Multivariate Analysis - Abalone Dataset

ZZSC5855 FINAL PROJECT | THOMAS FRANK (5385080) | DATE: 04/10/2024

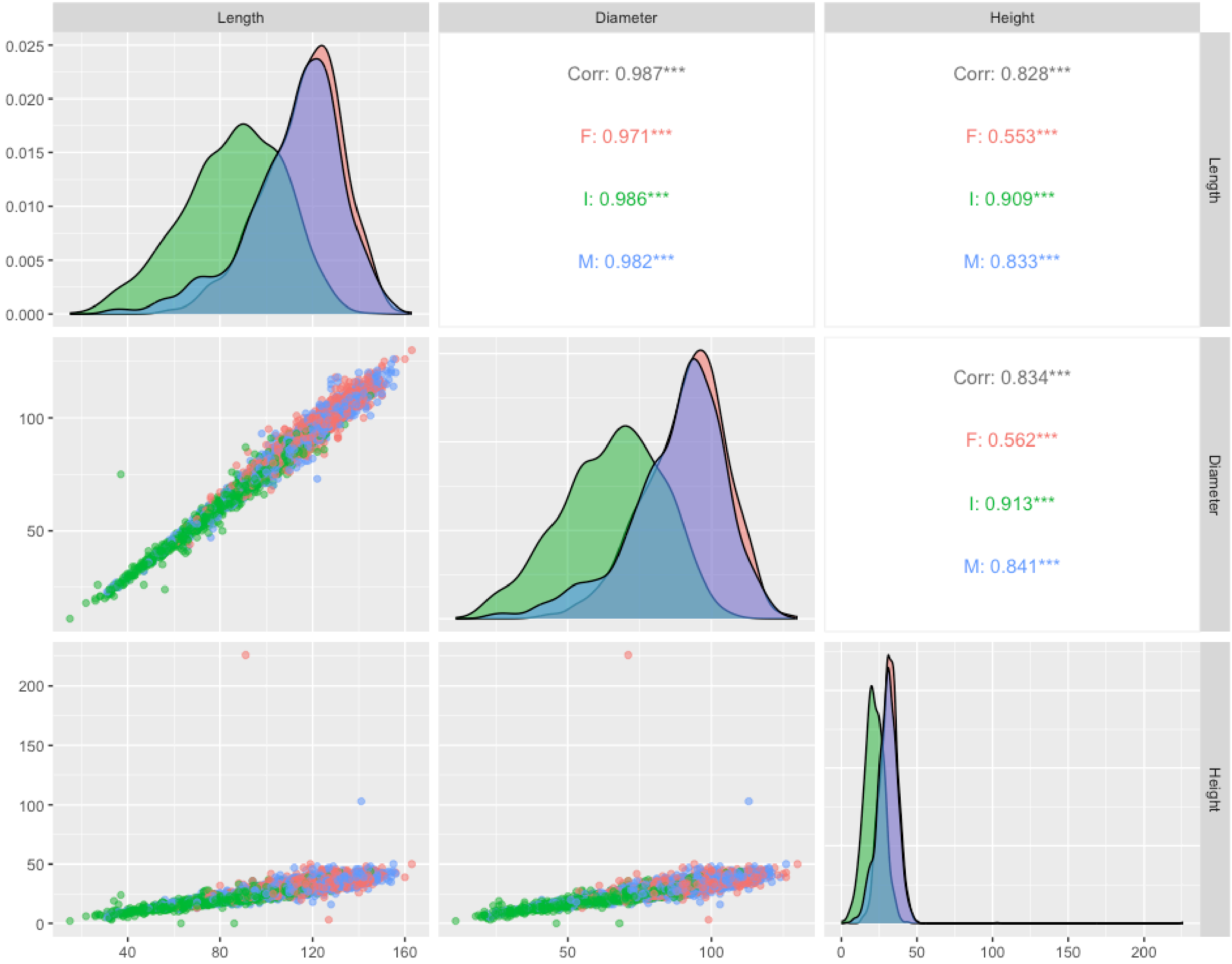
Question 1: Sustainability

The objective of this analysis is to develop methods for predicting the sex of abalone based on exterior measurements (length, diameter, and height), focusing on:

1. **General Sex Prediction:** Predicting whether an abalone is Male, Female, or Infant.
2. **Infant Detection:** Avoiding the harvesting of infants by specifically identifying them.
3. **Female Prediction:** Identifying females for potential profitability.
4. **Male Prediction:** Identifying males, prioritising sustainability by focusing on the preservation of females and infants.

