MSDS 6372 Project 1

Abalone Age Prediction Using Multiple Linear Regression

Team members:

Brian Kruse, Christopher Boomhower, Andrew Abbott, Johnny Quick

Date: 10/8/2016

Introduction

Abalones are a marine gastropod mollusk, or sea snail, found throughout the world. Abalone meat has been harvested for human consumption and the shells and pearls have been used for decorative purposes for thousands of years. In recent decades, abalone populations have been threatened by overfishing and mismanagement. To help countries determine the overall health of the abalone population it is important to be able to accurately determine the age of a specific specimen. The standard method for determining the age of an abalone has been to count the number of layers in its shell. This is a tedious process involving cutting a sample from the shell, staining it, and using a microscope to count the number of layers. Drilling into the shell requires the use of protective equipment to prevent inhalation of abalone shell dust which is unhealthy for human beings. A simpler and safer method for estimating age would contribute to the survival of the abalone population. With this in mind, the goal of this paper is to use more easily measurable physical characteristics of abalone to estimate the age of an abalone, which is the number of rings plus 1.5, using multiple linear regression (MLR).

This paper is organized as follows: the descriptive statistics section introduces the data set. Variables are defined, summary statistics are provided and outliers are addressed. The next section describes the analysis. Assumptions are addressed and variable reduction techniques described. The final model is derived here as well. Lastly, the interpretation section contains the statistical and contextual interpretation of the model.

Descriptive Statistics

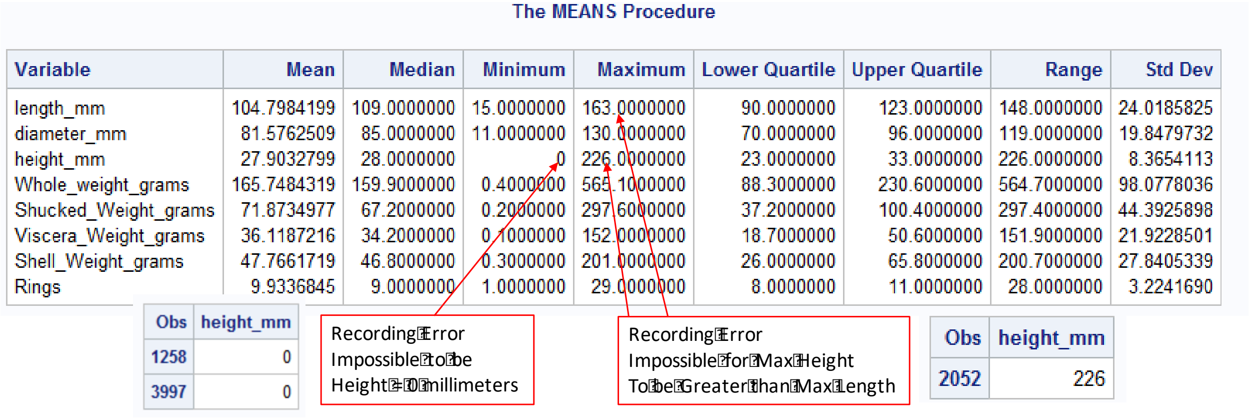
This Blacklip Abalone observational study comes from UCI Machine Learning laboratory. The data originate from a 1995 research study that contains 4,177 observations; 1,307 Female, 1,342 Infant, and 1,528 Male abalone were harvested in North Coast and Isles of Bass Straight Tasmania (DS1 in Appendix). The abalone data set contains variables below that are relevant to the Tasmanian shellfish industry.

Table 1. Abalone Data Set Variables

Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

The procedure to obtain the variables consists of two common processes for the shellfish industry. Licensed divers extract the abalone from the ocean floor, place the abalone in mesh containers, and transport the subjects to a processing facility. The processing facility measures and weighs the whole abalone. Next, the processor shucks and weighs the abalone meat. The processor removes and weighs the gut. The researcher would drill, stain and identify the number of rings in the shell (DS2).

Before analysis could commence, our team multiplied all continuous variables by 200 to reflect the data as represented within the original research (Original data divided by 200 for measurement tool compatibility). The summary statistics including recording errors are identified below (SAS_DS1).



The MEANS Procedure

Variable	Mean	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range	Std Dev
length_mm	104.7984199	109.0000000	15.0000000	163.0000000	90.0000000	123.0000000	148.0000000	24.0185825
diameter_mm	81.5762509	85.0000000	11.0000000	130.0000000	70.0000000	96.0000000	119.0000000	19.8479732
height_mm	27.9032799	28.0000000	0	226.0000000	23.0000000	33.0000000	226.0000000	8.3654113
Whole_weight_grams	165.7484319	159.9000000	0.4000000	565.1000000	88.3000000	230.6000000	564.7000000	98.0778036
Shucked_Weight_grams	71.8734977	67.2000000	0.2000000	297.6000000	37.2000000	100.4000000	297.4000000	44.3925898
Viscera_Weight_grams	36.1187216	34.2000000	0.1000000	152.0000000	18.7000000	50.6000000	151.9000000	21.9228501
Shell_Weight_grams	47.7661719	46.8000000	0.3000000	201.0000000	26.0000000	65.8000000	200.7000000	27.8405339
Rings	9.9336845	9.0000000	1.0000000	29.0000000	8.0000000	11.0000000	28.0000000	3.2241690

Obs	height_mm
1258	0
3997	0

Recording Error
Impossible to be
Height = 0 millimeters

Recording Error
Impossible for Max Height
To be Greater than Max Length

Obs	height_mm
2052	226

Figure 1. Abalone Data Set Descriptive Statistics

The following frequency distributions identify that all variables are not normally distributed (SAS_DS2).

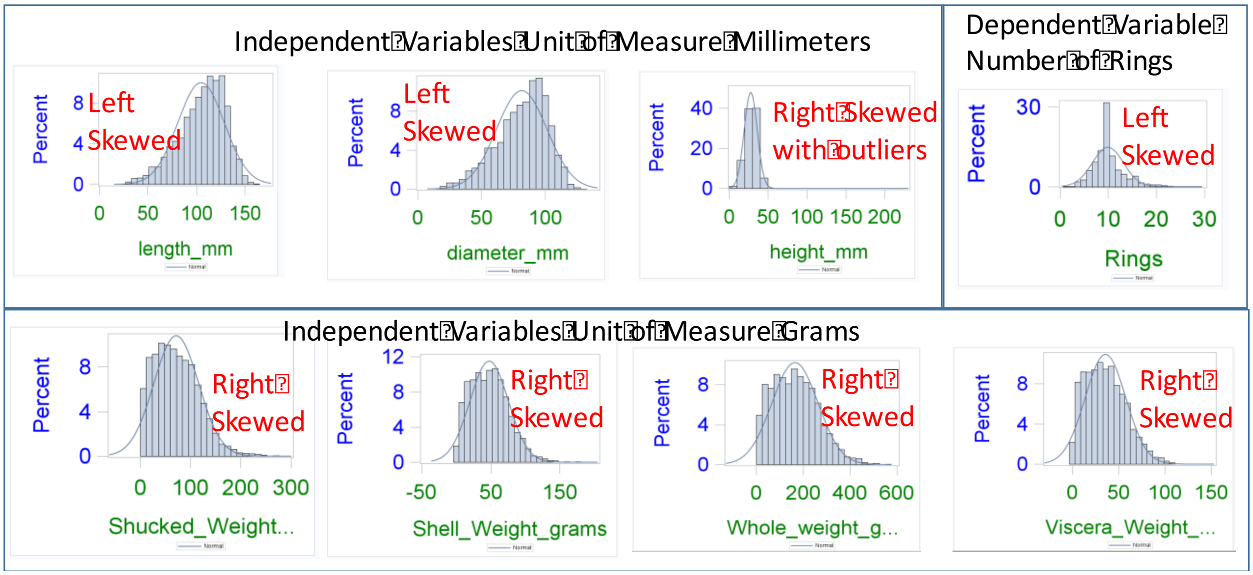


Figure 2. Frequency Distribution of Variables

Analysis

Now that relevant background has been provided and all variable details discussed, we continue our analysis to address whether these variables may be utilized for abalone age predictions, and if so, to determine an appropriate model. Before doing so, however, it is needful to check all assumptions for MLR. These assumptions are data normality, independence, linearity, and constant variance.

