Classification and Pricing of Farmed Abalone

A. Gill

December 12, 2022

Introduction

An enquiry has been made as to the potential use of software to help abalone farmers to quickly assess whether an abalone should be harvested or left in the water. The aim is to use non-intrusive measurements - length, diameter and height - to predict an abalone's sex and infancy status, as well as its weight (and thus market value), allowing farmers to return rejected abalone to the water without having harmed them.

1 Exploratory Analysis

Data on more than 4000 harvested abalone have been supplied. The dataset includes the following columns (variables):

- Sex
- Length, diameter and height
- Whole, shucked, viscera and shell weights
- Number of rings on the shell (an indicator of age)

The data have been checked for missing or impossible values. No entries are missing, but two abalone have a recorded height of 0, a physical impossibility. These two rows have been removed.

The data analysis techniques to be used depend on the supplied data meeting certain conditions:

- **Normality.** When the probability of each value occurring is graphed against the value itself, the curve (known as a density plot) should form the classical bell shape expected of "normal" data.
- **Linearity.** The relationship between each pair of variables should be linear.
- **Homoscedasticity.** The variation (spread) in one variable should remain the same, regardless of the value of another variable.

1.1 Normality

The density plots for all the numerical variables are shown in Figure 1.

All of the variables are skewed; that is, the bell curve is stretched either left or right. This can be repaired, to some extent, by transforming the data i.e. altering the values to force them into a more symmetrical shape.

