Predicting Pre-Molt Sizes of Crabs Using Recorded Post-Molt Sizes

Written by: Nick Paternostro

Abstract

Due to fishing restrictions on female crabs, nearly the adult male populations of crabs are fished each year. Based on previous studies, allowing female crabs to be fished has been a method of controlling the fluctuation of the crab population, especially during the fishing season. Biologists are interested in placing size restrictions on the male crab population to allow crabs to mate before they are fished. To determine suitable size restrictions for the crab populations, biologists must study the growth patterns of crabs, specifically, the size of the shell. Data is gathered from different crab populations: crabs from a laboratory and crabs captured and recaptured. Using existing pre-molt and post-molt sizes, a least-squares regression line is generated for these 2 crab populations as well as the entire crab population. In addition to creating a linear model, the residuals of each population are analyzed to determine if the linear model can be treated as an appropriate method for prediction. The linear model is then applied to the recorded post-molt sizes of section 2 and predicted pre-molt sizes are found. There is some variation amongst the predicted pre-molt sizes due to the fouled shells in the sample. To determine the effect of fouling and inaccurate measurements on the post-molt size, hence the predicted pre-molt size, the post-molt sizes from section 2 are divided into clean shells and fouled shells. Once again, using the linear model, the predicted pre-molt size is found and a comparison between sets indicates the clean shells have less variation in post-molt sizes. Lastly, a more stable method for predicting pre-molt sizes based on recorded post-molt sizes is introduced.

Introduction

Dungeness crabs are one of the most popular catches during the summer season in regions such as the Pacific Coast of North America. Nearly the entire adult male population is fished yearly, however, restrictions are placed on female crabs to prevent over-fishing of the crab population. Year after year there exists a smaller population of these Dungeness crabs in the ocean. Many biologists believe if female crabs are allowed to be fished, the population of crabs would increase. Biologists have stated that an imbalance in the sex ratio has been a factor in the decline of the crab population. If the male population would grow, there would-be long-term benefits on the Dungeness crab population. Previous studies have found allowing female crabs to be fished will have a positive effect on the crab popula-

tion such as decreasing fluctuations in the population during the fishing season. However, to place a restriction on the fishing of male crabs, a minimum size must be stated to prevent these crabs from being captured. To generate a restriction, biologists require specific information relating to molting to predict the growth patterns of crabs. One problem with measuring the size of crabs is that crabs molt regularly; meaning crabs will cast off and shell and grow a new one. This is problematic as the shell sizes are used to determine the growth patterns on crabs which will assist biologists in developing restrictions for the minimum allowable size for fishing of crabs. To study the growth patterns of these crabs, biologists are interested in generating a model using current data for pre-molt sizes and post-molt sizes of crabs that are either in a laboratory or have been captured and recaptured. Using this model, pre-molt sizes can be predicted, with some small error, based on a set of post-molt sizes recorded.

Methods

This data was imported from Mathematica as part of a study of the adult female Dungeness crab conducted by Hankin, Diamond, Mohr, and Ianelli with assistance from the California Department of Fish and Game along with commercial crab fishers from northern California. This experiment was conducted to help biologists develop an interpretive model for predicting pre-molt sizes of crabs based on recorded post-molt size. The data in this experimental study was a mix of some laboratory data and some capture-recapture data. This information was obtained from December to June during 1981,1982 and 1992. The size measurement of the external carapace was made along the widest part of the shell. The crabs examined in a laboratory setting were collected during the molting season of female crabs. The pre-molt carapace was measured when the crab was first collected and post-molt measurements were made three to four days after the crab left its old shell to ensure the new shell was fully developed. For the cases where the crab was captured and recaptured, the crabs were caught, measured, and then tagged with a method of identification before being released into the water. This was done during the months of January through March right before seasonal molting. After molting season, commercial fisheries would bring crabs they had caught with a tag to be measured for post-molt measurement. This data of pre-molt and post-molt measurements for crabs examined in a laboratory and crabs captured and recaptured will be used to generate a linear model to predict pre-molt sizes based on post-molt sizes in section 2 of this experimental study.

The focal point for this study was helping these biologists develop a method for predicting pre-malt sizes based on post-molt sizes. This will allow biologists to place restrictions on the fishing of crabs under a certain size to maintain a steady population of crabs during the fishing season. Using Mathemat ica, the mean, median, standard deviation, skewness, kurtosis were found for each set of pre-molt and post-molt sizes (all crabs, crabs examined in a laboratory, and crabs captured and recaptured) along with residuals of these data sets. These characteristics of each distribution were used for comparison of the pre-molt and post-molt sizes and the effect of different environments on the growth of crabs.

The mean of the data sets will provide an average pre-molt size and post-molt size which can be used for general comparison amongst the different populations of crabs being investigated. The median is used when comparing both distributions as whatever lies below the median will be 50% of the data and whatever lies above the median will also be 50% of the data. The standard deviation will indicate if the pre-molt and post-molt sizes are closer to the mean or if the data was spread out.
The skewness and the kurtosis are used in this statistical study to determine how the distribution deviates from a normal distribution. The skewness will be useful when dividing the data into clean shells and fouled shells. Here, the skewness can be used as an indicator for inaccurate measurements as there will likely be more outliers, hence a larger (or more negative) skewness. The kurtosis will be of the utmost importance when examining for the residuals For a visual representation of the post-molt sizes as well as the corresponding pre-molt sizes, histograms were used.

While it is important to understand the descriptive characteristics of the pre-molt and post-molt sizes, the descriptive statistics of the residuals are even more important. The residuals are the vertical distance from the best fit line to the actual data point (post-molt size, pre-molt size). A visualization of the vertical distance from the actual data point to the line of best fit is created. This will help statisticians understand the scedasticity of the residuals and further analyze where the linear model is no longer an appropriate method of prediction. Using a theorem known as least squares regression, a linear model can be generated by minimizing the sum of the squares of the residuals which will be used to obtain predictions for the pre-molt size based on the post-molt size. In order the assume the linear model is accurate, there is a certain extent to which the residuals can vary from a normal distribution. Using histograms and smooth histograms, the residuals are plotted and compared with a normal distribution curve. Using these figures, a visual understanding of how the residuals vary from a normal distribution can be generated. The descriptive statistics of the residuals will be used to understand the traits of the residuals and determine where there would be the most amount of error. The kurtosis and skewness of the residual will be used to understand where most of the error lies and how much the residuals deviate from a normal distribution. To further investigate the results of the kurtosis and skewness and to understand the residuals, the percentage of data within the one standard deviation of mean is calculated and compared to a normal distribution. This will provide the answer to whether the distribution has longer tails or a higher peak based on the kurtosis. Quantile plots will be used to understand where the linear model is no longer an accurate prediction for the pre-molt sizes. By comparing the variation from the normal quantile line, a conclusion can be made regarding how closely a certain distribution models a normal distribution. The residuals will be compared to this normal line to conclude whether the linear model is an appropriate tool for prediction.

Using the linear model and given post-molt sizes, Mathematica's Map function was used to find the pre-molt sizes for each post-molt size given in the second set of data. The linear model will be used as a prediction technique for the pre-molt size using an input of the post-molt size. Then, to further understand the relation between the pre-molt and post-molt sizes, the ListPlot tool plotted all the resulting pre-molt sizes based on the post-molt sizes. Tools for finding the descriptive statistics of the results

were again used to analyze the relation between post-molt size and the predicted pre-molt size. Just as before, histograms were used as a visual tool to understand the results given from the descriptive statistics. To further investigate the predicted pre-molt sizes and determine how fouled measurements affected the distribution, the data is split into 2 populations: clean shells and fouled shells. The goal was to determine if the fouled shells lead to a misinterpretation of the post-molt sizes, that is was there more variation amongst post-molt sizes for fouled shells. By using descriptive statistics, as well as histograms for visualization, a definitive conclusion can be made whether fouled shells corrupt the data and lead to inaccurate predictions of pre-molt size. Lastly, the line of averages is presented as an alternative method for generating a linear model. This method is introduced as it may be more stable than a least-squares regression line as outliers do not have a significant effect on the mean of a subinter val.

Results

Section 1

Data for Entire Population of Crabs

The population of crabs being examined in this statistical study will be used to develop a method for pre-molt prediction based on post-molt sizes recorded. The least-squares regression line will minimize the summation of the square of the residuals to obtain predictions of pre-molt size based on post-molt size with little error. The average pre-molt size of the first set of data is roughly 129 mm while the average corresponding post-molt size is approximately 143 mm. The standard deviation is used to measure the dispersion of the data points surrounding the mean. A larger standard deviation corresponds to a larger variation from the mean. The skewness is used to measure the symmetry of the distribution and is used in tandem with the kurtosis to compare distributions to a normal distribution. Both sets of data are negatively skewed as shown by Figures 2 & 3. The kurtosis measures the fatness of the tails and the peakiness of the distribution. A shown in Table 1, the kurtosis for both pre-molt and post-molt are very large indicating these sets of data vary from a normal distribution. However, the residuals of this data set have a large kurtosis, indicating this linear model has a large number of residuals close to 0, that is 0 error between the approximation and the actual value.

Figure 1: Post-Molt Sizes with Corresponding Pre-molt Sizes for all Entries(Blue)
and the Least Squares Regression Line (Red)
-25.2137+1.07316 x (equation 1)

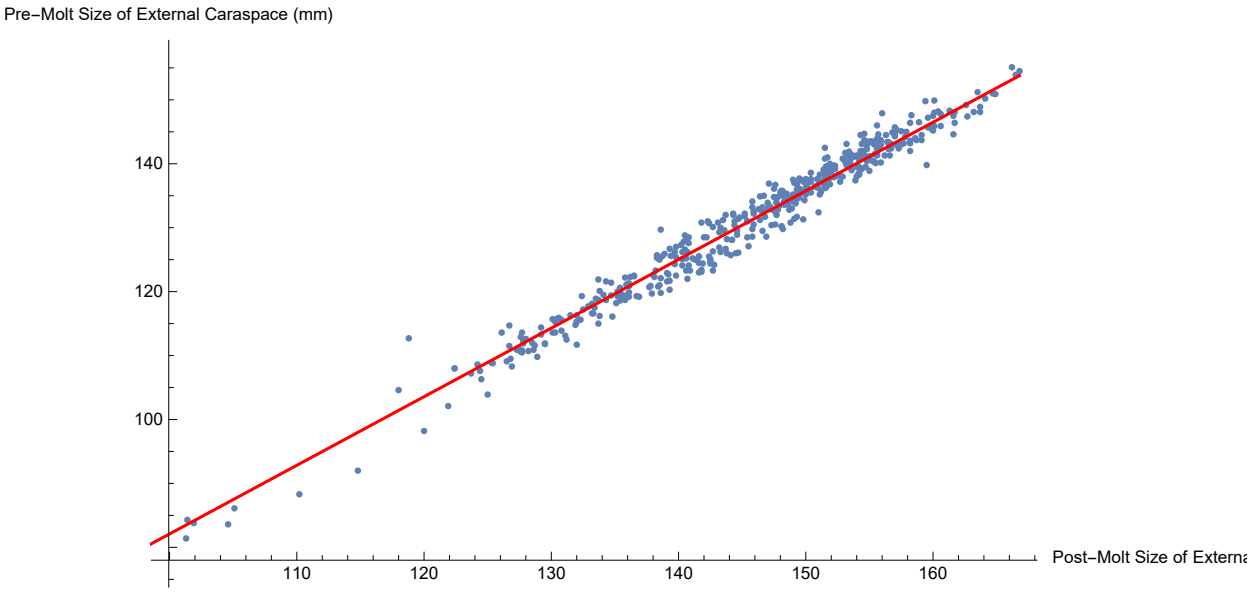


Table 1: Descriptive Statistics of Pre-Molt and Post-Molt sizes for all Crabs Being Examined

