Ethan Dao
UID: 206030580

Stats 101 Final Project: Predicting Abalone Age

1. Introduction and Cleaning/Transforming Data

The data set that I have used is the "Abalone" data set from the UCI machine learning repository, and it has nine variables; length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and number of rings. Each variable has 4,177 observations, and the dataset was taken from the UCI Machine Learning Repository.

Abalones are a group of marine gastropod mollusks, and we can deduce the age of abalones from counting the rings on their shells, as the number of rings + 1.5 gives the age of abalones in years. However, this is a tedious process, and instead of counting the rings for each individual abalone, my objective in this report is to see if there is a relationship between the ages of abalones and their physical characteristics. I will be using a multiple linear regression model to try and predict age from the eight other variables in the data set (length, diameter, height, whole weight, shucked weight, viscera weight, shell weight), which describe the physical characteristics of abalones.

My data had no missing values, so I did not have to clean my data or account for missing values. I decided to transform the sex and rings variable to fit a better linear model. I decided to split the sex into two categorical variables, infant and non-infant, as I would expect infants vs. non-infant abalones to have a greater difference in rings than non-infant males and females. We can visualize this using a box plot analyzing the descriptive statistics of abalones by category. As a result of this, we can change the 'Sex' column from M, F, and I to I (infant) and A (adult). Finally, I decided to change the 'Rings' column to 'Age' to better reflect my objective. Since an abalone's age is perfectly correlated to its number of rings (Age = Rings + 1.5), we can add 1.5 to every observation in the 'Rings' column to get the abalones' ages.

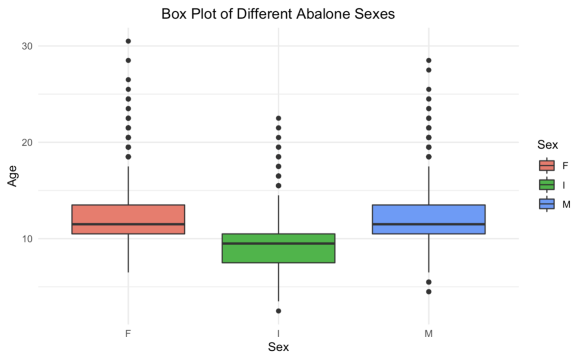


Figure 1.1: Box plot explaining descriptive statistics of different abalone sex

2. Descriptive Statistics

Now, I will analyze the descriptive statistics of my data set. Looking at the summary table of the cleaned data, we can get a better look at the central tendencies of the data, as well as an overview of the distribution and variables in the data. As for the age, the range is from 2.5 years to 30.5 years old, and the median and mean of the abalones' ages are 10.5 and 11.43, respectively. After looking at the scatterplots between each of the predictor variables and age, it appears that all of them have a positive correlation with age. However, we must transform and select our variables carefully in order to get a model that avoids multicollinearity, overfitting, and other problems so we can get the most accurate model to predict abalone age.

	Length	Diameter	Height	Whole.weight	Shucked.weight	Viscera.weight	Shell.weight	Age	Infant_Indicator
Min.	0.0750	0.0550	0.0100	0.0020	0.0010	0.0005	0.0015	2.50	0.000
1st Quad.	0.4500	0.3500	0.1150	0.4422	0.1862	0.0935	0.1300	9.50	0.000
Median	0.5450	0.4250	0.1400	0.8000	0.3360	0.1710	0.2340	10.50	0.000
Mean	0.5421	0.4079	0.1396	0.8290	0.3595	0.1807	0.2388	11.44	0.321
3rd Quad.	0.6150	0.4800	0.1650	1.1535	0.5020	0.2530	0.3287	12.50	1.000
Max	0.8150	0.6500	1.1300	2.8255	1.4880	0.7600	1.0050	30.50	1.000

