Modelling_Exercises_Week1

Guillem_Amat

September 2, 2019

Old Faithful

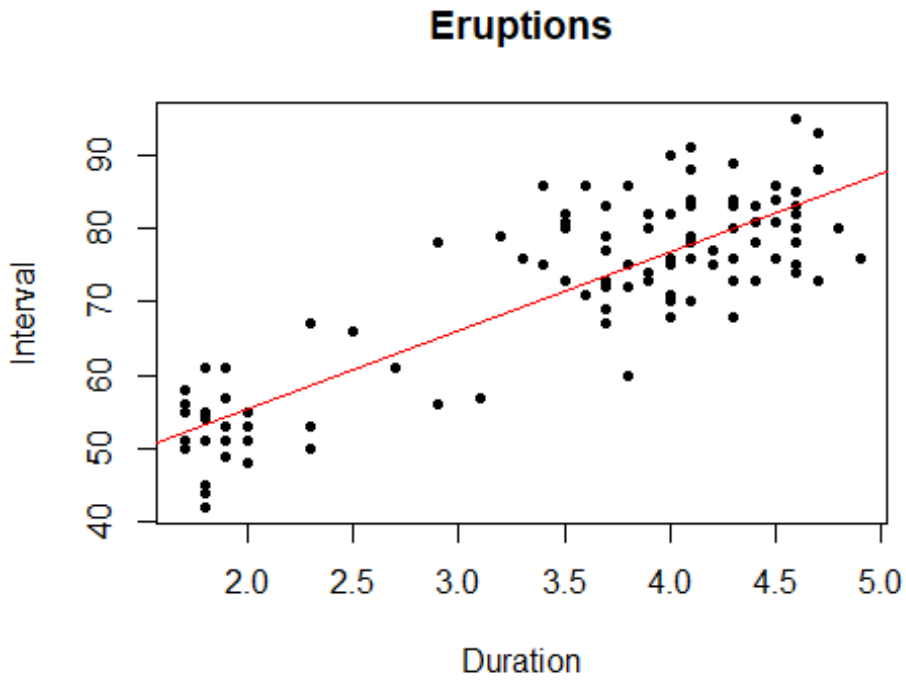
1. Regression Model

```
#We first import the data from the CSV file
Eruption_Data <- read.csv(file = "C:/Users/Guillem/Desktop/Duke
University/11_Modelling and Representation of Data/0_Homework
Exercises/0_Data/Homework 1/OldFaithful.csv", header = TRUE, sep = ",")


#We create a linear model to predict Interval from Duration
Eruption_Model <- lm(Interval ~ Duration, data = Eruption_Data)
```

2. Model Interpretation

```
#Inspecting the results
summary(Eruption_Model)

##
## Call:
## lm(formula = Interval ~ Duration, data = Eruption_Data)
##
## 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

#Fitting the model to the data
plot(Eruption_Data$Duration, Eruption_Data$Interval, xlab = "Duration",
```

```
ylab = "Interval", main = "Eruptions", pch = 20)
abline(Eruption_Model, col = "red")
```



- Intercept: 33.8282. It would be the value of the Interval between eruptions if there was an Eruption of Duration 0.
- Slope: 10.7410. For each unit of Duration (minutes), Interval increases by the slope value.
- Residual Error: 6.683. It measures the standard deviation of the value of the errors and how close the fit is to the points.
- R-Squared: 0.7369. It measures the proportion of the variance in the model that is explained by the data. This is a pretty high value.

