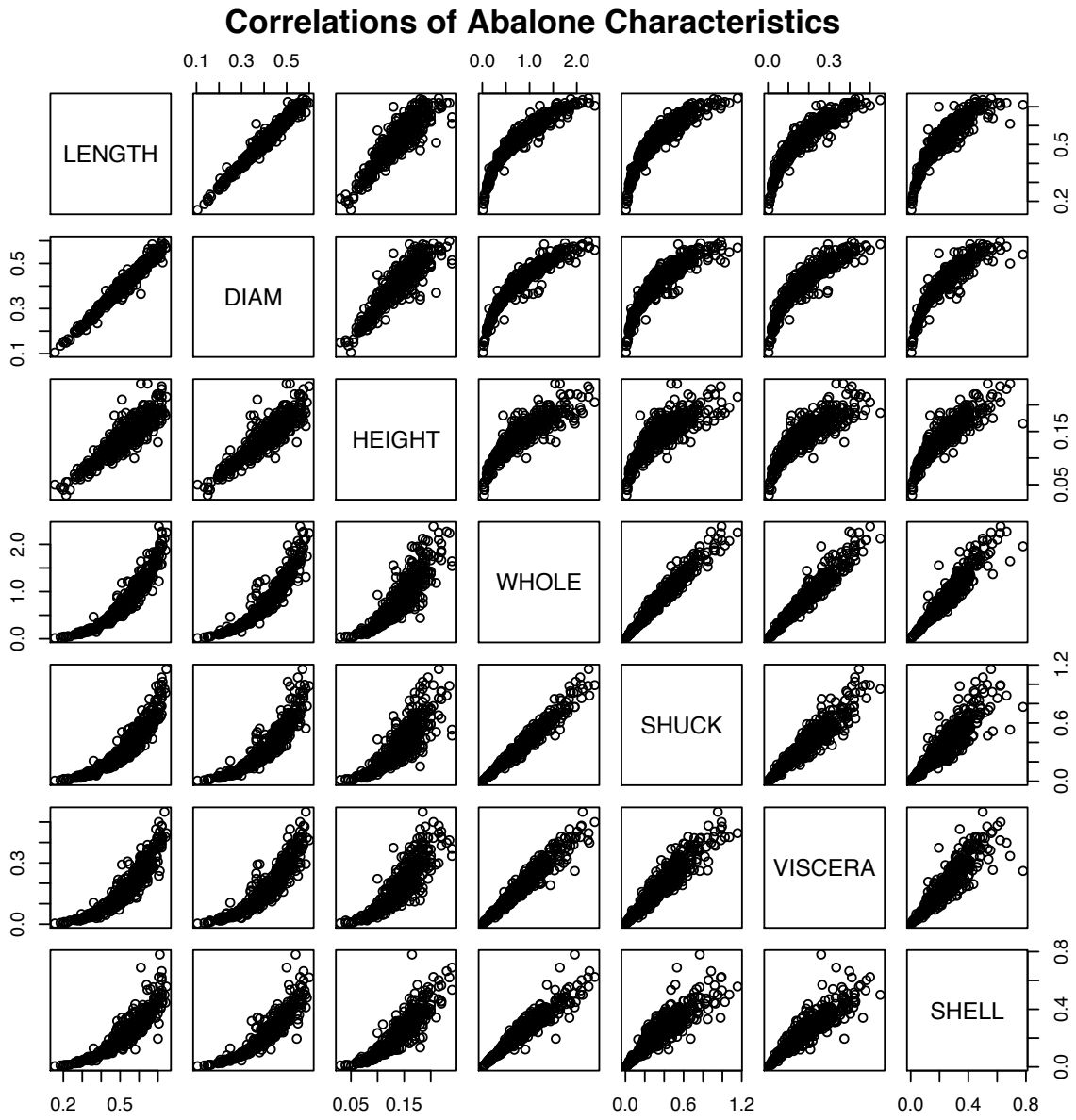
Data Analysis Assignment #2
Andrea Bruckner

Introduction

The first data analysis aimed to discover if abalone age could be discerned from physical measurements but concluded that it could not because of the considerable overlap of data values between adult and infant abalone. In this analysis, we are trying to determine if physical measurements can be used as a basis for harvesting abalone. By analyzing variance of physical measurements and looking at linear regression and other types of analysis, we will try to see if decision rules can be constructed. Establishing cut off values for certain physical measurements can theoretically enable harvesters to capture only the adult abalone while leaving the infant abalone alone to grow and reproduce so that there is a continued supply of abalone for future harvests. However, these cutoff values must yield enough abalone to make the effort and costs of harvesting worthwhile. Given that the first analysis determined that abalone age cannot be determined from physical measurements, establishing rules for harvesting abalone in this analysis will be challenging since abalone age is a major consideration for whether abalone are suitable to harvest.



A Pearson Correlation Coefficient is appropriate for the bivariate plots showing a linear relationship between the variables, while a Spearman Correlation Coefficient is appropriate when the relationship is curvilinear. Based on the bivariate plot matrix above, the following pairs of attributes are suitable for a Pearson Correlation Coefficient: LENGTH-DIAM, LENGTH-HEIGHT, DIAM-HEIGHT, WHOLE-SHUCK, WHOLE-VISCERA, WHOLE-SHELL, SHUCK-VISCERA, SHUCK-SHELL, and VISCERA-SHELL. All the rest of the attribute pairs are more suited for a Spearman Correlation Coefficient. The tables below contain the correlation coefficients generated in R. A correlation coefficient of 1 indicates a perfect correlation, and the pairs of attributes that have a coefficient as close to 1 as possible are the most strongly correlated.

