HW1_Echo_Liu

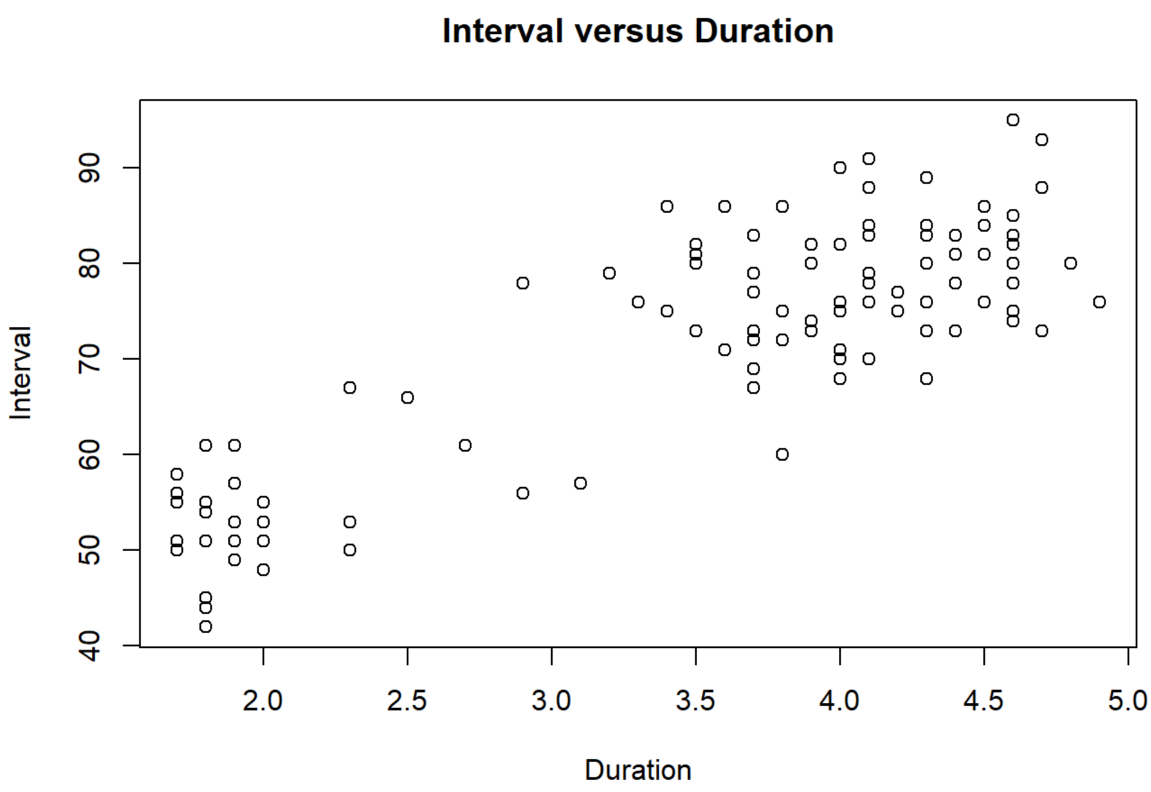
Bingying(Echo) Liu

September 4, 2018

Problem 1: Old Faithful

Step 1: Exploratory data analysis

```
#read in the OldFaithfulGeyser data-- data on durations of Old Faithful eruptions and intervals until subse
quend eruption
OldFaithful <- read.csv("C:/Users/Echo Liu/Downloads/Duke University/1st semester/702_Modeling_and_Represen
tation/HW1/OldFaithful.csv")

plot(OldFaithful$Interval ~ OldFaithful$Duration,
     xlab = "Duration",
     ylab = "Interval",
     main = "Interval versus Duration")
```



Step 2: Fit a linear regression model

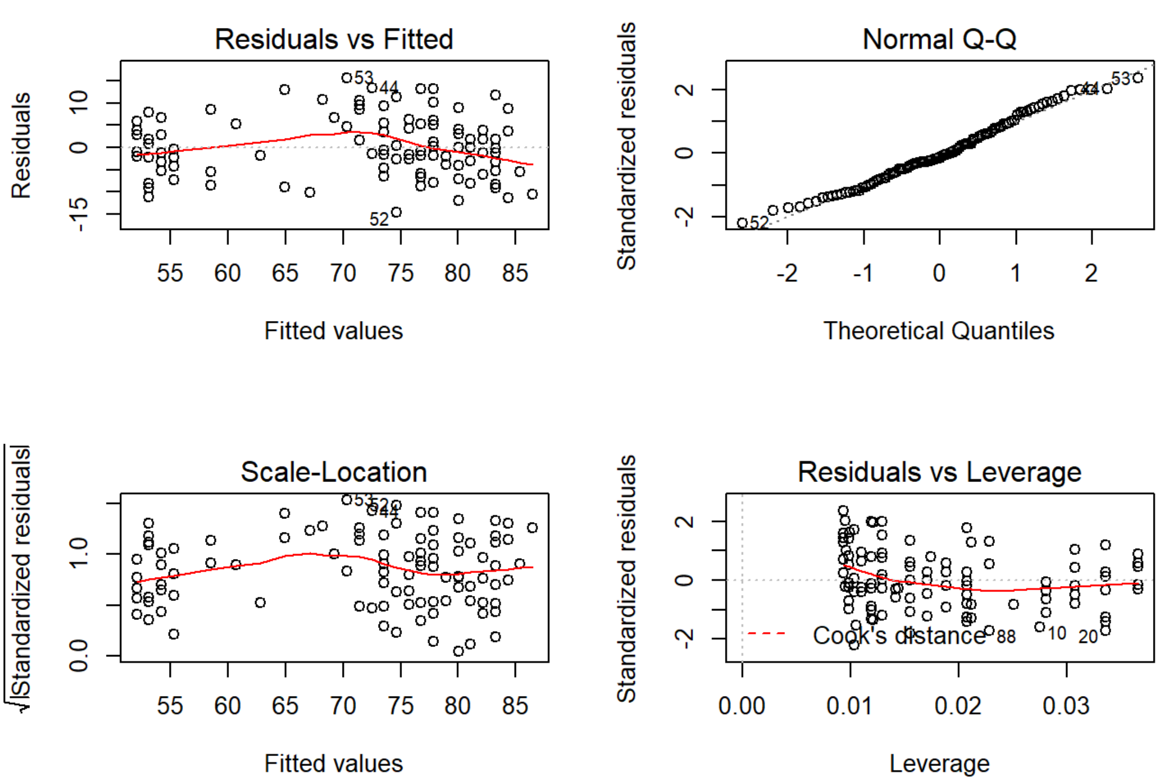
```
fit <- lm(Interval ~ Duration, data = OldFaithful)
summary (fit)
```

```
##
## Call:
## lm(formula = Interval ~ Duration, data = OldFaithful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.440  -1.088   4.467  15.652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8282     2.2618   14.96   <2e-16 ***
## Duration     10.7410     0.6263   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 105 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7344
## F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

Step 3: Use R to create the regression diagnostic plots

Since in the residual plot (upper-left), there is only random scatter and no discernable pattern (LOESS curve is almost a horizontal line), linearity assumption is met. Furthermore, points are noramlly distributed across all fitted values, thus constant variance assumption is met as well. From the normal Q-Q plot, we can see that all of the points fall near a line. Thus, normality assumption is satisfied. Finally, given how data is collected, we can conclude that independency assumption is also satisfied.

