Model Design, Data Analysis and Estimate for

Research of the Ages of the Abalone Shells

Yunke Zhou

Purdue University – Department of Statistics

Master of Applied Statistics Exam Spring 2020 Question 2

**Part 1. Introduction**

* 1. **Abstract**

In this study, the client offered a data set of the abalone shells, including the sex length, height, diameter and the weights of different parts of the shell. The researchers also recorded the number of rings, from which we can get the actual age of the shell.

The current method to check the age of the shells is cutting them through the cone, staining them and counting the number of rings with a microscope, which is a time-consuming task. The client want to achieve a model to get the ages from the other easily measurable factors of the shell.

**1.2 Introduction of the Study**

The following table includes the reported physical measurements of the abalone shells that the client offered. Please note that except for the sex, all the other variables other than rings are continuous. The study will use these data to check if the ages can be predicted with precision. According to the client’s claim, (Rings + 1.5) is exactly the ages we want, so we can either change its data to the number of years or directly predict the rings. In this report we will use the latter one.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data Type | Measurement Unit | Description |
| Sex | Nominal | - | M, F, I(infant) |
| Length | Continuous | mm | Longest Shell Measurement |
| Diameter | Continuous | mm | Perpendicular to length |
| Height | Continuous | mm | With meat in the shell |
| Whole Weight | Continuous | grams | Whole abalone |
| Shucked Weight | Continuous | grams | Weight of meat |
| Viscera Weight | Continuous | grams | Gut weight (after bleeding) |
| Shell Weight | Continuous | grams | After being dried |
| Rings | Integer | - | +1.5 gives the age in years |

**Table 1.** The Physical Measurements and the Explanation

**1.3 Client’s requirement**

1. Select meaningful factors and give a proper model for the ages of the abalone shells.
2. Analyze the model and give reasons of the model selection.
3. Use the model to simulate the data and check if it agrees with the actual measurement.

**Part 2. Preliminary Data Analysis**

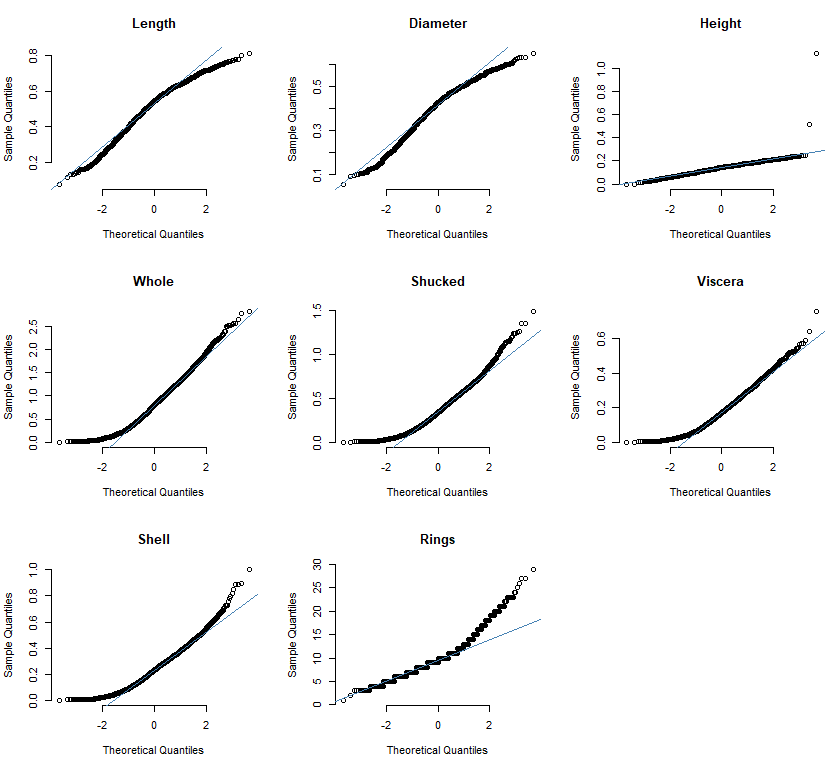
* 1. **Data collection and Potential Issues**

Since all the data sets have been collected by the client, it will not be practical to improve the collecting method. Bias from measuring process might happen during the data collection period, and this might cause a random effect. Besides, the individual shell’s variation in traits will also be a random effect.