The first assumption to be analyzed, which was discussed briefly in the previous section, is normality within the data. Based on previous discussions surrounding the variables present in this data set (See Figure 2), it is clear that all variables exhibit *some* form of skewness. *Length*, *Height* (with the two largest outliers excluded), and *Diameter*, while only slightly skewed, are mostly normally distributed; so for the purposes of this analysis no transformation is deemed necessary. Weight variables and the rings & years response, on the other hand, demonstrate right skewness that would benefit from a square root transformation. Said transformation is applied and data normalization, by variable, confirmed (See Figure 8 in Appendix). Though the square root transformation does not coerce the data into perfect normality, weight and rings & years distributions are far more normal than they are without the square root computation.

Normality is further assessed via a saturated MLR model Q-Q plot and histogram with fitted normal curve, as provided in Figure 3 (The saturated model includes all explanatory variables and all observations except those with height of 0 mm). The Q-Q plot follows a relatively straight line until it begins curving upward just above its positive t-quartile of 1. This shift in slope is most likely due to the still present, slight skewness observed in the explanatory variable data. Support for this theory is demonstrated in the histogram of studentized residuals which shows that while the model's data is mostly normal, there is some slight right-skewness present still. Though not perfect, these checks on normality are judged sufficient for our prediction on abalone age.

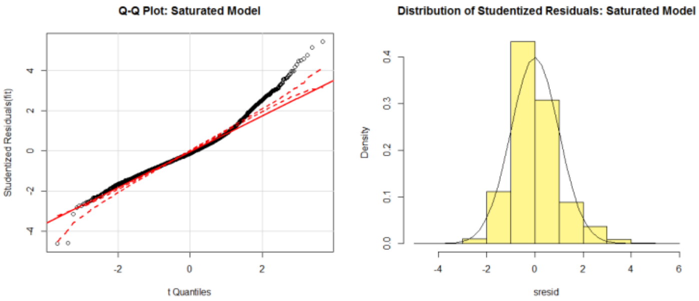


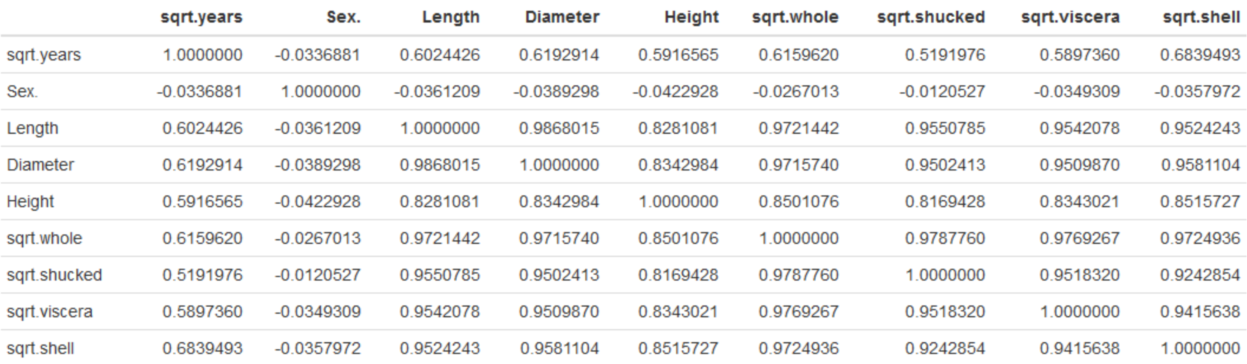
Figure 3. Q-Q Plot and Histogram of Studentized Resuiduals for Assumptions Check

The second assumption, independence between observations and independence between explanatory variables, is analyzed next. As discussed in the previous section, these measurements were collected on 4,177 independent abalone. However, one would expect dimensional and weight data to be rather correlated for any given abalone specimen. This is simply a logical assumption that, for example, as abalone sizes increase, their weights likely increases as well. As further example, if abalone shucked weight is large, whole weight is likely to be large too.

These correlation assumptions are assessed in Table 2 below. Meeting expectations, greatest correlation occurs between dimension and weight explanatory variables. So, even though observations are independent from one another, collinearity is clearly present in the data. Normally, this would be problematic as it limits the potential for model application to other abalone populations and datasets. Any changes in environmental conditions, food supplies, etc., could mean an over-fitted model. However, due to the scope of this study and the expected inferences to this abalone population alone, such collinearity must be accepted and the MLR analysis is allowed to proceed as planned.

It is worth noting that each individual explanatory variable in Table 2, except *Sex*, strongly correlates to the response, *sqrt.years*, on its own. In fact, *sqrt.shell* alone has a positive correlation with *sqrt.years* of r = 0.6839 (Coefficient of Determination, R^2, is 0.4677) . It is advised to bear this number in mind as MLR model refinement and final generation is described further.

Table 2. Independent-Independent and Independent-Dependent Correlation



	sqrt.years	Sex.	Length	Diameter	Height	sqrt.whole	sqrt.shucked	sqrt.viscera	sqrt.shell
sqrt.years	1.0000000	-0.0336881	0.6024426	0.6192914	0.5916565	0.6159620	0.5191976	0.5897360	0.6839493
Sex.	-0.0336881	1.0000000	-0.0361209	-0.0389298	-0.0422928	-0.0267013	-0.0120527	-0.0349309	-0.0357972
Length	0.6024426	-0.0361209	1.0000000	0.9868015	0.8281081	0.9721442	0.9550785	0.9542078	0.9524243
Diameter	0.6192914	-0.0389298	0.9868015	1.0000000	0.8342984	0.9715740	0.9502413	0.9509870	0.9581104
Height	0.5916565	-0.0422928	0.8281081	0.8342984	1.0000000	0.8501076	0.8169428	0.8343021	0.8515727
sqrt.whole	0.6159620	-0.0267013	0.9721442	0.9715740	0.8501076	1.0000000	0.9787760	0.9769267	0.9724936
sqrt.shucked	0.5191976	-0.0120527	0.9550785	0.9502413	0.8169428	0.9787760	1.0000000	0.9518320	0.9242854
sqrt.viscera	0.5897360	-0.0349309	0.9542078	0.9509870	0.8343021	0.9769267	0.9518320	1.0000000	0.9415638
sqrt.shell	0.6839493	-0.0357972	0.9524243	0.9581104	0.8515727	0.9724936	0.9242854	0.9415638	1.0000000

The Variable Inflation Factor (VIF) values of Table 3 also support this notion of multicollinearity within the saturated model. Essentially, a variable's high VIF indicates strong correlation to other explanatory variables and suggests redundancy in MLR explanatory variable selection. In this case, several variables exhibit high VIF values but *sqrt.whole* portrays the highest, implying redundancy in *sqrt.whole*'s inclusion. As will be explained during model refinement, one of the first orders of business will be to remove *sqrt.whole* from the model and observe the effects on the remaining variables' VIF values.

