Report on Abalone Data Set

Saikat Beta, Presidency University, KOLKATA

December 2, 2022

Abstract

Abalones, also called ear-shells or sea ears, are one type of reef-dwelling marine snails. The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked in a variety of cultures. It is difficult to tell the ages of abalones because their shell sizes 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. A research group are interested in using some of abalones' physical measurements, especially the height measurement to predict their ages. The research group 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, and particularly, that a larger height is associated with an older age. Our data set is abalone.csv.

The two variables in the dataset (abalone.csv) are Height (in mm) and the number of Rings.

Research Problem :

A research group are interested in using some of abalones' physical measurements, especially the height measurement to predict their ages.

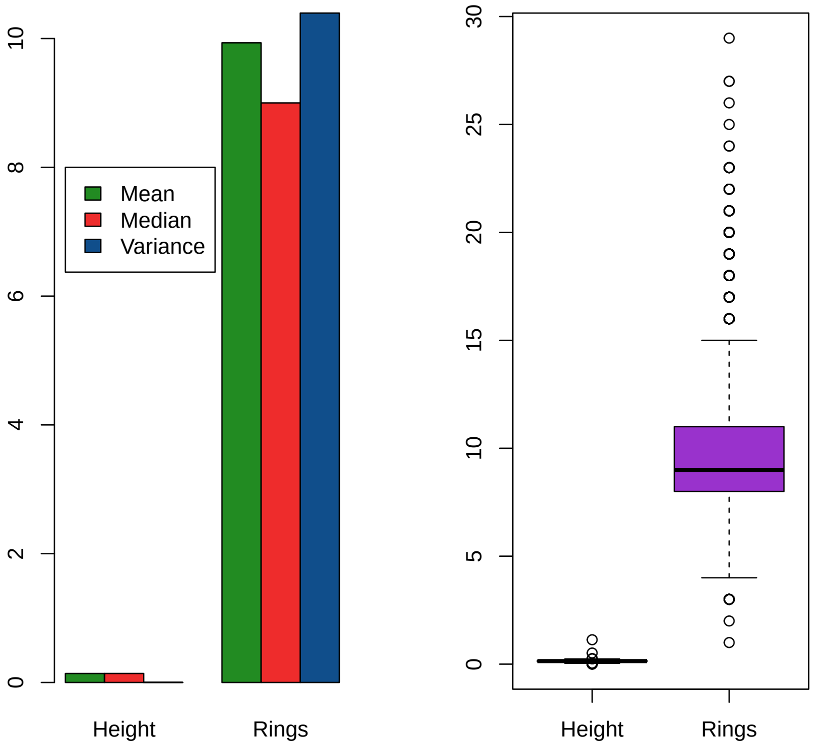
Research Hypothesis :

A simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages, and particularly, that a larger height is associated with an older age.

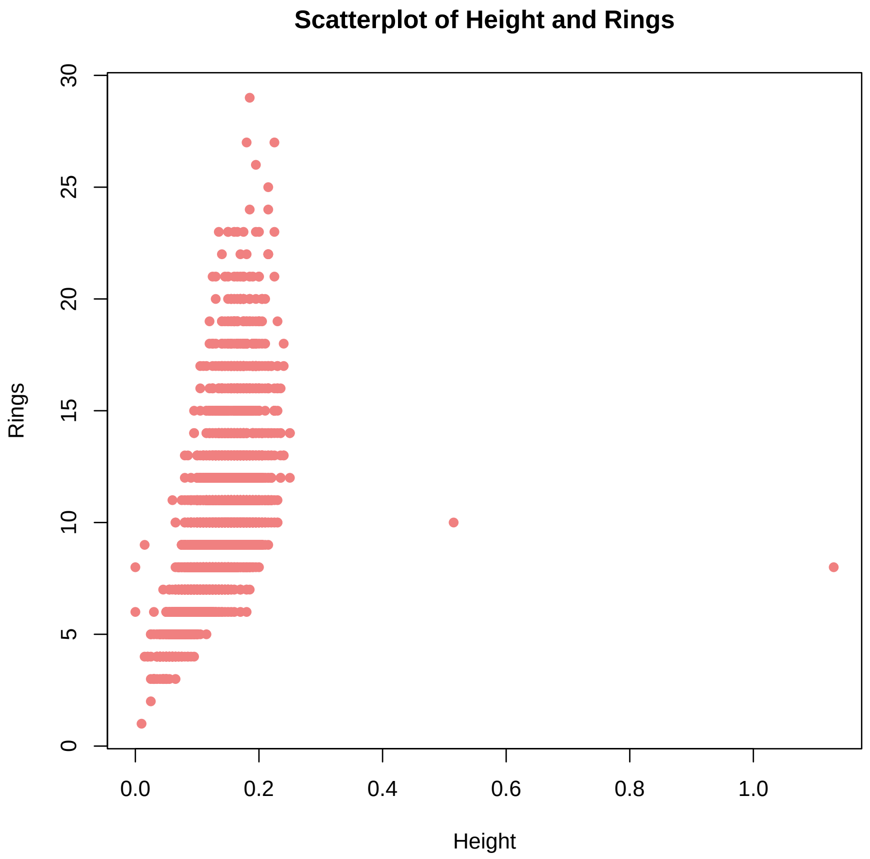
We first load the dataset and examine the two variables individually:

```
##   Height Rings
## 1  0.095    15
## 2  0.090     7
## 3  0.135     9
## 4  0.125    10
## 5  0.080     7
## 6  0.095     8
##      Height           Rings
##  Min.   :0.0000   Min.   : 1.000
##  1st Qu.:0.1150   1st Qu.: 8.000
##  Median :0.1400   Median : 9.000
##  Mean   :0.1395   Mean   : 9.934
##  3rd Qu.:0.1650   3rd Qu.:11.000
##  Max.   :1.1300   Max.   :29.000
```

