Jia-Wei (Jessie) Liang
Professor Gordon Linoff
Analytics Frameworks and Methods
September 2016

Analysis on the Economic Value of Abalone

❖ My Business Problem:

The economic value of abalone is positively correlated with its age. Therefore, to be able to distinguish the age of abalone is important for both abalone farmers and customers in order to determine its price. However, the current technology to estimate the age is quite costly and inefficient. Farmers usually cut the shells and count the rings through microscopes. My goal is to take the physical measurements of abalone, which are easier to obtain, and find the best indicators that could forecast the rings (age). If a statistical procedure proves to be reliable and accurate enough, working hours could be saved.

❖ About the Dataset:

The dataset, *Abalone*, contains detail information about the abalone's physical characteristics (sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight) and their rings (age). It contains 9 columns, along with 4179 observations. The dataset is provided by the University of California Irvine Machine Learning Reposition.

❖ Five Steps Toward Data Cleansing

Step 0: *Glance through the dataset.*
There is no missing data.

Step 1: *Take a look at the correlation matrix.*

Correlations

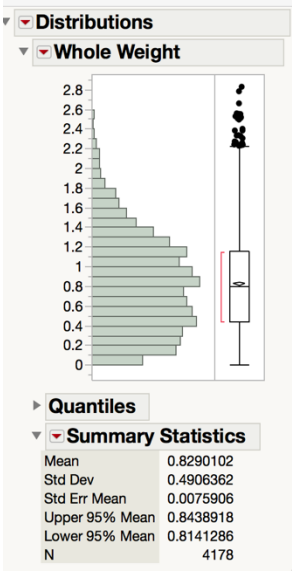
	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
Length	1.0000	0.9868	0.8276	0.9253	0.8979	0.9031	0.8978	0.5568
Diameter	0.9868	1.0000	0.8338	0.9254	0.8931	0.8998	0.9054	0.5747
Height	0.8276	0.8338	1.0000	0.8193	0.7750	0.7984	0.8174	0.5575
Whole Weight	0.9253	0.9254	0.8193	1.0000	0.9694	0.9664	0.9554	0.5404
Shucked Weight	0.8979	0.8931	0.7750	0.9694	1.0000	0.9320	0.8827	0.4209
Viscera Weight	0.9031	0.8998	0.7984	0.9664	0.9320	1.0000	0.9077	0.5039
Shell Weight	0.8978	0.9054	0.8174	0.9554	0.8827	0.9077	1.0000	0.6276
Rings	0.5568	0.5747	0.5575	0.5404	0.4209	0.5039	0.6276	1.0000

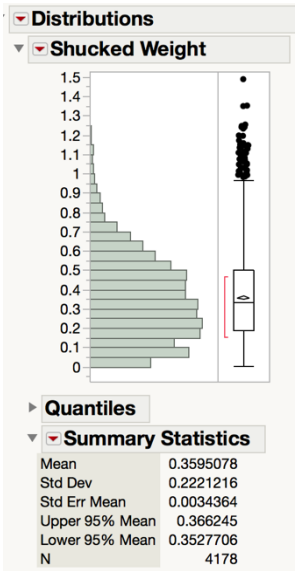
According to the correlation matrix, we could see that there are values higher than 0.5, which indicate that there are strong correlations between them, and this might cause the issue of collinearity. I would pay more attention to these variables in the later progress of data cleaning and processing.

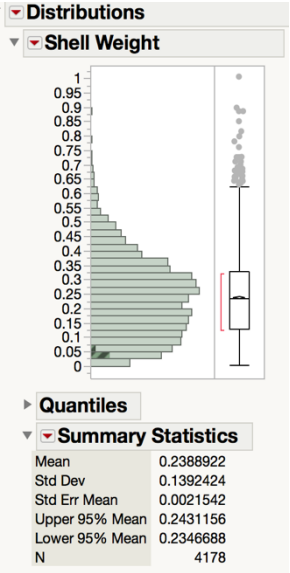
Step 2: *Look at the distribution plot of each independent variables and check if they are reasonable or not.*

By observing the distribution plot of every independent variable, I found some outliers that should be removed in order to make the (distribution) plot more normally distributed. There are three of

the independent variables that have significant results after removing the outliers, which are: "whole weight", "shucked weight" and "shell weight".