Table 3. Saturated Model VIF Values for Each Explanatory Variable

	male	**female**	**Length**	**Diameter**	**Height**	**sqrt.whole**	**sqrt.shucked**	**sqrt.viscera**	**sqrt.shell**
VIF	1.90	2.01	45.41	45.15	3.79	158.36	40.35	25.08	33.94

The third assumption is that the relationship between the explanatory variables and the response variable is linear. This assumption is best tested via scatterplots depicting the explanatory variables on the x-axis and response on the y-axis, as shown in Figure 4 (*Sex* variable levels excluded since *Sex* is an indicator variable). Not only do all variables elicit a linear relationship with *sqrt.years* (when excluding two largest *Height* outliers as shown in bottom two plots), which is the key assumption parameter to be met, they all demonstrate strong, positive correlation as mentioned above.

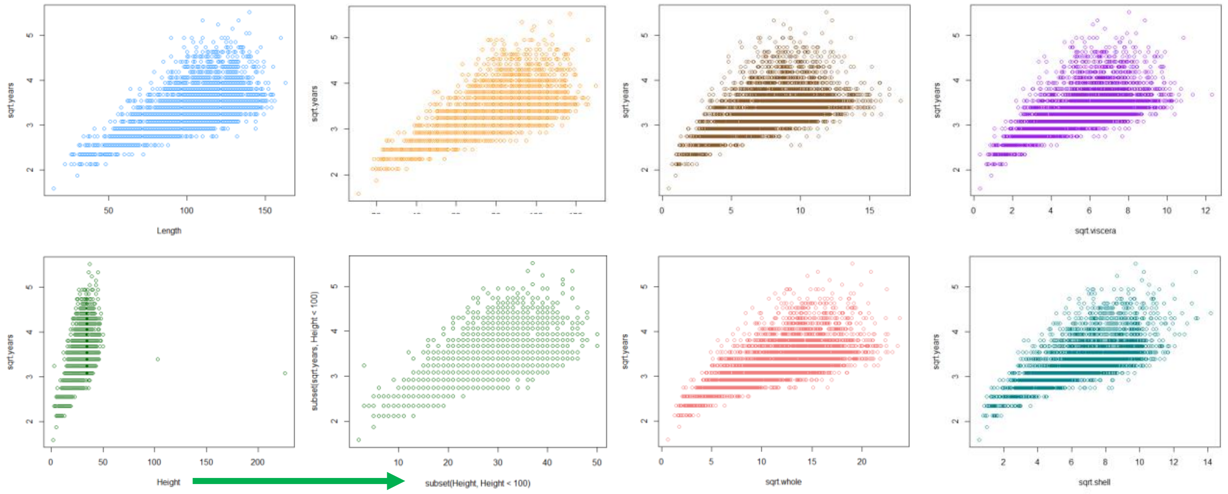


Figure 4. Scatterplots of Continuous Explanatory Variables vs. sqrt.years Response (Height Outliers Included in Plot)

The final assumption check is that of constant variance. This is a check to ensure that for every value of X, the spread of Y around the Y mean is the same. Partial residual plots are used to compare each individual explanatory variable variance as shown in Figure 5. While weight variance remains mostly constant, *Length*, *Diameter*, and *Height* variance seems to increase slightly with increasing parameter value. This is somewhat expected given the slight left-skewness discussed previously. Relieving this departure from assumption, the residuals vs. fitted or predicted values plot on the right within Figure 5 indicates that with the exception of some key outliers, saturated model variance as a whole is mostly constant. As will be observed, constant variance will improve as the MLR model is refined.

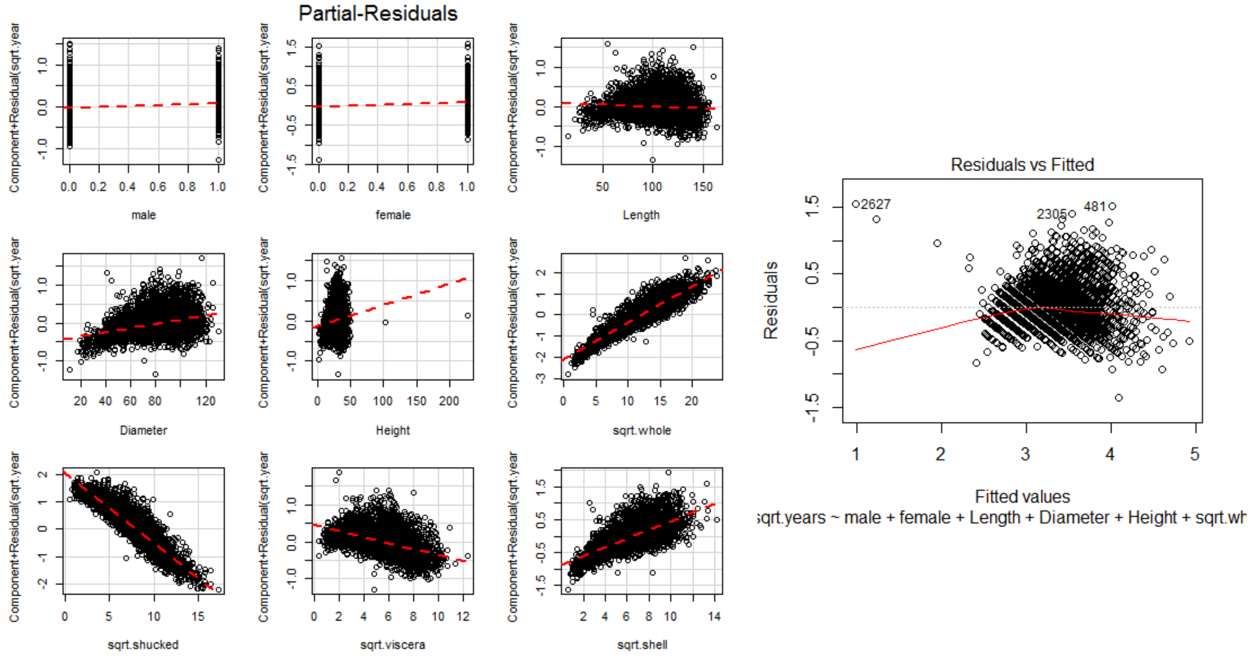
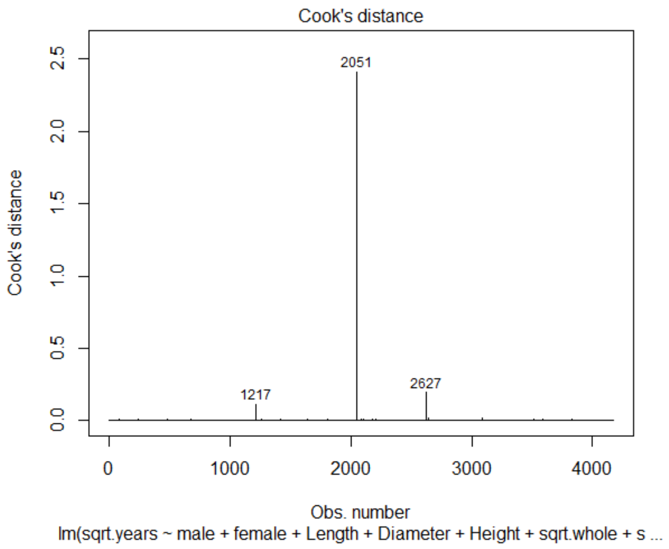
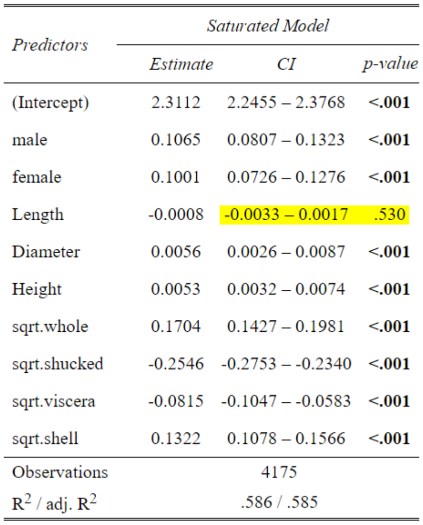