We then plot the values in graphs:



The unit of height is mm.
We plot the scatterplots of the given data:



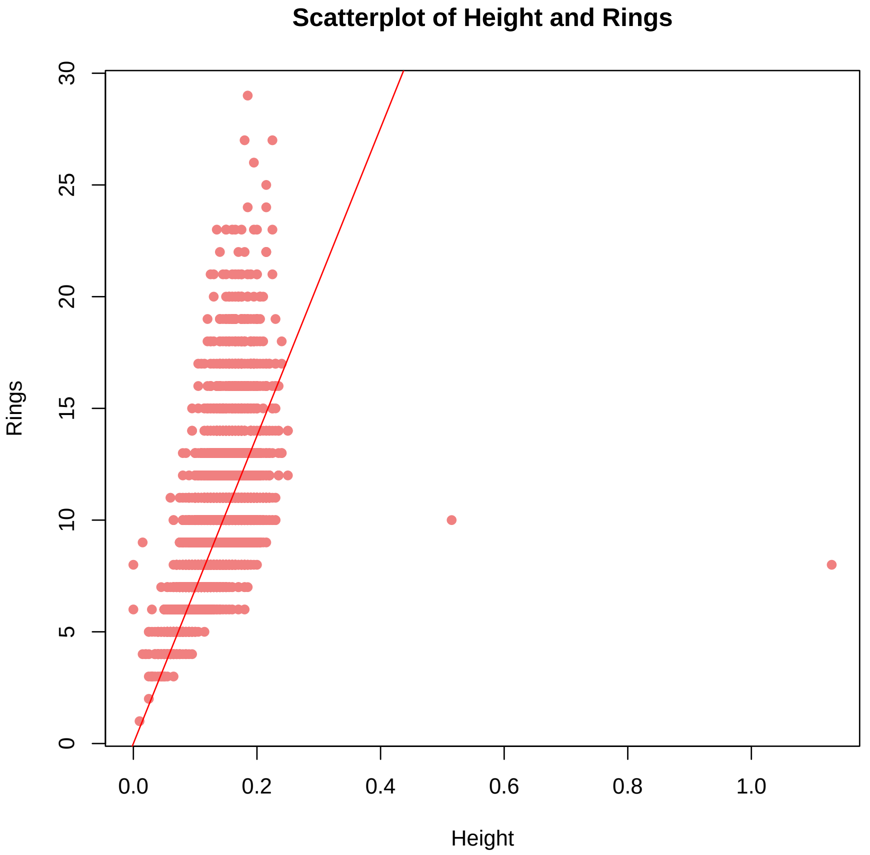
From the scatterplot, we find that the heights are concentrated in the range 0 − 0.2mm and that with a slight shift in the height, the number of rings can change. Moreover, there are some outliers for the heights. Also, as the number of rings increases, there are fewer data for the corresponding rings.