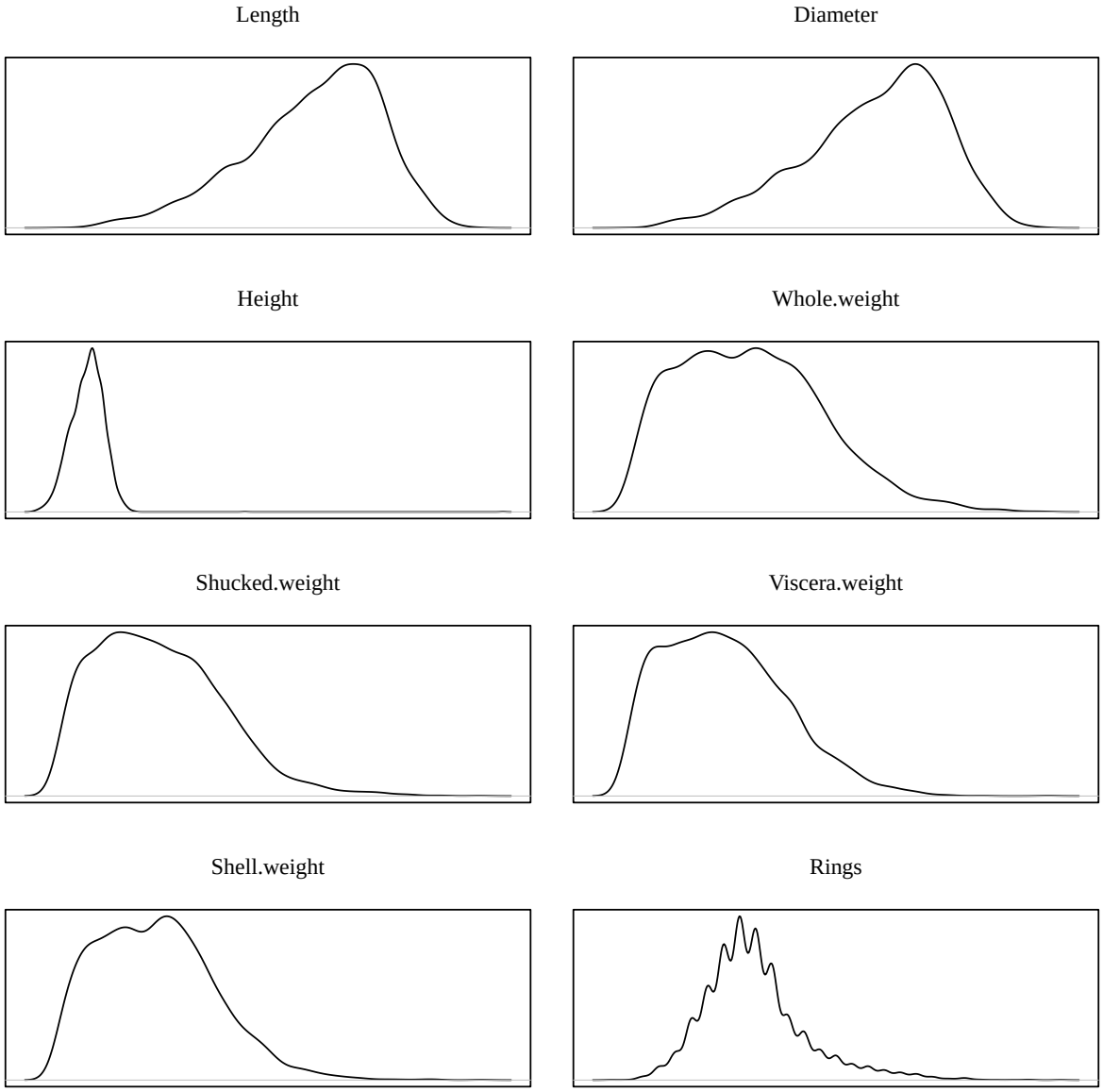

Figure 1: Density Plots

1.2 Linearity and Homoscedasticity

These two requirements can be checked by plotting pairs of the variables against one another, as shown in Figure 2. The following observations can be made:

- The relationships between all the weight variables, and the length and diameter, are non-linear.
- The slightly oval shape of the plots pairing the four weights suggests that these variables are normally distributed. Examining Figure 1 confirms that the four weight variables are the most "normal", with low kurtosis (pointiness) but some skew to the right.

1.3 Data Correction

Transformations have the benefit of pulling outliers (extreme values) in towards the centre, so that they are no longer considered extreme. Then, any remaining outliers may be removed from