Figure 5. Partial Residuals (Left) and Residuals (Right) for the MLR Saturated Model

The final steps before model refinement are to observe the true impact of Observation 2052 on the MLR model and to provide a preliminary assessment of parameter estimate significance. As displayed in the Cook's D plot of Figure 6, Observation 2052 (shown as 2051 here since the original observation numbers were shifted with the removal of the 0 mm height observations) is clearly a high influence data point. Based on discussions around its validity as a height value (226 mm), it will be removed during the first phase of model refinement.

The saturated model's explanatory variables' coefficients and associated p-values and confidence intervals (CI) are displayed on the right in Figure 6. With the exception of *Length*, all predictors portray statistically significant coefficients. As *Length* contributes little value to the model, it will be removed during refinement. However, multicollinearity has already been assessed and there is anticipated need for additional predictor removal beyond the removal of *Length* alone. Also noteworthy is the adjusted-R^2 value of 0.585. One goal throughout model reduction will be to keep this value as large as possible while simultaneously reducing VIF values to acceptable levels and remaining in check with MLR assumptions.





Predictors	Saturated Model		
	Estimate	*CI*	*p-value*
(Intercept)	2.3112	2.2455 – 2.3768	**<.001**
male	0.1065	0.0807 – 0.1323	**<.001**
female	0.1001	0.0726 – 0.1276	**<.001**
Length	-0.0008	-0.0033 – 0.0017	.530
Diameter	0.0056	0.0026 – 0.0087	**<.001**
Height	0.0053	0.0032 – 0.0074	**<.001**
sqrt.whole	0.1704	0.1427 – 0.1981	**<.001**
sqrt.shucked	-0.2546	-0.2753 – -0.2340	**<.001**
sqrt.viscera	-0.0815	-0.1047 – -0.0583	**<.001**
sqrt.shell	0.1322	0.1078 – 0.1566	**<.001**
Observations	4175		
R^2 / adj. R^2	.586 / .585		

Figure 6. Saturated Model Cook's D (Left) and MLR Predictor Coefficients, Confidence Intervals, and p-values (Right)

A summary for the steps taken during model refinement are provided in Figure 7. These steps are taken to reduce the effects of multicollinearity, to simplify the model while holding prediction integrity, and to better meet all MLR assumptions. The first step of reduction is to remove *Length* and Observation 2052. Doing so has little effect on VIFs or Adjusted-R^2 as depicted above the first grey box. Since *sqrt.whole* exhibits a VIF of 158.5, it is removed during the next phase. All the while, assumptions are re-checked to ensure compliance. Yet, after *sqrt.whole* removal, *Diameter* still has a large VIF of 19.7; therefore, it is removed during the next phase. After removing *Diameter*, assumptions still remain in check but *sqrt.viscera*'s VIF equals 15.6. While adjusted-R^2 has only decreased from 0.5854 to 0.5709, *sqrt.viscera*'s VIF suggests it is likely a redundant variable and may be removed. As such, it is removed during the next stage (shaded green). At this stage, the model has been reduced to include only *Sex* indicator levels, *Height*, *sqrt.shucked*, and *sqrt.shell*, and still has an adjusted-R^2 at 0.5709, but *sqrt.shell* exhibits a borderline-high VIF value of 10.8. When attempting to simplify the model further by removing this variable, adjusted-R^2 plummets to 0.4391 which is even lower than the R^2 value of 0.4677 for *sqrt.shell* on years discussed earlier. For this reason, the model after *sqrt.viscera* removal is selected as the final model.

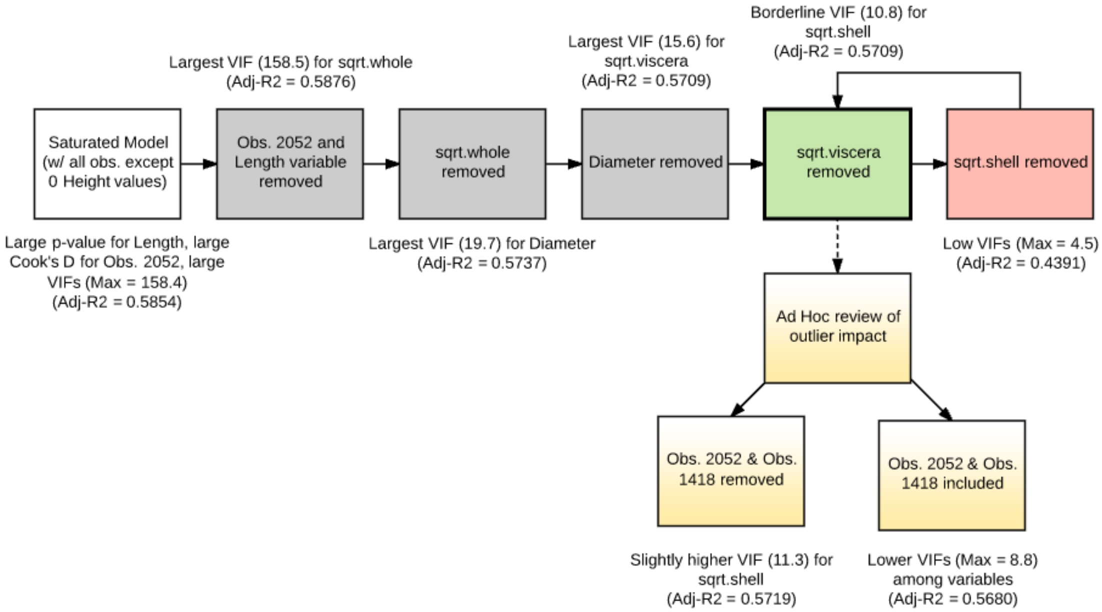
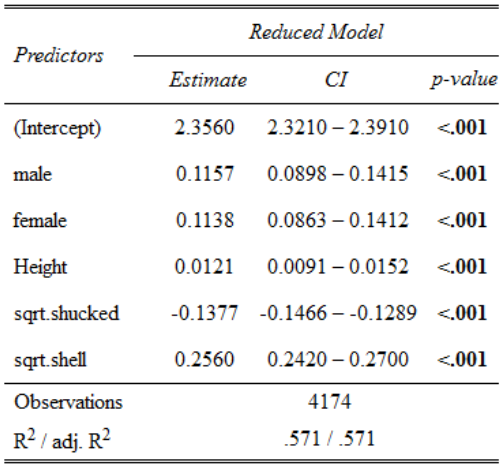