Pearson Correlation Coefficients for Linear Bivariate Plots Depicting a Spatial-Spatial Relationship			
	LENGTH	DIAM	HEIGHT
LENGTH	1	0.9839313	0.8946284
DIAM	0.9839313	1	0.8979687
HEIGHT	0.8946284	0.8979687	1

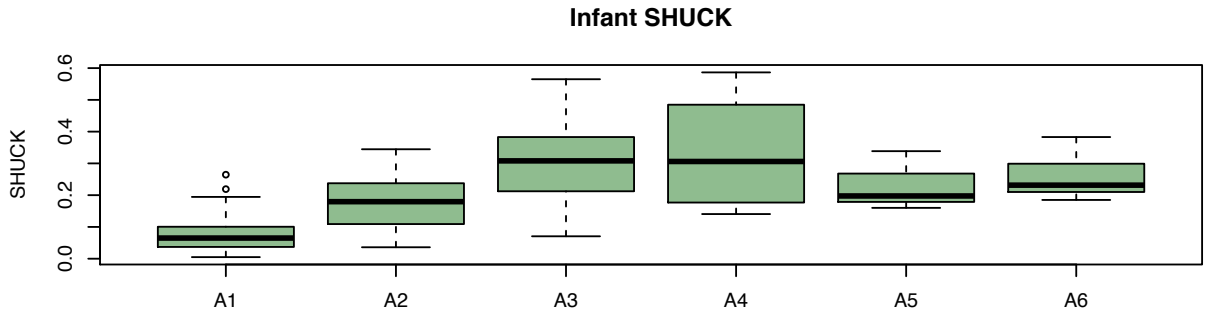
In the Pearson table showing the spatial-spatial relationships, the attribute pair with the strongest correlation is LENGTH-DIAM (or DIAM-LENGTH). All the pairs that include HEIGHT have correlation coefficient that is nearly 0.10 less than the LENGTH-DIAM coefficient. While all of the spatial-spatial attributes have a strong positive correlation, the LENGTH-DIAM relationship would be the best choice for analyzing abalone since these attributes have the strongest correlation.

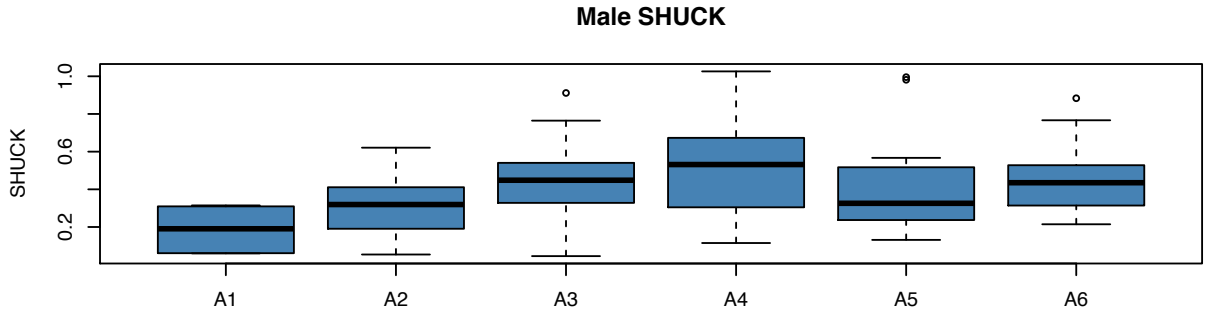
Pearson Correlation Coefficients for Linear Bivariate Plots Depicting a Weight-Weight Relationship				
	WHOLE	SHUCK	VISCERA	SHELL
WHOLE	1	0.9772859	0.9683481	0.9545941
SHUCK	0.9772859	1	0.9426071	0.8950214
VISCERA	0.9683481	0.9426071	1	0.9081349
SHELL	0.9545941	0.8950214	0.9081349	1

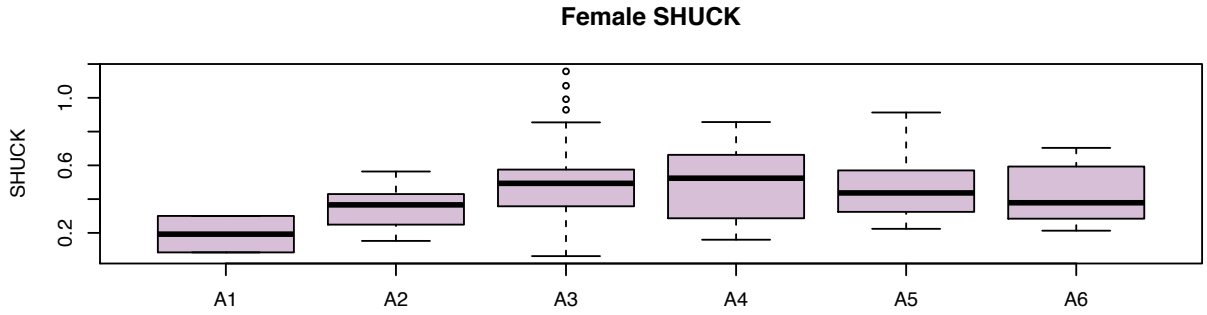
In the Pearson table showing the weight-weight relationships, the relationship between WHOLE and SHUCK shows the strongest correlation. The correlation for WHOLE-SHELL, WHOLE-VISCERA, and VISCERA-SHUCK is still very strong but just a little bit weaker than the relationship between WHOLE-SHUCK. SHELL-SHUCK and SHELL-VISCERA have the weakest correlation, although their relationship still shows a strong positive correlation since their coefficients are both close to 0.90.

Spearman Correlation Coefficients for Curvilinear Bivariate Plots Depicting a Spatial-Weight Relationship				
	WHOLE	SHUCK	VISCERA	SHELL
LENGTH	0.9709303	0.9597496	0.9529723	0.939822
DIAM	0.9656709	0.9492259	0.9465787	0.9441545
HEIGHT	0.9074207	0.8702185	0.8973609	0.918469

In the Spearman table showing the spatial-weight relationships, LENGTH-WHOLE shows the strongest relationship. All the relationships are strong since most of them have a coefficient greater than 0.90; the two weakest relationships are HEIGHT-WHOLE and HEIGHT-SHUCK, each of which have a coefficient that falls just below 0.90.







