

Predicting Size of Pre-Molt Crab Shells Based on the Size of Pre-Molt Crab Shells

McCord Murray

February 12, 2022

Abstract

The goal of this study is to determine if there is a high correlation between a crab's post-molt shell size and its pre-molt shell size such that a model can be used to predict the pre-molt size. In order to test our hypothesis, data was used from a study in 1989 on the growth and reproductive dynamics of adult female Dungeness crabs. The descriptive statistics show that the residuals from the lab-grown crabs follow a normal distribution more closely than the residuals for the ocean-caught crabs. The lab-grown crabs' residuals followed a normal distribution closely while the distribution of the residuals of ocean-caught crabs had a higher kurtosis. It seems possible that the environment of the lab had an impact on the growth of the lab-grown crabs.

Background and Significance

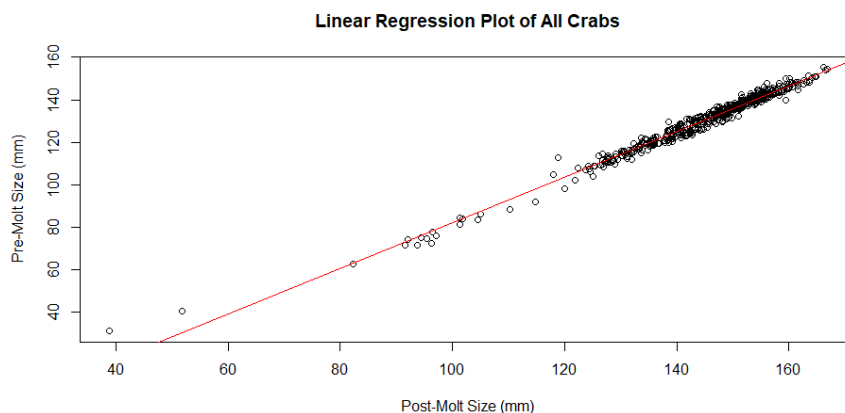
It has been shown that many aspects of nature follow different patterns, varying from symmetry to complex geometric shapes in living things. Depending on the complexity of these patterns, humans have been able to predict the outcomes of natural events. The natural event being explored in the study involves predicting the size of crab shells before and after molting. One big disadvantage in predicting patterns in nature is that sometimes we are not able to observe our results, especially if we are trying to predict what something **was like** instead of what it **will be**. The goal of this study is to determine if there is a high correlation between a crab's post-molt shell size and its pre-molt shell size such that a model can be used to predict the pre-molt size. The hypothesis of this study is that we can predict, with a high degree of accuracy, the size of a crab's previous shell based on the size of its current shell.

Methods

In order to test our hypothesis, data was used from a study in 1989 on the growth and reproductive dynamics of adult female Dungeness crabs. The data contains five variables. The first variable is 'presz', which is the size of the shell before molting. The second is 'postsz', which is the size of the shell after molting. The third is 'inc', which is the difference in size between the pre-molt and post-molt shells. The fourth is 'year', which is the year the data was collected. The last variable is 'lf', which says whether the crab was caught in the ocean or raised in a lab. Since we are trying to predict the size of the pre-molt shells based on the size of the post-molt shells, the post-molt shells are the independent variable and will be our x-axis in the linear model. Next, the crabs should be separated into three different groups. Crabs that were caught in the ocean, crabs that were grown in a lab, and the combination of both groups. For all three of these groups, a linear regression model will be created that gives us an equation for the line best fit. Then, for all three lines the residuals will be calculated and plotted. From the residuals we will determine whether or not the regression models for each group are heteroscedastic or homoscedastic. First, the residuals distribution should be tested for normality with a Q-Q plot. If the residuals follow a normal distribution, a Breusch-Pagan test can be used to test for heteroscedasticity. If the distribution is not normal, the variance of the residuals can be evaluated at multiple places in the distribution to determine whether or not the variance is changing by a significant amount. The result of this test will ultimately determine whether or not our models will be good at predicting data with a greater range.