Out[•]=

	Pre–Molt Size	Post–Molt Size
Mean	129.21	143.90
Median	132.8	147.4
Standard Deviation	15.86	14.64
Skewness	–2.01	–2.35
Kurtosis	9.77	13.12

Figure 2: Histogram of Pre-Molt sizes for all Crabs Being Examined

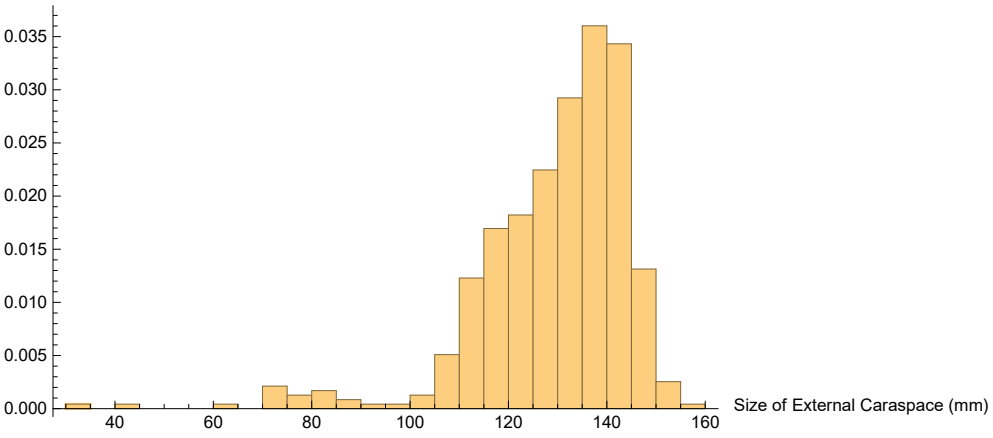
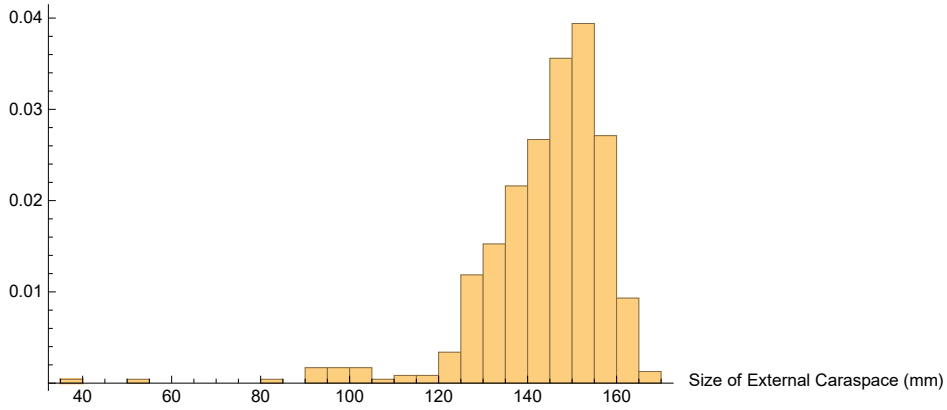


Figure 3: Histogram of Post-Malt sizes for all Crabs Being Examined



R-Squared value of 0.980833

Figure 4: Plot of Residuals for all Crabs Being Examined

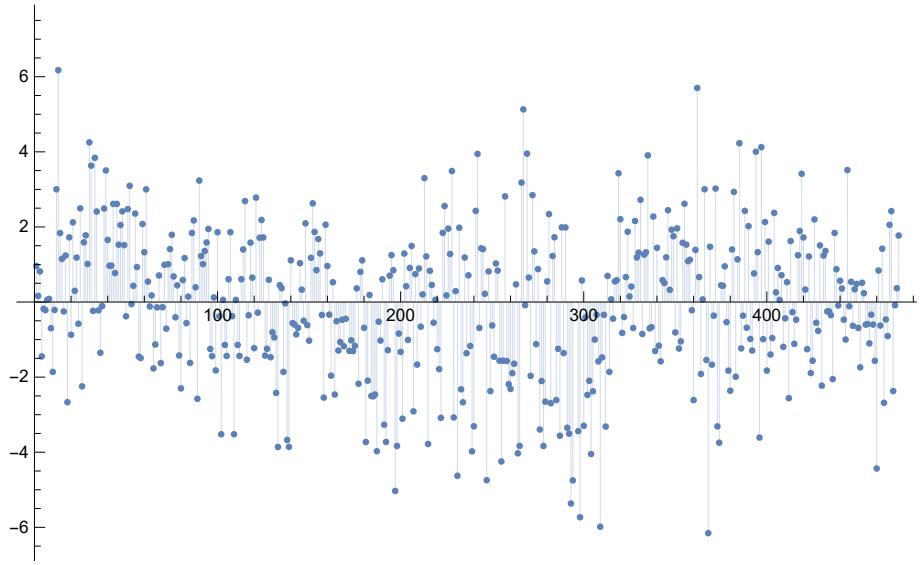


Figure 5: Histogram of Residuals for all Crabs Being Examined
in Comparison with a Normal Distribution of the Residuals

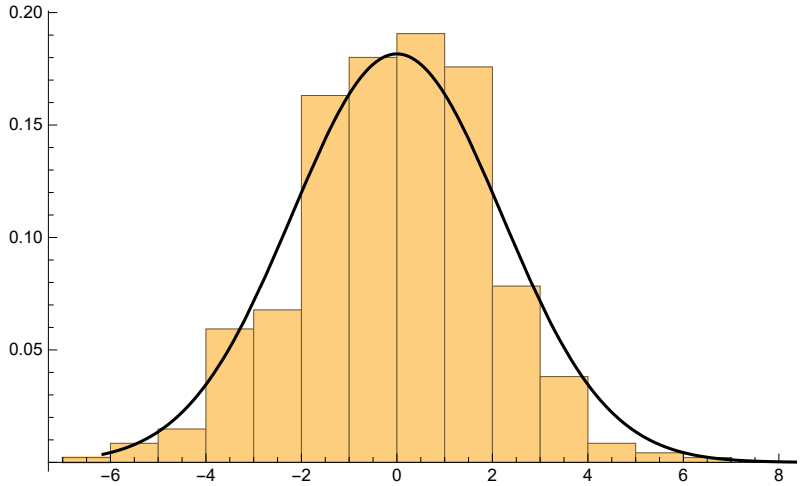


Figure 6: Smooth Histogram of Residuals for Pre-Molt and Post-Molt Sizes
for all Crabs Examined with a Normal Distribution of the Residuals

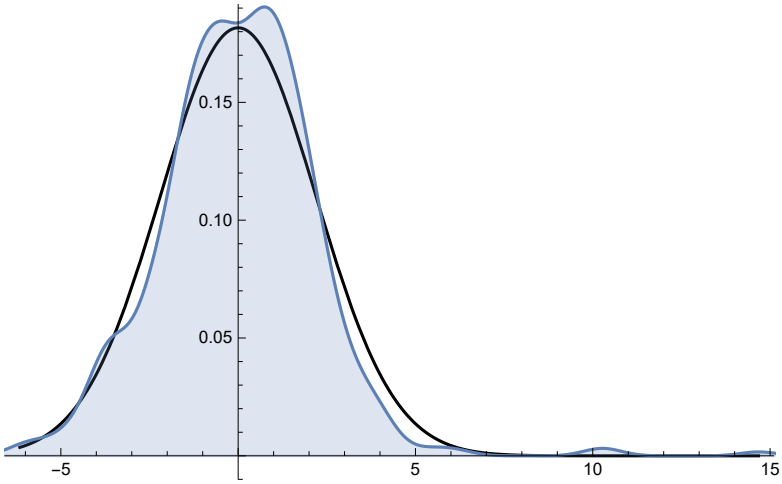


Table 2: Descriptive Statistics of Residuals for Entire Crab Population

Out[•]=

	Residuals for Pre−Molt and Post−Molt Data of all Crabs Being Examined
Mean	2.86E−15
Median	0.056
Standard Deviation	2.196
Skewness	0.845
Kurtosis	8.379

Figure 7: Quantile Plots of Residuals for Entire Crab Population

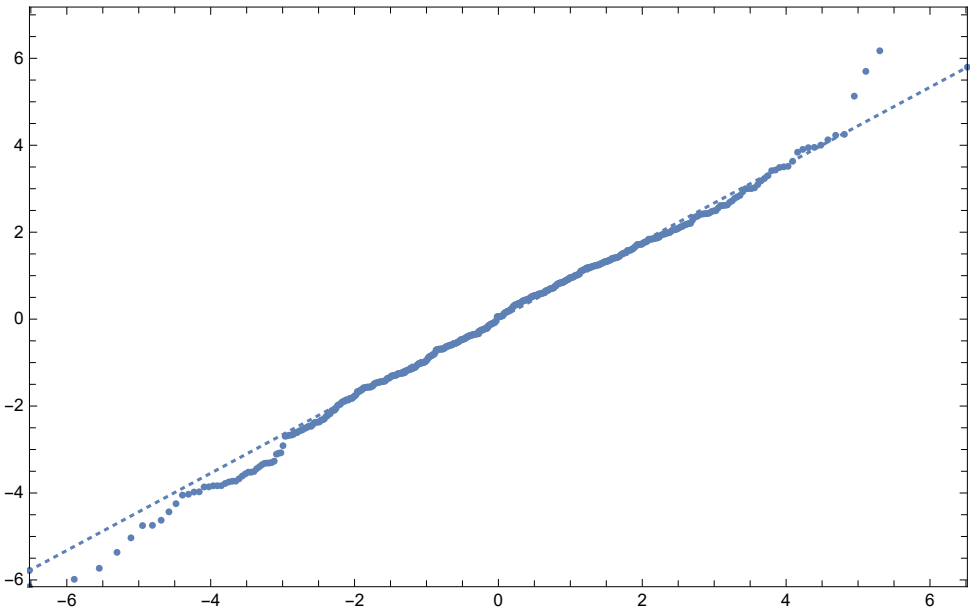


Table 3: Comparison of Residuals for Entire Crab Populations with a Normal Distribution

Out[ ]=

	Residuals	Normal Distribution
% of Cases in between 2 Standard Deviations of the Mean	0.9688	0.9545
% of Cases in between 1 Standard Deviations of the Mean	0.7458	0.6827

Data From Crabs Examined in a Laboratory

Next, the data is split into 2 populations, crabs raised and examined in a laboratory and crabs that are captured, measured, released, then recaptured again. In this section, the crabs examined in the laboratory are of interest. A line of best fit is desired to make predictions based on the pre-molt and post-molt size correlation found from laboratory results. Equation 2 can be used as the linear model for the prediction of pre-malt sizes based on post-molt sizes. The distribution of this data is similar to that of the previous case, however, there is a slightly higher kurtosis for the crab sizes when raised in a laboratory, this is a slight cause for concern when using a linear model. It is important to understand the residuals are not a normal distribution, but the residua Table 6 will indicate whether the linear model minimizes the residuals, meaning the kurtosis of the residuals should be large. The question becomes do the residuals deviate from a normal distribution in a way that would harm the accuracy of the linear model.

