Xuan Yu - HW1

Xuan Yu

9/3/2018

Problem 1

This is the summary for the model:

```r
oldfaithful <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW1/OldFaithful.csv")
lm_old <- lm(Interval~Duration, data = oldfaithful)
summary(lm_old)
```

```
##
## 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
```

This is the 95% confidence interval for the model:

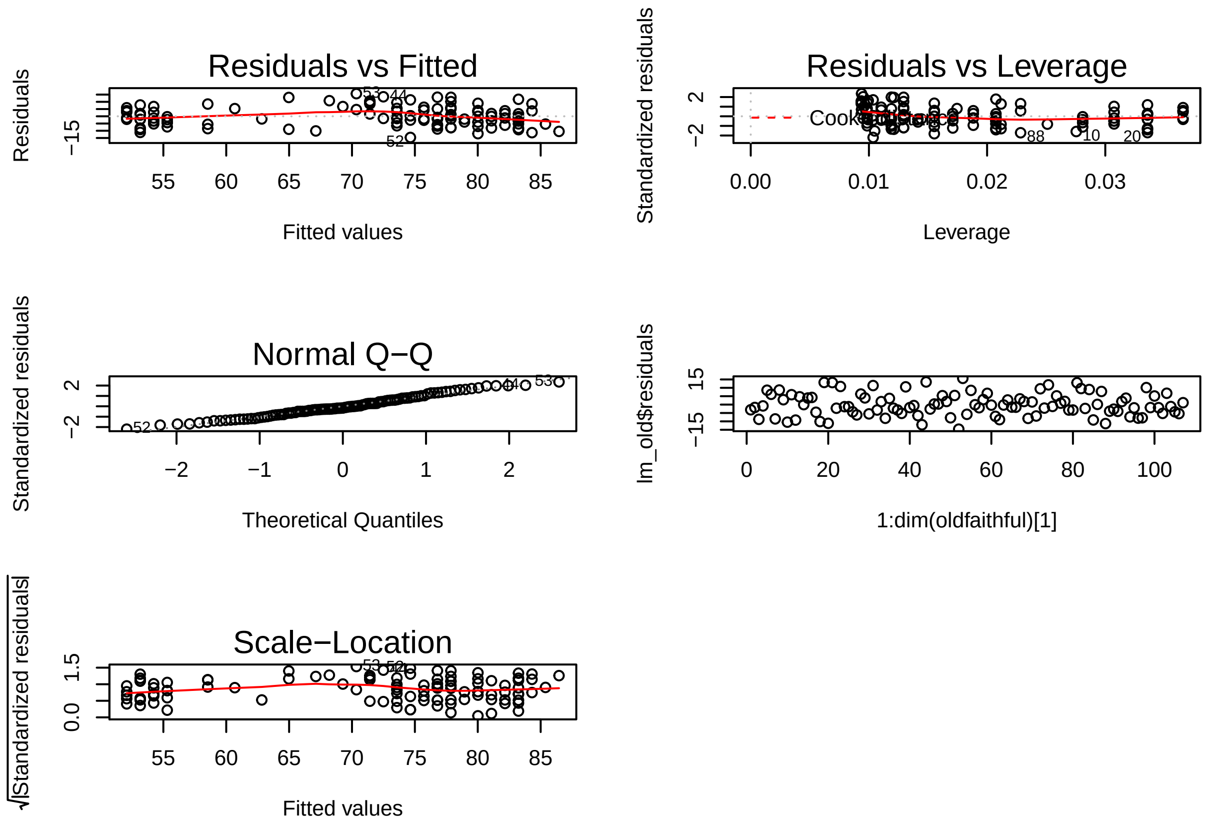
```r
confint(lm_old)
```

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

Check the assumption:

We are using the first three plots of the first commond and then the second commond to check the assumption.

```r
par(mfcol = c(3,2))
plot(lm_old)
plot(1:dim(oldfaithful)[1], lm_old$residuals) #for the independence assumption
```



Description:

All the assumptions are met.