Results

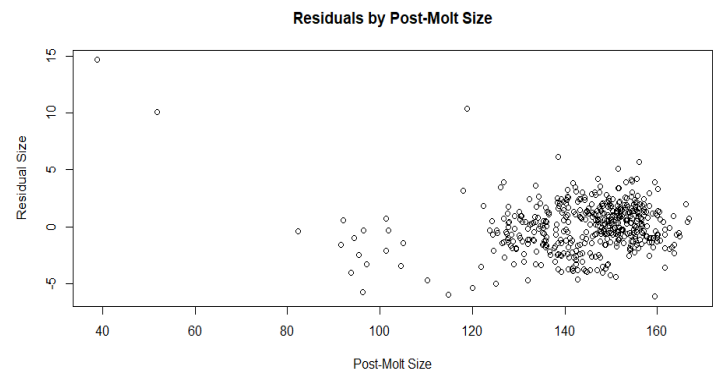
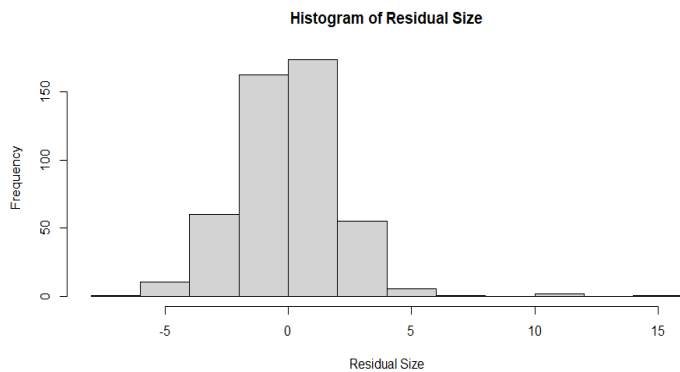
The linear regression plot of the sizes of crab shells with the post-molt size being the predictive variable yielded a line best fit of $1.073x - 25.214$ and a Pearson's R-Squared value of **0.9808326**.



When using the equation $1.073x - 25.214$ to predict the pre-molt sizes of other crabs, the following data was observed:

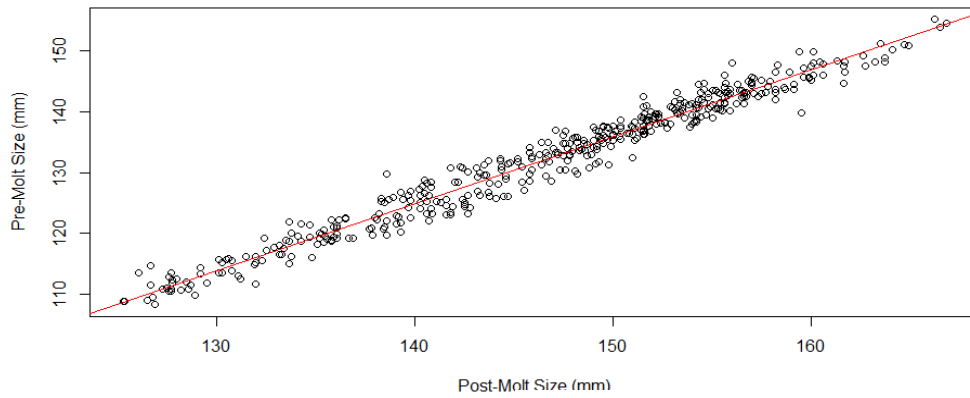
Post-molt	127.7	133.2	154.8	142.5	120.0	134.1	133.8
Actual Pre-molt	113.6	118.1	142.3	125.1	98.2	119.5	116.2
Predicted Pre-molt	111.8	117.7	140.9	127.7	103.5	118.7	118.4
Difference	1.8	0.4	1.4	2.6	5.3	0.8	2.2

Examining residuals shows that most of the residual sizes were close to 0 and the distribution had a very high kurtosis of 8.378684. However, constructing a histogram of the residuals and plotting out the residuals over post-molt size shows that there are some outliers causing unwanted noise in the data.

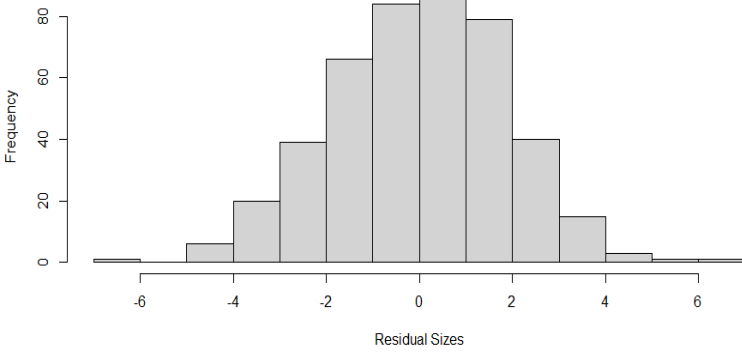


Replotting these graphs with all post-molt shell sizes less than 125 mm omitted, resulted in the following visualizations:

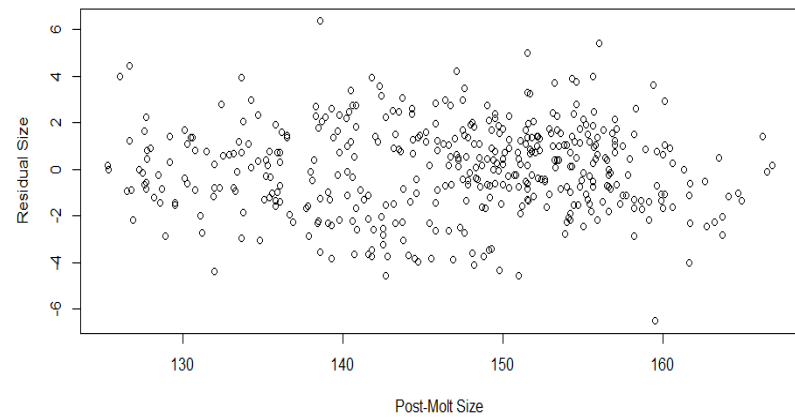
Linear Regression Plot of All Crabs



Histogram of Residual Sizes



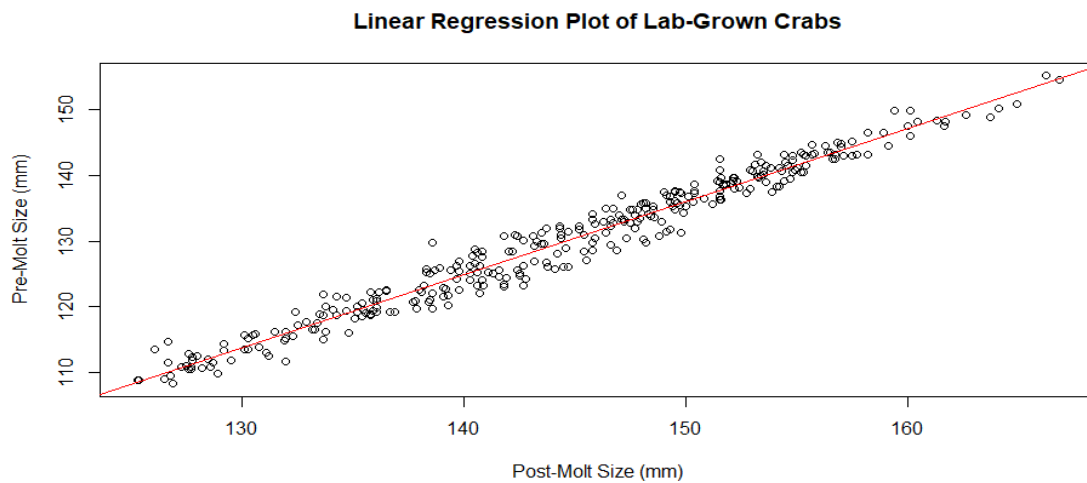
Residuals by Post-Molt Size



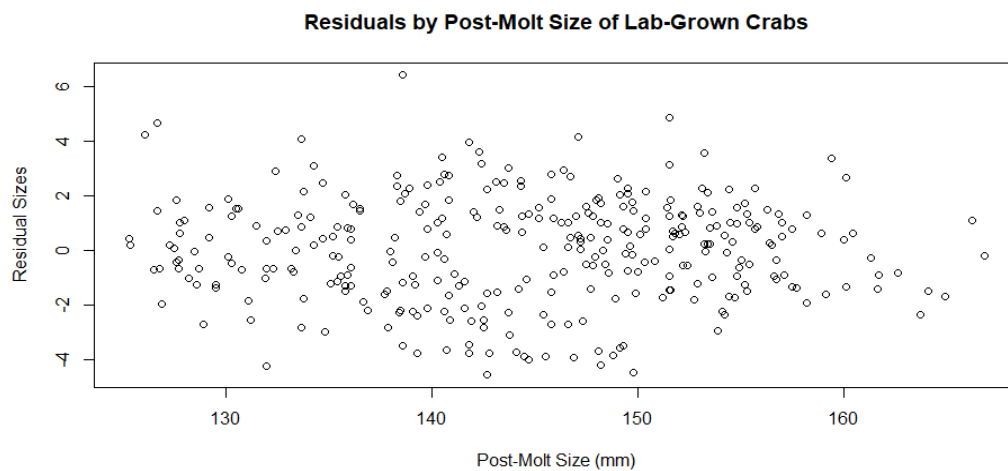
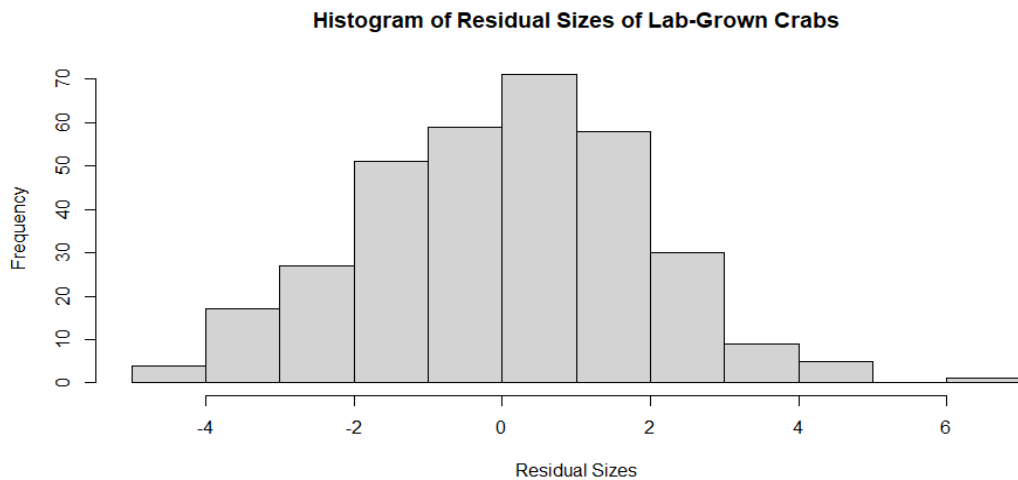
The resulting equation of the linear regression line for all crabs with all post-molt sizes less than 125 mm omitted is **$1.099x - 28.995$** with a Pearson's R-Squared Value of **0.967966**. When using this equation to predict the pre-molt sizes of other crabs, the following data was observed:

