STAT2402 ASSIGNMENT 1

Student Name: Aaminah Irfan Student Number: 23642166

Executive Summary

This report presents a comprehensive analysis of abalones, a marine mollusk species of economic significance. The analysis centers on estimating analysis of abalone age based on various physical characteristics and building a predictive linear model. The dataset under examination comprises information from 4,177 abalones, featuring continuous numerical variables such as length, diameter, weights, and the number of rings. Additionally, a categorical variable, "Sex," is included in the dataset. The primary goal of this analysis is to establish a practical and efficient method for predicting the age of abalones using linear regression models. Traditional methods for age estimation involve labor-intensive processes such as counting shell rings, with an adjustment of 1.5 years to improve accuracy. However, these methods are fraught with uncertainty. To address this challenge, we build upon the work of Hossain (2019), who introduced innovative approaches using linear regression models for abalone age prediction based on physical characteristics. We explore two primary models, Model 1 and Model 2, to determine which offers the most accurate age estimates. Model 2, which incorporates square root transformations for weight variables, emerges as the superior model. Moreover, we delve into Model 3, which includes interaction terms to capture complex relationships between predictor variables. This model, despite a slightly lower R-squared value, offers a nuanced understanding of the data and is chosen for its ability to account for these interactions. The main findings of this analysis confirm the significant relationship between abalone age and their physical characteristics except Sex. Model 3, with interaction terms, provides a more comprehensive understanding of this relationship, while Model 2 offers a simpler alternative for practical applications. This report highlights the importance of using advanced statistical techniques to estimate abalone age accurately, contributing to more informed decisions for both consumers and abalone farmers. However, it's essential to acknowledge the limitations of the analysis, particularly the assumptions of linearity and homoscedasticity in the regression models. Future research should explore more robust modeling techniques and further investigate abalone biology to enhance age prediction accuracy.

Introduction

Abalones are a single-shelled (gastropod) marine mollusc and a family of reef-dwelling marine snails *[1]*. Their significance extends beyond marine ecosystems, as they have been a valuable food sources across the world, especially in areas where they are abundant in population [3]. Economic value of abalones is positively correlated with its age, making it important for both consumers and farmers to determine its age to be able to predict its price [3],[2]. Traditionally, number of shell rings are counted to estimate the abalones' age. Rings grow in the inner shell of the abalone. Number of shell rings grow as the abalone ages, where one ring is formed in approximately one year. Counting the number of rings involves cutting the ring, polishing, staining, and observing under the microscope. Due to uncertainty of staining all rings, researchers have added a value of 1.5 to the ring count to improve the approximation of age [3].

Due to the complexity of the traditional method, one notable study (Hossain, 2019) sheds light on the economic importance of abalone age and presents an innovative adjustment to traditional age estimation methods. Their findings suggest that simplifying the age estimation process using readily available physical attributes can be a practical and effective approach. The research delved into using Least Square Estimation Model and Ordered Probit Model to predict the age of the abalone using physical characteristics. It used physical characteristics: Sex, Length, Whole Weight, and Height to estimate the age of abalone and results found that Ordinary Least Squares Model (OLS) or Least squares Estimation models tend to perform better in predicting age of the abalone, especially when categorical variables like Sex are involved. Ordered Probit model provides ordered classifications but has limitations due to the small number of classes. Analysis revealed that accurate predictions are possible for abalones

with rings between 3 and 14. Based on the analysis, it was also suggested that use of simple physical characteristics such as weight, height, diameter, and length can be used to estimate the abalone age. By incorporating insights from the research paper by Hossain, this research paper aims to devise a more simple and practical method than the traditional method for estimating the abalone age using linear regression models and all physical characteristics. In short, this study aims to use abalones data to analyse the relationship between number of rings and other physical characteristics and obtain a model that can be used to predict the age of abalone using physical characteristics.

Data set used had information on 4177 abalones and its physical characteristics. All variables in the data are continuous and numerical variables, except 'Sex' which is a categorical variable. Table 1 provides the summary of data.

