Predicting the Age of Abalone: Multiple Linear Regression Analysis

Salvatore Palmeri and John Rempe

MAT 402-1

12/8/2025

Introduction

Accurately estimating abalone age is vital in marine biology, fisheries management, and ecological monitoring, but the traditional method of cutting the shell, staining it, and examining the rings under a microscope is labor-intensive and destructive. Since shell rings reliably indicate age, statistical models predicting ring count offer a practical alternative for assessing large abalone populations. In this study, we examined regression techniques using dimensions and weight-based features from the Abalone dataset obtained from the UCI Machine Learning Repository. This dataset was created by researchers in a biological study where they collected physical measurements from blacklip abalone located along the North Coast and Islands of Bass Strait. With roughly four thousand observations in the dataset, our goal was to find the most significant attributes that accurately predict ring count, revealing their age.

The variables in the dataset are Sex, Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight, and Rings. The variable Sex is a categorical variable consisting of males, females, and infants. We later decided to remove infants from our study and focus on males and females as the age of infants' is not important in our focus. The variables Length, Diameter, and Height are all the physical size measurements of the abalones measured in millimeters. The weight measurements are all measured in grams with the whole weight being the weight of the abalone shell and meat included. The Shucked Weight variable is just the weight of the meat. The variable Viscera Weight is the measured weight of the gut after bleeding. The Shell Weight variable is the weight of the shell only after being dried. Researchers are able to obtain the age in years of the abalones by viewing the number of rings on their shell and adding 1.5. After removing infant abalones and treating sex as a binary variable, we sampled 500 random individuals to test if their physical attributes could replace the tedious process of determining age.

Graphical Presentation and Summary Statistics

Variable	Min.	Q1	Median	Q3	Max.	Mean	SD.
Length (mm)	0.155	0.515	0.585	0.635	0.815	0.5687	0.098
Diameter (mm)	0.115	0.4	0.455	0.5	0.65	0.4455	0.081
Height (mm)	0.025	0.13	0.155	0.175	0.25	0.1535	0.033
Whole Weight (grams)	0.024	0.6843	1.1015	1.2837	2.657	1.0117	0.459
Shucked Weight (grams)	0.009	0.2774	0.432	0.5594	1.488	0.4352	0.214

Viscera Weight (grams)	0.005	0.15	0.2145	0.2906	0.541	0.2226	0.105
Shell Weight (grams)	0.0075	0.195	0.2878	0.3688	0.7975	0.29	0.132
Rings	5	9	10	12	23	10.88	3.055

Table 1: Summary Statistics of Abalone Physical Attributes from Random Sample

The summary statistics in the table provide an overview of the physical characteristics of abalone in the random sample. The measurements of length, diameter, and height show relatively small variation given their mean values. The small interquartile ranges suggest the dimensions of the abalone are consistent throughout the sample. On the other hand, the weight-related variables display greater variation with their wider ranges. The wide spreads between the weight variables show more diversity in abalone mass as opposed to the exterior measurements of their shell.

