Multivariate Analysis

on Abalone for Sustainability and Profitability

Dylan Wu | z5019314

12/5/22

ZZSC5855 | Multivariate Analysis for Data Scientists

Contents

Executive Summary

Many countries regulate abalone harvesting due to its slow growth rates and high value, thus highlighting the importance of engaging in these practices in a sustainable way whilst simultaneously maximizing profitability. To help the client develop an integrated set of calipers and goggles with advanced measuring capabilities, multivariate analysis was conducted using different classification algorithms to assess the sustainability of harvesting specific types of abalone. Out of the models tested, it was found that support vector machines provided the greatest classification accuracy overall and performed specifically well in classifying infants with a score of 81.1%. Height was also identified as a statistically important physical measurement that helped distinguish between sexes. Additionally, a profitability algorithm was developed predicting the shucked and visceral weights of an abalone based on its physical measurements of length, diameter, and height. The multivariate linear model was able to explain 94.4% of the variance in shucked and visceral weight with the dimensions of length, height, and diameter.

Exploratory Data Analysis

Initial data exploration showed that there were 4,177 abalone with their key summary statistics being recorded in *Table 1.1*. Out of the 9 independent variables available within the dataset, 6 were used to assess the sustainability and profitability of abalone as described in Table *1.2*. Within the dataset, there were 1,528 males, 1,342 infants, and 1,307 females which represented 37%, 32%, and 31% of the abalone sample respectively (*Figure 1.3)*.

Data Preprocessing

After cleaning the dataset and removing abnormalities such as null values, outliers, and records where shucked weight was greater than the total weight of the abalone, there were 4,086 observations remaining. The data was separated into training and testing splits to validate the performance of the model where 80% of the data was used for training the algorithm and 20% of the data was used to test the performance of the model. Furthermore, the data was transformed and scaled to ensure that it met the normality assumptions needed for specific models to be executed.

Hypothesis Testing

Four tests were conducted to determine whether the abalone data met the normality assumptions required for model building. These included:

1. Homogeneity of covariance – whether the variances across different sexes were the same across length, diameter, and height
2. Multivariate Normality – whether the physical measurements were normal for each sex
3. Multicollinearity – whether the predictive power decreased in accordance with an increased correlation between the predicting variables of length, height, and diameter
4. Independence – whether the abalone was randomly and independently sampled.

Mardia's multivariate skewness and kurtosis measures were used to verify the multivariate normality assumption of the abalone dataset alongside Shapiro-Wilk's tests for univariate normality. The results of these assessments have been provided in *Tables 1.4* and *1.5*. Whilst all these tests were failed leading to some of the normality assumptions being violated, the scatter plots in *Figures 1.5* and *1.6,* revealed an elliptical appearance expected from normal populations. Additionally, the histograms closely resembled the shape of a normal distribution, and the Q-Q plots were fitted mostly on the diagonal line with minor deviation on the tail ends (*Figure 1.7)*. With these results, the data was considered adequate for our normality assumption.

Modelling

Whilst numerous models were experimented with, the best performing models have been included in this report for the company to consider.

Multiclass Classification Model

Three main models including quadratic discriminant analysis, support vector machines, and logistic regression were compared to find the best performing sex classification models. Given that covariances were unequal as found during hypothesis testing, it was deducted that quadratic discriminant analysis may be more suitable for classification over linear discriminant analysis because sex was not linearly separable. As shown in *Table 1.8*, hyperparameter tuning was applied to the support vector machine models and cross validation was applied across all models to prevent overfitting alongside a more accurate estimate of out-of-sample accuracy.

Quadratic Discriminant Analysis

The quadratic discriminant analysis model is a generative model that assumes each class follows a normal distribution and has a different covariance matrix. It uses a discriminant function to assign the abalone into two or more sexes.

Support Vector Machines

The support vector machine model took the abalone observations and outputted them onto a hyperplane to find the decision boundaries (the lines or non-linear boundaries that would best separate the sexes). Its ability to take data as input and transform them into the required form for processing, provided substantial flexibility in creating the decision boundaries, leading to greater classification performance.

Logistic Regression

The logistic regression model was based on modelling the probability that an abalone belonged to a specific sex using a default threshold of 0.5, where anything greater than 0.5 was identified as belonging to the target class and anything below to the other class.

