Regression Project

```r
library(ggplot2)
library(readr)
library(car)
```

```
## Loading required package: carData
```

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(Matrix)

columns <- c("Sex","Length", "Diameter", "Height", "Whole_wt", "Shuck_wt", "Visc_wt", "Shell_wt", "Ring

abalone <- read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data",col
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Sex = col_character(),
##   Length = col_double(),
##   Diameter = col_double(),
##   Height = col_double(),
##   Whole_wt = col_double(),
##   Shuck_wt = col_double(),
##   Visc_wt = col_double(),
##   Shell_wt = col_double(),
##   Rings = col_double()
## )
```

```
abalone$Sex <- as.factor(abalone$Sex)
abalone
```

```
## # A tibble: 4,177 x 9
##    Sex   Length Diameter Height Whole_wt Shuck_wt Visc_wt Shell_wt Rings
##    <fct>  <dbl>    <dbl>  <dbl>    <dbl>    <dbl>   <dbl>    <dbl> <dbl>
##  1 M      0.455    0.365  0.095    0.514   0.224   0.101     0.15     15
##  2 M      0.35     0.265  0.09     0.226   0.0995  0.0485    0.07      7
##  3 F      0.53     0.42   0.135    0.677   0.256   0.142     0.21      9
##  4 M      0.44     0.365  0.125    0.516   0.216   0.114     0.155    10
##  5 I      0.33     0.255  0.08     0.205   0.0895  0.0395    0.055     7
##  6 I      0.425    0.3    0.095    0.352   0.141   0.0775    0.12      8
##  7 F      0.53     0.415  0.15     0.778   0.237   0.142     0.33     20
##  8 F      0.545    0.425  0.125    0.768   0.294   0.150     0.26     16
##  9 M      0.475    0.37   0.125    0.509   0.216   0.112     0.165     9
## 10 F      0.55     0.44   0.15     0.894   0.314   0.151     0.32     19
## # ... with 4,167 more rows
```

```
set.seed(42)
#Splitting dataset in train and test using 70/30 method
indexes <- sample(1:nrow(abalone), size = 0.3 * nrow(abalone))
abalone_train <- abalone[-indexes,]
abalone_test <- abalone[indexes,]
```

```
#Q1
rankMatrix(abalone[,2:8])[1]
```

```
## [1] 7
```

$$Age = \beta_0 + \beta_1 Height + \epsilon$$

with P1-P4 and full rank assumption

with the rank assumption and under P1-P4 which are: P1: Errors are centered P2: The model is homoscedastic. Variance of all the error terms are same. P3: Errors are uncorrelated. P4: Errors are gaussian. We also assume that no high leverage outliers are present. In order to study those hypothesis, we'll be visualing the regression line and then the residuals graphically to observe if they satisfy our assumptions. We'll also build the following tests to further investigate the postulates 2,3 and 4: Breush-Pagan test for P2, Durbin-Watson test for P3, Shapiro-Wilks test for P4 We'll check if our full rank assumption is met. We will also be computing the Cook distances to detect if there are outliers that change too much our estimations for beta's. Finally, we'll build confidence intervals for the betas to study the efficiency of the model.

```r
#Q2

summary(abalone)
```

```
##  Sex          Length         Diameter          Height          Whole_wt     
##  F:1307   Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020  
##  I:1342   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415  
##  M:1528   Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995  
##           Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287  
##           3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530  
##           Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255  
##     Shuck_wt          Visc_wt          Shell_wt          Rings       
##  Min.   :0.0010   Min.   :0.0005   Min.   :0.0015   Min.   : 1.000  
##  1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.: 8.000  
##  Median :0.3360   Median :0.1710   Median :0.2340   Median : 9.000  
##  Mean   :0.3594   Mean   :0.1806   Mean   :0.2388   Mean   : 9.934  
##  3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:11.000  
##  Max.   :1.4880   Max.   :0.7600   Max.   :1.0050   Max.   :29.000  
```

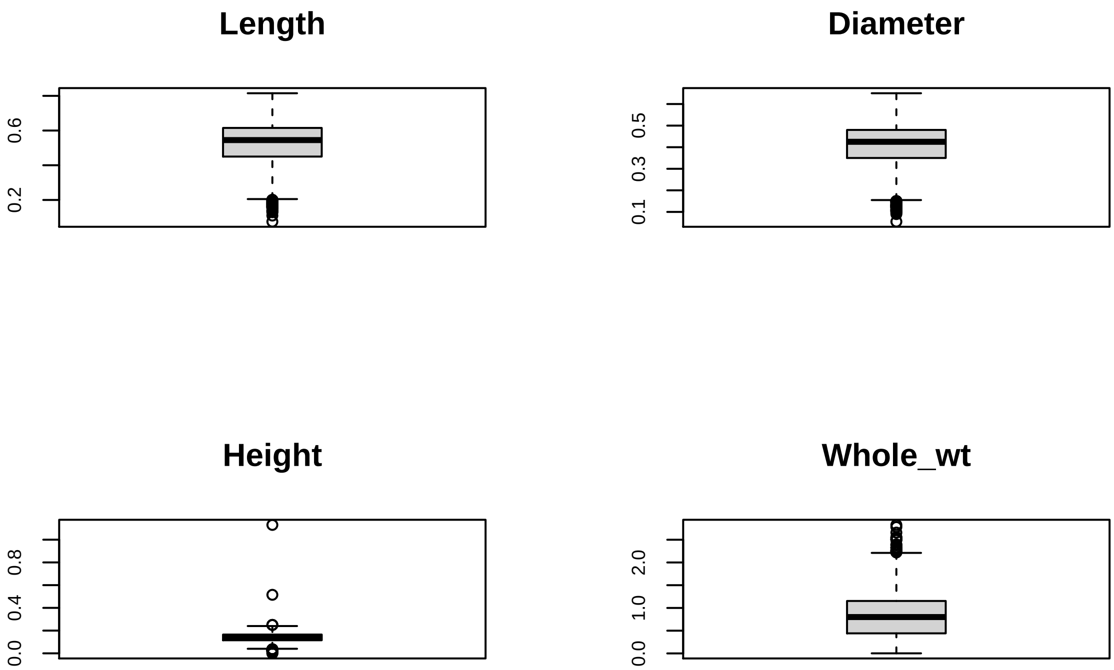
```r
diag(var(abalone))
```

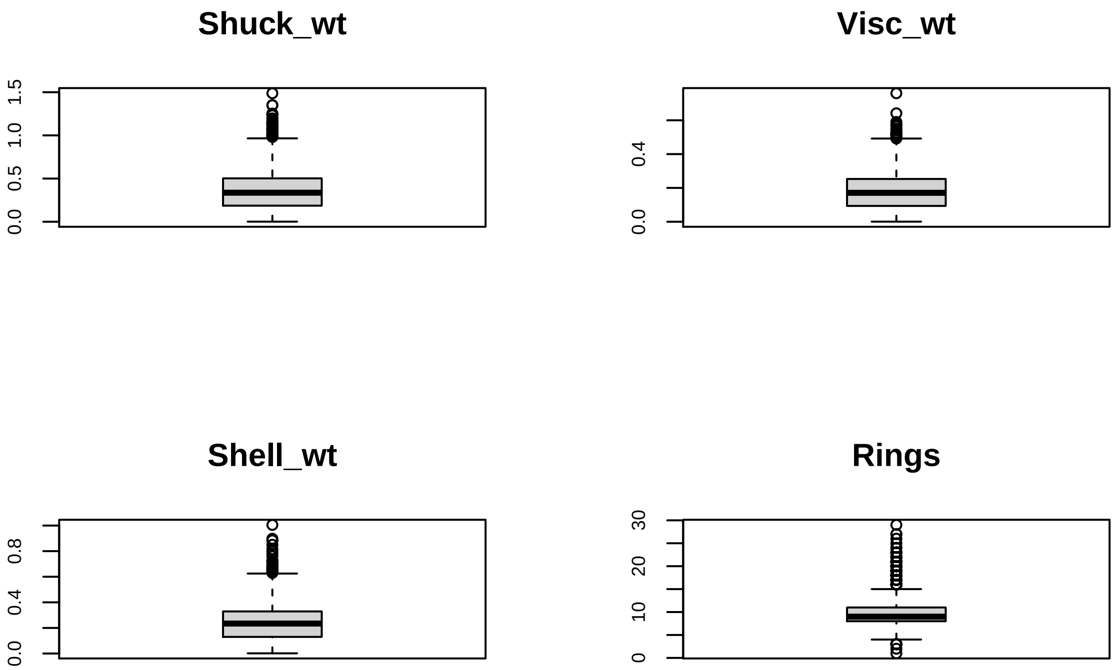
```
## Warning in var(abalone): NAs introduced by coercion

##          Sex       Length     Diameter       Height     Whole_wt     Shuck_wt 
##           NA  0.014422308  0.009848551  0.001749503  0.240481389  0.049267551 
##      Visc_wt     Shell_wt        Rings 
##  0.012015284  0.019377383 10.395265947 
```

```r
sqrt(diag(var(abalone)))
```

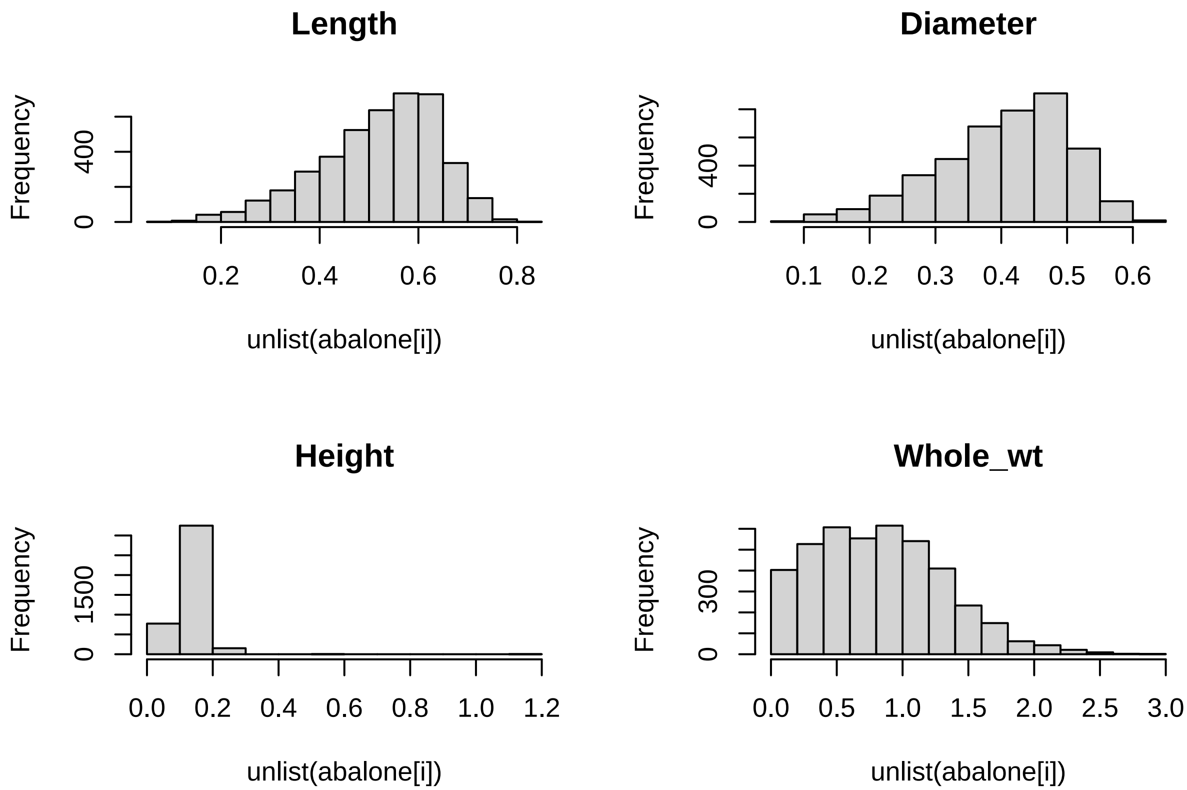
```
## Warning in var(abalone): NAs introduced by coercion