Figure 8: Post-Molt Sizes with Corresponding Pre-molt Sizes for Crabs Examined in a Laboratory (Blue) and the Least Squares Regression Line (Red);
- 25.3439+1.07394 x (equation 2)

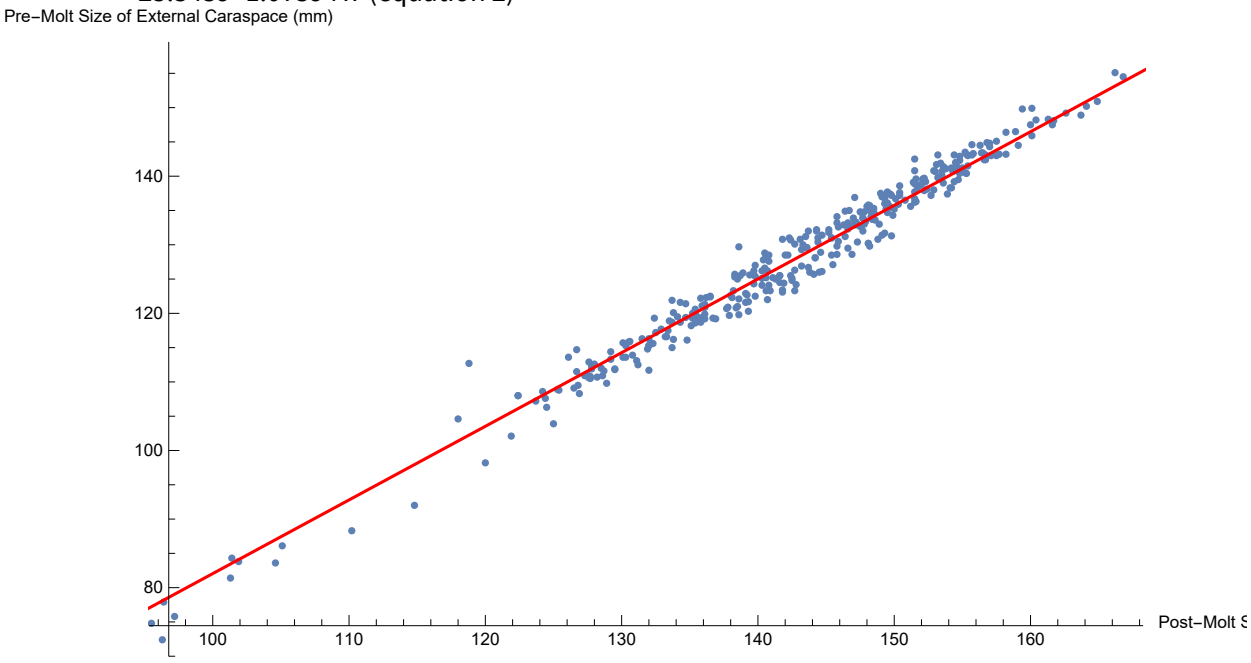


Table 4: Descriptive Statistics of Pre-Molt and Post-Molt

Sizes for Crabs Examined in a Laboratory

Out[•]=

	Pre–Molt Size	Post–Molt Size
Mean	126.20	141.11
Median	128.9	143.7
Standard Deviation	16.57	15.28
Skewness	–1.889	–2.288
Kurtosis	9.02	12.44

Figure 9: Histogram of Pre-Molt sizes for Crabs Examined in a Laboratory

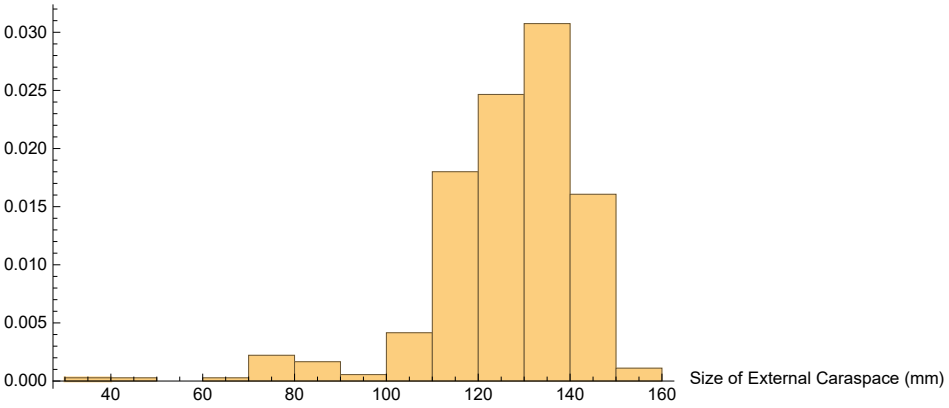
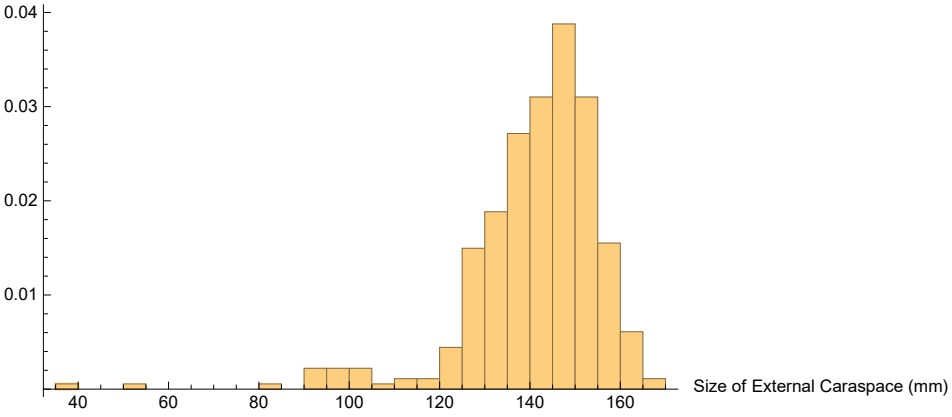


Figure 10: Histogram of Pre-Molt sizes for Crabs Examined in a Laboratory



R-Squared value of 0.980999

Figure 11: Plot of Residuals for Crabs Examined in a Laboratory

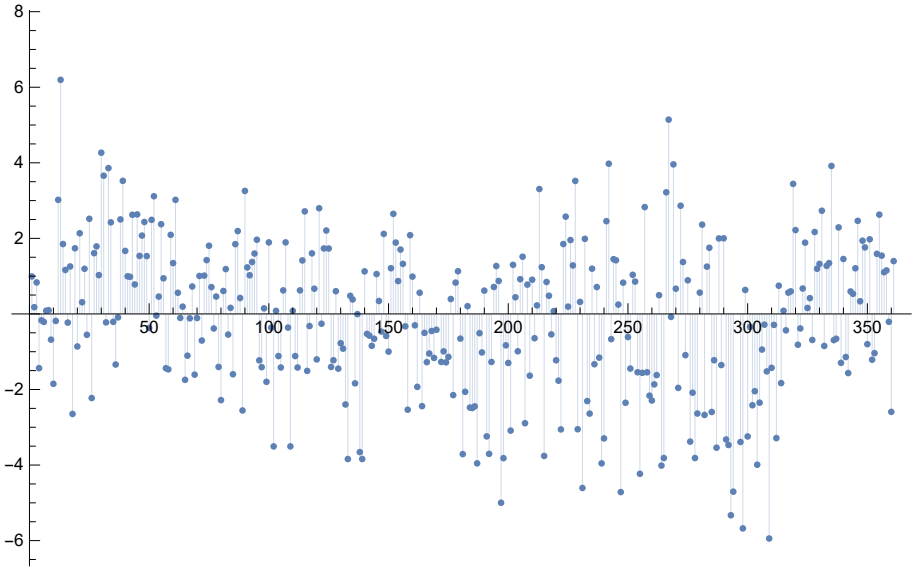


Figure 12: Histogram of Residuals for Crabs Examined in a Laboratory in Comparison with a Normal Distribution of the Residuals

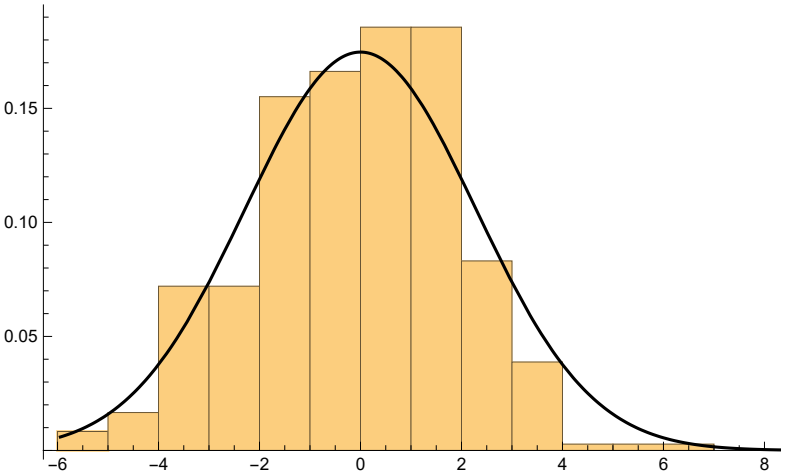


Figure 13: : Smooth Histogram of Residuals for Pre-Molt and Post-Molt Sizes for Crabs Examined in a Laboratory Overlayed with a Normal Distribution of the Residuals

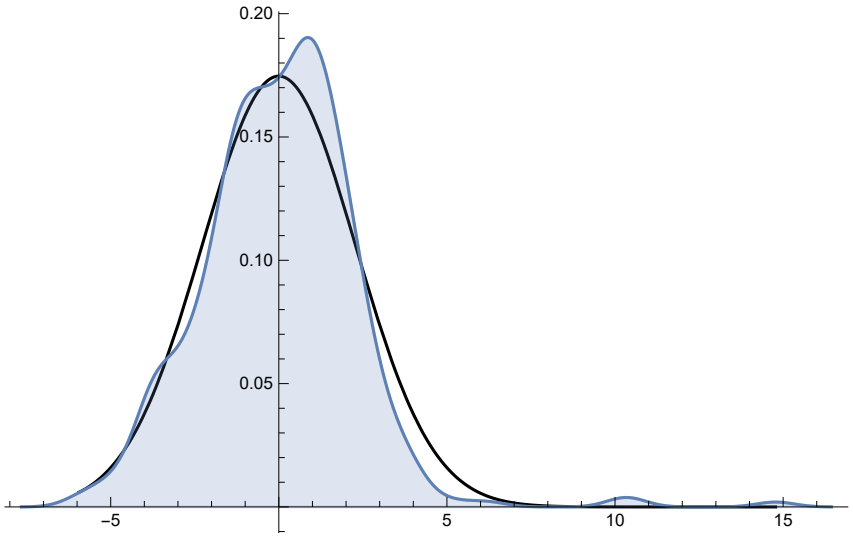


Table 5: Descriptive Statistics of Residuals for Crab Population Examined in a Laboratory

Out[•]=

	Residuals for Pre−Molt and Post−Molt Data of Crabs Being Examined in a Laboratory
Mean	4.787E−14
Median	0.0835
Standard Deviation	2.284
Skewness	1.012
Kurtosis	9.001

Figure 14: Quantile Plots of Residuals for Crab Population Examined in a Laboratory