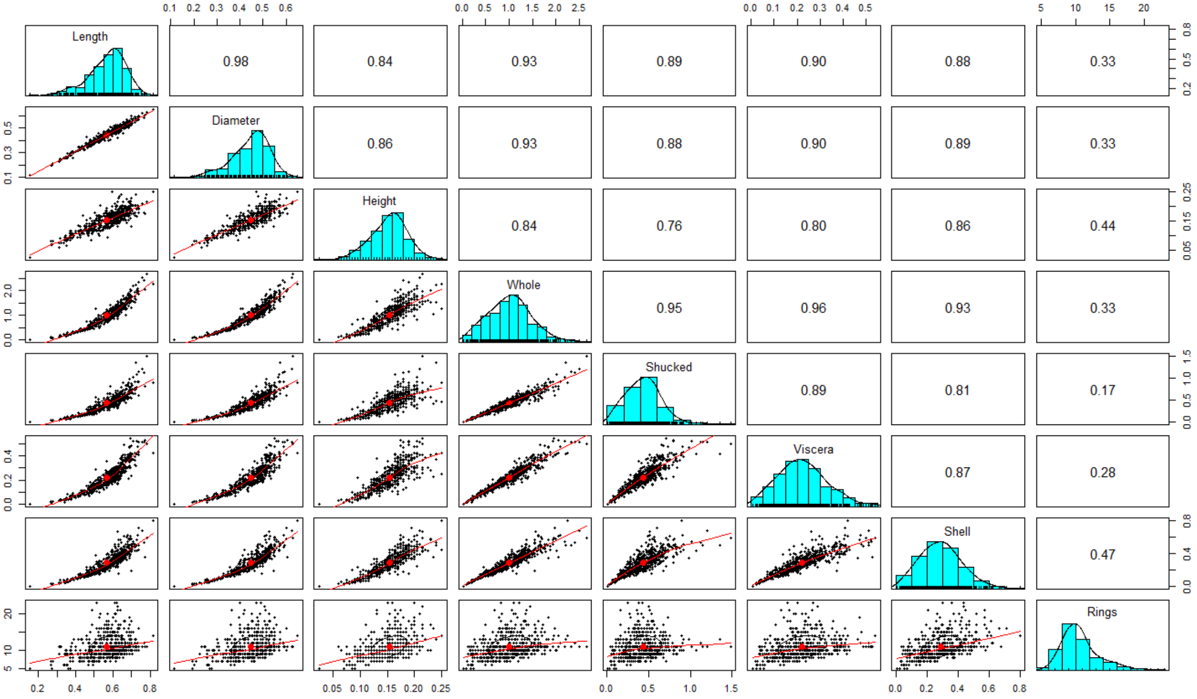

Figure 1: Pairwise Scatterplots, Distributions, and Correlations

The scatterplots for each pair of variables show that many of the relationships are linear. Specifically, the weight variables compared to the size variables show a tight upward pattern representing almost perfect linear relationships. This indicates that larger abalones predictably weigh more across each measure. In contrast, the scatterplots involving Rings are positive, but much weaker with more variability especially at higher ages. As for the histograms, the size variables show only mild skewness supporting that extreme sizes in abalone are uncommon. As for the weight variables, they display a clear visual representation of being right skewed. These histograms show that most abalone in the sample are light in weight. All physical attributes of the abalones including the size and weight variables show highly strong positive correlation between each other. These strong associations among the predictor variables are expected because the measurements reflect the size and will scale as the abalones grow larger. However, the correlation values of the physical characteristics to Rings show a much weaker correlation. While the values indicate age is positively related to size and weight, the relationship is moderate. Since Rings does not have a strong linear relationship with any predictor, we can assume the weak correlation is going to be represented by low R-squared values in the regression models.

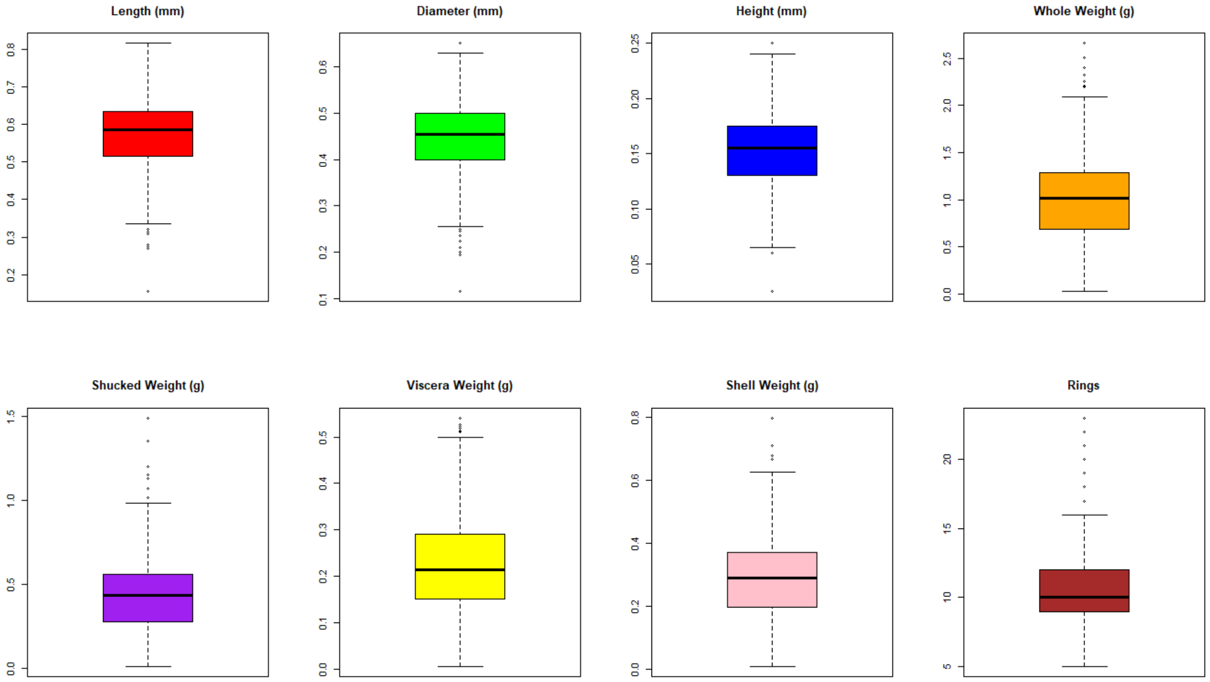

Figure 2: Boxplots of Physical Measurements and Rings

Outliers are present across all variables. For the size variables specifically, their distributions are fairly symmetrical with mostly a small number of low-end outliers. The greater variability and more prominent outliers in the weight variables are consistent with their right-skewed histograms.