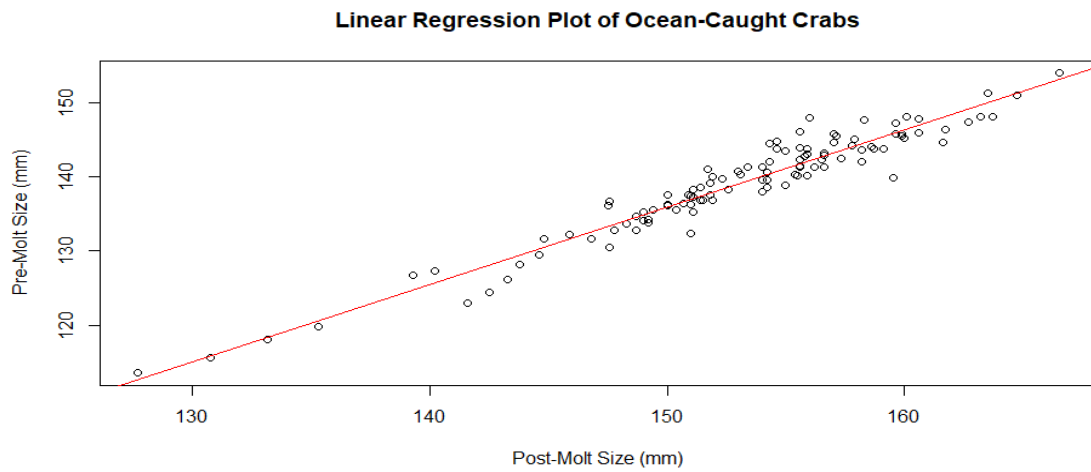
Post-molt	127.7	133.2	154.8	142.5	120.0	134.1	133.8
Actual Pre-molt	113.6	118.1	142.3	125.1	98.2	119.5	116.2
Predicted Pre-molt	111.4	117.4	141.1	127.6	102.9	118.4	118.1
Difference	2.2	0.7	1.2	2.5	3.7	1.1	1.9

Additional visualizations were plotted separately for crabs raised in labs and crabs caught in the ocean.

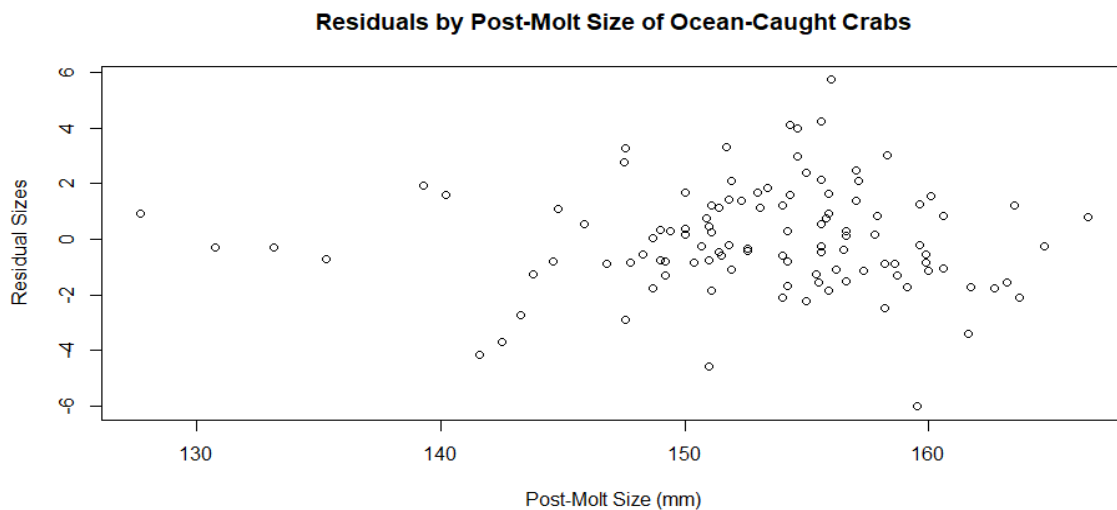
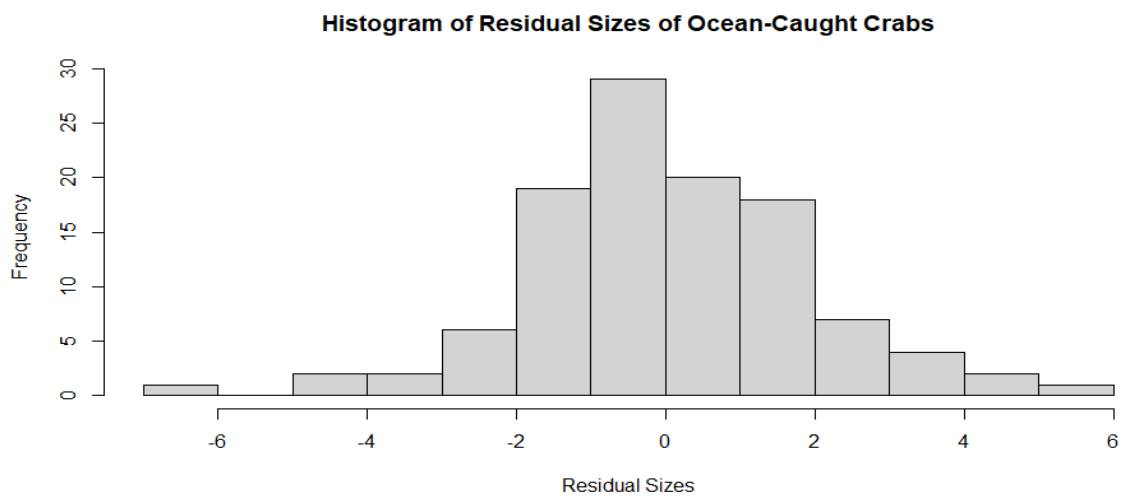


