Submitted to the Annals of Applied Statistics

STREAMLINING ABALONE AGE DETERMINATION WITH PREDICTIVE MODELS

BY HONGYU CHEN[1,a]

[1] *Faculty of Mathematics, University of Waterloo,* [a]*h542chen@uwaterloo.ca*

1. Introduction. Abalone is a type of mollusk with a convex, rounded to oval shell that can vary from highly arched to flattened. It is widely liked for its taste and nutritional value across the world. Beyond culinary appeal, abalone are farmed for their decorative shells, which are highly prized economically. The price of abalone correlates positively with its age, determined by counting growth rings that appear annually inside its shell. This process involves cutting the shell, polishing, staining, and counting rings under a microscope which is a labour-intensive task. The **research question** of this project aims to explore predicting abalone age using easily obtainable physical measurements, offering a faster alternative to microscopic ring counting. By analyzing attributes such as shell size and weight, predictive models can efficiently estimate abalone age, establishing correlations between these measurable features and the number of rings.

2. Data. The dataset used in this project is sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/1/abalone) and contains measurements of the physical attributes of abalone. This section will thoroughly explain data quality, encompassing univariate and multivariate analysis.

Variable	Description	Missing	Outliers	Min.	Mean	Median
Type	M, F, and I (infant)	no	-	-	-	-
LongestShell	Longest shell measurement	no	49	0.075	0.524	0.545
Diameter	Perpendicular to length	no	59	0.0550	0.4079	0.4250
Height	With meat in shell	no	29	0.0000	0.1395	0.1400
WholeWeight	Whole abalone	no	30	0.0020	0.8287	0.7995
ShuckedWeight	Weight of meat	no	48	0.0010	0.3594	0.3360
VisceraWeight	Gut weight (after bleeding)	no	26	0.0005	0.1806	0.1710
ShellWeight	After being dried	no	35	0.0015	0.2388	0.2340
Rings **(Target)**	+1.5 gives the age in years	no	278	1.000	9.934	9.000
OtherWeight	-	-	-	-0.44750	0.01800	0.03700

TABLE 1

Summary of variables in the abalone dataset

2.1. *Data Quality.* The abalone dataset includes 8 variables and Rings are the target variable. The abalone dataset exhibits no missing values, as shown in (Table 1). After reviewing the dataset summary, it was noted that the minimum value for "Height" is 0, which is biologically implausible. Additionally, considering that "Whole Weight" represents the total weight of the abalone, the sum of other weight predictors should always be less than or equal to the Whole Weight. To quantify this discrepancy, we introduced a new variable called "Other Weight", which captures the difference between the Whole Weight and the combined sum of the other weight predictors. Notably, a negative minimum weight was observed, which is also implausible. Given that these instances constitute a small portion of the dataset, they were removed to ensure dataset integrity and reliability for further analysis.

Fig 1: Relational Matrix

2.2. *Exploratory Data Analysis.* Through the relational matrix (Figure 1), we observed that all the variables exhibit positive correlations with the number of rings which indicates that older abalones tend to have larger shell sizes and higher weights. The "Type" column shows the data is evenly distributed among the three types: Male (Blue), Female (Red), and Infant (Green). After reviewing the plots on the diagonals, all variables appear normally distributed. The distributions of males and females are similar across predictors. The longest shell and diameter exhibit significant left skewness. Height and all four weight features show significant right skewness which may be because of outliers. It's very clear that infant abalones are generally small and weigh much less than adults.
The high correlation values indicate high multicollinearity among the predictors. This is particularly evident since the whole weight should include all other weight predictors. The length of the Longest shell and diameter also show a strong correlation, as the length largely determines the diameter. Given this high multicollinearity, it will be crucial to employ specific methods or models designed to handle multicollinearity during the modelling process. Techniques such as regularization or feature selection may be necessary to mitigate the effects of correlated predictors and enhance the robustness of the model.

3. Methods. To get a good predictive model for age determination, we will focus on several models which have different properties. First, we will separate the data set into training data (80%) and test data (20%). Test data is only used for evaluation.

3.1. *Baseline Model.* We will start by fitting a linear model using all variables as our baseline model. Next, we will employ feature selection techniques to simplify models to make them easier to interpret and avoid the curse of dimensionality.