##        Sex     Length   Diameter     Height   Whole_wt   Shuck_wt    Visc_wt 
##         NA 0.12009291 0.09923987 0.04182706 0.49038902 0.22196295 0.10961425 
##   Shell_wt      Rings 
## 0.13920267 3.22416903 
```

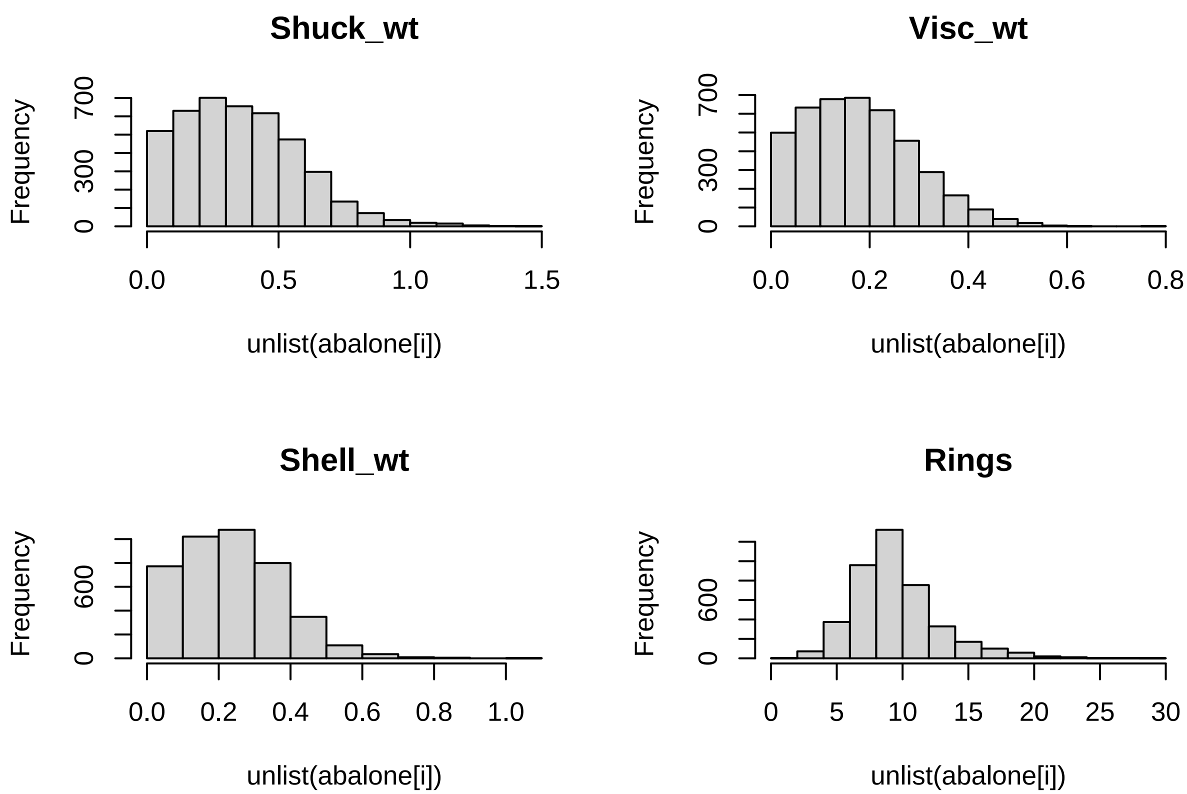
```r
par(mfrow=c(2,2))
for (i in 2:ncol(abalone)){
  boxplot(abalone[i], boxwex=0.5, cex.axis=0.75, main=colnames(abalone[i]))
}
```


Length
0.2
0.6
Diameter
0.1
0.3
0.5
Height
0.0
0.4
0.8
Whole_wt
0.0
1.0
2.0




```
par(mfrow=c(2,2))
for (i in 2:ncol(abalone)){
  hist(unlist(abalone[i]), main=colnames(abalone[i]))
}
```

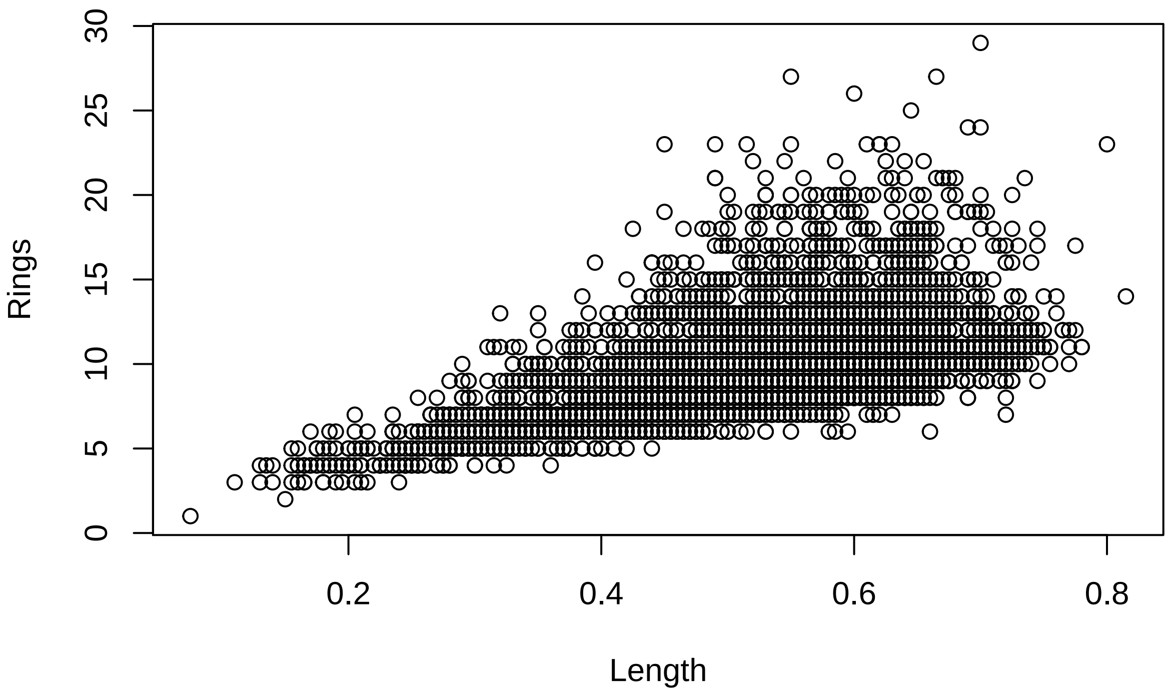

Length
Frequency
0
400
0.2
0.4
0.6
0.8
unlist(abalone[i])
Diameter
Frequency
0
400
0.1
0.2
0.3
0.4
0.5
0.6
unlist(abalone[i])
Height
Frequency
0
1500
0.0
0.2
0.4
0.6
0.8
1.0
1.2
unlist(abalone[i])
Whole_wt
Frequency
0
300
0.0
0.5
1.0
1.5
2.0
2.5
3.0
unlist(abalone[i])

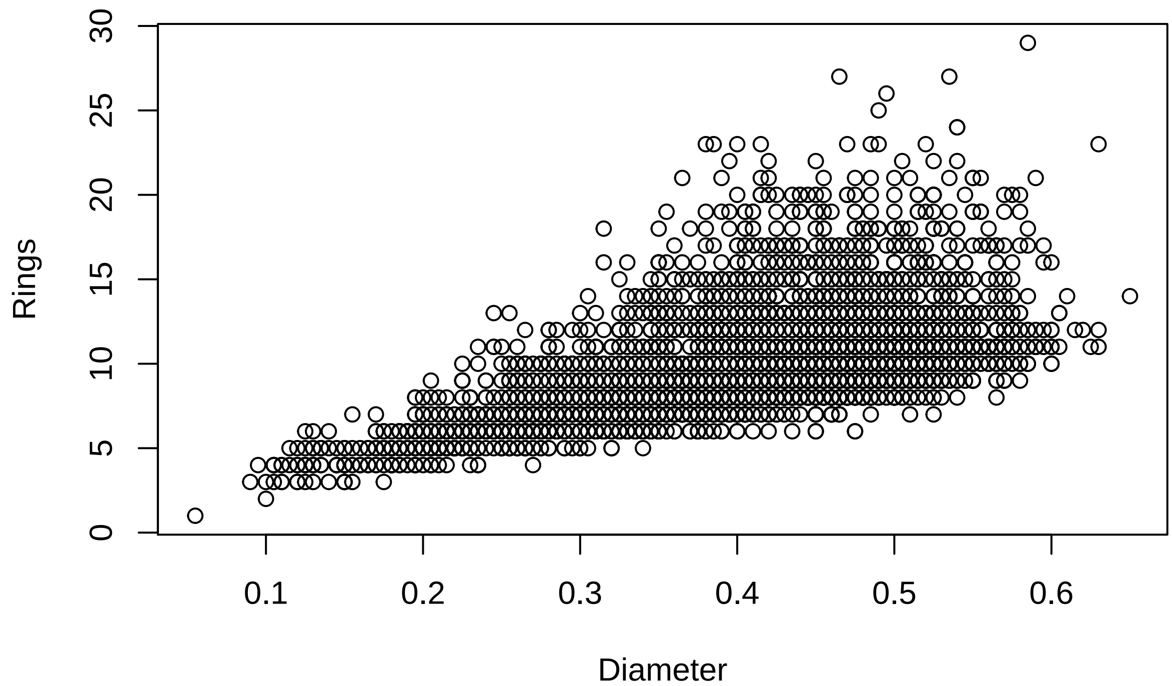


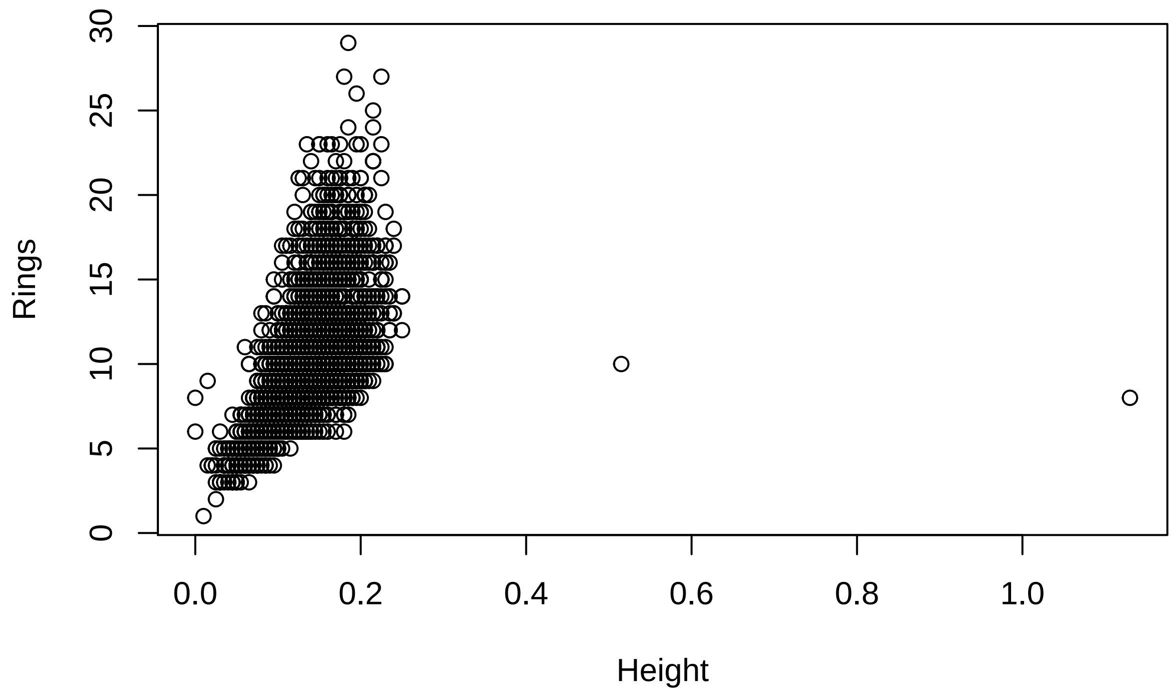
Considering the boxplots of all the features in the dataset, all the variables present outliers (by the definition of quantiles and interquantile range). Height has two significant outliers. In addition, The medians of the various different types of weights are more or less close to each other.

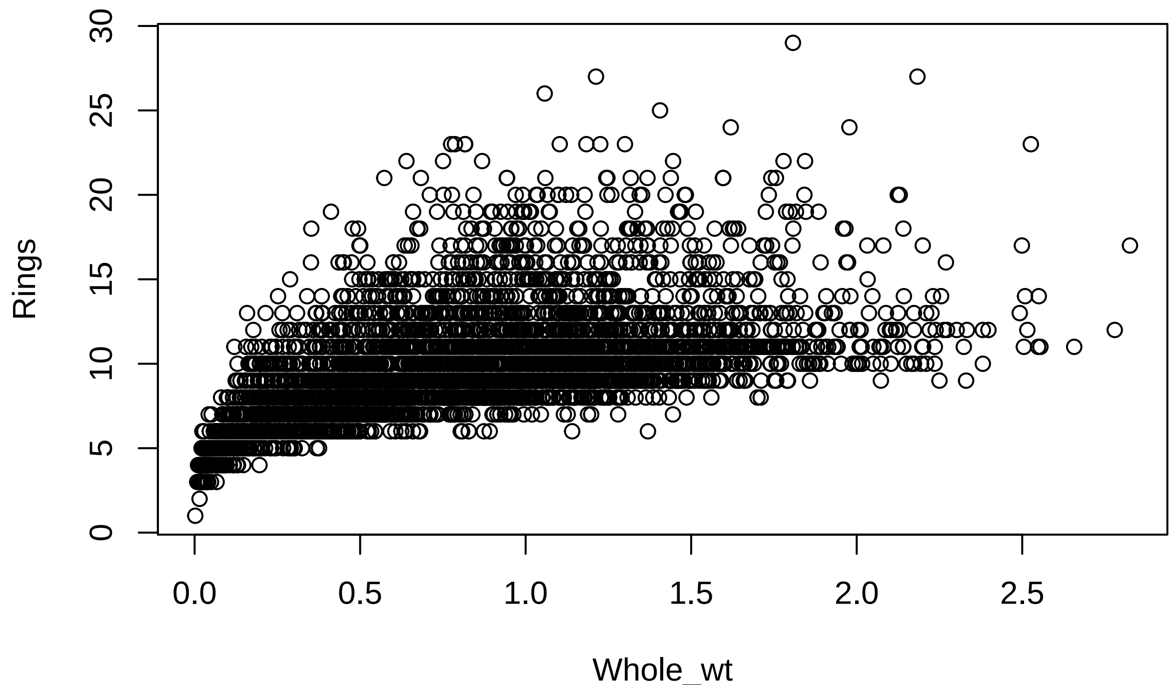
Considering the Histograms, It's easy to see how the distribution of Rings is more or less centered, the Length and the Diameter are left skewed (the frequency of larger values is bigger), while all the others present are right skewed (frequency of smaller values is bigger).

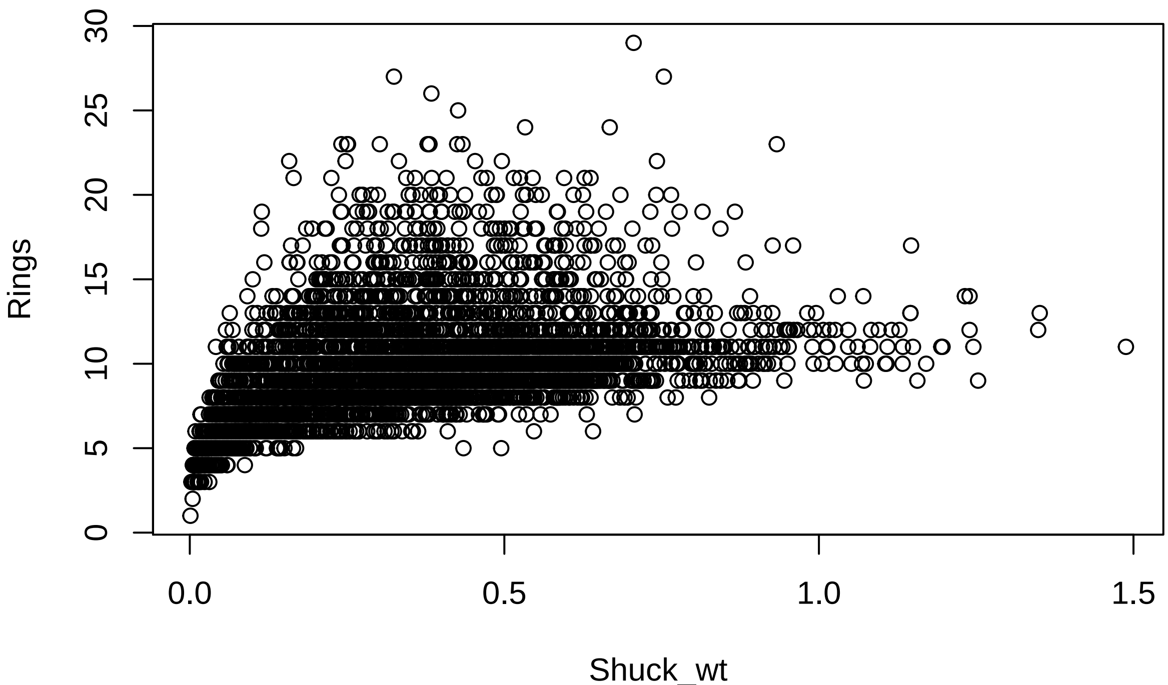
```r
#Q3