3. Confidence Interval

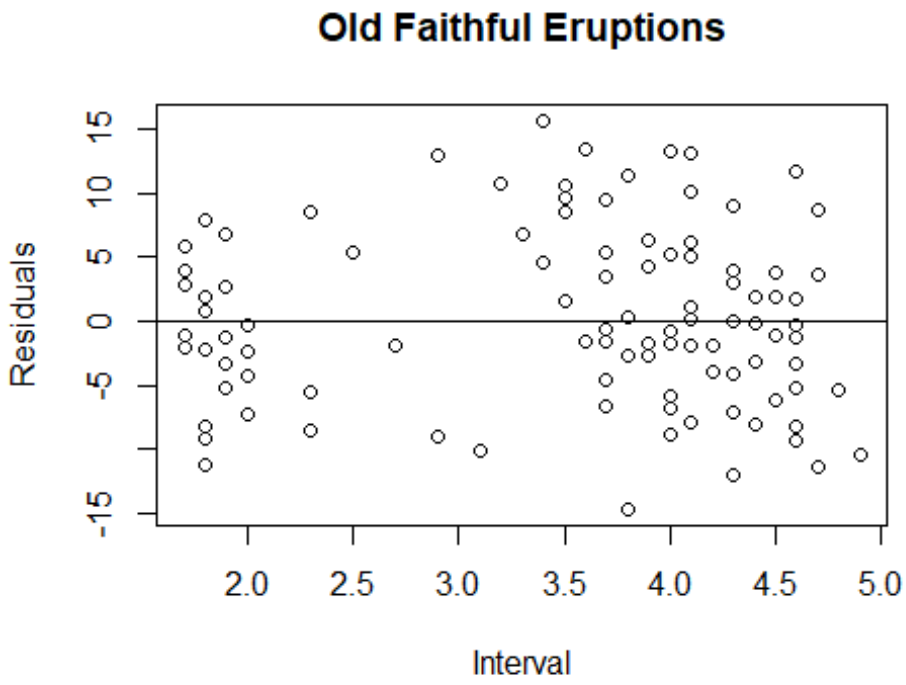
```
confint(Eruption_Model, level = 0.95)
```

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

The 95% confidence interval for the slope is between 9.4 and 11.98. We are confident that Duration and Interval are positively correlated; the longer the duration of an eruption the more tourists will need to wait for the next one.

4. Residuals

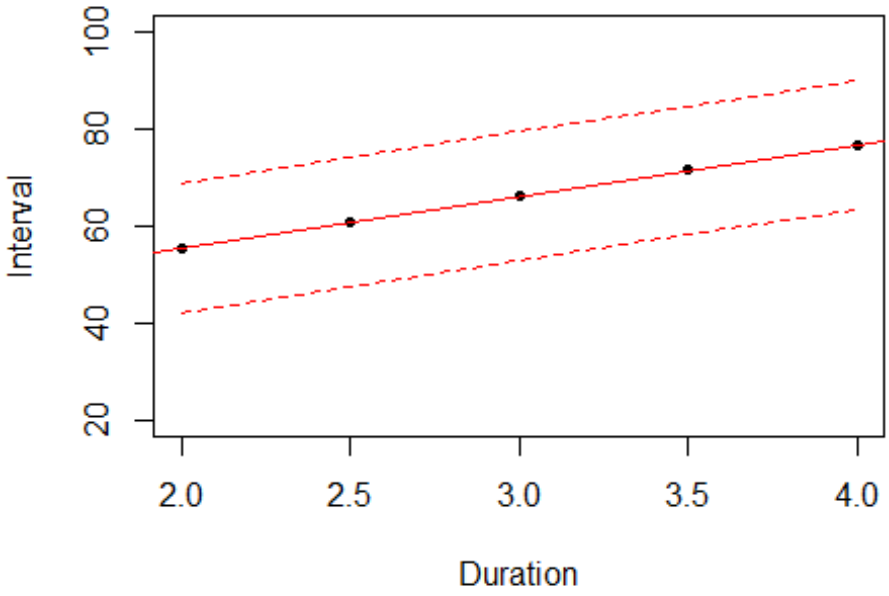
```
Eruption_Residual = resid(Eruption_Model)
plot(Eruption_Data$Duration, Eruption_Residual, ylab="Residuals",
xlab="Interval", main="Old Faithful Eruptions")
abline(0, 0)
```



Although the results seem a bit odd, we can not see a clear discernible pattern among the residuals. Based on the residual plot above, it seems the model fulfills the regression assumption.

5. Confidence Intervals for specific Durations

```
Durations <- data.frame(Duration = c(2, 2.5, 3, 3.5, 4))
Durations_Prediction <- data.frame(predict(Eruption_Model, Durations,
interval = "prediction"))

#We plot the three points in the line
plot(Durations_Prediction$fit ~ Durations$Duration, pch = 20, xlab =
"Duration", ylab = "Interval", xlim = c(2,4), ylim =c(20, 100))


abline(Eruption_Model, col = "red")
lines(Durations$Duration, Durations_Prediction$lwr, col = "red", lty = 2)
lines(Durations$Duration, Durations_Prediction$upr, col = "red", lty = 2)
```