Table 1: Abalone Data Summary

Variable	Number of Observations	Value Range	Mean
Length: Longest shell measurement (mm)	4177	0.075- 0.815	0.524
Diameter : Shell Diameter (mm)	4177	0.055 -0.650	0.407
Height: Shell Height (mm)	4177	0.0000 - 1.130	0.1395
Wholewt: : Weight of the abalone (g)	4177	0.0020-2.8255	0.8287
Shuckedwt : Weight of meat without shell (g)	4177	0.0010-1.4880	0.3594
Viscerawt : Gut weight (after bleeding) (g),	4177	0.0005-0.7600	0.1806
Shellwt : Shell weight after being dried (g)	4177	0.0015-1.0050	0.2388
Rings : Number of rings (+1.5 gives the age in years)	4177	1.000-29.000	9.934
Sex: F = Female M = Male I = Infant	F = 1307 I = 1342 M = 1528		Mean of Rings (Age) by Sex F = 11.13 I = 7.89 M = 10.71

Methodology

Statistical methods were employed to predict age of abalones based physical measurements in the data. Response variable (y) considered was: Rings (number of Rings) and Explanatory variables (x) were: Sex, Length, Diameter, Height, Wholewt , Shuckedwt, Viscerawt, Shellwt. The analysis utilizes linear regression model, model selection and use of interaction term to develop predictive models. To conduct linear regression analysis, the assumptions were: predictor variables are linear, residuals are normally distributed, residuals have constant variance (homoscedasticity), and there is no multi-collinearity among predictor variables. Significance level of 5% was used through out the analysis. If the p-value for the significance test was less than 0.05 the null hypothesis would be rejected. For all linear models: Null and Alternate hypothesis were as follows:
Null hypothesis: There is no significant relationship between the number of rings and the explanatory variables

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$

Alternative hypothesis: There is a significant relationship between the number of rings and the explanatory variables:

$$H_1 : \text{At least one of } (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_6, \beta_7, \beta_8) \text{ is not equal to } 0$$

Data was loaded and structure and summary results were analysed. Data pre-processing was performed to ensure the quality of the data set. Observations with Height values of 0 were removed as they were implausible. Outliers were

also removed from 'Height' variable. Moreover, square root transformations were applied to 'Wholewt', 'Shuckedwt' , 'Viscerawt' and 'Shellwt' to improve linearity with 'Rings'.

Model (Model 1) with no transformations on weight variables was regressed on Rings. Model 2 with with weights transformed was also regressed on Weights. Process of Model Selection was employed on both Model 1 and Model 2 was used to remove all insignificant variables. Variables with p-values < 0.05 were considered insignificant.

Model 1 and Model 2 were compared using Analysis of Variance (ANOVA) F-Test to choose the best model. ANOVA F-test is used to compare goodness of fit between different regression models. The F-statistic for ANOVA test is computed by comparing the reduction in the sum of squared residuals (RSS) between the two models with the corresponding degrees of freedom. A high-statistic with a low p-value (< 0.05), suggests a preference of one model over the other.

Model Diagnostics were performed on Model 2 to test the assumptions of linearity.

To explore an in-depth relationship between the explanatory variables, model involving interaction term was employed. Logarithmic transformation was employed on the response variable: Rings, to adjust the skeweness of the variable. Logarithm of Rings was regressed on all explanatory variables and all combinations of explanatory variables (interaction terms). Model for analysis of interaction term was:

$$log(y) = \beta_0 + (\beta_1 \times X_1) + (\beta_2 \times X_2) + (\beta_3 \times X_3) + ... + (\beta_i \times X_j) + (\beta_i \times X_j : X_k))$$

Backward Selection was used to remove insignificant variables (i.e., p-value < 0.05) and the final model i.e., Model 3 was chosen.

Model Diagnostics were performed on Model 3 to test the assumptions of linearity.