The median shucked weight for all abalone generally tends to increase with age until it decreases at age class A5. In comparison to age class A5, age class A6 shows a slight increase in median shucked weight for infant and male abalone but a continued decrease for female abalone. The boxplots in general reveal that the amount of harvestable meat is quite variable as an abalone ages. There is considerable overlap among the interquartile ranges—and ranges in general—for each age class. An abalone with a shucked weight of 0.3 g can possibly belong to any age and sex abalone when we consider the entire range and all the outliers for each class. If the boxplots had much tighter ranges and minimal outliers, we could infer that the harvestable meat in abalone correlates strongly with age. However, the wide variability and presence of outliers among all the abalone, regardless of sex, indicate that unknown factors are influencing the shucked weight of the abalone.

Below is a 2x2 contingency table with marginal totals. It reveals that abalone with a volume that falls below the median volume also tend to have a shucked weight that falls below the median shucked weight. Similarly, abalone with a volume above the median tend to have a shucked weight above the median. Only 49 abalone, or less than 10 percent of the sample size, do not follow this pattern.

		VOLUME		
SHUCK		Below	Above	Sum
	Below	226	25	251
	Above	24	225	249
	Sum	250	250	500

Chi-square Value: 323.2132
p-value: 2.89073e-72

When we run a test to calculate the Pearson chi-square statistic on the sample data, we are testing the null hypothesis that each cell in the 2x2 contingency table is equal, or contains 125 members. If the cells were all equal, then the test would show that VOLUME and SHUCK are independent. The expected value calculations are 124.5 and 125.5, but the actual values are quite different as evinced by the large chi-squared value of 323.2. Because the actual values are different than the expected values, the test indicates that there is a very strong statistical relationship. Furthermore, the p-value is very small, which indicates that there is very unlikely probability that the chi-square value is large by chance. Given the evidence that there is a strong statistical relationship, we must reject the null hypothesis and assume that SHUCK and VOLUME are strongly related.

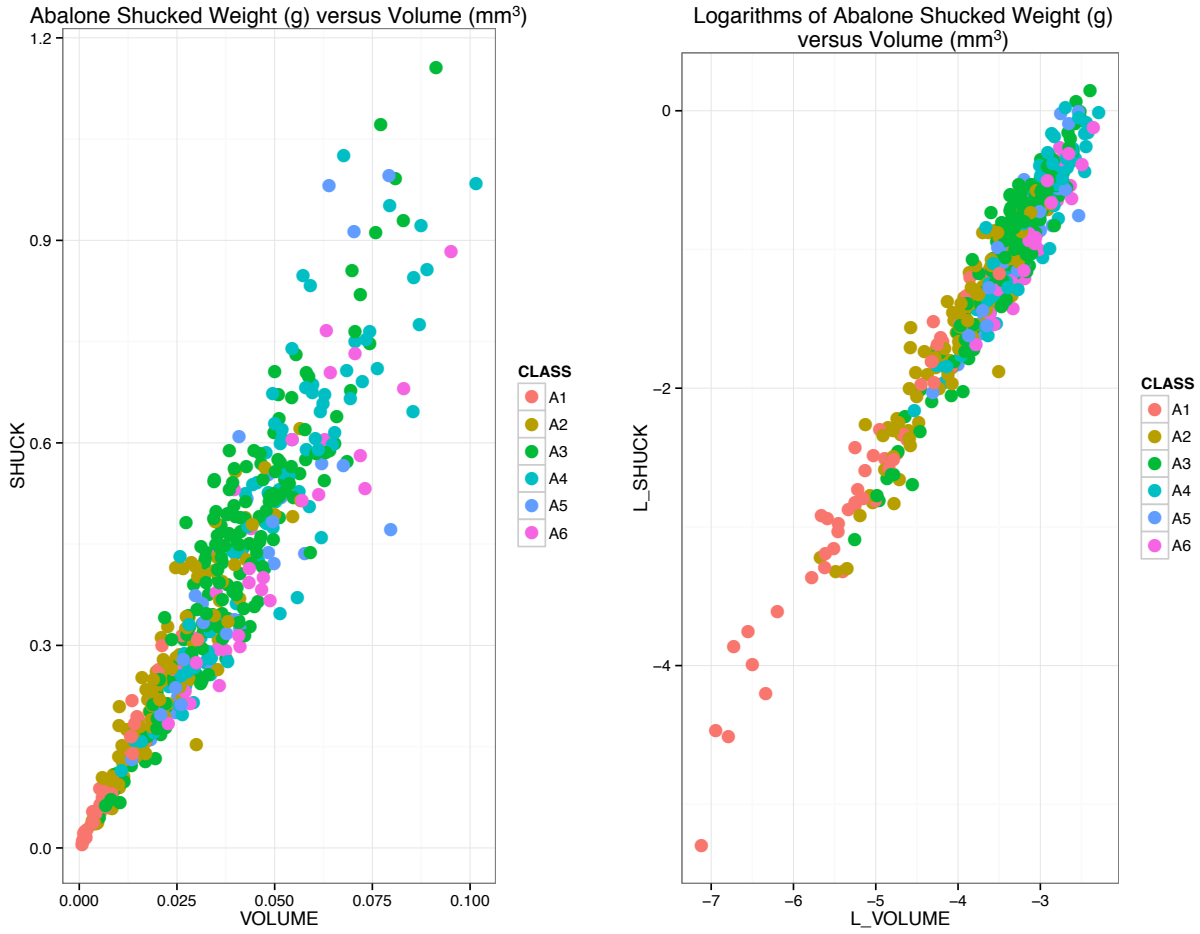
Below is an ANOVA table created using the interaction term CLASS*SEX. While the table reveals that the interaction term is not statistically significant, the high F-values and low p-values for CLASS and SEX show that both of these attributes are highly significant in relation to the SHUCK weight when they are viewed separately. The ANOVA table without the interaction term CLASS*SEX points to these same results. This means we should reject the null hypothesis that there is no difference in means for the variation of SHUCK among CLASS and SEX.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CLASS	5	7.15	1.4301	48.545	< 2e-16
SEX	2	1.914	0.9568	32.479	5.85E-14
CLASS:SEX	10	0.188	0.0188	0.637	0.783
Residuals	482	14.199	0.0295		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CLASS	5	7.15	1.4301	48.91	< 2e-16
SEX	2	1.914	0.9568	32.72	4.55E-14
Residuals	492	14.387	0.0292		

In order to evaluate which pairs of variables are different from the others, we can run the Tukey test. The table below shows that there is a significant difference for the SHUCK weight for all the classes that have a low adjusted p-value. Only pairs A5-A3, A6-A3, A5-A4, A6-A4, and A6-A5 are not significant since their lower and upper bounds of the confidence interval do not include 0. For SEX, we can see that there is a significant difference for pairs I-F and M-I but not for M-F. This indicates that SHUCK variance is significant when comparing adult (M and F) abalone to infant abalone.