Exploration:

The initial data exploration reveals that Infant abalones are likely to be more easily distinguishable from adult abalones based on their exterior measurements, as they tend to cluster at the lower end of the dataset. In contrast, distinguishing between Male and Female abalones may be more challenging, as their measurements overlap significantly in the upper ranges.

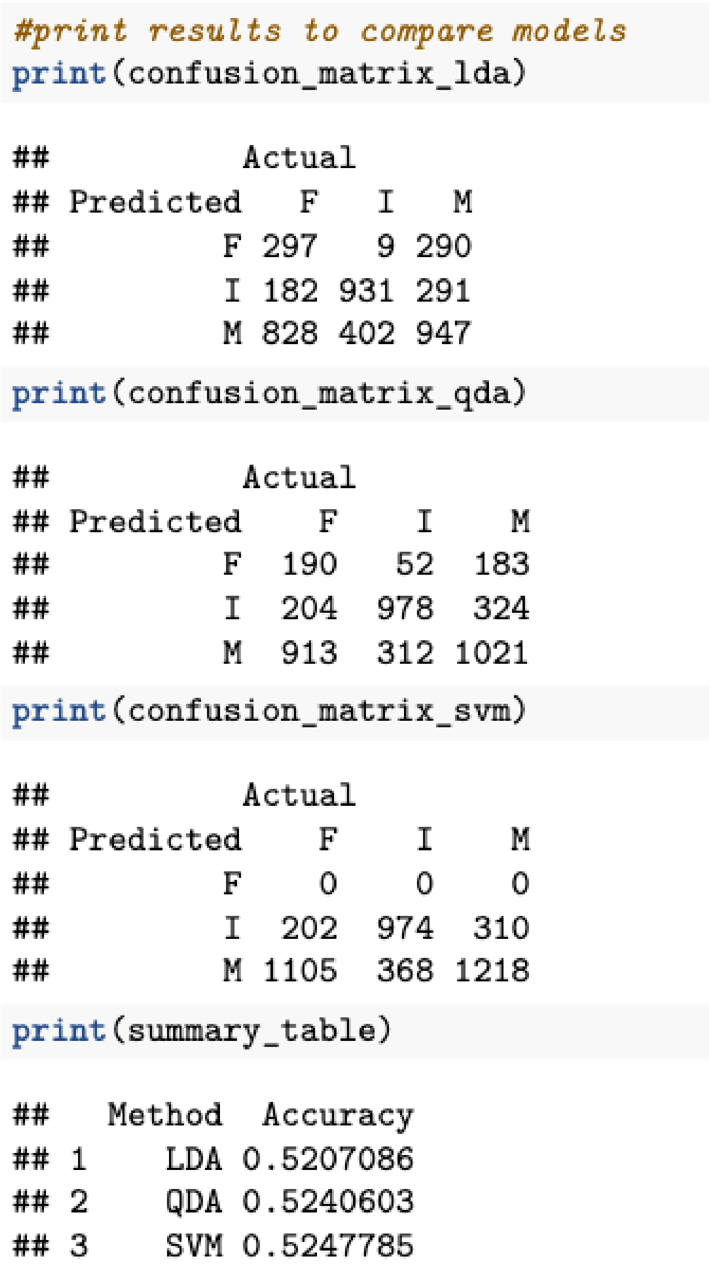


Methods:

- **Discriminant Analysis (LDA & QDA)**: LDA and QDA classify abalone based on exterior measurements by finding boundaries between sexes. LDA is effective for similar group spreads, while QDA handles differing covariances, making them both suitable for distinguishing between male, female, and infant abalones.
- **Support Vector Machines (SVM)**: SVM, with a linear kernel, is useful for the abalone dataset as it aims to identify optimal boundaries between abalone sexes, which is particularly valuable when the relationship between exterior measurements and sex classification is complex.

