Age of Abalone

470212871, 480435020, 480388575, and 480497305

The University of Sydney DATA 2002 - T14-05

This version was compiled on November 5, 2019

By utilising multiple linear regression, this project predicts the age of abalone that is commercial viable (4-8yrs) from external variables to avoid costly and intrusive alternative procedures. After optimising and transforming from an initial full model in accounting for mutlicolinearity and removing insignificant variables, the final model containing factors of shell/shucked weight, diameter, length and gender. It possessed a R-squared of 0.509 and Root Mean Square Error (RMSE) of 2.5605, being relatively more accurate in our target age range.

Abalone | Multiple Linear Regression | Prediction

1. Introduction. An abalone is sea snail with significant Australia commercial viability. According to the Department of Agriculture, Australian abalone production is projected at $209 million by 2024, becoming the third largest producer worldwide. Research conducted by the Regional Sea Farm Development and Demonstration evidenced the taste of Abalone is highly dependent upon its age, with the ideal age for consumption between 4 to 8 years. This present report utilises easily measured physical features (such as weight and size) as predictor variables in order to estimate an abalone's age. Despite the age of Abalone being directly linear (+1.5) to the number of rings, measuring rings involved cutting open the shell, staining it and counting under a microscope, which is expensive and tedious. If the model can become viably accurate, it avoids expensive processes in cutting the abalone shell to allow for mass harvesting optimisation in targeting abalones at their commercial prime.

2. Data Set. The dataset was originally collected by the Marine Resources Laboratories - Tarrona from the department of Primary Industry and Fisheries in Tasmania in 1994. However, the data set was subsequently sourced from the UCI Machine Learning Repository. Thus, further information regarding the process of abalones collection was not provided, thus potentially limiting assumptions of independence embedded in random sampling.

The variables in the data set included **Sex** (male, female or infant). The infant category denotes abalone before they express sexual features (~2-3yrs when farmed). The **Length** of the Shell, its **Diameter** and the **Height** of the meat were also provided, alongside weight variations, including **Whole weight**, **Shucked weight**, **Viscera weight** and **Shell weight**.

We also note all continuous values had been divided by 200 (as UCI indicated), and corrections to observed values aligned them with secondary research, as prior to that the max weight value was 2.2 whilst an average abalone weighed ~300grams.

3. Analysis.

3.1 Pre-processing data variables. To model **Age** as a multiple regression function, we first considered the feasibility of potential variables presented in the data.

- **Rings** is excluded as age is calculated directly from it: $age = rings+$ 1.5 (indicated in UCI source) and would be counterintuitive to our aim in predicting age independent of number of rings.

- Dummy variables were created to incorporate the **Sex** variable. However the Infant category, consisting of abalone approximately 2.5 years old *(Government of Western Australia Department of Fisheries, 2011)* was dropped as it fell outside our investigative boundary (limited predictions to 4-8years).

All other variables were considered within the initial full model.

Model-Optimsation outline.

a) A full model was created;
b) **Assumptions** regarding multiple regression then considered at first instance;
c) A **backwards stepwise approach** was taken for the initial model with formal hypothesis testing to optimise model;
d) **Forward and backward model selection** methods using AIC was uniform with our final model, and assumptions reconsidered.

At the outset, the full model contains:
$age = \beta_0 + \beta_1 length + \beta_2 diameter + \beta_3 height + \beta_4 skweight + \beta_5 vweight + \beta_6 shweight + \beta_7 allweight + \beta_8 male + \beta_9 female + \epsilon.$

3.2 Assumption check. We must first consider the assumptions underlying multiple linear regression to ensure that a valid initial model informs our variable selection towards a resultant model that retains its statistical validity. The assumptions include a) Independent errors, b) Normally distributed residuals, c) Linear relationship, d) Homoscedasticity and e) Multicollinearity.

1) The error term is **Independent** as the residuals are uncorrelated and the data sampled randomly.
2) **Normality** assessed referencing the QQPlot **(Figure 1)**, with observations failing to follow the diagonal line at extremities (especially upper tail). However, Central Limit Theorem applies due to large sample size (4177) to satisfy normality.
3) **Linearity**: the data looks symmetrically distributed above and below zero in the residuals plot (Figure 1), not forming a "bow" shape and hence reasonable linear as values aren't consistently negative/positive at extremities of fitted values.
4) **Homoscedasticity** is violated here (Figure 1). As the spread of residuals systematically "fans out"/increases as fitted values increase, inconsistent spread will hamper our model's accuracy at higher values. This coincides with latter observations of greater inaccuracies when $age > 12$ in the proposed model.

Hence, we consider transforming **Age**.