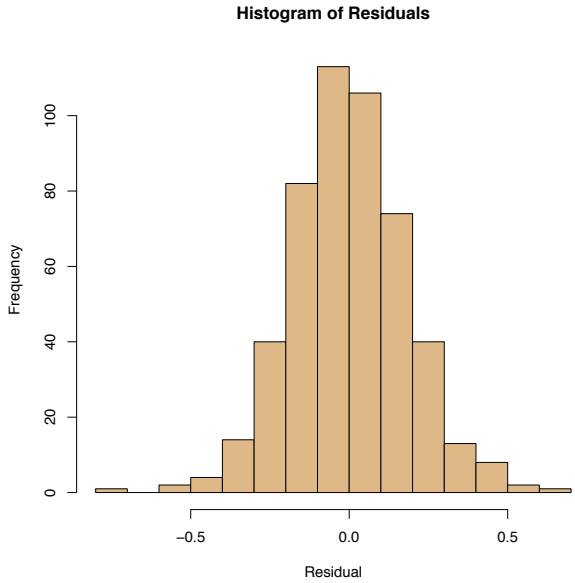
Tukey multiple comparisons of means: 95% family-wise confidence level				
Fit: aov(formula = SHUCK ~ CLASS + SEX, data = mydata)				
$CLASS				
	diff	lwr	upr	p adj
A2-A1	0.146726549	0.06315513	0.230297969	0.0000105
A3-A1	0.317324385	0.23830178	0.396346989	0
A4-A1	0.392685243	0.30599709	0.479373392	0
A5-A1	0.299591837	0.19131837	0.4078653	0
A6-A1	0.323234694	0.21496123	0.431508157	0
A3-A2	0.170597837	0.11178059	0.229415083	0
A4-A2	0.245958695	0.17718565	0.31473174	0
A5-A2	0.152865288	0.05832408	0.247406493	0.0000697
A6-A2	0.176508145	0.08196694	0.271049351	0.0000021
A4-A3	0.075360858	0.01219345	0.138528265	0.0090366
A5-A3	-0.017732549	-0.10827773	0.072812628	0.9934541
A6-A3	0.005910308	-0.08463487	0.096455485	0.9999686
A5-A4	-0.093093407	-0.19040061	0.004213799	0.0699212
A6-A4	-0.069450549	-0.16675776	0.027856656	0.3201162
A6-A5	0.023642857	-0.09330585	0.140591564	0.9924132
$SEX				
	diff	lwr	upr	p adj
I-F	-0.13046758	-0.17613478	-0.084800382	0
M-F	-0.03404416	-0.07740096	0.009312638	0.1558885
M-I	0.09642342	0.0527574	0.140089445	0.0000009

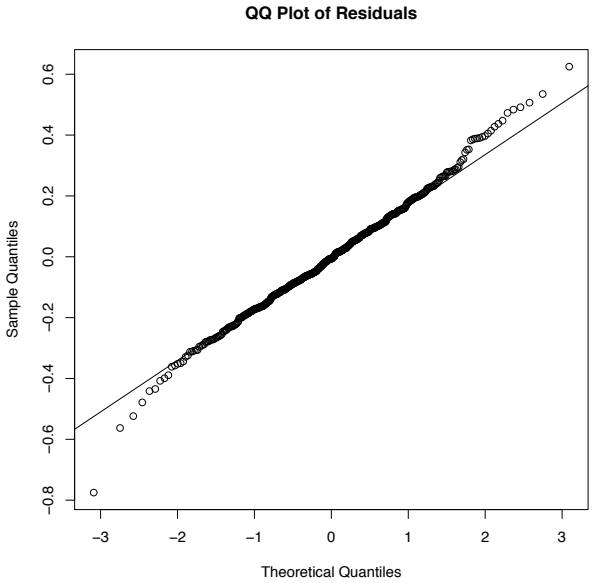


When plotting SHUCK versus VOLUME, we notice that age classes A1 and A2 are tightly gathered at the lower end of the graph, A3 occupies much of the middle of the graph, and the other classes are more spread out and tend to occur more frequently at the higher end of the graph. There is a distinct wedge shape to the data, which suggests greater variance for the relationship between SHUCK and VOLUME, especially at the higher volumes. However, when we plot the logarithms of these same two attributes, we lose the wedge shape and see a more consistent linear relationship. In the logarithms graphs, the classes also appear in more distinct groups. It is easier to see that SHUCK and VOLUME more or less increase along with CLASS, or abalone age. The oldest abalone, as represented by A5 and A6, do not quite follow this upward trend as they appear to occur near the top of the graph but do not have the highest values. This is consistent with the results shown in the boxplots on page 3.