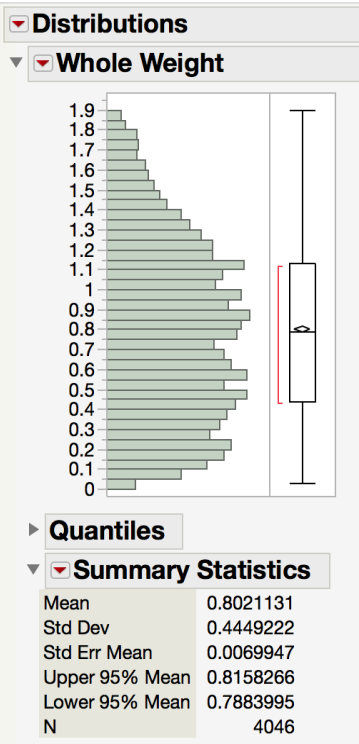
The distribution plots of whole weight, shucked weight and shell weight show that there are obvious outliers. I have two methods to fix this problem (remove the outliers).

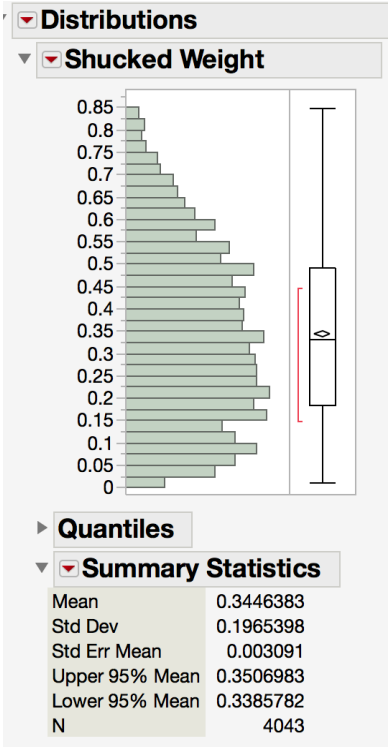
Methods: *Take the variable "whole weight" as an example:*

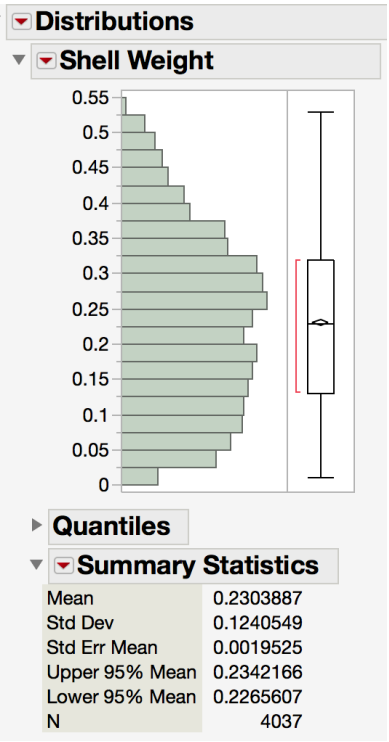
1.The first method is to change all the observations that differ a lot from the mean to the number that is closest to its value. Since observations more than 1.9 as well as observations less than 0.03 are far away from the mean, (considered to be outliers) I would make changes of the value of whole weight, which are more than 1.9 to 1.9 and observations less than 0.03 to 0.03. This method is useful when the problematic points are in large proportion of the original dataset. By conducting this method, we would not lose too much information(data) from the dataset.