Multivariate Regression Model

Before implementing the multivariate regression model, a Box-Cox transformation was applied to shucked and visceral weight to convert them into a normal shape. After removing additional multivariate outliers identified by Mahalonobi’s distance metric, there were 3,619 abalone observations in the dataset that were used in the profitability algorithm. The algorithm was based on a multivariate regression model and its output is described in *Tables 3.7 and 3.8* as well as the following equations:

$$weight_{shucked} = 12.67 + 3.28 * length^2 + 0.99 * diameter^2 + 0.51 * height$$

$$weight_{viscera} = 8.64 + 1.94 * length^2 + 0.65 * diameter^2 + 0.95 * height$$

The precomputed summaries of the multivariate regression model above were used to produce a prediction interval of the profitability algorithm. It used the predicted weights from the regression as well as the dollar value of 1 gram of shucked weight and the dollar value of 1 gram of viscera weight in the following equation:

$$Abalone\ Price = value_{shucked} * weight_{shucked} + value_{viscera} * \ weight_{viscera}$$

Results

Confusion matrices were used to assess the performance of different classification models. The Receiver Operating Characteristic (ROC) curve^ was used to visually summarise the performance of the classification models over all possible thresholds and the higher the Area Under the Curve (AUC) score was, the better the model was in predicting the specified sex.

Classification

The classification models struggled with distinguishing between all three sexes as well as females versus non-females and males versus non-males mainly because target classes were overlapping as illustrated in *Figure 1.6*. However, they performed well when identifying infants as revealed in the ROC curves in *Figures 2.5* and *2.6*. Support vector machines displayed the consistently strongest performance amongst the three models, highlighting their effectiveness in high dimensional spaces.

Predicting the Sex of Abalone

The best performing model was the support vector machine (SVM) which achieved a classification accuracy of 53.3% which was 1.6% and 0.9% better than quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA) respectively. The confusion matrix which depicts the performance of the SVM classification model such as the number of sexes predicted that were correct has been provided in *Table 2.1*.

Predicting Infants

While all models performed well with classification accuracies above 80%, the best performing model was the SVM with 81.1% accuracy as shown in *Table 2.3*. The QDA model correctly predicted the greatest number of infants but underperformed in correctly predicting the other sexes, misclassifying 89 as infants in *Table 2.2*.

Predicting Females

The best performing model was the SVM which achieved a classification accuracy of 68.5% which was more than double that of the logistic regression and 0.9% greater than the QDA model.

Predicting Males

The best performing model was the logistic regression which achieved a classification accuracy of 65% which was almost the same as the SVM model which had an accuracy of 64.9%.

Of the three physical measurements, height was the most important variable for distinguishing between sexes as it was the only statistically significant variable across all three logistic regression models. Length was statistically insignificant in determining whether the sex of an abalone was specifically male or female.

Prediction

The predictive model demonstrated its effectiveness with a high R-squared of 0.944 and relatively small residual standard errors of 1.063 and 7.057 (*Table 3.7 and 3.8)*. Adjusting for the number of variables

^Note that the support vector machine does not produce a predicted probability and thus an ROC curve could not be constructed for these models.

considered, the multivariate linear model revealed that 94.4% of the variance observed in shucked weight and viscera weight can be explained by length, height, and diameter. The residuals of the multivariate regression model appeared to be mostly normal (*Figure 3.9)* and randomly distributed with some slight deviation on the tail end of the Q-Q plot as shown in *Figure 4.0*. Hence, it was a good fit with the abalone data.

With the prediction interval, the model estimated that 90% of the time, an abalone with average measurements such as 27.91mm in height, 104.8mm in length, and 81.6mm in diameter has a market value between $727.55 and $757.90, given that the price is $1/gram of shucked weight and $1/gram of viscera weight.

Recommendations

It is recommended that the company develop diving technology that integrates classification algorithms based on support vector machine models because it will enable divers to accurately assess whether an abalone is an infant or not. However, they should avoid extending its usage to male or female detection and instead experiment with other classification algorithms such as K-Nearest Neighbours, Naiive Bayes, or Decision Trees to achieve better performance in these areas. This improvement will be subject to the computing powers available given these algorithms become significantly slower if the number of examples or predictor variables increase.

In terms of predicting the weight of an abalone, the multivariate regression model was very effective in predicting the shucked weight and viscera weight. It is recommended that the company incorporate the precomputed summary of the model which creates a prediction interval of the value of an abalone based on weights in their next generation product. The low standard error and ability to capture a majority of the variance in weights, will help divers make the instantaneous decision of whether to harvest an abalone or leave it alone for the time being, thereby meeting the profitability and sustainability requirements of the company.

Conclusion

Our findings highlighted that multivariate analysis models have the potential to assist in biodiversity and that predictions can be precomputed to extract information without computational constraints. Given the excellent performance of support vector machines with certain abalone and poor performance with others, future technological development should expand to include other modern machine learning algorithms. During data analysis, there were physical measurements demonstrating a strong relationship with sex, inferring that the incorporation of other physical measurements such as rings or age could outperform the support vector machine model.