3.1.1. *Akaike Information Criterion / Bayesian Information Criterion.* They serve to balance the goodness of fit of a model with its complexity by penalizing overly complex models. As the number of parameters (p) increases, the log-likelihood $(\log L_i)$ tends to rise, but the penalty term ensures that the overall evaluation considers both fit and model simplicity. Adding parameters requires their contribution to model fit to outweigh the penalty, thus achieving a balanced selection. (n) in BIC stands for the number of values in the data set.

$$\mathrm{AIC}_i = -2\log L_i + 2p_i \quad \mathrm{BIC}_i = -2\log L_i + p_i \log n$$

3.1.2. *Best Subset Regression.* Best subsets regression is an automated tool used in the exploratory stages of model building to identify a useful subset of predictors.

3.1.3. *Stepwise selection through GCV.* Tuning parameters are almost always selected using (generalized) cross-validation: the value of the tuning parameter is chosen to minimize the GCV score. Stepwise selection through GCV automates the process of choosing variables in a model by sequentially adding or removing them based on their impact on model fit, as evaluated by the GCV metric.

$$\mathrm{GCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - \mathrm{trace}(\mathbf{S})/N}\right]^2$$

3.2. *Multicollinearity.* Having previously discussed the issue of multicollinearity, we have now explored various feature selection techniques and aim to evaluate these alongside additional regularization techniques. The Variance Inflation Factor (VIF) serves as a measure of multicollinearity in regression analysis. We will also assess and adjust our previous models using feature selection techniques to determine if we can effectively reduce VIF.

3.2.1. *Ridge Regression / Elastic Net.* As many of variables may be collinear or highly dependent, we will initially try Ridge regression, which effectively reduces the influence of inputs to values close to zero. It is known for its ability to handle collinearity effectively. However, it does not produce sparse models where coefficients are exactly set to zero. LASSO regression, which could potentially offer sparse solutions by setting coefficients to zero, is not suitable here due to its known struggles with collinearity. Considering these factors, we will explore Elastic Net regression, which combines the strengths of the Ridge and LASSO by handling collinearity while allowing coefficients to be set to zero. Aim to minimize:

$$\text{Ridge:} \quad ||y - X\beta||_2^2 + \lambda||\beta||_2^2 \quad \text{Elastic Net:} \quad ||y - X\beta||_2^2 \quad +\lambda(\frac{1-\alpha}{2}||\beta||_2^2 + \alpha||\beta||_2)$$

3.3. *Nonlinear Regression.* The method above was used under the assumption of a linear relationship between input variables and the target variable. However, real-world scenarios often involve relationships that are not strictly linear. In this section, we will explore nonlinear regression techniques.

3.3.1. *Additive Model.* Additive models enable us to assess and interpret non-linear relationships independently, providing a clearer understanding of the data's complexity. An additive model is the following:

$$E(y|X_1 = x_1, ..., X_p = x_p) = \alpha + f_1(x_1) + ... + f_p(x_p)$$

Each $f_j(x)$ is individual cubic B-splines and the whole thing is fitted through ridge regression.

3.3.2. *Polynomial Regression.* Next, we will explore the non-linear relationships using polynomial regression. Polynomial regression involves modelling the relationship between the independent variable x and the dependent variable y as an nth-degree polynomial in x. This method fits a curved line to the data, capturing non-linear patterns in the relationship between x and the conditional mean of y. We will investigate various polynomial degrees and use ANOVA tests to analyze differences between each model.

3.3.3. *Support Vector Machine Regression.* In SVM regression, the algorithm seeks to identify a hyperplane within a high-dimensional space that effectively represents the relationship between input variables (features) and the continuous output variable. Unlike traditional regression methods focused on minimizing prediction errors, SVM regression aims to find a hyperplane that maximizes the margin $||\mathrm{w}||_2^2$ which is the distance between the hyperplane and the nearest data points (support vectors). In the function below, y_i is 1 or -1.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n}(1 - \mathbf{y}_i\hat{y}_i)^+ \quad \text{s.t.}\hat{y}_i = \langle \mathbf{x}_i, \mathbf{w}\rangle + b, \forall i$$

For regression scenarios where a linear model is inadequate, nonlinear SVM regression can be employed using a nonlinear kernel function that maps x into a higher-dimensional space. This approach allows SVM regression to capture complex, nonlinear relationships between input and output variables. Some common kernel includes Polynomial, Gaussian Radial Basis function, and Sigmoid.