2. The second method is to delete all the problematic points since they could be viewed as unreasonable or even wrong points. By deleting them, we could prevent them from ruin our prediction model. However, we must be careful while conducting this method because once we delete too much data, we would not have abundant information to build our model. Take the whole weight example again, if we count the number of observations more than 1.9 and observation less than 0.03, there are only 4178-4046=132 observations, which is only about 3% of the dataset (small proportion). Therefore, in this example, we could use method 2.

By using the the above methods in the three independent variables, I could now obtain new distribution plots, which are shown below:







From the "summary statistics", we could see that the means are almost the same as the former unfixed plots, which indicates that we do not lose too much information (data), especially the necessary ones. In addition, we could see the standard deviation is much smaller than before, which makes the data more reasonable.

Whole weight	Delete observations that are more than 1.9 and less than 0.03. *Number of observations*: 4178-4046=132 (3.16% of the whole data)	Change of standard deviation: 0.491 lower to 0.445
Shucked weight	Delete observations that are more than 0.85 and less than 0.01. *Number of observations:* 4178-4043=135 (3.23% of the whole data)	Change of standard deviation: 0.222 lower to 0.197
Shell weight	Delete observations that are more than 0.53 and less than 0.01. *Number of observations:* 4178-4037=141 (3.37% of the whole data)	Change of standard deviation: 0.139 lower to 0.124

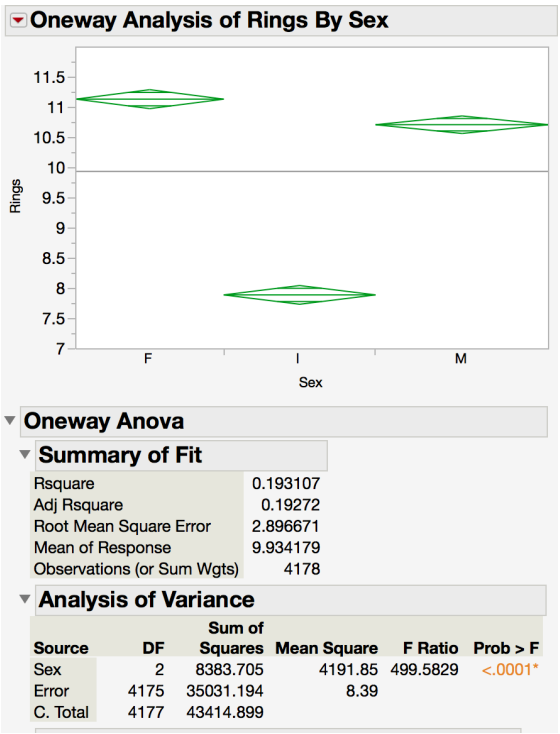
Step 3: *Draw scatter plots for each independent variable along with the dependent variable.*

This is a very important step because it would give us a general understanding of the correlation between each dependent variable and independent variable. If there is no obvious correlation between them (either positive or negative), we should think twice before putting the independent variables into the model.

After I drew the scatter plots of each independent variable with dependent variable, I found out that most of them have a clear positive correlation with the ring (age).

Below are some of the interesting plots that I would like to discuss:

1.



Oneway Anova

Summary of Fit

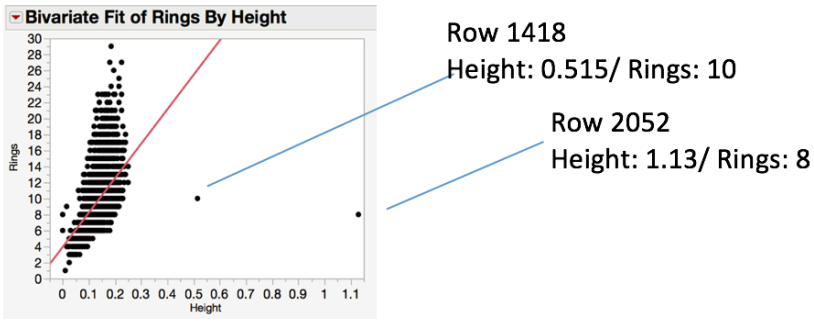
Rsquare	0.193107
Adj Rsquare	0.19272
Root Mean Square Error	2.896671
Mean of Response	9.934179
Observations (or Sum Wgts)	4178