Appendix

Table 1.1 Description of Abalone Data

Name	Data Type	Measurement Unit	Description
Sex	Nominal	--	Male (M), Female (F), and Infant (I)
Length	Continuous	mm	Longest shell measurement
Diameter	Continuous	mm	Perpendicular to length
Height	Continuous	mm	With meat in shell
Shucked Weight	Continuous	grams	Weight of meat
Viscera Weight	Continuous	grams	Gut weight (after bleeding)

Table 1.2 Summary Statistics of Abalone Data

Variable	Mean	Median	Standard Deviation	Range
Length	104.80	109.00	24.00	148.00
Diameter	81.60	85.00	19.83	119.00
Height	27.91	28.00	8.36	226.00
Shucked Weight	71.87	67.20	44.41	297.40
Viscera Weight	36.15	34.20	21.92	151.90

Figure 1.3 Frequency Distribution of Abalone Sex

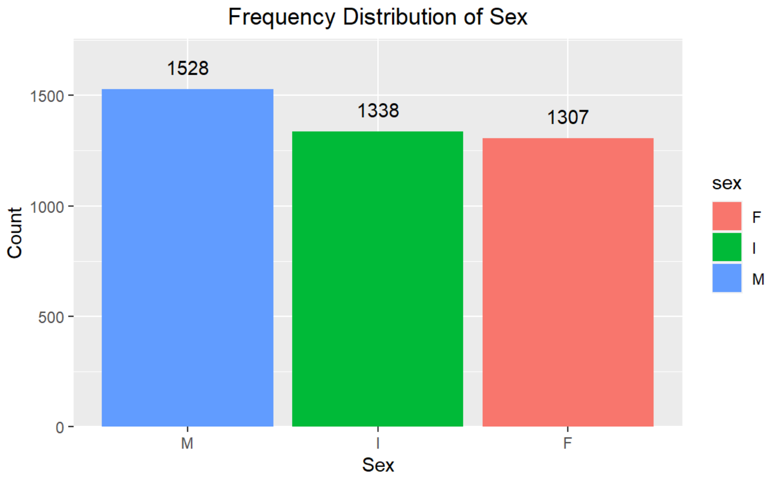


Table 1.4 Hypothesis tests ***before*** *transforming and scaling the abalone*

Assumptions	Tests	p-value	Result
Covariance Equality	Box’s M	2.2e-16	No
Multivariate Normality	Mardia Skewness	0	No
	Mardia Kurtosis	0	No
Univariate Normality	Shapiro-Wilk (Height)	<0.001	No
	Shapiro-Wilk (Length^2)	<0.001	No
	Shapiro-Wilk (Diameter^2)	<0.001	No

Table 1.5 Hypothesis tests ***after*** *transforming and scaling the abalone*

Assumptions	Tests	p-value	Result
Covariance Equality	Box’s M	2.2e-16	No
Multivariate Normality	Mardia Skewness	0	No
	Mardia Kurtosis	0	No
Univariate Normality	Shapiro-Wilk (Height)	<0.001	No
	Shapiro-Wilk (Length^2)	<0.001	No
	Shapiro-Wilk (Diameter^2)	<0.001	No

Figure 1.6 A pairwise plot matrix of transformed and scaled abalone for classification modelling

