Project: Abalones age prediction

Ali Haidar, Sébastien Meyer

Introduction

Abalones are one type of reef-dwelling marine snails. It is difficult to tell the ages of abalones because their shellsizes not only depend on how old they are, but also depend on the availability of food. The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. We are interested in using some of abalones physical measurements, especially the height measurement to predict their ages. Biologists believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages. In particular, that a larger height is associated with an older age.

The dataset and its description are available at https://archive.ics.uci.edu/ml/datasets/Abalone.

Global libraries and parameters

Firstly, we import the necessary libraries.

```
library(readr)
library(carData)
library(car)
library(knitr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggfortify)
```

Then, we set up global parameters.

```
set.seed(42)
```

Importing the data

```
# Read the csv file
abalone <- read.csv2("abalone_data.csv", header = T, sep = ",")

# Lowercase column names
names(abalone) <- tolower(names(abalone))

# Data types
abalone$sex <- as.factor(abalone$sex)
abalone$whole_weight <- as.double(abalone$whole_weight)
abalone$shucked_weight <- as.double(abalone$shucked_weight)
abalone$viscera_weight <- as.double(abalone$viscera_weight)
```

```r
abalone$shell_weight <- as.double(abalone$shell_weight)

# Splitting dataset in train and test using 70/30 method
indices <- sample(seq_len(nrow(abalone)), size = 0.3 * nrow(abalone))
abalone_train <- abalone[-indices, ]
abalone_test <- abalone[indices, ]
```

Part I: EDA and Model validation

Question 2. Find summary measures of each variables (mean, variance, range, etc). Examine the variables individually (univariate). Graphically display each. Describe what you see.

```r
str(abalone)
```

```
## 'data.frame':    4176 obs. of  9 variables:
##  $ sex           : Factor w/ 3 levels "F","I","M": 3 1 3 2 2 1 1 3 1 1 ...
##  $ length        : int  70 106 88 66 85 106 109 95 110 105 ...
##  $ diameter      : int  53 84 73 51 60 83 85 74 88 76 ...
##  $ height        : int  18 27 25 16 19 30 25 25 30 28 ...
##  $ whole_weight  : num  45.1 135.4 103.2 41 70.3 ...
##  $ shucked_weight: num  19.9 51.3 43.1 17.9 28.2 47.4 58.8 43.3 62.9 38.8 ...
##  $ viscera_weight: num  9.7 28.3 22.8 7.9 15.5 28.3 29.9 22.5 30.2 29.5 ...
##  $ shell_weight  : num  14 42 31 11 24 66 52 33 64 42 ...
##  $ rings         : int  7 9 10 7 8 20 16 9 19 14 ...
```

In the Abalone dataset, we have the following variables:

- **sex**: *factor* corresponding to the sex of the snail, which can be male (M), female (F) and infant (I)
- **length**: *integer* corresponding to the length of the shell
- **diameter**: *integer* corresponding to the diameter of the shell, perpendicular to length
- **height**: *integer* corresponding to the height of the meat inside the shell
- **whole_weight**: *double* corresponding to the weight of the whole abalone
- **shucked_weight**: *double* corresponding to the weight of the meat inside the shell
- **viscera_weight**: *double* corresponding to the weight of the gut after bleeding
- **shell_weight**: *double* corresponding to the weight of the shell alone
- **rings**: *integer* corresponding to the number of rings on the shell, +1.5 gives the age of the abalone in years

```r
summary(abalone)
```

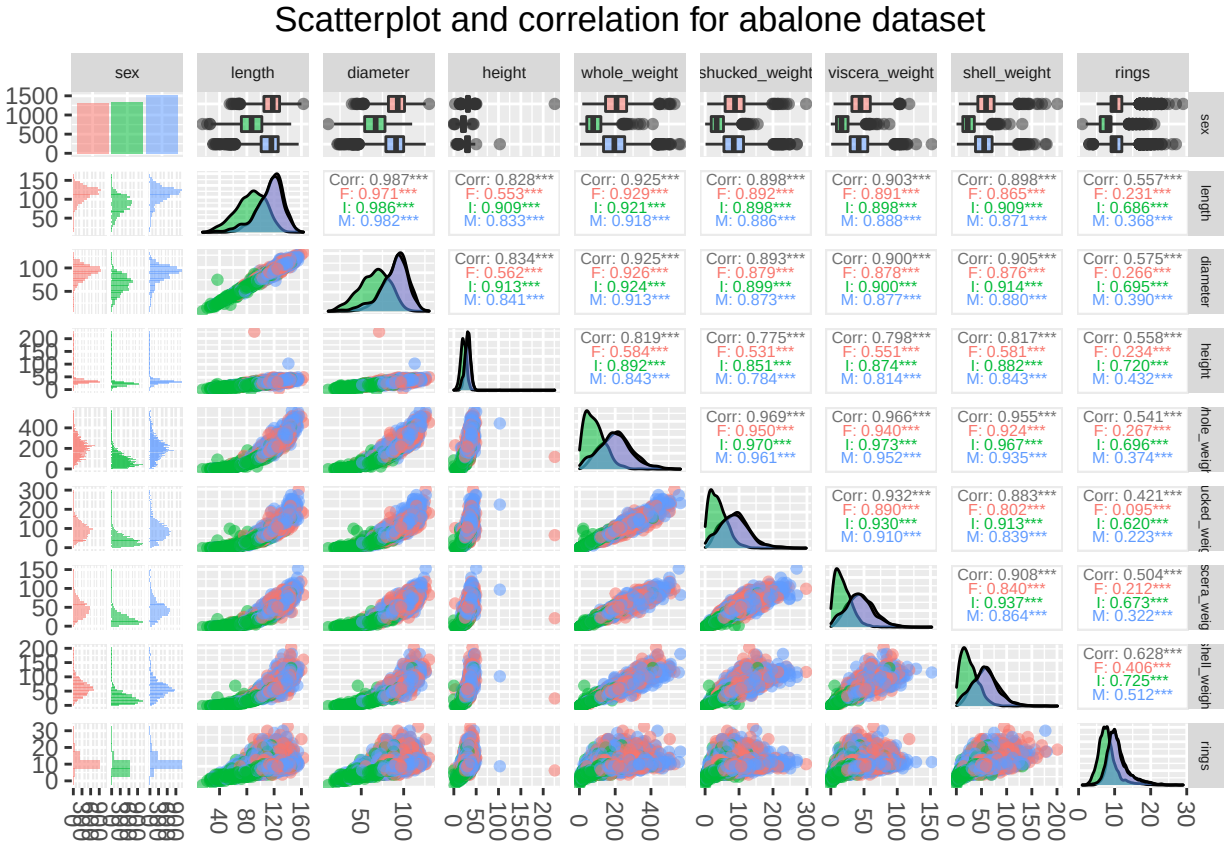
```
##  sex          length         diameter          height        whole_weight
##  F:1307   Min.   : 15.0   Min.   : 11.00   Min.   :  0.00   Min.   :  0.4
##  I:1342   1st Qu.: 90.0   1st Qu.: 70.00   1st Qu.: 23.00   1st Qu.: 88.3
##  M:1527   Median :109.0   Median : 85.00   Median : 28.00   Median :159.9
##           Mean   :104.8   Mean   : 81.58   Mean   : 27.91   Mean   :165.8
##           3rd Qu.:123.0   3rd Qu.: 96.00   3rd Qu.: 33.00   3rd Qu.:230.7
##           Max.   :163.0   Max.   :130.00   Max.   :226.00   Max.   :565.1
##  shucked_weight   viscera_weight    shell_weight        rings
##  Min.   :  0.20   Min.   :  0.10   Min.   :  0.30   Min.   : 1.000
##  1st Qu.: 37.20   1st Qu.: 18.68   1st Qu.: 26.00   1st Qu.: 8.000
##  Median : 67.20   Median : 34.20   Median : 46.80   Median : 9.000
##  Mean   : 71.88   Mean   : 36.12   Mean   : 47.77   Mean   : 9.932
##  3rd Qu.:100.40   3rd Qu.: 50.60   3rd Qu.: 65.80   3rd Qu.:11.000
##  Max.   :297.60   Max.   :152.00   Max.   :201.00   Max.   :29.000
```