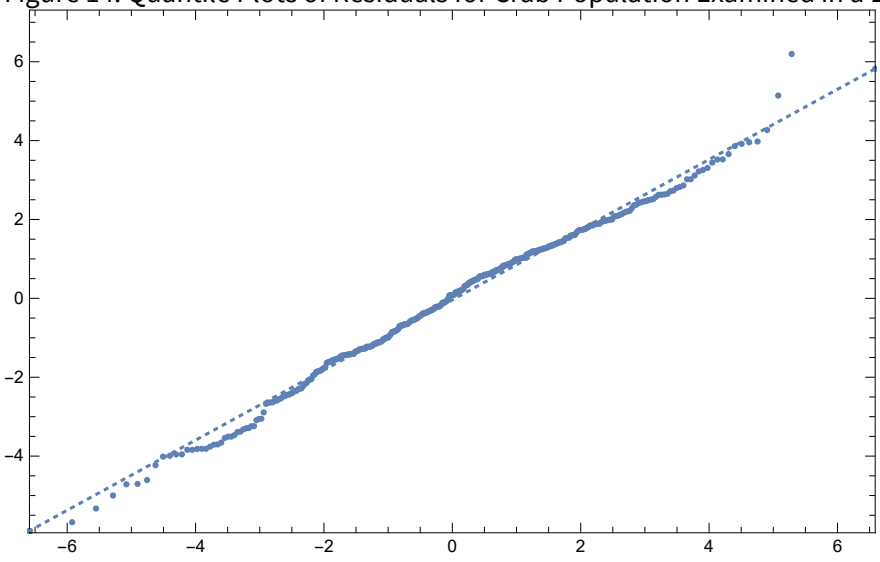


Table 6: Comparison of Residuals for Crab Population Examined in a Laboratory with a Normal Distribution

Out[•]=

	Residuals of Laboratory Data	Normal Distribution
% of Cases in between 2 Standard Deviations of the Mean	0.9668	0.9545
% of Cases in between 1 Standard Deviations of the Mean	0.7396	0.6827

Data from Crabs Captured-Recaptured

Just as the populations of crabs examined in laboratories are interpreted, so must the crabs that are captured and recaptured. Equation 3 is the linear model for this set of data, minimizing the sum of the square of the residuals. Pre-molt and post-molt sizes for this set of data are closer to a normal distribution than previous results. The skewness and kurtosis of the residuals are a closer model to a normal distribution than in the previous case, however, there is still large peakiness around the mean of the residuals. Using the residuals, the linear model can be verified as an appropriate prediction method.

Figure 15: Post-Molt Sizes with Corresponding Pre-Molt Sizes for Crabs Captured-Recaptured (Blue) and the Least Squares Regression Line (Red)
-20.4016+1.04215 x (Equation 3)

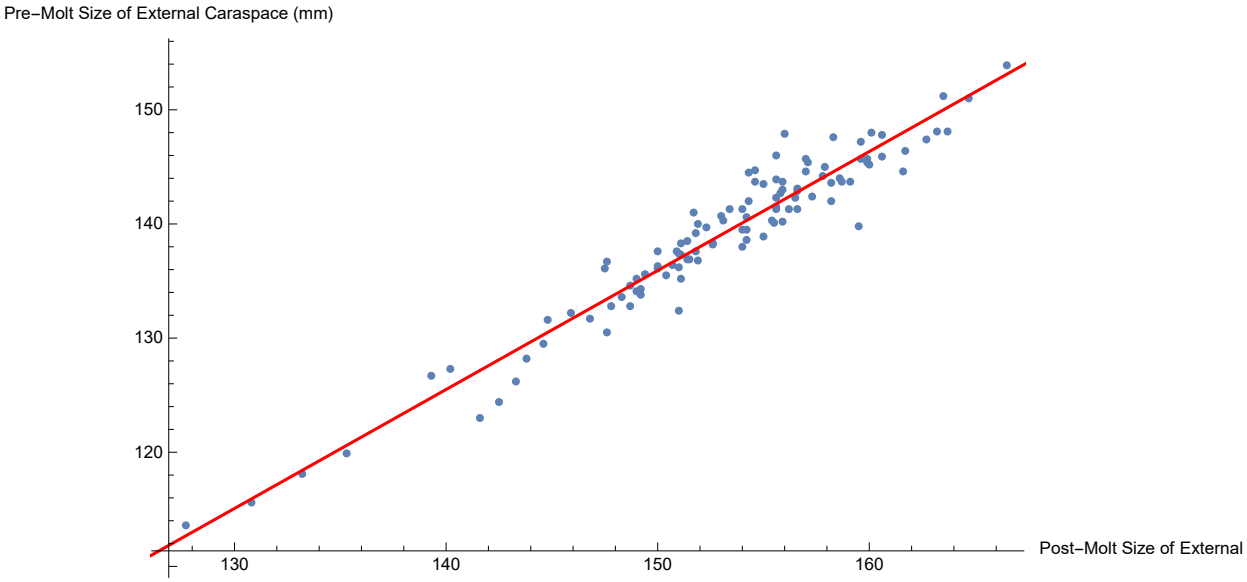


Table 7: Descriptive Statistics of Pre-Molt and Post-Molt Sizes for Crabs Captured and Re-captured

Out[•]=

	Pre–Molt Size	Post–Molt Size
Mean	139.01	152.96
Median	140.1	154
Standard Deviation	7.25	6.72
Skewness	–1.11	–1.12
Kurtosis	4.76	5.24

Figure 16: Histogram of Pre-Molt Sizes for Crabs Captured and Re-captured

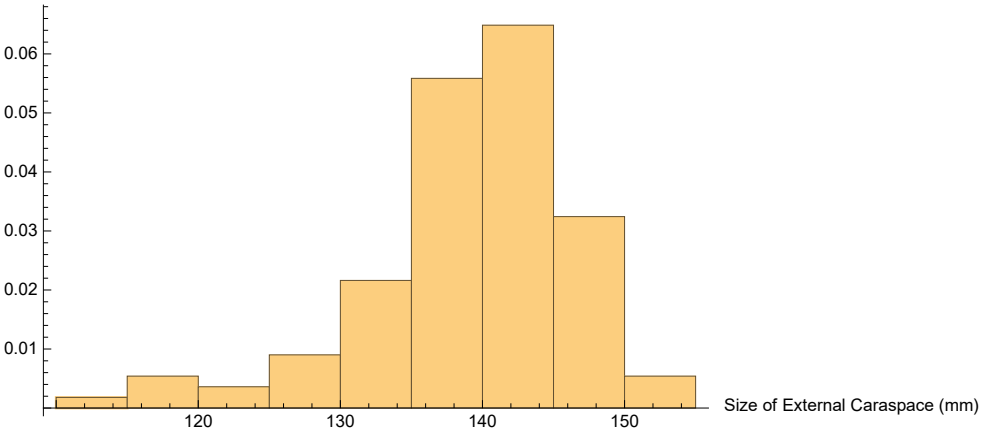
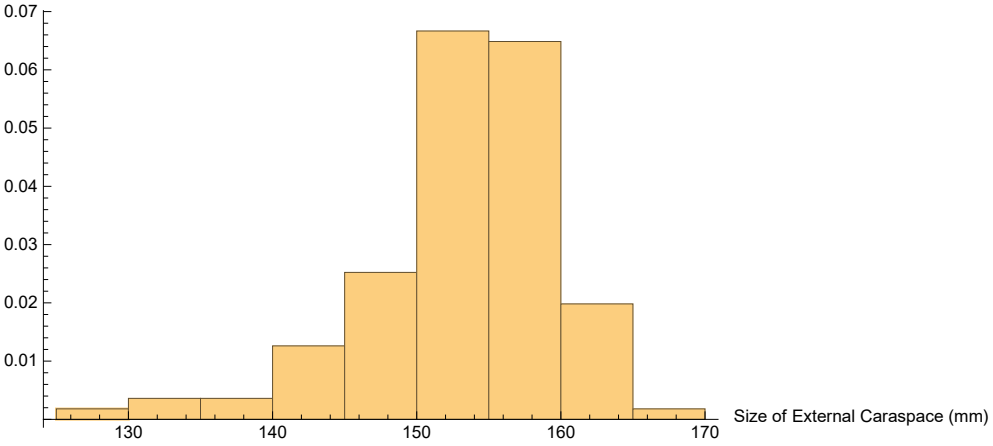


Figure 17: Histogram of Post-Molt Sizes for Crabs Captured and Re-captured



R-Squared value of 0.932775

Figure 18: Plot of Residuals for Crabs Captured and Re-captured

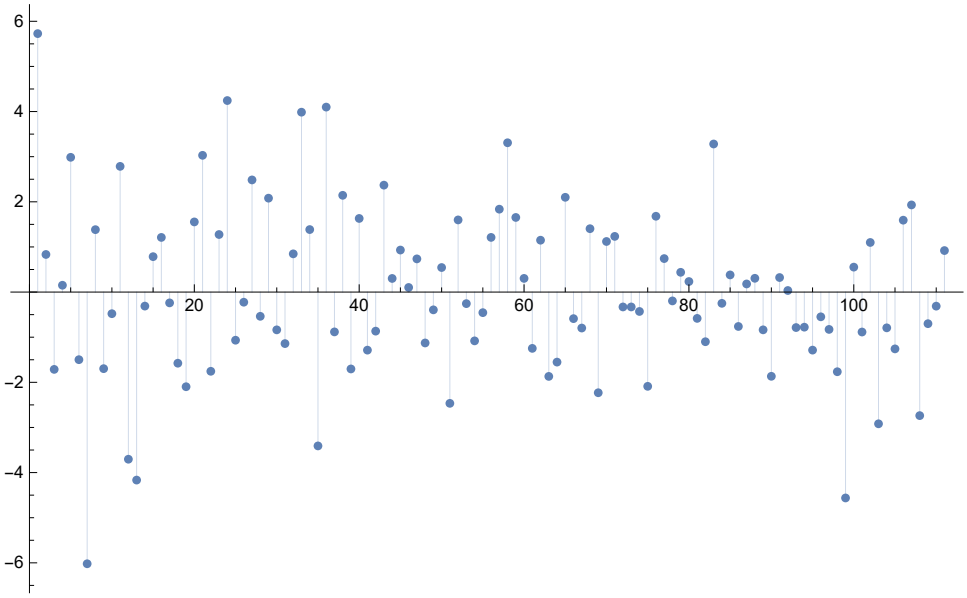


Figure 19: Histogram of Residuals for Crabs Captured and Re-captured in Comparison with a Normal Distribution of the Residuals

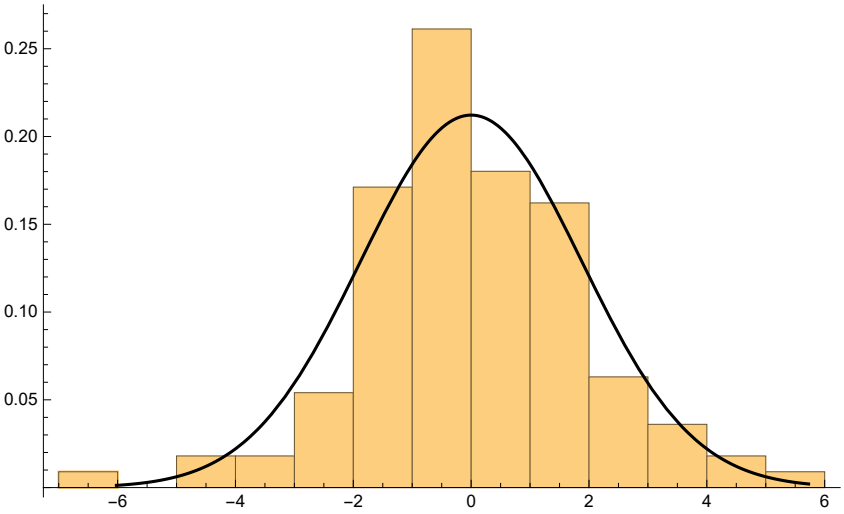


Figure 20: Smooth Histogram of Residuals for Pre-Molt and Post-Molt Sizes for Crabs Captured and Re-captured Overlayed with a Normal Distribution of the Residuals