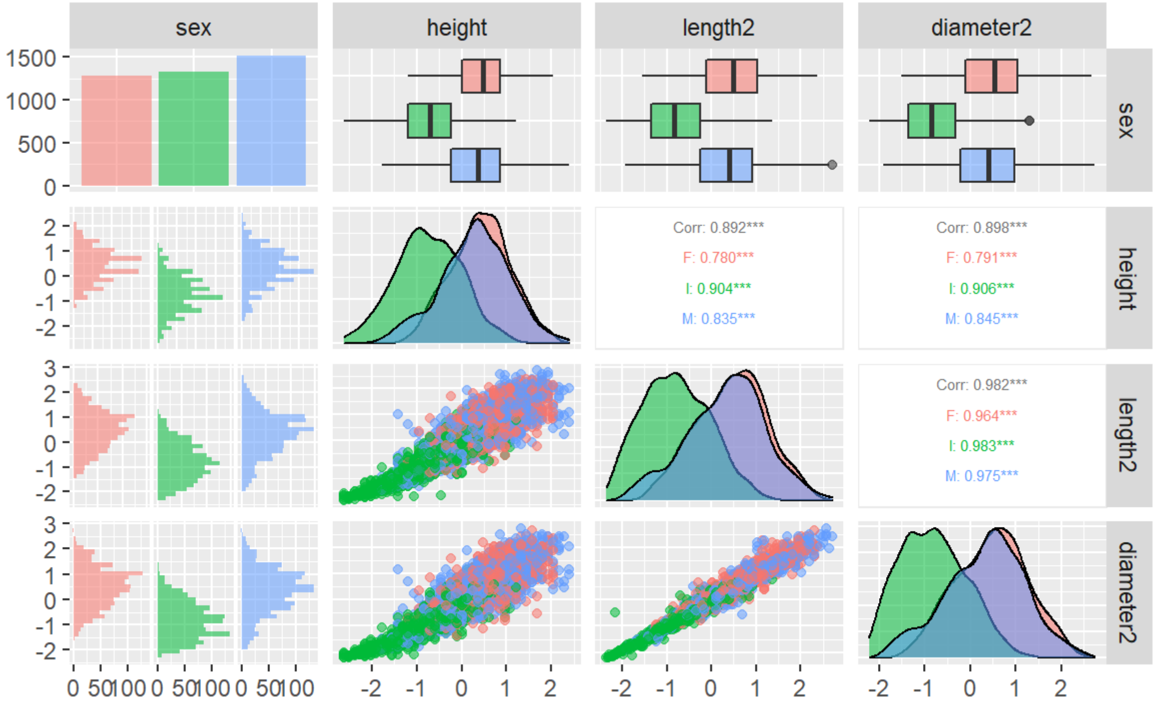


Figure 1.6 A pairwise plot matrix of transformed and scaled abalone for multivariate regression