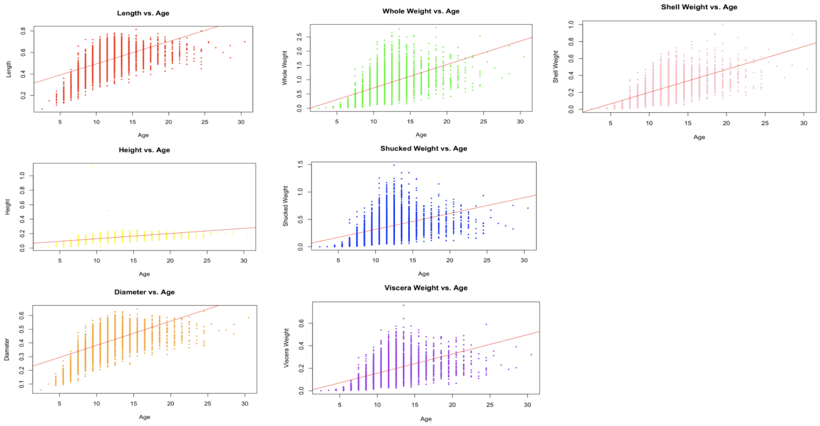


Figure 2.1: Table of descriptive statistics for each predictor variable in dataset
Figure 2.2: Scatter plot of each predictor variable vs. abalone age

3. Choosing our predictors and building our model

Method 1: Correlation Matrix

To choose my predictors, I will be using a correlation matrix to weed out variables highly correlated to each other, then using a backwards stepwise regression method to choose only predictor variables that will minimize AIC to prevent overfitting and multicollinearity in the model. Looking at the correlation between each of the variables and age, each variable has a correlation greater than +-0.4, so we will keep all of the predictor variables.

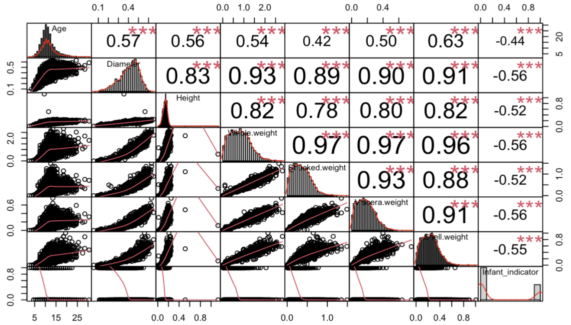
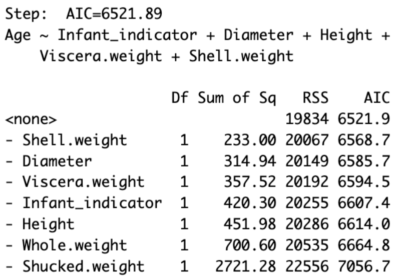


Figure 3.1: Correlation graph between each predictor variable

However, we run into a problem where many of the variables are highly correlated with each other. After doing some more investigation, it looks like the whole weight, shucked weight, viscera weight, and shell weight are very highly correlated, likely due to the fact that they are all components of an abalone's weight. We can keep the shell weight, since it has the highest correlation with age when looked at individually. A similar instance happens between the length and diameter variables, since they are both measures of the size of an abalone. We keep the diameter variable, since it has a higher individual correlation with age.

Method 2: Backwards Stepwise Regression

When we perform backwards stepwise regression on the initial model, length is removed as a predictor variable to minimize the AIC, which is consistent with what we did when we selected our variables using the correlation matrix. However, after removing the length, shucked weight, shell weight, and viscera weight variables, we want to make sure that there are no other variables that increase the AIC. After performing the stepwise regression again on our filtered data, we see that we do not have to remove any more variables, and although it is still not ideal, we see a drastic decrease in the VIF when compared to the VIF before removing variables with high correlations, which improves the multicollinearity issue we had in our data.



```
Step:  AIC=6521.89
Age ~ Infant_indicator + Diameter + Height + Whole.weight + Shucked.weight + 
    Viscera.weight + Shell.weight

                   Df Sum of Sq   RSS    AIC
<none>                           19834 6521.9
- Shell.weight      1    233.00 20067 6568.7
- Diameter          1    314.94 20149 6585.7
- Viscera.weight    1    357.52 20192 6594.5
- Infant_indicator  1    420.30 20255 6607.4
- Height            1    451.98 20286 6614.0
- Whole.weight      1    700.60 20535 6664.8
- Shucked.weight    1   2721.28 22556 7056.7
```

Diameter	Height	Shell Weight	Infant_indicator
5.538278	4.229334	4.558694	1.495239

Figure 3.2: Backwards stepwise regression of initial linear regression model, VIF after variable selection

4. Model Diagnostics and Transformations

When running model diagnostics on the original multiple linear regression model after variable selection, the graph of the residuals vs fitted values and the standardized residuals vs leverages followed a pattern, as there seemed to be non-constant variances and outliers in the data the way that the data was modeled.