Results

To begin the analysis, data structure and summary results were examined of abalone data which contained information on 4,177 abalones and their physical characteristics. Table 1 provides a summary of the abalone data. Note 'Height' variable contains value of 0. Hence, it was removed as it is unlikely to observe the Height of 0 and is likely to be an error in data entry. Individual plots for all variables were examined for any outliers, un-explained data and skewness. Figure 1 {Appendix} shows the box plot for Sex variable against Rings. It was noted that most abalones ranged between 2 and 18 number of Rings. Outliers from Sex were not removed as they could potentially be real data and not error in data points. Figure 2 displays Frequency histograms and scatter plots for Length, Diameter and Height. Height variable was sliced so it only include data with Height less than 0.4 mm to exclude outliers. Even though Length and Diameter frequency histograms seem to be skewed to the right, no transformations could make it more linear than when not transformed. Figure 3 displays linear plots for variables: Wholewt , Shuckedwt, Viscerawt and Shellwt before and after square root transformations. Note, square root transformations make the variables more linear. Hence square root transformations were used in the analysis for the variables: Wholewt , Shuckedwt, Viscerawt and Shellwt. Analysis involved developing predictive models for estimating abalone age based on physical characteristics. Initially simple model (Model 1) without transformations on weight variables was regressed on the response variable (Rings) . Model 2 i.e., Model with transformations on weight variables, was also regressed on the response variable. Process of Backward selection was used to remove insignificant variables i.e., variables with p-values < 0.05. Backward Selection involves including all predictor variables and iteratively removing the least significant variable , one with highest p-value, until all variables are statistically significant.

Final Model 1 and Model 2 were as follows:

Model 1 Equation:

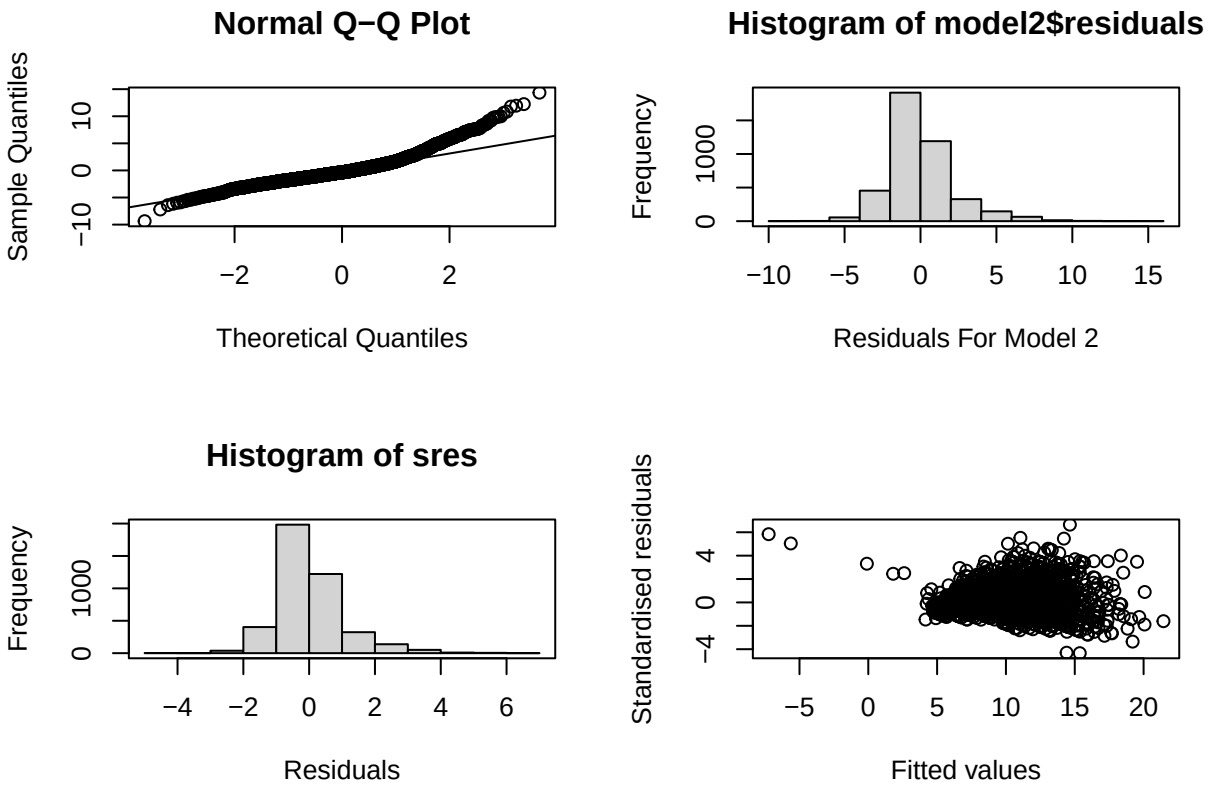
$$Rings = 2.55 + (8.75 \times Diameter) + (25.20 \times Height) + (9.07 \times Wholewt) - \quad (1)$$
$$(19.81 \times Shuckedwt) - (10.55 \times Viscerawt) + (7.58 \times Shellwt) \quad (2)$$