Analysis of Variance

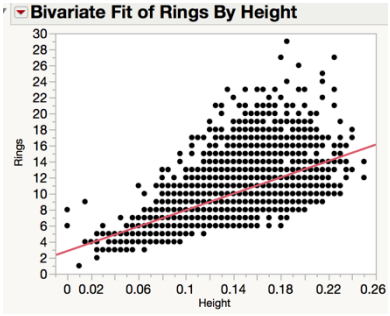
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Sex	2	8383.705	4191.85	499.5829	<.0001*
Error	4175	35031.194	8.39		
C. Total	4177	43414.899			

From the ANOVA table above, we could see that P-value is less than 0.05, which indicates that sex might have effects on the rings (age). However, we could not find a linear correlation between them. Therefore, I would rather remove this variable away or do some transformation to the variable in order to make it more valid. Details will be discussed in step 4.

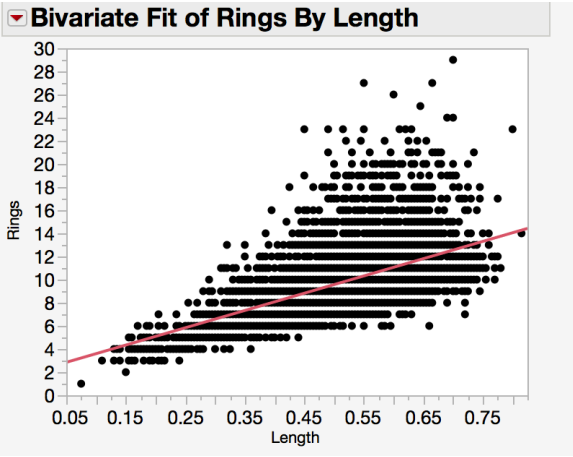
2.

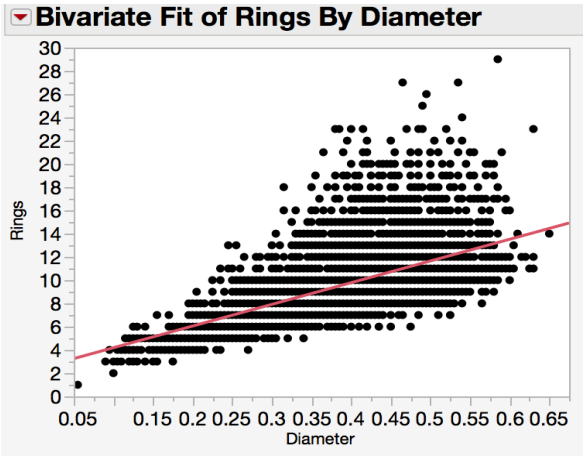


We could see that there are two outliers that would necessary affect the fit line. After removing them, the scatter plot becomes as below, which the line is flatter.

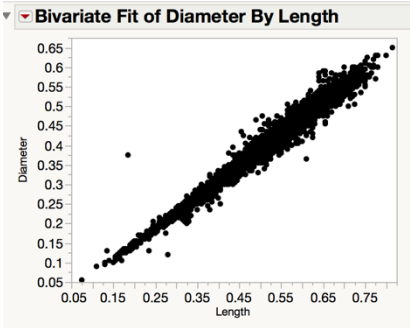


3.





Since abalones are mostly round, the number of length[1] and diameter[2] are mostly the same. We could see from the above plots that the scatter plots of length and diameter looks really similar. By conducting another scatter plot between this two variables, we could see a strong correlation.