plot(Rings ~ Length + Diameter + Height + Whole_wt + Shuck_wt + Visc_wt + Shell_wt, data=abalone)
```


Rings
0
5
10
15
20
25
30
0.2
0.4
0.6
0.8
Length

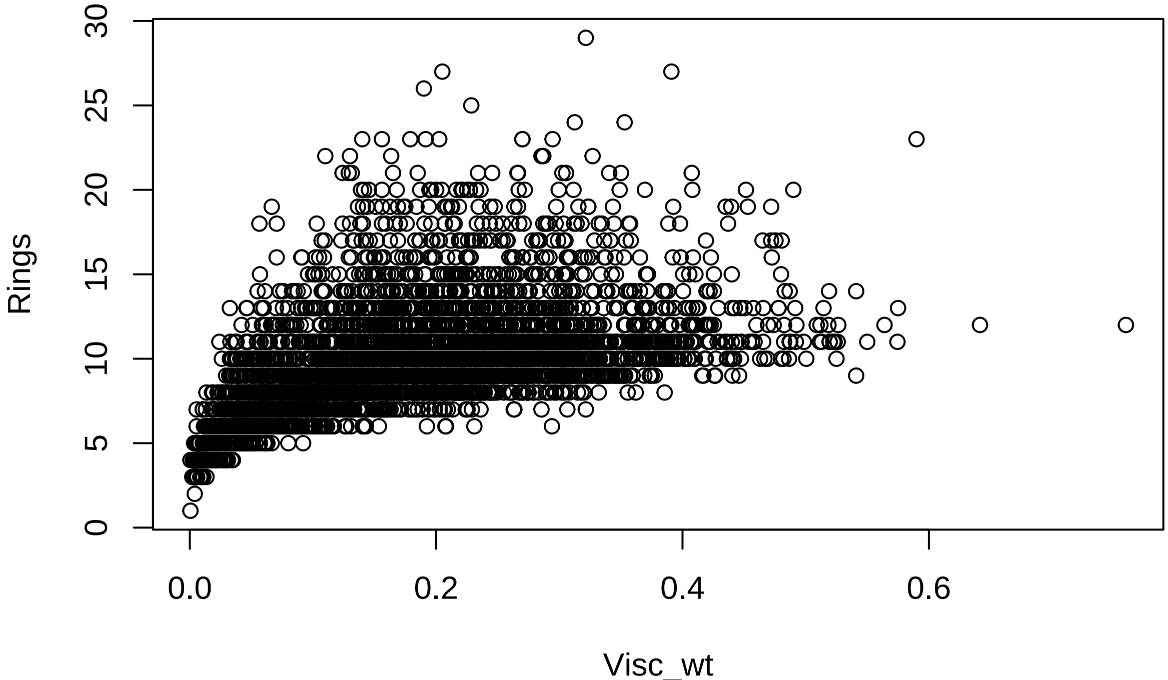

Rings
0
5
10
15
20
25
30
0.1
0.2
0.3
0.4
0.5
0.6
Diameter

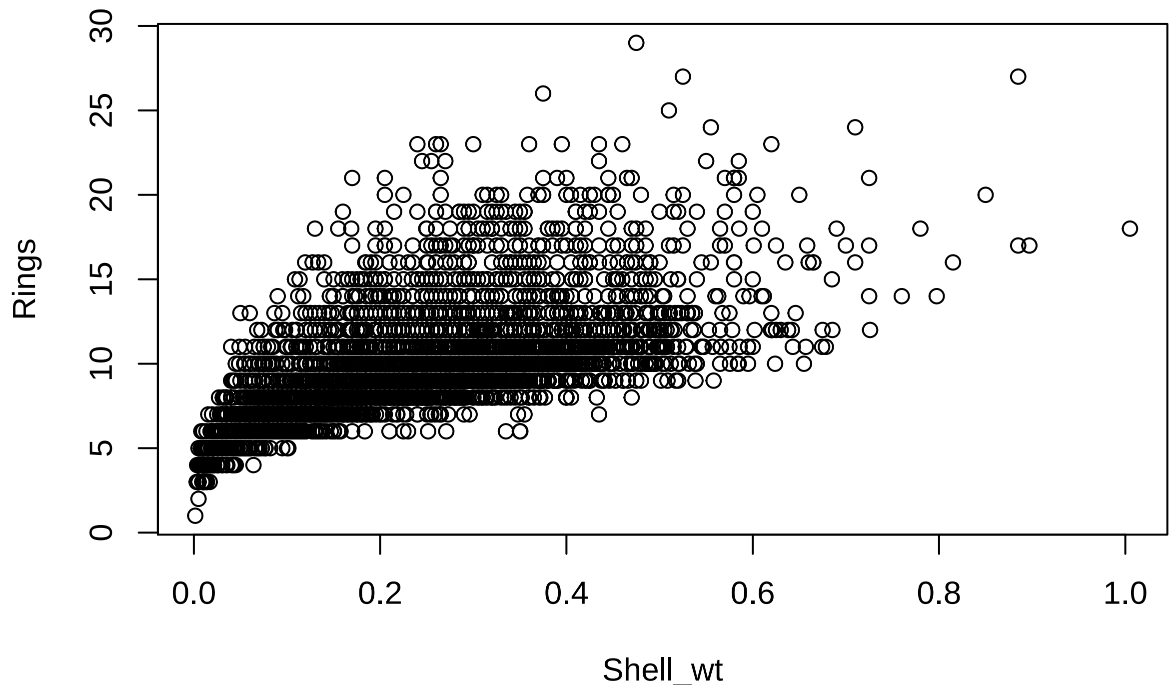

Rings
Height
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0


Rings
Whole_wt
0
5
10
15
20
25
30
0.0
0.5
1.0
1.5
2.0
2.5



Rings
Shuck_wt
0
5
10
15
20
25
30
0.0
0.5
1.0
1.5


Rings
Visc_wt
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6



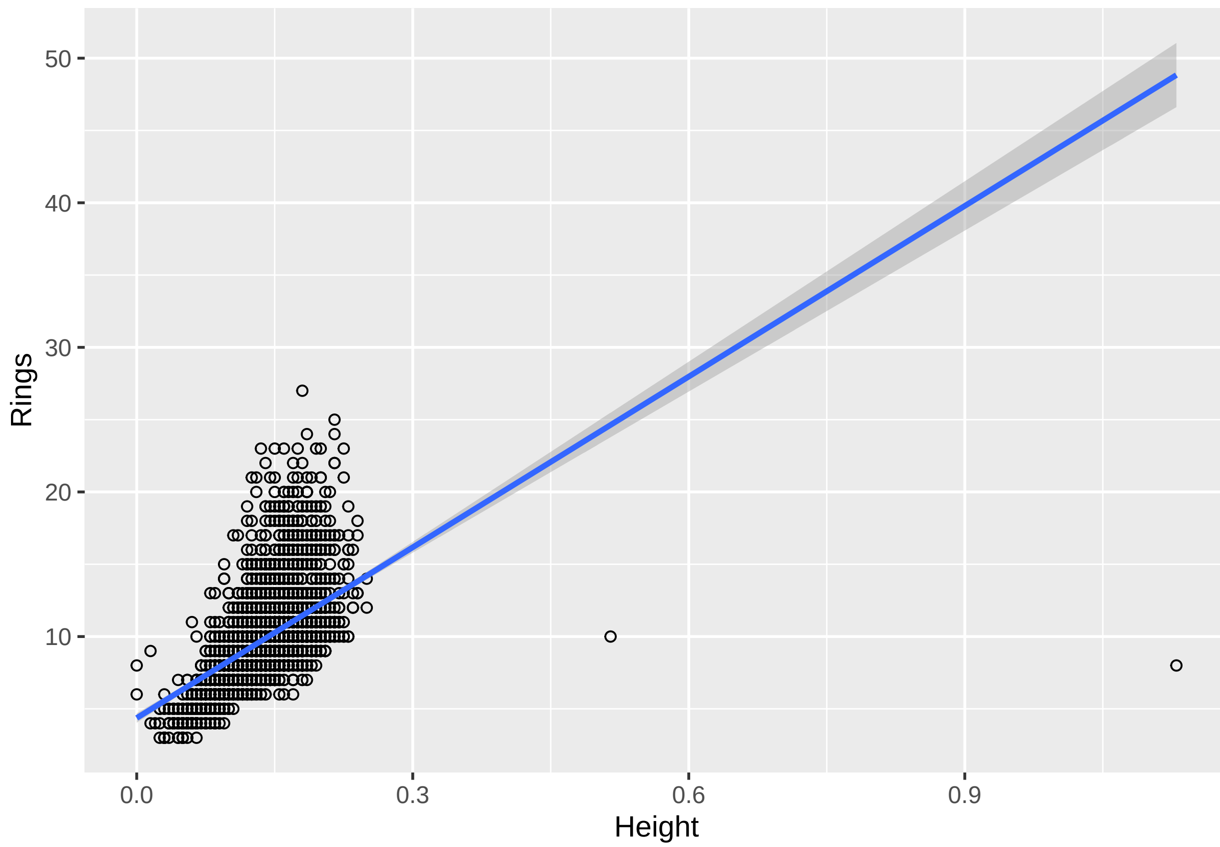
We can graphically see the positive correlation between Rings (and consequently, age of Abalone) and Height, confirming the biologists' hypothesis. In general, from the scatter plots, we can also see that there are linear correlations between Rings and other variables such as Length and Shell Weight.