Results:

1. General Sex Prediction:



```
#print results to compare models
print(confusion_matrix_lda)
```

```
##          Actual
## Predicted   F   I   M
##         F 297   9 290
##         I 182 931 291
##         M 828 402 947
```

```
print(confusion_matrix_qda)
```

```
##          Actual
## Predicted    F    I    M
##         F  190   52  183
##         I  204  978  324
##         M  913  312 1021
```

```
print(confusion_matrix_svm)
```

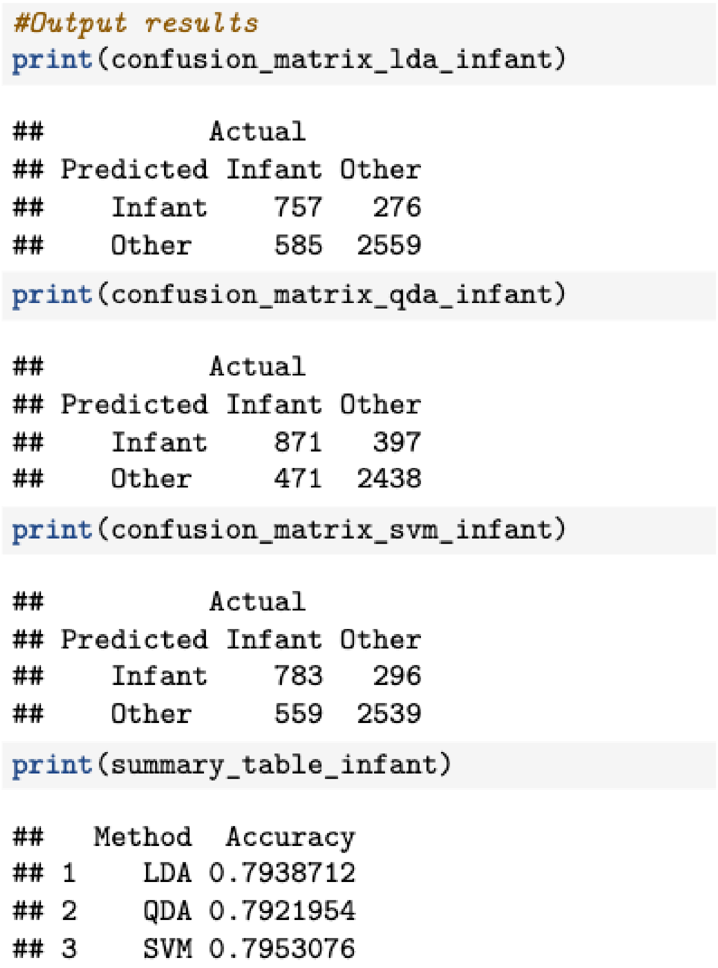
```
##          Actual
## Predicted    F    I    M
##         F    0    0    0
##         I  202  974  310
##         M 1105  368 1218
```

```
print(summary_table)
```

```
##   Method  Accuracy
## 1    LDA 0.5207086
## 2    QDA 0.5240603
## 3    SVM 0.5247785
```

Both LDA and QDA yielded moderate prediction accuracies, around **52% for LDA and 52.4% for QDA**, indicating that while exterior measurements do provide some information about sex, they may not be the sole determinants. SVM performed similarly but mis-classified all females.

2. Infant Classification:



```
#Output results
print(confusion_matrix_lda_infant)
```

```
##           Actual
## Predicted Infant Other
##    Infant    757   276
##    Other     585  2559
```

```
print(confusion_matrix_qda_infant)
```

```
##           Actual
## Predicted Infant Other
##    Infant    871   397
##    Other     471  2438
```

```
print(confusion_matrix_svm_infant)
```

