Multivariate Analysis of Tasmanian Blacklip Abalone Gender and Weights

ZZSC5855 – Multivariate Analysis for Data Scientists
Amy Raggatt – z5294732
Assessment Item 7 – Project

I. Introduction

Multivariate analysis provides key insights into datasets which contain more than one response variable, or variable to be predicted. Usually, when studying a complex real-life situation, many variables are required. The 1994 abalone dataset [1] is a well-known multivariate dataset which contains information on the dimensions, weight, sex and number of rings for 4177 Blacklip abalone caught off the coast of Tasmania. Tasmania's abalone fishery produces the highest amount of wild caught abalone in the world, accounting for 25% of global production [2]. It is vital that decisions made regarding sustainability limit the impact of fishing on long-term abalone population and the wider environmental ecosystem, and ensure the continued success of the industry for years to come. The abalone industry also provides income and investment capital for a large number of Tasmanians, therefore, the economic impact of harvesting abalone is also of extreme importance and a focus is placed on profitability. This project will aim to use multivariate analysis to answer questions posed about sustainability and profitability.

II. Exploratory Data Analysis

After loading the data and removing the variables which were not of interest to the client (Rings, Whole weight and Shell weight), we interpreted the pairs plot in order to understand the relationships between the variables (see Fig. 1). Females are indicated in pink, males in blue and infants in green. The boxplots show that there are outliers among all numeric variables. The bar chart shows that the distribution for each gender is approximately equal, so class imbalance is not considered to be an issue in this analysis. Some other points of note include the positive skew for Height, which appears to be effected by some extremely large outlier values, and the non-linearity seen in the curved plots between the weight variables and length, width and height.

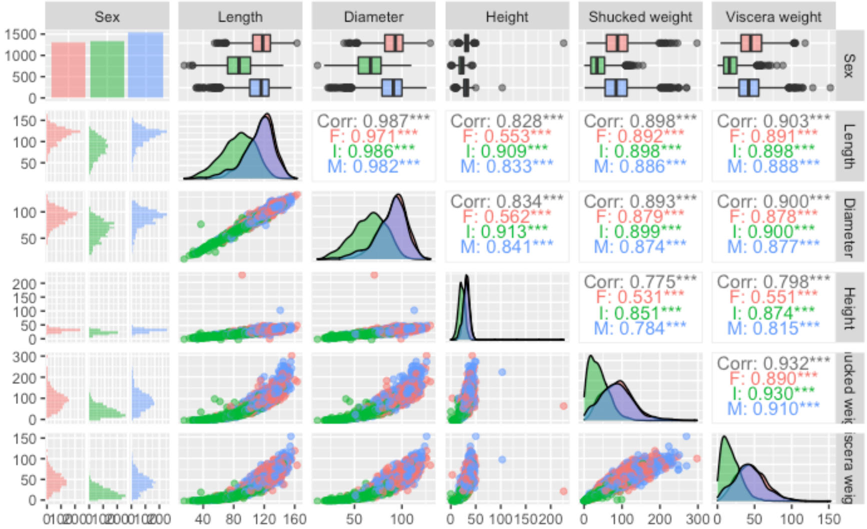

Fig. 1. Pairs plot of Abalone data, separated by gender (M = blue, F = pink, I = green).

III. Data Cleaning and Transformations

First, the data was checked for missing values. All 4177 observations include all measurements. The validity of some height measurements was considered. Two observations had height values of 0 mm, which seemed unlikely given the other measurements present (see Fig. 2). The two largest height values (103 mm and 226 mm) were also studied (see Fig. 3). These very small and large height values are not consistent with the other measurements taken for that particular observation and seem unreasonable. They appear to potentially be the result of improper data entry, and were removed from the dataset.