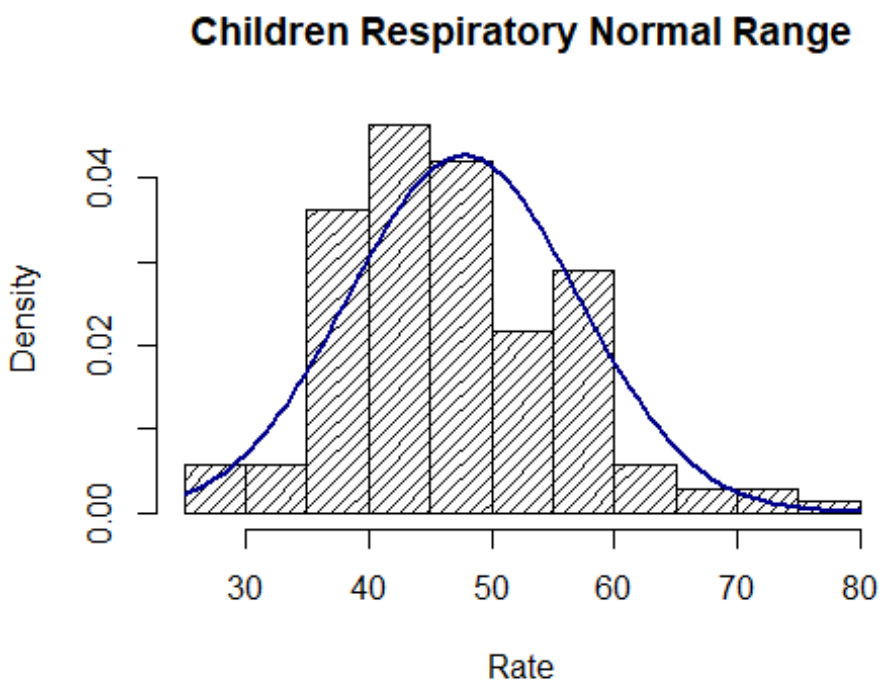


Respiratory Rates for Children

```
#We first import the data from the CSV file
Respiratory_Data <- read.csv(file = "C:/Users/Guillem/Desktop/Duke
University/11_Modelling and Representation of Data/0_Homework
Exercises/0_Data/Homework 1/Respiratory.csv", header = TRUE, sep = ",")
```

1. Normal Range for Children

```
Respiratory_Kids_Data <- Respiratory_Data[which(Respiratory_Data$Age <=
3),]
m <- mean(Respiratory_Kids_Data$Rate)
s <- sqrt(var(Respiratory_Kids_Data$Rate))
x <- Respiratory_Kids_Data$Rate

hist(Respiratory_Kids_Data$Rate, density = 20, freq = FALSE, main =
"Children Respiratory Normal Range", xlab = "Rate")
curve(dnorm(x, mean = m, sd = s), add = TRUE, col = "darkblue", lwd = 2)
```

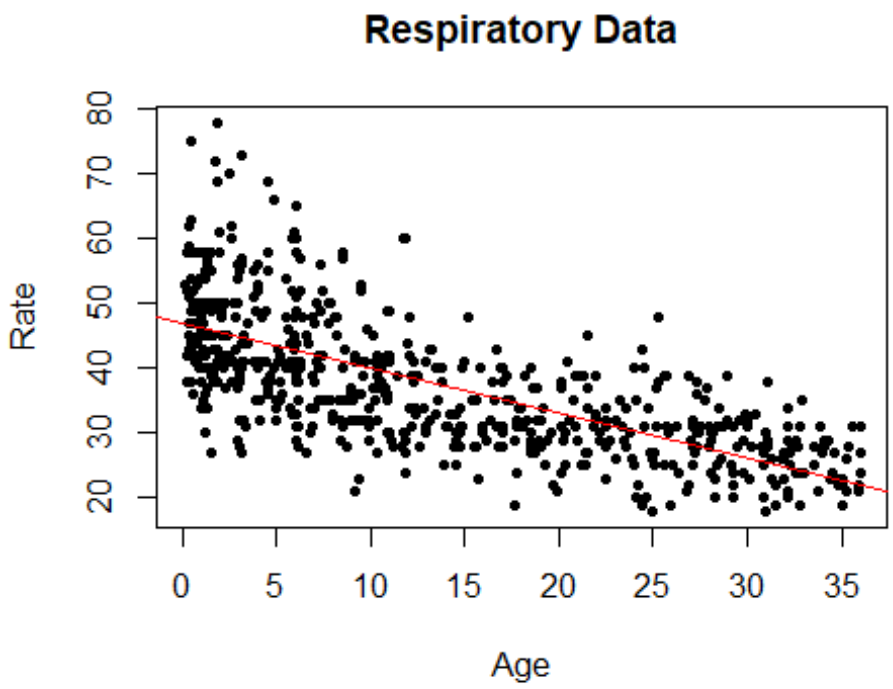


2. Model and Linear Transformations

As a first step we fit the model to the data, inspect summary statistics, plot the data as a scatterplot and fit the model to it.