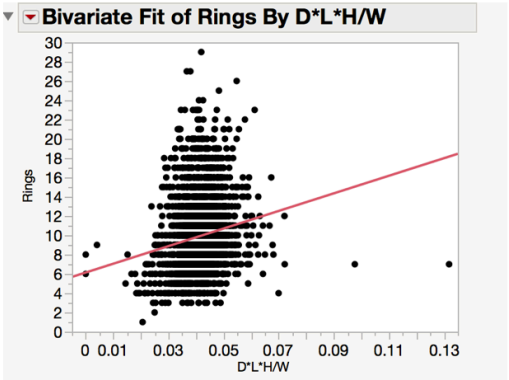


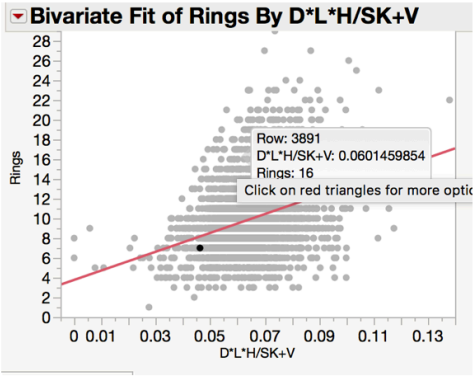
Therefore, in step 4, I will try to combine these two variables together to see if there could make any differences. In addition, the same technique applies to "whole weight and shucked weight", "viscera weight and shell weight".

Step 4: *Create some new variables or make some transformations to solve the problem in step 3.*

To create a new variable, we need to make sure that the new variable makes senses, and I would like to create a formula that would solve problems in step 3.

$$\text{Formula=}\begin{cases}\text{If the data belongs to Infants}=\dfrac{Diameter\times Length\times Height}{Whole\ Weight}\\ \text{Else,}\quad =\dfrac{Diameter\times Length\times Height}{(shucked\ Weight+Viscera\ Weight)}\end{cases}$$





[1] Length: Longest shell measurement (mm)
[2] Diameter: Perpendicular to length (mm)

Both of the plots show that the new variable has a positive linear correlation with rings (age). The new variable contains information about sex, diameter, length and weights, which fixed all the problem in step 3.

Correlations

	D*L*H/W	Rings
D*L*H/W	1.0000	0.1728
Rings	0.1728	1.0000

Correlations

	D*L*H/SK+V	Rings
D*L*H/SK+V	1.0000	0.3473
Rings	0.3473	1.0000

Step 5: *Take a deep look at the correlation between the independent variables.*

I found that when I fit rings (age) with "shell weight", "(D*L*H)/W", "(D*L*H)/(SK+V)" into the linear regression model, the coefficient of (D*L*H)/W will become negative. In addition, the correlation matrix shows that there is correlation between these variables.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	21278.778	7092.93	1337.032
Error	4172	22132.375	5.30	Prob > F
C. Total	4175	43411.153		<.0001*

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.619831	0.247102	18.70	<.0001*
D*L*H/W	-167.6526	9.982464	-16.79	<.0001*
D*L*H/SK+V	139.70185	5.213507	26.80	<.0001*
Shell Weight	13.584424	0.259561	52.34	<.0001*

Correlations

	D*L*H*SH/SK+V	D*L*H/W
D*L*H*SH/SK+V	1.0000	0.3215
D*L*H/W	0.3215	1.0000

In order to decrease the effect of correlation, I would like to create a new variable by combining the shell weight to one of the exist variables. Below is the new correlation matrix:

The issue of collinearity has been solved.

Correlations

	D*L*H/SK+V	D*L*H/W	Shell Weight
D*L*H/SK+V	1.0000	0.8115	0.1633
D*L*H/W	0.8115	1.0000	0.1316
Shell Weight	0.1633	0.1316	1.0000

❖ Outline of My Previous Work on Data Cleansing:

I created some new variables and made some transformations in order to solve the problems encountered while doing data cleansing. The new variables are "D/L" and "new". D/L (Diameter/Length) is created because most abalones are rounded, the numerical value of length[3] and diameter[4] are almost alike.