Considering that random effects from the measuring process is not reported by the client and we are not sure about its existence and level, we will only consider the shell individuals’ difference as the random effect.

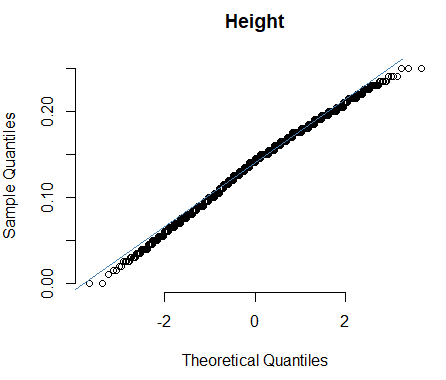
* 1. **The Normality Check of the Data**

The first step is to check the normality of all the factors. The following Q-Q plots give the results of the check.



**Figure 1.** Q-Q Plots of all the factors for the normality check

We can find that there are two obvious outliers, and they could be easily recognized by arranging the factor “Height”. They can be easily ruled out by R, and after that the Q-Q plot changes to the following:



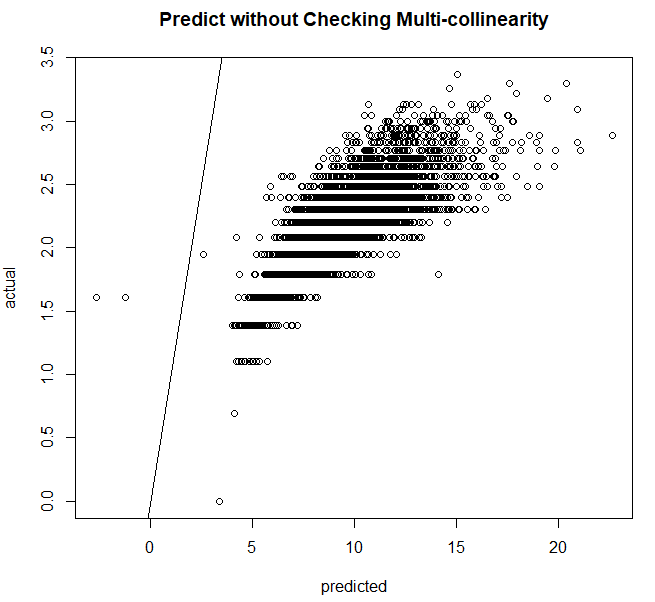
**Figure 2.** Q-Q Plots of all the factors

Now the data of the height factor is highly normally distributed according to the figure 2, and its normality is the best out of all the factors. Thus the height factor should be very helpful for the predicting model. However, we still need the multi-collinearity check between the factors before we finally confirm our model. The next part will explain this in detail

* 1. **Multi-collinearity Check**

The multi-collinearity is the phenomenon that one variable in the regression can easily predict by the other or several others [1]. In another word, large multi-collinearity effect means high correlation between two factors. Usually we should us only one of these highly related variables instead in the regression because the serious multi-collinearity will largely influence the prediction result.

For example, if we ignore the multi-collinearity and use the full model, the final outcome will be like the following graph:



**Figure 2.** Predict results without checking multi-collinearity

We can see that the predict result is bad. In this study, we can tell from common sense that diameter and length could be related, and all the four weight factors could also possibly be highly related. To solve this issue, we will need to use the Variance inflation factor (VIF).