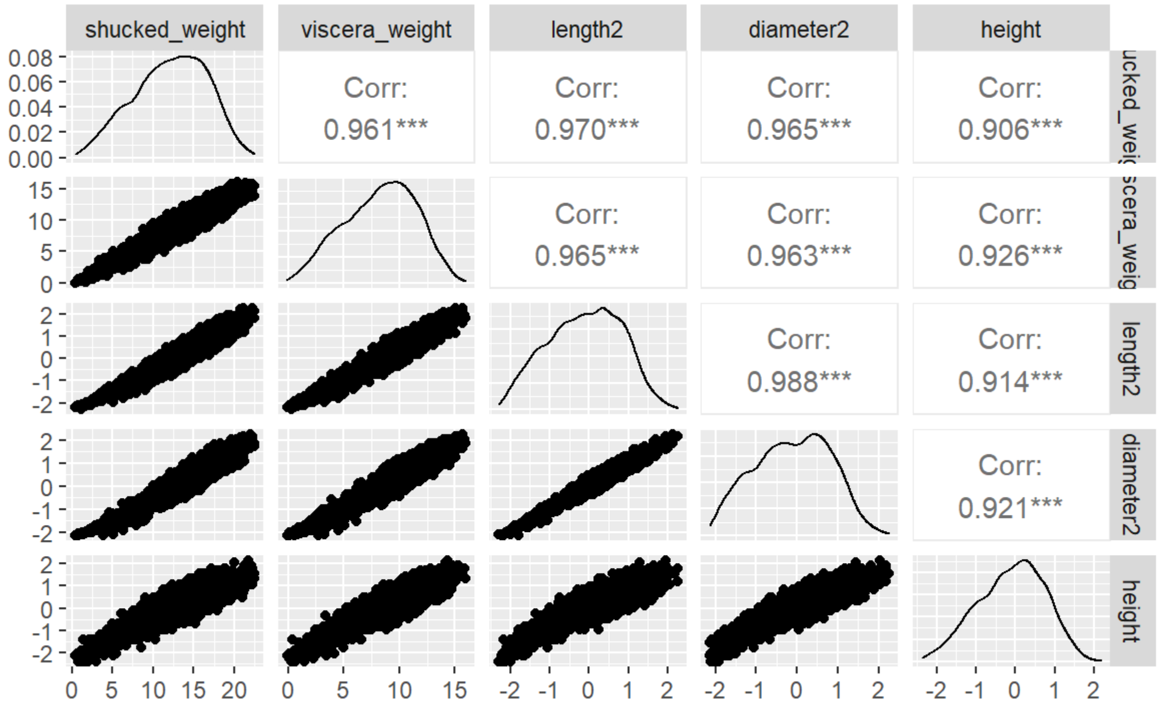


Figure 1.7 Univariate plots of the physical measurements and weights of abalone

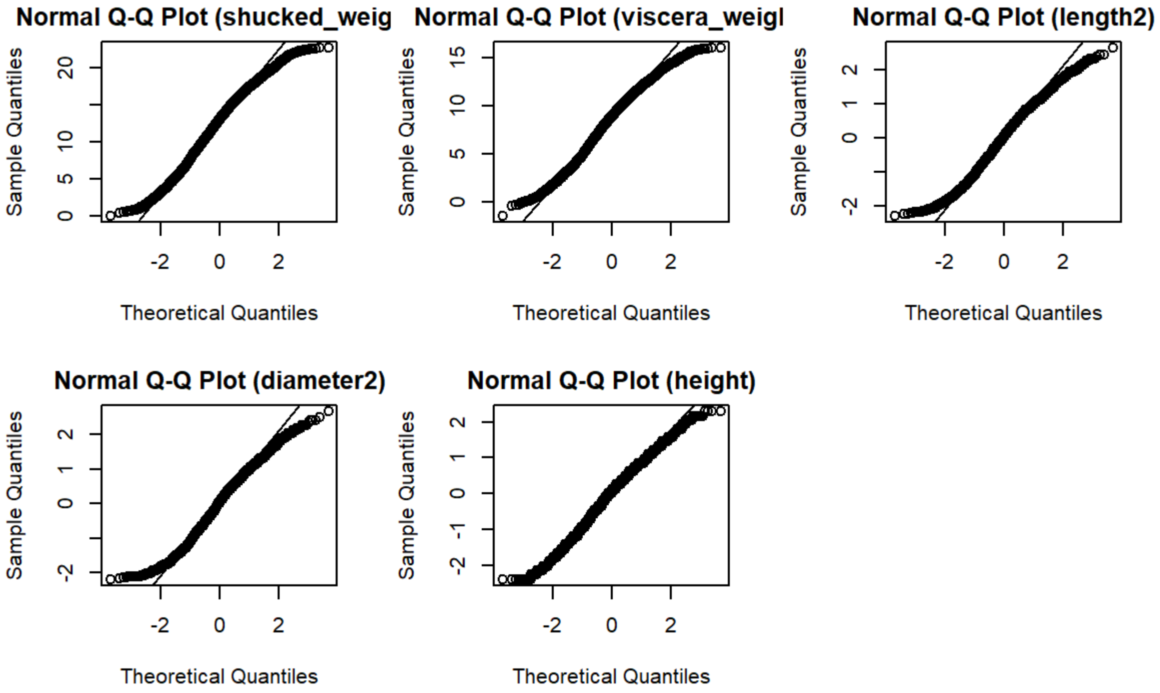


Table 1.8 Parameter tuning of SVM using 10-fold cross validation

gamma	cost	error	dispersion
0.1	0.1	0.4773947	0.02399332
1.0	0.1	0.4744779	0.02549835
10.0	0.1	0.4771549	0.01887699
0.1	1.0	0.4764244	0.02259902
1.0	1.0	0.4737486	0.02390483
10.0	1.0	0.4822508	0.01789743
0.1*	**10.0**	**0.4727730**	**0.02463882**
1.0	10.0	0.4747242	0.02144611
10.0	10.0	0.4851593	0.02690557

***Best parameters were gamma = 0.1, cost = 10**

Table 1.9 Confusion matrix of the linear discriminant analysis model for predicting abalone sex

	Prediction		
Truth	**Female**	**Infant**	**Male**
Female	270	200	811
Infant	5	970	347
Male	281	315	917
Accuracy: 52.4%			

Table 2.0 Confusion matrix of the quadratic discriminant analysis model for predicting abalone sex

	Prediction		
Truth	**Female**	**Infant**	**Male**
Female	296	241	744
Infant	47	1,016	259
Male	320	374	819
Accuracy: 51.8%			

Table 2.1 Confusion matrix of the support vector machine model for predicting abalone sex

	Prediction		
Truth	**Female**	**Infant**	**Male**
Female	110	250	921
Infant	0	1,023	299
Male	81	371	1,061
Accuracy: 53.3%			

Table 2.2 Confusion matrix of the quadratic discriminant analysis model for predicting infants

	Prediction	
Truth	Others	Infant
Others	484	89
Infant	79	200
Accuracy: 80.3%		

Table 2.3 Confusion matrix of the support vector machine model for predicting infants

	Prediction	
Truth	Others	Infant
Others	513	60
Infant	101	178
Accuracy: 81.1%		

Table 2.4 Confusion matrix of the logistic regression model for predicting infants

	Prediction	
Truth	Others	Infant
Others	543	30
Infant	132	147
Accuracy: 81.0%		

Figure 2.5 ROC and AUC of the logistic regression model predicting infants

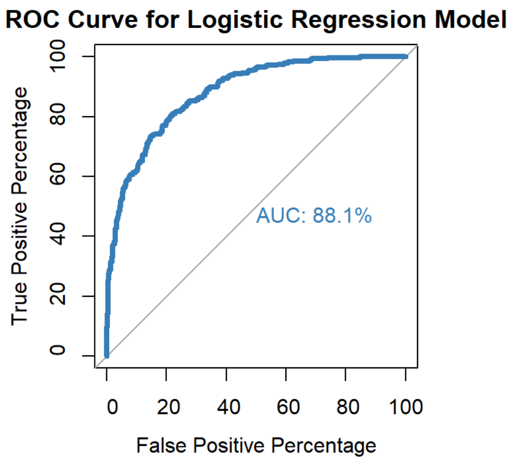


Figure 2.6 ROC and AUC of the quadratic discriminant analysis model predicting infants