```
par(mfrow=c(2,2))
plot(fit)
```



Step 4: Get 95% confidence intervals for the coefficients

Interpretation: For every minute increase in duration, we expect an increase in interval by 10.74 minutes (95% CI: 9.499, 11.983).

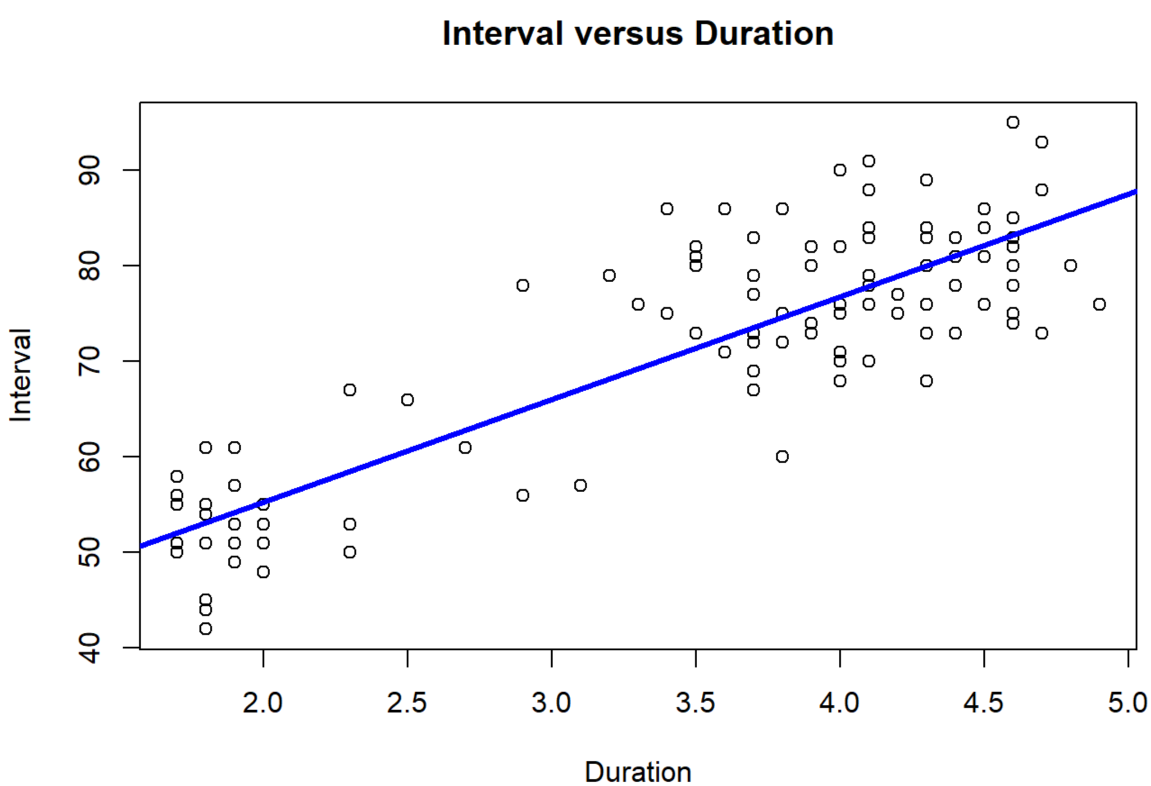
```
confint(fit,level = 0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 29.343441 38.31297
## Duration     9.499061 11.98288
```

Step 5: Add the regression onto the scatter plot

```
plot(OldFaithful$Interval ~ OldFaithful$Duration,
     xlab = "Duration",
     ylab = "Interval",
     main = "Interval versus Duration")+
  abline(fit,col='blue',lwd=3, data = OldFaithful)
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "data"
## is not a graphical parameter
```




```
## integer(0)
```

Step 6: Make predictions

The 95% prediction interval for the waiting time until the next eruption if the duration of the previous one was 4 mins is (63.4631, 90.12108).