3.3 Transformation. **Box-cox, square root and reciprocal** transformation were applied (Figure 2):

Inverse and square-root transformations failed to improve homoscedasticity, instead creating linearity issues within the model. And whilst we see heteroscedasticity decreased with Box-cox, a notable negative parabolic trend emerged.

As linearity is the most important assumption within any linear regression model and our overall aim predicts the age of commercially viable abalone (around 4-8yrs), heteroscedasticity will not overly influence our predictions as we predict on the lower age range where model remains accurate. Also, complex transformations will decrease the interpretation of our results. Hence, the original **Age** variable is *preserved without any transformation.*

5) **Multicollinearlity** in factors like allweight, skweight and shweight that correlate significantly will like be observed as it varies proportionally with abalone growth. The correlation matrix (Figure 3) confirms this, with almost all variable combinations posessing correlation > 0.8. Further evaluation utilisng VIF (Variance Inflation Factor) found **Allweight** as an large outlier (109.59 compared to mean of 37.53). Thus, we remove it and test if there are substantial performance impairments, as well as applying further optimisations to remove insignificant variables.

Modelling.

4.1 Model optimisation. We firstly split the data into training (90%) and test(10%) subsets.

1) Addressing multicolinearity

We then evaluate whether removing the **allweight** variable will significant hamper model performance for both in-sample and out-of-sample predictions. For outer sample performance we can see no substantial loss of performance.**(Refer to appendix-out/in sample)**

In considering the differences in **Root Mean Square Error** (standardised differences between prediction and observations) and R^2 (percentage of variance explained by model), we note that overall there is an

statistically insignificant decrease in the accuracy of prediction. Thus, the amended model is retained to better address multicolinearity. **(Refer to appendix- Test after dropping all weight)**

2) Further Optimisation: Backwards variable selection

We now undergo a formal hypothesis test for any variables that should be further excluded to optimise the model.

From regression analysis (Figure 4), both length and vweight are individually insignificant at the 5% level of significance. However, as the p-values only test individual coefficients, we test if these coeffecients (first vweight, then length (Figure 5)) are significant separately as follows: **(Refer to appendix- Final test)**

First we formally define the pre-existing model with population parameters:

$age = \beta_0 + \beta_1 length + \beta_2 diameter + \beta_3 height + \beta_4 skweight + \beta_5 vweight + \beta_6 shweight + \beta_7 male + \beta_8 female + \epsilon$.

Hypothesis: $H_0 : \beta_5 = 0$ vs $H_1 : \beta_5 \neq 0$ [for vweight]; then $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ [for length]

Assumptions: The residuals ϵ_i are iid $N(0, \sigma^2)$ and there is a linear relationship between y and x. This has been previously considered and results are consistent under the new model(s)

Test statistic: $T = \frac{\hat{\beta_5}}{SE(\hat{\beta_5})} \tilde{} t_{n-p}$ for vweight and $T = \frac{\hat{\beta_1}}{SE(\hat{\beta_1})} \tilde{} t_{n-p}$ for length under H_0 with 4170 degrees of freedom.

Observed test statistic: $t_0 = 0.345$ for vweight and $t_0 = -0.987$ for length (Figure 5)

P-value:

$2P(t_4170 \geq |0.235|) = 0.730$ (for weight)

$2P(t_4170 \geq |-0.987|) = 0.324$ (for length) (Figure 5)

Conclusion:

We do not reject either H_0 at the 5% level of significance as both p-value is greater than 0.05. Hence, there is no evidence to suggest that there is a significant linear relationship between age and vweight or length and both are dropped from the model.

Our findings were also consistent with applications of forward and backward searches using AIC, which both selected an identical model that removed "length" and "vweight" as insignificant predictors for model optimisation.

4.1 Test of the final model. Before concluding on the final model, a reconsideration of the residual and QQplots made sure we satisified our assumptions again. The problems of linearity, homoscedasticity and normality have not been exacerbated and the model remains valid.

We then utilised 10 datapoints from the original sample to visualize the prediction of our model, and it generally appeared accurate when age < 12 by evaluating both the prediction and confidence interval. Then a more formal evaluation was also undertaken to find the AIC and R^2 to evaluate the model at each stage, with the full, amended (removing allweight) and final model. **(Refer to appendix- Prediction)**

5. Conclusion. Therefore, the final fitted model is:

age = 4.40 + 0.05 × diameter + 0.06 × height – 0.06 × shucked weight + 0.10 × shell weight + 0.92 × male + 0.87 × female + ϵ.

Shucked weight cannot be considered separately as it leads to increasing abalone weight decreases age), and thus is considered together with shell weight in the interpretation of estimated coefficients.

On average, holding the other variables constant,

1) A 1-mm increase in diameter and in height leads a 0.05-unit increase and a 0.06 year increase in age of abalone respectively.

