An Abalone-Age Investigation

470408957 | 490443251 |

This version was compiled on November 18, 2020

In this project we investigated two predictive models for the predicitng the number of rings in an abalone which is used to find the age by adding 1.5 to the number. In order for our models to satisfy the linearity assumptions we had to perform transformations on the data set, where we then could conduct multiple linear regression through backward and forward selection using Akaike Information Criterion (AIC). The resulting model selection was further manipulated as there was strong observed collinearity with some independant variables which led to our final model.

1. Introduction

The orginal article for this data set was utilised to explore the age of an abalone. However, the classical method for determining age is to cut the shell, stain it, and count the number of rings under a microscope. This is a very tedious process, which provides reasoning to encompass easier ways in calculating the number of rings without manually counting the rings. We therefore aim to construct a regression model that can predict the number of rings an abalone has (calculating the age), using only an abalones physical attributes which are easily and quickly measured. To do so, linear relationships need to be considered between the variables and the value they aim to predict.

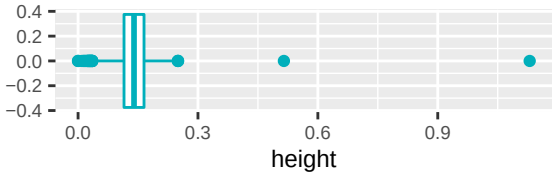
2. Data Set

This biological study looks at abalones which are a common type of sea snail, where the majority of abalone species are found in cold waters, such as off the coasts of New Zealand, South Africa, Australia, and Tasmania. The data set is provided by the University of California Irvine Machine Learning Repository which explores variables describing physical characteristics for abalones. The original data was collected by Marine Resources Division in Taroona, Tasmania, containing 4177 observations with 9 different variables which has no missing values.

2.1 Variables.

- Sex; factor; male, female or infant
- Length (mm); continous; longest shell measurement
- Diameter (mm); continous; perpendicular to length
- Height (mm); continous; with meat in shell
- Whole Weight (g); continous; whole abalone
- Shucked Weight (g); continous; weight of meat
- Viscera Weight (g); continous; gut weight (after bleeding)
- Shell Weight (g); continous; after being dried
- Rings; integer; number of rings. +1.5 gives age in years

2.2 Outliers. In exploring the different variables there appeared to be two values identified as clear outliers in the height variable. Thus, concluding they are likely to be errors induced when entering the data and will be filtered out.



3. Analysis

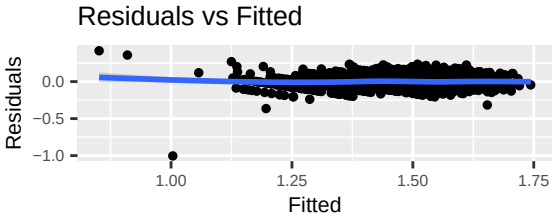
3.1 Transformations. Prior to selecting an appropriate model, with reference to appendix 1, clearly the indepedant variables do not possess a linear relationship with number of rings. The variables clearly rise rapidly and reach a plateau, thus it was found best to perform log transformations of the variable rings, length, diameter, weight shucked, weight viscera, weight shell and a square root transformation was applied to height and log of rings. With reference to appendix 2, the data set now adopts a linear relationship with the predictive variable, allowing for a linear regressive model to work appropiately.

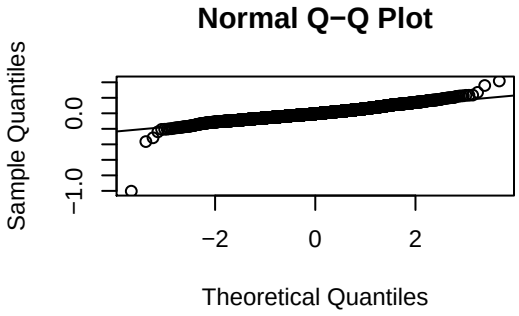
3.2 Model Selection. Provided the linearity assumptions with respect to the dependant variable are satisfied, a model search was justfiied by the Akaike information criterion through a backward and forward variable seleciton. After conducting the relevant seacrch it was found that the forward model did not include log of diameter and log of length while the backward approach included all variables, shown in the table below.