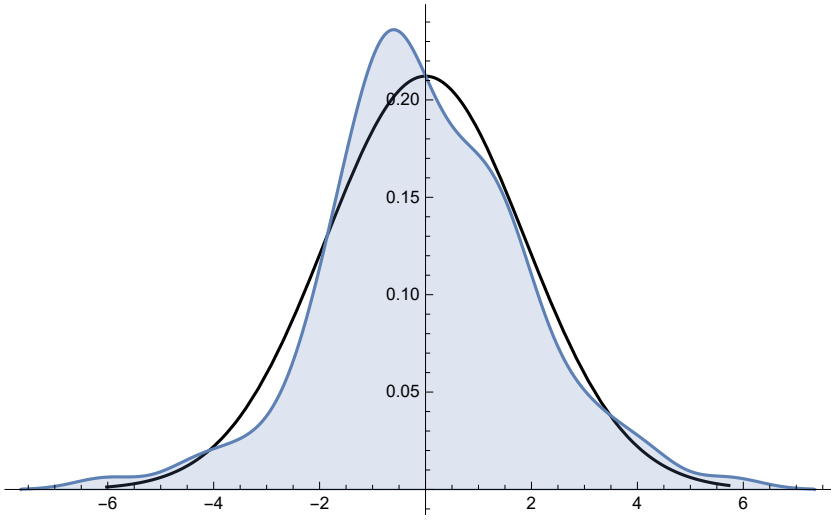


Table 8: Descriptive Statistics of Residuals for Crab Population Captured and Re-captured

Out[ ]=

	Residuals for Pre−Molt and Post−Molt Data of Crabs Captured and Re−captured
Mean	2.04E−14
Median	−0.249
Standard Deviation	1.88
Skewness	0.0355
Kurtosis	3.913

Figure 21 : Quantile Plots of Residuals for Crab Population Captured and Re-captured

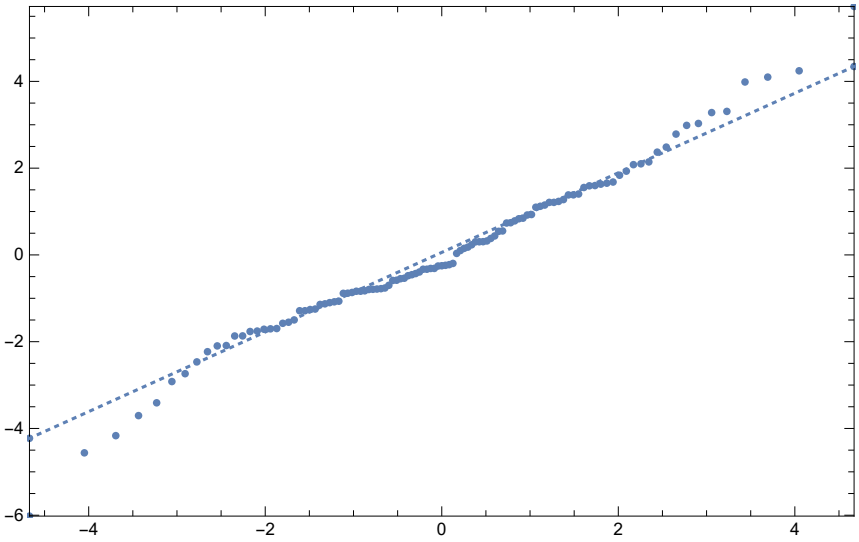


Table 9: Comparison of Residuals for Crab Population Captured and Re-captured with a Normal Distribu tion

Out[•]=

	Residuals of Captured–Recaptured Data	Normal Distribution
% of Cases in between 2 Standard Deviations of the Mean	0.9369	0.9545
% of Cases in between 1 Standard Deviations of the Mean	0.7658	0.6827

Section 2

Predicting Pre-Molt Sizes From Post-Molt Sizes

Using the linear model generated above, the pre-malt sizes can now be predicted based on the given post-molt sizes of this section. The predicted pre-malt sizes plotted with the corresponding post-molt sizes will lead to a linear correlation between these results. Besides, there may be some error in the measurements which effect the linear model. To determine the possible error due to measurements, the post-molt values from section 2 are split into fouled shells and clean shells. Using the linear model found from section 1, the predicted pre-molt sizes were found. Using descriptive statistics and analyzing the least squares line, a conclusion can be made regarding the effect of fouling on inaccurate calculations leading to inaccurate predictions.

Table 10: Descriptive Statistics for Predicted Pre-Molt Sizes Based on Recorded Post-Molt Sizes

Out[•]=

	Predicted Pre–Molt Sizes Based on Recorded Post–Molt Sizes	Corresponding Post–Molt Sizes
Mean	130.64	145.23
Median	132.60	147.05
Standard Deviation	12.72	11.85
Skewness	–0.763	–0.763
Kurtosis	3.74	3.74

Figure 22: Predicted Pre-Molt Sizes Based on Recorded Post-Molt Sizes

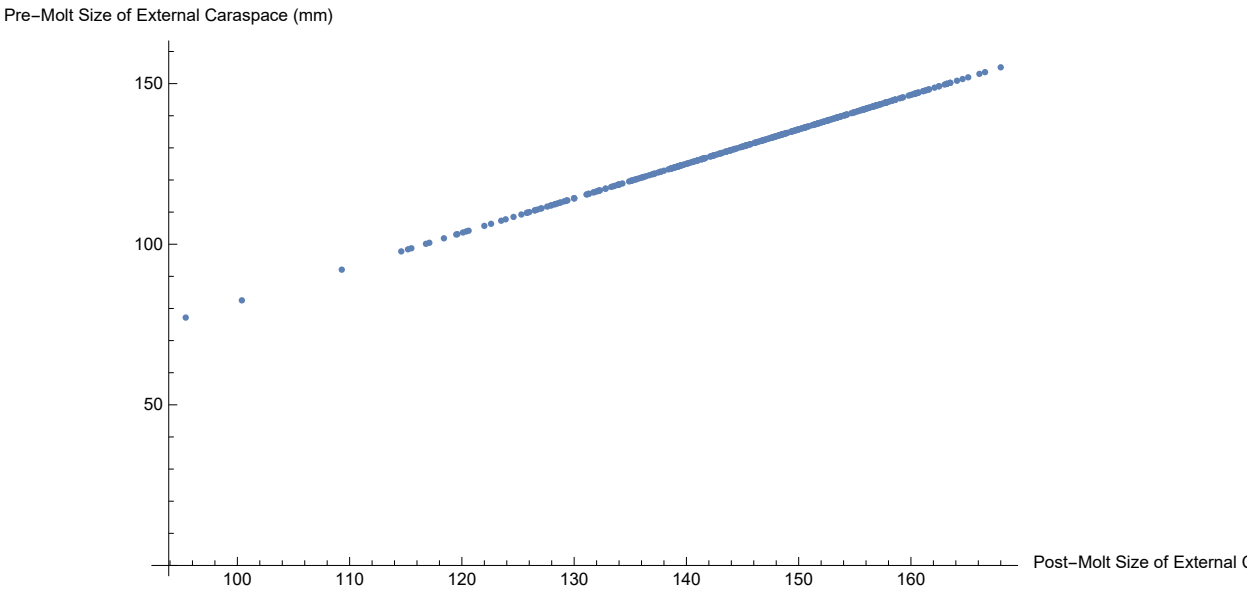


Figure 23: Histograms of Predicted Pre-Molt Sizes (mm)

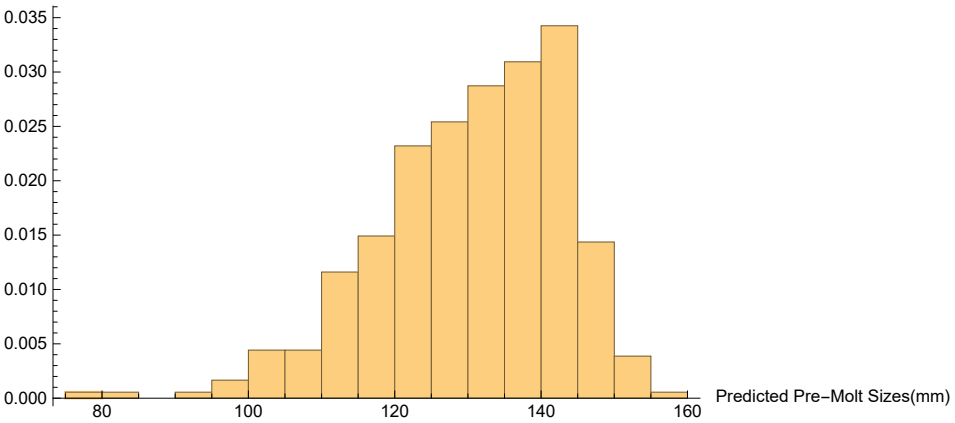


Figure 24: Smooth Histogram of Predicted Pre-Molt Sizes

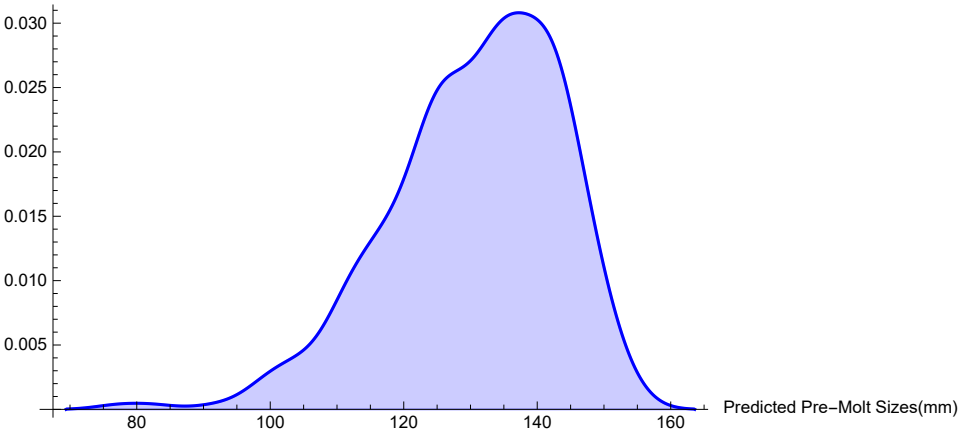
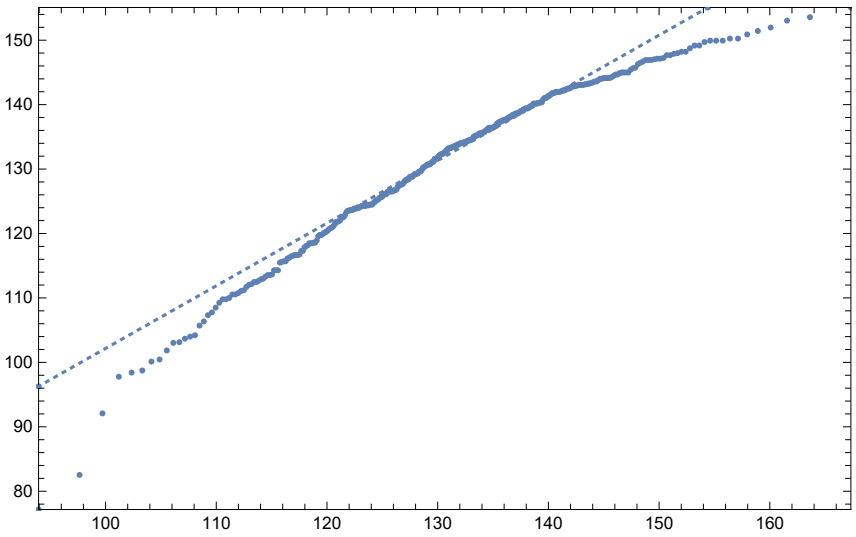


