1. Introduction

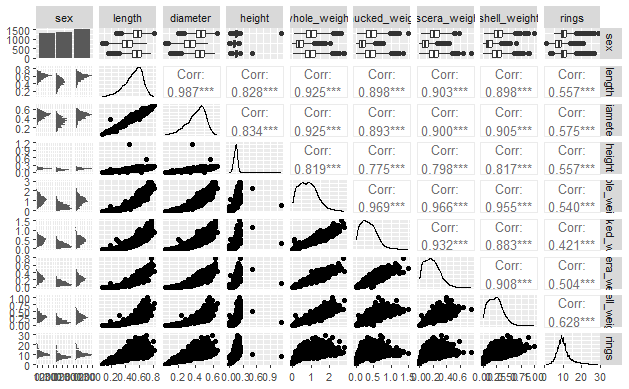
Determining the age of abalone is a cumbersome as well as expensive process which increases the cost and limits its popularity. It usually is done through microscopic counting of its rings relative to the process of counting rings to find the age of a tree. Therefore, there is an intent to determine the age of abalones using econometric methods to reduce the costs of producers as well as consumers. The economic value of abalone is positively correlated with its age. Estimating the age of abalone accurately is important for both farmers and customers to determine its price (Hossain and Chowdhury, 2019)

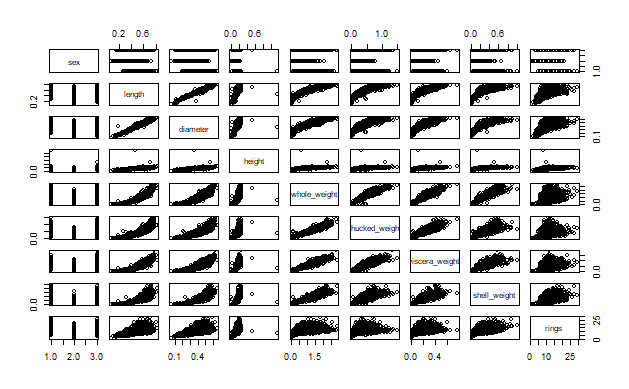
In our dataset we have 8 IV’s namely sex, length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight.

We have 1 DV namely: “Rings”, rings would tell us how to predict the age of the Abalone so we use this as our dependent variable.

We have one categorical variable namely: “Sex” that has 3 levels (1) Male denoted in M (2) Female denoted in F (3) Infant denoted in I.

1. Descriptive statistics



Shows length is highly correlated.

1. Model creation – LM and Stepwise

* Initial LM model:

model <- lm(rings ~ sex + length + diameter + height + whole\_weight + shucked\_weight + viscera\_weight + shell\_weight, data = df)

Initial model that has not been regressed

* Model1 (Regressed model):

lm(rings ~ sex + diameter + height + whole\_weight + shucked\_weight + viscera\_weight + shell\_weight, data = df)

This model has been regressed and Length is removed as it fails to reject the null hypothesis at 0.05 pvalue

* Model2 (Final model):

lm(rings ~ sex + diameter + height + whole\_weight, data = df)

This model we have removed shucked\_weight, viscera\_weight, and shell\_weight as they can be represented by whole\_weight and makes our model linear. Therefore, this is our **final model**.

1. Discussion of results

In our initial model we have removed the “length variable” as this was not statistically significant at 0.05pvalue. Then we have two models, model1 which has only length removed and model2 which has shucked\_weight + viscera\_weight + shell\_weight removed. To choose from model1 and model2 which first check linearity assumption as this has to be satisfied first and model2 yields a better linearity than model1. Therefore, we use model2 even as it has a lower R2(0.3) as compared to model1 (0.5) we still need to use this as our R2 score would be useless if our model does not follow linearity assumption.

Further, all other variables in our model2 are statistically significant at 0.05 pvalue and follows a normal distribution of residuals as even though the distribution of the error terms are not random on our diagnostics tests, we assume a normal distribution due to the central limit theorem that states with a sufficiently large enough sample size we can assume a normal distribution from the dataset (Frost, 2022). Here we have 4000 observations which is a sufficiently large enough sample size.

1. Model diagnostics

Assumptions:

1) Linearity: The expected (mean) outcome is a linear function of the predictors. Also, the parameters must be linear.

2) Homoscedasticity: The residuals, conditional on X, have constant variance.

3) Independence: Errors do not covary after conditioning on X.

4) Normality: Errors follow a normal distribution.

5) No multicollinearity: Independent variables cannot be predicted from each other.

6) No outlier effects: The model represents the data well, and no observations disproportionately influence the model fit.

7) No measurement error: Predictors are measured perfectly.

8) No specification error: The statistical model matches the data generating process in its functional form and in which variables are included.

1. Here we can observe linearity in our model represented by the red line.

Chart, scatter chart

Description automatically generated

1. The QQ plot also shows us the residuals follow a exponential distribution after 1 unit on the x axis. However, the central limit states that given enough sample size we can assume a normal distribution as explained by Frost, (2022).

Chart, line chart

Description automatically generated

1. Correlation plot tells us highly correlated IV’sTable

   Description automatically generated with low confidence
2. Outliers shows us specific observations that can skew our model

Chart, histogram

Description automatically generated

1. Residuals vs leverage and the normality of residuals, tells us the distribution of our error terms do not follow a normal distribution. However, this can be assumed to be normal based on the central limit theorem.

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

1. Conclusion / Recommendations

Our model performs well to describe our Dependent variable however there is still room for improvement namely in our model the Male sex dummy variable is not statistically significant maybe we can use other regression methods that would remove or categorize this further so we can have a better model, however in doing so, we are unable to predict the age of a Male abalone.

1. Appendix

Male categorical variable is not statistically significant.

Text

Description automatically generated

Bibliography:

Hossain, Md. Mobarak, & Chowdhury, Md Niaz Murshed. (2019, January 3). Econometric ways to estimate the age and price of abalone. Munich Personal RePEc Archive. <https://mpra.ub.uni-muenchen.de/91210/1/MPRA_paper_91210.pdf>

Frost, J. (2022, January 30). Central limit theorem explained. Statistics By Jim. https://statisticsbyjim.com/basics/central-limit-theorem/