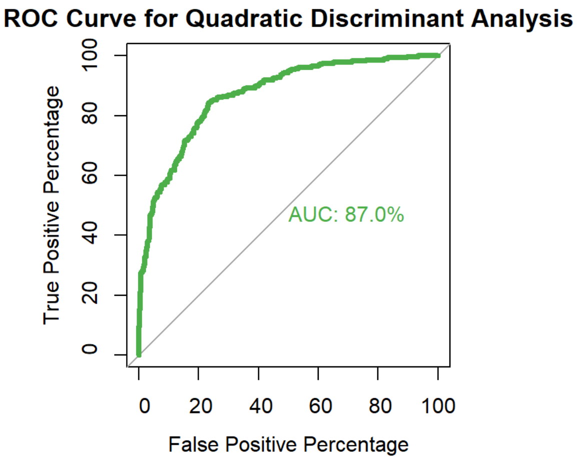


Table 2.7 Confusion matrix of the quadratic discriminant analysis model for predicting females

	Prediction	
Truth	**Female**	**Others**
Female	76	193
Others	83	500
Accuracy: 67.6%		

Table 2.8 Confusion matrix of the support vector machine model for predicting females

	Prediction	
Truth	**Female**	**Others**
Female	10	259
Others	9	574
Accuracy: 68.5%		

Table 2.9 Confusion matrix of the logistic regression model for predicting females

	Prediction	
Truth	**Female**	**Others**
Female	129	140
Others	436	147
Accuracy: 32.4%		

Figure 3.0 ROC and AUC of the logistic regression model predicting infants

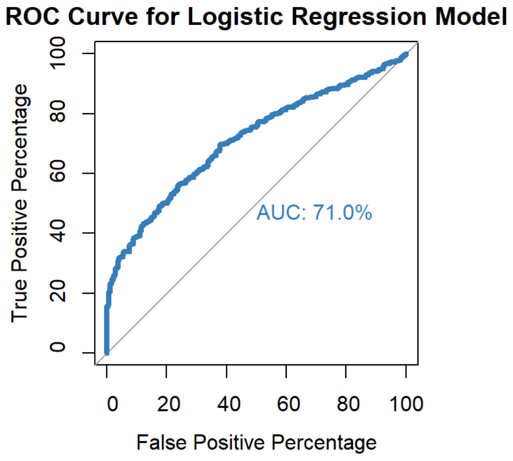


Figure 3.1 ROC and AUC of the quadratic discriminant analysis model predicting infants

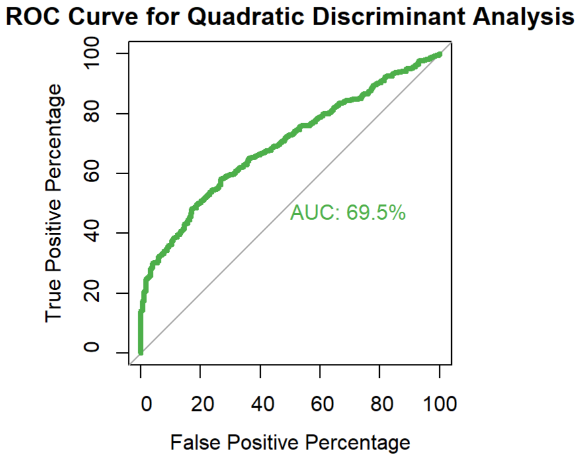


Table 3.2 Confusion matrix of the quadratic discriminant analysis model for predicting males

	Prediction	
Truth	**Others**	**Male**
Others	459	89
Male	222	82
Accuracy: 63.5%		

Table 3.3 Confusion matrix of the support vector machine model for predicting males

	Prediction	
Truth	Others	Male
Others	544	4
Male	295	9
Accuracy: 64.9%		

Table 3.4 Confusion matrix of the logistic regression model for predicting males

	Prediction	
Truth	Others	Male
Others	544	4
Male	294	10
Accuracy: 65.0%		

Figure 3.5 ROC and AUC of the logistic regression model predicting males

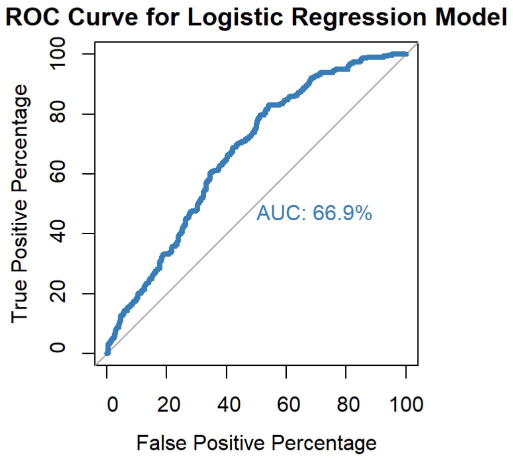


Figure 3.6 ROC and AUC of the quadratic discriminant analysis model predicting males