From this summary, we can deduce that the data points that we have are evenly distributed among the

sex variable. Regarding the dependent variable **rings**, we observe that values range from 1 to 29, which correspond to ages from 2.5 to 30.5 years. Median and mean are relatively close, with a small skew for the dependent variable.

Question 3. Generate a labeled scatterplot of the data. Describe interesting features trends. Does it agree with the biologists' hypothesis?

```
# Correlation and distribution plots
ggpairs(
    abalone,
    aes(color = sex, alpha = 1),
    title = "Scatterplot and correlation for abalone dataset",
    lower = list(combo = wrap("facethist", binwidth = 5)),
    upper = list(continuous = wrap(ggally_cor, size = 2))
) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = -90, vjust = 0.5),
    strip.text.x = element_text(size = 6),
    strip.text.y = element_text(size = 5)
)
```



We recall from the introduction that biologists believe there is a linear dependence between the **height** of abalones and their age. From the last line of the plot, we observe that there is indeed a correlation between large height values and large numbers of rings, however a simple linear dependence does not seem clear.

Secondly, there are very high correlations between the different explicative variables. For instance, there is a correlation of 0.987 between diameter and length, which can make the explanation of our models more difficult.

In addition, the separation between infants and adults is clear in almost all plots. However, distributions of variables for both male and female abalones are very similar. This indicates that the major difference is between infants and adults and might be a better variable for modeling.

Question 1. Write a mathematical formula modelling the several assumptions in the above description. Describe what kind of statistical techniques you are going to use to study these hypothesis (confidence intervals, test, ...).

The simple linear model can be described as follows, where Y is associated to **rings** and X is associated to **height**: $Y = X\beta + \epsilon$. Under the normal error assumption, we have to check if the following properties are indeed verified:

[**P0**] X is of full rank.

[**P1**] Errors are centered: $\forall i = 1..n, \mathbb{E}_\beta(\epsilon_i) = 0$.

Possible assessment:

- The mean curve in the *Residuals vs. Fitted* plot should be close to zero and straight

[**P2**] Errors have homoscedastic variance: $\forall i = 1..n, \text{Var}_\beta(\epsilon_i) = \sigma^2$.

Possible assessments:

- The *Scale-Location* plot shows the repartition of residuals among observations, which should be uniform
- The *Breush-Pagan* test allows to assess the $\mathcal{H}_0$ hypothesis of homoscedasticity, which is rejected if the p-value is smaller than 0.05

In particular, a square or log transformation of the dependent variable Y might improve the model in case the homoscedastic assumption is rejected.

[**P3**] Errors are uncorrelated: $\forall i \neq j, \text{Cov}(\epsilon_i, \epsilon_j) = 0$.

Possible assessments:

- The *auto-correlation* function should not exceed the confidence interval around 0
- The *Durbin-Watson* test allows to assess the $\mathcal{H}_0$ hypothesis of uncorrelation, which is rejected if the p-value is smaller than 0.05

[**P4**] Errors are gaussian: $\forall i = 1..n, \epsilon_i \hookrightarrow \mathbb{N}(0, \sigma^2)$.

Possible assessments:

- The Q-Q plot shows the comparison between the quantiles of the standardized residuals and a true normal distribution, which should be close enough
- The *Shapiro-Wilk* test allows to assess the $\mathcal{H}_0$ hypothesis of gaussianity, which is rejected if the p-value is smaller than 0.05

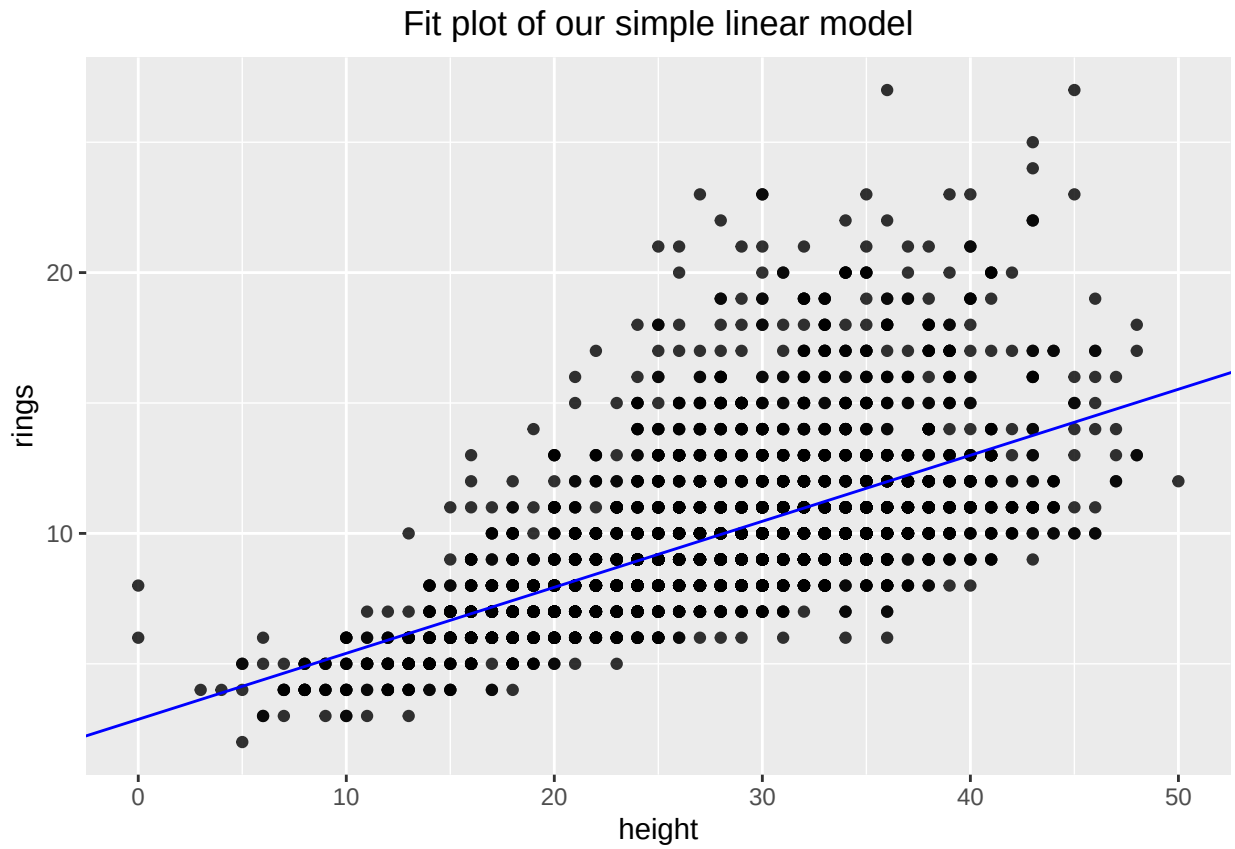
Question 4. Fit a simple linear regression to the data predicting number of rings using height of the abalones.

```
# Simple linear model using only height
lm_height <- lm(rings ~ height, data = abalone_train)
```