```r
#Q4

linear_mod = lm(Rings ~ Height, data=abalone)
summary(linear_mod)
```

```
##
## Call:
## lm(formula = Rings ~ Height, data = abalone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.496  -1.657  -0.607   0.839  17.112
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9385     0.1443   27.30   <2e-16 ***
## Height       42.9714     0.9904   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.677 on 4175 degrees of freedom
## Multiple R-squared:  0.3108, Adjusted R-squared:  0.3106
## F-statistic:  1882 on 1 and 4175 DF,  p-value: < 2.2e-16
```

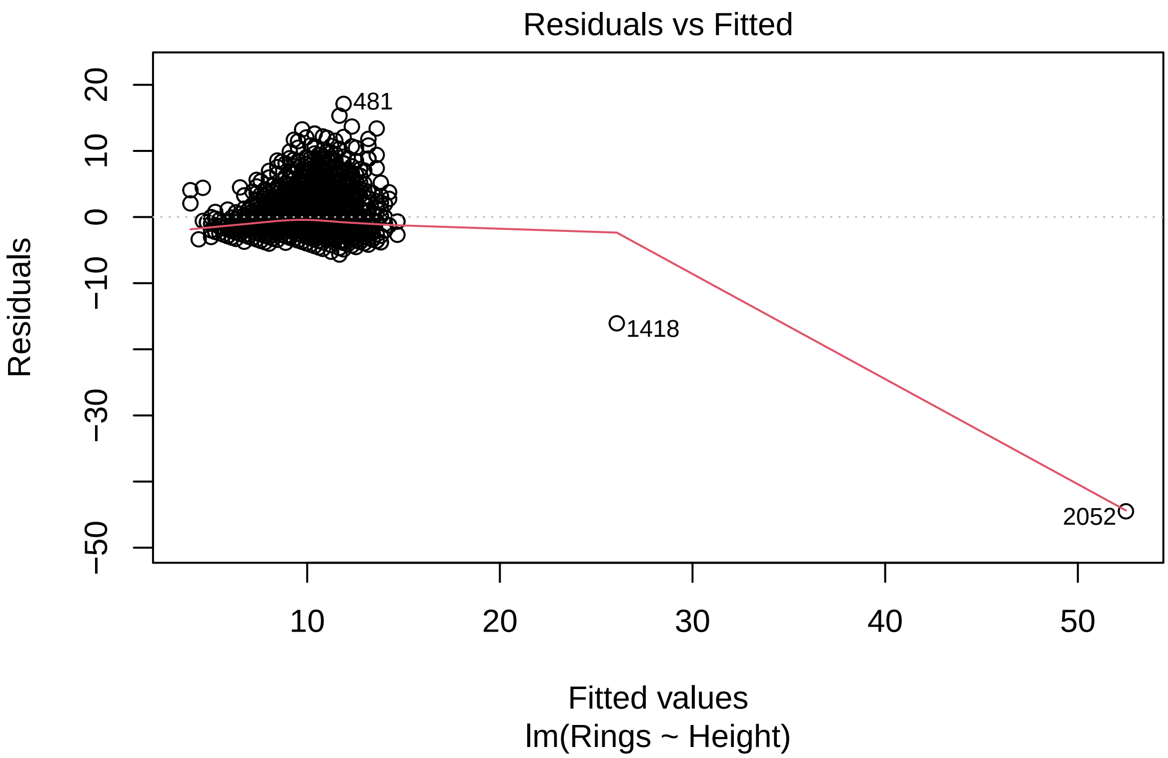
```r
#Q5

ggplot(abalone_train, aes(x=Height, y=Rings)) + geom_point(shape=1) + geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

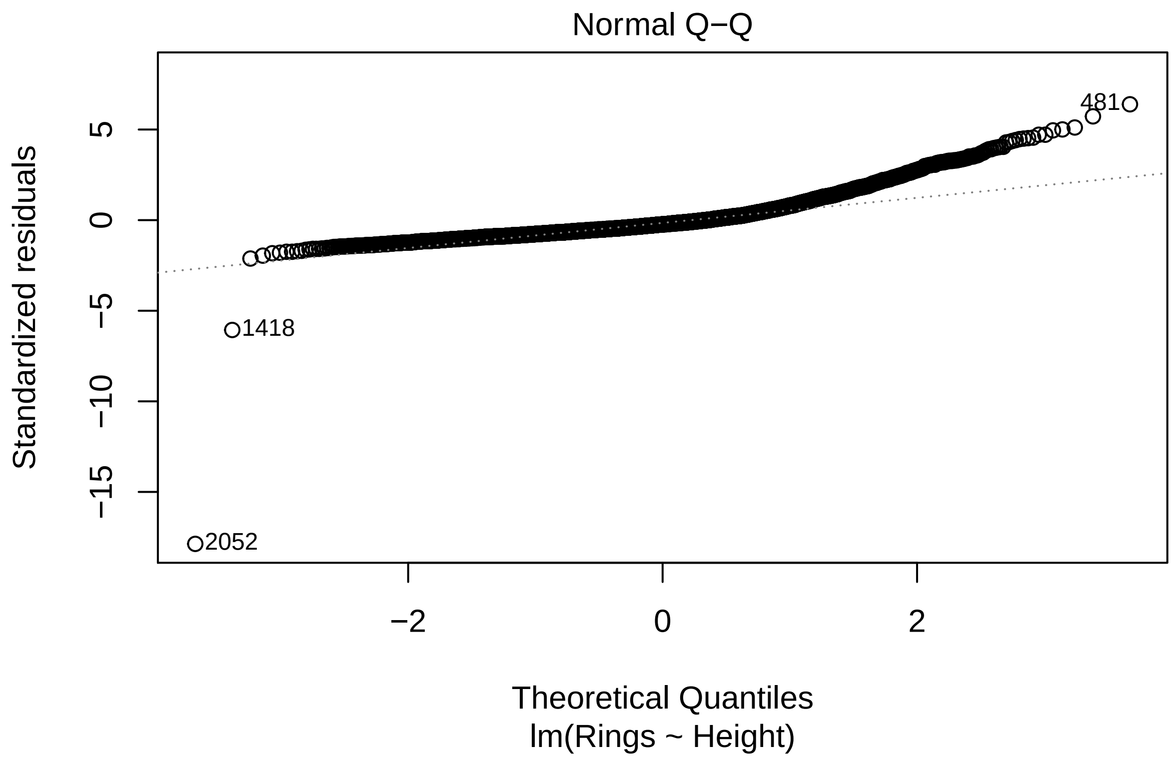


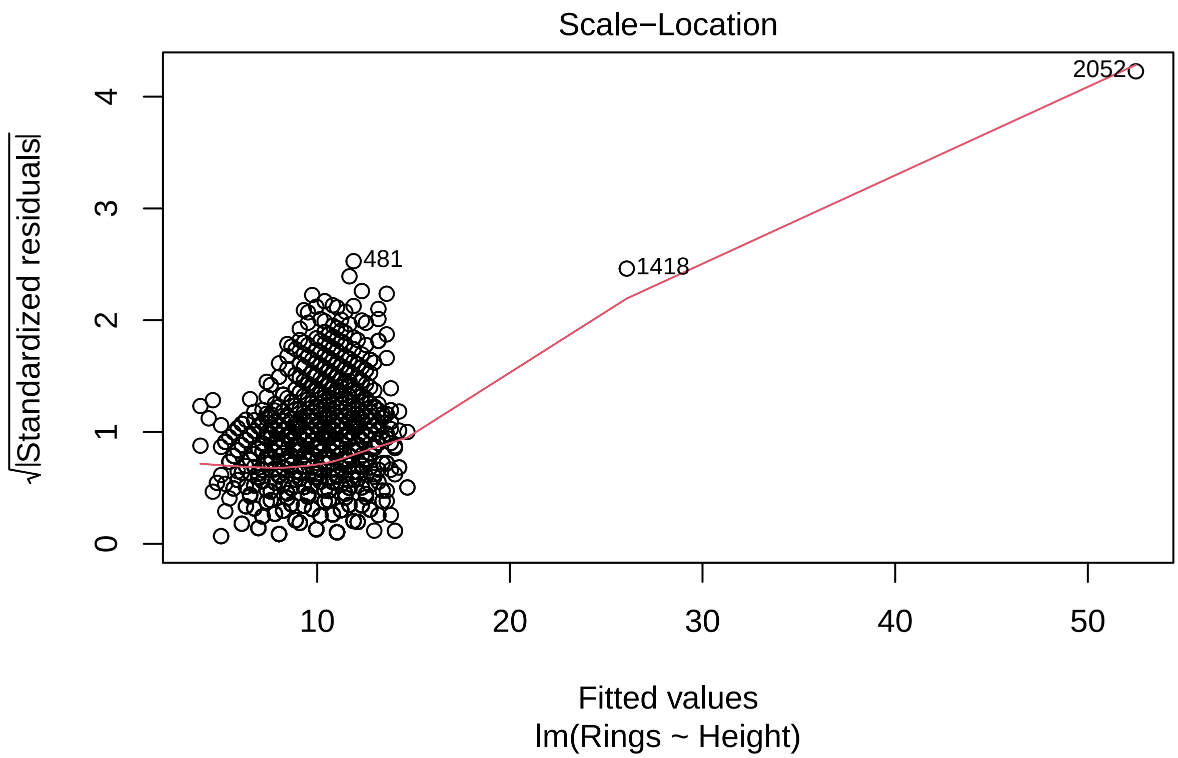
From the graph can be seen that are present two outliers of the predictor Height. Those two points are high leverage and are affecting the fit of the line. The line doesn't seem to be the best fit. Taking a polynomial or exponential function of Height might provide a better fit.

```r
#Q6

plot(linear_mod)
```