Sex	Length (mm)	Diameter (mm)	Height (mm)	Shucked weight (g)	Viscera weight (g)
I	86	68	0	41.3	17.2
I	63	46	0	11.5	5.7

Fig. 2. Data entries where Height = 0mm

Sex	Length (mm)	Diameter (mm)	Height (mm)	Shucked weight (g)	Viscera weight (g)
F	91	71	226	66.4	23.2
M	141	113	103	221.5	97.3
F	163	130	50	178.1	84.0

Fig. 3. Data entries in descending Height order

Many multivariate techniques assume variables are drawn from a multivariate normal distribution and that relationships are linear in nature, and therefore data transformation is used to transform data which does not fit these assumptions. The scatterplots (see Fig.1) show non-linear relationships between the three predictor variables and the response variables, and skewness can be seen in some of the probability distribution functions along the diagonal. After removing outliers, length and diameter appear to be positively skewed, while both weight variables appear to be negatively skewed. As a result, we transformed these variables. Replotting the pairs plot shows that our dataset now looks far more normal and linear in nature (Fig. 4). It should be noted that transforming variables can change the interpretation of our results, and predictions made from models involving transformed data should be treated with care to avoid errors.

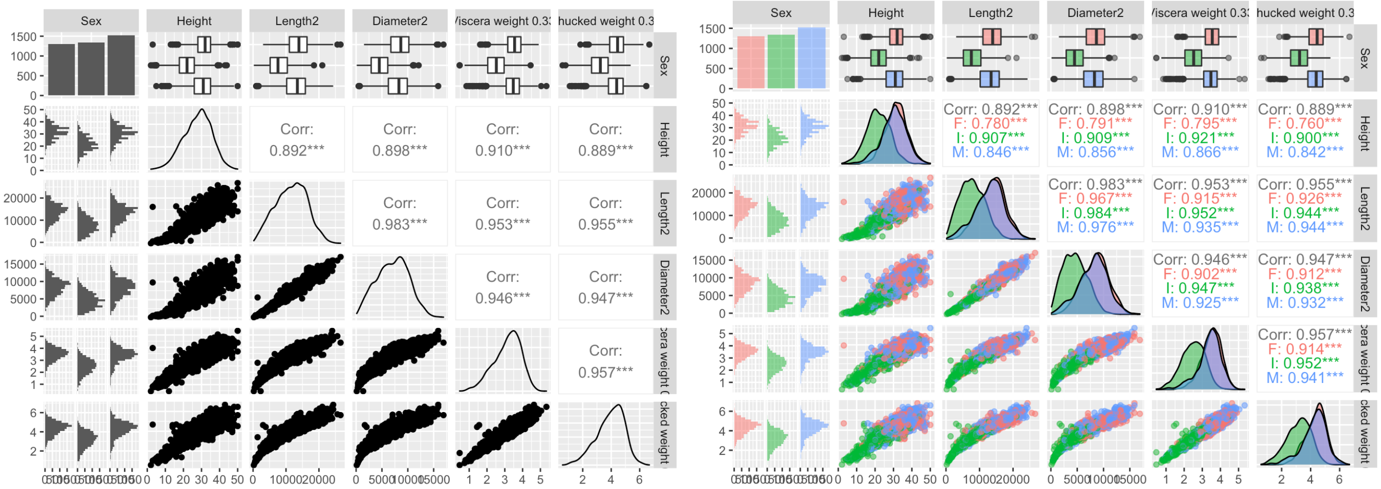

Fig. 4. Pairs plots of the transformed data. The right hand side graphs shows results separated by gender (M = blue, F = pink, I = green).

Despite our transformations, hypothesis tests which test for multivariate normality reject the null hypothesis that our transformed data are MVN or even univariate normal. However, we can see that our probability distributions are much more "normal" in shape than they were originally. They appear to be fairly symmetrical and bell-shaped (see Fig. 4) and skewness has decreased for all variables, even if only slightly.

IV. SUSTAINABILITY – CLASSIFYING SEX

Classifying the sex of an abalone was broken into four parts:

A. Multiclass classification (male, female or infant),
B. predicting infants as opposed to others (to avoid harvesting them),
C. predicting females as opposed to others (when profitability is prioritised) and
D. predicting males as opposed to others (when sustainability is prioritised).