Prediction

Here is the 95% prediction interval when the duration of the previous one is 4 minutes:

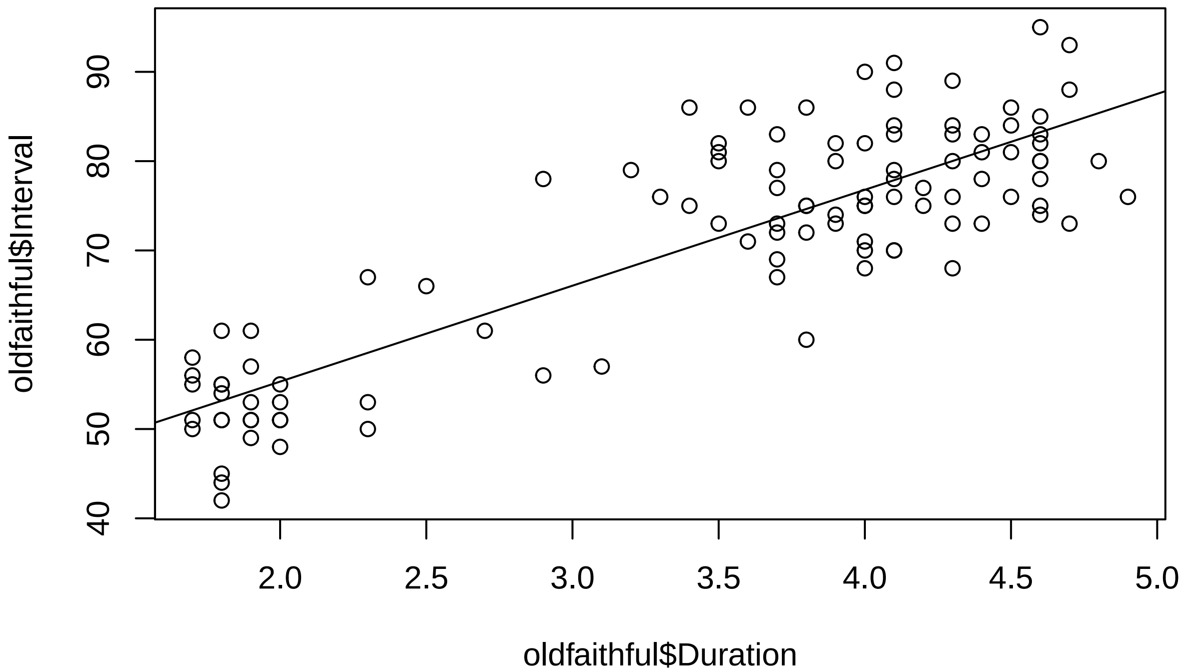
```
newdata.old <- data.frame(Duration = 4)
predict.lm(lm_old, newdata.old, interval = "prediction")
```

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

Conclusion:

When the duration of the previous eruption increases by 1 minute, the interval time until the next eruption will increase 10.741 minutes. The following plot shows this relationship.

```
plot(oldfaithful$Duration, oldfaithful$Interval)
abline(33.8282, 10.7410)
```



Problem 2

Load in the data and check the assumptions.

```
respiratory <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW1/Respiratory.csv")
```

We found that the linearity and normality assumption are not met.

Do the transformation and then the linear regression:

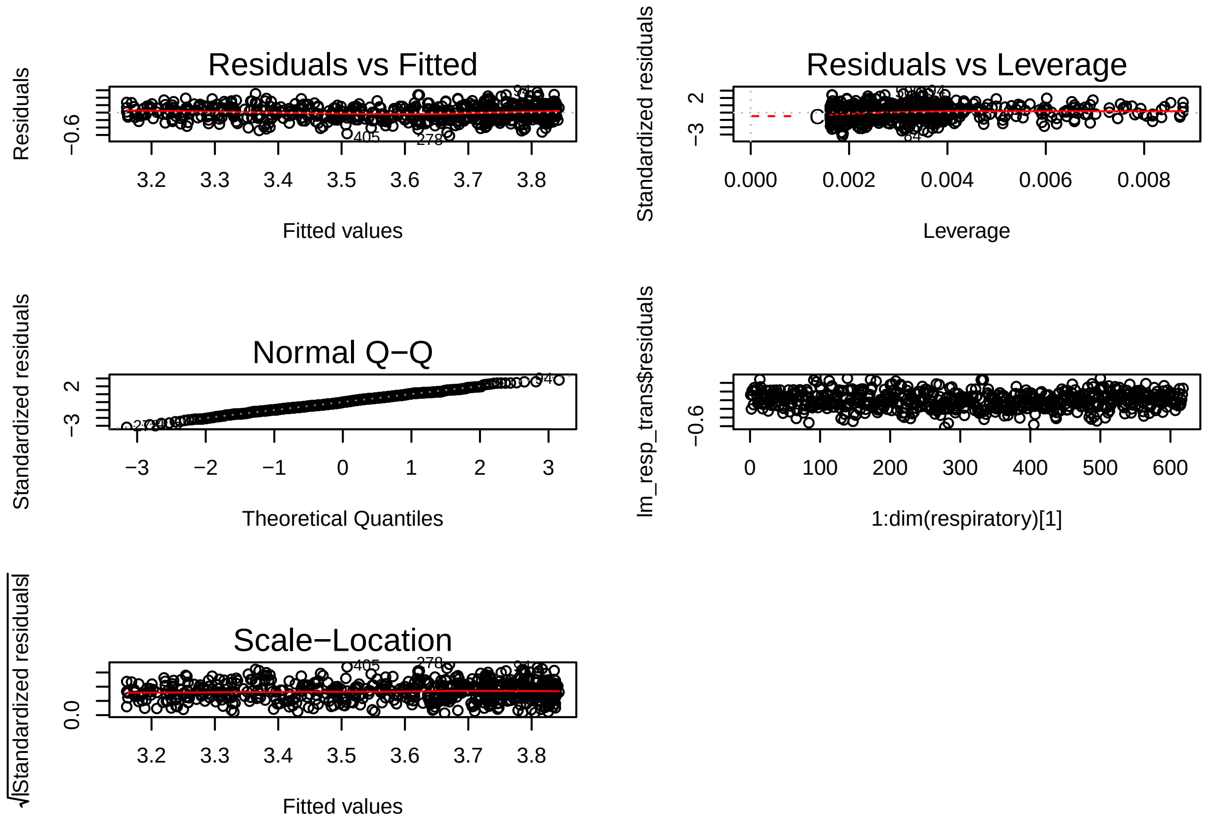
```
respiratory$LogRate <- log(respiratory$Rate)
lm_resp_trans <- lm(LogRate~Age, data = respiratory)
summary(lm_resp_trans)
```

```
##
## 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
```

Check the assumptions again:

```r
par(mfcol = c(3,2))
plot(lm_resp_trans)
plot(1:dim(respiratory)[1], lm_resp_trans$residuals) #for the independence assumption
```



Description:

All the assumptions are met.

This is the 95% confidence interval for the model:

```r
confint_log <- confint(lm_resp_trans)
exp(confint_log)
```

```
##                   2.5 %     97.5 %
## (Intercept) 45.6188072 47.9384076
## Age          0.9797541  0.9825891
```

Prediction

Here is the 95% prediction rate for three individual children: a 1 month old, an 18 months old, and a 29 months old:

```
newdata.resp <- data.frame(Age = c(1, 18, 29))
predict_confint_resp <- predict.lm(lm_resp_trans, newdata.resp, interval = "prediction")
predict_confint_resp <- exp(predict_confint_resp)
cbind(newdata.resp, predict_confint_resp)
```

```
##   Age      fit      lwr      upr
## 1   1 45.88368 31.17725 67.52721
## 2  18 33.21353 22.57614 48.86302
## 3  29 26.94664 18.30537 39.66714
```

Conclusion:

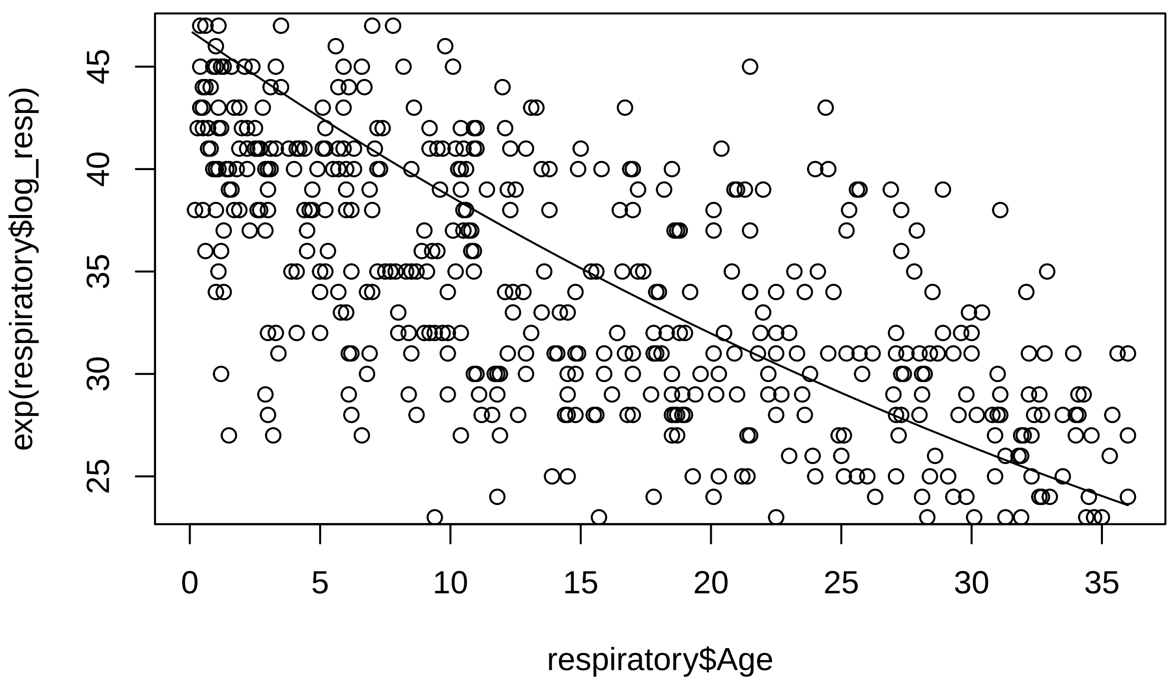
exp(-0.0190090) = 98.117%.