2) A 1-gram increase in shucked weight and shell weight combined leads to a 0.04 year increase in age of abalone.

3) Male abalones are older than female abalones by 0.05 years.

We also note that diameter, height, shucked weight, shell weight and genders are the significant variables that influence age.

In terms of model performance, the adjusted R-square of our model is 0.509. This means all predictor variables accounts for 50.9% of the total sum of squares (variance) explained by the regression. The root mean square error is 2.5605, indicating the standardised mean of all variance between the predicted and the actual abalone age. This means our model is relatively accurate and can be valid for our inquiry question provided enough predictions are made.

6. Limitations. The measured variables available in the dataset focused primarily on physical features of abalones that were naturally highly correlated, and contributed to the extensive multicollinearity that was somewhat mitigated by removing **allweight**. Whilst more variables could have been excluded,removing too many variables will stop the mutliple linear regression in generating an applicable and useful model.It also potentially contributed to issues in linearity and homoscedasticity that could not be addressed with transformations.

There could also be potential issues with random sampling and unintended dependence introduced by the method of abalone sampling, as they are difficult to soure in the wild and seasonal changes may affect their dimensions and weight. It is also important that despite the research being conducted on Blacklip abalone, Greenlip abalones are the predominant abalones currently farmed in Australia. These species have substantial biological differences, potentially rendering the prediction of age for Greenlip abalones invalid.

More data comprising of the features of Greenlip abalone will allow for the model to accommodate for both species. Another definite improvement could be the addition of environmental variables (ie. water temperature and population density) that are independent of an abalone's growth. Furthermore, as our inquiry question delved into the commercial optimum of farmed abalone, sampling farmed abalones instead of wild abalones (as they exhibit different ecological traits and growth behaviours) will develop a more relevant model.

Nonetheless, our final model remains generally accurate within our target range and our final model allows us to address our inquiry question in predicting the age of commercially viable abalones (around 4-8 year old) using measurable variables with meaningful accuracy.

7. Appendix.

Tables And Plots. Test after dropping all weight

```
#
#  Call:
#  lm(formula = age ~ length + diameter + height + skweight + vweight +
#      shweight + sex_F + sex_M, data = nab)
#
#  Residuals:
#       Min       1Q   Median       3Q      Max
#  -10.8897  -1.3299  -0.3379   0.8607  15.7426
#
#  Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#  (Intercept)  4.409092   0.271384  16.247  < 2e-16 ***
#  length      -0.003549   0.009209  -0.385     0.70
#  diameter     0.058149   0.011335   5.130 3.03e-07 ***
#  height       0.055457   0.007819   7.093 1.54e-12 ***
#  skweight    -0.057041   0.002330 -24.481  < 2e-16 ***
#  vweight     -0.003759   0.005198  -0.723     0.47
#  shweight     0.099027   0.003473  28.513  < 2e-16 ***
#  sex_F        0.872555   0.104173   8.376  < 2e-16 ***
#  sex_M        0.923806   0.097406   9.484  < 2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#  Residual standard error: 2.234 on 4168 degrees of freedom
#  Multiple R-squared:  0.5209, Adjusted R-squared:   0.52
#  F-statistic: 566.5 on 8 and 4168 DF,  p-value: < 2.2e-16
```

Final Test

```
#
#  Call:
#  lm(formula = age ~ diameter + height + skweight + shweight +
#      sex_F + sex_M, data = nab)
#
#  Residuals:
#       Min       1Q   Median       3Q      Max
#  -10.7530  -1.3221  -0.3387   0.8657  15.7933
#
#  Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#  (Intercept)  4.401425   0.247111  17.812  < 2e-16 ***
#  diameter     0.053707   0.004986  10.772  < 2e-16 ***
#  height       0.054925   0.007793   7.048 2.12e-12 ***
#  skweight    -0.058182   0.001884 -30.876  < 2e-16 ***
#  shweight     0.098245   0.003276  29.994  < 2e-16 ***
#  sex_F        0.865481   0.103229   8.384  < 2e-16 ***
#  sex_M        0.920104   0.096925   9.493  < 2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#  Residual standard error: 2.233 on 4170 degrees of freedom
#  Multiple R-squared:  0.5208, Adjusted R-squared:  0.5201
#  F-statistic: 755.4 on 6 and 4170 DF,  p-value: < 2.2e-16
```

Out-of-sample Predictions

```
#            model     RMSE        R2
#  1    full model 2.501869 0.4491533
#  2 amended model 2.517688 0.4412246
```

In-sample Prediction

```
#            model        R2
#  1    full model 0.5456839
#  2 amended model 0.5271141
```