Each part had three types of models fit to it:

- linear discriminant analysis (LDA)
- quadratic discriminant analysis (QDA) and a
- support vector machine (SVM).

Both LDA and QDA assume multivariate normality, and LDA has the added assumption of equal variance-covariance matrices. SVMs may be considered nonparametric, since they do not require us to make distributional assumptions about the data. Cross validation was used for all models in order to assess how well each performed on unseen data points.

A. Multiclass classification

Predicting the sex of an abalone as either male, female or infant may be challenging, when considering the colour coded scatterplots in Fig. 4. There appears to be a large amount of overlap for male and female abalone shells, which makes separating the data into different groups difficult. As a result, the models that were created for this task were not particularly informative or reliable. The linear discriminant analysis model was tested first and produced the results seen in Fig. 5.

Raw data confusion matrix:

		Actual		
		Female	**Infant**	**Male**
Prediction	**Female**	303	12	307
	Infant	180	938	285
	Male	823	390	935

Overall Accuracy = 52.15%

Transformed data confusion matrix:

		Actual		
		Female	**Infant**	**Male**
Prediction	**Female**	312	12	316
	Infant	209	979	321
	Male	785	349	890

Overall Accuracy = 52.26%

Fig. 5. Confusion matrices summarising performance of LDA for multiclass classification

It appears that the linear discriminant model predicts the correct sex (M, F or I) only a little more than half of the time (52.14% accuracy for raw data). The transformed data has a similar overall accuracy (52.26%) compared to the LDA for the raw data. The model which uses transformed data seems to correctly classify infants and females more often, while the raw data model does better at predicting males correctly.

The QDA gave similar results to those above. This may be because the dataset does not follow a normal distribution, even in its transformed form. Instead of assuming normality, we next fit an SVM, which unfortunately did not perform significantly better than either LDA or QDA. SVMs are known to underperform when response classes are overlapping. The large dataset also meant that the training time for this model was much higher.

B. Predicting Infants as opposed to others

After running LDA, QDA and SVM models aimed at classifying an abalone as either infant or not infant, the SVM model based on the transformed data was chosen as the preferred model for this task, since it had the highest classification accuracy (approximately 80.16%). Now that we are not trying to separate the males from the females, our classification accuracy has improved dramatically across the board. All models gave similar accuracy levels for this task. The QDA may also be an appropriate model if ease and speed of training is prioritised (see Fig. 6).

Transformed data confusion matrix:

		Actual	
		Not infant	**Infant**
Prediction	**Not infant**	2524	543
	Infant	309	797

Overall accuracy = 79.58%

Fig. 6. Confusion matrix summarising performance of QDA for infant VS non-infant

C. Predicting Females as opposed to others

Once again, our results across the three different methods for this task produced similar results, with the transformed SVM model again giving the highest classification accuracy (approximately 68.70%, see Fig. 7). As we would expect, the percentage of cases which were classified correctly decreased to around 68%. This is potentially due to the overlapping nature of the males and females.

D. Predicting Males as opposed to others

Interestingly, our results for this task produced one clear winner, unlike the previous 3 tasks that we completed. The SVM using raw data performed best with a classification accuracy of approximately 68.7%, compared to accuracy scores of approximately 62.5% for most other models.

A summary of the accuracy results across all models is shown below. The highlighted cells indicate best performance for that task.