Saturated Model (w/ all obs. except 0 Height values)
Large p-value for Length, large Cook's D for Obs. 2052, large VIFs (Max = 158.4) (Adj-R2 = 0.5854)
Obs. 2052 and Length variable removed
Largest VIF (158.5) for sqrt.whole (Adj-R2 = 0.5876)
sqrt.whole removed
Largest VIF (19.7) for Diameter (Adj-R2 = 0.5737)
Diameter removed
Largest VIF (15.6) for sqrt.viscera (Adj-R2 = 0.5709)
sqrt.viscera removed
Borderline VIF (10.8) for sqrt.shell (Adj-R2 = 0.5709)
sqrt.shell removed
Low VIFs (Max = 4.5) (Adj-R2 = 0.4391)
Ad Hoc review of outlier impact
Obs. 2052 & Obs. 1418 removed
Slightly higher VIF (11.3) for sqrt.shell (Adj-R2 = 0.5719)
Obs. 2052 & Obs. 1418 included
Lower VIFs (Max = 8.8) among variables (Adj-R2 = 0.5680)

Figure 7. MLR Model Refinement Process Summary

The selected model, which includes *Sex* indicator levels with *infant* treated as reference, *Height*, *sqrt.shucked*, and *sqrt.shelled*, meets model assumptions of constant variance, normality, and linearity with weakened effects of multicollinearity per final VIF values and low Cook's Distance values (Max Cook's D < 0.5). Checks for these assumptions are provided in the diagnostic plots and VIF table of Figure 9 and Table 6 in the Appendix.

The final model represents an acceptable, simplified reduction to the fully saturated, over-fitted model discussed initially. The final model's adjusted-R^2 is 0.571 as indicated in Table 4, which is only 0.014 less than the saturated model's of 0.585 as indicated in Figure 6. Additionally, an adjusted-R^2 of 0.571 is an equivalent Pearson's R of approximately 0.756 – which is larger than the single largest Pearson's R of 0.684 described using the correlation values in Table 2. This confirms that even though all numeric explanatory variables correlate strongly with *Years*, no single explanatory variable can predict abalone age as well as our derived model. This derived model is described by the predictor estimates of Table 4, and when put into formula form, is as follows: $\sqrt{Years} = 2.356 + 0.116(male) + 0.114(female) + 0.012(Height) - 0.138(\sqrt{Shucked}) + 0.256(\sqrt{Shell})$. Since the square root transformation was applied to the *Years* response variable, back-transformation is necessary to place the model into context. Therefore, the model becomes: $Years = \left[2.356 + 0.116(male) + 0.114(female) + 0.012(Height) - 0.138(\sqrt{Shucked}) + 0.256(\sqrt{Shell})\right]^2$.

Table 4. Reduced Model MLR Predictor Coefficients, Confidence Intervals, and p-values



Predictors	*Reduced Model*		
	Estimate	*CI*	*p-value*
(Intercept)	2.3560	2.3210 – 2.3910	<**.001**
male	0.1157	0.0898 – 0.1415	<**.001**
female	0.1138	0.0863 – 0.1412	<**.001**
Height	0.0121	0.0091 – 0.0152	<**.001**
sqrt.shucked	-0.1377	-0.1466 – -0.1289	<**.001**
sqrt.shell	0.2560	0.2420 – 0.2700	<**.001**
Observations	4174		
R^2 / adj. R^2	.571 / .571		

To review, Observation 2052 was removed due to suspicions of mismeasurement and high influence on model outcome. Returning to the model refinement summary in Figure 7, notice additional action is taken to review the effects of Observation 2052 (height = 226 mm) and also the second largest height at Observation 1418 (height = 103 mm). Though there is insufficient evidence to suggest Observation 1418 is invalid, its impact on final model parameters is worth considering. When removing Observation 1418 in addition to Observation 2052, there is only a slight improvement in adjusted-R^2 but simultaneously a worsening of *sqrt.shell* VIF. Including both outliers lowers the adjusted-R^2 slightly while improving *sqrt.shell*. Since MLR assumptions are impacted to some degree by their inclusion but only Observation 2052 may be justified for removal, the final model was chosen to remain at the *sqrt.viscera* removal step while excluding only Observation 2052.

<u>Interpretation</u>

Since our model is

$$Years = \left[2.356 + 0.116(male) + 0.114(female) + 0.012(Height) - 0.138\left(\sqrt{Shucked}\right) + 0.256\left(\sqrt{Shell}\right)\right]^2$$

we break it down into the following models for each of the Abalone tested: Infant, Male, and Female.

Infant: $Years = [2.356 + 0.012(Height) - 0.138\left(\sqrt{Shucked\ Weight}\right) + 0.256\left(\sqrt{Shell\ Weight}\right)]^2$

Male: $Years = [2.472 + 0.012(Height) - 0.138\left(\sqrt{Shucked\ Weight}\right) + 0.256\left(\sqrt{Shell\ Weight}\right)]^2$

Female: $Years = [2.47 + 0.012(Height) - 0.138\left(\sqrt{Shucked\ Weight}\right) + 0.256\left(\sqrt{Shell\ Weight}\right)]^2$

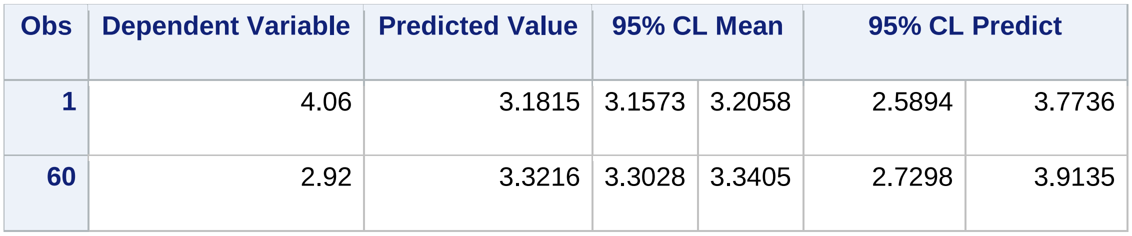
Our adjusted R^2= 0.57, so 57% of the variation in age in years is explained by Sex, Height, Shucked Weight, and Shell Weight. While this is a fairly large adjusted R^2, it means that 43% of the variation is left unexplained.

From our 95% confidence levels which can be seen in Table 4, we are 95% confident that our baseline (which is for the infant abalone) is between 2.32 and 2.39 years, 95% of the time the positive effect that male abalone have on age is between 0.09 and 0.142 years, and we are 95% confident that the positive effect that female abalone have on age is between 0.086 and 0.141 years. 95% of the time the positive effect that height has on age is between 0.009 and 0.015 years. We are 95% confident that the negative effect that shucked weight has on age is between 0.129 and 0.147 years. Finally, 95% of the results will show the positive effect that shell weight has on age is between 0.242 and 0.27 years.