Figure 25: Quantile Plot for Predicted Pre-Molt Sizes



Clean Shell Predicted Pre-Molt Sizes From Post-Molt Sizes

Table 11: Descriptive Statistics for Predicted Pre-Molt Sizes Based on Recorded Post-Molt Sizes for Clean Shells

Out[•]=

	Predicted Pre−Molt Sizes Based on Recorded Post−Molt Sizes for Clean Shells	Corresponding Post−Molt Sizes
Mean	127.30	142.11
Median	125.67	140.6
Standard Deviation	12.23	11.40
Skewness	0.041	0.041
Kurtosis	2.18	2.18

Figure 26: Predicted Pre-Molt Sizes Based on Recorded Post-Molt Sizes for Clean Shells

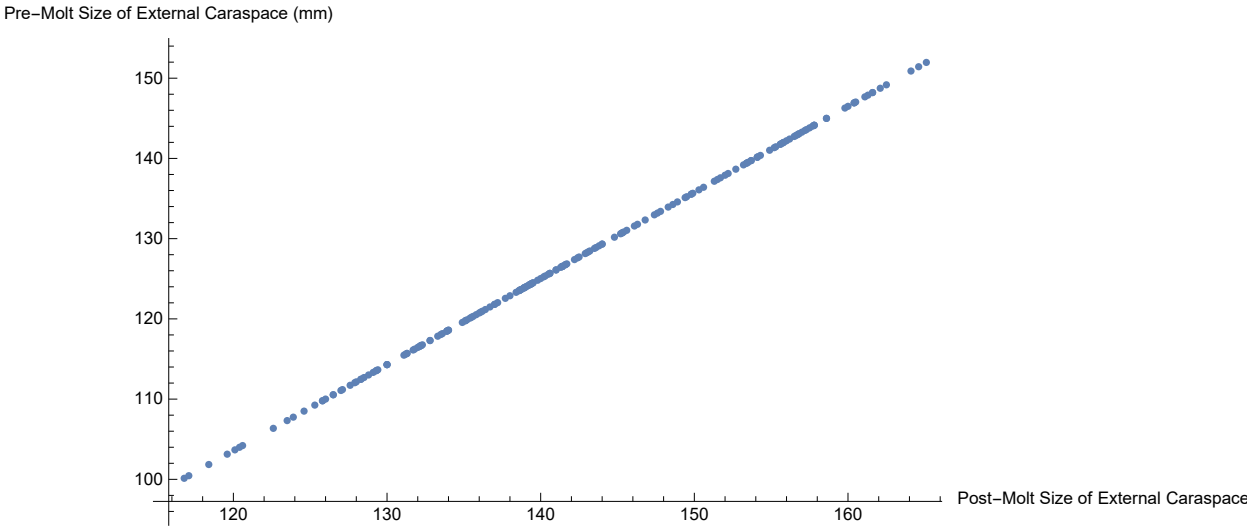


Figure 27: Histograms of Predicted Pre-Molt Sizes (mm) for Clean Shells

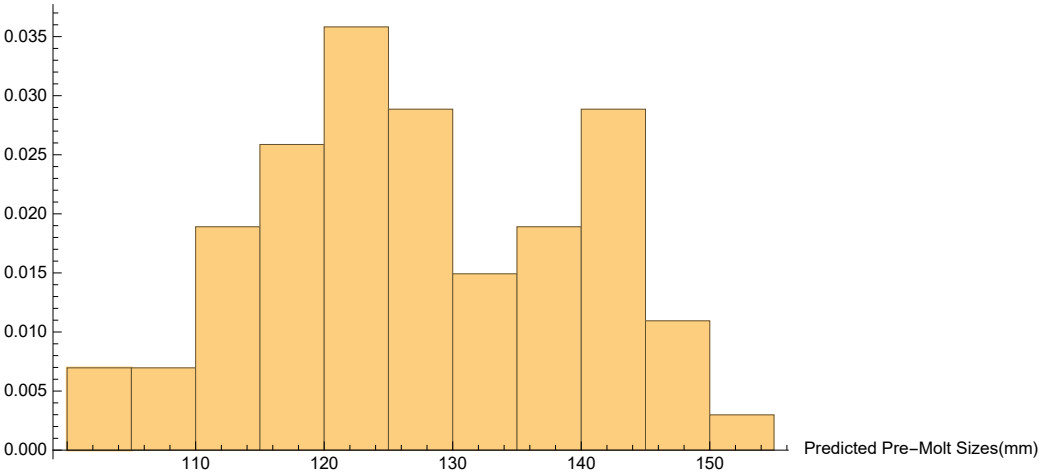


Figure 28: Smooth Histogram of Predicted Pre-Molt Sizes for Clean Shells

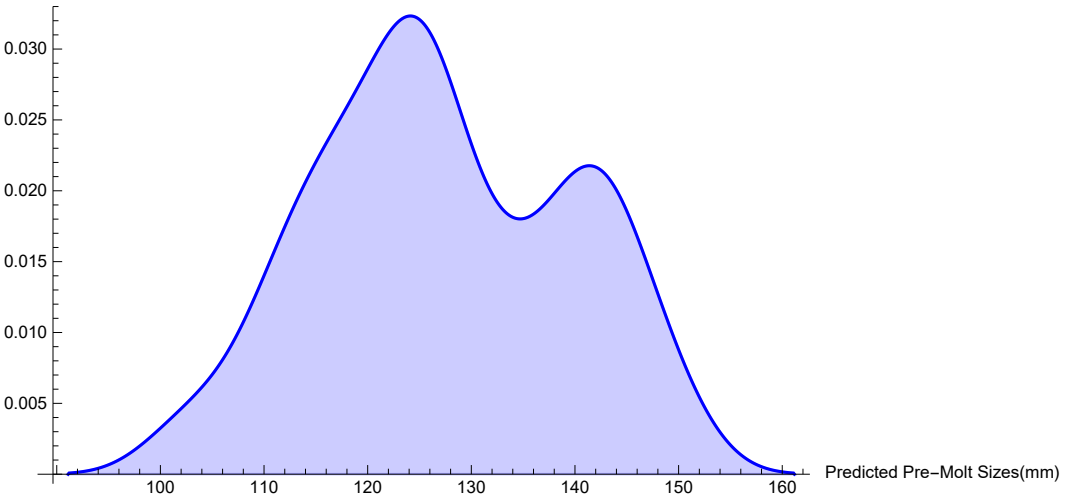
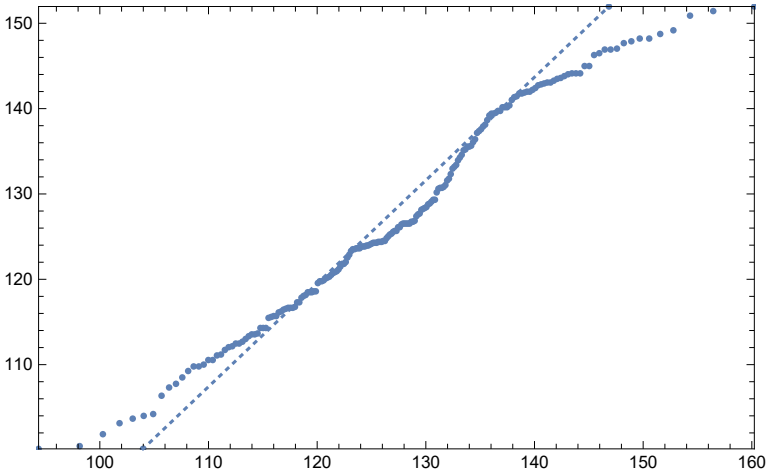


Figure 29: Quantile Plot for Predicted Pre-Molt Sizes for Clean Shells



Fouled Shells Predicted Pre-Molt Sizes From Post-Molt Sizes

Table 12: Descriptive Statistics for Predicted Pre-Molt Sizes Based on Recorded Post-Molt Sizes for

Fouled Shells

Out[ ]=

	Predicted Pre–Molt Sizes Based on Recorded Post–Molt Sizes for Fouled Shells	Corresponding Post–Molt Sizes
Mean	134.81	149.11
Median	136.41	150.6
Standard Deviation	12.09	11.27
Skewness	–2.046	–2.046
Kurtosis	9.059	9.059

Figure 30: Predicted Pre-Molt Sizes Based on Recorded Post-Molt Sizes for Fouled Shells