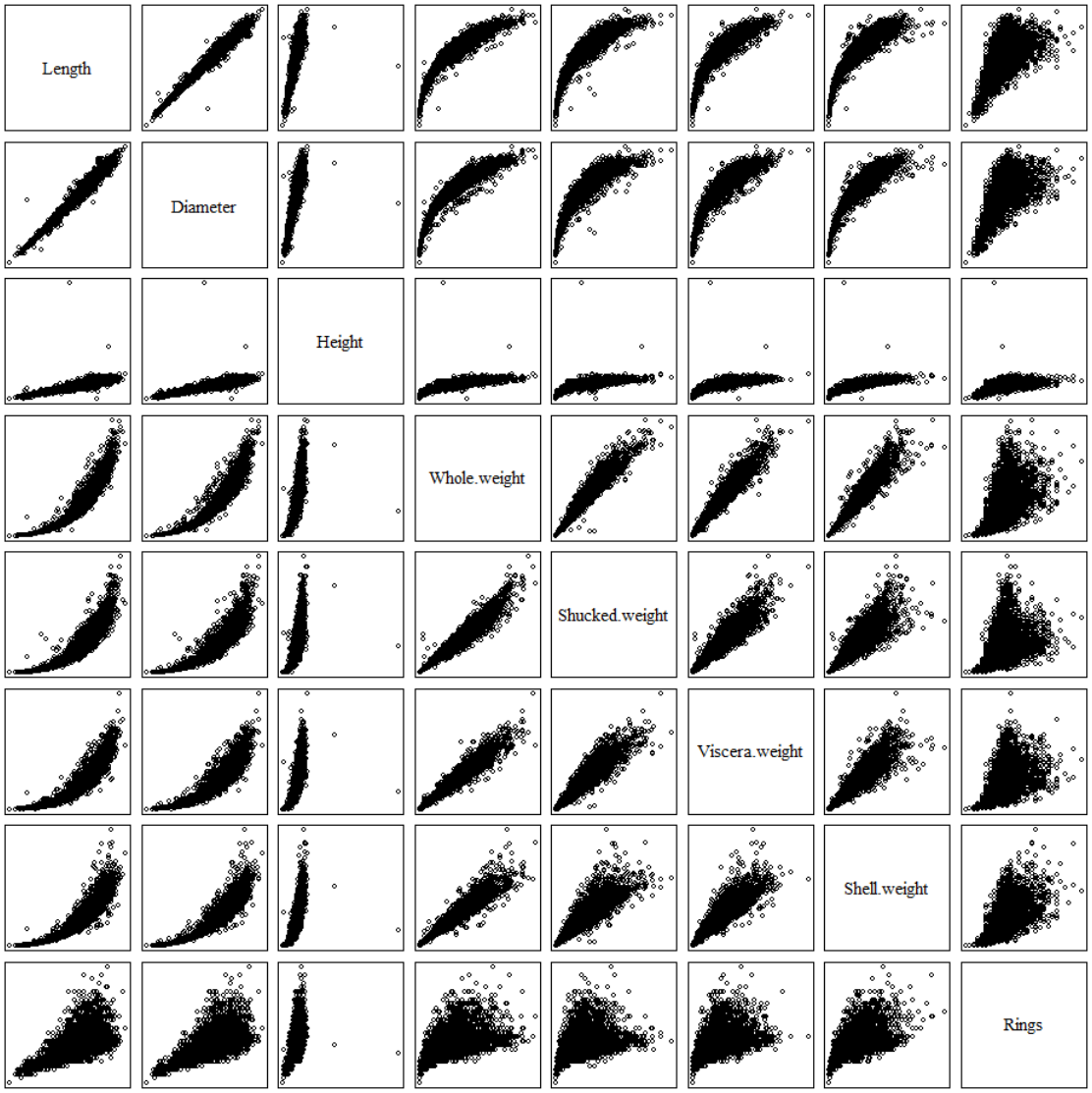

Figure 2: Scatter Plots of Variable Pairs

the dataset. Transformations also help to improve the linearity of the relationships between the variables (XXXX).

Table 1 shows that most outliers have been successfully addressed by transformations, leaving only a small number to be removed from the dataset (with the exception of height). Figure 3 illustrates a substantial improvement in every variable, after applying a transformation and removing the remaining outliers.

Having confirmed, through analysis of skewness and kurtosis, that the variables are now much closer to "normal", data modelling to predict sex and infancy status can now proceed.

2 Predicting Sex and Infancy

The supplied data have been fed into two machine learning algorithms in order to develop a "classifier" that labels an abalone an infant, female or male, based on non-intrusive measurements (length, diameter and height).

The first algorithm involves **discriminant analysis**; the second involves a **support vector**

Variable	Initial Outliers	Remaining Outliers
Length	49	8
Diameter	59	9
Height	27	156
Whole weight	30	1
Shucked weight	48	3
Viscera weight	26	2
Shell weight	35	8
Rings	278	268