The VIF of each factor can be easily achieved by R, and will be very useful to check to the multi-collinearity effects between them. First, when we check all the factors:



**Figure 3.** VIF of all factors

We usually don’t accept the VIF value above 10, but we also cannot delete all the factors except “Height” at once according to the figure. For this situation, we can delete two variables with largest VIF, and see the result again:



**Figure 4.** VIF of Length, Height, Shucked, Viscera and Shell

The result becomes much more acceptable, but we should still delete the factor “Viscera”:

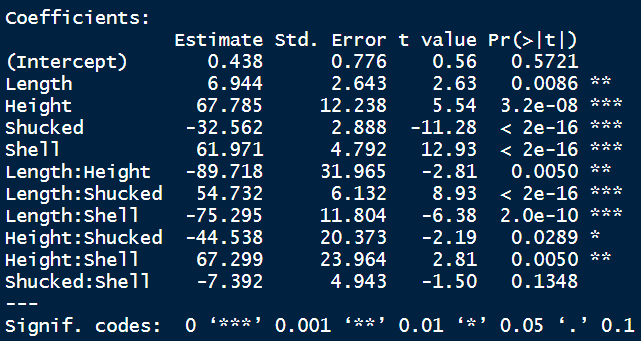


**Figure 5.** VIF of Length, Height, Shucked and Shell

This this an acceptable result for us. We can check the significance of each factor and their interactions, and start building up the model.

* 1. **Significance and Interaction**

We can achieve the significance of each term that might be added to the model in R as follow:



**Figure 6.** The significance of the coefficients

In figure 6, we can see many coefficients with high significance, but their p-value are actually different by orders of magnitude over 10. It will be good to include all the factors with (\*\*\*) and the Length factor since two of the interactions including this factor are extremely significant, although adding this factor might only influence the result with nuance.

Up till now, we have decided the terms we will take into consideration when building the model, and in the next part, we will use the data to check how well the model actually works.

**Part 3. Model**

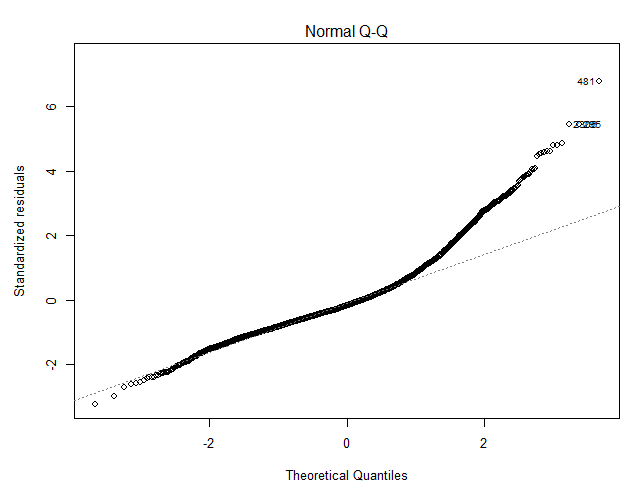
**3.1 Model**

The model of this study will be like the following formula:

**Formula 1.** Final Linear Model

Here represents for the random effect of the shell individuals, and represents for the residual. is the effect of length, and is the effect of interaction between length and shell.