```
newduration = c(4)
newdata1 = data.frame(Duration = newduration)
predict.lm(fit, newdata1, interval = "prediction")
```

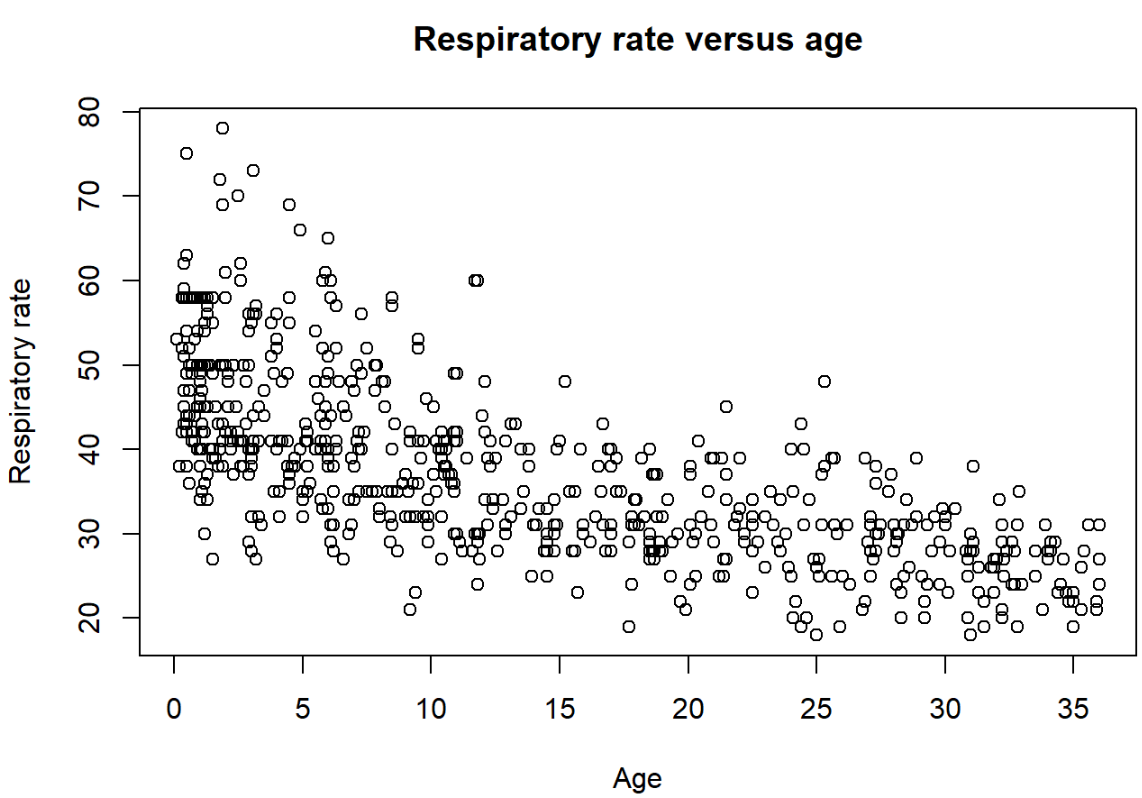
```
##        fit     lwr      upr
## 1 76.79209 63.4631 90.12108
```

Problem 2: Respiratory Rates for Children

Step 1: Exploratory data analysis

From the scatter plot, we could see that the point cloud has a curve/quadratic tendency instead of linear property, thus linearity assumption is not met. Secondly, variance across x becomes smaller as x gets bigger. Therefore, constant variance assumption is also not satisfied.

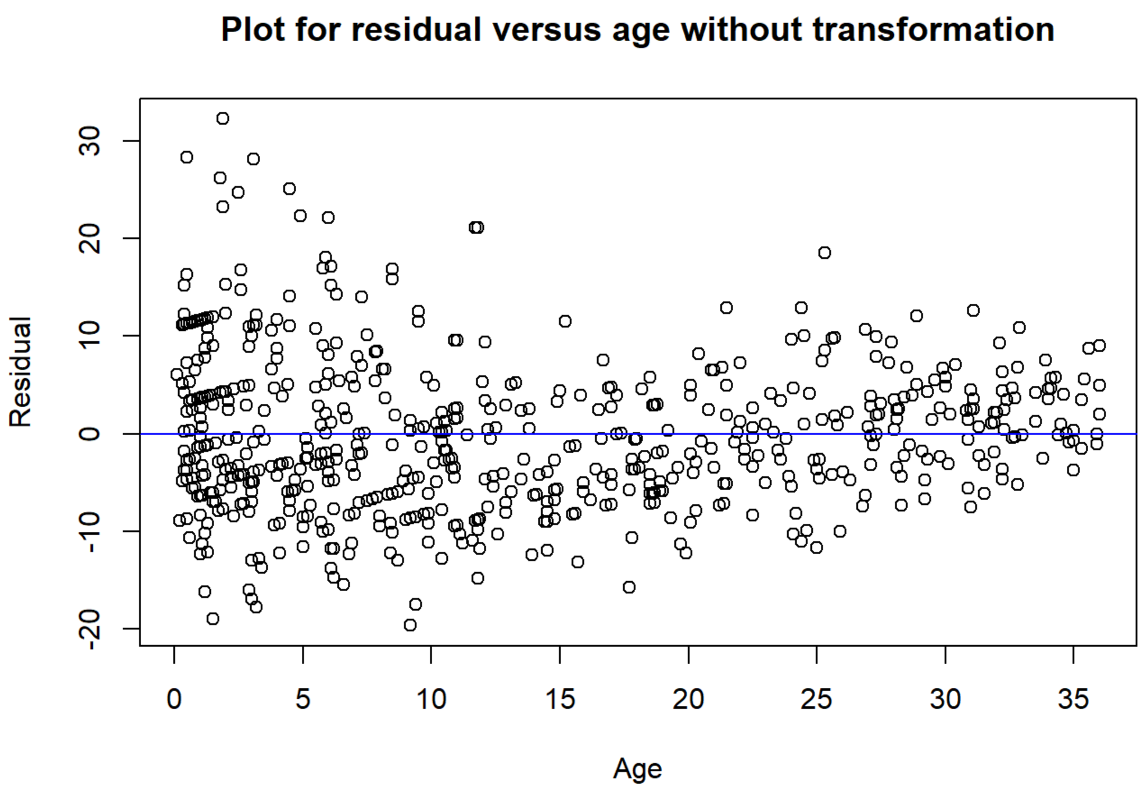
```
Respiratory <- read.csv("C:/Users/Echo Liu/Downloads/Duke University/1st semester/702_Modeling_and_Represen
tation/HW1/Respiratory.csv")

#Exploratory data analysis
plot(Respiratory$Rate ~ Respiratory$Age,
     xlab = "Age",
     ylab = "Respiratory rate",
     main = "Respiratory rate versus age"
     )
```



Step 2: Residual plot (non-transformation)

The resivual plot verifies that my observation above is correct. The points are not randomly scattered and variance across x becomes smaller, which means we need to do some transformations.

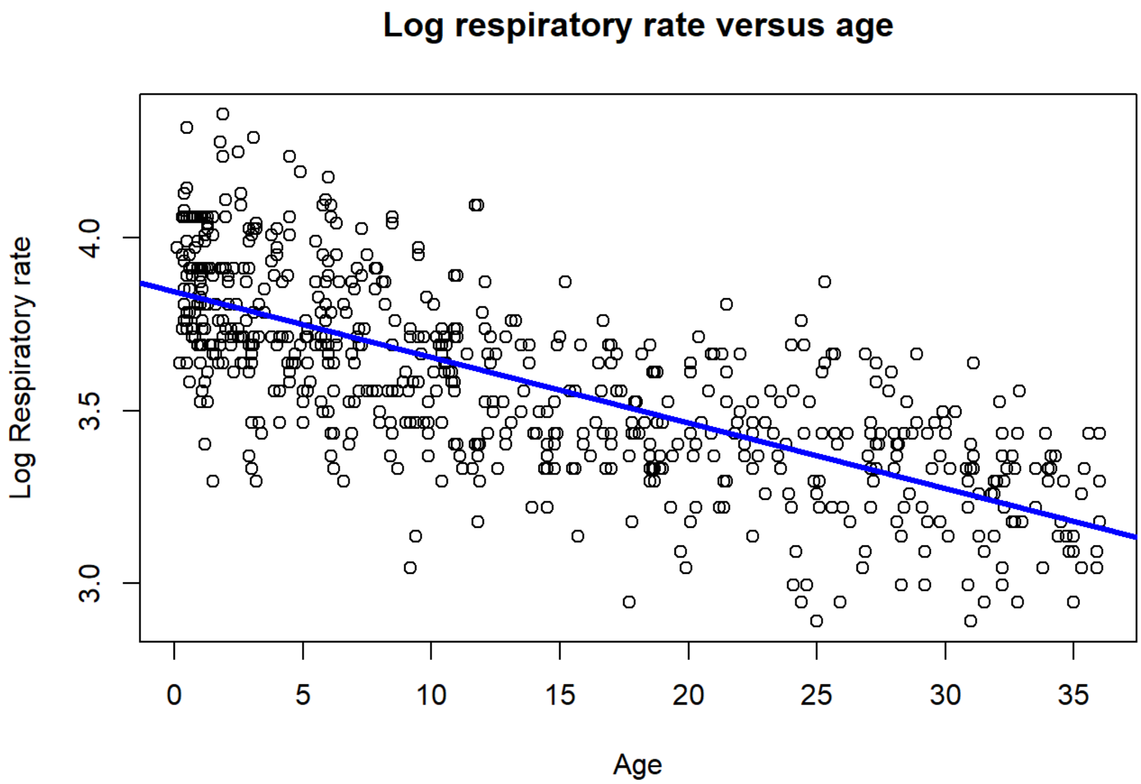
```
resfit <- lm(Rate ~ Age, data = Respiratory)