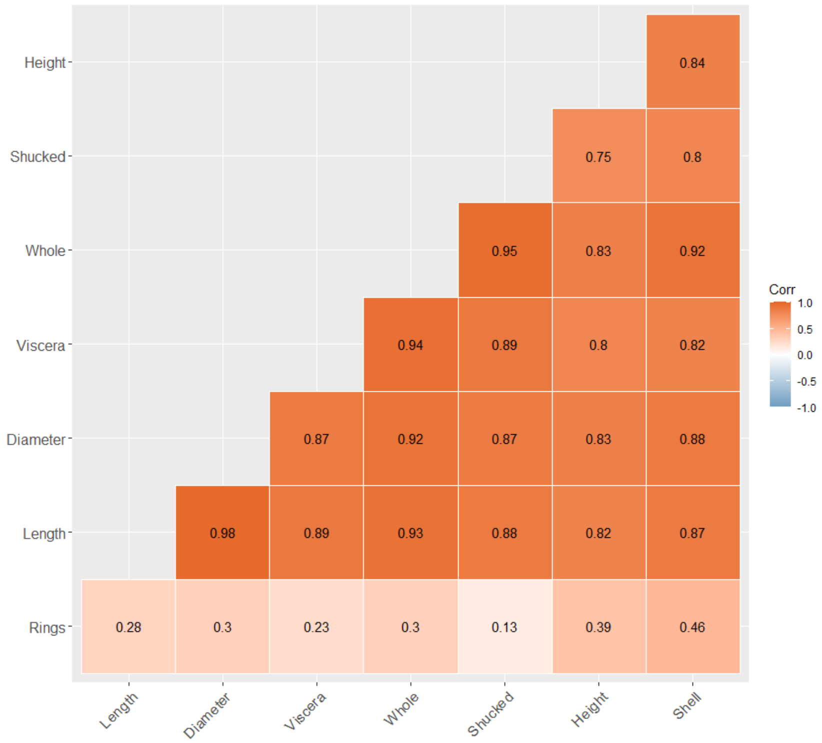

Figure 3: Correlation Heatmap

This heatmap is able to provide a clearer visualization of the strength of the linear relationships between the numerical variables in the dataset. As seen, the size variables and weight variables exhibit very strong positive correlations with one another, with most coefficients exceeding 0.85. This reflects reality that as abalones grow larger, every dimension and weight component increases. The high correlations between the weight and size variables raise a multicollinearity concern, while age continues to show weak correlations with the physical measurements.

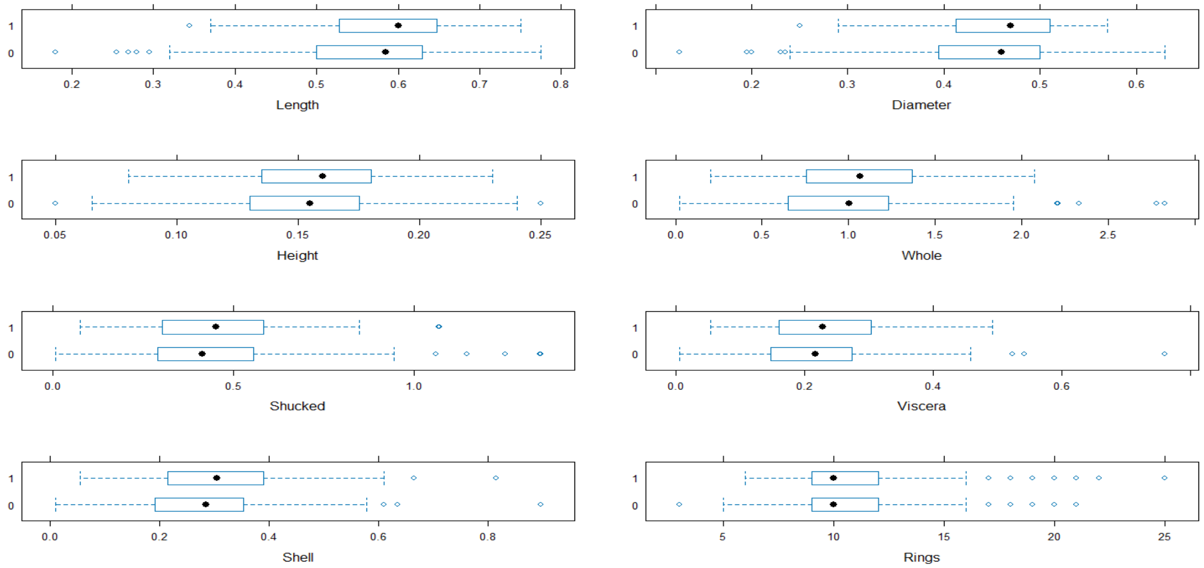

Figure 4: Boxplots by Sex (0=Male and 1=Female)

These boxplots are comparing the distributions of each of the numerical variables between male and female abalones. The plots provide insight into whether Sex may meaningfully contribute to predicting the response variable Rings. These plots also can show if Sex plays a major role in the varying sizes and weights of abalones. However, for the size measurements males and females have similar distributions with the medians and interquartile ranges being almost identical and both having a few outliers. As for the weight measurements, females actually appear only sightly heavier in some of the measures. This is only an extremely small difference, meaning it’s certainly not enough to demonstrate strong predictive power in the model. As for our response variable, Rings, the boxplot demonstrates nearly identical distributions between males and females. This suggests that Sex is not strongly associated with age.