	LDA		QDA		SVM	
	raw	transformed	raw	transformed	raw	transformed
A. Multiclass classification	52.14%	52.26%	51.52%	51.74%	52.36%	52.50%
B. Predicting Infants as opposed to others	79.46%	79.10%	79.58%	78.82%	80.01%	80.16%
C. Predicting Females as opposed to others	68.7%	68.92%	67.75%	67.94%	68.68%	68.70%
D. Predicting Males as opposed to others	62.59%	62.66%	62.50%	62.04%	68.68%	63.29%

Fig. 7. Summarised accuracy results for classifcation problems

V. PROFITABILITY – PREDICTING WEIGHT

A. Multivariate multiple regression mode analysis

This task required us to create a model which took as its inputs length, diameter and height, and predicted both shucked and viscera weights. Since we have two response variables and since this is no longer a classification problem (we will get a numeric value for weight, rather than grouping males, females and infants), a multivariate multiple regression model was appropriate. The model resulted in two formulas, one which calculates shucked weight and the other viscera weight. All predictors (height, length squared and diameter squared) are classed as significant for all models. On this basis, we will not be removing any of these predictors from the model. The adjusted R^2 values indicate that 91.95% of the variation in shucked weight and 92.53% of the variation in viscera weight can be explained by the variation in height, length squared and diameter squared. Unfortunately though, the patterns that can be seen in the residual plots (Fig. 8) indicate that our model may be biased. A good model has residual plots with points randomly scattered above and below the zero mark (shown here by a horizontal grey line). This curved pattern indicates non-linearity, which we identified as a potential issue. Linearity is an assumption of the multivariable regression model. To remedy this, more complex transformations may be necessary. For the sake of simplicity and ease of use, we will not be transforming our variables any further.

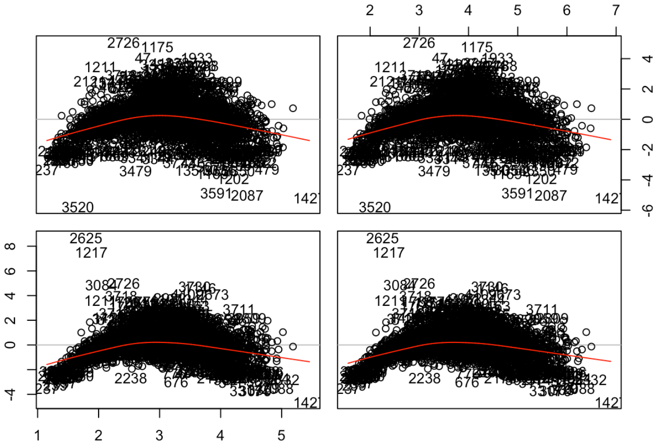

Fig. 8. Residual versus fitted values for the multivariate regression model

B. *Profitability index*

We have created a function which provides a predicted sale price for a new abalone given its measurements and that day's prices. The function also provides a prediction interval that contains the true dollar value some specified percent of the time (e.g. 90%). This function works with the transformed data rather than the original raw data. As a result, we acknowledge that this interval may not provide accurate results and would advise against using this function in its current state. Further research and investigation would need to be carried out to improve the performance and reliability of using these confidence intervals if the client was still interested in pursuing this feature.

VI. Conclusion

This project has used multivariate analysis to answer questions posed about sustainability and profitability of harvesting abalone with varying success. The dataset in its raw form posed challenges to the data scientist as many multivariate models have assumptions which must be met in order for the model to be valid. Predicting infant versus not infant seems to be an achievable goal with the use of an SVM. It was challenging for the models to detect differences between mature male and female abalones and therefore other techniques may need to be investigated to achieve those goals. The multivariate linear model used to predict the sale price of an abalone shows promise, however, the non-linearity of the data is concerning and may prompt us to modify the structural form of the model in order to improve reliability.

VII. References

[1] Archive.ics.uci.edu. 2021. *UCI Machine Learning Repository: Abalone Data Set.* [online] Available at: <http://archive.ics.uci.edu/ml/datasets/Abalone> [Accessed 16 October 2021].

[2] Tasmanian Abalone Council. 2021. *MANAGING THE INDUSTRY AND THE ABALONE RESOURCE*. [online] Available at: <https://www.tasabalone.com.au/about-industry/fishery-management/> [Accessed 16 October 2021].