plot(y=resfit$residuals, x=Respiratory$Age,
     xlab = "Age",
     ylab = "Residual",
     main = "Plot for residual versus age without transformation")
abline(0,0,col="blue")
```



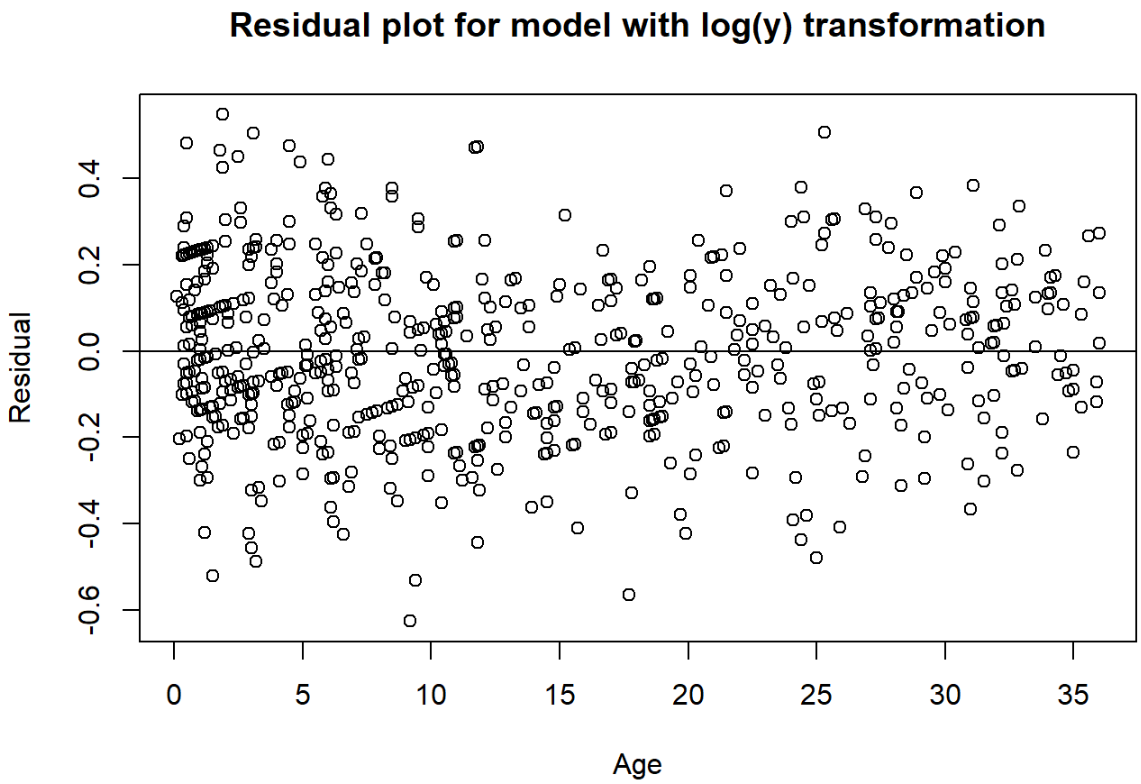
Step 3: Try tranforming with log(y)

When we transfrom y to log(y), points in the residual plot are randomly scattered around 0 and variance of residuals are almost constant across all Xs. Therefore, linearity and constance variance assumptions are satisfied. The norm qq plot shows that points fall near a line. Thus, noramlity assumption is also satisfied. Finally, given how data was collected ,we can conclude that independency assumption is met.

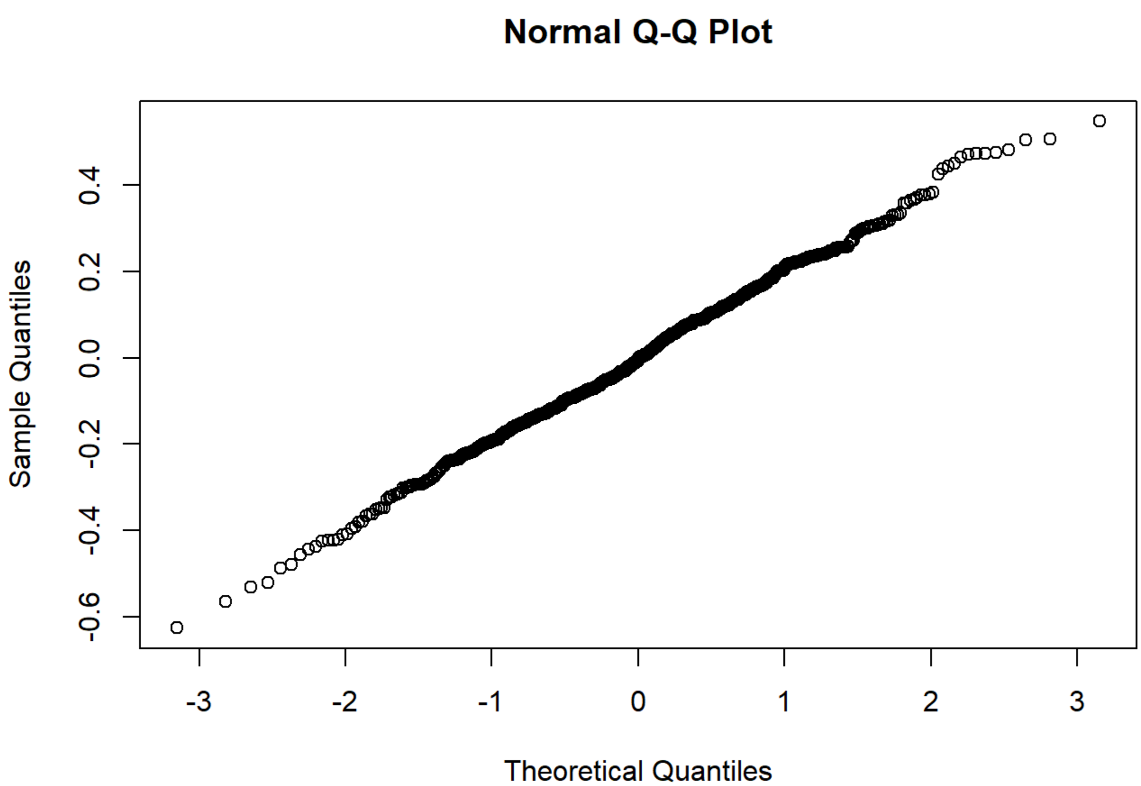
```
Respiratory$lograte = log(Respiratory$Rate)
plot(y= Respiratory$lograte, x=Respiratory$Age,
     xlab = "Age",
     ylab = "Log Respiratory rate",
     main = "Log respiratory rate versus age"
     )

reslogratefit <- lm(lograte ~ Age, data = Respiratory)
abline(reslogratefit,col='blue',lwd=3)    #add the regression onto the scatter plot
```




```
plot(y=reslogratefit$residuals, x=Respiratory$Age,
     xlab = "Age",
     ylab = "Residual",
     main = "Residual plot for model with log(y) transformation")
abline(0,0)
```