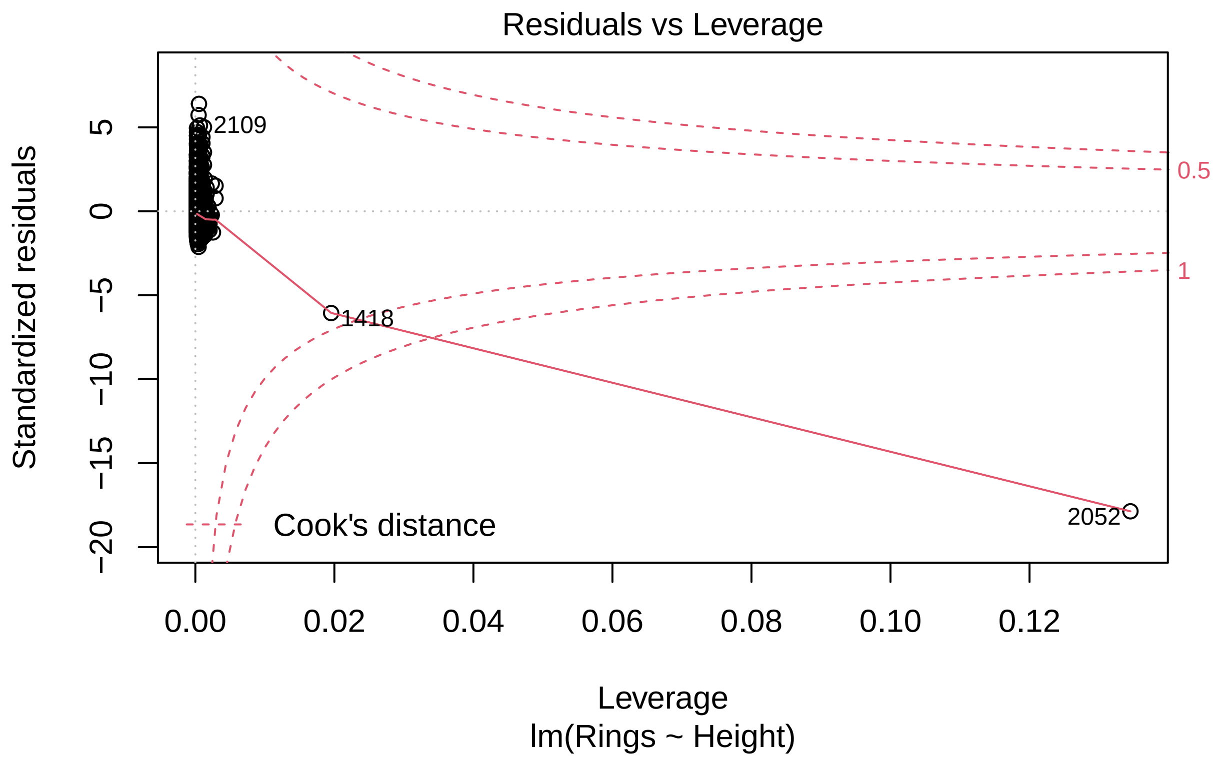


Residuals vs Fitted
Residuals
Fitted values
lm(Rings ~ Height)
481
1418
2052
20
10
0
−10
−30
−50
10
20
30
40
50


Normal Q−Q
Standardized residuals
5
0
−5
−10
−15
481
1418
2052
−2
0
2
Theoretical Quantiles
lm(Rings ~ Height)



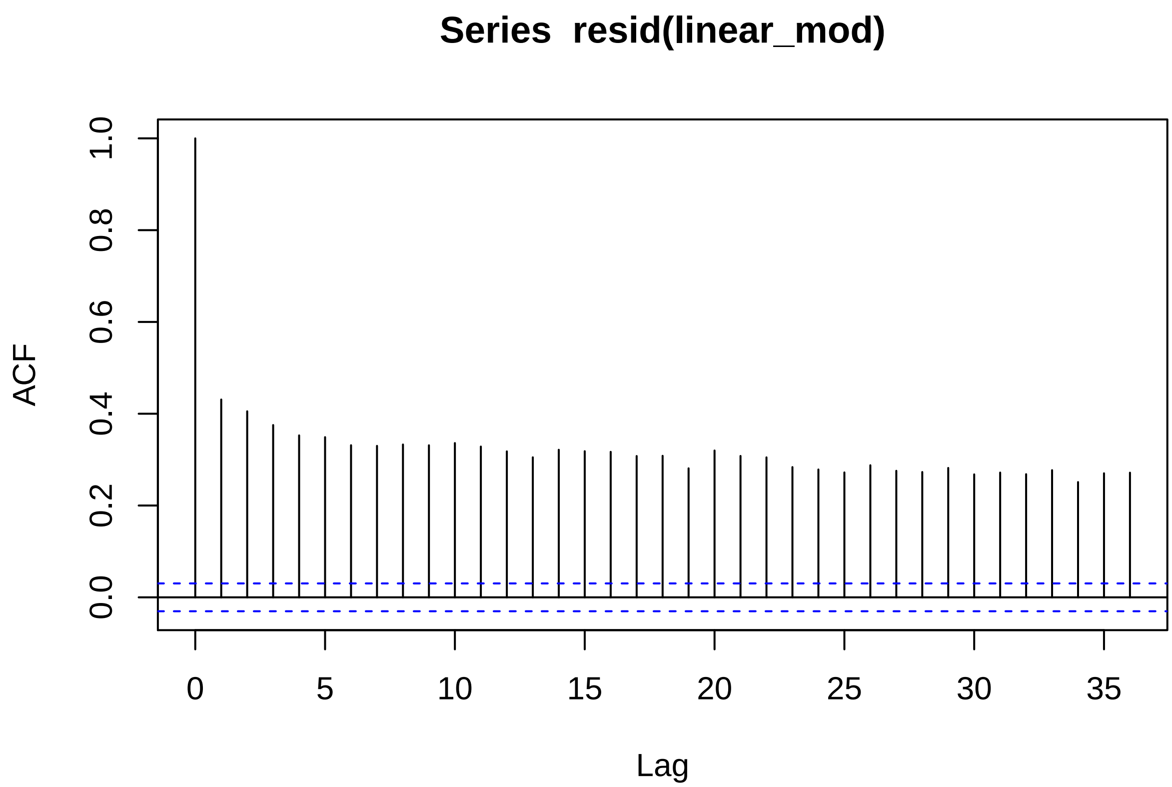
Scale−Location
√|Standardized residuals|
0
1
2
3
4
481
1418
2052
10
20
30
40
50
Fitted values
lm(Rings ~ Height)




```r
durbinWatsonTest(linear_mod, max.lag=10)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4309901      1.136388       0
##    2       0.4053487      1.187644       0
##    3       0.3753891      1.247076       0
##    4       0.3528196      1.292197       0
##    5       0.3490177      1.299796       0
##    6       0.3313801      1.334845       0
##    7       0.3299767      1.334554       0
##    8       0.3329437      1.327089       0
##    9       0.3312710      1.330374       0
##   10       0.3361000      1.318232       0
##  Alternative hypothesis: rho[lag] != 0
```

```r
acf(resid(linear_mod))
```




```r
bptest(linear_mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  linear_mod
## BP = 678.35, df = 1, p-value < 2.2e-16
```

```r
shapiro.test(resid(linear_mod))
```

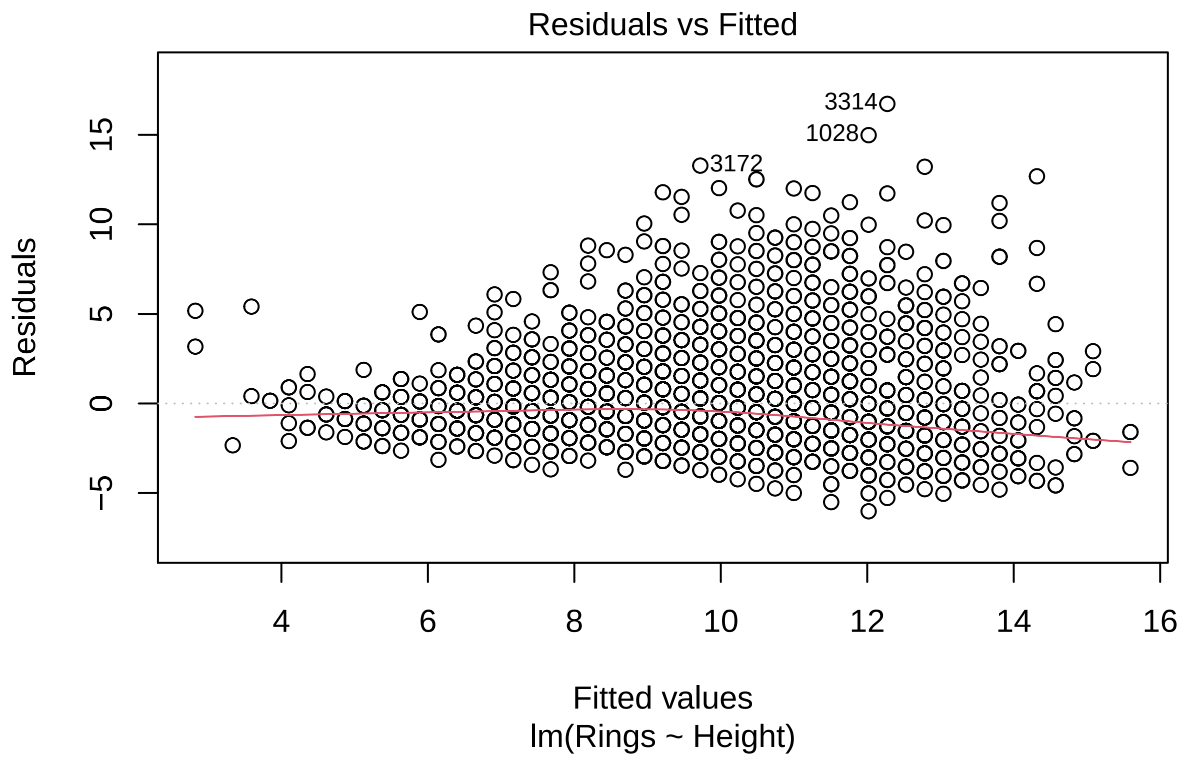
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(linear_mod)
## W = 0.83379, p-value < 2.2e-16
```

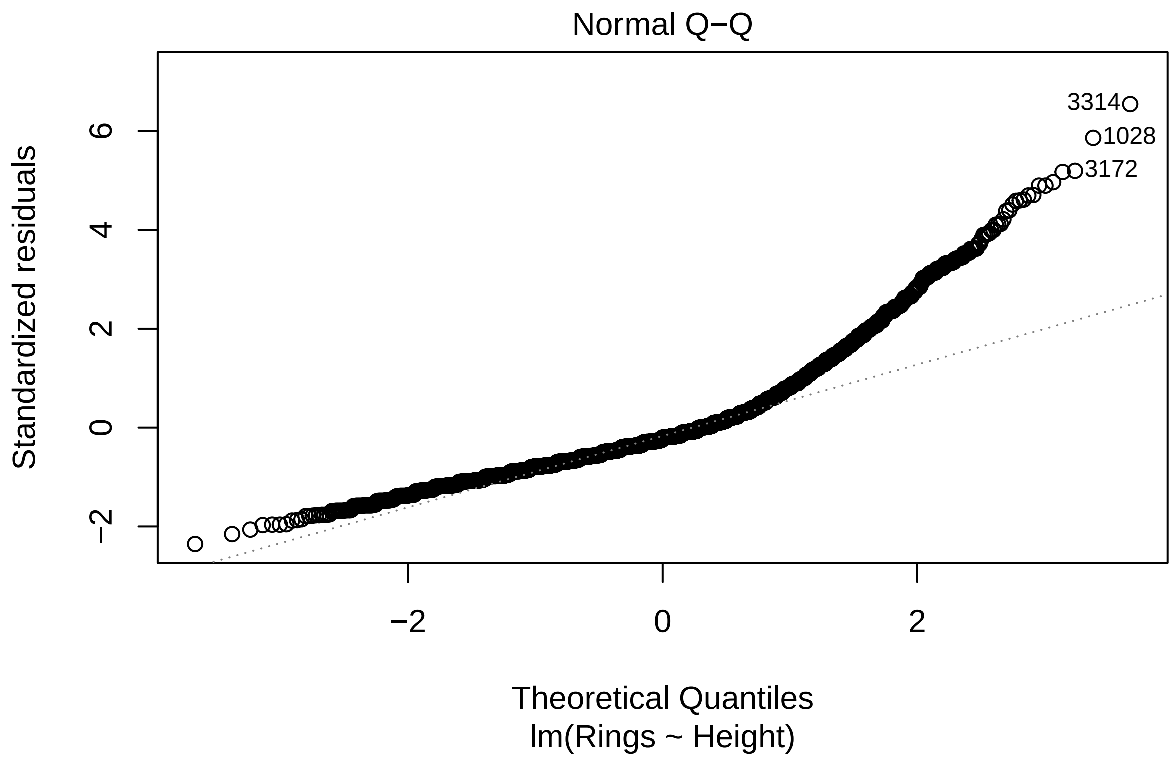
The errors are not centered since the Residuals-Fitted graph does not have a line which on average is zero due to the presence of two outliers in the data. The errors are Gaussian in the lower quantiles since in the Normal Q-Q plot more or less lies on the line that represent the quantiles of the standard normal. The plot diverges at higher quantiles, suggesting that we could perform feature engineering. The results of the Shapiro-Wilkes test also do not suggest Gaussian distribution of residuals. Possibly due to the presence of outliers, there is heteroskedasticity since the line in the Scale-Location plot is really far from being horizontal. In addition, the studentized Breusch-Pagan test has a very low p-value, so there is high probability of heteroskedasticity. The results of Durbin-Watson test suggest autocorrelation. This may be due to ordering in the data.

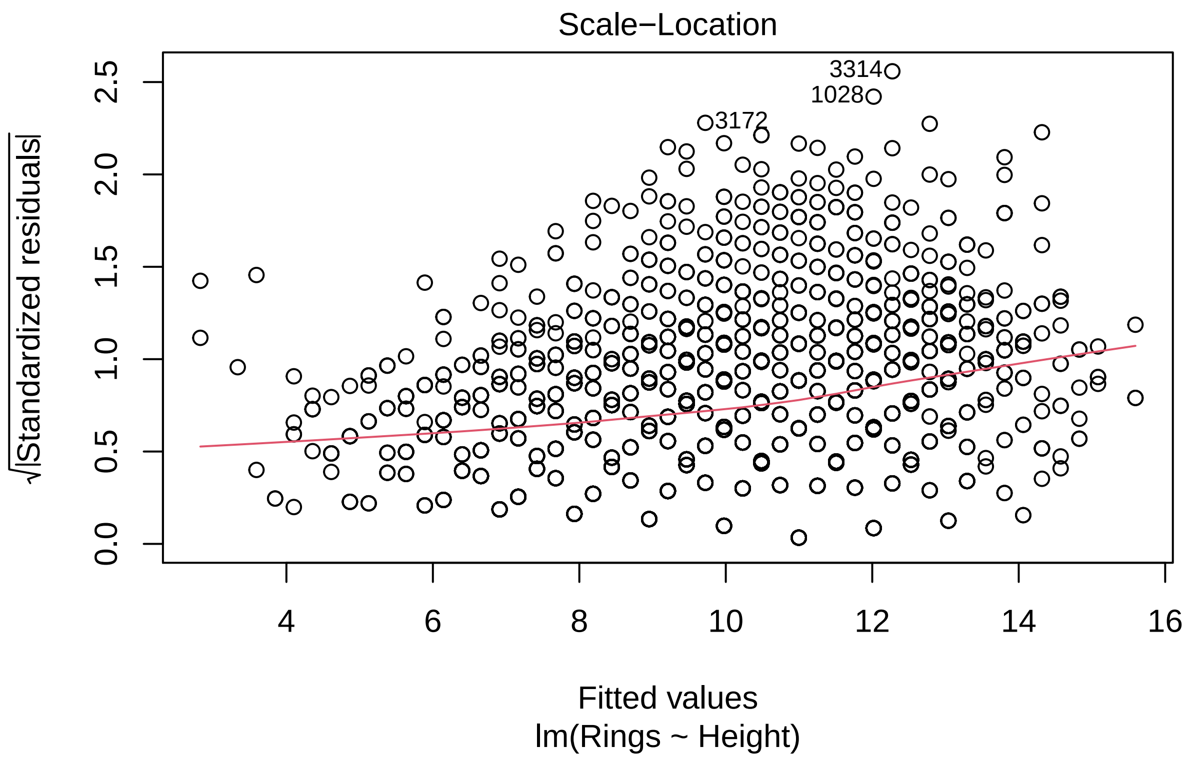
```r
#we remove the two outliers and sort the data randomly
new_abalone = abalone[-c(1418, 2052),]