Model 2 Equation:

$$Rings = 3.49 - (5.09 \times Length) + (6.70 \times Diameter) + (16.22 \times Height) + \quad (3)$$
$$(19.96 \times \sqrt{Wholewt}) - (27.02 \times \sqrt{Shuckedwt}) - (9.00 \times \sqrt{Viscerawt}) + (12.53 \times \sqrt{Shellwt}) \quad (4)$$

For both Models, all the variables and the overall model were statistically significant (i.e., p-value < 0.05). Hence the null hypothesis of no significant relationship between the number of rings and the explanatory variables was rejected. Model 1 and Model 2 were compared using Analysis of Variance (ANOVA) F-test to choose the best model. F-test revealed a p-value of almost 0 i.e., p-value < 0.05. Consequently, Model 2 was chosen as the best model. Multiple R-squared for Model 2 is 55.35% and an F-statistic of 737.7 with a p-value < 0.05.

Model Diagnostics were performed on Model 2 to test for linearity. Plots for Model Diagnostics are following:



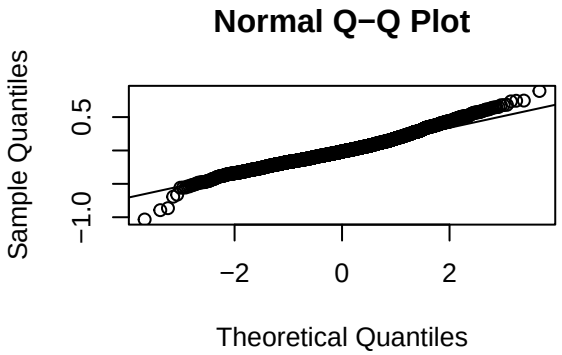
For analysis of model involving interaction terms, similar approach of Backward Selection was used. Insignificant variables (variables with highest p-value and less than 0.05) were iteratively removed until only significantly variables remained. Multiple R-squared for Model 3 is 64.4% and an F-statistic of 469.9 with a p-value < 0.05.

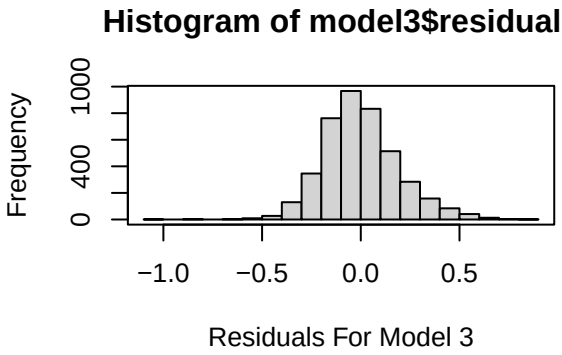
Final model i.e., Model 3 was chosen.