Selecting a couple of the observations as seen in Table 5, we can see the prediction intervals for those particular observations. The data values for the following interpretations have been squared in order to back transform the data into years. For observation 1, we can see that for a randomly selected male abalone with height of 19 mm, shucked weight of 0.225 g and shell weight of 0.15 g, the predicted value for age in years is 10.125, and based on the prediction interval, we are 95% confident that the age in years is between 6.705 and 15.315. Based on the confidence interval, for the population of male abalone with height of 19 mm, shucked weight of 0.225 g, and shell weight of 0.15 g, we are 95% confident that the mean age in years is between 9.967 and 10.278. For observation 60, we can see that for a randomly selected female abalone with height of 25 mm, shucked weight of 0.246 g, and shell weight of 0.175 g, the predicted value of age in years is 11.033, and based on the prediction interval, we are 95% confident that the age in years is between 7.452 and 15.315. Based on the confidence interval, for the population of female abalone with height of 25 mm, shucked weight of 0.246 g, and shell weight of 0.175 g, we are 95% confident that the mean age in years is between 10.91 and 11.162.

The measurements were meticulously controlled which formed a great strength for this study, but a lack of controlling the environment could be considered a glaring weakness which hinders the ability for inference beyond this particular population of abalone. Another weakness in this study is the multicollinearity between the variables used to measure the response variable. Much of this collinearity is unavoidable due to the inherent collinearity between measurements such as height, length, and diameter. Some confounding variables which prevent inference beyond the scope of this particular population of abalone include but are not limited to: climate differences, types and amounts of food available to the abalone, harvesting differences in amount and/or timing.

Table 5. Predicted Values & Intervals



Obs	Dependent Variable	Predicted Value	95% CL Mean		95% CL Predict	
1	4.06	3.1815	3.1573	3.2058	2.5894	3.7736
60	2.92	3.3216	3.3028	3.3405	2.7298	3.9135

Conclusion

This paper used multiple linear regression to provide a model for predicting the age of an individual abalone from easily measured physical characteristics. Seven physical measurements of abalone were evaluated for inclusion in the model as independent variables. Our analysis produced a final model that included gender, height, shucked weight and shell weight with an adjusted R^2 of 0.57.

Appendix:

DS1) Study Reference: http://archive.ics.uci.edu/ml/datasets/Abalone

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (_Haliotis_species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, TechnicalReport No. 48 (ISSN 1034-3288)

There are additional assumptions that need to be made regarding measurement;

1) The scale was calibrated to accurately measure the weight of the abalone.
2) 4177 observations were taken and it is assumed that the research team instituted a protocol that would minimize data entry errors, mismeasurement, and counting of rings.
   a) Three recording errors were identified. Observation 2052 – height outlier that was greater than maximum length value. Abalone length will always be greater than its height. Two other observations (1258 and 3997) had no height measures and were removed due to a recording error.
      a. Abalone Reproduction and Growth. "**Abalone may grow rounded or flat shells depending on their environment" (http://www.marinebio.net/marinescience/06future/abrepro.htm)**
3) We are further assuming that the people shucking the abalone were skilled enough in the capacity to not leave statistically significant amounts of meats within the shell to affect the observation.
4) There is an assumption that the time to measure the abalone did not cause evaporation that would significantly affect the measurement of the weight of the abalone.
5) The exact timeframe for when the experiment was conducted is unknown and this student assumes that weather did not play a factor in any measurements.
6) We are assuming that the research team used a precise enough instrument like a micrometer to measure the dimensions of the abalone and that the recorders had sufficient training in their use.

DS2) Procedure Reference https://www.youtube.com/ watch?v=OfODE3DnBiss

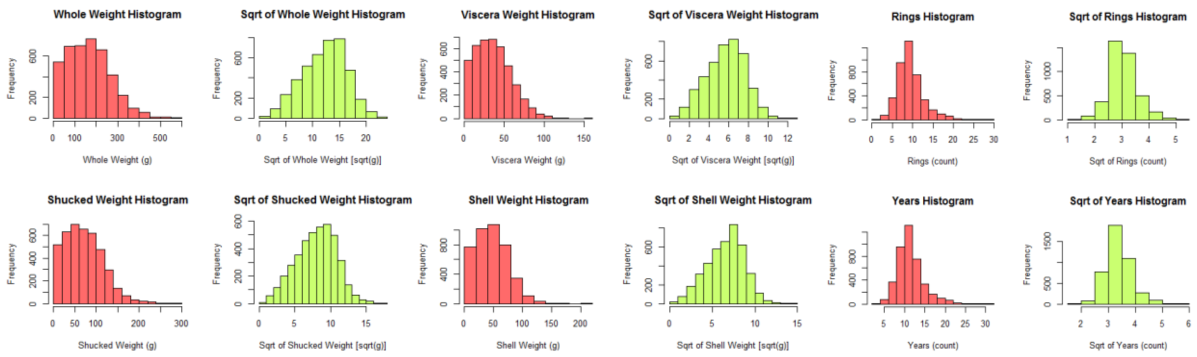


Figure 8. Weight Variable Distributions Pre- and Post-sqrt Transformation

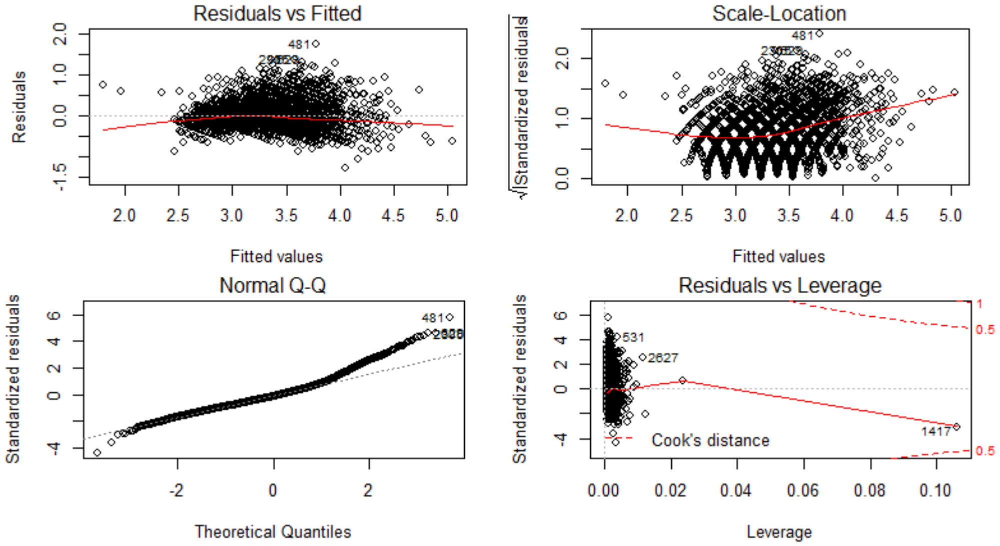


Figure 9. Plot Diagnostics for Reduced MLR

Table 6. Reduced Model VIF Values for Each Explanatory Variable

	male	female	Height	sqrt.shucked	sqrt.shell
VIF	1.85	1.93	6.75	7.15	10.82

Summary Statistics SAS Code

SAS_DS1)

```
* SAS Code revert abalone data from modified machine
  learning data set to actual values recorded in study;
data sasuser.abalone_Multiply_200;
set sasuser.abalone;
length_mm = length*200;
```

```
height_mm = height*200;
diameter_mm = diameter*200;
Whole_weight_grams = Whole_Weight*200;
Shucked_Weight_grams = Shucked_weight*200;
Viacera_Weight_grams = Viscera_weight*200;
Shell_Weight_grams = Shell_weight*200;
run;

* SAS Code to create summary statistics;
proc means data=sasuser.abalone_Multiply_200 mean median min max q1 q3 range
stddev;
var length_mm diameter_mm height_mm Whole_weight_grams Shucked_weight_grams
Viscera_weight_grams Shell_weight_grams Rings;
run;


* SAS Code to Identify Recording Errors;
proc print data = sasuser.abalone_Multiply_200;
var height_mm;
where height_mm = 0;
run;
proc print data = sasuser.abalone_Multiply_200;
var height_mm;
where height_mm > 200;
run;
```