Multiple Linear Regression

Fitted Model 1 (No Categorical Variable):

$$\widehat{Rings} = 6.991 - 3.991 Length + 2.864 Diameter + 21.556 Height + 10.537 WholeWeight \\ - 19.257 ShuckedWeight - 12.116 VisceraWeight + 6.815 ShellWeight$$

Variable	Estimate	Standard Error	T-value	P-value	Significance
Intercept	6.991	1.194	5.857	< 0.0001	***
Length	-3.991	6.008	-0.664	0.5068	
Diameter	2.864	7.265	0.394	0.6936	
Height	21.556	6.632	3.25	0.0012	**
Whole	10.537	1.966	5.36	< 0.0001	***
Shucked	-19.257	2.142	-8.992	< 0.0001	***
Viscera	-12.116	3.407	-3.557	0.0004	***
Shell	6.815	3.113	2.189	0.0291	*
R-squared: 40.9% Adjusted R-squared: 40.06%			F-statistic: 48.63 on 7 and 492 DF P-value: < 0.0001		

Table 2: Original Model without Categorical Variable

Fitted Model 2 (With Categorical Variable):

$$\widehat{Rings} = 6.9456 - 3.3385 Length + 2.5138 Diameter + 21.7161 Height \\ + 10.6687 WholeWeight - 19.5803 ShuckedWeight \\ - 12.1521 VisceraWeight + 6.7106 ShellWeight - 0.3092 Sex$$

Variable	Estimate	Standard Error	T-value	P-value	Significance
Intercept	6.9456	1.1926	5.824	< 0.0001	***
Length	-3.3385	6.0164	-0.555	0.5792	

Diameter	2.5138	7.2599	0.346	0.7293	
Height	21.7161	6.6249	3.278	0.0011	**
Whole	10.6687	1.9654	5.428	< 0.0001	***
Shucked	-19.5803	2.1497	-9.108	< 0.0001	***
Viscera	-12.1521	3.4023	-3.572	0.0004	***
Shell	6.7106	3.1101	2.158	0.0314	*
Sex	-0.3092	0.2065	-1.497	0.135	
R-squared: 41.16%			F-statistic: 42.94 on 8 and 491 DF		
Adjusted R-squared: 40.21%			P-value: < 0.0001		

Table 3: Original Model with Categorical Variable

The first regression model was fit using only continuous predictors to explain the variation in Rings. The significant predictors in this model are Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight. These predictors show meaningful linear associations with Rings. For example, the predictors Height and Whole Weight display positive effects suggesting the taller and heavier abalones tend to be older. Even though the variables Length and Diameter have strong correlation with the other physical measurements, they are not statistically significant in the model, meaning they do not contribute to predicting age. The first model explains approximately 40.9% of the variability in Rings, indicating that it does not do the best job. However, with an F-statistic value of 48.63 and low p-value, the regression model as a whole is highly significant. As for the second model, the same five predictors remained statistically significant. Also, the coefficients of predictors remained somewhat the same meaning the inclusion of Sex did not meaningfully alter the structure of the model. The Sex variable did not show up as statistically significant suggesting that female abalones do not differ significantly in age from males further supporting the boxplots. This model's R-squared value is only slightly higher than the first at 41.16% and adjusted R-squared similar. The F-statistic value of 42.94 and low p-value shows the regression model remains highly significant but the improvement from adding Sex is minimal. Overall, the models are significant, but it is likely that multicollinearity influences both models in the way that there is overlapping information among the predictors.