The resulting equation of the linear regression line for lab-grown crabs is $1.113x - 30.985$ with a Pearson's R-Squared Value of **0.967887**.





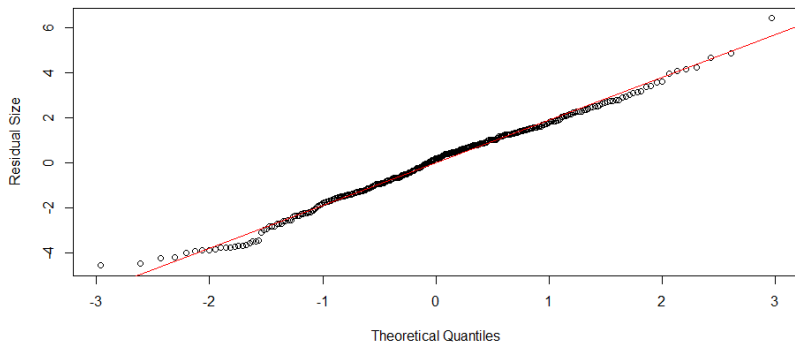
The resulting equation of the linear regression line for ocean-caught crabs is $1.042x - 20.402$ with a Pearson's R-Squared Value of **0.9327747**.



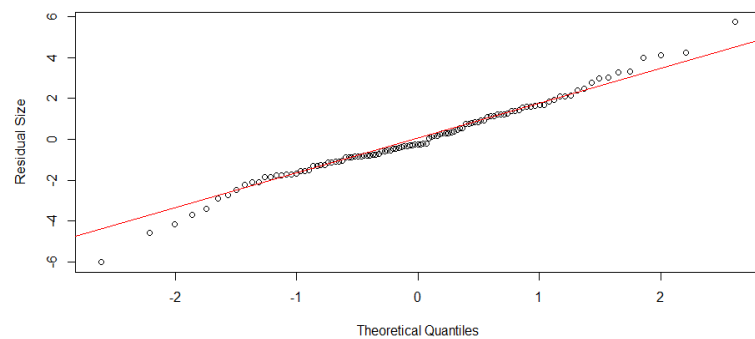
The descriptive statistics show that the residuals from the lab-grown crabs follow a normal distribution more closely than the residuals for the ocean-caught crabs. Namely, the kurtosis of the residuals for the lab-grown crabs is closer to 3 than the kurtosis for the ocean-caught crabs. Constructing QQ-Plots of the residuals for the lab-grown and ocean-caught crabs supports this.