SAS_DS2)

```
* SAS Code to create frequency distribution of variables;
%macro proc_hist(DSN);
proc sgplot data=sasuser.abalone_Multiply_200;
histogram &DSN;
yaxis
labelattrs=(size=33pt color=blue)
valueattrs=(size=33pt color=blue);
xaxis
labelattrs=(size=33pt color=green)
valueattrs=(size=33pt color=green);
density &DSN/ type = normal;

run;
%mend Proc_hist;
%Proc_hist(length_mm);
%Proc_hist(diameter_mm);
%Proc_hist(height_mm);
%Proc_hist(Shucked_Weight_grams);
%Proc_hist(Whole_weight_grams);
%Proc_hist(Viscera_Weight_grams);
%Proc_hist(Shell_Weight_grams);
%Proc_hist(rings);
```

Code for Full Analysis Using R:

```
#############################################################
## Chris Boomhower, Andrew Abbott, Brian Kruse, Johnny Quick
## MSDS 6372-401
## Project 1: Multiple Linear Regression Analysis
## 10/03/2016
##
## Makelike.R
##############################

setwd("Analysis/Data/")
source('abalone_clean.R', echo = TRUE)
setwd('../')
source('abalone_EDA.R', echo=TRUE)
source('abalone_ModelRefinement.R', echo=TRUE)
setwd('../')


#####################
## abalone_clean.R ##
#####################

require(dplyr)
require(formattable)

##############################################
## Import raw data from disk
##############################################
aby <- read.table("abalone.data", sep = ",")

##############################################
## Clean-up / Preliminary EDA to see what
## transformations may be required
##############################################
aby.clean <- aby
names(aby.clean) <- c("Sex.", "Length", "Diameter", "Height", "Whole",
"Shucked", "Viscera", "Shell", "Rings")
aby.clean[,2:8] <- aby.clean[,2:8] * 200 # Muliply continuous variables by
200 to rescale to actual values (See UCI study data description)

# Create dummy variables to match SAS methodology
aby.clean$male <- ifelse(aby.clean$Sex. == "M", 1, 0)
aby.clean$female <- ifelse(aby.clean$Sex. == "F", 1, 0)
aby.clean$Years <- aby.clean$Rings + 1.5

str(aby.clean)
head(aby.clean)

# Review distributions
par(mfrow = c(1, 1))
barplot(table(aby.clean$Sex.)) # Abalone sex barplots
table(aby.clean$Sex.) # Combine values with barplot using ggplot

par(mfrow = c(2, 2))
hist(aby.clean$Length) # Height Histograms
hist(aby.clean$Diameter)
hist(aby.clean$Height)
boxplot(aby.clean$Height, main = "Boxplot of Height")
```

```
#hist(log(aby.clean$Height+1))

# View potential outliers in Height data
outliers = boxplot(aby.clean$Height, plot=FALSE)$out
aby.clean[aby.clean$Height %in% outliers,]

# Remove 0 values for Height
sum(aby.clean$Height == 0)
aby.clean <- aby.clean %>% filter(Height > 0)
sum(aby.clean$Height == 0)

# Observe what the Height distribution would look like without the two large
outliers
par(mfrow = c(1, 1))
hist(aby.clean[aby.clean$Height < 100,]$Height)

# Continue reviewing distributions
par(mfrow = c(2, 2))
hist(aby.clean$Whole, main = "Whole Weight Histogram", col = "indianred1",
xlab = "Whole Weight (g)")
hist(sqrt(aby.clean$Whole), main = "Sqrt of Whole Weight Histogram", col =
"darkolivegreen1", xlab = "Sqrt of Whole Weight [sqrt(g)]")
hist(aby.clean$Shucked, main = "Shucked Weight Histogram", col =
"indianred1", xlab = "Shucked Weight (g)")
hist(sqrt(aby.clean$Shucked), main = "Sqrt of Shucked Weight Histogram", col
= "darkolivegreen1", xlab = "Sqrt of Shucked Weight [sqrt(g)]")
hist(aby.clean$Viscera, main = "Viscera Weight Histogram", col =
"indianred1", xlab = "Viscera Weight (g)")
hist(sqrt(aby.clean$Viscera), main = "Sqrt of Viscera Weight Histogram", col
= "darkolivegreen1", xlab = "Sqrt of Viscera Weight [sqrt(g)]")
hist(aby.clean$Shell, main = "Shell Weight Histogram", col = "indianred1",
xlab = "Shell Weight (g)")
hist(sqrt(aby.clean$Shell), main = "Sqrt of Shell Weight Histogram", col =
"darkolivegreen1", xlab = "Sqrt of Shell Weight [sqrt(g)]")
hist(aby.clean$Rings, main = "Rings Histogram", col = "indianred1", xlab =
"Rings (count)")
hist(sqrt(aby.clean$Rings), main = "Sqrt of Rings Histogram", col =
"darkolivegreen1", xlab = "Sqrt of Rings (count)")
hist(aby.clean$Years, main = "Years Histogram", col = "indianred1", xlab =
"Years (count)")
hist(sqrt(aby.clean$Years), main = "Sqrt of Years Histogram", col =
"darkolivegreen1", xlab = "Sqrt of Years (count)")

# Provide descriptive summary statistics
min <- round(sapply(aby.clean[,2:8], min, na.rm = TRUE),2)
max <- round(sapply(aby.clean[,2:8], max, na.rm = TRUE),2)
mean <- round(sapply(aby.clean[,2:8], mean, na.rm = TRUE),2)
median <- round(sapply(aby.clean[,2:8], median, na.rm = TRUE),2)
std.dev <- round(sapply(aby.clean[,2:8], sd, na.rm = TRUE),2)
variance <- round(sapply(aby.clean[,2:8], var, na.rm = TRUE),2)

var.summary <- rbind(min, max, mean, median, std.dev, variance) # Combine
rows for summary table
```

```
formattable(var.summary)


##############################################
## Transform the data
##############################################
aby.clean$sqrt.whole <- sqrt(aby.clean$Whole)
aby.clean$sqrt.shucked <- sqrt(aby.clean$Shucked)
aby.clean$sqrt.viscera <- sqrt(aby.clean$Viscera)
aby.clean$sqrt.shell <- sqrt(aby.clean$Shell)
aby.clean$sqrt.rings <- sqrt(aby.clean$Rings)
aby.clean$sqrt.years <- sqrt(aby.clean$Years)

###################
## abalone_EDA.R ##
###################

require(MASS)
require(dplyr)
require(car)
require(sjPlot)

##############################################
## Perform EDA
##############################################

## Generate initial model for EDA
fit <- lm(sqrt.years ~ male + female + Length + Diameter + Height +
sqrt.whole + sqrt.shucked + sqrt.viscera + sqrt.shell, data = aby.clean)

par(mfrow = c(2,2))
attach(aby.clean)
plot(Length,sqrt.years, col = "dodgerblue")
plot(Diameter,sqrt.years, col = "darkorange")
plot(Height,sqrt.years, col = "darkgreen")
plot(subset(Height,Height < 100),subset(sqrt.years,Height < 100), col =
"darkgreen")
plot(sqrt.whole,sqrt.years, col = "lightcoral")
plot(sqrt.shucked,sqrt.years, col = "tan4")
plot(sqrt.viscera,sqrt.years, col = "purple")
plot(sqrt.shell,sqrt.years, col = "turquoise4")

## Check for covariance
formattable(cor(cbind(sqrt.years, Sex., Length, Diameter, Height, sqrt.whole,
sqrt.shucked, sqrt.viscera, sqrt.shell)))
detach(aby.clean)