```
qqnorm(reslogratefit$residuals)
```




```
summary(reslogratefit)
```

```
##
## Call:
## lm(formula = lograte ~ Age, data = Respiratory)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62571 -0.13201 -0.00402  0.13489  0.54771
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.8451185  0.0126277  304.50   <2e-16 ***
## Age         -0.0190090  0.0007357  -25.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1964 on 616 degrees of freedom
## Multiple R-squared:  0.5201, Adjusted R-squared:  0.5193
## F-statistic: 667.6 on 1 and 616 DF,  p-value: < 2.2e-16
```

Step 4: Make predictions

95% prediction intervals for the rate of children who are individually 1 month old, 18 months old and 29 months old are (31.177,67.527), (22.576,48.863), (18.305,39.667) respectively.

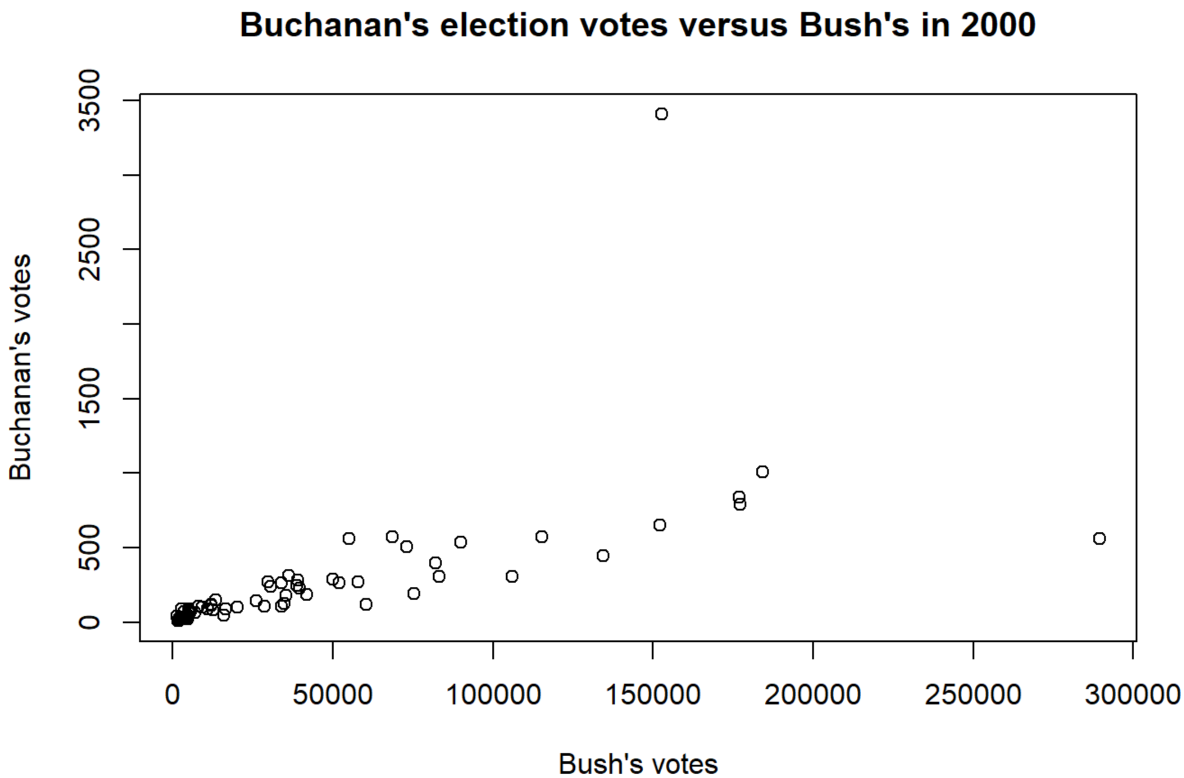
```
newage = c(1,18,29)
newdata2 = data.frame(Age = newage)
exp(predict.lm(reslogratefit, newdata2, interval = "prediction"))
```

```
##        fit      lwr      upr
## 1 45.88368 31.17725 67.52721
## 2 33.21353 22.57614 48.86302
## 3 26.94664 18.30537 39.66714
```

Problem 3: The Dramatic U.S. Presidential Election of 2000

Step 1: Exploratory data analysis

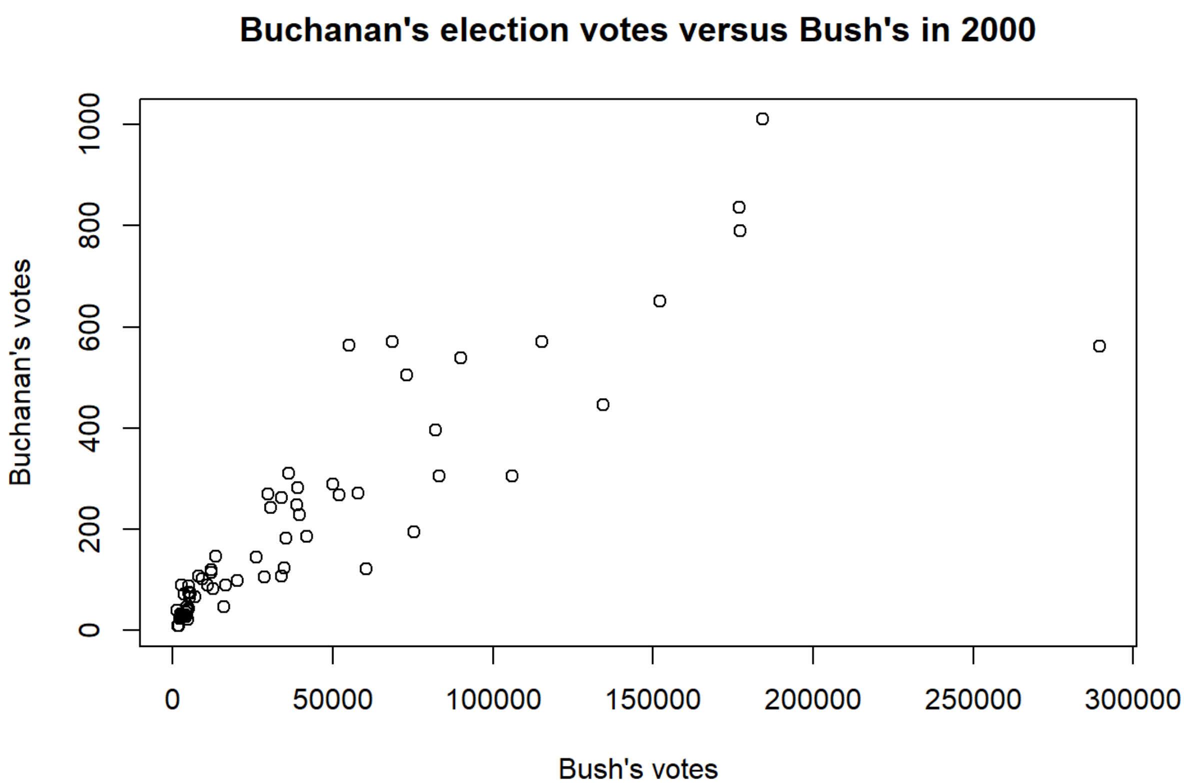
```
#let's read in the data and plot an innitial scatter plot
Elections <- read.csv("C:/Users/Echo Liu/Downloads/Duke University/1st semester/702_Modeling_and_Representa
tion/HW1/Elections.csv", header=TRUE, stringsAsFactors=FALSE)

plot(Elections$Buchanan2000 ~ Elections$Bush2000,
     xlab = "Bush's votes",
     ylab = "Buchanan's votes",
     main = "Buchanan's election votes versus Bush's in 2000")
```