Table 1: Outliers Before and After Transformations

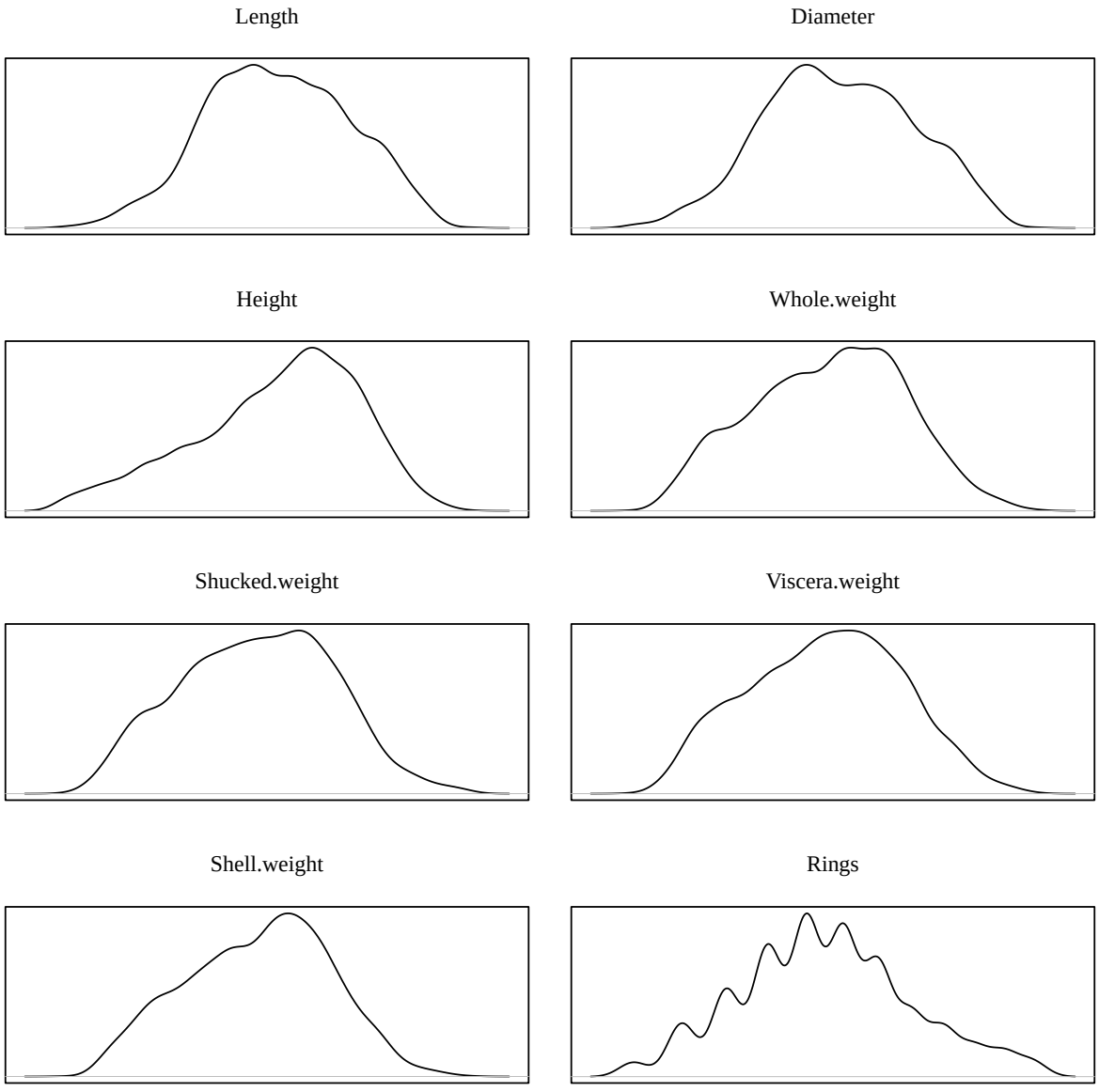

Figure 3: Density Plots After Transformations and Removal of Outliers

machine (SVM). Detailed explanations for these algorithms can be found in XXXX and XXXX.

2.1 Three-Way Classifier

Table 2 reports the accuracy for each of the attempted algorithms.

Model	Accuracy %
quadratic discriminant analysis (QDA)	0.5050
linear discriminant analysis (LDA)	0.5153
linear SVM	0.5147
radial SVM	0.5168

Table 2: Accuracy of Three-Way Classifiers

The overall accuracy for all models is poor, but is slightly better for SVM. Table 3 shows the accuracy of the two SVM models by sex.

Class	Accuracy (Linear) %	Accuracy (Radial) %
Infant	0.7206	0.7121
Female	0	0.0025
Male	0.7820	0.7926

Table 3: Accuracy of SVM Classifiers by Sex and Infancy

Nearly all of the female abalone are labelled incorrectly by these models. By contrast, their accuracy is acceptable (above 70%) for infants, and approaching good (nearly 80%) for males. In order to understand this phenomenon, the three input variables can be grouped by sex and infancy status, as shown in Figure 4.

The boxplots show that the values of length, diameter and height for males and females overlap, while the values for infants are substantially separated from the sexes. This explains why the models are unable to differentiate between males and females.

Because of this, **none** of the models is suitable as a three-way classifier.

2.2 Binary Classifiers

Having failed to specify a three-way classifier, binary classifiers have been attempted, to differentiate between:

- Infants and non-infants (to avoid harvesting infant abalone)
- Females and non-females (to selectively harvest females for increased profits)
- Males and non-males (to selectively harvest males to preserve female populations)

A performance summary for the best models in these scenarios is given in Table 4. While it appears that all three models perform reasonably well, closer inspection indicates otherwise. From the **confusion matrix** shown in Table 5:

- **I vs Not I**: Good at identifying non-infants, but poor at positively identifying infants.
- **F vs Not F**: Classifies nearly all abalone as non-female, regardless of true sex.
- **M vs Not M**: Classifies nearly all abalone as non-male, regardless of true sex.

Therefore, it is recommended that a binary classifier for infants and non-infants **only** be used to identify non-infants, not for positively identifying infants. In addition, it is recommended that a binary classifier **not** be used at all for females and non-females, and males and non-males.

Unfortunately, this means that the objective to use non-intrusive measurements to predict the sex of abalone cannot be met, because males and females are not differentiable by their physical proportions.