Question 5. Generate a labeled scatterplot that displays the data and the estimated regression function line. Describe the line's fit.

```
# Fitted values
ggplot(data = abalone_train) +
ggtitle("Fit plot of our simple linear model") +
geom_point(aes(x = height, y = rings), alpha = 0.8) +
geom_abline(
```

```
    intercept = lm_height$coefficients[1],
    slope = lm_height$coefficients[2],
    colour = "blue"
) +
theme(plot.title = element_text(hjust = 0.5))
```

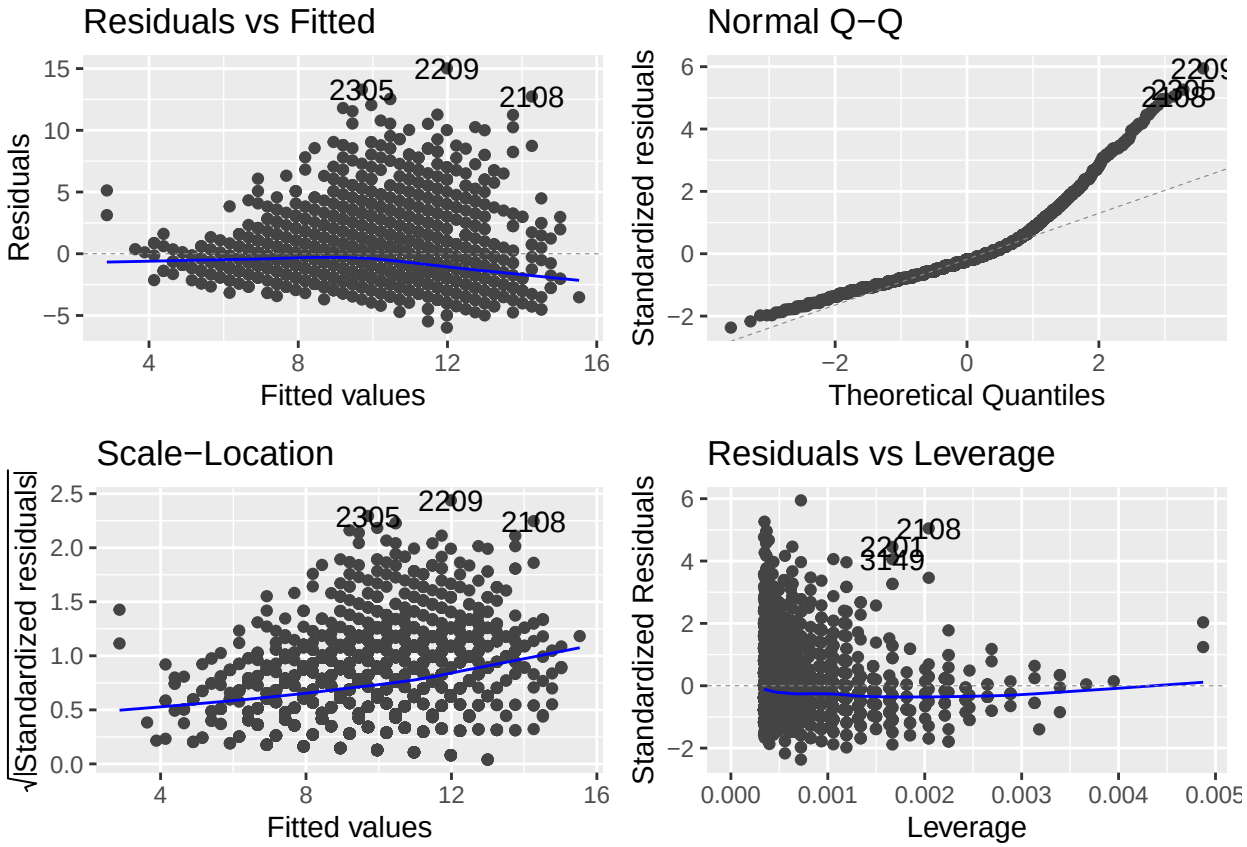


From the fitted line and scatter plot, we can deduce that the linear model might not be sufficient to describe the relationship between **height** and **rings** variables.

Question 6. Do diagnostics to assess whether the model assumptions are met; if not, appropriately transform height and/or number of rings and refit your model. Justify your decisions (and recheck your diagnostics).

First, we assess the initial assumptions for the linear regression of **rings** using **height** only.

```
autoplot(lm_height)
```



[**P0**] Since we are only using one feature, X is undoubtedly of full rank.

[**P1**] It is clear that, for higher values of **rings**, the residuals are negative in average. This shows that the model predicts too high values and indicates for a transformation of the dependent variable.