According to the fomula, when the age of the baby increases by 1 month, the respiratory rate will become 98.117% of itself.

In other word, when the age of the baby increases by 1 month, the respiratory rate will decrease by 1.883%.

The following plot shows this relationship between the respiratory rate and the age:

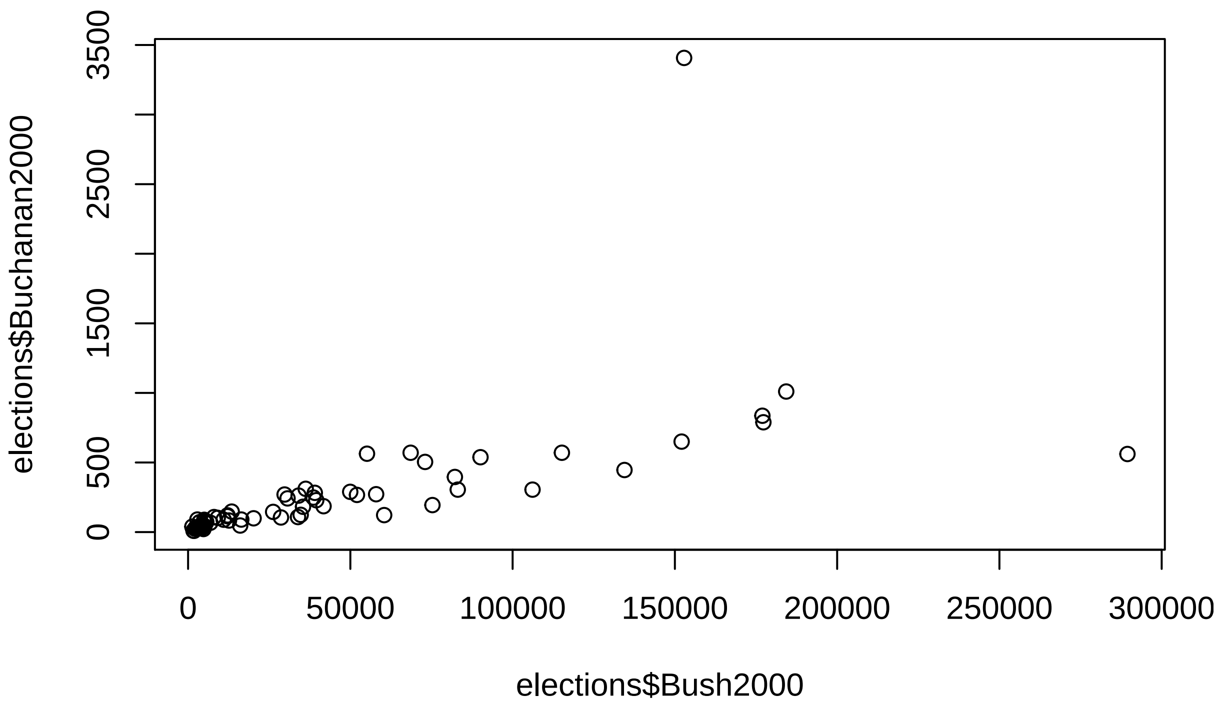
```
respiratory$log_resp <- (-0.0190090 * respiratory$Age) + 3.8451185
plot(respiratory$Age, exp(respiratory$log_resp), type = 'l')
points(respiratory$Age, respiratory$Rate)
```