Descriptive Statistics of Residuals	All Crabs	Lab-Grown Crabs	Ocean-Caught Crabs
Mean	0	0	0
Median	0.1036037	0.172492	-0.2496302
Standard Deviation	1.8921	1.879751	1.880067
Skewness	-0.0663074	-0.06161526	0.03555623
Kurtosis	3.1165	2.959804	3.912915
5-Number Summary	-6.5156, -1.2919, 0.1036, 1.3107, 6.3561	-4.5681, -1.2840, 0.1725, 1.2692, 6.3960	-6.0205, -1.0908, -0.2496, 1.2111, 5.7270

Q-Q Plot for Lab-Grown Crabs

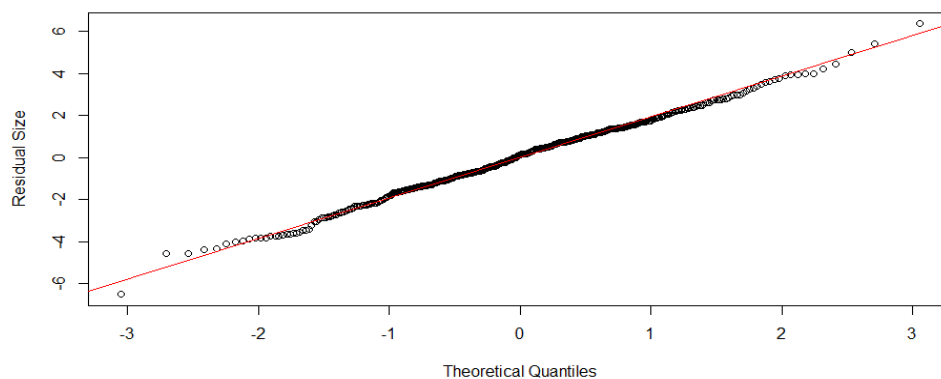


Q-Q Plot for Ocean-Caught Crabs

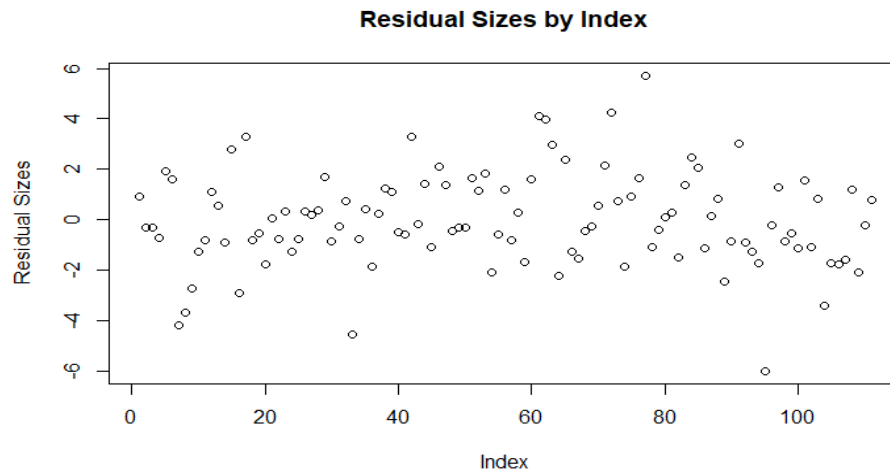


The distribution of the residuals for all of the crabs in this experiment is also close to normal.

Q-Q Plot for All Crabs



Since the distributions of the residuals for all crabs and lab-grown crabs are close to normal, Breusch-Pagan tests were used to test for heteroscedasticity in the distributions. The p-value for the lab-grown crabs was 0.5922 and the p-value for all crabs was 0.2153, neither of which are low enough to reject the null hypothesis that the distribution is homoscedastic. For the ocean-caught crabs, the variance of the residuals had to be measured to test for scedasticity.



The variance of the residuals in the ocean-caught crabs plot starts at around 3.5, increases up to 4.5 towards the middle, and decreases back down to 3.3 shortly after. The variance of residuals towards the very end was not considered as the data may be misleading. As a result, there is not enough evidence to suggest that any of the 3 distributions of residuals are heteroscedastic.

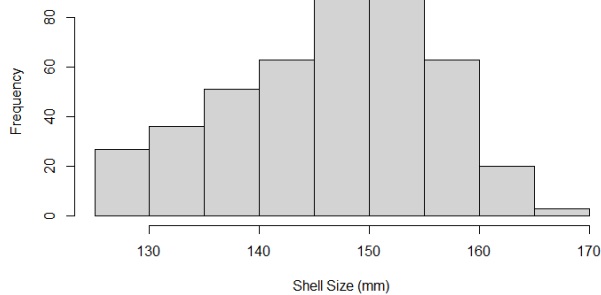
Descriptive Statistics of Post-Molt Shell Size	All Crabs	Lab-Grown Crabs	Ocean-Caught Crabs
Mean	146.5122	144.3551	152.964
Median	148	145.2	154
Standard Deviation	9.462834	9.270429	6.719967
Skewness	-0.3308025	-0.09535862	-1.119064
Kurtosis	2.317466	2.272376	5.240706
5-Number Summary	125.3 139.8 148.0 154.0 166.8	125.3 137.8 145.2 151.5 166.8	127.7 150.0 154.0 157.0 166.5

Descriptive Statistics of Pre-Molt Shell Size	All Crabs	Lab-Grown Crabs	Ocean-Caught Crabs
Mean	132.0404	129.7105	139.009