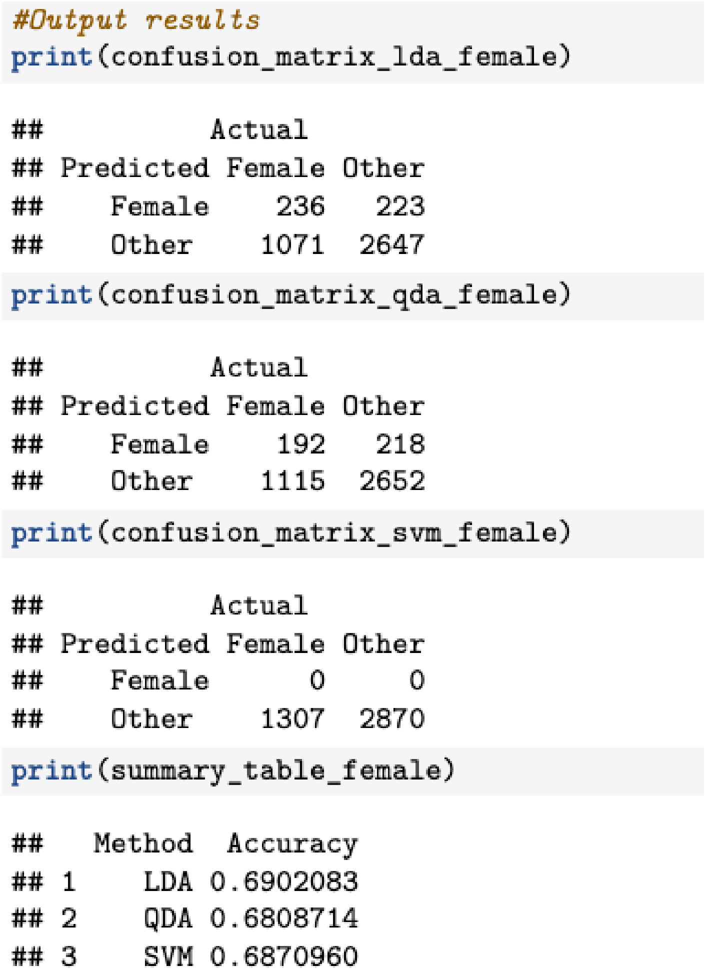
```
##           Actual
## Predicted Infant Other
##    Infant    783   296
##    Other     559  2539
```

```
print(summary_table_infant)
```

```
##   Method  Accuracy
## 1    LDA 0.7938712
## 2    QDA 0.7921954
## 3    SVM 0.7953076
```

For the binary classification between infants and others, LDA and SVM yielded accuracies of **79.4% and 79.5%** respectively. These results suggest that discriminating infants from other abalone based on size measurements is feasible with fairly high accuracy.

3. Female Classification:



```
#Output results
print(confusion_matrix_lda_female)
```

```
##          Actual
## Predicted Female Other
##    Female    236   223
##    Other    1071  2647
```

```
print(confusion_matrix_qda_female)
```

```
##          Actual
## Predicted Female Other
##    Female    192   218
##    Other    1115  2652
```

```
print(confusion_matrix_svm_female)
```

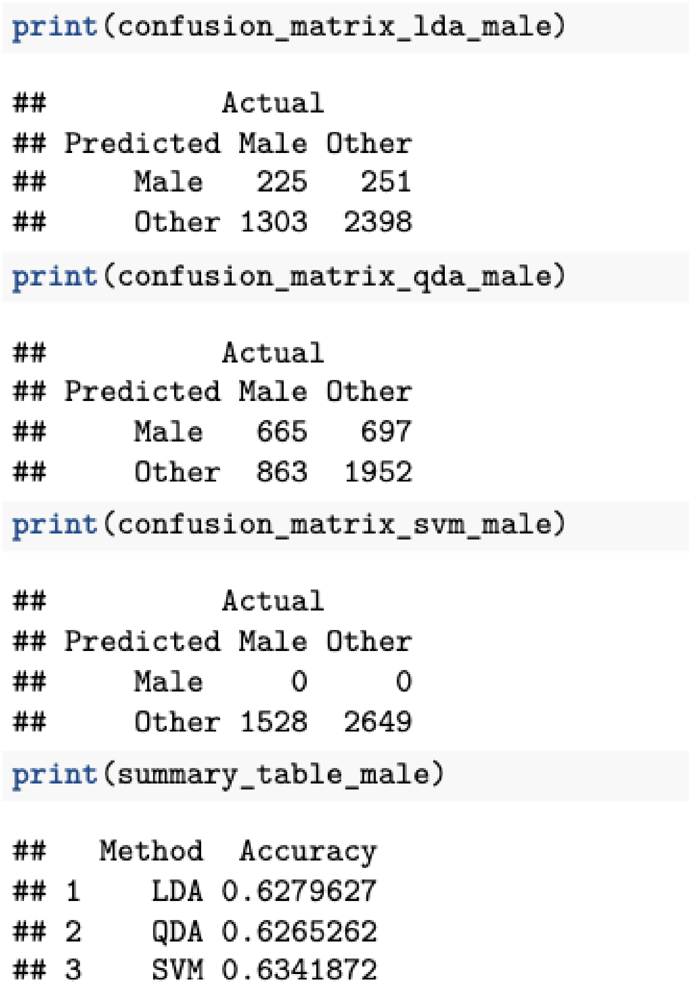
```
##          Actual
## Predicted Female Other
##    Female      0     0
##    Other    1307  2870
```

```
print(summary_table_female)
```

```
##   Method  Accuracy
## 1    LDA 0.6902083
## 2    QDA 0.6808714
## 3    SVM 0.6870960
```

When predicting females, LDA yielded an accuracy of **69%**, while QDA slightly underperformed at **68%**. Although SVM performed comparably with an accuracy of **68.7%**, we can see from the confusion matrix that SVM is unsuitable as it misclassifies all Females.