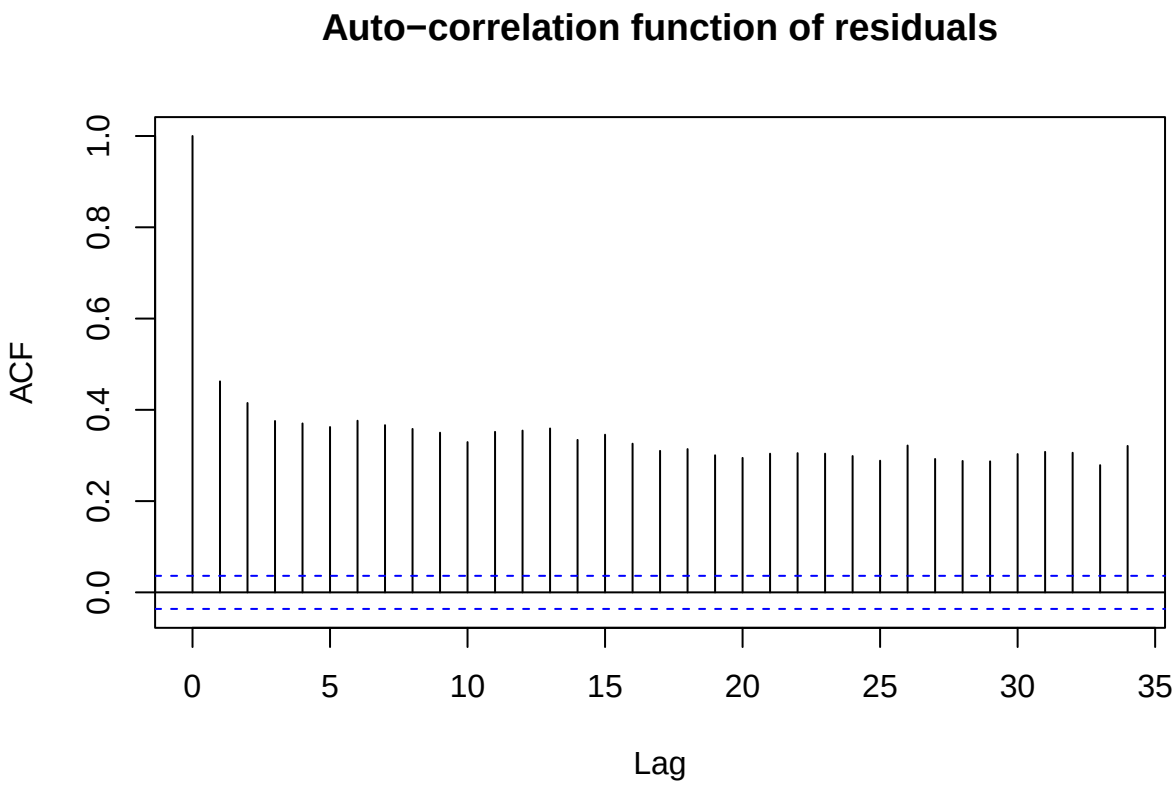
[**P2**] Below, we show the *Breush-Pagan* test's p-value. The p-value is 2.22e-16, therefore we can reject $\mathcal{H}_0$. In addition, we see from the plot that the residuals are clearly not equally spread.

```
# Breush-Pagan test
ncvTest(lm_height)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 237.6428, Df = 1, p = < 2.22e-16
```

[**P3**] Below, we show the $auto - correlation$ plot as well as the *Durbin-Watson* test's p-value. The p-value is negligeable, therefore we can reject $\mathcal{H}_0$. Moreover, the auto-correlation function of residuals is clearly not close to zero.

```
# Auto-correlation function
acf(lm_height$residuals, main = "Auto-correlation function of residuals")
```




```
# Durbin-Watson test
durbinWatsonTest(lm_height)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4623555      1.074311       0
##  Alternative hypothesis: rho != 0
```

[**P4**] Below, we show the *Shapiro-Wilk* test's p-value. The p-value is negligeable, therefore we can reject $\mathcal{H}_0$. In addition, the plot shows that there is a clear deviation from a normal distribution for higher quantiles.

```
# Shapiro-Wilk test
shapiro.test(lm_height$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_height$residuals
## W = 0.8842, p-value < 2.2e-16
```

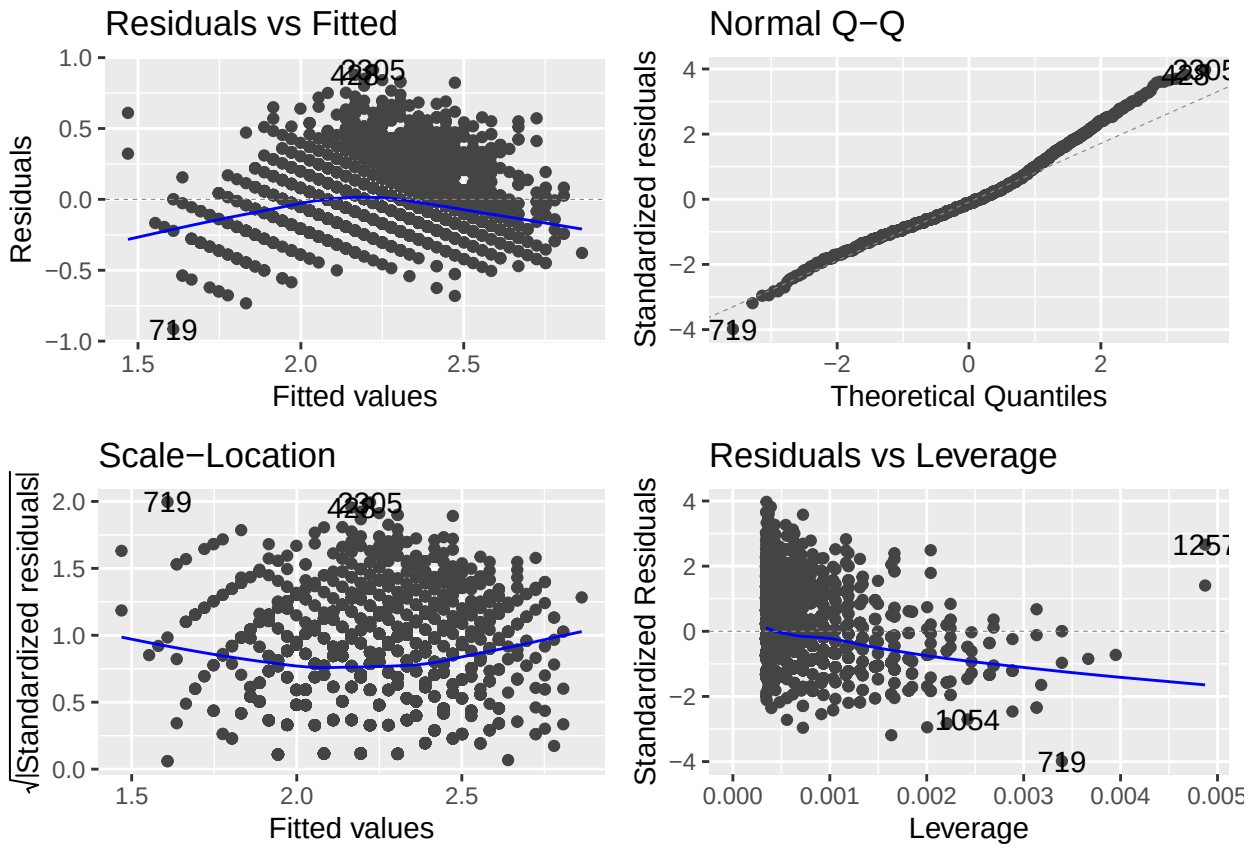
All in all, from our assumptions, only [**P0**] seems to be verified and [**P1**] to a certain extent. All [**P2**], [**P3**] and [**P4**] are not verified. From our observations, a transformation of the dependent variable Y might improve our results. The fact that the predictions made by our model are larger than actual values, we will thus try the following transformation: $\log(Y)$. The model is now: $\log(Y) = X\beta + \epsilon$.

```
# Linear model using only height w. log transformation
abalone_train$log_rings <- log(abalone_train$rings)
abalone_test$log_rings <- log(abalone_test$rings)

lm_log_height <- lm(log_rings ~ height, data = abalone_train)
```