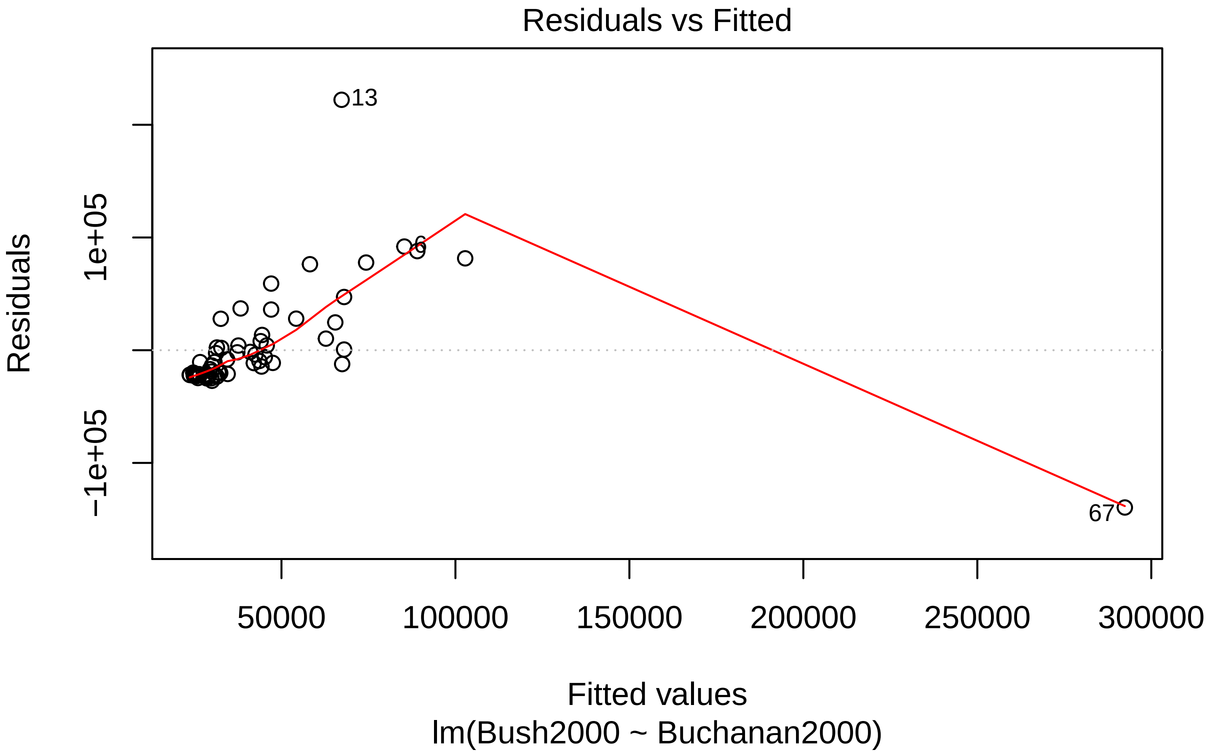
Problem 3

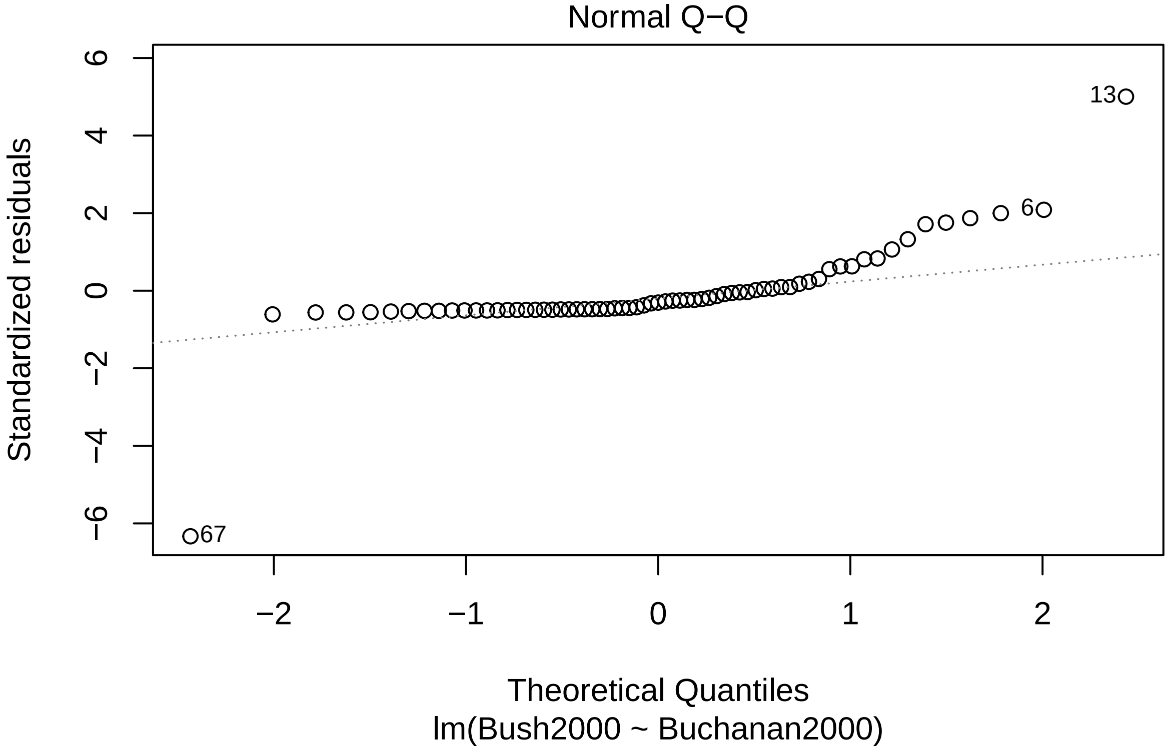
Load in the data and plot it, we see an outlier, we run the model with the outlier, and linearity assumption is not met:

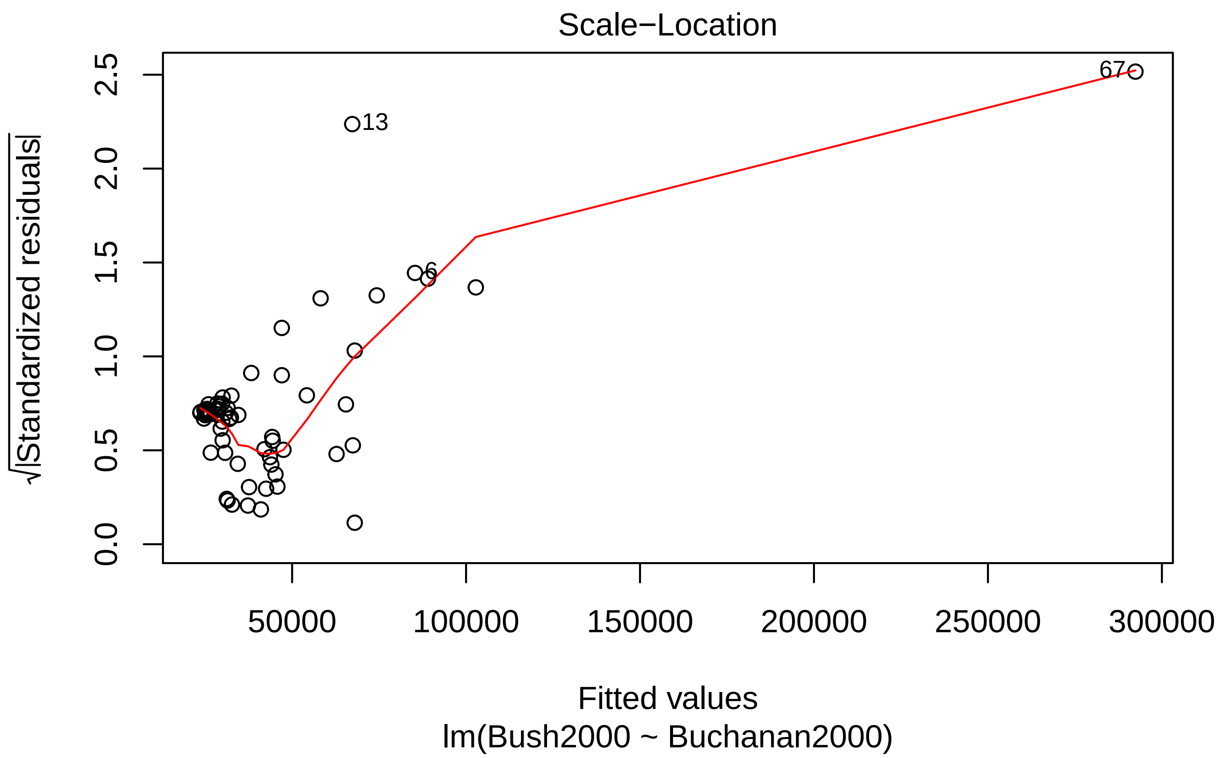
```
elections <- read.csv("/Users/xuanyu/Desktop/MIDS courses/data modeling/HW/HW1/Elections.csv")
plot(elections$Bush2000, elections$Buchanan2000)
```