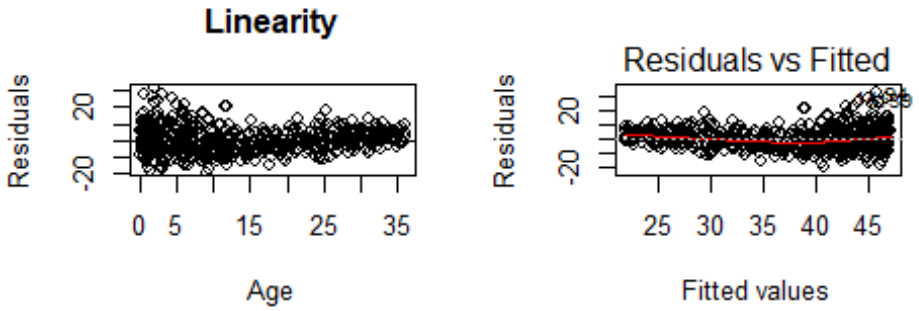
```
#We fit the data to a model and we check summary statistics
Respiratory_Model <- lm(Rate ~ Age, data = Respiratory_Data)
summary(Respiratory_Model)

##
## Call:
## lm(formula = Rate ~ Age, data = Respiratory_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.652  -5.432  -0.608   4.589  32.270
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.05216    0.50422   93.32   <2e-16 ***
## Age         -0.69571    0.02938  -23.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.842 on 616 degrees of freedom
## Multiple R-squared:  0.4766, Adjusted R-squared:  0.4758
## F-statistic: 560.9 on 1 and 616 DF,  p-value: < 2.2e-16
```

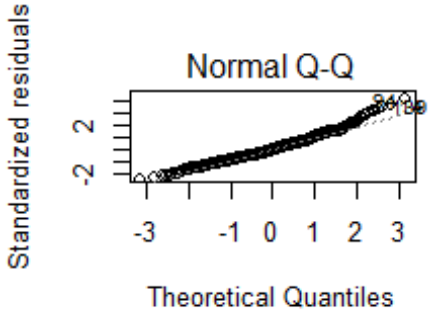
```r
#We plot the values as a scatter plot and fit the model in the plot
plot(Respiratory_Data$Age, Respiratory_Data$Rate, xlab = "Age", ylab =
"Rate", main= "Respiratory Data", pch = 20)
abline(Respiratory_Model, col="red")
```



We compute and display the different model diagnostics to check for Normality, Linearity, Independence and Equal Variance.

```r
par(mfrow = c(2,2))

#Checking Linearity
plot(Respiratory_Model$residuals, x = Respiratory_Data$Age, xlab = "Age",
ylab = "Residuals", main = "Linearity")
abline(0, 0)

#Checking Independence and Equal Variance
plot(Respiratory_Model, which = 1)

#Checking Normality
plot(Respiratory_Model, which = 2)
```

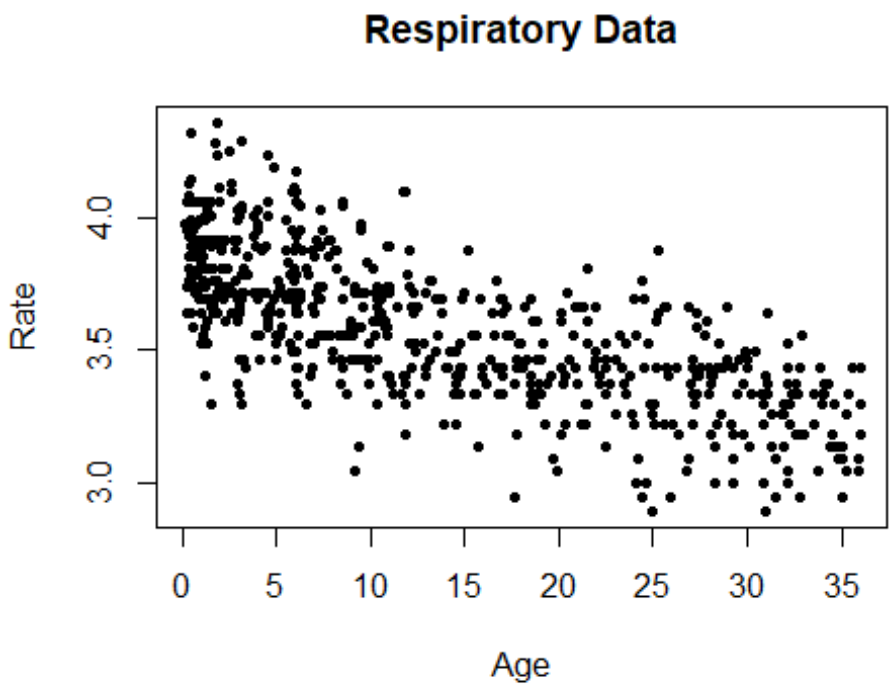




It seems the model assumptions might not be reasonalble for this dataset:

- In the Linearity plot, we should observe no discernible pattern of any type. Instead we can clearly see the data forming a funnel.
- In the Residuals vs Fitted plot, which we use to check for Independence and Equal Variance, we can observe the data forming a curve.
- The Normality does not seem to hold as a considerable number of residuals in the qqplot along the edges stray away from the line.

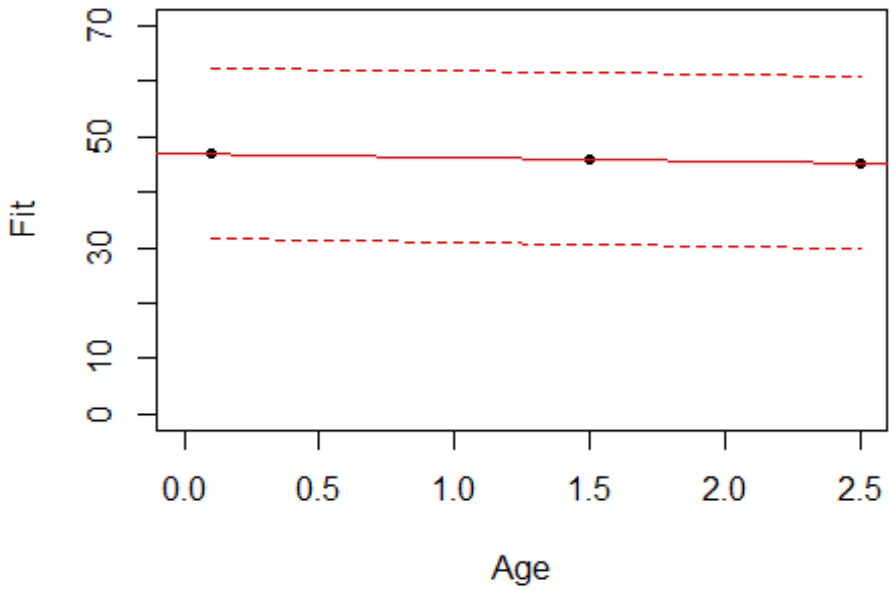
```
plot(Respiratory_Data$Age, log(Respiratory_Data$Rate), xlab = "Age", ylab
= "Rate", main= "Respiratory Data", pch = 20)
abline(Respiratory_Model, col="red")
```



If we apply a logarithmic transformation to the y axis the data seems to have a stronger linear pattern.

3. Confidence Intervals

```
Kids_Age <- data.frame(Age = c(0.1, 1.5, 2.5))
Kids_Prediction <- data.frame(predict(Respiratory_Model, Kids_Age,
interval = "prediction"))

#We plot the three points in the line
plot(Kids_Prediction$fit ~ Kids_Age$Age, pch = 20, xlim = c(0,2.5), ylim
=c(0, 70), xlab = "Age", ylab = "Fit")


abline(Respiratory_Model, col = "red")
lines(Kids_Age$Age, Kids_Prediction$lwr, col = "red", lty = 2)
lines(Kids_Age$Age, Kids_Prediction$upr, col = "red", lty = 2)
```