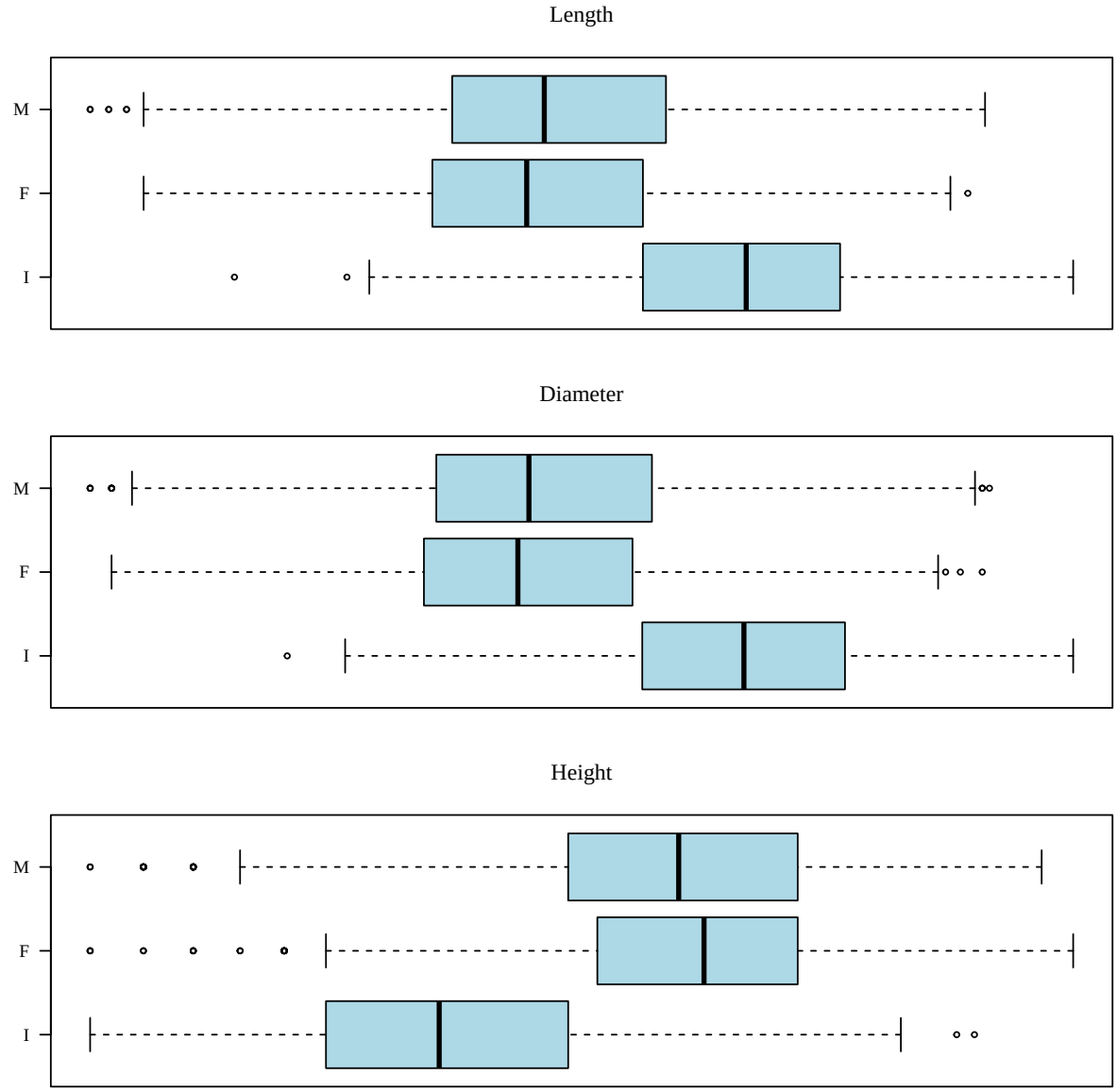

Figure 4: Input Variables by Sex Class

3 Estimating Value

By returning smaller (and likely younger) abalone to the water and allowing them to mature and reproduce, farmers can ensure sustainability of abalone populations, while maximising profits by selectively harvesting larger (presumably older) abalone with more meat.