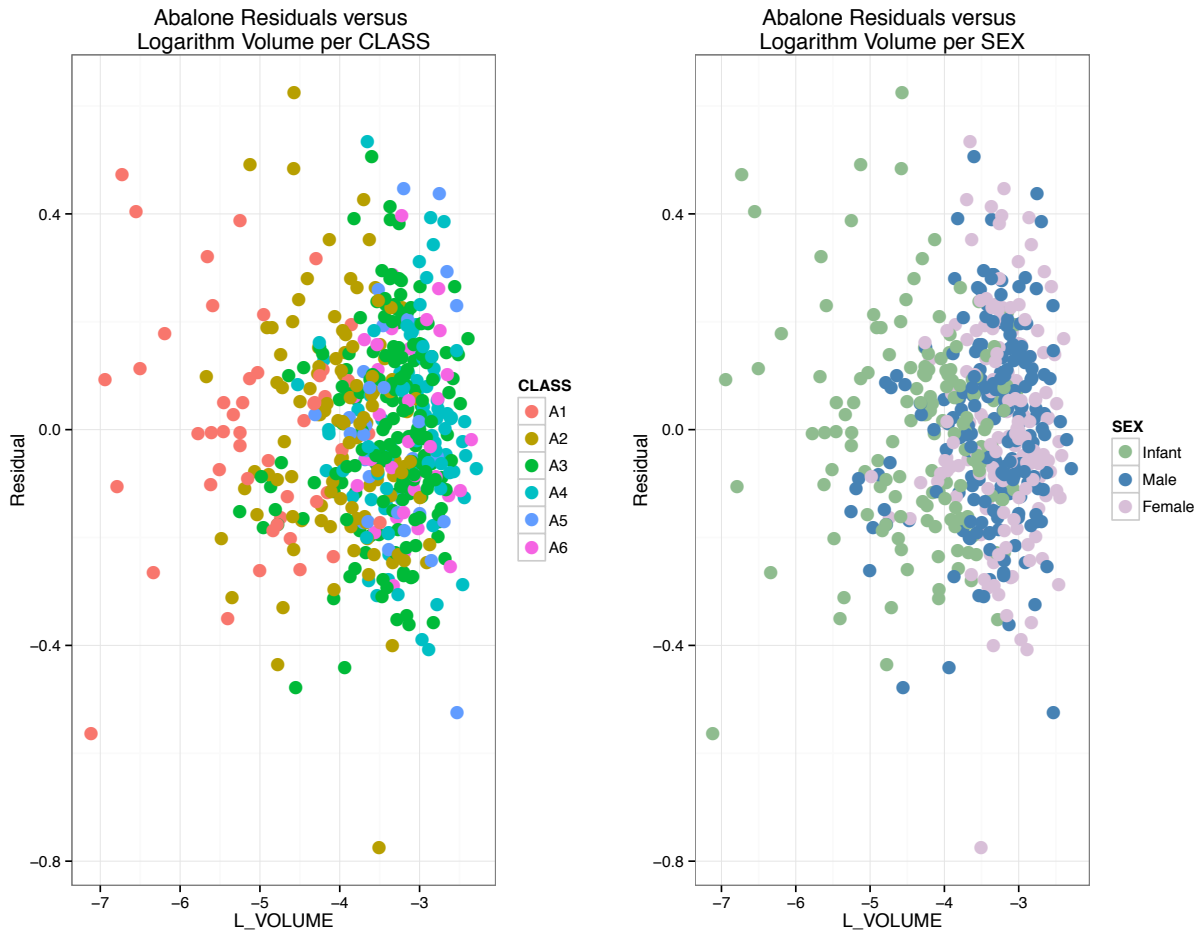
Call: lm(formula = L_SHUCK ~ L_VOLUME + CLASS + SEX, data = mydata)					
Residuals:					
Min	1Q	Median	3Q	Max	
-0.77497	-0.11656	-0.00626	0.1116	0.62489	
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	2.56468	0.08012	32.009	< 2e-16	***
L_VOLUME	1.02839	0.01585	64.89	< 2e-16	***
CLASSA2	-0.06531	0.03583	-1.823	0.06898	.
CLASSA3	-0.12701	0.03961	-3.207	0.00143	**
CLASSA4	-0.18302	0.04418	-4.143	4.04E-05	***
CLASSA5	-0.21944	0.04946	-4.436	1.13E-05	***
CLASSA6	-0.27904	0.05029	-5.549	4.71E-08	***
SEXI	0.01564	0.02551	0.613	0.54013	
SEXM	0.02665	0.02029	1.313	0.18979	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 0.1861 on 491 degrees of freedom					
Multiple R-squared: 0.9475,					
Adjusted R-squared: 0.9466					
F-statistic: 1107 on 8 and 491 DF, p-value: < 2.2e-16					

The high F-value and very low p-value shown in the table above reveals that the model is significant and that at least one of the variables in the model has a significant influence on the logarithm of SHUCK. The R-squared value is also very large, thus indicating that the model can provide some explanations for why there is a lot of variability in the SHUCK logarithm graph. The table above reveals that the variables CLASS A1 and SEX F are the basis for the model. These variables happen to be the first in their class, which is why they appear as the baseline. The other variables included in the model all have some level of significance, except for SEXI and SEXM. Although the table above does not display the significance level for SEXF and CLASS A1, it is unlikely they are significant given that the other two sexes are not significant and the other younger age classes, A2 and A3, have less significance than the three older age classes.