While "new" is a column that is formed by the below formula:

$$\text{Formula=}\begin{cases}\text{If the data belongs to Infants}=\dfrac{Diameter\times Length\times Height}{Whole\ Weight}\\ \text{Else,}\quad =\dfrac{Diameter\times Length\times Height}{(shucked\ Weight+Viscera\ Weight)}\end{cases}$$

I created this formula because from my previous work, I could see that "sex" (female/male/infants) has effects on the rings (age). However, we could not find a linear correlation between them. Despite from removing this variable away, I did some transformation on the variable in order to make it more valid.

**Interpretation: "New" is the volume per each weight. From my perspective, infants are not yet mature so I use the add-up total weight (Whole Weight). While adult (grown) abalones have more obvious body structures, we could discuss each in detail. (Shucked Weight + Viscera Weight)
In addition, the number of whole weight is almost the total of adding shucked weight, viscera weight, and shell weight together. Therefore, in this final project, I would choose to take "whole weight", "Height", "D/L", and "new" as my variables into my further analysis.

❖ My Analysis:

1. Ordinary Least Square

Summary of Fit

RSquare	0.426743
RSquare Adj	0.426194
Root Mean Square Error	2.442616
Mean of Response	9.934626
Observations (or Sum Wgts)	4176

AICc	BIC
19316.91	19354.91

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	18525.418	4631.35	776.2431
Error	4171	24885.735	5.97	Prob > F
C. Total	4175	43411.153		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-1.569904	0.791165	-1.98	0.0473*	.
Whole Weight	1.2766417	0.180719	7.06	<.0001*	5.4932716
Height	19.87467	2.649588	7.50	<.0001*	7.279381
D/L	5.226551	1.067053	4.90	<.0001*	1.1729954
new	63.01238	3.361129	18.75	<.0001*	1.8267966

[3] Length: Longest shell measurement (mm)
[4] Diameter: Perpendicular to length (mm)

From the regression result, we could see that all of the independent variables have a P-value less than 0.05, which means all the predictors are significant. The Analysis of Variance table indicates that the model is significant because the P-value of F is less than 0.05. In addition, we could see from the Parameter Estimates table that all of the VIF value[5] of the variables are less than 10, which means the dataset do not have the effect of multicollinearity[6].

Then let's take a look at the coefficient of each variables. Whole Weight, Height, D/L, and new are positive, which are reasonable because in the reality, these factors do have a positive relationship with rings (age). Adult (grown) abalones are considered to be bigger, larger and heavier.

2. Lasso Regression

Measure	Training
Number of rows	4178
Sum of Frequencies	4176
-LogLikelihood	9652.4439
BIC	19354.91
AICc	19316.908
Generalized RSquare	0.4267433

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-1.569904	1.4013789	1.2549756	0.2626	-4.316556	1.1767477
Whole Weight	1.2766417	0.2179219	34.319135	<.0001*	0.8495226	1.7037607
Height	19.87467	3.1346388	40.199888	<.0001*	13.730891	26.018449
D/L	5.226551	1.9460399	7.2131825	0.0072*	1.412383	9.040719
new	63.01238	4.2679337	217.9799	<.0001*	54.647384	71.377376
Scale	2.4411529	0.0443636	3027.8594	<.0001*	2.3542018	2.5281041

Lasso regression could be used as a variable selection model. Nevertheless, since I have cleaned the data and finished the data selection process, the Lasso Regression would not be a relevant method in this case because we could see from the table that Lasso Regression has high AIC[7] and high BIC[8], and the R square has not even change.