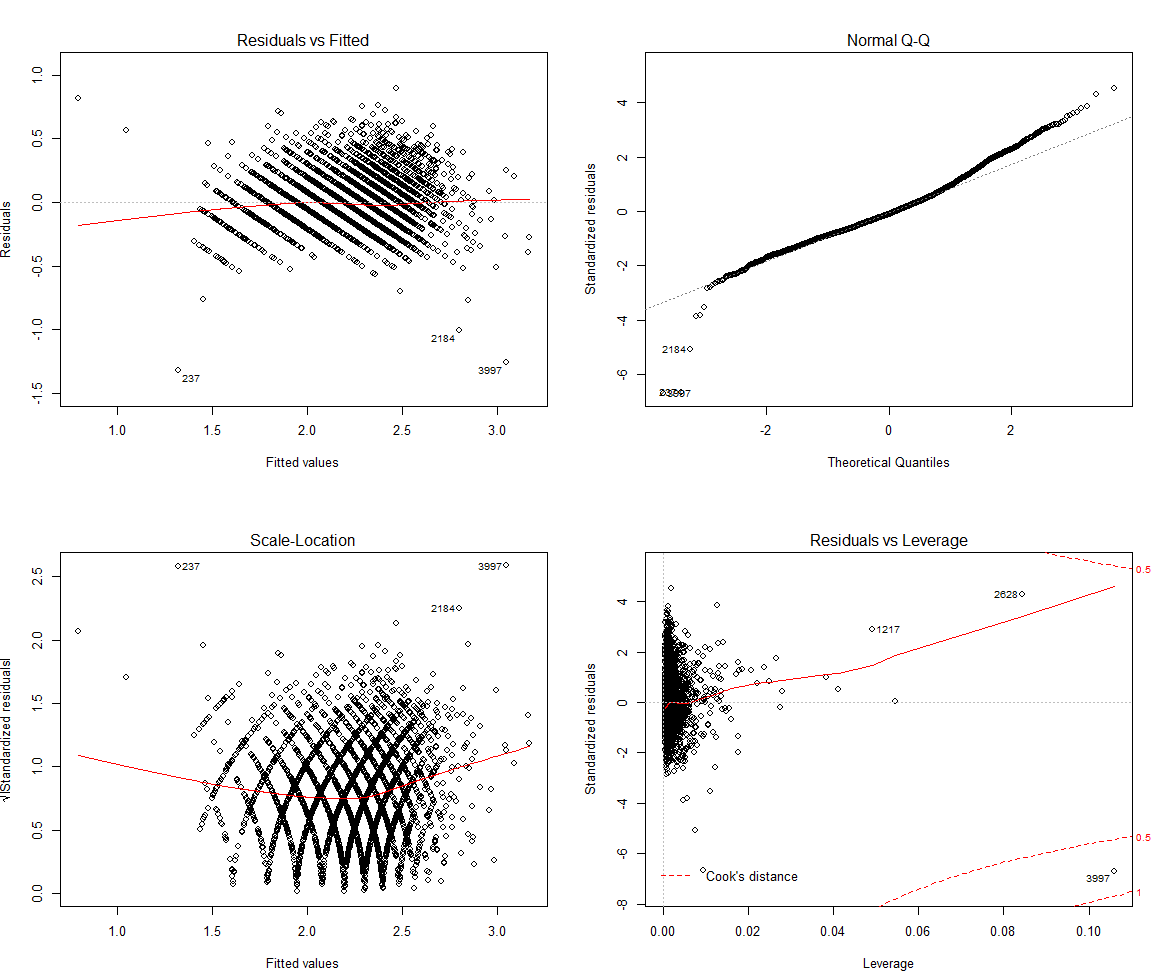
We also need to check the normality check is not satisfying, and whether we should do the log transformation of the model to solve this issue. The Q-Q plot of the model is as follow:



**Figure 7.** Normal Q-Q plot of the model

Obviously there is a deviation at the tail part, which encourages us to do the log transform. The difference after the transform will be shown in the following. Before that, we will analyze the model first.

**3.2 Model Analysis**



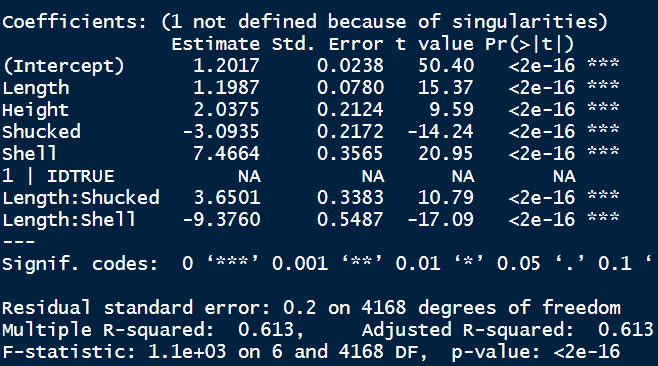
**Figure 8.** Residual, Q-Q, Scale-Location and Residual vs. Leverage Plot

Figure 8 gives a detailed analysis of the model. Each plot offers important information of the model so the explanation will be given by plots.