```
#In this scatter plot, we can see a major outlier with Buchanan's votes over 3000. Compared to the second l
argest amount of Buchanan's votes which is approximately 1000, this is a 2000 difference. Thus, we look int
o the data again to figure out which row of data is causing the outlier problem.
Elections[Elections$Buchanan2000 > 3000,]
```

```
##     X     County Buchanan2000 Bush2000
## 67 67 Palm Beach         3407   152846
```

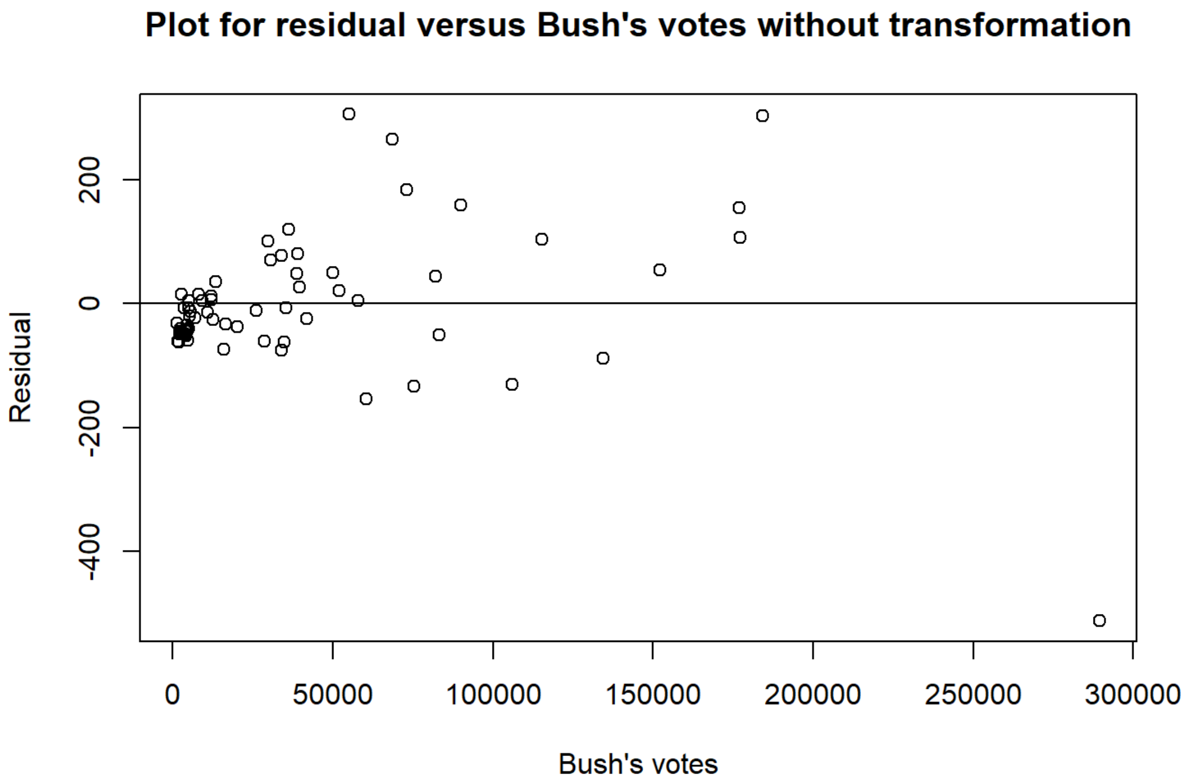
```
#Let's exclude the palm beach result and try to predict Buchannan's votes according to Bush's votes in palm
 county.
TestElections <- Elections[-67,]

plot(TestElections$Buchanan2000 ~ TestElections$Bush2000,
     xlab = "Bush's votes",
     ylab = "Buchanan's votes",
     main = "Buchanan's election votes versus Bush's in 2000")
```



Step 2: A linear model will not fit these data (see the redidual plot)

Residual plot shows that neither plots are randomly scattered nor are variance of the points constant through the x-axis. Therefore, we need to do some transformations.

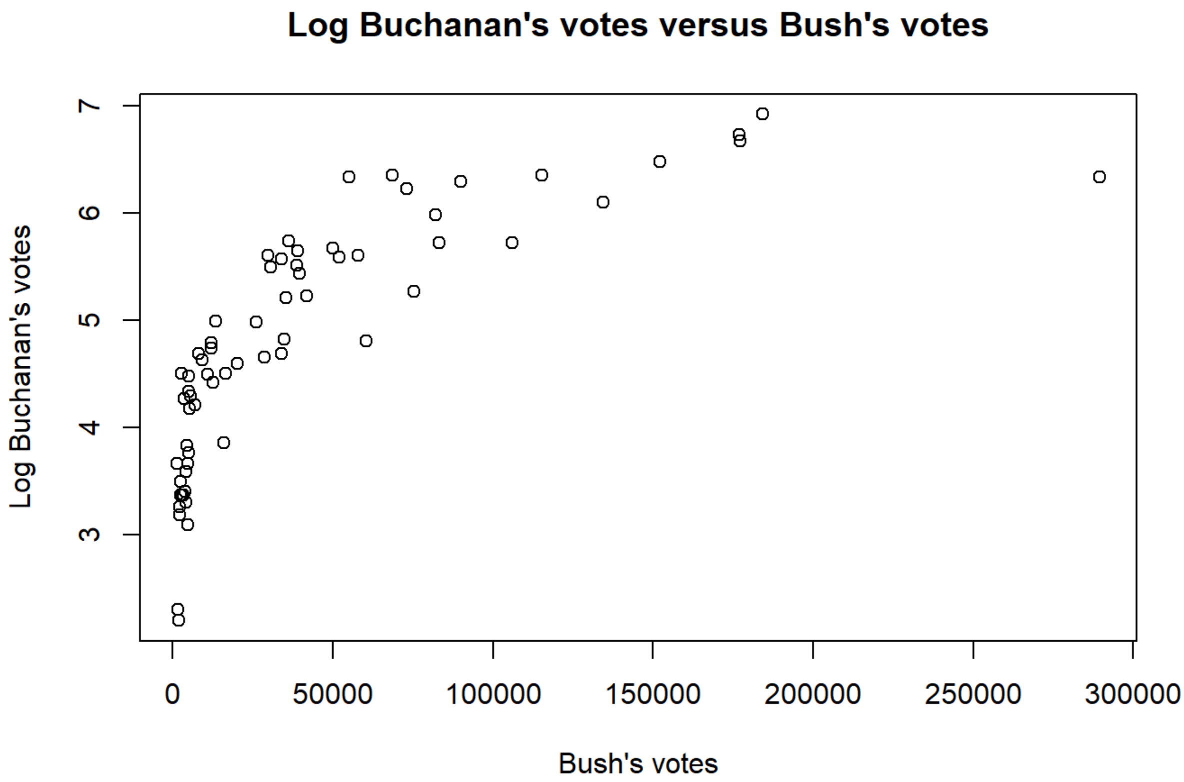
```
elefit <- lm(Buchanan2000 ~ Bush2000, data = TestElections)
plot(y=elefit$residuals, x=TestElections$Bush2000,
     xlab = "Bush's votes",
     ylab = "Residual",
     main = "Plot for residual versus Bush's votes without transformation")
abline(0,0)
```



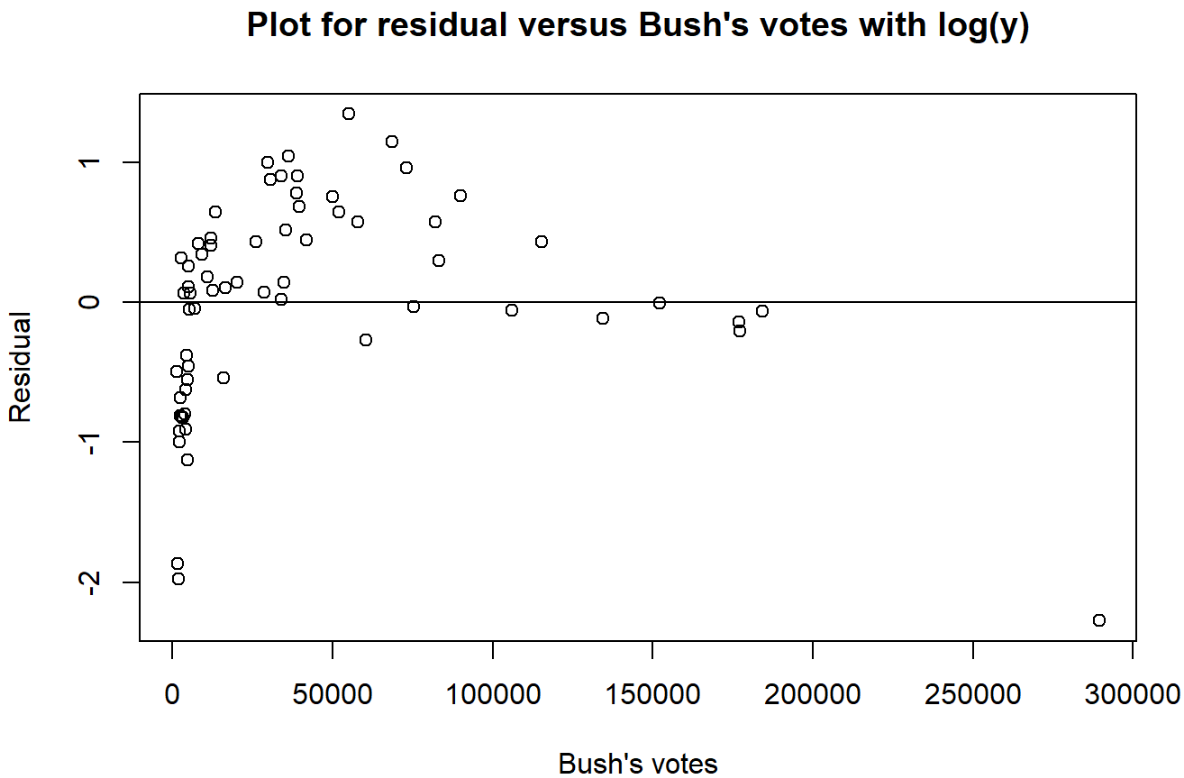
Step 4: Try tranforming with log(y)

The scatter plot of log(y) versus x shows a non-linear trend, so we need to do figure out other transformations to build a linear model.

