8001 1399.164
```

My model predicts 592 votes for Buchanan in Palm Beach county, with a 95 percent prediction interval of between 251 and 1399 votes. The observed number of votes for Buchanan in Palm Beach county was 3407, a value far above the upper limit of the prediction interval.

8. (10 points) *Carefully* interpret your statistical findings. On what assumptions do they rely? What does this analysis tell you about what Gore's vote would have been without the butterfly ballot?

These findings are assuming that errors are normaly distributed, and the variance of errors is independent and constant across observations. After all of our error checking and our variable transformations, these assumptions appear to be reasonable for my final model. Finally, this prediction assumes that Palm Beach is not fundamentally different than the counties used to build the model. When predicting the number of votes for Buchanan in Palm Beach, we expect the true number of buchannan votes in palm beach to fall within our prediction interval 19 out of 20 times. In reality the number of votes for Buchanan in Palm Beach county was much, much higher than the limits of our prediction interval.

Based on this, we should reject the assumption that the rest of Florida's counties are representative of Palm Beach. Something about voting in Palm Beach is different. The butterfly ballot is a very plausible explantion for why Buchannan has so many more votes, but I don't know that we have the data necessary to test that explanation, or to predict how much higher Gore's total would have been.

Even if we did, 16 years later, it might be too painful to answer that question. How about we ignore politics and focus on biology!