Elections

1. Scatterplot of Bush vs Buchanan
As always, we start by importing the data.

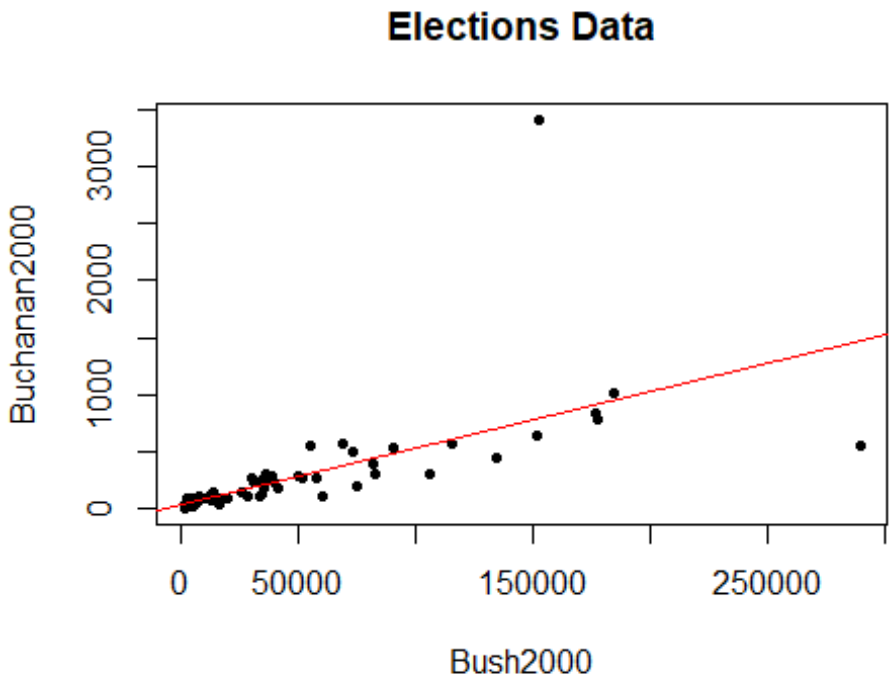
```
#Importing the data
Elections_Data <- read.csv(file = "C:/Users/Guillem/Desktop/Duke
University/11_Modelling and Representation of Data/0_Homework
Exercises/0_Data/Homework 1/Elections.csv", header = TRUE, sep = ",")
```

We create a linear model by predicting Buchanan votes from Bush votes and plot the values as a scatter plot while fitting the model to the data.

```
Elections_Model <- lm(Buchanan2000 ~ Bush2000, data = Elections_Data)
summary(Elections_Model)

##
## Call:
## lm(formula = Buchanan2000 ~ Bush2000, data = Elections_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -907.50  -46.10  -29.19   12.26 2610.19
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 4.529e+01  5.448e+01   0.831    0.409
## Bush2000    4.917e-03  7.644e-04   6.432 1.73e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 353.9 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
```

```
plot(Elections_Data$Bush2000, Elections_Data$Buchanan2000, xlab =
"Bush2000", ylab = "Buchanan2000", main= "Elections Data", pch = 20)
abline(Elections_Model, col="red")
```



From what we can see in the scatterplot, there is a strong linear relationship between the number of votes Buchanan got per county and the number of votes Bush got. There is a clear outlier, which is Palm Beach county that does not yield the expected result.

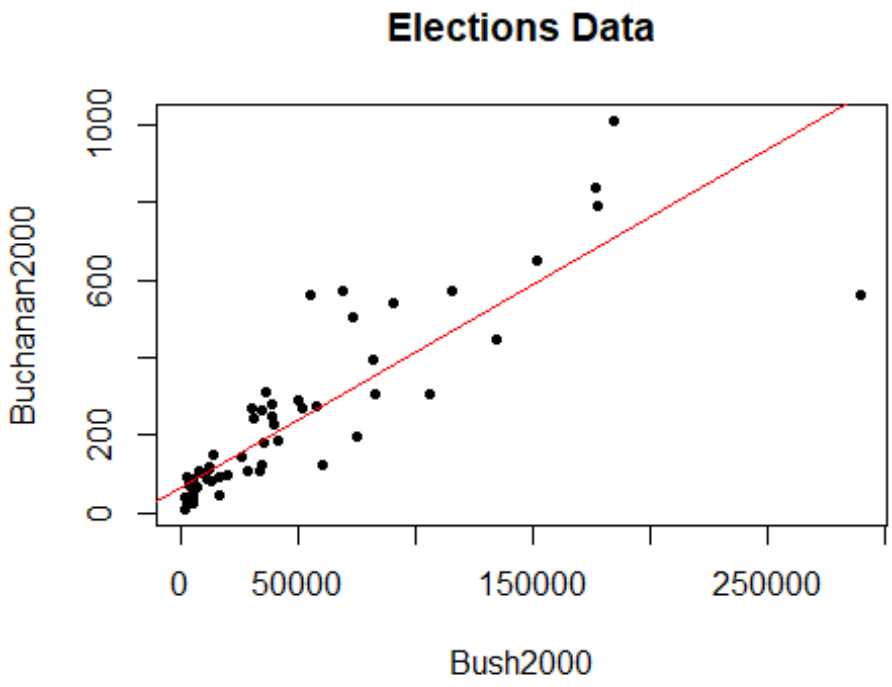
2. Scatterplot and Model of Bush vs Buchanan without Palm Beach County

```
Elections_Data_WPM <- Elections_Data[which(Elections_Data$County != "Palm
Beach"), ]

Elections_Model_WPM <- lm(Buchanan2000 ~ Bush2000, data =
Elections_Data_WPM)
```

```r
summary(Elections_Model_WPM)
```

