Q1: Data Analysis Report

Introduction

As one of the most frequently consumed marine snails, the size of *abalone* varies from small to very large (i.e. from 20mm to 200mm). In determination of the age (in years) of abalones, cutting the shell through the cone is included, in order to count the number of rings on the shell through a microscope. It is believed that there exists a correlation between rings and age of the abalone, from the given number of rings plus 1.5 years will be the age. However, due to the time-consuming process and inconvenience of actual measuring number of rings, it is suggested to use other measurement in order to make the process easier. In general, other physical appearance including length, sex, height and diameter are all attribute that can be used. In this case, we are going to build a linear regression model of abalone's age from its height (in mm).

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

Summary of individual data set & Directed modeling

Table 1 gives the basic information of height and number of rings of 4177 abalones studied by *Marine Research Laboratories-Taroona.* By formula [Q1-1.5*IQR, Q3+1.5*IQR], we have most majority of observed height falls into range [0.04, 0.24mm]; observed number of rings falls into range [3.5, 15.5]. Any other observed value not in this range is considered to be outliers.

The fully marginal distribution of both height and number of rings are presented in Figure 2 and Figure 3.

Table 1. Summary of Height& Rings

	Min	Max	Mean	Median	Variance	1st Quantile	3rd Quantile	IQR
Rings	1.0000	29.000	9.9340	9.0000	10.395	8.0000	11.000	3.0000
Height (mm)	0.0000	1.1300	0.1395	0.1400	0.0017	0.1150	0.1650	0.0150

Figure 2. Histogram of Height

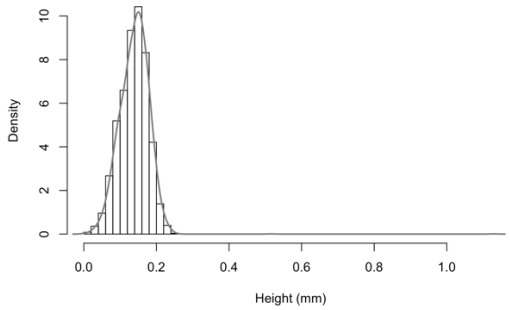
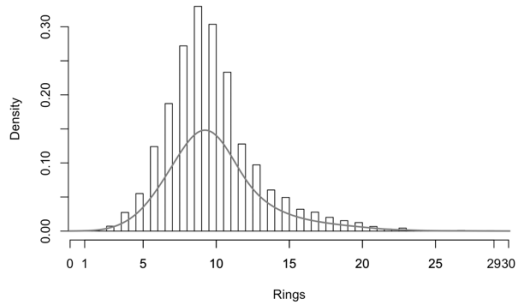


Figure 3. Histogram of number of Rings



The direct linear regression model between abalone height and counted rings is showed in Figure 4, indicates that there exists positive correlation between these two variables. However, as showed in Figure 5, under the effect of two extreme outliers, the model built on this is mis-specified.

Figure 4. scatterplot and direct LR model

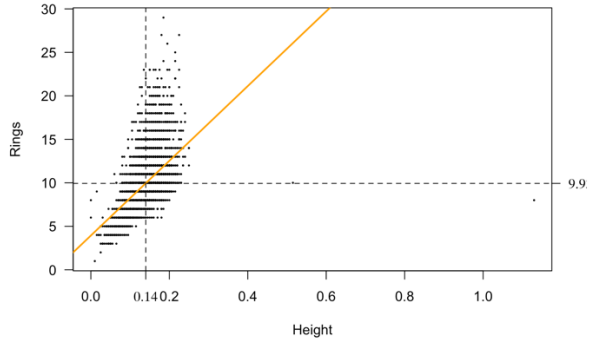
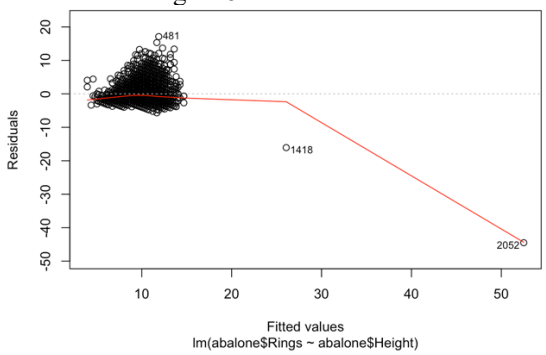


Figure 5. Residuals VS Fitted



Diagnostic modeling

Given the initial raw dataset, both of the outliers in height and ring counted have significant influence on the accuracy of the model we built. In order to maintain a better fitted linear regression model, we modify the model by omitting all the outliers from initial dataset (i.e. for height outside from [0.04, 0.24mm] as well as ring counted outside from [3.5, 15.5]) and just focusing on the vast majority data. Figure 6-9 show the diagnostic modeling fit (i.e. final model fit) and related plots.