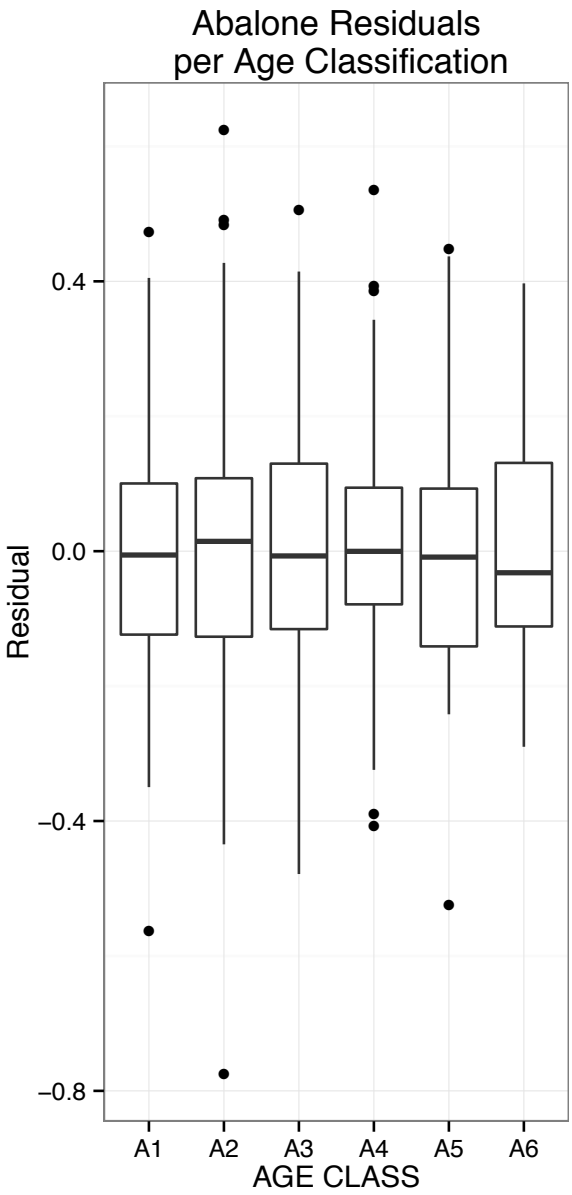


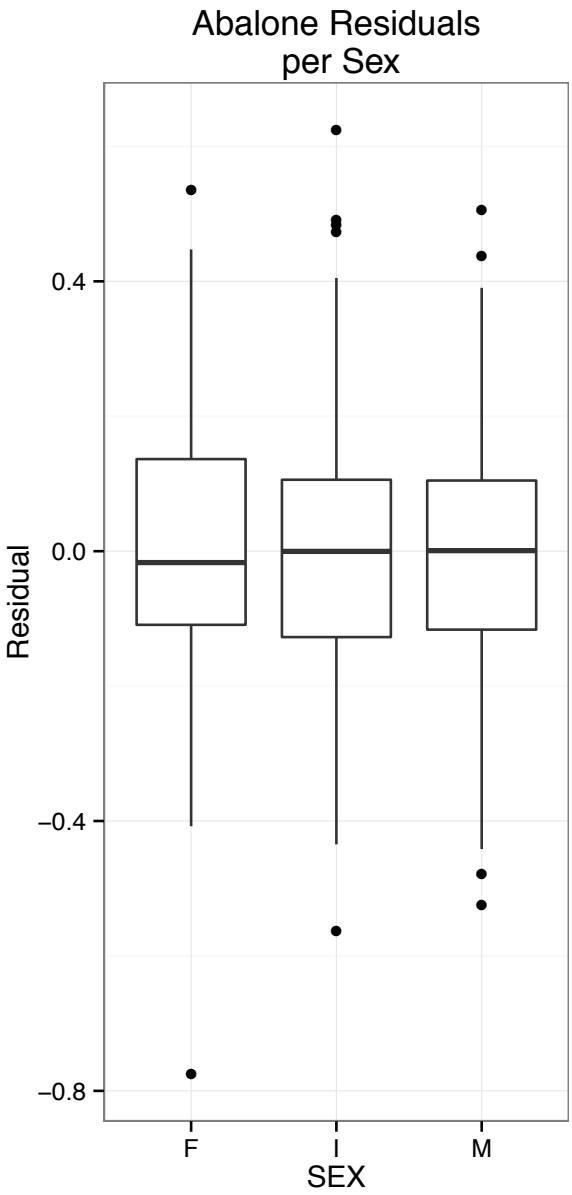


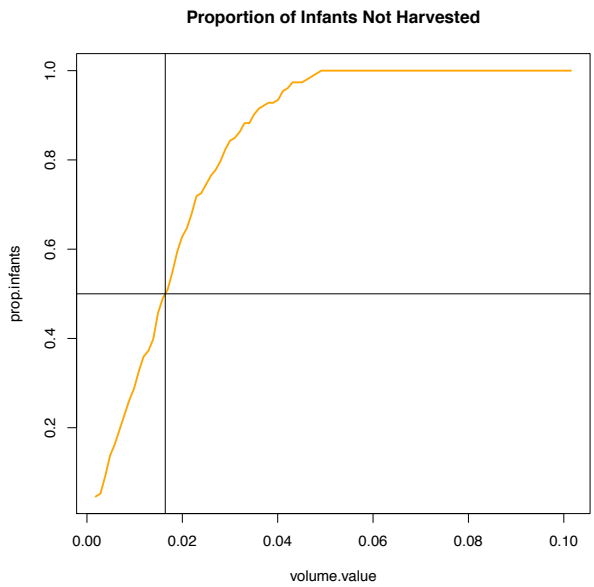
Both of the graphs above reveal that the residuals form a fairly normal distribution with a slight skew. There are a few deviant data points, as indicated by the trailing points in the QQ Plot and the stray bars in the histogram. The skewness value for the data above is approximately 0.05, which reveals that the data is skewed toward the right. The kurtosis value is approximately 3.74, which reveals that the distribution is Leptokurtic, or sharper than normal.

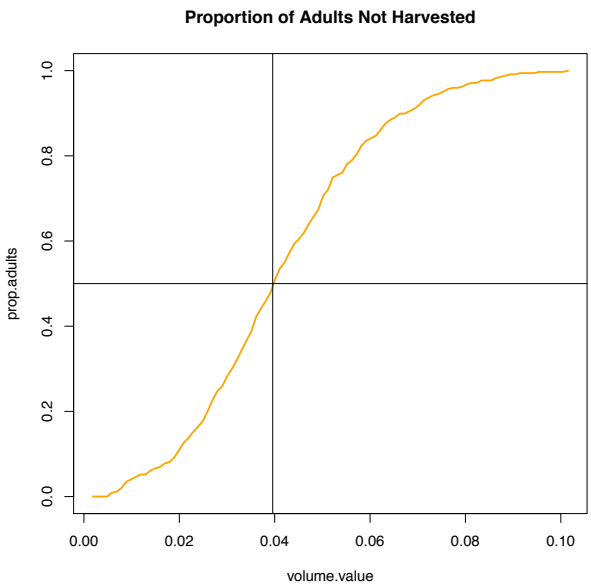


When looking at the plots above of the residuals in relation to the logarithmic VOLUME, we can see that the residuals do not follow any sort of pattern. Since residuals represent the difference between each real point and the fitted "version" of that point, the graphs should show a random plotting of points, which both graphs do. The only sort of grouping evident from the graphs above is that the younger age classes, A1 and A2, and the Infant abalone appear to occur more toward the left side of the graphs while the older age classes and adult abalone are more toward the right side of the graphs. Regardless, the plots still indicate that the model fits the data pretty well since the residuals do not contain any predictive information and they tend to be centered on zero throughout the graphs. The boxplots on the following page make it even clearer that the model fits the data well since the medians are all very close to zero, and the residuals have a fairly consistent spread. The boxplots do clearly show the outliers, but this information is consistent with the other graphs in the paper thus far and therefore does not indicate that the model is a poor fit.