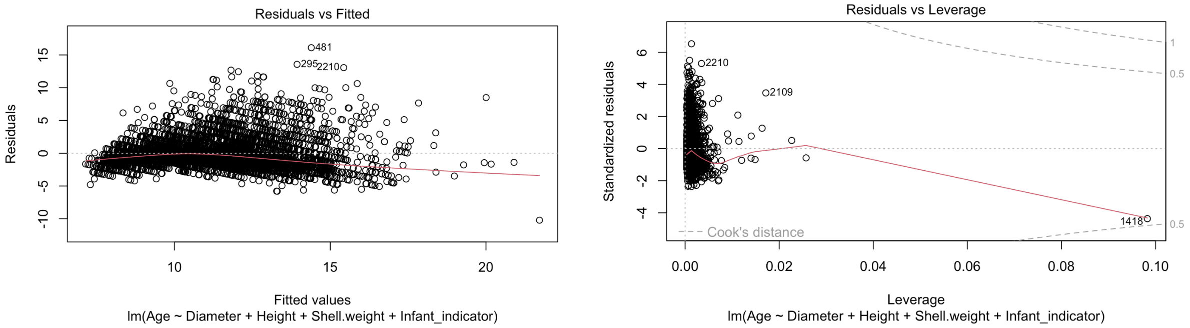


Figure 4.1: Residual graphs before log transformation

However, after transforming both the explanatory and predictor variables (with the exception of the indicator function) using a log transformation and running the model diagnostics again, significant improvements were seen in the plots of the graphs. Although the plot of the standardized residuals vs leverages was still not quite a straight line, the leverage plot followed less of a significant downwards pattern, and I was able to make the leverage point less of an outlier, resulting in a "nicer" graph. Additionally, the other three model diagnostics (QQ plot to check normality, residuals graphs to check for constant variance and normal distribution of error terms of error terms) followed a straight line, and therefore my model was able to follow the assumptions.

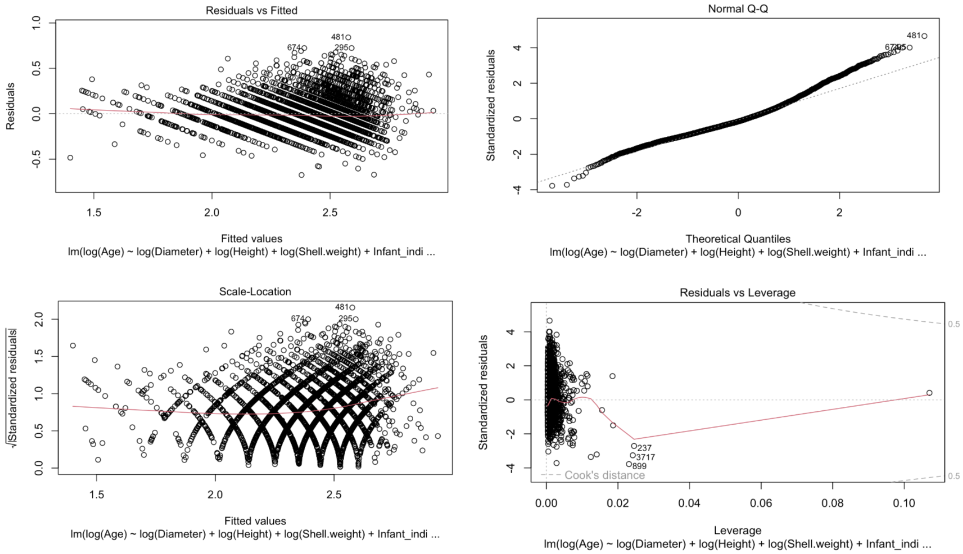
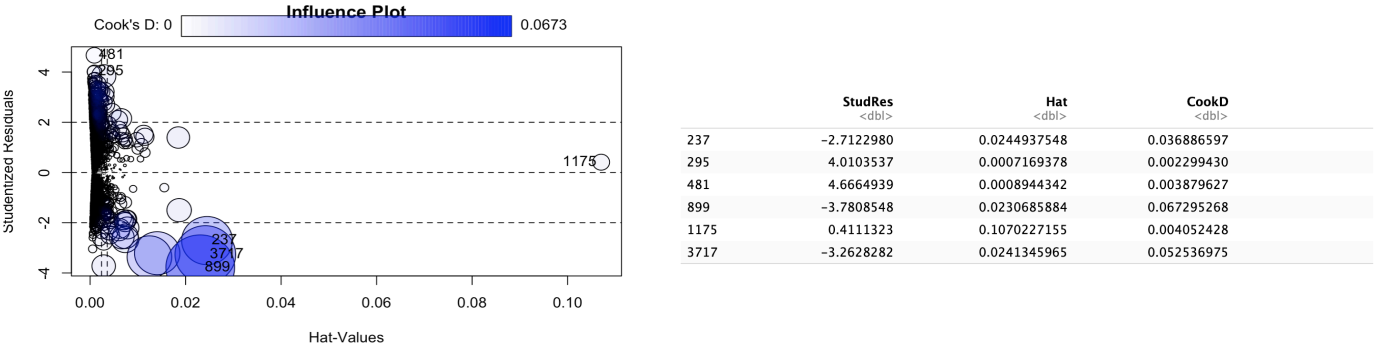