set.seed(1234)
new_abalone <- new_abalone[sample(nrow(new_abalone)), ]

linear_mod_new = lm(Rings ~ Height, data=new_abalone)
summary(linear_mod_new)
```

```
##
## Call:
## lm(formula = Rings ~ Height, data = new_abalone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0187 -1.6770 -0.5294  0.8122 16.7259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8246     0.1485   19.02   <2e-16 ***
## Height       51.0780     1.0281   49.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 4173 degrees of freedom
## Multiple R-squared:  0.3717, Adjusted R-squared:  0.3715
## F-statistic:  2468 on 1 and 4173 DF,  p-value: < 2.2e-16
```

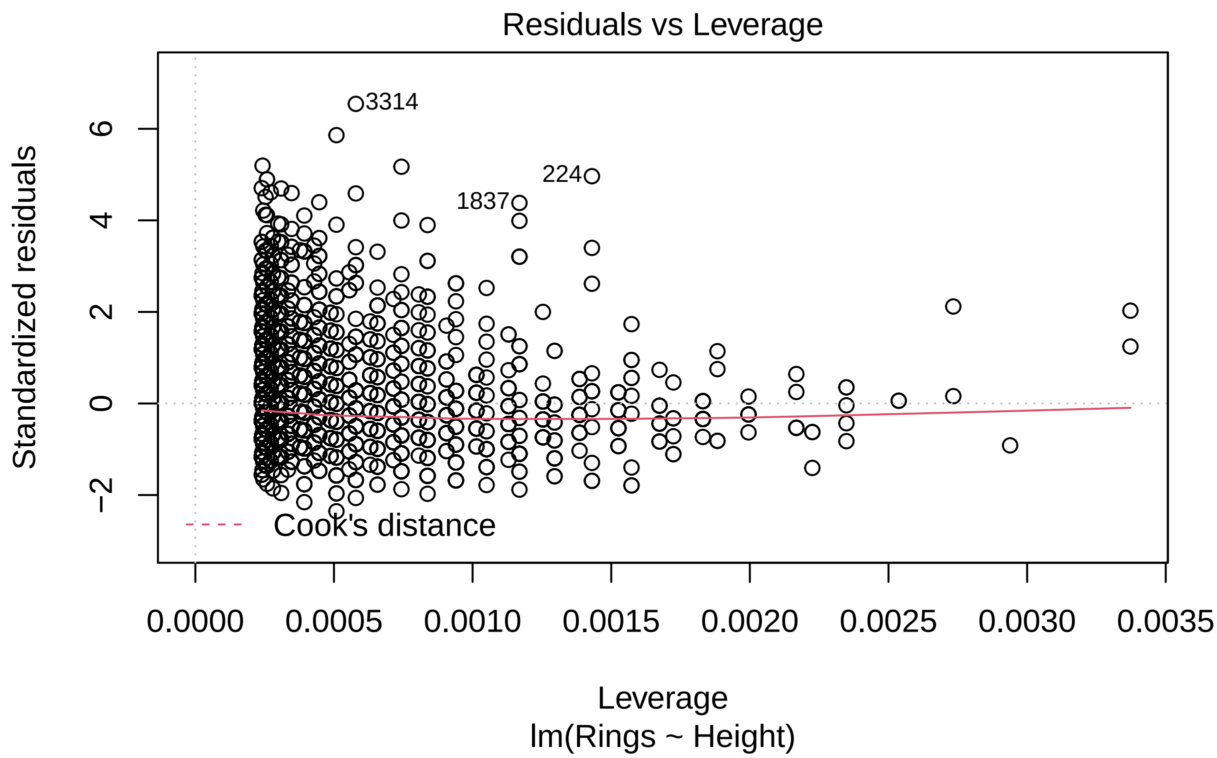
```r
plot(linear_mod_new)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Rings ~ Height)
3314
1028
3172
15
10
5
0
−5
4
6
8
10
12
14
16


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Rings ~ Height)
3314
1028
3172
−2
0
2
4
6



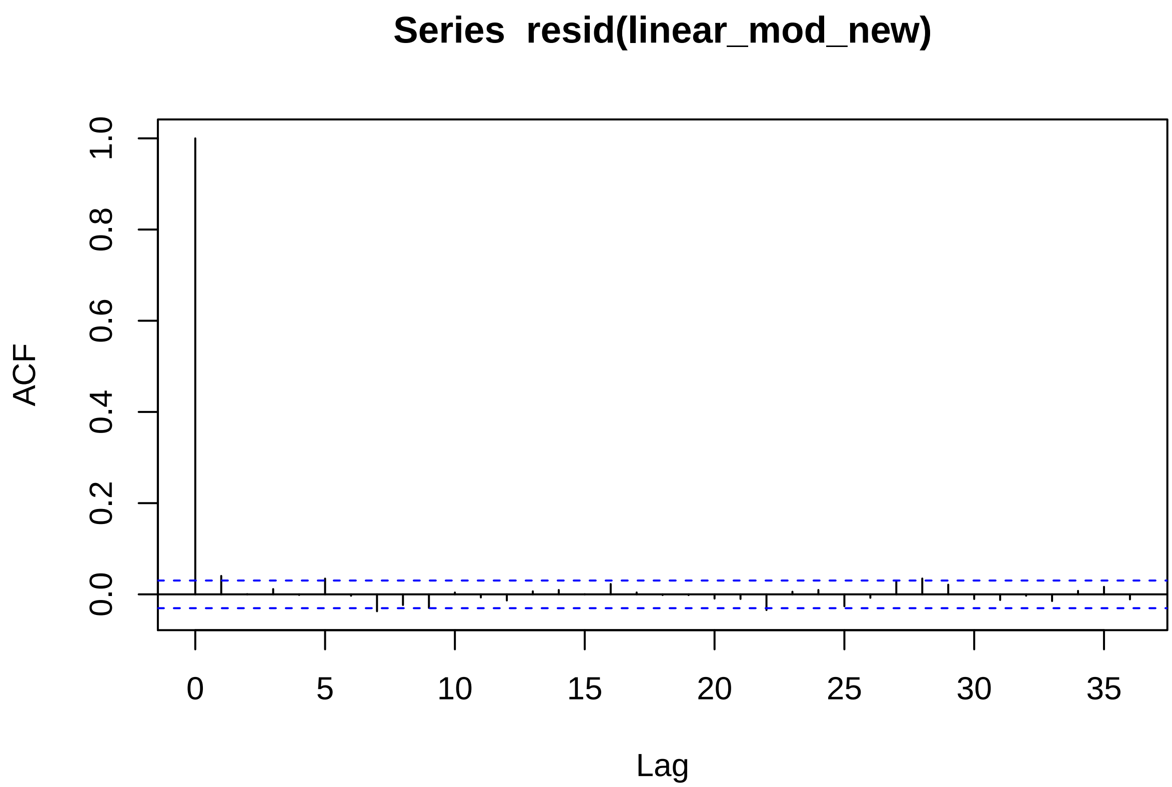
Scale−Location
3314
1028
3172
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
4
6
8
10
12
14
16
Fitted values
lm(Rings ~ Height)




```r
durbinWatsonTest(linear_mod_new, max.lag=10)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    0.0404600551      1.917993   0.008
##    2    0.0004429449      1.997446   0.890
##    3    0.0116268036      1.974962   0.450
##    4   -0.0012948336      2.000263   0.954
##    5    0.0346083421      1.928456   0.022
##    6   -0.0030990245      2.003624   0.872
##    7   -0.0370314207      2.071192   0.016
##    8   -0.0233494914      2.043751   0.130
##    9   -0.0298373590      2.056468   0.070
##   10    0.0040797957      1.988193   0.808
##  Alternative hypothesis: rho[lag] != 0
```

```r
acf(resid(linear_mod_new))
```




```r
bptest(linear_mod_new)
```