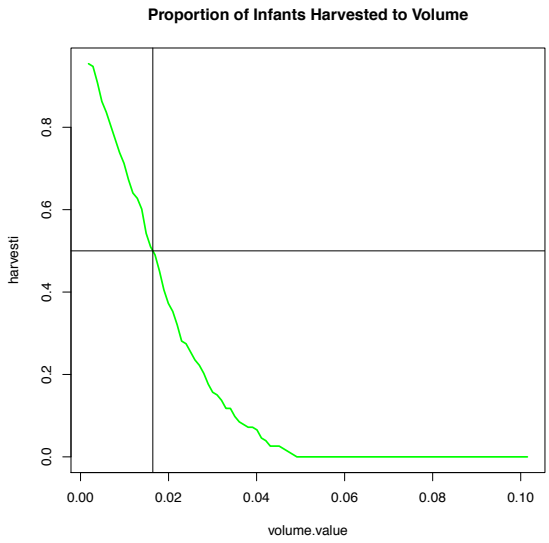

Abalone Residuals
per Age Classification
Residual
0.4
0.0
−0.4
−0.8
A1
A2
A3
A4
A5
A6
AGE CLASS

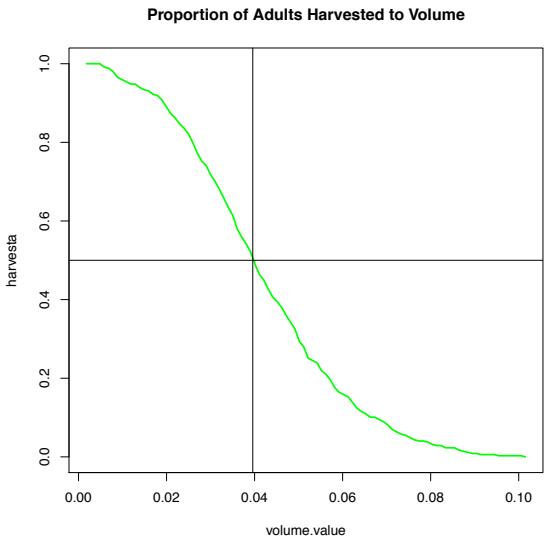

Abalone Residuals
per Sex
Residual
0.4
0.0
−0.4
−0.8
F
I
M
SEX

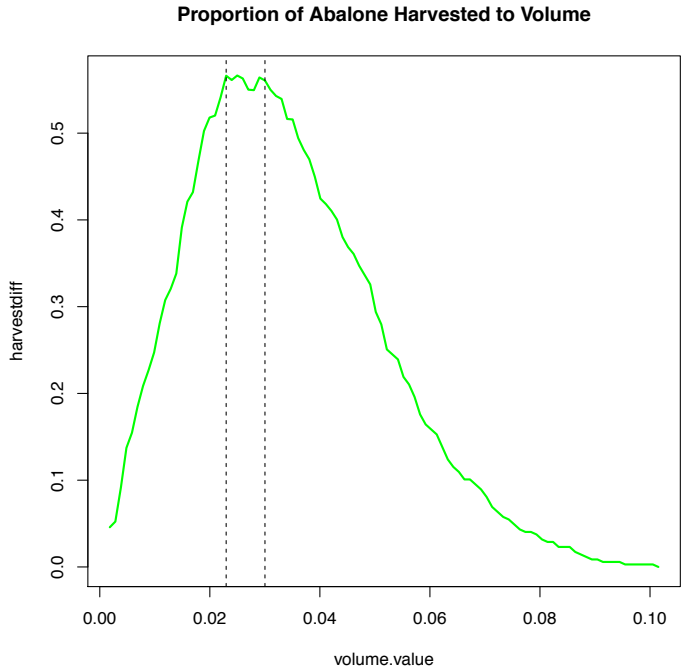




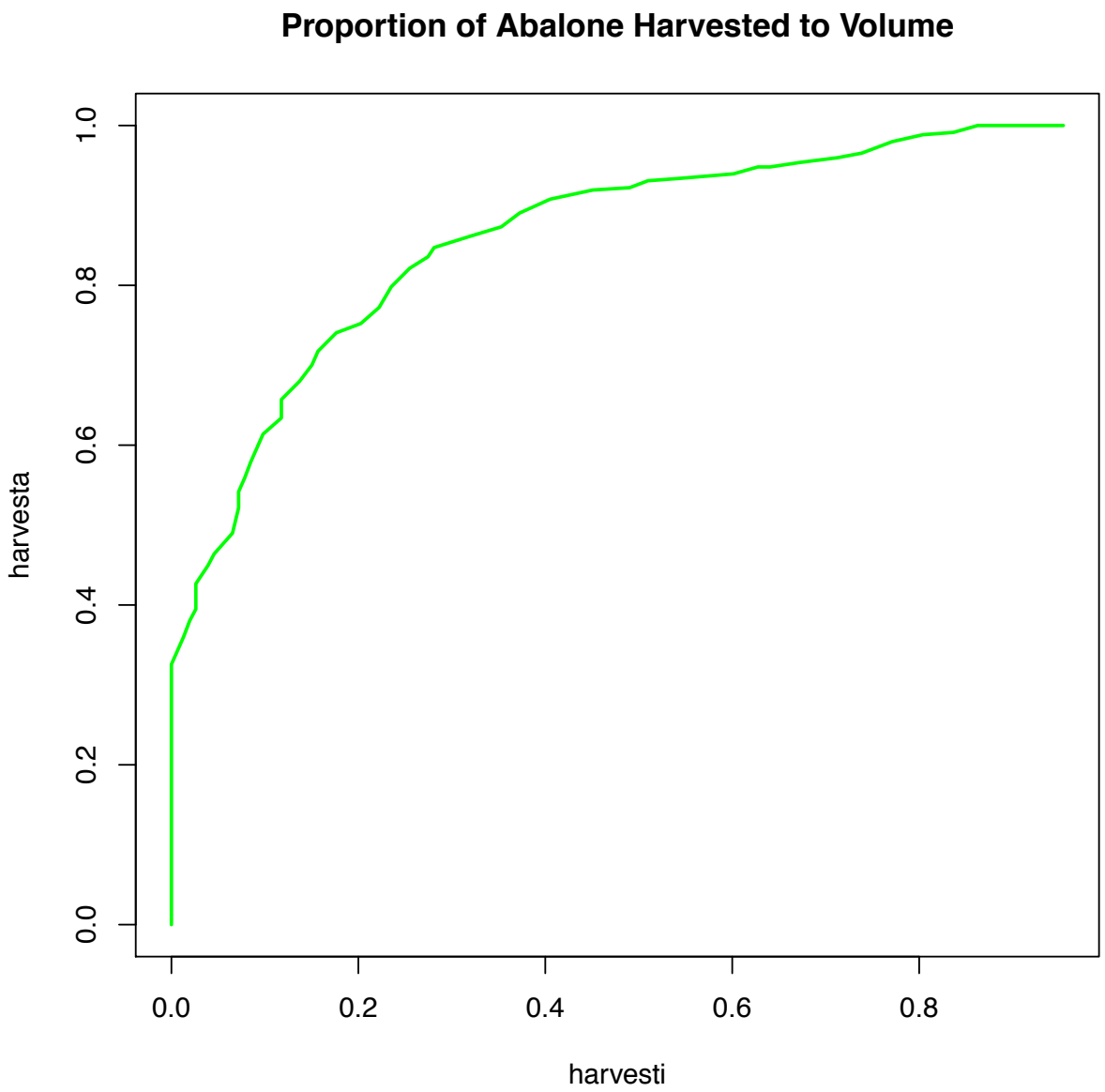
The graphs above depict the proportions of infant and adult abalone not harvested. The desired volumes for harvest differ between infant and adult abalone. The volume level that splits the infant population at 50 percent harvest is approximately 0.016, while the volume level for adults at the same percent harvest is approximately 0.040. The difference in volume is perhaps due to the fact that adult abalone have a wide variety volumes and the lower volumes of the adults often overlap with many of the volumes of the infants.







The plots above show the proportions of abalone harvested according to volume. In the top left graph, the proportion of infant abalone harvested tapers off sooner than the adult abalone since, as they grow in volume, they cease to be infants and are instead considered adults. Infant abalone tend to have smaller volumes while adult abalone have larger volumes, but since there is overlap between the two abalone groups, we want to maximize the harvest proportions for each abalone group. In the bottom graph, the two vertical dotted lines represent the range of values that can possibly maximize the difference in harvest proportions: 0.023 to 0.030.



In the graph above, the proportion of adult abalone harvested (“harvesta”) is plotted against the proportion of infant abalone harvested (“harvesti”). A threshold of zero would allow us to harvest all abalone, but this would cause the abalone populations to deplete more rapidly than they can be replaced. However, setting too high of a threshold makes for a poor harvest that would not be worth the time and effort necessary. The quick upward line at the beginning of the graph indicates that we can harvest some adults without harvesting any infants, but as the graph curves, we are more likely to unintentionally harvest infants as we aim to harvest adults. Using R, we can see that there are decreasing proportions of volume values for infant abalone until we reach a point where the point where the values are zeros. The zeros indicate that we have reached a volume that is larger than all of the infants. The smallest volume for which we do not harvest any proportion of infants is approximately 0.049. At this cutoff volume, we are harvesting approximately a proportion of 0.326 adult abalone. By using this volume threshold, we are able to harvest approximately 32.6 percent of adults while not harvesting any infants. In some ways this does seem to be a reasonable choice for a decision rule since we are able to