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

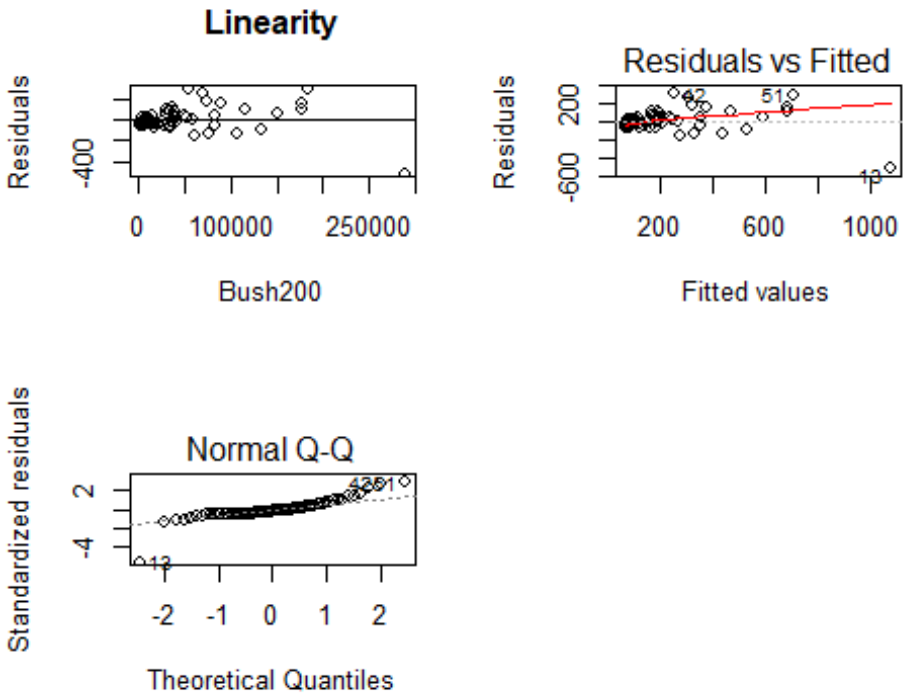
```r
plot(Elections_Data_WPM$Bush2000, Elections_Data_WPM$Buchanan2000, xlab =
"Bush2000", ylab = "Buchanan2000", main= "Elections Data", pch = 20)
abline(Elections_Model_WPM, col="red")
```



By removing Palm Beach County, R-Squared increases dramatically from ~40% to ~80%. This means that the variance that the model explains increases by a factor of two.

3. Model Assumptions and Output

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

#Checking Linearity
plot(Elections_Model_WPM$residuals, x = Elections_Data_WPM$Bush2000, xlab
= "Bush200", ylab = "Residuals", main = "Linearity")
abline(0, 0)

#Checking Independence and Equal Variance
plot(Elections_Model_WPM, which = 1)