Prediction Table

```
#  # A tibble: 11 x 6
#     predicted_age actual_age lwr_prediction upr_prediction lwr_confidence
#             <dbl>      <dbl>          <dbl>          <dbl>          <dbl>
#   1          12.0       10.5           8.30           15.6           11.9
#   2          13.8       10.5           10.1           17.4           13.6
#   3          10.9        9.5           7.24           14.6           10.8
#   4          11.6       12.5           7.94           15.3           11.5
#   5          12.4        9.5           8.77           16.1           12.3
#   6          13.0       12.5           9.34           16.7           12.9
#   7          12.4       11.5           8.72           16.1           12.3
#   8          10.8       11.5           7.15           14.5           10.7
#   9          12.5       12.5           8.80           16.2           12.3
#  10          14.7       11.5          11.0            18.4           14.6
#  11          11.4       11.5           7.77           15.1           11.3
#  # ... with 1 more variable: upr_confidence <dbl>
```


References. Agriculture AGD (2019). Annual Fisheries Outlook 2019. URL https://archive. ics.uci.edu/ml/datasets/abalone.

Kassambara, A. (2018, 03). Multicollinearity Essentials and VIF in R. Retrieved from Statistical tools for high-throughput data analysis: http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). (n.d.). Retrieved from http://www.eumetrain.org/data/4/451/english/msg/ver_cont_var/uos3/uos3_ko1.htm

Nash W, Sellers T, Talbot S, Cawthorn A, Ford W (1994). Abalone Data Set. URL https://archive.ics.uci.edu/ml/datasets/abalone.

Project RSDD (1990). BIOLOGY AND CULTURE OF ABALONE. URL http://www.fao.org/3/AB731E/AB731E01.htm.

Yobero, C. (2016, June). Methods for Detecting and Resolving Heteroskedasticity. Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/187387_3ca34c107405427db0e0f01252b3fbdb.html

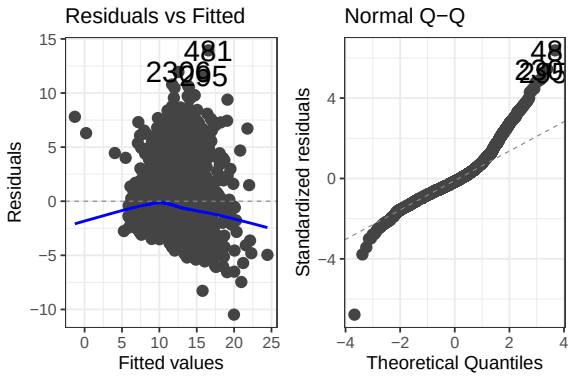

Fig. 1. Assumption First

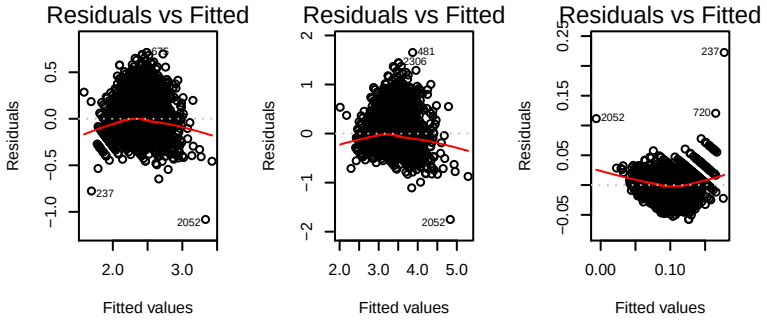

Fig. 2. Transformation

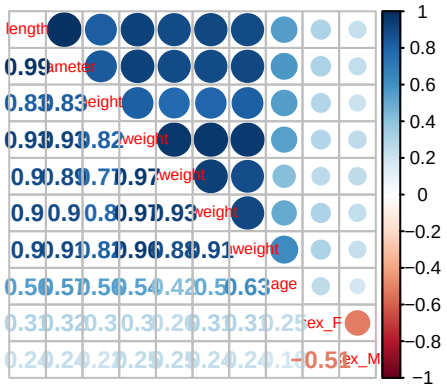

Fig. 3. Correlation Matrix

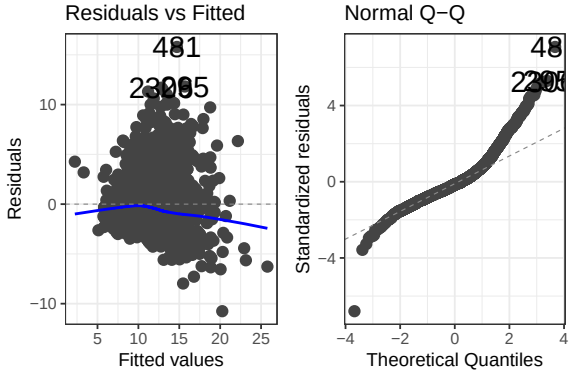

Fig. 4. Assumption Final