harvest a decent proportion of adults. If we were to harvest fewer adults, the costs associated with harvesting might not make it worth it. While harvesting more adults might be more desirable and cost effective from a harvesting perspective, a higher proportion of harvested adults could lead to more rapid population depletion.

Because classes A1 and A2 represent the youngest abalone, we want to avoid harvesting these classes. By using a cutoff proportion of 0.035, we are able to harvest the largest volume without harvesting any abalone in age classes A1 and A2. This value is slightly lower than 0.049, which is the smallest volume for which we do not harvest any proportion of infants. By lowering the cutoff, we end up harvesting some infant abalone, but we still avoid the abalone that fall in the two youngest age classes.

Conclusions

Using physical measurements as a basis for harvesting abalone is difficult. There is no perfect model that can help identify how best to harvest adult abalone while largely leaving the infant abalone unharvested. As some of the graphs in this paper have shown, some physical measurements, such as shucked weight, vary too greatly to be of much use for distinguishing adult from infant abalone. Harvesting abalone based on volume is more promising, but it does come with its own set of challenges. The largest infant abalone overlap with the smallest adult abalone. If we want to avoid harvesting any infant abalone, we are able to harvest slightly less than a third of adult abalone. While this would certainly ensure that the abalone population could continue to reproduce, this proportion might not be attractive to harvesters. There is an increased level of effort and cost involved in attempting to harvest just the adults, and a harvest of less than a third may not be financially worth it. However, harvesting more abalone could put the population at risk and jeopardize future abalone harvests if the population gets too low because too many infants were harvested.

Before putting too much reliance on the volume cutoff, we would be wise to research abalone more. There seem to be unknown factors, environmental or otherwise, that influence abalone growth. Understanding what some of these factors are can help us make a better model. This highlights a major challenge in working with observational study data. Because there are no controls with gathering the data, we cannot necessarily apply the model we create based on this sample data to other abalone samples. Working with more carefully collected data can help us create better models that might produce a higher threshold for abalone harvesting that would be more economically attractive to harvesters while still being biologically conscious of the abalone populations.

I have learned a lot about abalone from this analysis, but I have come away more with the conviction that clean data collection and a healthy skepticism about where the data is coming from is necessary if we want to create effective and relevant models.