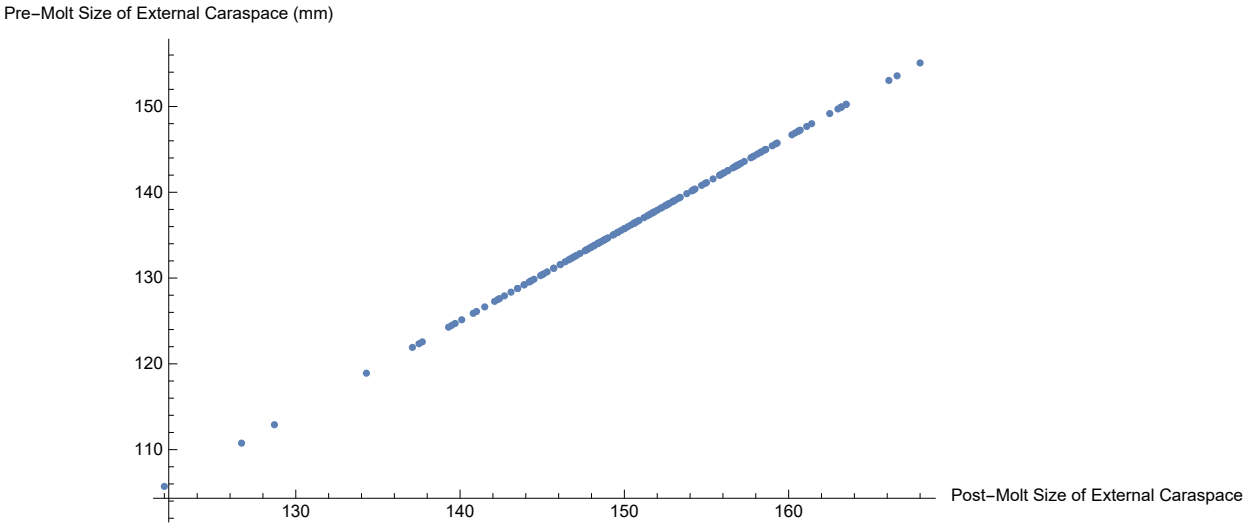


Figure 31: Histograms of Predicted Pre-Molt Sizes (mm) for Fouled Shells

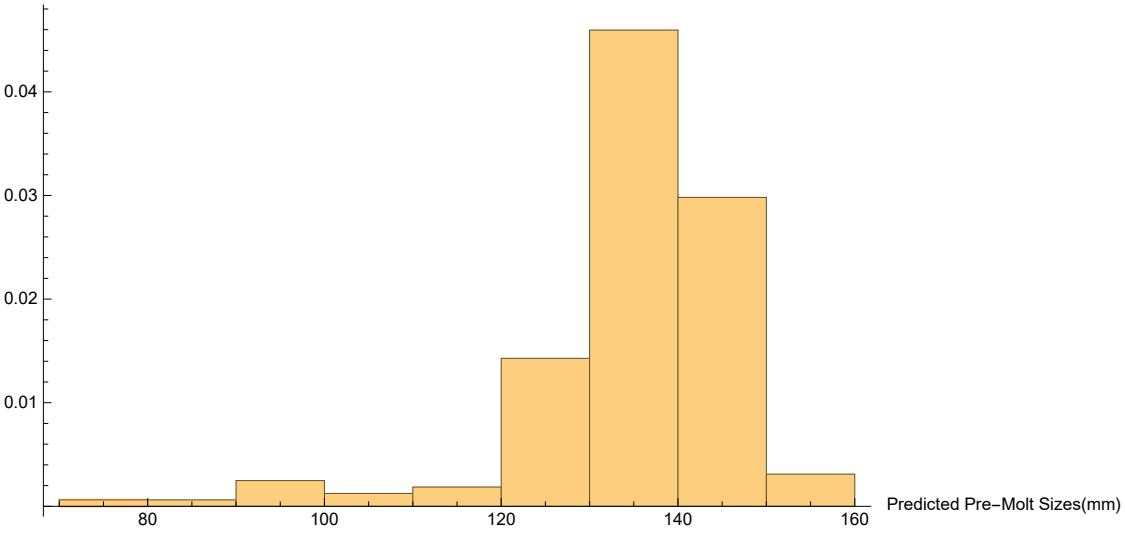


Figure 32: Smooth Histogram of Predicted Pre-Molt Sizes for Fouled Shells

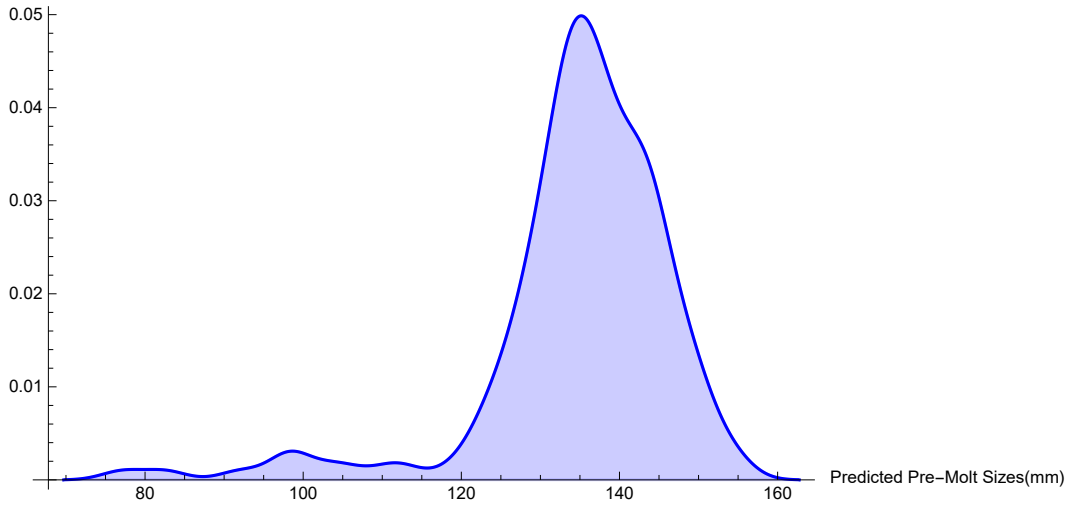
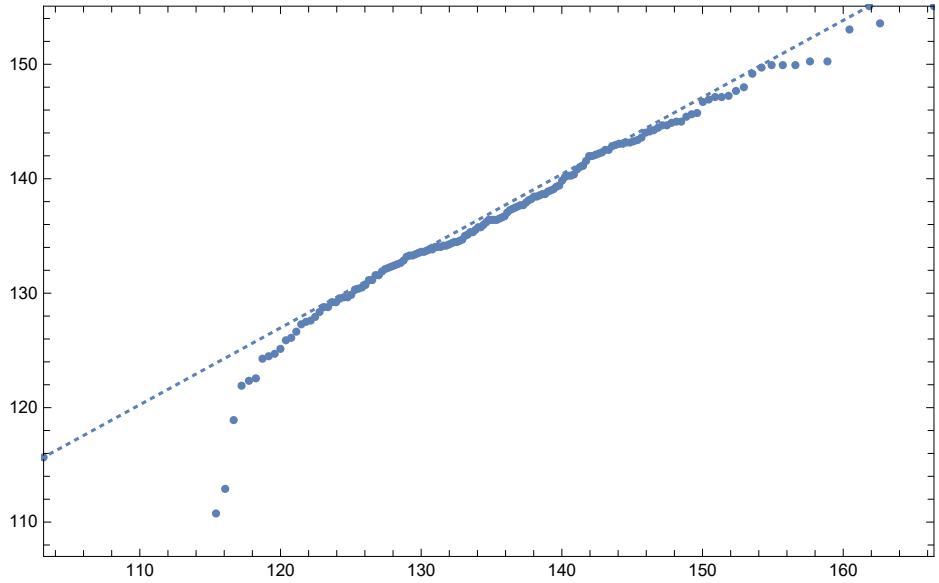


Figure 33: Quantile Plot for Predicted Pre-Molt Sizes for Fouled Shells



Line of Averages

Figure 34: Bundles of Averages for Post-Molt Data

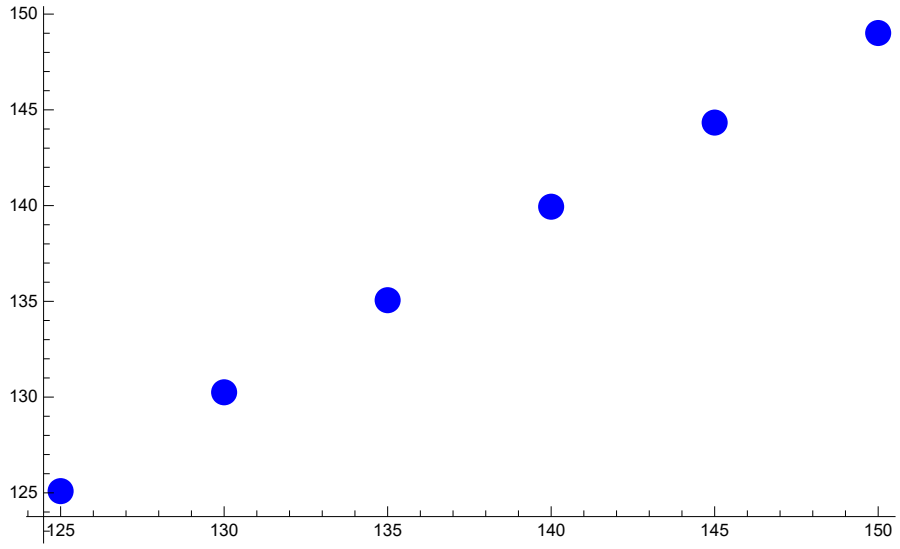
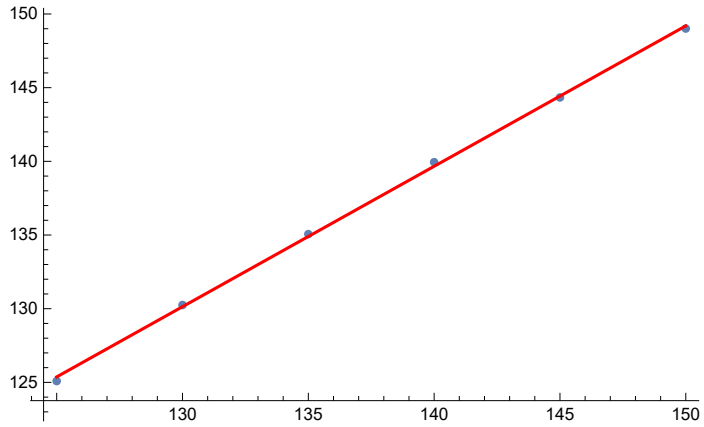


Figure 35: Line of Averages for Predicting Pre-Molt Sizes from Post-Molt Sizes



Discussion and Conclusion

Biologists are interested in studying the growth patterns of crabs to place proper restrictions on the fishing of crabs. These biologists are in need of a model to study the growth patterns of these crabs, which can be generated by using current data for pre-molt sizes and post-molt sizes of crabs that are either in a laboratory or have been captured and recaptured. However, for biologists to accurately report estimations of pre-malt sizes based on post-molt sizes, an accurate linear model is needed. The linear model is found by minimizing the square of the residuals, essentially minimizing the square of the error of each data from the predicated value given by the linear model to the actual recorded value. The model of least squares (red line) is plotted with the data points (Blue) in Figures 1,8, and 15. As can be seen, the line of best fit seems to minimize the distance from each data point to the line of best fit. Using the linear model generated by the post-molt and pre-molt data given, the post-molt sizes in section 2 of the results can be used to find the pre-molt sizes of crabs.

Considering all crabs evaluated in this study, regardless of habitat, both the pre-molt sizes and post-

molt sizes varied from a normal distribution. Both pre-molt sizes and post-molt sizes have a smaller mean than the median, meaning it is assumed the majority of data will lie above the mean. This is also referred to as the skewness. As can be seen in Table 1, the skewness of both pre-molt and post-molt sizes is negative; it can be seen in the histograms of these data sets (Figure 2 & Figure 3) that have outliers that lie far below the mean which contributes to the negative skewness. This is common amongst crab shells, as fouling often occurs which leads to inaccurate measurements of the carapace. In addition to the large skewness, fouling also contributes to the large kurtosis of the distribution of pre-molt and post-molt sizes. In these distributions, the tails of the post-molt size and pre-molt size are especially long and have a rather large peak as well. This can be seen clearly in the histogram plot of these distributions (Figure 2 & Figure 3) which corresponds with the large Kurtosis of the distributions shown in Table 2. This preliminary data for pre-molt sizes and post-molt sizes will be used to make predictions for the pre-molt sizes based on given post-molt sizes. A numerical indication for how the predicted linear model fits the data(Figure 1) can be demonstrated using the value of r-squared. When considering the entire crab population being investigated, the r-squared value is 0.9808, meaning the linear model is close to an accurate representation of the actual pre-molt size. While this r-squared value would be groundbreaking in psychology studies, usually for physics and other sciences, r squared can reach higher values, meaning more accurate models can be generated. As mentioned, the line of least squares is generated by minimizing the square of the residuals; obviously, the smaller the error the more accurate the linear model. These errors, distance from the predicted value to the actual value, are known as the residuals. These residuals are plotted in Figure 4 along with the distance from the line of least squares (vertical axis). Since there is not a fanned-out shape at either end of the residual plot, the residuals are said to be homoscedastic, meaning the residuals have no significant variation towards one end or another. This scedasticity is a helpful way to determine where the linear model would fail and no longer be accurate. For the post-molt and pre-molt sizes of all crabs being evaluated, the residuals seem close to a normal distribution as demonstrated in FIgures 5,6, and 7. The mean of the residuals is approximately 0 as shown in Table 2, while the median indicates 50% of the data lies above and below this value. The mean of the residuals is 0 as that is how a residual is defined; the line of least squares minimizes the distance of the residuals, meaning the summation of the residuals should be approximately 0. It is important to have the residuals approximately represent a normal distribution to validate the accuracy of the prediction of the linear model beyond the given data. However, in the case of pre-molt and post-molt sizes of all crabs, the skewness and kurtosis deviate from a normal distribution by a rather large scaling. The skewness is about 0.8 while the kurtosis is roughly 8.5, while for a normal distribution, the skewness is 0 because the distribution is symmetric and the kurtosis is approximately 3. Checking the quantile plot is a sufficient way of inspecting why the residual distribution varies from that of a normal distribution. According to Figure 7, the residuals seem to follow the linear line representing the normal distribution, except falling off slightly at the tails, although this is no sign of a large deviation from a normal distribution. Instead, it seems as peakiness is the main contributor to the large kurtosis. To verify this assumption is correct, Table 3 was generated. Looking within 2 standard deviations of the mean, the percentage of data within a normal distribution is approximately 95% while for this case, roughly 97% of the residuals lie within 2 standard deviations