3.4. *Advanced Model.* Aim to prioritize interpretability and performance.

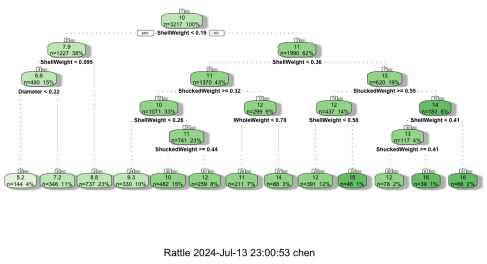

Fig 2: Regression Tree

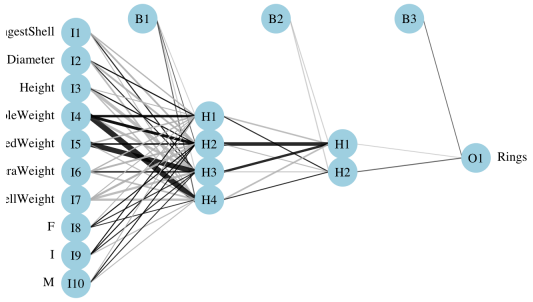

Fig 3: Neural Network

3.4.1. *Random Forest / Regression trees.* A regression tree (Figure 2) is a model that provides discrete predictions by partitioning the sample space. Decision trees have a naturally intuitive structure consisting of branches and nodes that make decisions based on feature values. This structure facilitates easy interpretation as a series of if-then-else conditions. The fundamental algorithm for constructing a tree is recursive partitioning. Random forests are similar models that aggregate predictions from multiple decision trees. And it's a kind of ensemble model.

3.5. *Nerual Network.* A neural network (Figure 3) is a computational model inspired by the way biological neural networks in the human brain process information. It consists of interconnected nodes (neurons) organized in layers. Each neuron receives input, processes it using an activation function, and then passes the output to the next layer of neurons. We will be using 4x2 hidden layer architecture. The network assessed its predictions against actual outputs, calculated loss or error, and refined its weights through backpropagation. Neural networks leverage interconnected layers of nodes to learn complex patterns and relationships within data, making them versatile tools for various predictive tasks.

4. Results. To assess model fit, we will use RMSE (Root Mean Square Error). The results for each model/method are presented in (Table 2) in order. Our initial approach involved fitting a baseline linear model using all variables, resulting in a test RMSE of 2.23.
We explored best subset regression and found that using 4 variables yielded performance comparable to using more variables, while also providing a straightforward model.
Next, we attempted to streamline the model selection process using AIC and BIC criteria. This simplified model yielded performance comparable to the initial linear model. Notably, both criteria removed the variable "LongestShell" due to its high correlation with diameter. Additionally, BIC suggested excluding the "Type" variable, possibly due to its stricter penalty for model complexity compared to AIC. We constructed a residuals plot and a normal QQ plot suggesting that the model violates the constant variance and normality assumptions. Residuals increase with fitted values, and the QQ plot shows fat tails. To address these issues, we use log transforming the response. Although it does not improve the performance at all, the model assumptions look much better.

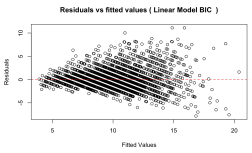

Fig 4: Residual plot

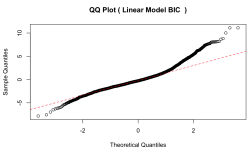

Fig 5: QQ plot

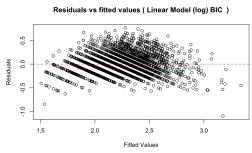

Fig 6: log(response)

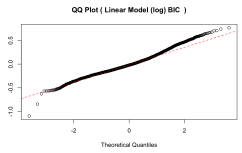

Fig 7: log(response)