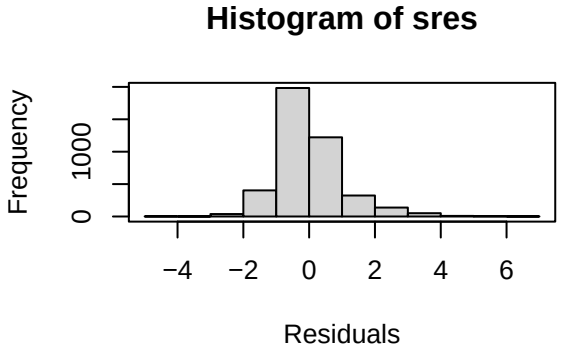
Model 3 Equation:

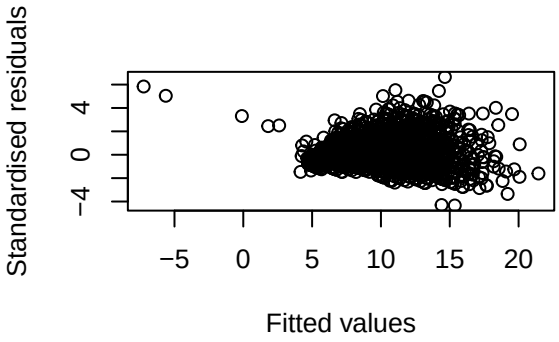
$$\begin{aligned}
\log(Rings) =& 0.43 + (2.3 \times \text{Length}) + (4.80 \times \text{Diameter}) + (3.44 \times \text{Height}) && (5)\\
& - (2.47 \times \text{Shuckedwt}) - (2.14 \times \text{Viscerawt}) + (4.32 \times \text{Shellwt}) && (6)\\
& - (10.34 \times \text{Length:Diameter}) + (3.41 \times \text{Length:Shuckedwt}) && (7)\\
& + (6.88 \times \text{Diameter:Viscerawt}) + (6.27 \times \text{Height:Wholewt}) && (8)\\
& - (11.31 \times \text{Height:Shuckedwt}) - (12.88 \times \text{Height:Shellwt}) && (9)\\
& + (0.68 \times \text{Shuckedwt:Wholewt}) - (2.44 \times \text{Viscerawt:Wholewt}) && (10)\\
& + (2.53 \times \text{Shuckedwt:Viscerawt}) - (2.18 \times \text{Shuckedwt:Shellwt}) && (11)
\end{aligned}$$

Model Diagnostics were performed on Model 3 to test for linearity. Plots for Model Diagnostics are following:









Discussion

Results for Model 2 and 3 revealed all statistically significant variables (p-value < 0.05). This suggest that there is a significant relationship between number of rings and the explanatory variables.

Interpretation for Model 2 is: For each unit increase in Length of abalone, the predicted number of rings decreases by 5.09 units. For each unit increase in diameter, number of rings increases by 6.70 units, implying diameter is positively related to Age. For each unit increase in Height, Rings increases by 16.22 units. For each unit increase in square root of Whole weight, predicted number of rings increases by 19.96 units. For each unit increase in square root of weight of meat without shell, Rings decreases by 27.02. For each unit increase in square root of gut weight, Rings decreases by 9. For each unit increase in square root of Shell weight, Rings increases by 12.53.

Interpretation for Model 3 is: For each unit increase in Length of abalone, the expected natural logarithm of Rings is expected to increase by 2.3 units, holding all other factors constant. For each unit increase in Diameter, the expected natural logarithm of Rings is expected to increase by 4.79 units, holding all other factors constant. For each unit increase in Height, the expected natural logarithm of Rings is expected to increase by 3.44 units, holding all other factors constant. For each unit increase in Weight of meat without shell, the expected natural logarithm of Rings is expected to decrease by 2.47 units, holding all other factors constant. For each unit increase in Gut Weight, the expected natural logarithm of Rings is expected to decrease by 2.14 units, holding all other factors constant. For each unit increase in Shell Weight, the expected natural logarithm of Rings is expected to increase by 4.32 units, holding all other factors constant. For each unit increase in combined effect of Length and Diameter, Rings decreases by 10.34. For each unit increase in combined effect of Length and Weight of meat (without shell), Rings increases by 3.41 For each unit increase in combined effect of Diameter and Gut Weight, Rings increases by 6.88 For each unit increase in combined effect of Height and Whole Weight, Rings increases by 6.27. For each unit increase in combined effect of Height and Weight of meat, Rings decreases by 11.31. For each unit increase in combined effect of Height and Shell Weight, Rings decreases by 12.88. For each unit increase in combined effect of Weight of meat and Whole weight, Rings increases by 0.68. For each unit increase in combined effect of Gut weight and Whole weight, Rings decreases by 2.44. For each unit increase in combined effect of Weight of meat and Gut

Weight, Rings increases by 2.53. For each unit increase in combined effect of Weight of meat and Shell Weight, Rings decreases by 2.18.