	Forward Model		Backward Model	
Predictors	Estimates	p	Estimates	p
(Intercept)	1.43	**<0.001**	1.45	**<0.001**
logshell	0.11	**<0.001**	0.11	**<0.001**
logshucked	-0.19	**<0.001**	-0.19	**<0.001**
logwhole	0.19	**<0.001**	0.20	**<0.001**
sexi	-0.02	**<0.001**	-0.01	**<0.001**
logviscera	-0.03	**<0.001**	-0.02	**<0.001**
sqrtheight	0.13	**0.007**	0.12	**0.012**
			0.07	**0.005**
			-0.08	**0.005**
Observations	4175		1.45	
R^2/R^2 adjusted	0.647 / 0.647		0.648 / 0.647	
AIC	-10882.310		-10887.886	

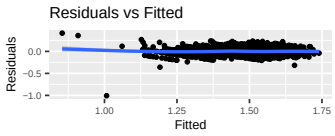
3.3 Assumption Checking.

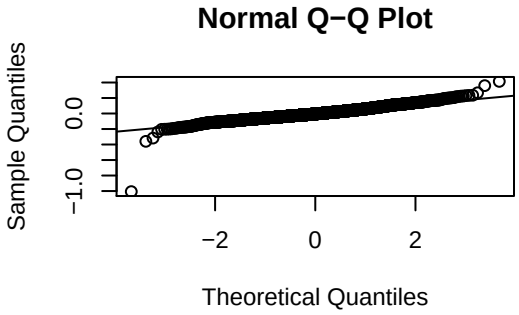
Forward Residual Verse Fitted/QQ Plot.





Backward Residual Verse Fitted/QQ Plot.



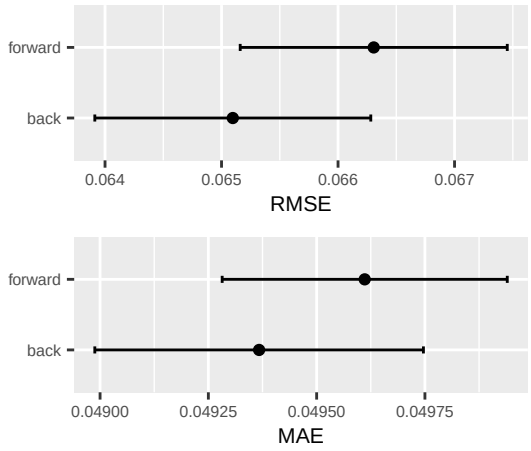


Assumption checks for both forward and backward models;

- **Linearity**: The residual vs fitted values plot indicates no obvious curvature for both thus the linearity assumption is satisfied.
- **Independence**: Referencing to appendix 3 the data was collected over 5 different regions in the Tasman Sea, hence there will be independence between the observations from the differing locations that the data was pulled from.
- **Homoskedasticity**: The residuals do not appear to be fanning out or changing over the range of the fitted values for both, thus the constant error variance assumption is met.
- **Normality**: The normality assumption is at least approximately satisfied. In the QQ plot, the points are reasonably close to the diagonal line, however the sample size is large enough to rely upon the central limit theorem ensuring the inferences are approximately valid.

4. Results

They had the same adjusted R^2 thus we need to encompass the RMSE and MAE to identify the better performing model. Cross validation was used to compute RMSE and MAE of our forward and backward models. This is done to account for and minimise overfitting. From comparing the two models' RMSE and MAE we could see that they only differ very slightly, indicating the better performing model will be the backward model over the forward model, due to having a lower RMSE and MAE.



Talk about Given the strong multicollinearity observed between some variables it is good to note the multicollinearity among some variables, does it make sense to include all the weight variables in the model? There is no mention of the observed collinearity and its effect on the model.

5. Discussion/Conclusion

All standard LaTeX environment are directly usable if needed, including of course all mathematical environments and symbols such as, say, the greek lettering: α, β, γ, and so on.

References

Allaire J, R Foundation, Wickham H, Journal of Statistical Software, Xie Y, Vaidyanathan R, Association for Computing Machinery, Boettiger C, Elsevier, Broman K, Mueller K, Quast B, Pruim R, Marwick B, Wickham C, Keyes O, Yu M (2017). *rticles: Article Formats for R Markdown*. R package version 0.4.1, URL https://CRAN.R-project.org/package=rticles.

MacFarlane J (2017). *Pandoc: A Universal Document Converter*. Version 1.19.2.1, URL http://pandoc.org.

Xie Y (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17, URL https://yihui.name/knitr/.