#Checking Normality
plot(Elections_Model_WPM, which = 2)
```



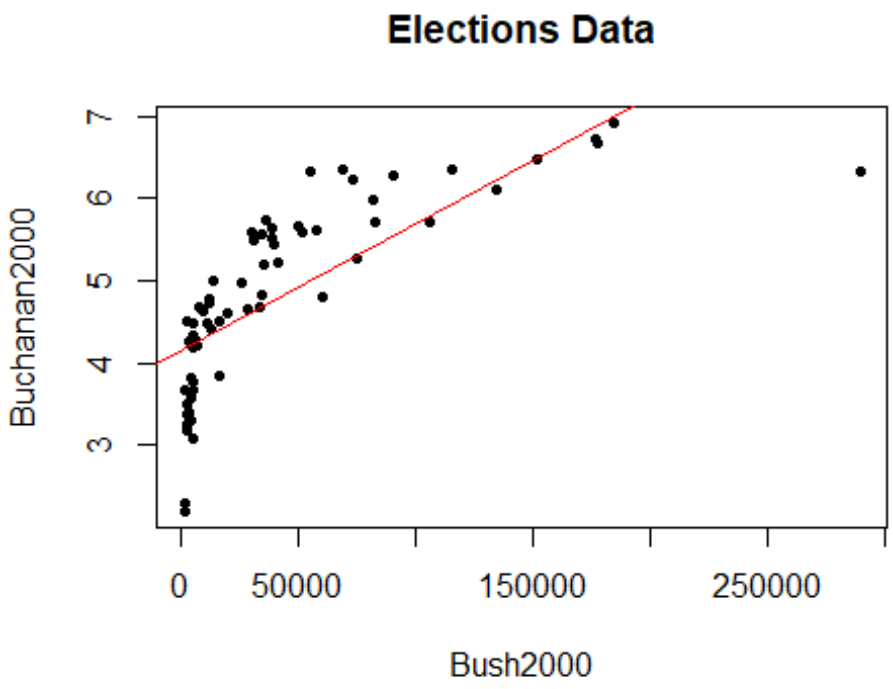
It seems the model assumptions might not be reasonalble for this dataset:

- In the Linearity plot, we should observe no discernible pattern of any type. Instead we can clearly see the data values diverging as the variable in the x-axis increase.

- In the Residuals vs Fitted plot, which we use to check for Independence and Equal Variance, we can observe how the fitted values that are higher also have higher residuals.
- The Normality does not seem to hold as a considerable number of residuals in the qqplot along the right edge stray away from the line.

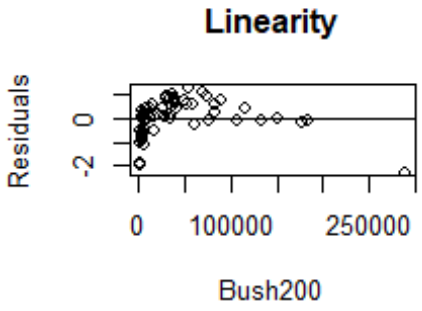
We proceed to try to apply logarithmic transformations to see if fit improves and assumptions become reasonable.

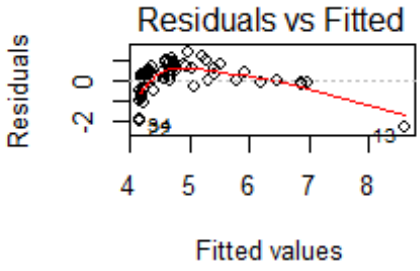
```
Elections_Model_log_WPM <- lm(log(Buchanan2000) ~ Bush2000, data =
Elections_Data_WPM)

summary(Elections_Model_log_WPM)

##
## Call:
## lm(formula = log(Buchanan2000) ~ Bush2000, data = Elections_Data_WPM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27274 -0.48845  0.07695  0.50187  1.34188
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.142e+00  1.157e-01  35.808  < 2e-16 ***
## Bush2000    1.541e-05  1.669e-06   9.233 2.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7505 on 64 degrees of freedom
## Multiple R-squared:  0.5712, Adjusted R-squared:  0.5645
## F-statistic: 85.25 on 1 and 64 DF,  p-value: 2.222e-13

plot(Elections_Data_WPM$Bush2000, log(Elections_Data_WPM$Buchanan2000),
xlab = "Bush2000", ylab = "Buchanan2000", main= "Elections Data", pch =
20)
abline(Elections_Model_log_WPM, col="red")
```

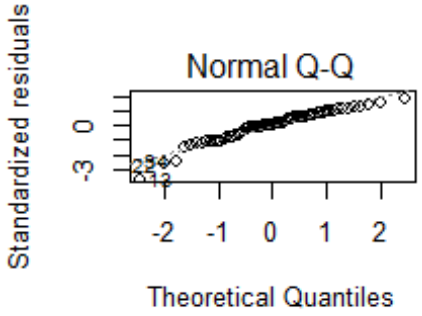


We then check for the different model assumptions.

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

#Checking Linearity
plot(Elections_Model_log_WPM$residuals, x = Elections_Data_WPM$Bush2000,
xlab = "Bush200", ylab = "Residuals", main = "Linearity")
abline(0, 0)

#Checking Independence and Equal Variance
plot(Elections_Model_log_WPM, which = 1)

#Checking Normality
plot(Elections_Model_log_WPM, which = 2)
```







It does not seem that logarithmic transformations help improve the model assumptions. Neither Linearity, Variance or Normality seem reasonable after applying a log transformation and R-Squared dimnishes.

4. Confidence Intervals

```
Palm_Beach_County <- data.frame(Bush2000 =
c(Elections_Data[which(Elections_Data$County == "Palm Beach"),
]$Bush200))
predict(Elections_Model_WPM, Palm_Beach_County, interval = "prediction")

##        fit     lwr      upr
## 1 597.7677 364.709 830.8264
```