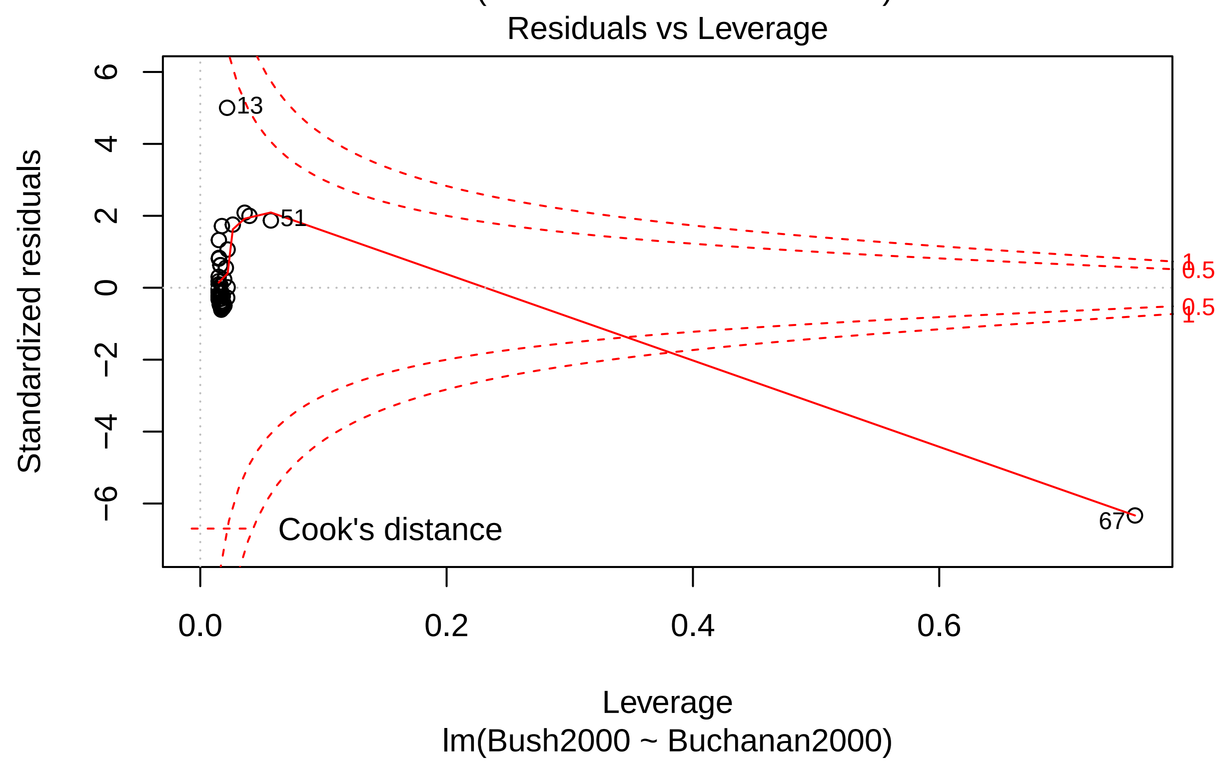



```
lm_ele <- lm(Bush2000~Buchanan2000, data = elections)
plot(lm_ele)
```


Residuals vs Fitted
13
6
67
Residuals
1e+05
−1e+05
50000
100000
150000
200000
250000
300000
Fitted values
lm(Bush2000 ~ Buchanan2000)


Normal Q−Q
13
6
67
Standardized residuals
6
4
2
0
−2
−4
−6
−2
−1
0
1
2
Theoretical Quantiles
lm(Bush2000 ~ Buchanan2000)





So we decied to remove the outlier to do the prediction:

After checking the linear regression assumptions, we found that the linearity and normality assumption are not met.