Finally, we perform our tests again, with our modified model.

```
autoplot(lm_log_height)
```



[**P0**] X is still of full rank.

[**P1**] The results are not perfect, with an inversed U-shape. This might indicate a relationship with square of **height**.

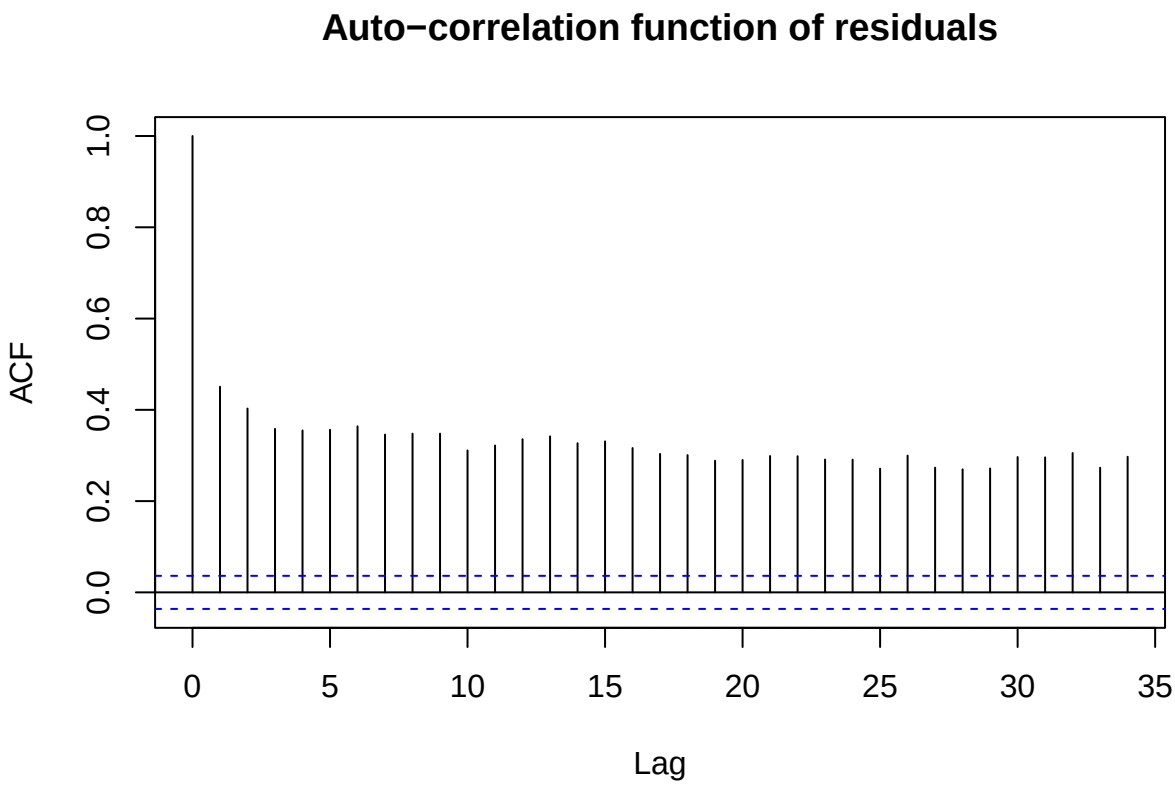
[**P2**] Below, we show the *Breush-Pagan* test's p-value. The plot is much better than for the simple linear model, with a red curve close to 1. Also, the p-value is 0.72724, which is larger than 0.05 and $\mathcal{H}_0$ cannot be rejected.

```
# Breush-Pagan test
ncvTest(lm_log_height)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1216648, Df = 1, p = 0.72724
```

[**P3**] Below, we show the $auto - correlation$ plot as well as the *Durbin-Watson* test's p-value. The auto-correlation is very similar to the simple linear model.

```
# Auto-correlation function
acf(lm_log_height$residuals, main = "Auto-correlation function of residuals")
```




```
# Durbin-Watson test
durbinWatsonTest(lm_log_height)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4506149      1.097652       0
##  Alternative hypothesis: rho != 0
```

[**P4**] Below, we show the *Shapiro-Wilk* test's p-value. The plot is much closer to the normal distribution with our log transformation of the dependent variable. However, the p-value of the test is still smaller than 0.05.

```
# Shapiro-Wilk test
shapiro.test(lm_log_height$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_log_height$residuals
## W = 0.97447, p-value < 2.2e-16
```

Therefore, we see that [**P2**] and [**P4**] are now verified, at least to a certain extent. With our new model, [**P3**] is still not verified.