Stepwise selection through GCV resulted in a model with performance similar to the linear model. Interestingly, it selected the same model as indicated by AIC.
To tackle multicollinearity, we identified high variance inflation factors (VIF) among the variables. While the previous feature selection models failed to address this issue, high VIF variables still exist. Potentially because of all of the feature selection methods decide to keep all four weight predictors. As an alternative approach, we attempted brute-force removal of highly collinear (VIF) variables, which reduced VIF but also increased prediction error due to the loss of important variables. We then attempted LASSO regression with minimum lambda (with smallest GCV) and elastic net, hoping to address collinearity without sacrificing performance. However, these methods did not significantly improve results in this case, possibly indicating that the assumed linear relationships in our data do not hold well.
We then proceeded with an additive model to identify the optimal model (i.e., achieving the smallest RMSE). The stepwise selection was employed, revealing that the model using all parameters performed best. The approximate significance of smooth terms indicated that when all variables were used as covariates, every variable (except diameter) was considered a highly significant smooth term, suggesting non-linear relationships. In essence, as the measurable features of abalones change, their effects on ring counts do not vary consistently, for instance, these effects may increase more rapidly within certain ranges of height and less so in

others. Then we explored polynomial regression to further discover non-linear relationships. We conducted polynomial transformations (except diameter) up to degree 10 and employed ANOVA tests and AIC/BIC, ultimately finding that a polynomial degree of 4 provided optimal results as the ANOVA test shows that the addition of these higher degree polynomials ($\geq$5) does not significantly improve the model fit. To further address nonlinear relationships, we also experimented with Support Vector Regression (SVR), using the Radial Basis Function (RBF) to map the input data into a higher-dimensional space which shows to be optimal for our dataset. In the end, all of our non-linear models perform better than the previous linear regression models in RMSE scores where our additive model has the lowest test RMSE with good model assumptions. Despite high multicollinearity, we still selected a model with all parameters. Interestingly, we found out that weight itself cannot predict age effectively, including other weight variables is necessary to maintain model performance.
We also tested Random Forest with parameter tuning, which aggregates outputs from multiple decision trees. After parameter tuning, we found out that 2200 trees would provide the best performance on the test dataset. Its reasonable results (close to the baseline model) and interpretability provide advantages compared to other methods. Pruning regression trees was attempted to reduce complexity and enhance accuracy but exhibited worse performance.
Finally, Neural Networks were also explored, and configured with 4x2 hidden layers. The neural net package in R uses the logistic (sigmoid) activation function. This approach outperformed all other methods. This highlights the formidable capability of neural networks to learn complex functions effectively.

Model / Method	**Details**	**RMSE (Train)**	**RMSE (Test)**
Linear Model	Baseline Model	2.213922	2.229341
Linear Model	AIC / BIC	2.213955 / 2.21531	2.228616 / 2.226061
Linear Model	Stepwise Selection (GCV)	2.213955	2.228616
Linear Model	Remove high VIF	2.262283	2.286281
Lasso Regression	Cross-validation, Min_λ	2.214153	2.230656
Elastic Net	Linear Model	2.214103	2.230499
Additive Model	Stepwise Selection	2.063426	2.16213
Polynomial Regression with degree of 4	AIC	2.074328	2.193002
Support Vector Machine	Radial Basis Function	2.08917	2.197734
Random Forest	2200 Trees	1.977292	2.226882
Regression Tree	With Pruning	2.272167	2.42185
Neural Network	4x2 Hidden Layers	0.07	0.169

TABLE 2
Result

5. Conclusions. In conclusion, our study shows the feasibility of predicting abalone age based on physical attributes. Through data analysis and model evaluation, neural networks are the most accurate method for age estimation. Additionally, both the additive model and polynomial regression unveiled insightful non-linear relationships, opening avenues for deeper exploration. Looking ahead, refining neural network architectures and exploring ensemble methods offer promising paths to enhance predictive accuracy further. Techniques such as varying architectures, trying different activation functions, hyperparameter tuning, transfer learning and ensemble methods hold the potential for improving model robustness. Furthermore, decision trees/Random Forest retain their utility in scenarios prioritizing interpretability, where decisions can be understood and explained by humans. Future research can advance the precision of abalone age estimation, thereby contributing significantly to the fields of aquaculture management and environmental conservation. These advancements are crucial for sustainable resource management and biodiversity preservation in marine ecosystems.