We estimate a linear regression model to fit with the datset:

```
##
## Call:
## lm(formula = y ~ x - 1)
##
## Coefficients:
##     x
## 68.87
```

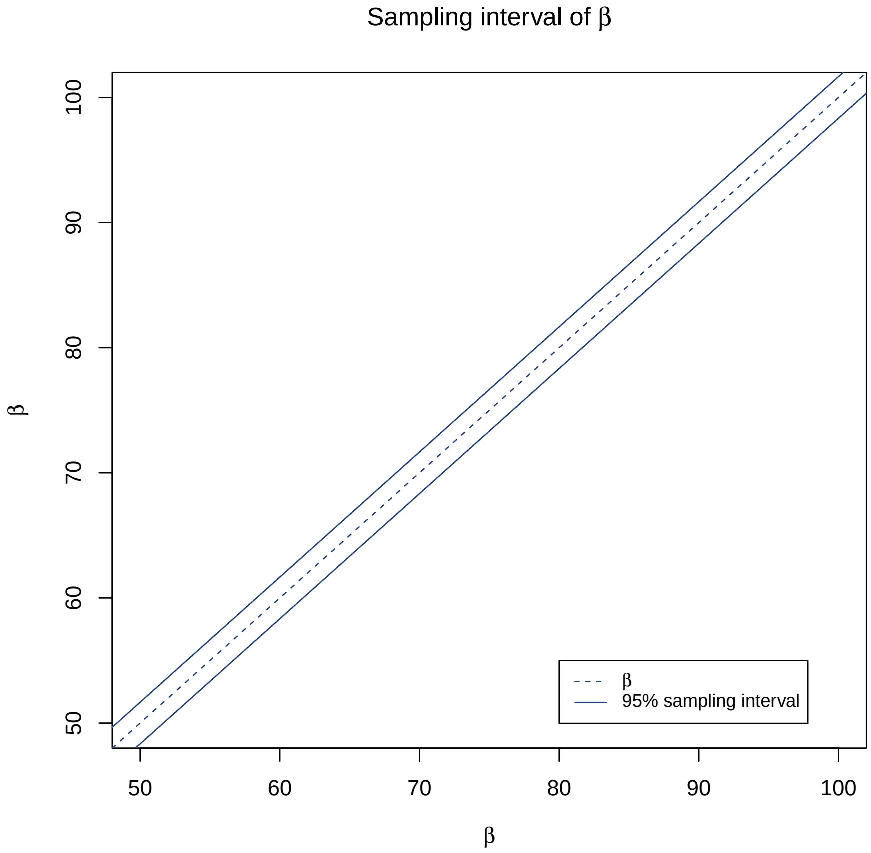
Here, the intercept is forcibly taken to be 0, as it's meaningless to measure the number of rings of an abalone which has it's height as 0mm.

We fit a linear regression model to the scatterplot:



The line does not fit nicely to the given scatterplot, especially as the height is increased, the number of rings, as predicted by the regressed line does not match with the given data.

In the regressed line, there is only one parameter which can be interpreted as if we choose two sets of heights of abalone which differ by 1mm, then, we can expect the difference in the number of rings of the abalones in the two sets to be 68.87 $\approx$ 69 on an average.



The 95% confidence interval gives us the margin of error for the true linear relationship between the height and the number of rings of the abalones. There is no intercept, and thus, there is no question of confidence interval of the intercepts.

We test a null hypothesis that the relationship is not statistically significant against the possible alternative that it is significant. Thus, if we reject our null hypothesis, then we would be accepting the fact that the relationship is statistically significant.

```
##     x
## FALSE
```

We find, from the sample data, that the value of the relationship between the height and the number of rings of abalones lies in the critical region of the the test. Thus, with $\alpha = 0.05$, we reject the null hypothesis that there is no statistically significant relationship between the height and the number of rings of abalones.

We want to find the point estimate and a 95% confidence interval when the height is given to be 0.128mm.

```
##        fit      lwr      upr
## 1 8.815786 8.738319 8.893253
```

We find the point estimate of the number of rings of an abalone to be 8.815786 $\approx$ 9. This implies that on an average, if we select an abalone to be of height 0.128mm, we can expect it's number of rings to be close to 9 on an average. We also find the 95% confidence interval of the prediction to be [8.738319, 8.893253].

```
##        fit     lwr      upr
## 1 9.091279 1.60215 16.58041
```

We are interested in predicting the number of rings for an abalone with height at 0.132. We find the predicted value to be 9.091279 and the 99% prediction interval to be [1.60215, 16.58041].

Finally, we conclude the fact that a linear regression model with a normal error function does not fit the data. However, we found that the prediction interval is too wide which implies that the margin of error while predicting with the help of the given model is large. This is a direct consequence of the fact that the linear model is not a good fit for the dataset. Further, looking at the scatterplot, it seems that a quadratic model would be a better fit than a linear model for this data.

Reference:

https://archive.ics.uci.edu/ml/datasets/Abalone