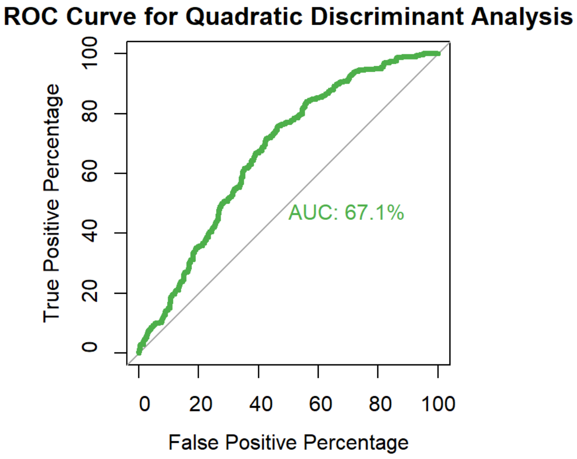


Table 3.7 Coefficients of the shucked weight multivariate linear model

	Estimate	Standard Error	Test Statistic	P-value
(Intercept)	12.66746	0.01782	711.015	2e-16
Length2	3.22707	0.12491	26.235	2e-16
Diameter2	0.99177	0.13103	7.569	4.74e-14
Height	0.51372	0.05404	9.505	2e-16
Residual standard error: 1.063 on 3,615 degrees of freedom				
Multiple R-squared: 0.9436				
Adjusted R-squared: 0.9435				

Table 3.8 Coefficients of the viscera weight multivariate linear model

	Estimate	Standard Error	Test Statistic	P-value
(Intercept)	8.64068	0.01283	673.382	2e-16
Length2	1.93997	0.08997	21.563	2e-16
Diameter2	0.64658	0.09437	6.852	8.55e-12
Height	0.95466	0.03892	24.526	2e-16
Residual standard error: 0.7657 on 3,615 degrees of freedom				
Multiple R-squared: 0.944				
Adjusted R-squared: 0.944				

Figure 3.9 Residual diagnostics of shucked and viscera weight from multivariate regression

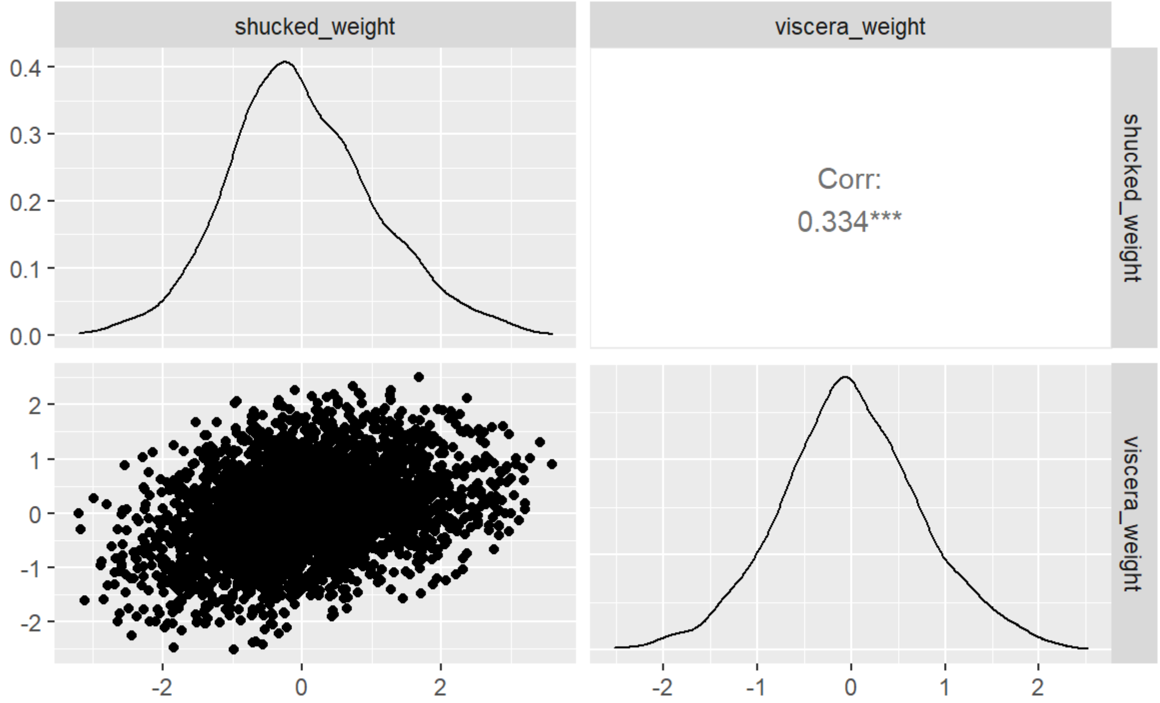
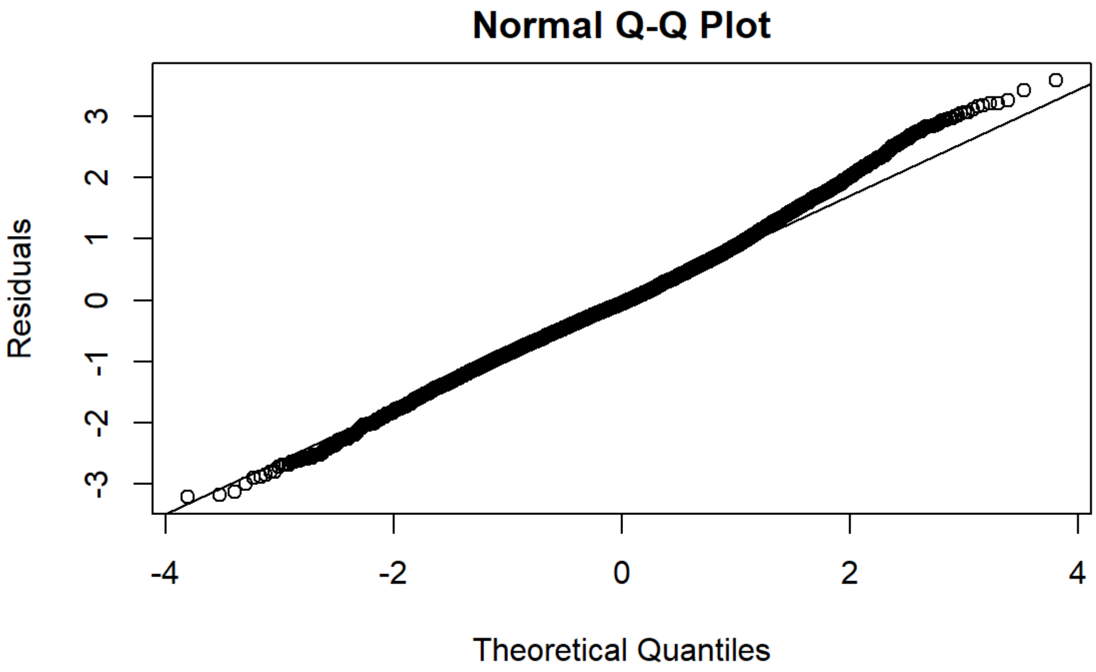