Figure 6. Diagnostic modeling

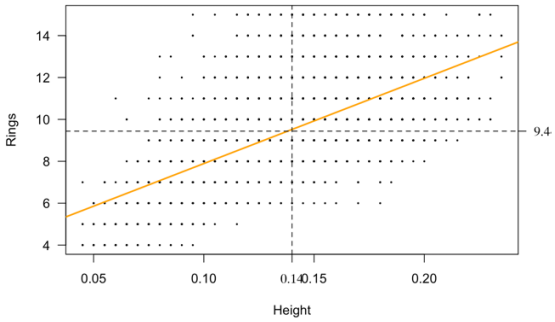


Figure 7. Diagnostic residual VS. fitted

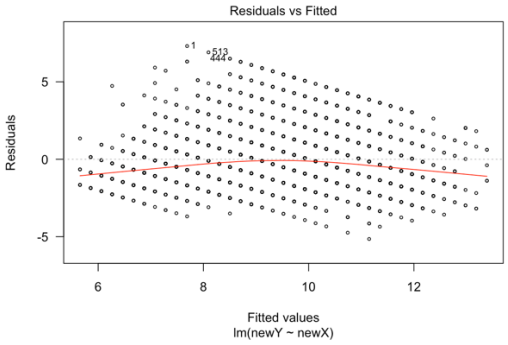


Figure 8. Diagnostic residual distribution

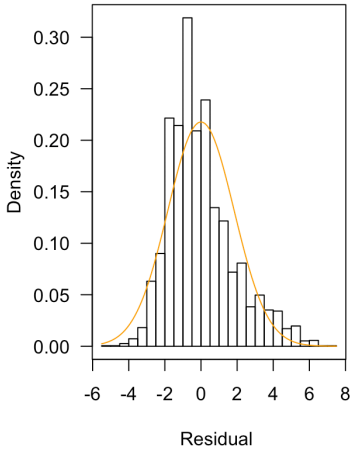
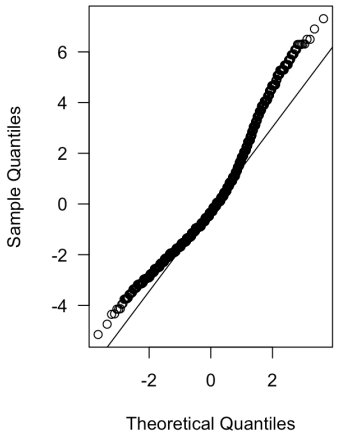


Figure 9. Normal Q-Q plot



Inference and results

Table 2 shows the estimated parameter and it's corresponding confidence interval at a level of α=0.05. We notice that since our estimated slope β1_hat=40.7, it is considered that an increase of 0.1mm in height is associated with 0.4 additional of ring counted. Observed we have P-value is less than 2e-16, indicates that there exist a significant relationship between height and ring counted.

Table 2. Parameter estimate and 95% C.I of diagnostic modeling

	Estimate	Lower C.I.	Upper C.I.
Intercept	3.8235	3.5969	4.050157
Slope	40.6995	39.1111	42.287887

Table 3 shows the two predicted C.I with level of 95% at height=0.128 (mm) and 99% at height=0.132 (mm) respectively. During the process of diagnostic modeling, since we omitted the outliers outside from the quantile intervals, we have the two C.I.s pretty close to each other, which indicates a more precise interval for the future prediction.

Table 3. Fitted value and corresponding C.I. at level 95% and 99%

Height (mm)	Sig level	Fitted value	Lower C.I.	Upper C.I.
0.128	95%	9.195858	9.118885	9.272831
0.132	99%	9.195858	9.137299	9.254416

Figure 10 shows the detailed information about the diagnostic modeling. We see that with df=3862, β0 and β1 follows t-distribution; R^2=0.3952 and adjusted R^2=0.3951, which means this diagnostic regression model can explain at most 39.5% of Y variance. However, since we obtain pretty low P-value, we still have enough confidence that our model is suitable.

Figure 10. data of diagnostic modeling

```
##
## Call:
## lm(formula = newY ~ newX)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1494 -1.3005 -0.3180  0.8855  7.3100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8235     0.1156   33.08   <2e-16 ***
## newX         40.6995     0.8102   50.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.832 on 3862 degrees of freedom
## Multiple R-squared:  0.3952, Adjusted R-squared:  0.3951
## F-statistic:  2524 on 1 and 3862 DF,  p-value: < 2.2e-16
```

Conclusion

By constructing the simple linear regression model between height of abalone's and ring counted, we conclude that there exists a linear relationship between these two variables. Simply in the form: Y=40.6995X+3.8235, where Y represent height in mm and X represent ring counted. By observation of p-values of estimated parameters of β0 and β1, we say that the model we construct after omitting the outlier is suitable. However, with only ~40% variance of Y we can capture which is not high enough, there is quite big space we can improve our model in order to fit all the requirements. Instead of simple linear regression model, researchers could construct multiple linear regression model by including abalones' other physical attributes, as predictor of rings counted.