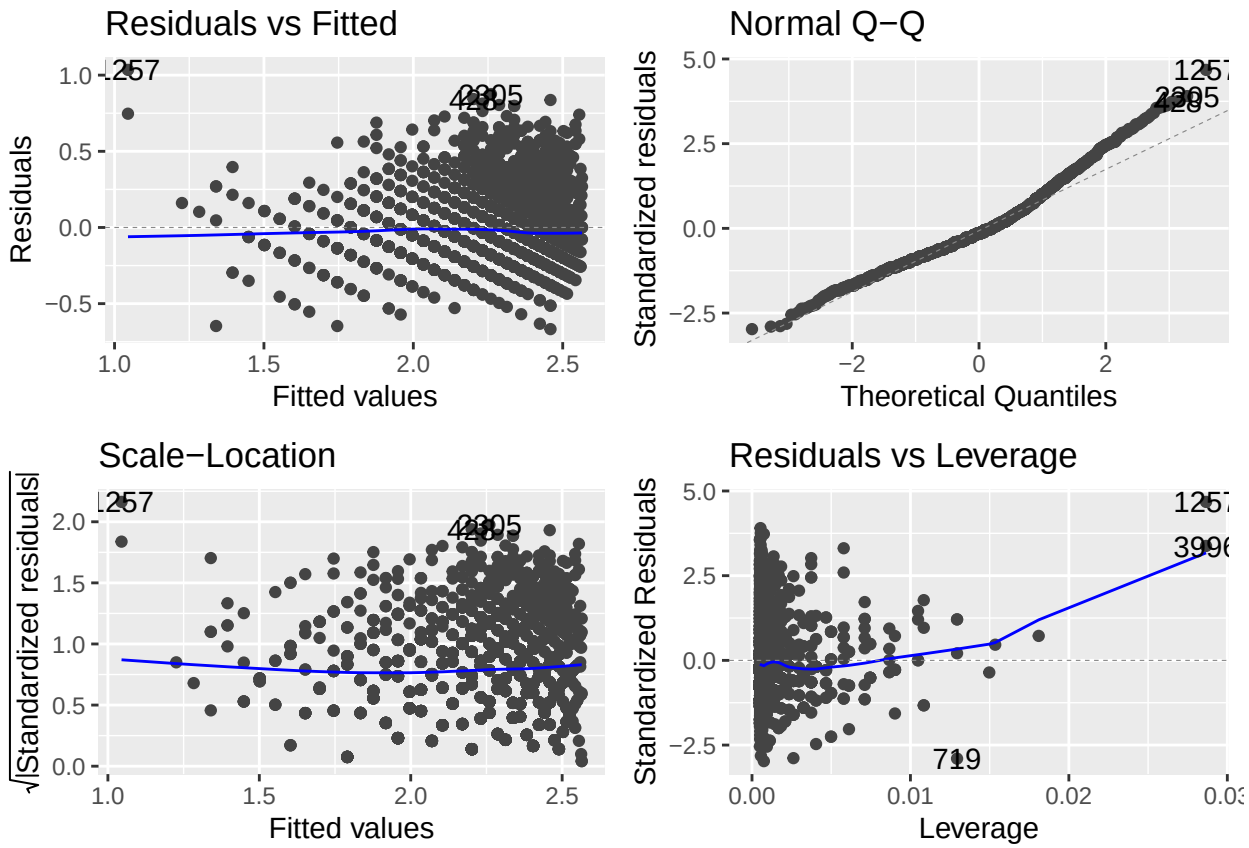
Question 7. Interpret your final parameter estimates in context of the problem. Is there a statistically significant relationship between the height and the number of rings (and hence, the age) of abalones?

We will now try out the following model: $\log(Y) = X\beta + \epsilon$ where X contains the square of **height** plus **height**.

```
# Linear model using only height w. log transformation and square of height
abalone_train$height2 <- abalone_train$height**2
abalone_test$height2 <- abalone_test$height**2

lm_log_height2 <- lm(log_rings ~ height + height2, data = abalone_train)
```

```
autoplot(lm_log_height2)
```



[**P0**] X is still of full rank.

[**P1**] The expectation of residuals is getting closer to zero again.

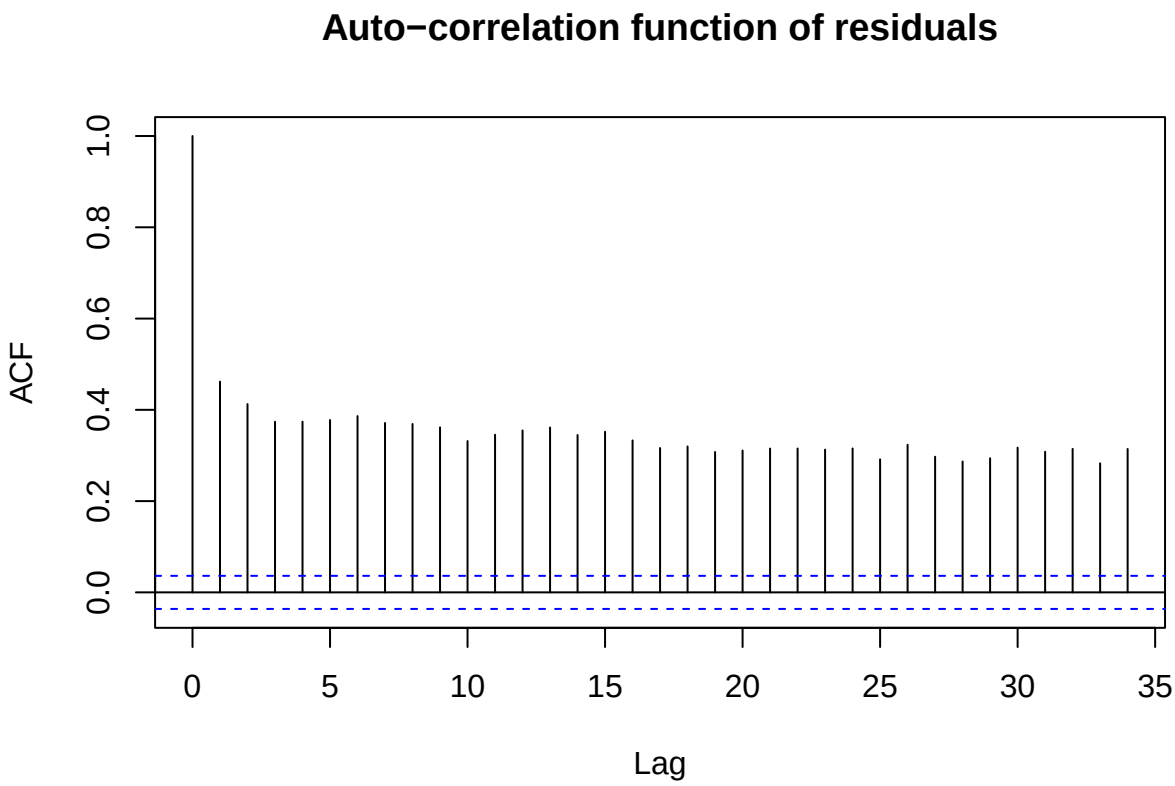
[**P2**] Below, we show the *Breush-Pagan* test's p-value. Results are slightly better with this model.

```
# Breush-Pagan test
ncvTest(lm_log_height2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1200643, Df = 1, p = 0.72896
```

[**P3**] Below, we show the *auto − correlation* plot as well as the *Durbin-Watson* test's p-value. Results are very similar to the model containing only the **height**.

```
# Auto-correlation function
acf(lm_log_height2$residuals, main = "Auto-correlation function of residuals")
```




```
# Durbin-Watson test
durbinWatsonTest(lm_log_height2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4618527      1.075562       0
##  Alternative hypothesis: rho != 0
```