Figure 4.2: Model diagnostic graphs after log transformation

In the model diagnostic process, I also scanned for high leverage points, influential points, and outliers using various thresholds according to my data. Since I was dealing with a bigger dataset, standardized residuals outside the range (-4, 4) would be considered as outliers, and I used the leverage threshold and Cook's distance to find out the leverage and influential points. According to my calculations, as well as the influence plot in the "car" package, there were six "unusual" points in my data after scanning for leverage and influential points and outliers.



	StudRes <dbl>	Hat <dbl>	CookD <dbl>
237	-2.7122980	0.0244937548	0.036886597
295	4.0103537	0.0007169378	0.002299430
481	4.6664939	0.0008944342	0.003879627
899	-3.7808548	0.0230685884	0.067295268
1175	0.4111323	0.1070227155	0.004052428
3717	-3.2628282	0.0241345965	0.052536975

Figure 4.3: Influence plot, unusual observations in data set

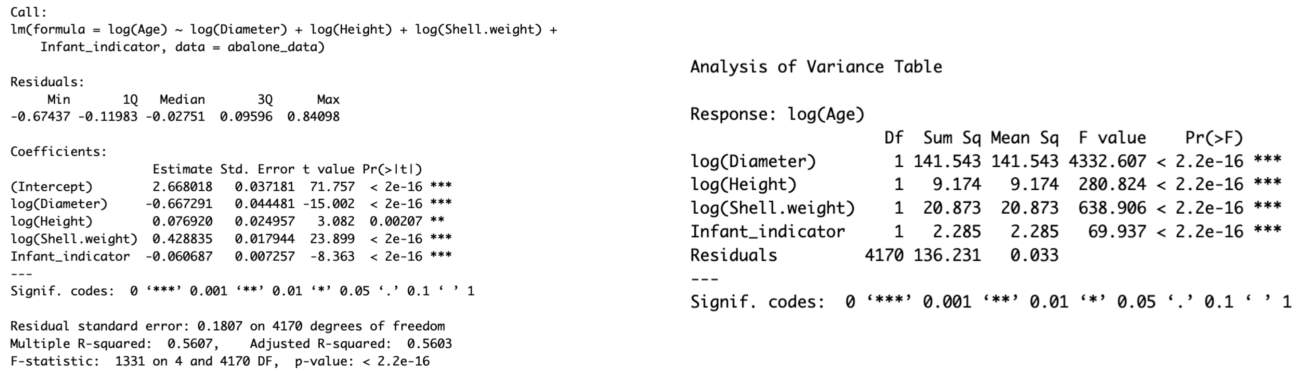
After further investigation of these points in the dataset, I came to the conclusion that three of the data points were infants below the first quartile in terms of age, which meant deviations in physical characteristics would have a greater impact on age, leading to the data points being more prone to being outliers. The other three data points were genuine outliers in terms of age vs. physical characteristics, as they had bigger measurements compared to their age or vice versa.