Regression Diagnostics in Multiple Linear Regression

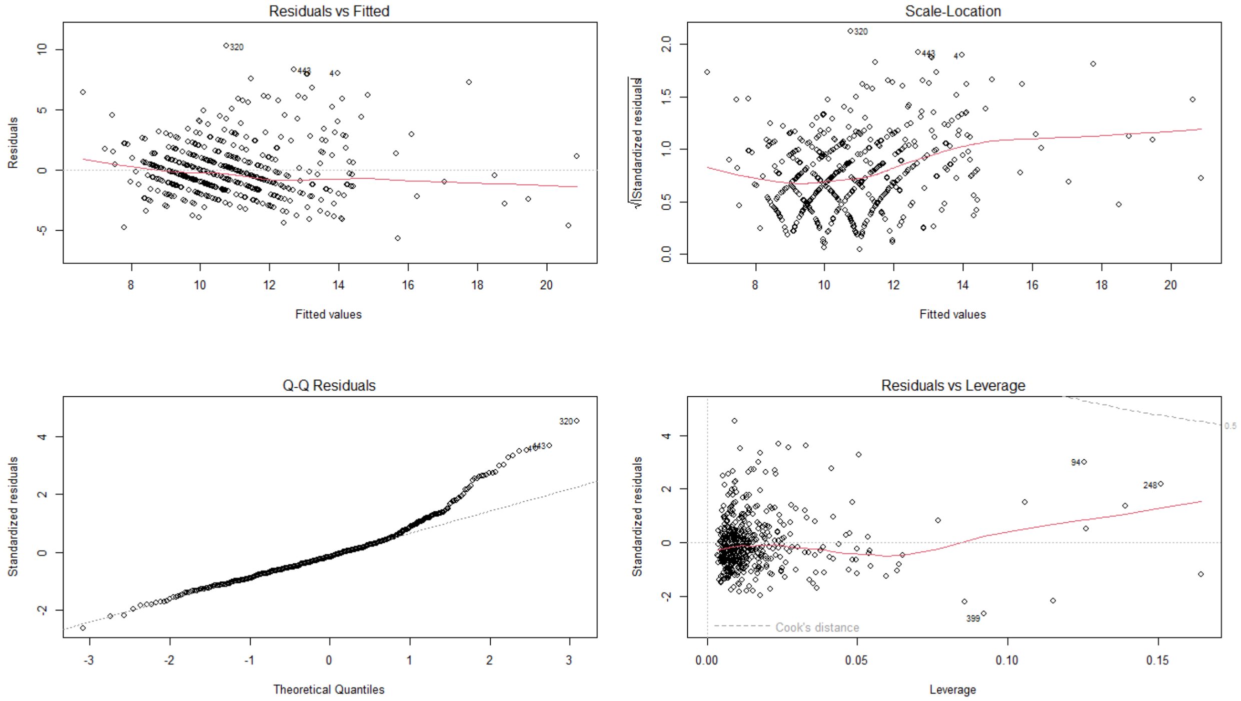

Figure 5: Diagnostic Plots for Model 1

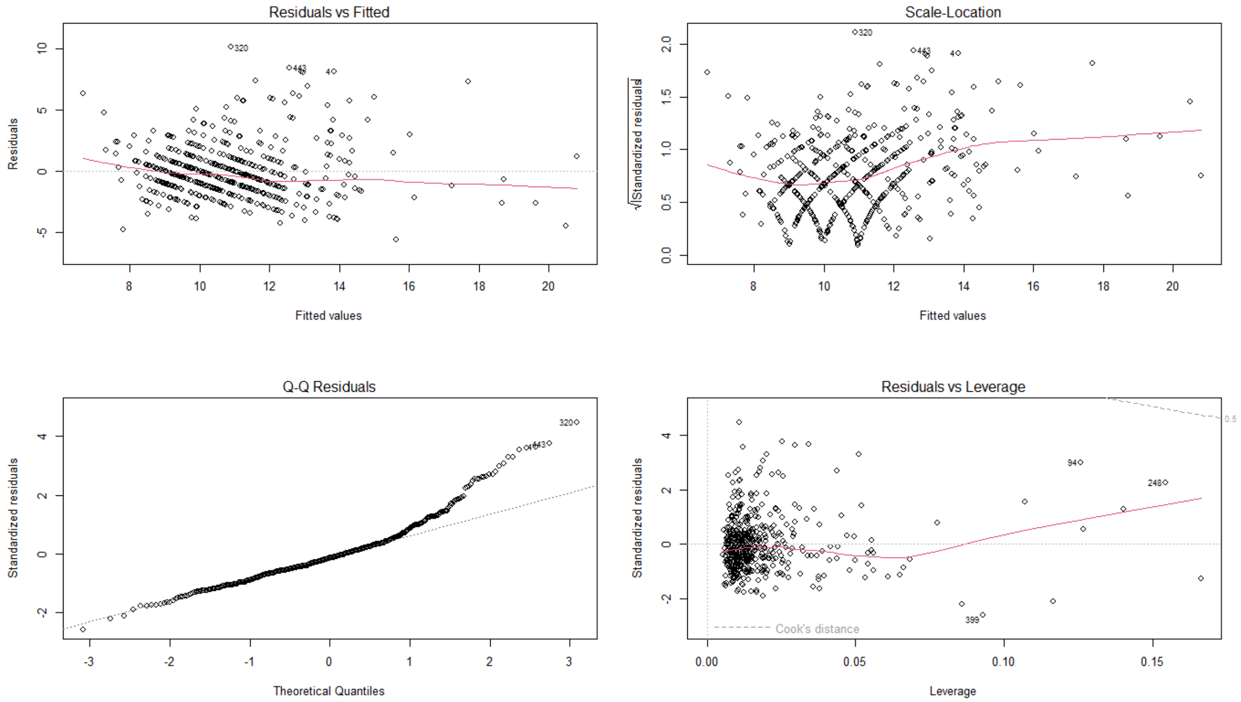

Figure 6: Diagnostic Plots for Model 2

These diagnostic plots for model 1 and model 2 are very similar with the one difference of the variable Sex being included in model 2. In the residuals vs. fitted plot, a cloud of residuals is apparent with a downward bending line. This pattern suggests the relationship between the predictors and Rings is not entirely linear. It shows a transformation of the response variable may be required. The scale-location plot confirms heteroscedasticity because the residuals have larger variability for larger predicted age values. In the Q-Q plot, residuals do follow the center line, but there is lots of deviation in the right tail confirming right skewness in the residuals and normality violations. The comparison between these two models based on the diagnostic plots confirm Sex does not address any of the regression assumption violations, and a transformation is needed.