pairs(~sqrt.years + Sex. + Length + Diameter + Height + sqrt.whole +
sqrt.shucked + sqrt.viscera + sqrt.shell, data = aby.clean)

# QQ Plot
par(mfrow = c(1,1))
qqPlot(fit, main = "Q-Q Plot: Saturated Model")
```

```
# Studentized residuals distribution
sresid <- studres(fit)
hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals:
Saturated Model", col = "khaki1")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)

# Outlier assessment
outlierTest(fit) # Bonferonni p-value for most extreme obs

# Cook's D plot: identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
par(mfrow = c(1,1))
plot(fit, which=4, cook.levels=cutoff)

# Evaluate Collinearity
vif(fit) # variance inflation factors

# Evaluate Nonlinearity: component + residual plot (Partial Residuals)
crPlots(fit, main="Partial-Residuals")

# Model summary
summary(fit)
sjt.lm(fit, string.est = "Estimate", string.dv = "Saturated Model",
       string.p = "p-value", digits.est = 4, digits.ci = 4, show.header =
TRUE)

## Generate diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)

## Perform stepwise selection for assessment
step <- stepAIC(fit, direction="forward")
step$anova # display results

###############################
## abalone_ModelRefinement.R ##
###############################

ModelEDA <- function(modelfit, df, pairsIn){
    ## Check for covariance
    pairs(pairsIn, data = df)

    # QQ Plot
    par(mfrow = c(1,1))
    qqPlot(modelfit, main = "QQ Plot for Model Reduction")

    # Studentized residuals distribution
    sresid <- studres(modelfit)
    hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals")
    xfit<-seq(min(sresid),max(sresid),length=40)
    yfit<-dnorm(xfit)
    lines(xfit, yfit)

    # Outlier assessment
    writeLines("\n-----Bonferroni p-values for most extreme outliers-----\n")
    print(outlierTest(modelfit)) # Bonferonni p-value for most extreme obs

    # Cook's D plot: identify D values > 4/(n-k-1)
    cutoff <- 4/((nrow(mtcars)-length(modelfit$coefficients)-2))
    par(mfrow = c(1,1))
    plot(modelfit, which=4, cook.levels=cutoff)

    # Evaluate Collinearity
    writeLines("\n-----VIF Values-----\n")
    print(vif(modelfit)) # variance inflation factors

    # Evaluate Nonlinearity: component + residual plot (Partial Residuals)
    crPlots(modelfit, main="Partial-Residuals")

    # Model summary
    writeLines("\n-----Model Summary-----\n")
    print(summary(modelfit))

    ## Generate diagnostic plots
    layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
    plot(modelfit)
}

###############################################################################
## Model Reduction Attempt #1 (observation 2052 [obs.2051 after removing
## 0 Height values] and Length variable removed)
###############################################################################
aby.clean2 <- subset(aby.clean, rownames(aby.clean) != 2051) # Remove Outlier
2051 since mis-measurement presumed

fit2 <- lm(sqrt.years ~ male + female + Diameter + Height + sqrt.whole +
sqrt.shucked + sqrt.viscera + sqrt.shell, data = aby.clean2)

## Check for covariance
attach(aby.clean2)
writeLines("\n-----Covariance Matrix-----\n")
cor(cbind(sqrt.years, Sex., Diameter, Height, sqrt.whole, sqrt.shucked,
sqrt.viscera, sqrt.shell))
detach(aby.clean2)
pairsIn <- eval(~sqrt.years + Sex. + Diameter + Height + sqrt.whole +
sqrt.shucked + sqrt.viscera + sqrt.shell)

ModelEDA(fit2,aby.clean2,pairsIn)

###############################################################################
## Model Reduction Attempt #2 (sqrt.whole removal)
###############################################################################
fit3 <- lm(sqrt.years ~ male + female + Diameter + Height  + sqrt.shucked +
sqrt.viscera + sqrt.shell, data = aby.clean2)
```

```
## Check for covariance
attach(aby.clean2)
cor(cbind(sqrt.years, Sex., Diameter, Height, sqrt.shucked, sqrt.viscera,
sqrt.shell))
detach(aby.clean2)
pairsIn <- eval(~sqrt.years + Sex. + Diameter + Height + sqrt.shucked +
sqrt.viscera + sqrt.shell)

ModelEDA(fit3,aby.clean2,pairsIn)

#############################################################################
## Model Reduction Attempt #3 (Diameter removal)
#############################################################################
fit4 <- lm(sqrt.years ~ male + female + Height + sqrt.shucked + sqrt.viscera
+ sqrt.shell, data = aby.clean2)

## Check for covariance
attach(aby.clean2)
cor(cbind(sqrt.years, Sex., Height, sqrt.shucked, sqrt.viscera, sqrt.shell))
detach(aby.clean2)
pairsIn <- eval(~sqrt.years + Sex. + Height + sqrt.shucked + sqrt.viscera +
sqrt.shell)

ModelEDA(fit4,aby.clean2,pairsIn)

#############################################################################
## Model Reduction Attempt #4 (sqrt.viscera removal) *** Selected model ***
#############################################################################
fit5 <- lm(sqrt.years ~ male + female  + Height  + sqrt.shucked + sqrt.shell,
data = aby.clean2)

## Check for covariance
attach(aby.clean2)
cor(cbind(sqrt.years, Sex., Height, sqrt.shucked, sqrt.shell))
detach(aby.clean2)
pairsIn <- eval(~sqrt.years + Sex. + Height + sqrt.shucked + sqrt.shell)

ModelEDA(fit5,aby.clean2,pairsIn)
sjt.lm(fit5, string.est = "Estimate", string.dv = "Reduced Model",
       string.p = "p-value", digits.est = 4, digits.ci = 4, show.header =
TRUE)

## SOME QUICK AD HOC REVIEW OF OUTLIER INFLUENCE ON FINAL MODEL
aby.clean3 <- subset(aby.clean2, rownames(aby.clean2) != 1417)
fitT <- lm(sqrt.years ~ male + female  + Height  + sqrt.shucked + sqrt.shell,
data = aby.clean)
vif(fitT) # variance inflation factors
summary(fitT)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fitT)

#############################################################################
```

```
## Model Reduction Attempt #5 (sqrt.shell removal)
#############################################################################
fit6 <- lm(sqrt.years ~ male + female  + Height  + sqrt.shucked, data =
aby.clean2)

## Check for covariance
attach(aby.clean2)
cor(cbind(sqrt.years, Sex., Height, sqrt.shucked))
detach(aby.clean2)
pairsIn <- eval(~sqrt.years + Sex. + Height + sqrt.shucked)

ModelEDA(fit6,aby.clean2,pairsIn)
```