4. Male Classification:



```
print(confusion_matrix_lda_male)
```

```
##          Actual
## Predicted Male Other
##     Male   225   251
##     Other 1303  2398
```

```
print(confusion_matrix_qda_male)
```

```
##          Actual
## Predicted Male Other
##     Male   665   697
##     Other  863  1952
```

```
print(confusion_matrix_svm_male)
```

```
##          Actual
## Predicted Male Other
##     Male     0     0
##     Other 1528  2649
```

```
print(summary_table_male)
```

```
##   Method  Accuracy
## 1    LDA 0.6279627
## 2    QDA 0.6265262
## 3    SVM 0.6341872
```

For identifying males, LDA achieved an accuracy of **63%**, while QDA reached a similar accuracy of **62.7%**. Again, although SVM performed comparably with an accuracy of **63.4%**, we can see from the confusion matrix that SVM is unsuitable as it misclassifies all Males.

Conclusion:

LDA and QDA are suitable for predicting the sex of abalone, with a particular strength in identifying infants, which is crucial for sustainable harvesting. While the models achieve moderate success in classifying males and females, further refinement or different techniques are recommended for more accurate predictions.

Question 2: Profitability

Introduction

The goal of this analysis is to predict the shucked and visceral weights of abalone using exterior measurements (length, diameter, and height) and develop an algorithm that provides profitability estimates based on fluctuating market prices for shucked meat and viscera. The model must allow flexibility in estimating profitability without retraining for every price change, relying on precomputed coefficients to meet computational constraints.