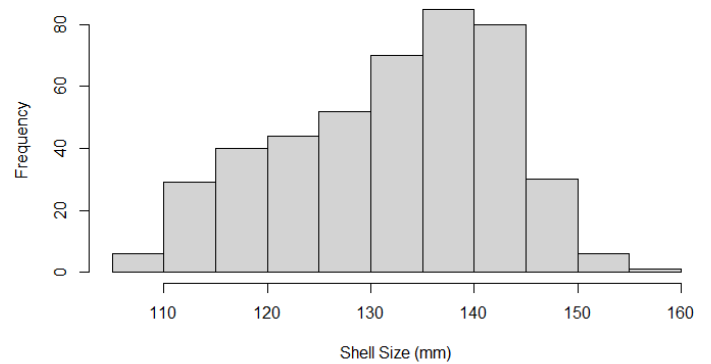
Median	133.7	130.6	140.1
Standard Deviation	10.57155	10.48962	7.251151
Skewness	-0.3435447	-0.08527832	-1.110875
Kurtosis	2.230345	2.176644	4.761443
5-Number Summary	108.3 124.2 133.7 140.4 155.1	108.3 121.5 130.6 137.9 155.1	113.6 136.1 140.1 143.8 153.9

The histograms for the different distributions of pre-molt and post-molt sizes show that the ocean-caught crabs have a higher peak than the other two distributions.

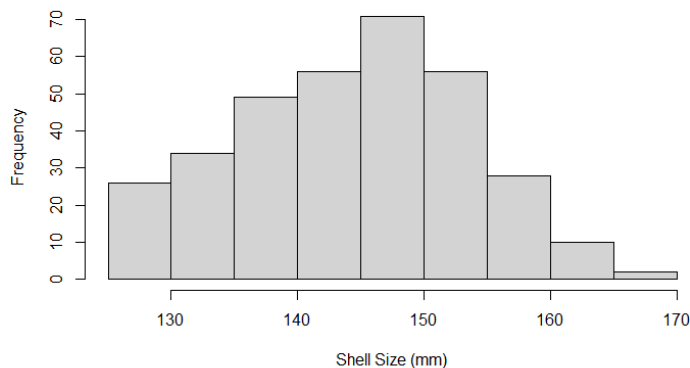
Histogram of Post-Molt Shell Size



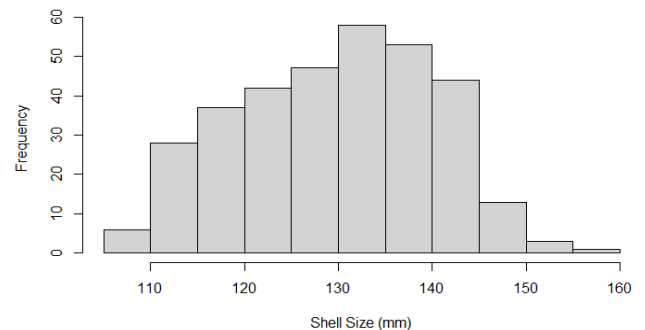
Histogram of Pre-Molt Shell Size



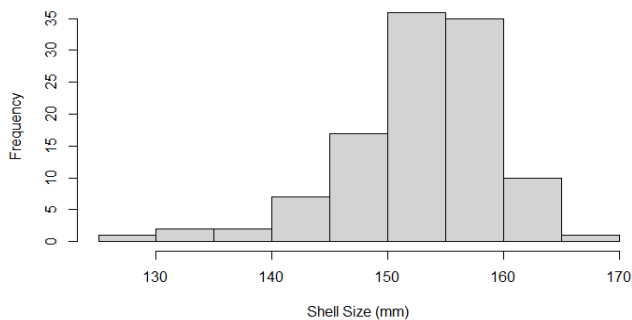
Histogram of Post-Molt Shell Size for Lab-Grown Crabs



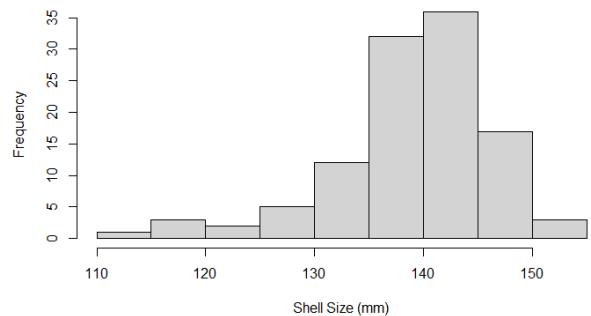
Histogram of Pre-Molt Shell Size for Lab-Grown Crabs