Figure 4.0 Normal Q-Q plot of residuals from multivariate regression



References

Alice, M. 2015, ‘How to perform a Logistic Regression in R’, *R-bloggers*, 13 September, viewed 1 December 2022, <https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/>.

Berman H.B. 2022, ‘Sums of Squares and Cross Products Matrix’, *Stat Trek*, viewed 29 November 2022, <https://stattrek.com/matrix-algebra/sums-of-squares>.

Date, S., ‘Time Series Analysis, Regression and Forecasting’, *timeseriesreasoning*, viewed 5 December 2022, <https://timeseriesreasoning.com/contents/overview-of-the-variance-covariance-matrices-used-in-linear-regression/>.

Dua, D. & Graff, C. 2019, ‘UCI Machine Learning Repository’, *University of California, Irvine, School of Information and Computer Sciences*, viewed 25 November 2022, <http://archive.ics.uci.edu/ml>.

Helwig, N.E. 2017, ‘Multivariate Linear Regression’, *University of Minnesota*, 16 January, viewed 3 December 2022, <http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf>.

Jim 2022, ‘Box Cox transformation in R’, *R-bloggers*, 23 October, viewed 3 December 2022, <https://www.r-bloggers.com/2022/10/box-cox-transformation-in-r/#:~:text=Box%20Cox%20transformation%20in%20R%2C%20The%20Box%2DCox%20transformation%20is,distribution%20is%20therefore%20very%20helpful.l>.

Johnson, R.A. & Wichern, D. 2008, ‘Applied Multivariate Statistical Analysis’, *Pearson Prentice Hall*.

Kassambara, A. 2020, ‘Transform Data to Normal Distribution in R’, *Datanovia*, viewed 4 September 2022, <https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/>.

Korkmaz, S., Goksuluk, D. & Zararsiz G. 2021, ‘MVN: An R Package for Assessing Multivariate Normality’, 29 June, viewed 27 November 2022, <https://cran.r-project.org/web/packages/MVN/vignettes/MVN.html>.

Le, J. 2018, ‘Support Vector Machines in R’, *DataCamp*¸ August, viewed 2 December 2022, <https://www.datacamp.com/tutorial/support-vector-machines-r>.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. 2014, ‘Cluster Analysis’, *GitHub*, viewed 30 November 2022, <https://easystats.github.io/parameters/reference/cluster_analysis.html>.