Variable Transformation and Handling of Influentials

Fitted Model 3 (With Log Transformation of Response Variable):

$$\begin{aligned}Ri\hat{n}gs = 1.84003 - 0.28043Length + 0.71935Diameter + 1.92391Height \\ + 0.72701WholeWeight - 1.51841ShuckedWeight \\ - 0.76688VisceraWeight + 0.56785ShellWeight - 0.02946Sex\end{aligned}$$

Variable	Estimate	Standard Error	T-value	P-value	Significance
Intercept	1.84003	0.10187	18.063	< 0.0001	***
Length	-0.28043	0.51389	-0.546	0.5855	
Diameter	0.71935	0.62010	1.16	0.2466	
Height	1.92391	0.56586	3.4	0.0007	***
Whole	0.72701	0.16788	4.331	< 0.0001	***
Shucked	-1.51841	0.18362	-8.269	< 0.0001	***
Viscera	-0.76688	0.29061	-2.639	0.0086	**
Shell	0.56785	0.26565	2.138	0.033	*
Sex	-0.02946	0.01764	-1.67	0.0956	.
R-squared: 40.93% Adjusted R-squared: 39.97%			F-statistic: 42.53 on 8 and 491 DF P-value: < 0.0001		

Table 4: Log Transformed Model with Categorical Variable

To address the violations that we saw in the diagnostic plots from the previous two models, a log transformation was performed on the response variable Rings, and we refit a new model continuing to incorporate the categorical variable. The variables Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight remain statistically significant in the model. Length, Diameter, and Sex are not significant. Even after the transformation, the variables' significance in the model are unchanged. The transformed model yields almost similar R-squared and adjusted R-squared values to the first two models. This shows that the transformation does not increase the explanatory power of the model for the variation. However, the transformation was done not just to improve the R-squared values but also correct the assumption violations from the diagnostic plots. The F-statistic and p-value from this model confirm that the model remains statistically significant overall.