[5] VIF (Variance Inflation Factor) provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.
[6] Multicollineraity exists when two or more of the predictors in a regression model are moderately or highly correlated. When it exists, it would wreak havoc on the analysis and thereby limit the research conclusion.
[7] AIC (Akaike Information Criterion) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model relative to each of the other models.
[8] BIC (Bayesian Information Criterion) is a criterion for model selection among a finite set of models. The model with the lowest AIC and BIC is preferred. *Wikipedia*

3. *Ridge Regression*

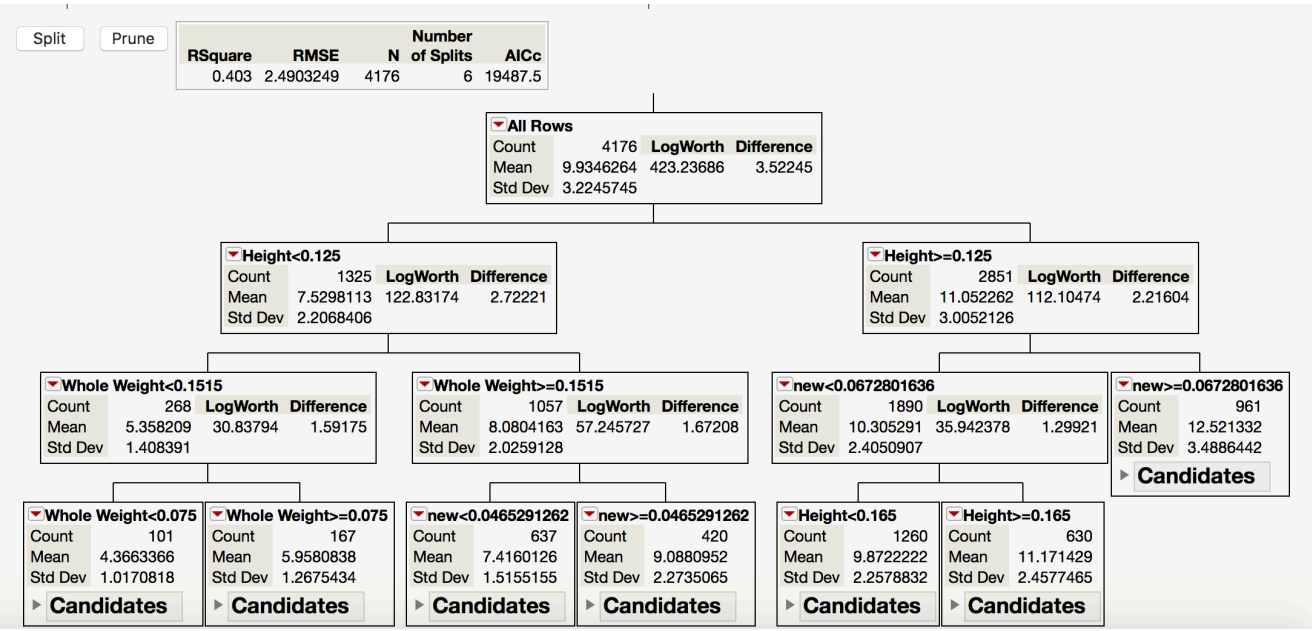
Measure	Training	Validation
Number of rows	3758	418
Sum of Frequencies	3758	418
-LogLikelihood	8725.7624	927.71925
BIC	17500.915	1891.6514
AICc	17463.547	1867.6429
Generalized RSquare	0.4264234	0.4280749

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-1.817064	1.5980575	1.2928724	0.2555	-4.949199	1.3150711
Whole Weight	1.2223553	0.230479	28.127561	<.0001*	0.7706247	1.6740859
Height	21.156399	3.327662	40.420815	<.0001*	14.634301	27.678496
D/L	5.4975541	2.2133177	6.1695217	0.0130*	1.159531	9.8355771
new	61.894118	4.5111247	188.24744	<.0001*	53.052476	70.73576
Scale	2.4669388	0.0471693	2735.2521	<.0001*	2.3744886	2.5593889

From the results, we could conclude that all the variables are are significant, and all of the coefficients are reasonable. Comparing this method with the Ordinary Least Square method as well as Lasso Regression, although the R square has not change very much, the value of AIC and BIC are decreasing.

4. *Decision Tree*



Comparing to other method, Decision Tree has the lowest R square and the highest AIC, which is not as good as the former models. However, it has its own advantages since it provides a clear and understandable data visualization format. We could easily find the main factor using Decision Tree Method.