1. **Residual vs. Fitted plot:** The fitted red line is almost a straight line, which usually appears in a good model. Most residual points are within the [-0.5, 0.5] interval, which is very satisfying.
2. **Normal Q-Q plot:** The Q-Q plot doesn’t have the large deviation on the tail any more, meaning that the log transform is effective. Although we have exclude two outliers in the previous part, there are still few outliers on the tail, but they will not strongly influence the result.
3. **Scale-Location plot:** This plot is used to check the assumption of equal variance. [2] The result is good since the red line is relatively horizontal, and the points are spread randomly along the red line.
4. **Residual vs. Leverage plot:** This plot helps us to find influential outlying cases if any. Not all outliers are influential in linear regression analysis, even though data have extreme values, they might not be influential to determine a regression line. [2] In this plot we can only see very few part of the 0.5 border line, and the 1 border line is almost not visible, which means there is not influential points since not point is out of the cook distance. Please note that if we didn’t delete the points in the previous part, they will be shown as outliers here.

**3.3 Significance Check**

Using R, we can achieve the detailed summary of the model after log transform as following:



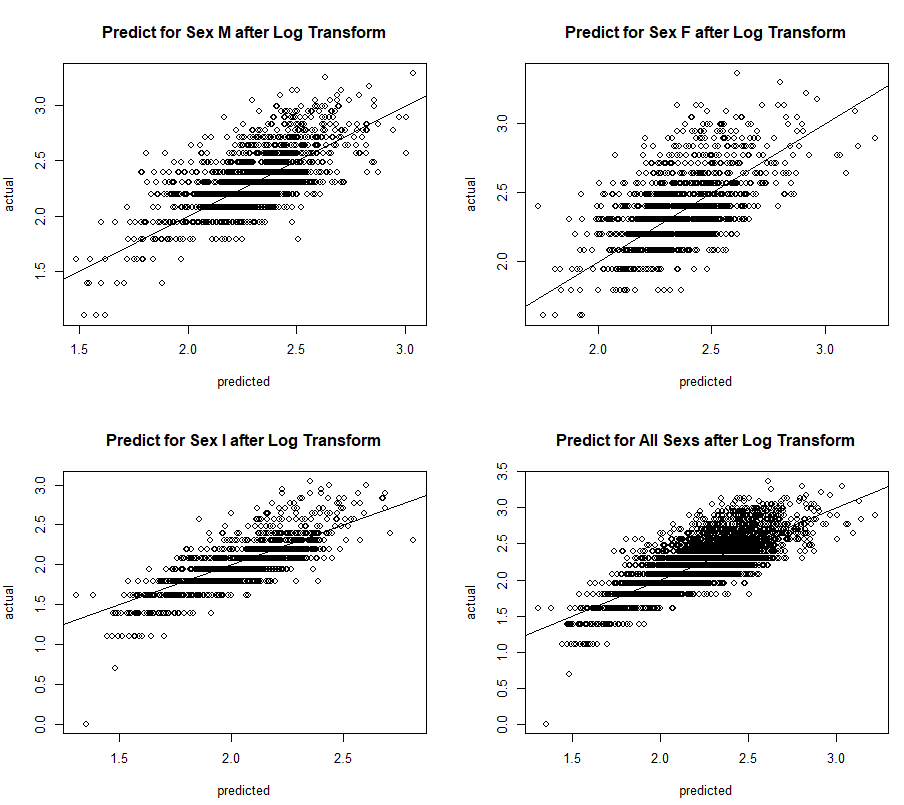
**Figure 7.** Model summary of the model after log transform

**Figure 3.** Use VIF to reduce factors

In the figure we can see that the p-value of the model is very small. If the p-value of a factor is much less than 0.05, we reject the null hypothesis that β = 0. Thus there is a significant relationship between the variables in the linear regression model of the data set faithful. [3] Similarly, we can see that all the factors in the model are very significant. Overall, this model is good enough for us to do the prediction.