To this end, a function has been developed by which a farmer can quickly predict the shucked and viscera weights of an abalone, based only on its measurements, and thus estimate the market value. A **multiple linear regression** (MLR) model has been fed summaries of the transformed data for length, diameter, height and the two weights of interest. The resulting prediction equations are:

$$Shucked\_weight\_t = 15.768 - 1.14 * Length\_t - 0.432 * Diameter\_t + 1.164 * Height\_t$$

$$Viscera\_weight\_t = 7.145 - 0.698 * Length\_t - 0.256 * Diameter\_t + 1.691 * Height\_t$$

Classifier	Best Model	Accuracy %
I vs Not I	SVM radial	0.7950
F vs Not F	SVM radial	0.6926
M vs Not M	SVM radial	0.6824

Table 4: Accuracy of Binary Classifiers

	Predicted	Predicted Not
Actual I	675	506
Actual Not I	273	2346
Actual F	84	1122
Actual Not F	46	2548
Actual M	1	1412
Actual Not M	0	2387

Table 5: Confusion Matrix for Binary Classifiers

Figure 5 shows that there is very good agreement between actual abalone weights and the weights predicted by these equations. An alternative model involving **principal component analysis** (XXXX) also yields good results, but the above MLR equations are slightly superior.

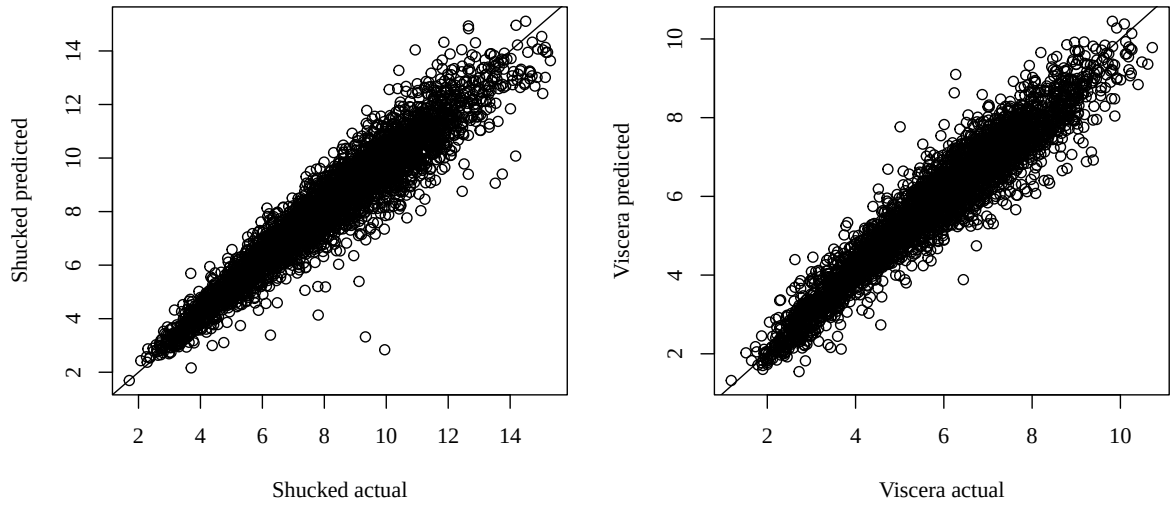

Figure 5: Predicted vs Actual Weights

The weights in the supplied data were originally transformed by taking the square root, so the opposite operation (square) converts them to their original units:

$$Shucked\_weight = Shucked\_weight\_t^2$$

$$Viscera\_weight = Viscera\_weight\_t^2$$

The market price for one abalone is thus estimated:

$$Est\_Price = Shucked\_weight * Price\_per\_gram_{sh} + Viscera\_weight * Price\_per\_gram_{vi}$$

This price estimate is accompanied by a prediction interval computed in accordance with XXXX, the bounds of which are determined by the end user (e.g. for a 90% prediction interval, alpha is set to 0.1). The final function thus has inputs and outputs as shown in Table 6.

Inputs	Outputs
length (mm)	estimated price ($)
diameter (mm)	lower bound ($)
height (mm)	upper bound ($)
price/gram, shucked ($)	
price/gram, viscera ($)	
alpha (default 0.05)	

Table 6: Inputs and Outputs of Price Estimator

Conclusion

The aims of this study have partially been met. Using dimensional measurements alone, it is currently not possible to balance competing interests in sustainability and profitability by selectively harvesting for sex. However, there is scope, through the infant vs non-infant classifier in conjunction with the price estimation function, to improve farming practices by selectively harvesting larger and more mature abalone.