Histogram of Post-Molt Shell Size for Ocean-Caught Crabs



Histogram of Pre-Molt Shell Size for Ocean-Caught Crabs



Conclusions

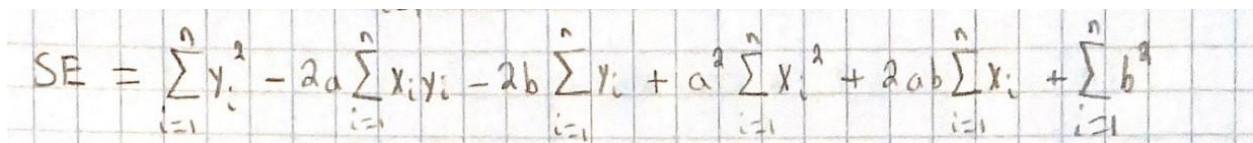
The goal of this study was to determine if there is a high correlation between a crab's post-molt shell size and its pre-molt shell size such that a model can be used to predict the pre-molt size. It was shown that for all crabs with a post-molt shell size greater than 125 mm, the correlation between the post-molt shell size and pre-molt shell size had a Pearson's R-Squared value of 0.967966 and residual sizes that closely followed a normal distribution. A Breusch-Pagan test of the residual distribution resulted in a p-value of 0.2153 which was not low enough to reject homoscedasticity. When separating the crabs into groups of lab-grown and ocean-caught, the distributions of the residuals were slightly different. The lab-grown crabs' residuals followed a normal distribution closely while the distribution of the residuals of ocean-caught crabs had a higher kurtosis. The reasoning for this difference may be due to most of the post-molt sizes of the ocean-caught crabs being between 150 mm and 160 mm whereas the range of shell sizes for lab-grown crabs is larger; mostly between 130 mm and 160 mm. It seems possible that the environment of the lab had an impact on the growth of the crabs. This study would be more meaningful if it focused only on ocean-caught crabs who grew up in their natural habitat. Since the ocean-caught crabs did not follow a normal distribution it's difficult to test for heteroscedasticity. For better results, a large sample of ocean-caught crabs should be tested. The ocean-caught crabs' shell size distribution has a high peak, meaning most of the data falls into the same region, and subsequently, is easier to predict.

Appendix

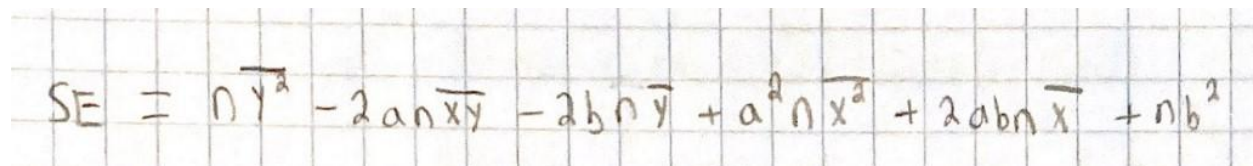
For a data set $(x_1, y_1), \dots, (x_n, y_n)$ show how to find the values for a, b that minimize the sum of squares

$$S(a, b) := \sum_{i=1}^n (y_i - (ax_i + b))^2$$

First the equation needs to be expanded.


$$SE = \sum_{i=1}^n y_i^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + a^2 \sum_{i=1}^n x_i^2 + 2ab \sum_{i=1}^n x_i + \sum_{i=1}^n b^2$$

Next, SE needs to be rewritten in terms of the averages of x and y (\bar{x} & \bar{y}), and n .


$$SE = n\bar{y}^2 - 2an\bar{x}\bar{y} - 2bn\bar{y} + a^2n\bar{x}^2 + 2abn\bar{x} + nb^2$$

Now SE is in a better form to find the partial derivatives with respect to a and b .

$$\frac{\partial SE}{\partial a} = -2n\overline{xy} + 2an\overline{x^2} + 2bn\overline{x} = 0$$

$$a\overline{x^2} + b\overline{x} = \overline{xy}$$

$$\frac{\partial SE}{\partial b} = -2n\overline{y} + 2an\overline{x} + 2nb = 0$$

$$a\overline{x} + b = \overline{y}$$

Now we have a system of equations that we can use to solve for a and b.

$$\begin{array}{l} a\overline{x^2} + b\overline{x} = \overline{xy} \\ a\overline{x} + b = \overline{y} \end{array} \rightarrow \begin{array}{l} \frac{a\overline{x^2}}{\overline{x}} + b = \frac{\overline{xy}}{\overline{x}} \\ a\overline{x} + b = \overline{y} \end{array}$$

The top equation is modified by dividing the whole equation by \overline{x} so that b can be eliminated.

$$\frac{a\overline{x^2}}{\overline{x}} - a\overline{x} = \frac{\overline{xy}}{\overline{x}} - \overline{y}$$

$$a = \frac{\frac{\overline{xy}}{\overline{x}} - \overline{y}}{\frac{\overline{x^2}}{\overline{x}} - \overline{x}} \cdot \frac{\overline{x}}{\overline{x}}$$

Then using algebra, the equation can be solved for a.

$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

Finally, using one of the equations we used earlier, we can solve for b.

$$a\bar{x} + b = \bar{y} \rightarrow b = \bar{y} - a\bar{x}$$