Methodology

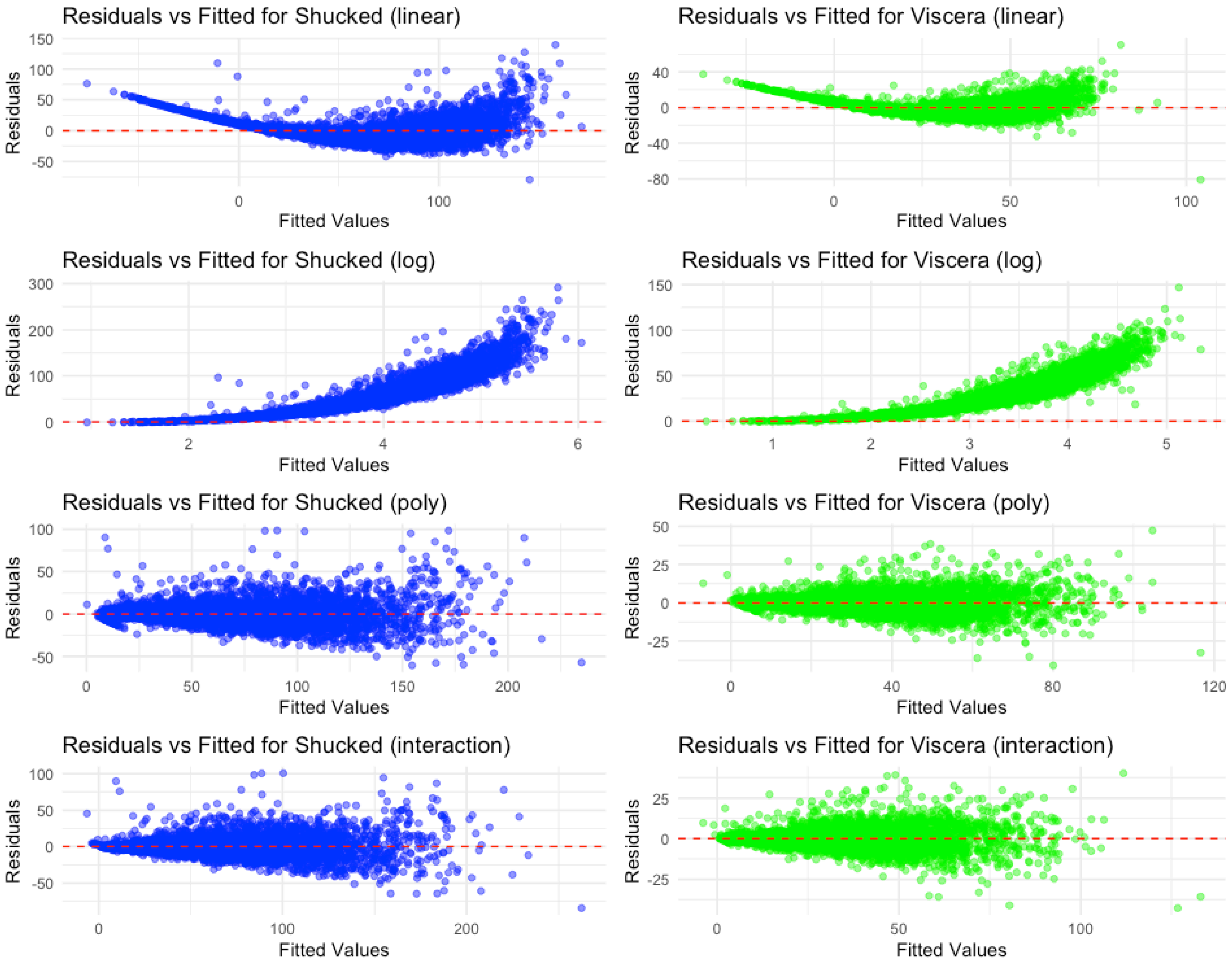
Step 1: Model Evaluation and Selection

We initially tested four different multivariate models to predict shucked and visceral weights based on the abalone dataset. These models were:

1. **Multivariate Linear Model (MLM)**
2. **Log-Transformed Model**
3. **Polynomial Model**
4. **Interaction Model (Selected Model)**

Each model was evaluated using residuals, RMSE, and R-squared metrics. The interaction model (fit_mlm_interaction) was chosen as the best performer, as its residuals showed no discernible pattern, unlike the linear and log models. Additionally, it had the lowest residuals and highest R-squared values, effectively capturing the complex interactions between the abalone's exterior measurements, making it more accurate than the simpler models.

```
Model:  linear
RMSE Shucked: 19.31654
RMSE Viscera: 9.168356
R-squared Shucked: 0.8106168
R-squared Viscera: 0.8250585
-----------------------------
Model:  log
RMSE Shucked: 80.69033
RMSE Viscera: 39.04218
R-squared Shucked: 0.8098055
R-squared Viscera: 0.8237192
-----------------------------
Model:  poly
RMSE Shucked: 14.71611
RMSE Viscera: 6.899468
R-squared Shucked: 0.8900819
R-squared Viscera: 0.9009302
-----------------------------
Model:  interaction
RMSE Shucked: 14.51208
RMSE Viscera: 6.909523
R-squared Shucked: 0.8931088
R-squared Viscera: 0.9006413
-----------------------------
```



Residuals vs Fitted for Shucked (linear)

Residuals vs Fitted for Viscera (linear)

Residuals vs Fitted for Shucked (log)

Residuals vs Fitted for Viscera (log)

Residuals vs Fitted for Shucked (poly)

Residuals vs Fitted for Viscera (poly)

Residuals vs Fitted for Shucked (interaction)

Residuals vs Fitted for Viscera (interaction)

Step 2: Precompute Coefficients

After selecting the interaction model, the coefficients were extracted and stored to summarize the relationships between length, diameter, height, and predicted weights. This approach ensures fast, efficient predictions using precomputed summaries, meeting the algorithm's computational constraints without the need to retrain the model.

Step 3: Predicting Shucked and Visceral Weights and Value with Confidence Intervals.

Using the precomputed coefficients, a function (predict_abalone_value_with_precomputed) was developed to predict shucked and visceral weights based on new input values for length, diameter, and height. This function allows users to input new abalone measurements and receive immediate weight predictions without needing to rerun the model.