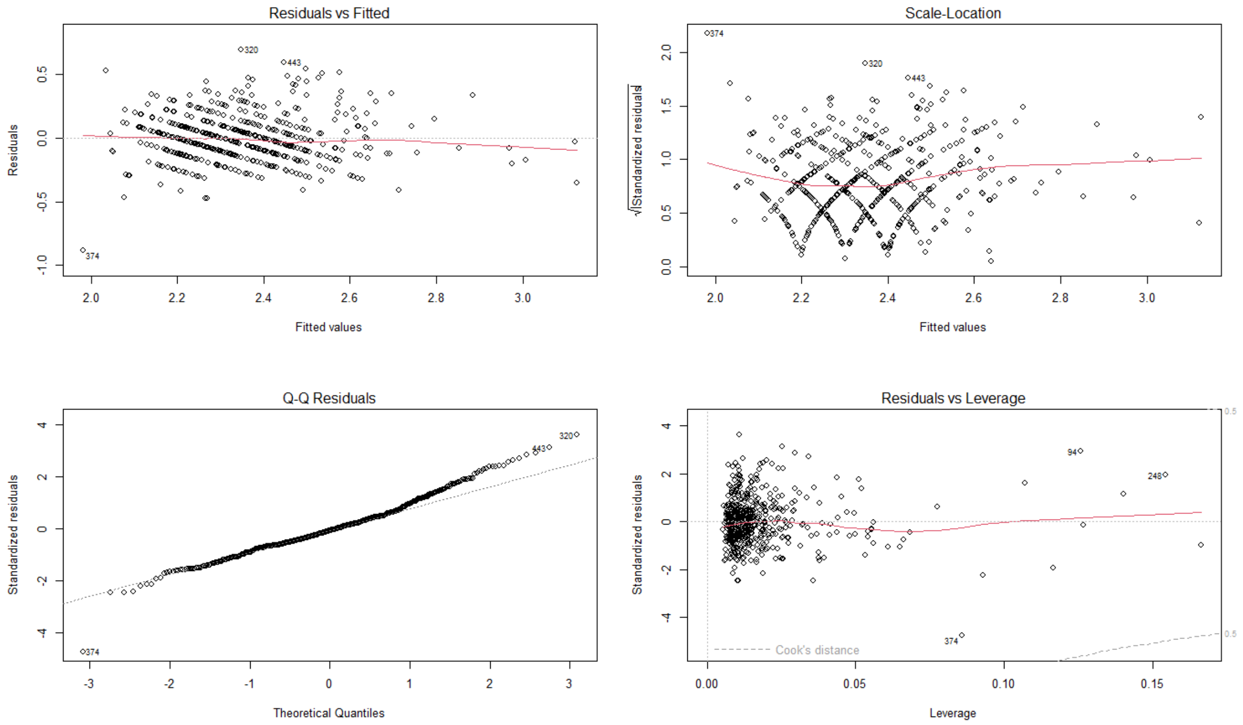

Figure 7: Diagnostic Plots for Model 3

After the log transformation, we can see a more centered and tighter plot of residuals around zero in the residuals vs. fitted plot. While there is an improvement, the relationship between predictors and age might not be fully linear. As for the scale-location plot, there is a noticeable reduction in heteroscedasticity from the log transformation resulting in a more constant variance across the fitted values. The transformation improved the residual normality slightly as seen in the Q-Q plot, but the heavy tail remains. The log transformation did not eliminate influential observations, suggesting the need for further change to the model for improvement.

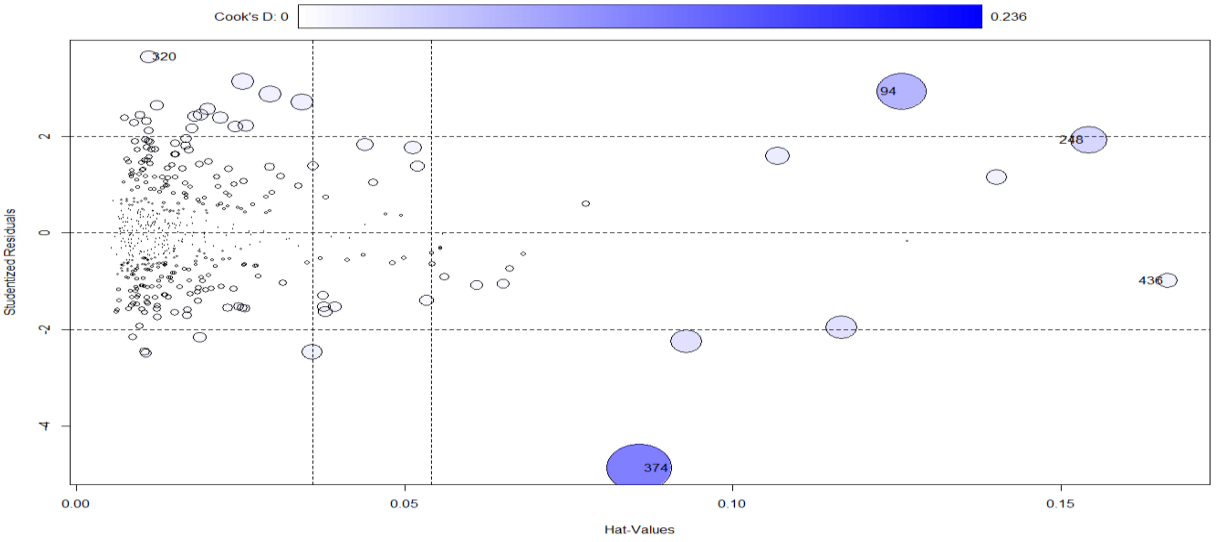

Figure 8: Influential Values Plot

From the plot of influential values, there are some observations that clearly stand out. Specifically, observation 374 that has a very large cook's distance significantly distorting the model. Furthermore, observations 94, 248, and 436 all represent high leverage and negatively influence the model's ability to predict. Removing the influential observations will allow the final model to better satisfy regression assumptions.

Fitted Model 4 (With Log Transformation of Response Variable and Influentials Removed):

$$\begin{aligned} Ri\hat{n}gs = 2.14221 - 0.67395 Length + 0.26813 Diameter + 1.62706 Height \\ + 0.86141 WholeWeight - 1.52609 ShuckedWeight \\ - 0.87776 VisceraWeight + 0.80345 ShellWeight - 0.04864 Sex \end{aligned}$$

Variable	Estimate	Standard Error	T-value	P-value	Significance
Intercept	2.14221	0.11062	19.365	< 0.0001	***
Length	-0.67395	0.48548	-1.388	0.1658	
Diameter	0.26813	0.59758	0.449	0.6539	
Height	1.62706	0.54415	2.99	0.0029	**
Whole	0.86141	0.21505	4.006	< 0.0001	***
Shucked	-1.52609	0.22573	-6.761	< 0.0001	***
Viscera	-0.87776	0.34064	-2.577	0.0103	*
Shell	0.80345	0.31173	2.577	0.0103	*
Sex	-0.04864	0.01558	-3.121	0.0019	**
R-squared: 40.95% Adjusted R-squared: 39.88%			F-statistic: 38.49 on 8 and 444 DF P-value: < 0.0001		

Table 5: Log Transformed Model without Influentials and with Categorical Variable

The removal of influential observations affected the model by having more stable standard errors for several predictors like Whole Weight and Shell Weight. The significance of variables remained the same with the level of significance varying slightly. There was the addition of the Sex predictor in becoming significant. Unfortunately, the R-squared and adjusted R-squared values remained extremely similar to the previous models indicating the model fit does not change drastically after removing influentials. While it would have been nice, the goal of removing the influential observations was not entirely to improve R-squared values, but to improve model validity, reduce bias, and strengthen the regression assumptions. Once again, the model remained highly significant overall proven in the F-statistic and p-value.