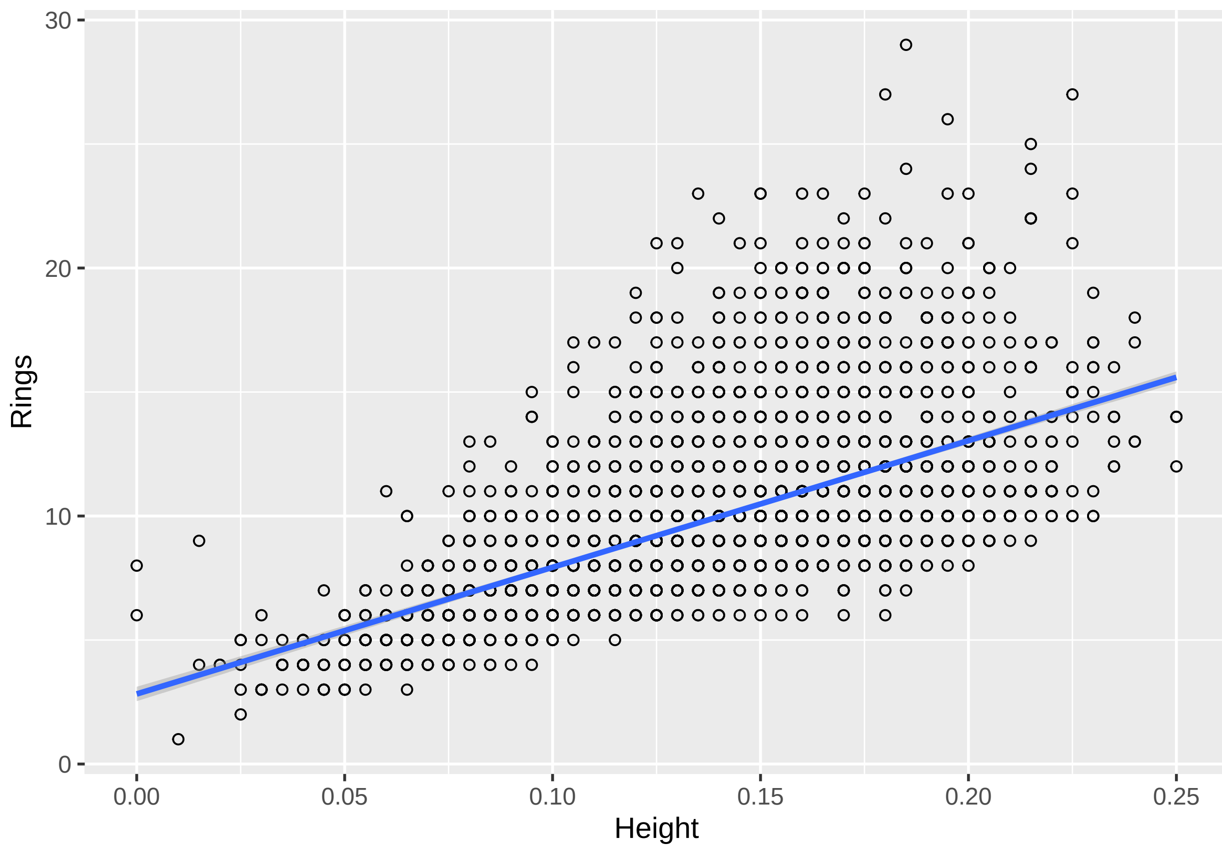
```
##
##  studentized Breusch-Pagan test
##
## data:  linear_mod_new
## BP = 120.98, df = 1, p-value < 2.2e-16
```

```r
shapiro.test(resid(linear_mod_new))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(linear_mod_new)
## W = 0.88304, p-value < 2.2e-16
```

```r
ggplot(new_abalone, aes(x=Height, y=Rings)) + geom_point(shape=1) + geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



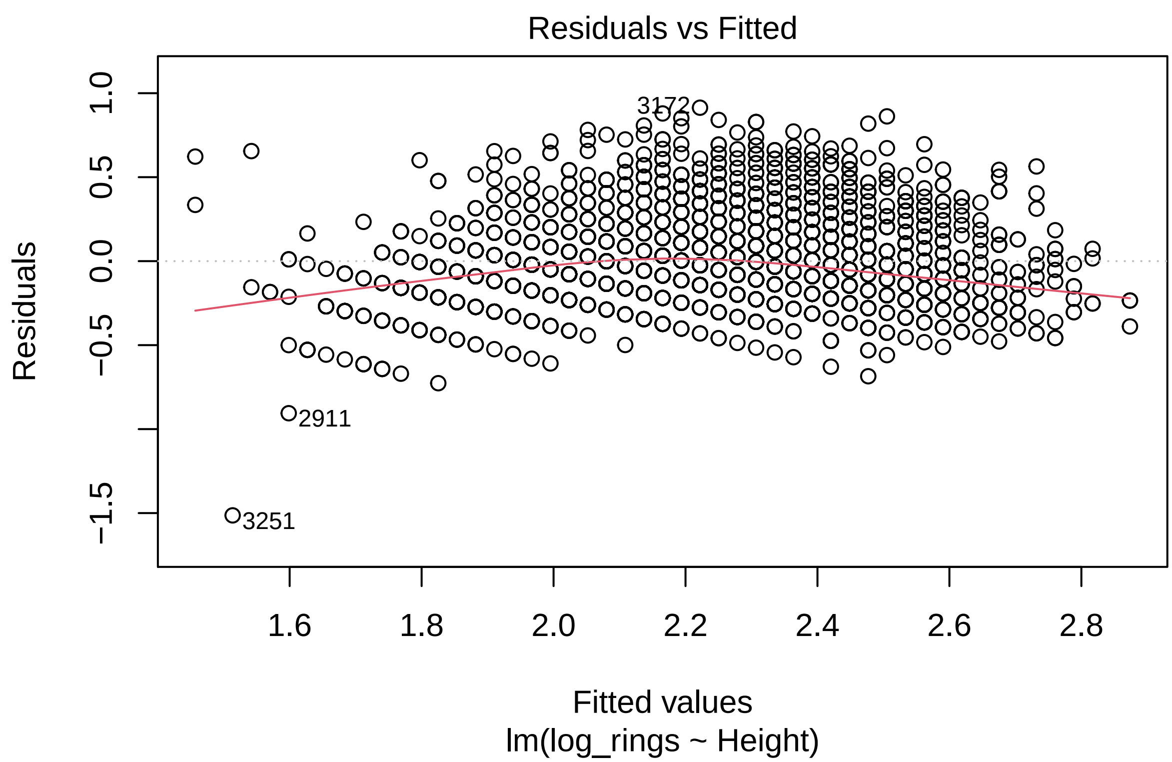
We removed the outliers sequentially till none of the points have Cook's distance greater than 1 From the new graph we can see that the elimination of the outliers allow us to better satisfy the postulates. The errors are more centered since the red line in the residuals vs fitted plot is on average more close to 0. However, we still see some trend in the variance of the residuals. Furthermore, the results of the B-P test also suggest that there exists heteroskedasticity. The results of the Q-Q plot and the S-W test suggest that the residuals do not follow a Gaussian Distribution. Sorting the data seems to have removed the apparent autocorrelation in the residual terms as seen from the results of the D-W test.

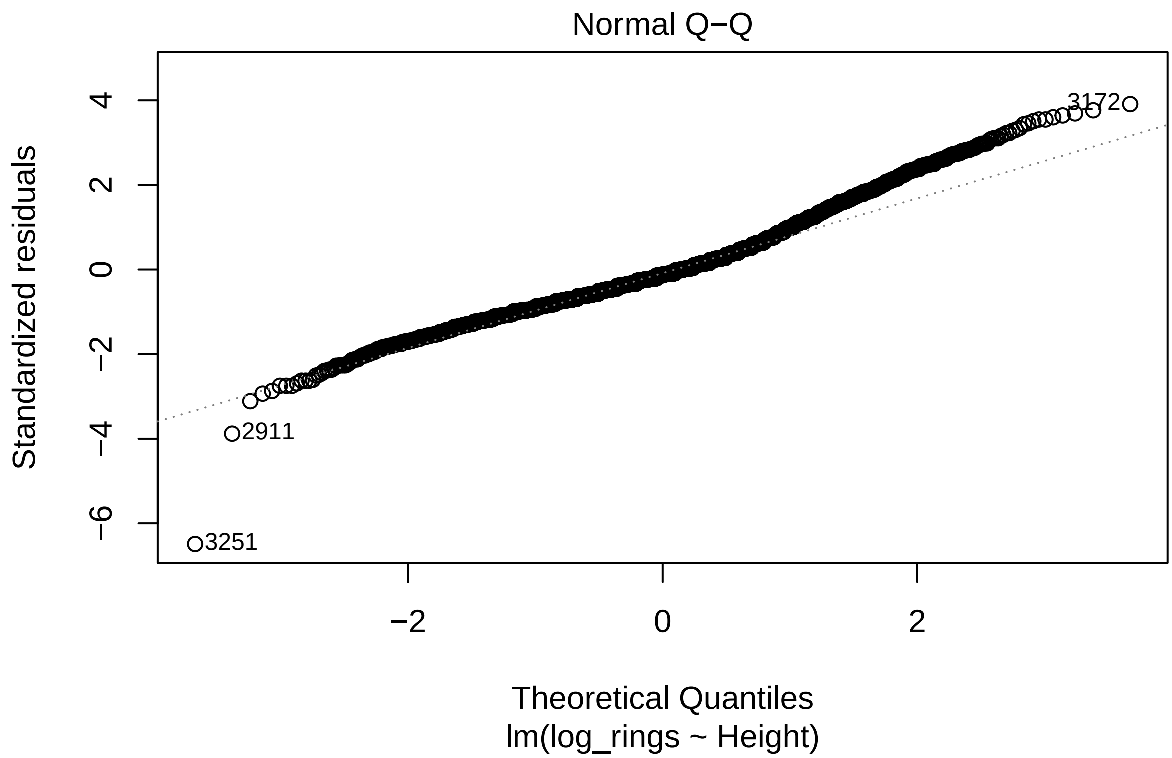
```
#here we used the logarithm of the number of Rings to get a better fit
new_abalone$log_rings = log(new_abalone$Rings)