```
TestElections$logBuchanan = log(TestElections$Buchanan2000)
plot(y= TestElections$logBuchanan, x=TestElections$Bush2000,
     xlab = "Bush's votes",
     ylab = "Log Buchanan's votes",
     main = "Log Buchanan's votes versus Bush's votes"
)
```



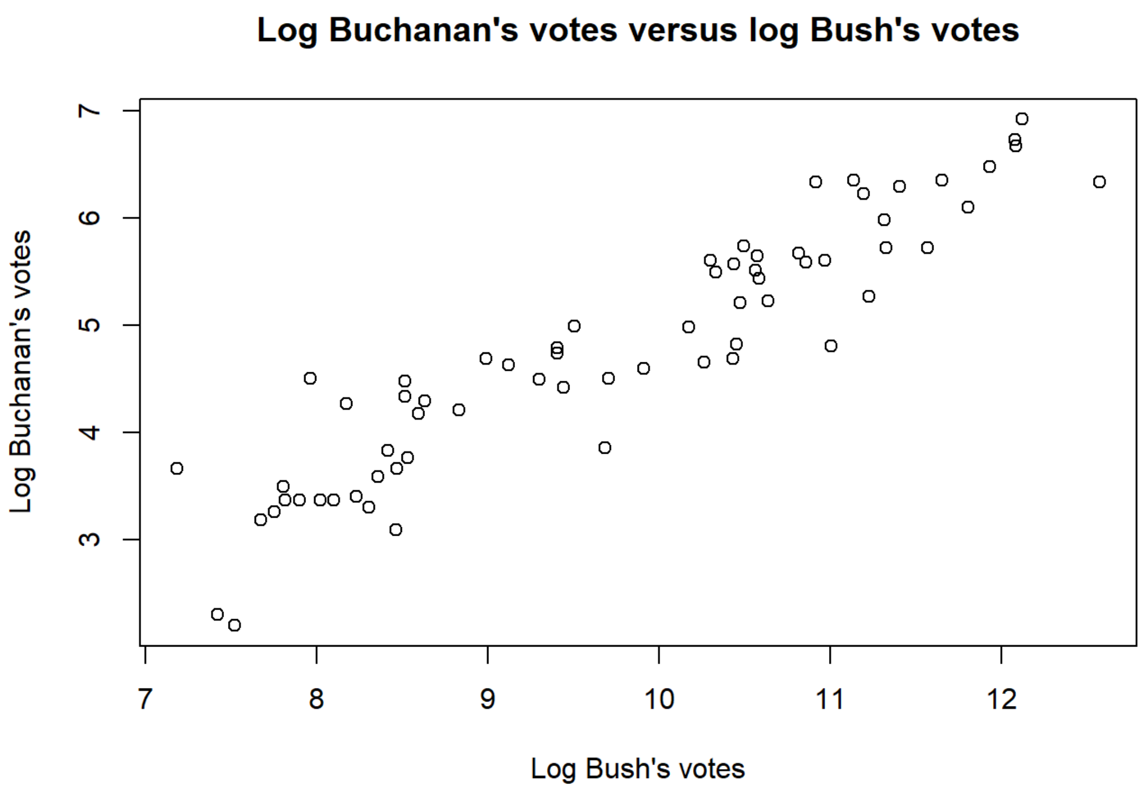