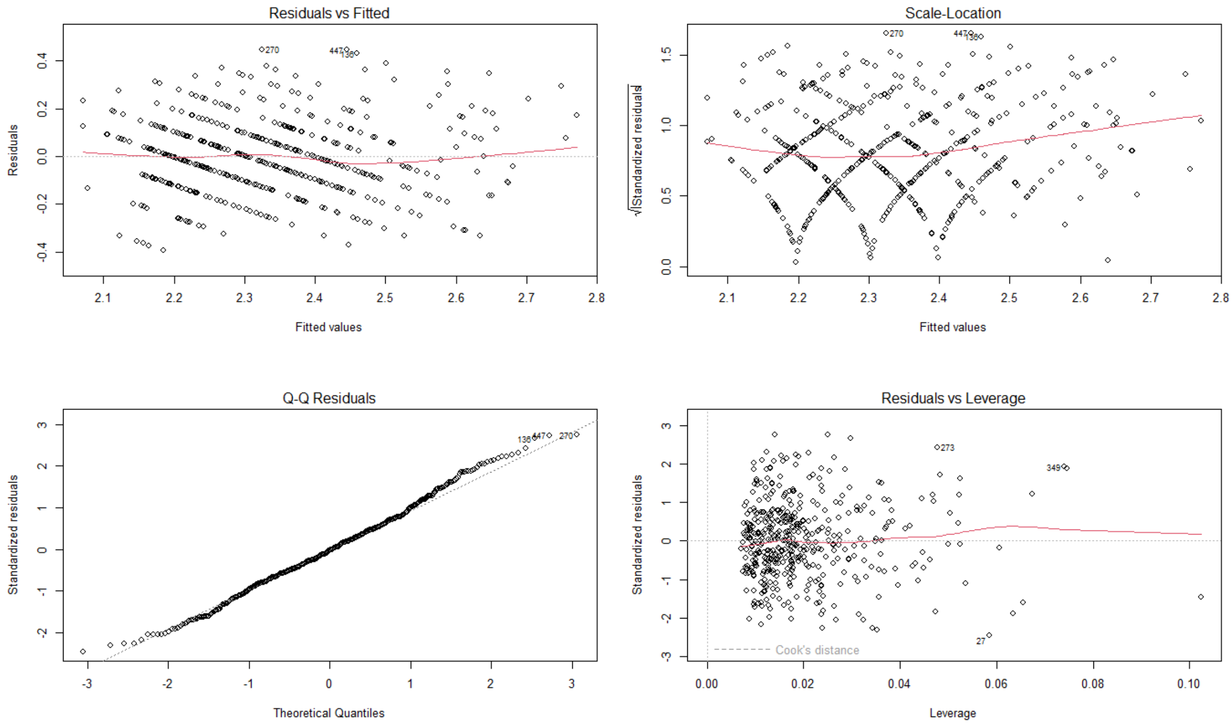

Figure 9: Diagnostic Plots for Model 4

After refitting the final model, an improvement in the diagnostic plots is clearly visible. The residuals vs. fitted plot shows a more symmetric distribution of residuals around zero. The linearity between predictors and the log transformed response variable has improved greatly. Although minor heteroscedasticity remains, there is a significant improvement in the final model. There is a major improvement in the Q-Q plot, which was the major focus throughout our study. The residuals fall closely along the line for majority of the distribution. The tail is notably better aligned compared to the previous models. After the removal of influential observations, the distribution is now recognized as close to normal. For the residuals vs. leverage plot, there are no observations that exceed the threshold of cook's distance. The plot recognizes that the high leverage influential points being removed helped eliminate the instability of the model. Overall, the log transformation and removal of influential observations helped improve the regression model greatly.

Conclusion

In summary, although our original and final model did not produce the highest R-squared values, it is completely normal given that a trait such as age is influenced by many unmeasured biological and environmental factors. The model still identifies several statistically significant predictors meaning it captures meaningful relationships even if it doesn't explain all the variability. Overall, this provides a strong starting model for predicting age of abalones without relying on the traditional time-consuming methods. With additional data beyond physical features such as genetics, environment, or growth conditions, this model could be further improved to make age prediction more accurate and reliable.

Works Cited

Monterey Bay Aquarium. (n.d.). Abalone. Monterey Bay Aquarium. Retrieved December 7, 2025, https://www.montereybayaquarium.org/animals/animals-a-to-z/abalone.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). Abalone [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.