[**P4**] Below, we show the *Shapiro-Wilk* test's p-value. Again, results are very similar here.

```
# Shapiro-Wilk test
shapiro.test(lm_log_height2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_log_height2$residuals
## W = 0.9679, p-value < 2.2e-16
```

Regarding the relationship between **rings** and **height**, our final model will be between the logarithm of **rings** and both **height** and squared **height**.

```
summary(lm_log_height2)
```

```
##
## Call:
## lm(formula = log_rings ~ height + height2, data = abalone_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66726 -0.15190 -0.03412  0.12168  1.03447
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0449720  0.0379841   27.51   <2e-16 ***
## height       0.0621415  0.0028466   21.83   <2e-16 ***
## height2     -0.0006351  0.0000518  -12.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2244 on 2921 degrees of freedom
## Multiple R-squared:  0.4898, Adjusted R-squared:  0.4894
## F-statistic:  1402 on 2 and 2921 DF,  p-value: < 2.2e-16
```

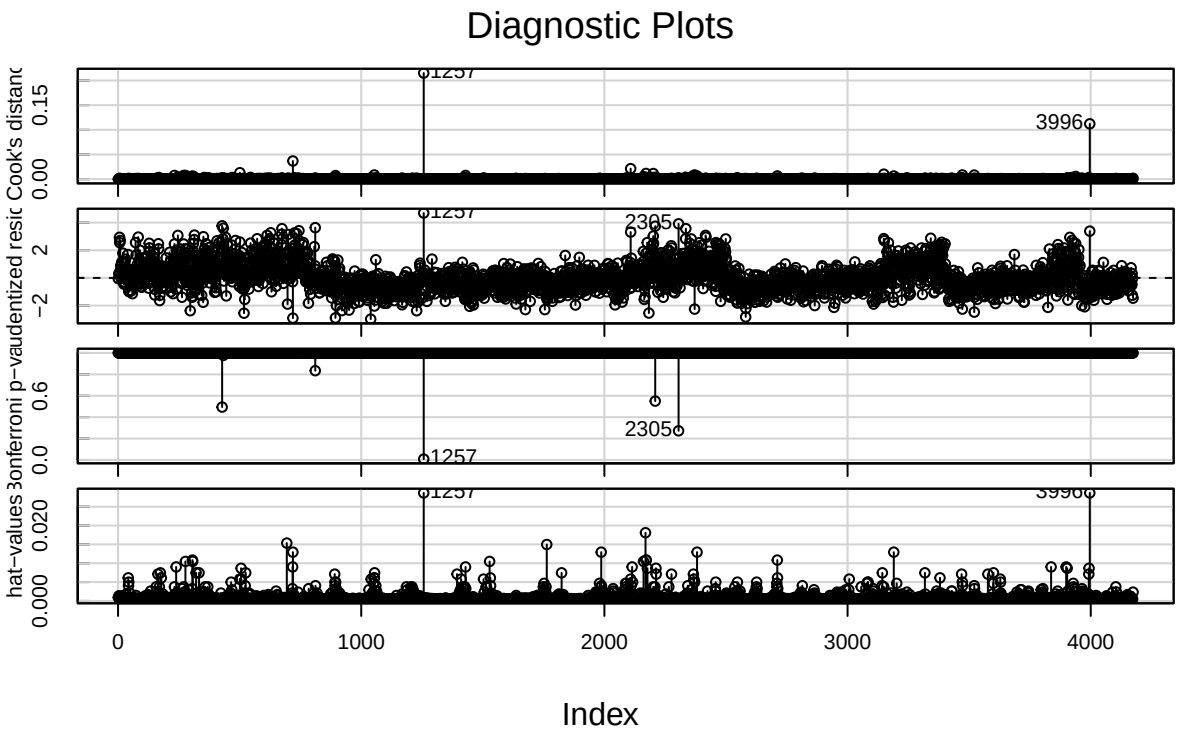
From the summary of our linear model, we observe thatt the p-value associated to the intercept, **height** and squared **height** are close to zero, thus we can say that there is a statistically significant relationship between **log(rings)** and **height**, squared **height**.

Question 8. Consider now all variables. Look at the scatterplot of the data (`GGally:ggpairs`). Look for correlations between predictors. Select some additional variables to add to the simple linear model in order to better predict number of rings. Justify your choices (keep in mind that we want a practical method to predict number of rings). Perform a multiple linear regression. Check the validity of the model. If validity conditions are not met, transform some variables, add/delete some variables, check for outliers and recheck until you find an acceptable model.

Regarding the other variables, we see that **diameter** is very correlated with **length**. Therefore, we will only consider **length**. Likewise, all the weights variables are very correlated, we will only consider **whole_weight** for now. We will also add a new variable **adulthood** that will be used in replacement of **sex**.

Now, let's have a look at possible outliers. The influence plots allow us to detect influential points. More specifically, Cook's distance on the first graph highlights which observations are more likely to influence the values of β. Then, the two next plots evaluate outliers. Finally, the last graph also indicates which points have a high leverage. The studentized and Bonferroni plots can be completed with a test of hypothesis $\mathcal{H}_0$ that the observation is not an outlier.

```
influenceIndexPlot(lm_log_height2)
```



```
outlierTest(lm_log_height2)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 1257 4.694671         2.7935e-06    0.0081682
```