```
elelogBuchananfit <- lm(logBuchanan ~ Bush2000, data = TestElections)  #Doesn't satisfy linearity assumptio
n, again we can verify this in the scatter plot

plot(y=elelogBuchananfit$residuals, x=TestElections$Bush2000,
     xlab = "Bush's votes",
     ylab = "Residual",
     main = "Plot for residual versus Bush's votes with log(y)")
abline(0,0)
```



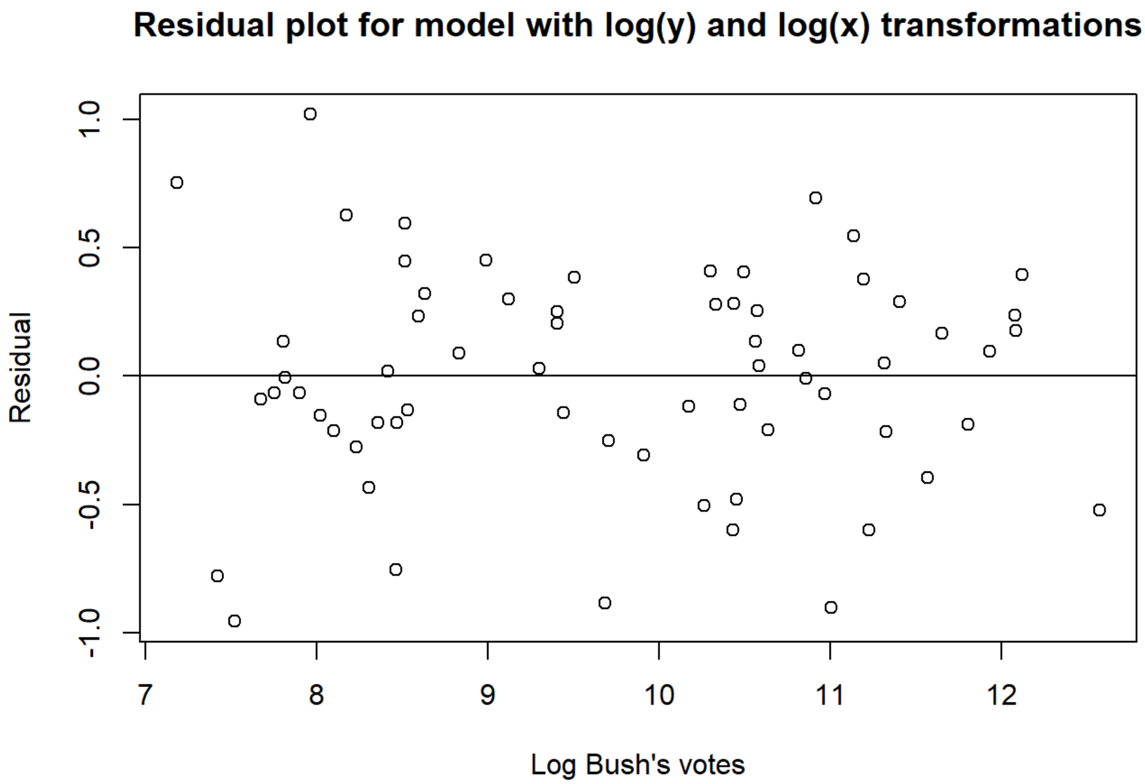
Step 5: Try transforming with log(y) and log(x)

The residual plot of log(y) and log(x) transformations satisfy both linearity and constant variance assumptions. Norm qq plot show that all residuals falls along a line, therefore normality assumption is satisfied. Finally, given how data was collected, we can also conclude that independency assumption is met.

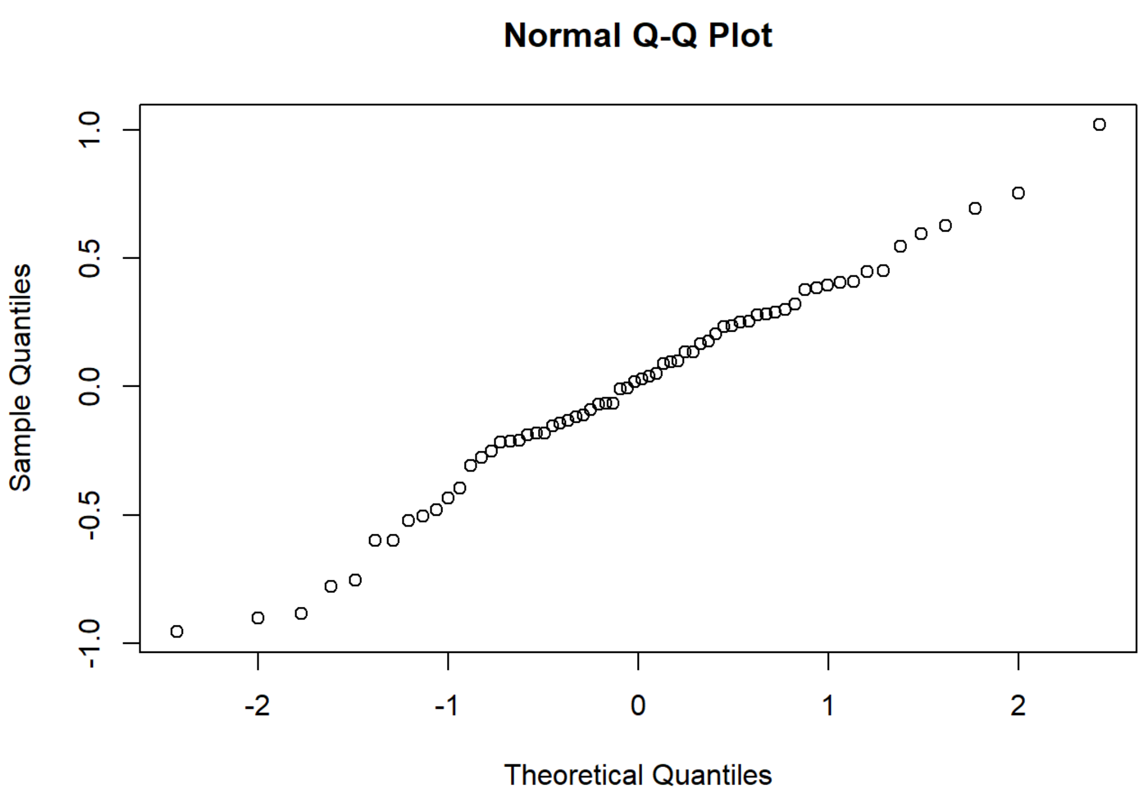
```
TestElections$logBush = log(TestElections$Bush2000)
plot(y= TestElections$logBuchanan, x=TestElections$logBush,
     xlab = "Log Bush's votes",
     ylab = "Log Buchanan's votes",
     main = "Log Buchanan's votes versus log Bush's votes"
     )
```




```
elelogBuchananBushfit = lm(logBuchanan~logBush, data = TestElections)

#residual plot
plot(y=elelogBuchananBushfit$residual, x=TestElections$logBush,
     xlab = "Log Bush's votes",
     ylab = "Residual",
     main = "Residual plot for model with log(y) and log(x) transformations")
abline(0,0)
```




```
##norm q-q plot
qqnorm(elelogBuchananBushfit$residual)
```




```
summary(elelogBuchananBushfit)
```

```
##
## Call:
## lm(formula = logBuchanan ~ logBush, data = TestElections)
##
## 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 ***
## logBush      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
```

Step 6: Make Predictions

95% prediction intervals for the Buchanan’s votes of palm county is in the range (250.80,1399.1). The actual result 3407 is way beyond the prediction interval, which means that Buchanan’s votes contain some part of Al Gore’s votes.

```
newdata3 = data.frame(logBush = log(152846))
exp(predict.lm(elelogBuchananBushfit, newdata3, interval = "predict"))
```

```
##        fit      lwr      upr
## 1 592.3769 250.8001 1399.164
```