Transformation:

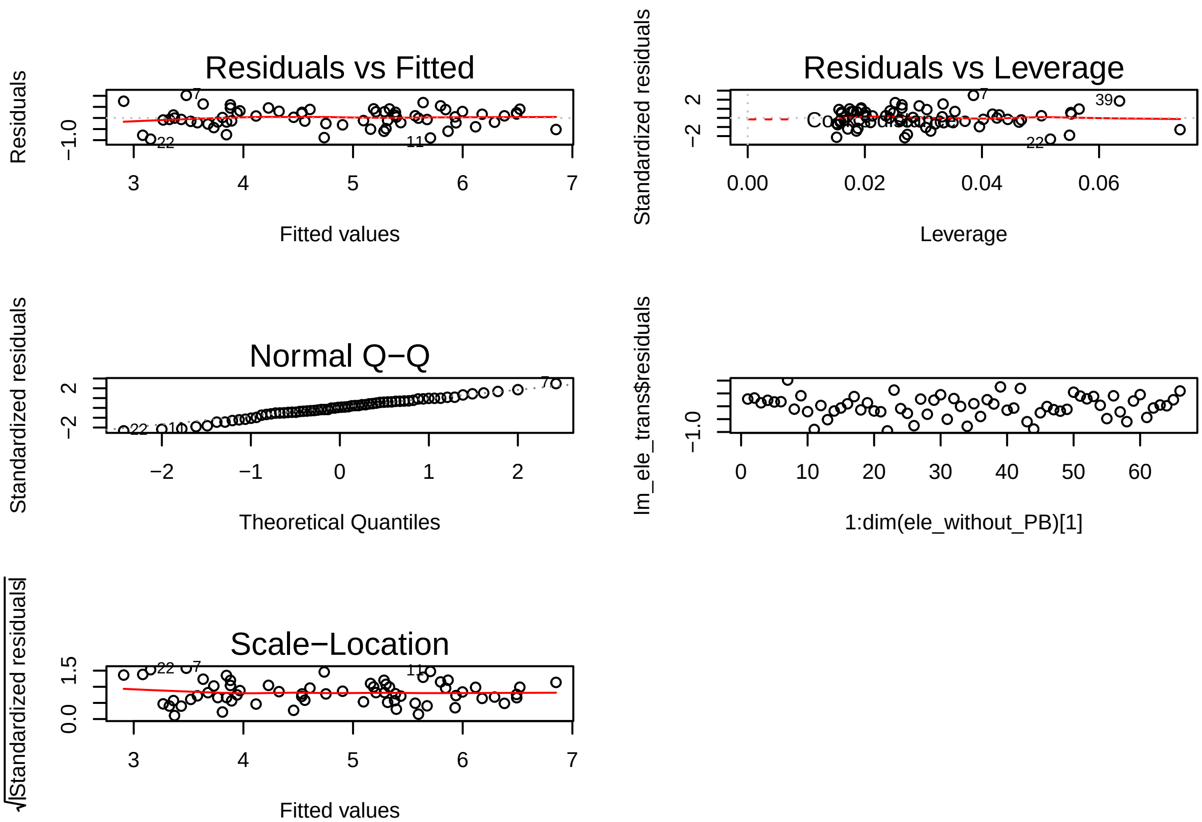
Because the linear regression doesn't met the assumptions, we need to do transformation for it.

Here is the transformation and the linear regression for it:

```r
ele_without_PB <- elections[elections$County != "Palm Beach",]
ele_without_PB$LogY <- log(ele_without_PB$Buchanan2000)
ele_without_PB$LogX <- log(ele_without_PB$Bush2000)
lm_ele_trans <- lm(LogY~LogX, data = ele_without_PB)
```

Check the assumption again:

```r
par(mfcol = c(3,2))
plot(lm_ele_trans)
plot(1:dim(ele_without_PB)[1], lm_ele_trans$residuals) #for the independence assumption
```



Description:

All assumtions are met.

This is the summary and confidence interval for the transformed model:

```
summary(lm_ele_trans)
```

```
##
## Call:
## lm(formula = LogY ~ LogX, data = ele_without_PB)
##
## 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 ***
## LogX         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
```

```
confint_log <- confint(lm_ele_trans)
exp(confint_log)
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.04738205 0.1952527
## LogX        1.93306916 2.2318146
```

The prediction for Buchanan's votes in Palm Beach County:

```
newdata.ele <- data.frame(LogX = log(152846))
predict_confint_ele <- predict.lm(lm_ele_trans, newdata.ele, interval = "prediction")
predict_confint_ele <- exp(predict_confint_ele)
predict_confint_ele
```

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

Votes intended for Gore:

```
c(3407 - predict_confint_ele[3], 3407 - predict_confint_ele[2])
```

```
## [1] 2007.836 3156.200
```

Conclusion:

So Buchanan's votes in Palm Beach county should be in the range of (250.8001, 1399.164), but he got 3407 votes. So there were (2007.836, 3156.200) votes which are intended for Gore.