5. Fitting and Interpreting Our Multiple Linear Regression Model, F-Test and T-Test

After cleaning and investigating the data, choosing our predictors and variables, and assessing model diagnostics for our multiple linear regression model, we get the final model:

$$\log(Age) = 2.66802 - 0.66729\log(Diameter) + 0.07692\log(Height) + 0.428835\log(Shell\_weight) - 0.060687(Infant\_indicator)$$

To further assess our model and the significance of our model, we can look at the model summary, or conduct a T-test, and analyze the different metrics within the summary. Additionally, we can conduct an F-test on our model to assess the significance of the variance in the response variable.



```
Call:
lm(formula = log(Age) ~ log(Diameter) + log(Height) + log(Shell.weight) + 
    Infant_indicator, data = abalone_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67437 -0.11983 -0.02751  0.09596  0.84098 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        2.668018   0.037181  71.757  < 2e-16 ***
log(Diameter)     -0.667291   0.044481 -15.002  < 2e-16 ***
log(Height)        0.076920   0.024957   3.082  0.00207 ** 
log(Shell.weight)  0.428835   0.017944  23.899  < 2e-16 ***
Infant_indicator  -0.060687   0.007257  -8.363  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1807 on 4170 degrees of freedom
Multiple R-squared:  0.5607,	Adjusted R-squared:  0.5603 
F-statistic:  1331 on 4 and 4170 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: log(Age)
                    Df  Sum Sq Mean Sq  F value    Pr(>F)    
log(Diameter)        1 141.543 141.543 4332.607 < 2.2e-16 ***
log(Height)          1   9.174   9.174  280.824 < 2.2e-16 ***
log(Shell.weight)    1  20.873  20.873  638.906 < 2.2e-16 ***
Infant_indicator     1   2.285   2.285   69.937 < 2.2e-16 ***
Residuals         4170 136.231   0.033                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.1: Model Summary (T-Test) and F-Test Pertaining to Multiple Linear Regression Model

Looking at the model summary/T-test, we see that each of the predictor variables have their respective coefficients (as seen in the equation), and that each predictor variable also has a very small P-value, indicating that the logarithms of diameter, height, shell weight, and infancy are all individually statistically significant in their relation to the logarithm of abalone age. Additionally, the adjusted R^2 value of the model is 0.5603, meaning that 56.03% of the variation in an abalone's age can be explained by our model after adjusting to account for the number of predictors. We can also look at the residuals and residual standard error in the model summary. The residuals in our model range from -0.67437 to 0.84098, which indicate the extent to which the model's predictions deviate from the actual values.

Looking at the ANOVA/F-test for our model, we see that all predictor variables are highly significant predictors of abalone age. The model also provides a good fit to the data, which can be shown by the highly significant F-values and low P-values for each predictor variable. Additionally, the residual mean square is 0.033, which is the average unexplained variance in the response variable (abalone age) after accounting for all the predictor variables.

6. Conclusion

Overall, I was able to construct a decent multiple linear regression model that predicted an abalone's age based on its physical characteristics. The greatest obstacle during this process was the multicollinearity of my dataset. Although individually each predictor variable was successful in predicting abalone age, there was a lot of correlation between the predictor variables. This meant variable selection was a very tedious, difficult task given the chosen dataset–even after variable selection, all VIF values were not below 5. However, I was able to reduce the VIF and correlation coefficients drastically compared to the initial linear regression model with all the predictors, and by transforming my variables using a log transformation, I was able to greatly improve the model diagnostics and create a model that fit the data and completed the objective at hand. After building our model and analyzing the summary and ANOVA tables, we can conclude that the model is significant and valid, and is sufficient for predicting abalone age using our chosen predictor variables (diameter, height, shell weight, and infancy).