of the mean. Investigating further, the percent of data within 1 standard deviation of the mean for a normal distribution is approximately 68%. However, for the residuals of pre-molt sizes and post-molt sizes, roughly 75% of the residuals lie within one standard deviation, resulting in a very peaky distribution. Further concluding, the residuals have fewer data in the tails than does a normal distribution, however, the peakiness leads to a larger kurtosis. While there is some error and the residuals deviate from a normal distribution, the question becomes how far do these residuals deviate from a normal distribution so this linear model can be used for the prediction of pre-molt sizes based on the given post-molt sizes in Section 2. Within one standard deviation of the mean of the residuals, Figures 5 & 6 demonstrate that the linear model is an accurate prediction. Due to the large kurtosis (by the peakiness of the distribution) and Table 3, it is found that roughly 75% of the residuals lie within one standard deviation of the mean. Therefore, since the majority of the residuals lie within one standard deviation of the mean, the majority of the residuals are close to 0. Furthermore, since the majority of the residuals are within one standard of 0 (residuals representing the error of the linear model) the linear model can be justified as an appropriate method for prediction.

Once a linear model was found for the entire crab population, the pre-molt sizes and post-molt sizes were investigated for 2 populations: crabs raised in a laboratory and crabs caught and recaptured seasonally. First, the pre-molt and post-molt sizes for crabs kept in a laboratory were analyzed. The pre-molt size was measured when the crab was first collected and the post-molt measurements were made three to four days after the crab left its old shell to ensure the new shell was fully developed. As Figure 1 was a linear model created by minimizing the summation of the square of the residuals for all crabs, regardless of the habitat, Figure 8 is a visual representation of the pre-molt and post-molt sizes of crabs examined in a laboratory and the corresponding linear of least squares (the line of best fit). As shown in Table 4, the mean is less than the median for pre-molt and post-molt sizes, corresponding to a negative skewness. This can be visualized as shown in Figures 9 & 10 where outliers are lying far below the mean which shifts the mean towards a smaller size for both pre-molt and post-molt cases. Again, the pre-molt and post-molts sizes for crabs examined in a laboratory have a much larger kurtosis than that of a normal distribution kurtosis, meaning these distributions deviate from a normal distribution. A correlation coefficient of 0.9809 is used to represent the accuracy of the linear model generated in FIgure 8. Similar to the correlation coefficient found for all crabs, 0.98 indicates there is a strong linear relationship between the pre-molt sizes and the post-molt sizes for crabs raised in a laboratory. As demonstrated in Figure 11, the residuals seem to be more slightly more of a heteroscedastic than for the residual plot of all the crabs, however, overall the residuals and linear model seem to be homoscedastic yet again. As the residuals take on the same definition, regardless of the data begin examined, the residuals still have a mean of approximately 0 as shown in Table 5. However, the skewness and kurtosis continue to suggest that this sample deviates from a normal distribution, for these cases both are larger. Using Figures 12 and 13, there seems to a more peakiness around the mean, suggesting the kurtosis is higher. Additionally, in Figure 13, it can be clearly seen the tails of this distribution are longer than those for the residuals Figure 6. The skewness of the residuals can be seen clearly as this smooth histogram reveals outliers are lying above the mean which making this distribution positively skewed.

Lastly, Figure 14 can be used to demonstrate how the residuals for pre-molt and post-molt sizes for crabs held in a laboratory have tails towards the positive end of the distribution that is much larger than what would be expected from a normal distribution. Similar to the previous case, Table 6 demonstrates that most of the residuals lie within 2 standard deviations of the mean, while roughly 74% of the residuals lie within one standard deviation of the mean. This corresponds with a higher kurtosis, as pre-molt and post-molt sizes of crabs measured in the laboratory have larger tails than the previous case. Since 74% of the residuals lie within one standard deviation for the mean, the linear model can be deemed an appropriate approximation method. In addition, this linear model(equation 2) seems to be almost identical to the linear model of the previous case (equation 1).

On the other hand, the other population of crabs was captured and recaptured. Crabs we caught, measured, tagged with a unique identification number, and returned to the water. These crabs were later caught in traps for post-molt measurements. Using the pre-molt sizes and post-molt sizes for the set of crabs caught and captured, a linear model was generated (equation 3). This linear model is plotted with the pre-molt and post-molt sizes of crabs that have been captured and recaptured in Figure 15. As can be seen, this linear model tries to minimize the error amongst all data points. This line is once again the least-squares regression line for the pre-molt and post-molt sizes of crabs that have been caught and recaptured. This line minimizes the summation of the squares of the residuals to predict accurate results based on post-molt sizes. As shown in Table 7, the mean of the post-molt sizes for this distribution is 152.9 mm and the corresponding mean for the pre-molt sizes is 139 mm. Unlike the previous 2 populations of crabs examined, this data converges slightly more towards a normal distribution. This conclusion is made based on the kurtosis of the crabs that have been caught and re-caught. While the skewness of this distribution is strongly negatively skewed, the kurtosis indicates this distribution varies slightly less from a normal distribution. Since the kurtosis is closer to 3, which is the kurtosis for a normal distribution, we can assume the pre-molt and post-mot sizes for crabs that have been captured and recaptured is a closet fit to a normal distribution. However, based on Figures 16 & 17, it is difficult to come to this conclusion as the distribution looks rather peaky with long tails. This lower kurtosis is not to say that this distribution is similar to that of a normal distribution, just that it adheres to a normal distribution slightly more than the previous cases. for the case of pre-molt and post-molt sizes of crabs captured and recaptured, the residuals are plotted in Figure 18. Based on this figure, the residuals and linear model can be said to be homoscedastic, meaning there are an even variation and no funneling towards one end or the other. However, based on Figures 19 & 20, the distribution of the residuals of pre-molt and post-molt sizes of crabs that have been captured and recaptured seem to follow a normal distribution more than the previous cases. There one large peak above the normal distribution which would leave some deviation from the normal curve, but overall, this distribution seems to follow a normal distribution. The mean of the residuals is still approximately 0, as required by the definition of a residual. The skewness of the residuals, as well as the standard deviation, is much smaller than the previous cases as shown in Table 8. This is one sign of this distribution adhering to a normal distribution. Additionally, the kurtosis of the residuals for this case is 4, which is larger than 3, the kurtosis for a normal distribution, but much closer to 3 than the previous

two cases. As shown in Figure 21, the residuals seem to follow the line that represents a normal distribution for the most part except towards the tails. Table 9 solidifies the statements that have been made regarding this distribution. Within 2 standard deviations from the mean, 93.6% of the residuals lie, meanwhile in a normal distribution, roughly 95% of the data lies within this range. While the previous two cases had very few tails, this distribution seems to have slightly longer tails that would be similar to that of a normal distribution. However, within one standard deviation of the mean roughly 77% of the residuals lie, while in a normal distribution 68% of the data lies within this same range. One downfall of this dataset is the correlation coefficient. The correlation coefficient for the cases where the crabs are captured and re-captured is roughly .93 a much lesser value than what was discovered in previous cases. In conclusion, this linear model is a less accurate prediction of the pre-molt sizes given the post-molt sizes of these crabs that have been captured and recaptured.

Using the linear model generated from the entire population of crabs, the pre-molt sizes can be predicted from the post-molt sizes, given some error will occur (residual). Figure 22 demonstrates the prediction of pre-molt sizes based on corresponding post-molt sizes using the linear model given by Eqn 1. As expected, the pre-molt sizes and post-molt sizes correlate linearly and are plotted along the line of least squares. The mean and median for the predicted pre-molt size and corresponding post-molt size are listed in Table 10. Notice the skewness and kurtosis are the same for both these data sets. This is due to the linear correlation between pre-molt sizes and post-molt sizes. Figures 23 & 24 are used to visualize how the distributions are negatively skewed as the majority of the data is above the mean, but outliers are pulling down on the mean. The predicted pre-molt data generated can be seen to deviate from a normal distribution as shown in the quantile plot in Figure 25.

In addition to predicting the pre-molt values, the second set of data was also split up into clean shells and fouled shells. Differentiating from these shells will help rule out mistakes, particularly in measurements. Some of the outliers from predicting the pre-malt size from the post-molt size may have been due to inaccurate measurements due to fouled shells. Figure 26 contains the post-molt sizes and predicted pre-molt sizes based on the linear model found for all crabs. As before, since the model is linear, the skewness and kurtosis of the pre-molt size and post-molt size are the same, shown in Table 11. One interesting point is the mean of clean shells is smaller than the mean of fouled shells. Fouled shells could potentially have barnacles, dents, fractures, or other deformations which would make it harder to get an accurate post-molt measurement. Similar to clean shells, the pre-malt sizes of fouled shells lie on the line represented by the linear model. This is expected as the linear model is used as the method for prediction. Table 12 lists the descriptive statistics for fouled shells. The kurtosis and skewness for both pre-molt and post-molt sizes for fouled shells deviate from a normal distribution, unlike that of clean shells. In addition, the average post-molt shell size is much larger than that for clean shells. It is believed this is due to the deformation of the shell and other sources of fouling. As can be seen in Figure 31 & Figure 32, the fouled data is much more negatively skewed, meaning the majority of the data is above the mean; this confirms our conclusions regarding the variations in the pre-malt size based on the linear model. Also, the skewness and kurtosis for clean shells is a closer approximation to

a normal distribution than fouled shells. From these statements, a further conclusion can be made and it is that the predicted pre-malt value is more accurate for clean shells than for fouled shells. This seems intuitive, but it can now also be justified by the figures listed in this section. While it cannot be confirmed, it is assumed the residual for fouled shells would be much larger than for clean shells. This is because clean shells provide accurate measurements while fouled shells tend to vary from the linear model.

One problem with regression is that it can be highly unstable because outliers will affect the least-squares model and deform the line of best fit. A different way to make a linear model is to divide the data into subintervals. Taking the mean of these subintervals for pre-molt and post-molt values, the cloud of the data points in that one subinterval can be replaced by a point. The averages of the subintervals can be shown in Figure 34. A line through these points, minimizing the error by each value will make them more stable because the variation is averages and outliers will have a much smaller effect on the data. This line is shown in Figure 35 and is an alternative way to predict pre-malt values given post-molt values.

References

1) Nolan, D., & Speed, T. P. (2000). Stat labs: Mathematical statistics through applications. Springer Science & Business Media.

Scientific writing made easy: A step-by-step guide to undergraduate writing in the biological sciences. (2016, October 3). The Ecological Society of America.

2)Heteroskedasticity. (n.d.). Investopedia
https://www.investopedia.com/terms/h/heteroskedasticity.asp

3) Introduction to residuals (article). (n.d.). Khan Academy.
https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/regression-library/a/introduction-to-residuals

4) Rencher, A. C., & Schaalje, G. B. (2008). Linear models in statistics. John Wiley & Sons.