It was also revealed that 55.35% of the variation in the data is explained by the model 2, while 64.4% of the data is explained by Model 3. However, Model 2 has a higher F-statistic, suggesting that it may have a stronger overall statistical significance. Since, the main objective of the analysis was to analyse the relationship between Rings and all other variables in the data, Model 3 is chosen to be a better model. This is because it has a higher R-squared indicating more variation in the data is explained by the model. Although, Model 3 has a slightly lower F-statistic, it is still statistically significant indicating the model and all the variables are significant. Model 3 also includes interaction terms, which can help capture complex relationship between predictor and response variables. Model Diagnostics plot for Model 3 also seems to be more linear and normal than for Model 2. However, Model 2 can be used due to its simplicity to interpret and implement. Both Model 2 and 3 can be used based on the depth of analysis required i.e., Model 3 can be used to predict age while accounting for interaction terms and Model 2 can be used for quicker and effective calculation of age.

While the study developed models to analyse relationship between Age and physical characteristics, there are several limitations. Firstly, the study uses linear regression to build models to analyse the relationship and assumes homoscedasticity and linearity. However, Residuals Plot in Model Diagnostics show a pattern and issues like multi-collinearity. Hence, a better and more robust models can be used besides linear model to analyse the relationship between the models. Another limitation is the lack of theoretical research conducted and lack of research on marine organisms like abalones, which prevented a better analysis of the model. Hence, more research on abalones should be conducted to build better and robust linear models.

References:

[1]: Abalone. (n.d.). www.fish.wa.gov.au. https://www.fish.wa.gov.au/Species/Abalone/Pages/default.aspx

[2]: Hossain, M. M., & Chowdhury, M. N. M. (2019). *Econometric Ways to Estimate the Age and Price of Abalone.* Federal Reserve Bank of St Louis.

[3]: Misman, M. F., Samah, A. A., Aziz, N. A. A., Majid, H. A., Shah, Z. A., Hashim, H., & Harun, M. F. (2019, September 1). *Prediction of Abalone Age Using Regression-Based Neural Network.* IEEE Xplore. https://doi.org/10.1109/AiDAS47888.2019.8970983