linear_mod_log = lm(log_rings ~ Height, data=new_abalone)
summary(linear_mod_log)
```

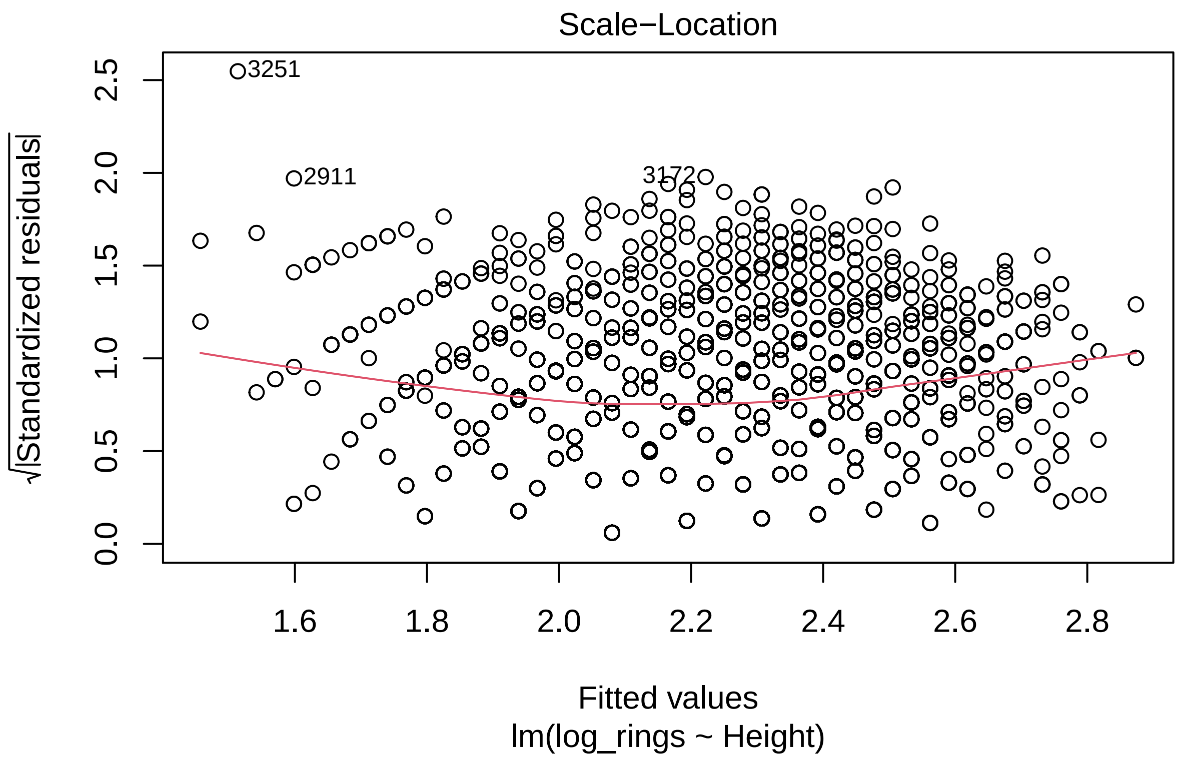
```
##
## Call:
## lm(formula = log_rings ~ Height, data = new_abalone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51358 -0.15916 -0.02918  0.11926  0.91353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45691    0.01357  107.39   <2e-16 ***
## Height       5.66708    0.09394   60.33   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2336 on 4173 degrees of freedom
## Multiple R-squared:  0.4658, Adjusted R-squared:  0.4657
## F-statistic:  3639 on 1 and 4173 DF,  p-value: < 2.2e-16
```

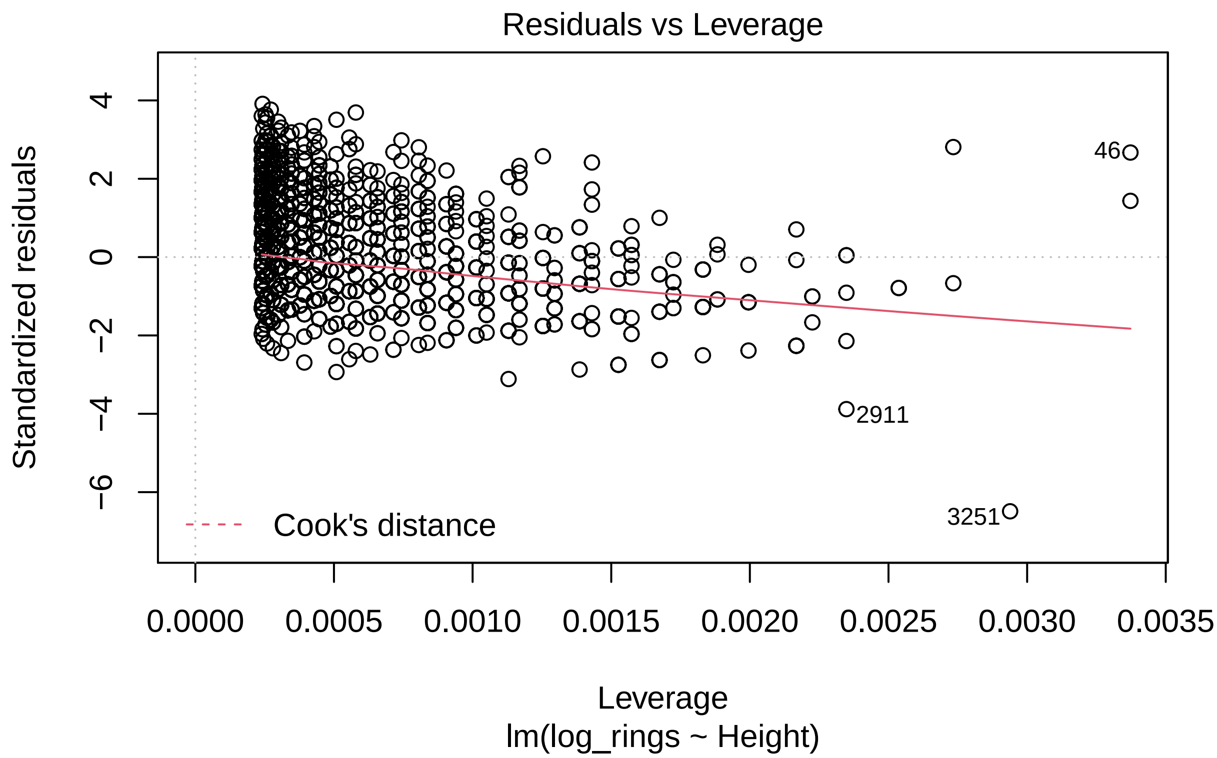
```
plot(linear_mod_log)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_rings ~ Height)
3172
2911
3251
4
2
0
−2
−4
−6
−2
0
2



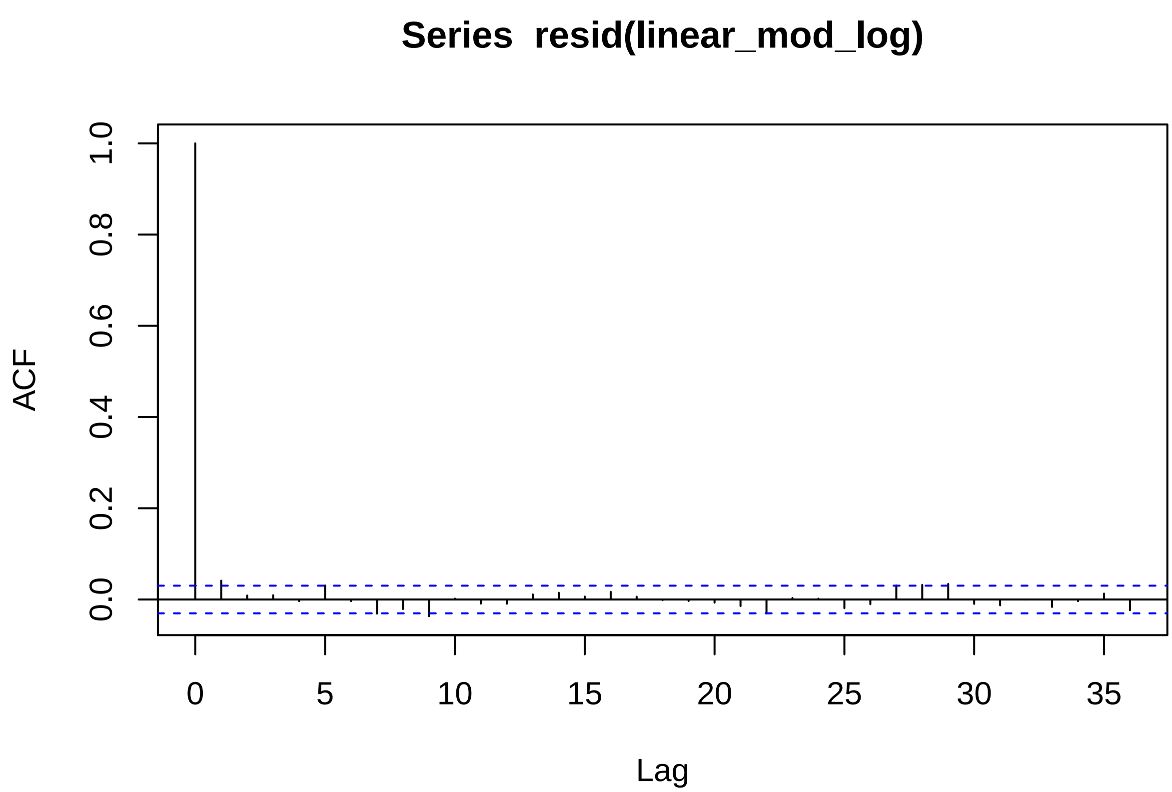
Scale−Location
3251
2911
3172
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Fitted values
lm(log_rings ~ Height)




```r
durbinWatsonTest(linear_mod_log, max.lag=10)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.041343992      1.916610   0.010
##    2     0.008777594      1.980946   0.516
##    3     0.009320182      1.979765   0.544
##    4    -0.003743386      2.005099   0.878
##    5     0.030866439      1.935863   0.046
##    6    -0.003602773      2.004438   0.822
##    7    -0.030868389      2.058518   0.046
##    8    -0.020888574      2.038555   0.152
##    9    -0.036724065      2.069696   0.014
##   10     0.001973726      1.991522   0.878
##  Alternative hypothesis: rho[lag] != 0
```

```r
acf(resid(linear_mod_log))
```




```r
bptest(linear_mod_log)
```

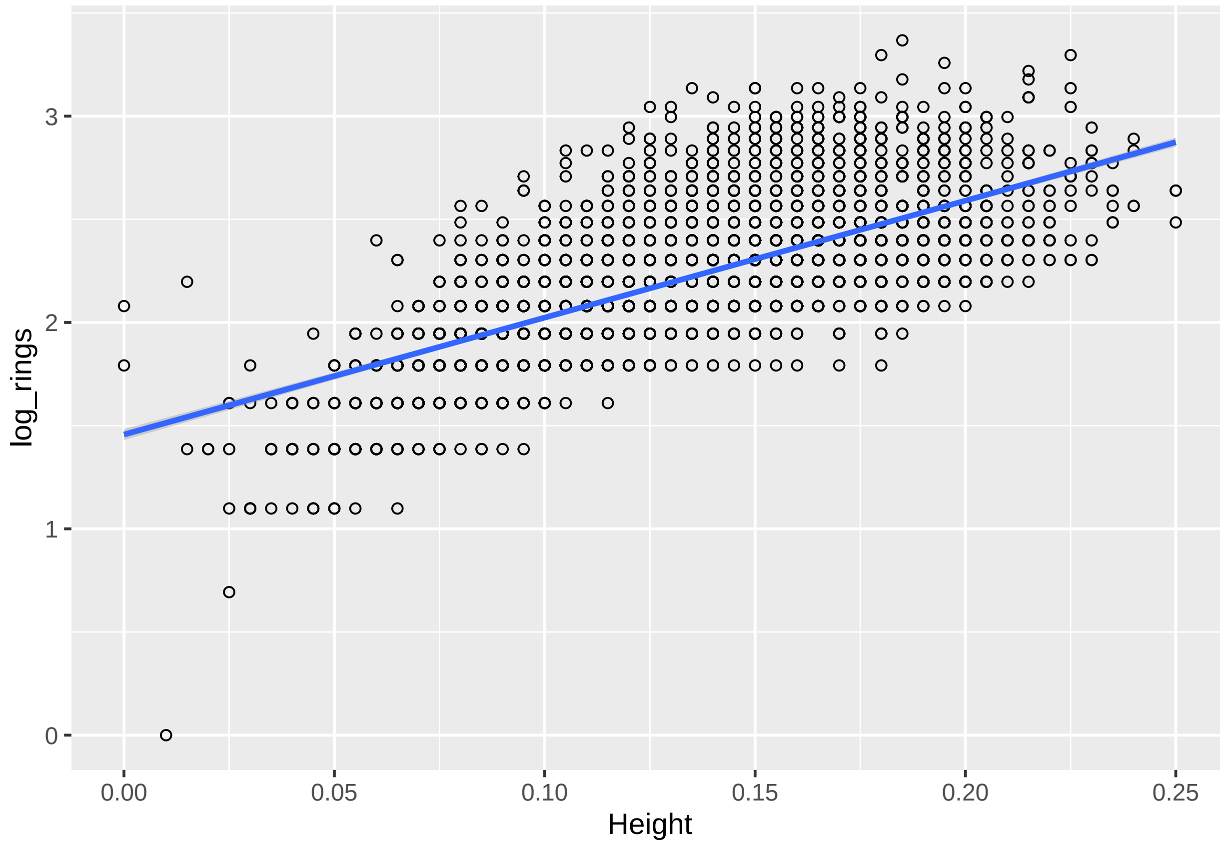
```
##
##  studentized Breusch-Pagan test
##
## data:  linear_mod_log
## BP = 2.9155, df = 1, p-value = 0.08773
```

```r
shapiro.test(resid(linear_mod_log))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(linear_mod_log)
## W = 0.97196, p-value < 2.2e-16
```

```r
ggplot(new_abalone, aes(x=Height, y=log_rings)) + geom_point(shape=1) +geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We decided to use the logarithm of the Rings as the response variable. This appears to better satisfy the postulate of homoskedasticity as seen from the results of the results of the B-P test. It also seems to better satisfy the condition of Gaussian distribution of residuals as we get a better value of the S-W statistic. Lastly, we observe a better fit as seen from the graph.

```
#Q7
confint(linear_mod_log, level=0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 1.430312 1.483506
## Height      5.482899 5.851252
```

In the context of the problem, these confidence intervals (of the coefficients) means that an additional unit change in Height will change the response variable (number of rings or its logarithm) by a value present in the confidence interval 95% of the times (so with 95% confidence).

```
#Q8
```

As the p-value is much less that an hypothetical 0.05 alpha, we reject the null hypothesis that $\beta_1 = 0$. Hence, there is a statistically significant relationship between the Height and the number of rings.