So far, I have built four different models, and each model has its own feature. Ordinary Least Square (OLS) is the simplest method in linear regression, we could easily run this model in any

software. Ridge and Lasso Regression are methods based on OLS, which could be used as variable selection methods. In addition, Ridge and Lasso Regression could sometimes have a better result than OLS method. As for Decision Tree, it is another kind of model. It provides clear split and it let people to better understand the relationship between dependent variable and independent variables.

Based on the above results, Ridge Regression has the lowest AIC and BIC. Therefore, I would choose this as my prediction model.

❖ **Conclusion:**

The formula of the regression model will be:

Rings (Age)= -1.817 + 1.222*Whole Weight + 21.156*Height + 5.498*D/L + 61.894*new

As we could see in this model, the largest coefficient belongs to "new". "New" is the abalone's volume (size)[9] per weight. Surprisingly, I found out that "Whole Weight" is not that important. In addition, "Height" is also a crucial factor.

<u>Suggestions for people raising and selling abalones:</u>

- Previously, as of now, in order to know the abalone's age, farmers have to take a sample of shell, straining it, and counting the number of rings under the microscope. It is a very boring and time consuming task. With my new proposal, there is no need to measure the numbers of layers of shell (rings) on the abalone's shell. We could tell the abalone's age from their physical measurement.

- The numerical value of their size divided by their weight is highly correlated to their age. *(Use "whole weight" as if it is an infant abalone, and use "shucked weight + viscera weight" for adult abalones, no matter they are male or female)*

- When you are in a hurry and do not have the time to do all the measurements, the abalone's height is a good way to predict their age too! The higher the abalones are, the older they would be.

<u>Suggestion for people buying abalones:</u>

- Abalones are an excellent source of iron and pantothenic acid. Because of them containing highly nutrients, abalones are very expensive. People often have the misunderstanding that the heavier the products, the better they are. However, in the abalone case it is not that true. The "Whole Weight" did not play an important role as well as other factors.

- Next time, when choosing abalones, do not rely entirely on the weight shown on the scale too much. The abalone's height is a better predictor. Higher abalones mean that they are more mature, which contains more nutrition, and it is worth buying them with higher prices.

[9] Volume (size)=Length*Diameter*Height

Appendix: *Logistic Regression*

I would like to perform another regression model that would tell the *possibility* that whether the abalone is old or young. Since logistic regression is used when the dependent variable in question is categorical, I divide the "Rings" dataset into two categories: 0 and 1 and I put them into a new column: Age

0 is for rings larger than 10, which means they are adult (grown) abalones; while 1 is for rings less than 10, which means they are not mature enough.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-8.8136295	0.8546136	106.36	<.0001*
Whole Weight	1.44544381	0.1869367	59.79	<.0001*
Height	5.48939901	2.8778077	3.64	0.0565
D/L	3.83960471	1.1044201	12.09	0.0005*
new	51.1853008	3.6100812	201.03	<.0001*

For log odds of 0/1

RSquare (U)	0.2385
BIC	4146.58

However, from the table above, an unexpected result showed up. The "Height" variable has the P-value higher than 0.05.

After I remove the variable "Height" and run the regression again:

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-8.6202269	0.848001	103.33	<.0001*
Whole Weight	1.75972736	0.0907201	376.26	<.0001*
D/L	3.969913	1.1030911	12.95	0.0003*
new	54.9032377	3.0551307	322.95	<.0001*

For log odds of 0/1

RSquare (U)	0.2378
BIC	4141.89

Now all the variables are significant, and the coefficients are still reasonable. While the R square is almost the same as the previous one, the BIC value is smaller, which seemed to be a better model.

Conclusion:
The formula of the logistic regression model will be:

Probability of being an adult (grown) abalone$=\frac{1}{1+e^{-z}}$

while, Z= -8.62 + 1.76 Whole Weight + 3.97 D/L + 55 new