The total value of the abalone is calculated based on the predicted shucked and visceral weights. The formula is:

$S=v_{\text{shucked}}\times X_{\text{shucked}}+v_{\text{viscera}}\times X_{\text{viscera}}$, where

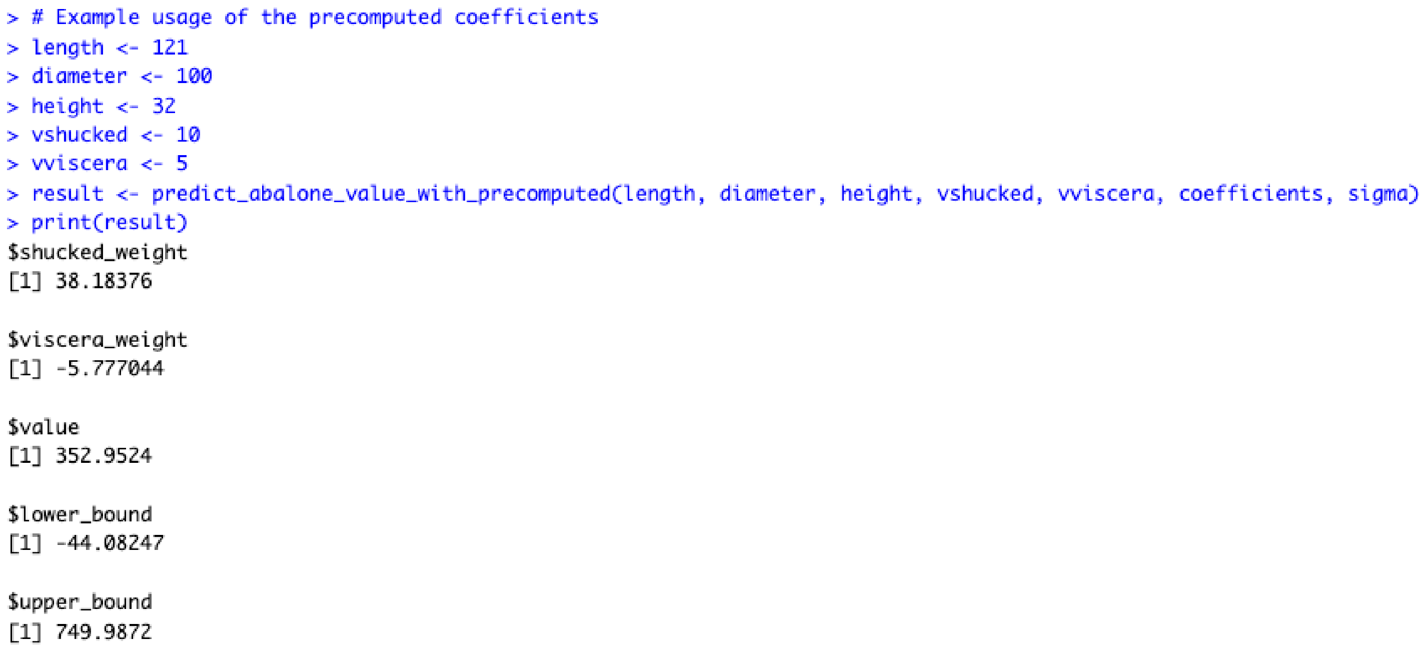
- X_{shucked} is the abalone’s shucked weight in grams
- X_{viscera} is the abalone’s viscera weight in grams
- v_{shucked} , the dollar value of 1 gram of shucked weight
- v_{viscera} , the dollar value of 1 gram of viscera weight;

In addition to predicting the abalone's value, a 90% prediction interval was calculated to account for the uncertainty in the estimates. This interval provides a range within which the true value is likely to fall, helping users make more informed decisions. The prediction interval is calculated using the residual standard deviation (sigma) from the interaction model and a t-distribution.

Step 4: Example Usage on Single Instance

The function successfully enables flexible profitability estimates, allowing users to input daily fluctuating prices for shucked meat and viscera. This flexibility ensures that the model remains relevant under changing market conditions. The interaction model can be applied to a single instance as shown below.