**Part 4. Prediction**

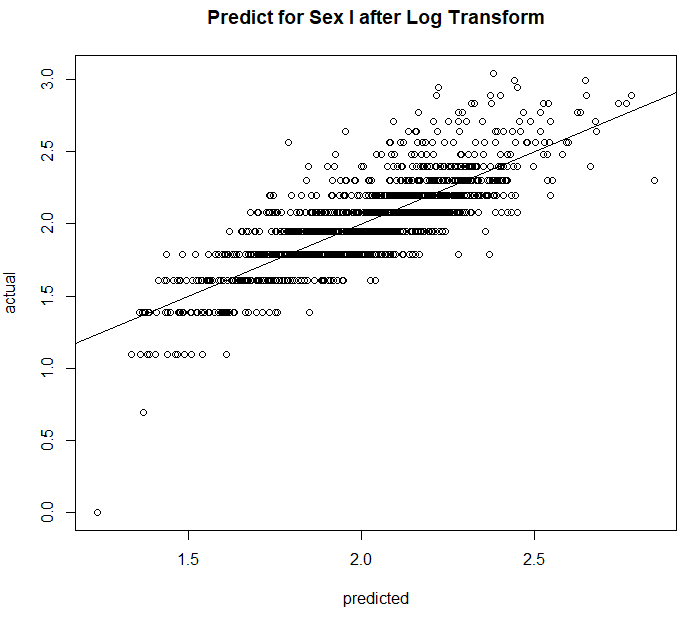
We haven’t introduced the sex factor since it’s a nominal variable and we are not sure how to quantify its influence on the prediction. Thus, in this part the prediction will be four parts: Male, Female, Infants and all. We can compare these different results to get the influence from the Sex.



**Figure 10.** Predict results by sex

From Figure 9, the predict result is good overall. The x-axis is the predicted values and the y-axis is the real value, while the line is the line y = x. The closer the points are distributed along the line, the better the predict results are. Note that the values in the figures are the rings value after log transform. Thus if the client wants values of the year, we can use to get the result.

The predict result of infants seems to have room for improvement and also influence the predict result for all genders. This might due to the difference between the data of infants and others. To solve this, we can build the model again with the infant shell data only, and we can achieve the following results:



**Figure 9.** Predict results of infants after modifying the model

Now the result is better. Through this phenomenon we can see that it’s not reasonable to predict all the genders together. Instead, estimating the number of rings separately is a better idea.

**Part 5. Conclusion**

Generally, we completed all the task the client requested in **Part 1.3**. For the effect of factors, we came to the conclusion that length, height, shucked weight and shell weight and some of their interactions are significant factors when estimating the factor rings. For the factor sex, we decided to do the prediction by sex separately.

However, there are still some parts of the study could be improved. For example, it’s necessary to know if there is any random effect during the measuring process, and we also need discuss with the idea of grouping by sex is practical.

Nevertheless, we can clearly see the high significance of the model, and the good predict result. They both proved that the ages of the abalone shell is highly related with those factors we took into the consideration, and also confirm us that the study is meaningful.

**Reference**

[1] Multi-collinearity - Wikipedia

https://en.wikipedia.org/wiki/Multicollinearity

# [2] Understanding Diagnostic Plots for Linear Regression Analysis https://data.library.virginia.edu/diagnostic-plots/

# [3]Significance Test for Linear Regression

# http://www.r-tutor.com/elementary-statistics/simple-linear-regression/significance-test-linear-regression