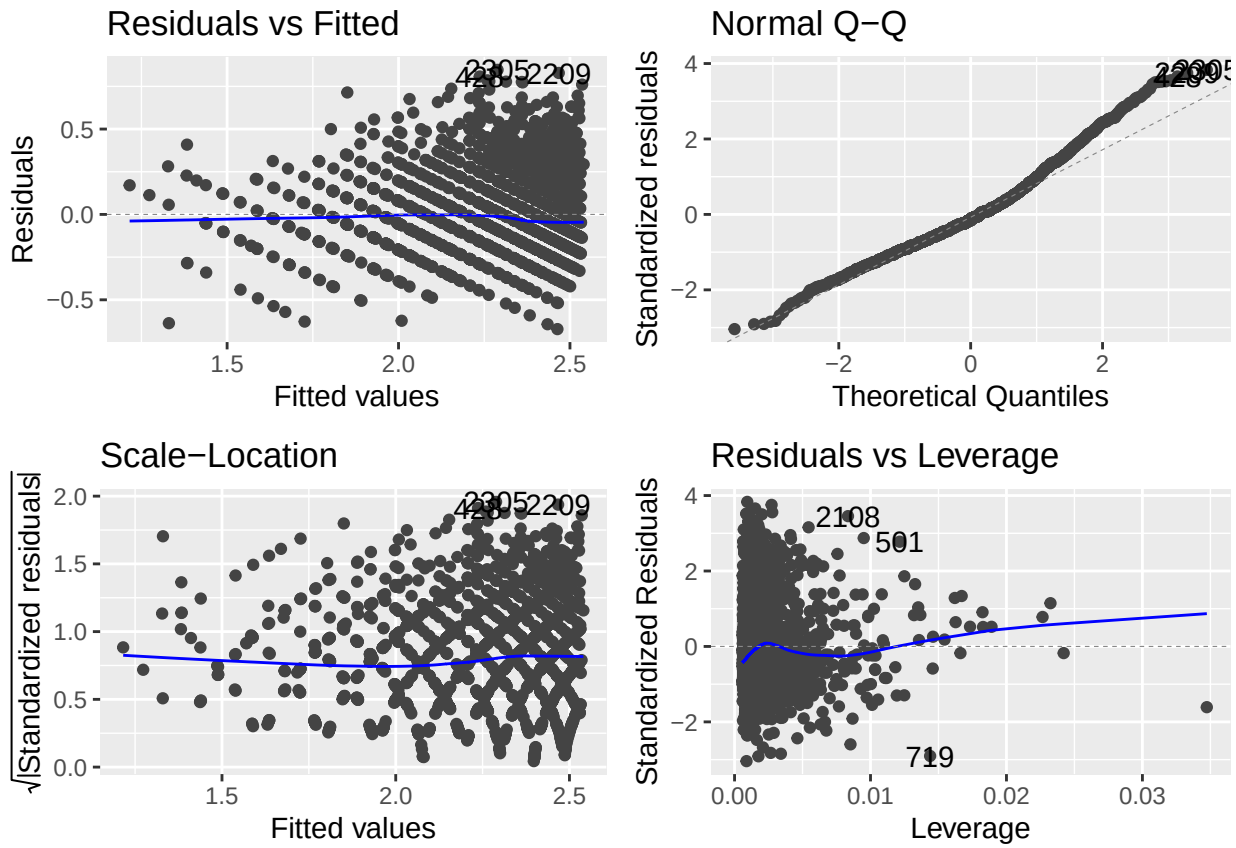
The influence plots help us detect outliers. As we could observe from the *Residuals vs. Leverage* plot of our last model, there are two points which can be pointed out as outliers. The points that we decide to remove from our study will be points number 1257 and 3996.

```
# Multiple linear regression model
abalone_train$adulthood <- 1
abalone_train$adulthood[abalone_train$sex == "I"] <- 0
abalone_test$adulthood <- 1
abalone_test$adulthood[abalone_test$sex == "I"] <- 0

# Remove outliers
abalone_train$cooksd <- cooks.distance(lm_log_height2)
abalone_train_out <- subset(abalone_train, cooksd < 0.1)

lm_multi <- lm(
    log_rings ~ height + height2 + adulthood + length + whole_weight,
    data = abalone_train_out
)
```

```
autoplot(lm_multi)
```



[**P0**] X is still of full rank given the success of the algorithm.

[**P1**]

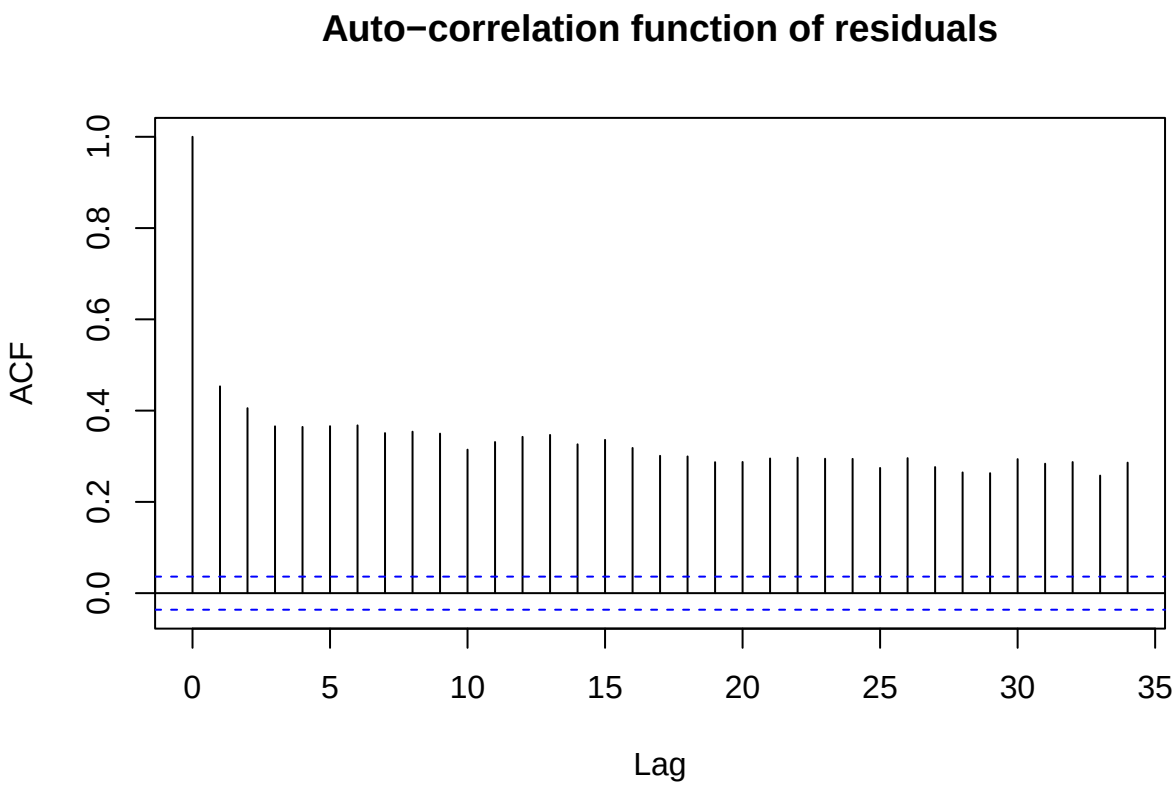
[**P2**] Below, we show the *Breush-Pagan* test's p-value.

```
# Breush-Pagan test
ncvTest(lm_multi)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 13.6673, Df = 1, p = 0.00021822
```

[**P3**] Below, we show the $auto-correlation$ plot as well as the *Durbin-Watson* test's p-value.

```
# Auto-correlation function
acf(lm_multi$residuals, main = "Auto-correlation function of residuals")
```




```
# Durbin-Watson test
durbinWatsonTest(lm_multi)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4531655      1.092943       0
##  Alternative hypothesis: rho != 0
```

[**P4**] Below, we show the *Shapiro-Wilk* test's p-value.

```
# Shapiro-Wilk test
shapiro.test(lm_multi$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_multi$residuals
## W = 0.9735, p-value < 2.2e-16
```