Appendix

Figure 1: Number of Rings (Age) by Sex

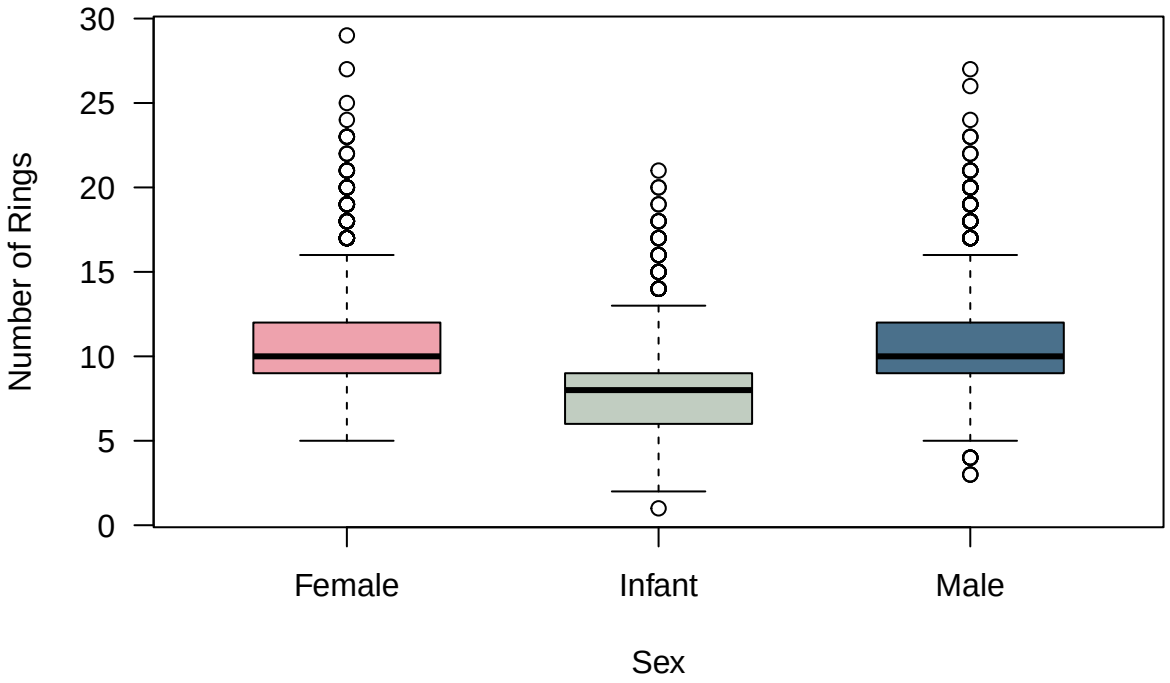


FIGURE 2:

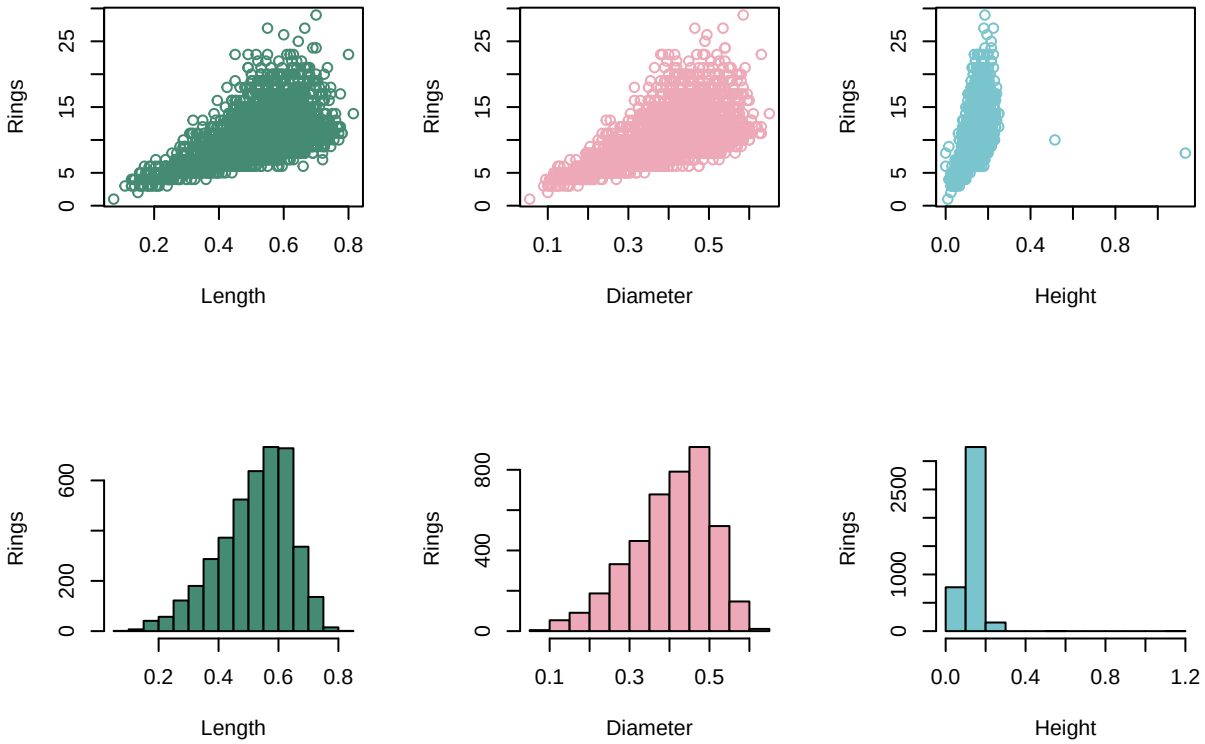
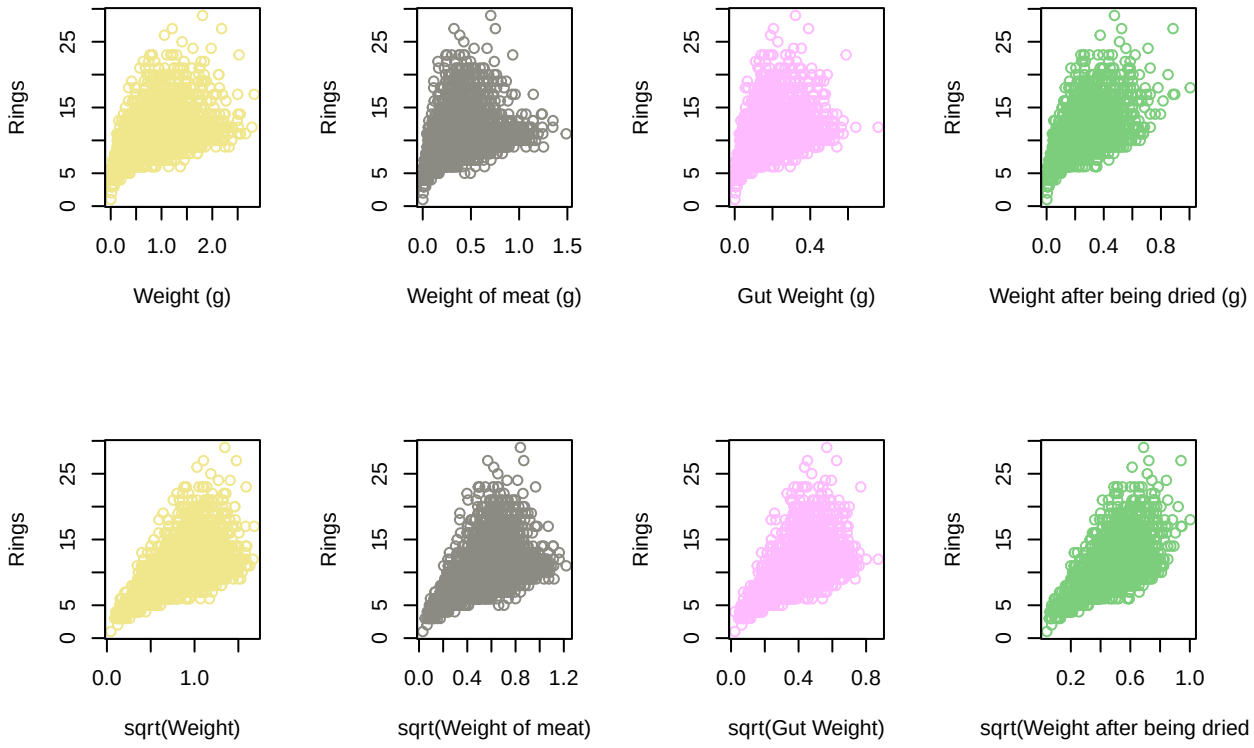


Figure 3:


Rings
0 5 15 25
0.0 1.0 2.0
Weight (g)
Rings
0 5 15 25
0.0 0.5 1.0 1.5
Weight of meat (g)
Rings
0 5 15 25
0.0 0.4
Gut Weight (g)
Rings
0 5 15 25
0.0 0.4 0.8
Weight after being dried (g)
Rings
0 5 15 25
0.0 1.0
sqrt(Weight)
Rings
0 5 15 25
0.0 0.4 0.8 1.2
sqrt(Weight of meat)
Rings
0 5 15 25
0.0 0.4 0.8
sqrt(Gut Weight)
Rings
0 5 15 25
0.2 0.6 1.0
sqrt(Weight after being dried

“‘