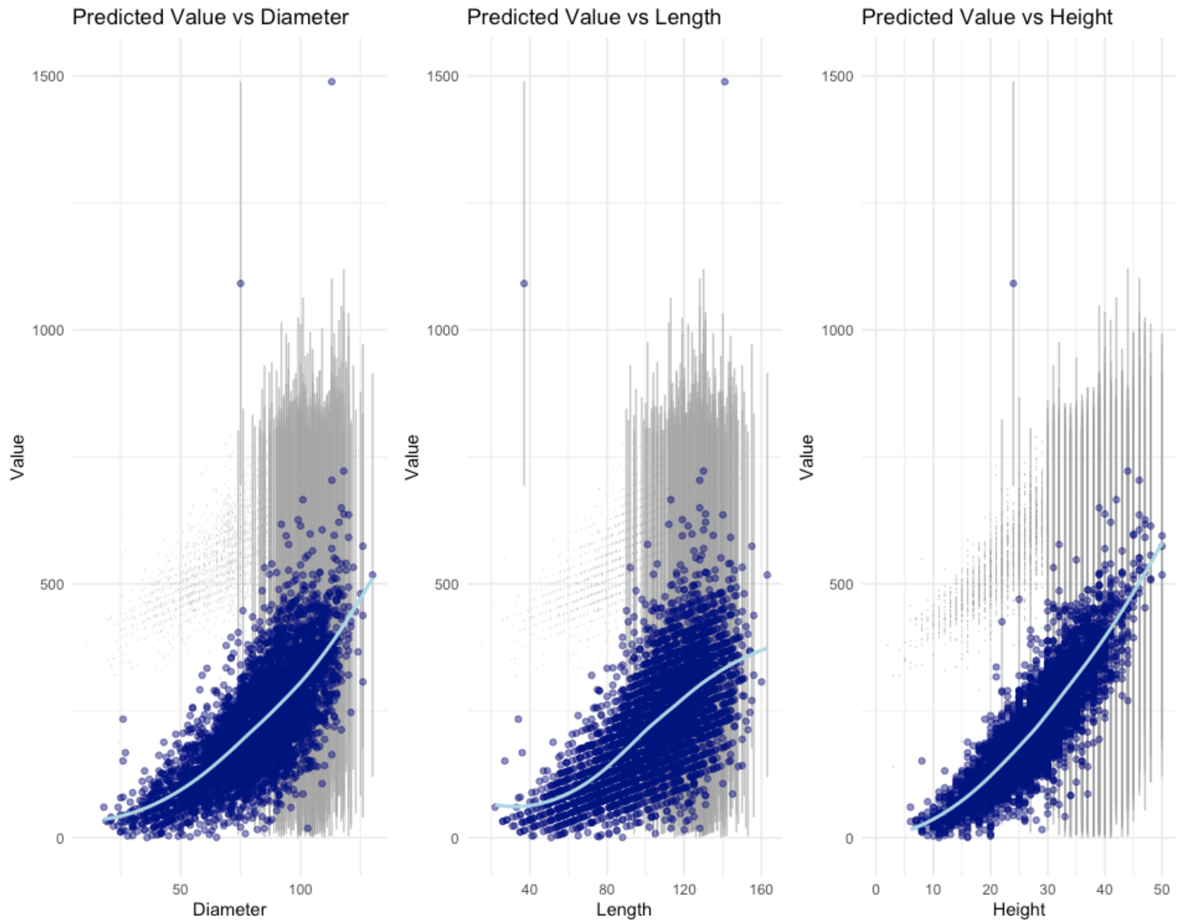
NOTE: The below assumes a predicted value of v(shucked) = 10 and v(viscera) = 5, however there are no prices given so these are completely arbitrary numbers and can be adjusted according to the actual price.



```
> # Example usage of the precomputed coefficients
> length <- 121
> diameter <- 100
> height <- 32
> vshucked <- 10
> vviscera <- 5
> result <- predict_abalone_value_with_precomputed(length, diameter, height, vshucked, vviscera, coefficients, sigma)
> print(result)
$shucked_weight
[1] 38.18376

$viscera_weight
[1] -5.777044

$value
[1] 352.9524

$lower_bound
[1] -44.08247

$upper_bound
[1] 749.9872
```

Step 5: Example Usage on Dataset (Multiple Instances)

Or it can be applied to an entire dataset as shown below to the entire Abalone dataset as an example. The predicted values were plotted against the input variables (length, diameter, and height) with 90% prediction intervals. The results demonstrated that larger abalones have higher predicted values. The prediction intervals were wider for larger abalones, reflecting the greater uncertainty in their predictions.



Conclusion

The interaction model was the best choice for predicting abalone weights, providing both accuracy and flexibility. The use of precomputed coefficients enables fast predictions, while prediction intervals offer insights into potential variability. This approach meets the computational requirements and provides reliable predictions for abalone profitability in dynamic market environments.