Prediction of the size of premolt crabs from the size of post molt crab shells

Jefferey Lin

The Issues:

The molting process is extremely important in understanding how these crabs grow, the catchability and breeding potential, and ensuring the overall health of the crab populations and viability of the crab fisheries. Being a crab fisherman is one of the most dangerous jobs in the world, which can be seen in Deadliest Catch on discovery. Since the early 2000s, there has been a noticeable rise in popularity of crab hatcheries and farms. In these controlled environments, the scientists or fisherman are able to measure crab sizes before and after molt.

However, doing the same measurements in the ocean, presents more challenges. It's impossible to calculate pre molt and post molt crab sizes, as the crab would either go to distributors to be sold worldwide or would be tossed back into the ocean. Consequently, when the fisherman catches a crab and records its size, there is a need to predict the crab's size prior to its molting. This predictive capability has immense importance for fisheries, as it enables them to calculate predicted growth rate, enhancing the understanding of crab biology and to establish informed catch quotas for those crab boats. With this predictive approach, fisheries can continue sustainable practices and ensure the long term health of the crab population.

The Findings:

In this study, I conducted an analysis of a dataset comprising 472 crabs, each measured twice, once before molting and once after molting. These measurements were taken either in the lab or out at sea, in the field. Out of the 472 crabs, 111 were measured in the field and 361 crabs were measured in the labs. This data set shows distinct growth trends among the crabs, and by using statistical methods, we can predict the pre molt size of a given crab based on its post molt size. This is a valuable tool for those catching crabs from the boat.

On average, the size of a crab after molting was 147.4 units while the average size of a crab pre molt was 121.29 units. The dataset exhibited a large variation of crab sizes as there were a number of crabs that were much smaller than the others, the smallest being 31.1 units pre molt, and 38.3 units post molt. The largest crab was 155.1 units pre molt and 166.8 units post molt. There was also a noticeable difference between the crabs that were measured in the field and crabs that were measured in the labs. The average size of the crabs that were measured in the

field was 139.01 units pre molt and 152.96 units post molt. Meanwhile, the average size of the crabs measured in the lab pre molt was 126.12 and 141.11 post molt.

Through graphical representation and calculation, I derived the predictive equation to figure out the size of a crab before molting, given that you know the size of the crab after molting: Y= 1.0732(X) -25.214, where X represents the size of the crab measured after molting, and Y is the predicted pre molt size. Furthermore, I also examined if there was a difference in the predicted sizes of crabs pre molt if they were measured in the lab or the field. For instance, if a crab was measured to be 100 units post molt, in the field, the predicted pre molt size of the crab would be 83.808 units, conversely, the predicted pre molt size of the lab measured crab would be 82.406. This subtle difference underscores the importance of considering where the crab was caught or measured.

Discussion:

The results of this study provide valuable insights into the growth patterns of crabs before and after molting, offering a predictive equation of, Y = 1.0732X - 25.214, as a practical tool for estimating pre-molt sizes based on post-molt measurements. The observed variability in crab sizes shows the importance of accounting for individual differences and environmental factors. The small distinction between crabs measured in the field and in the lab emphasizes the influence of habitat on crab growth, with potential implications for resource management. The subtle differences in predicted pre molt sizes based on the two measurement locations highlight the need for considerations in applying the predictive model. Which reinforces the importance of understanding the environmental impact in crab harvesting practices. Overall, this study contributes to the field by providing a valuable tool for crab size prediction while emphasizing the complexity of crab growth variability influenced by individual growth and environmental factors.

Appendix A: Method

Initially, a comparison between pre molt and post molt data points was conducted. The descriptive statistics of the two variables were initially calculated to simply describe the variables. These statistics include min and max, median, mean, standard deviation, skewness and kurtosis, and were calculated in Excel and are detailed in Figure 1. Then, in R studios, quantile plots were generated for both variables to check for data normality, which can be seen in figures 1 and 2. While these two plots offer insight into the normality of the data, an additional normality test was used for confirmation, called the Anderson-Darling test. The null hypothesis in this test is that the data is normally distributed. The alternative hypothesis is that the data is not normally distributed. This test is used to strengthen the assessment of normality in the crab molt data set.

To enhance data visualization, both a box histogram created in Excel and a smooth histogram created in R were exported for the two variables, to illustrate the distribution of the size of the crabs post molt and pre molt. These histograms can be seen in figures 3 and 4. Given that we want to predict what the size of the crabs were pre molt, given post molt data, a scatter plot was created(Figure 5). The X variable represents post molt data, which serves as the predictor, and Y variable denotes the pre molt data, acting as the predicted variable. A trend line was added and the linear least squares model was calculated to predict pre molt crab sizes based on post molt crab sizes. A Pearson's R-squared value was calculated to figure out how much of the variation in the predicted variable is accounted for by the predictor variable.

To know how well the linear model fits the provided data, the residuals or errors were visualized in Figure 6. This plot illustrates the difference between the actual data and the values predicted by the model. Additionally, a Q-Q plot of residuals (Figure 8) was generated in R to check the normality of the residuals. The skewness and kurtosis of the residuals were also calculated to further examine their distribution. To rigorously examine the normality of residuals, the Anderson Darling test was conducted once again, this time based on residuals The null hypothesis posits that the residuals are consistent with a normal distribution, while the alternative hypothesis suggests that it is not consistent with a normal distribution. Furthermore, a visual inspection of the plot of residuals was done to identify heteroskedasticity or homoskedasticity. Additionally, the Breusch-Pagan test was used to quantitatively determine heteroskedasticity or homoskedasticity.

To show the distinction between lab measured crabs versus field measured labs, the two original columns were now split into four, still maintaining the pre molt and post molt measurements. This resulted in pre molt and post molt data sets for both field and lab crabs. Subsequently, two scatter plots were generated for both the fields and lab crabs. For each scatter plot, a linear least squares model was computed to capture the predictive framework for estimating pre molt sizes based on post molt sizes, given whether the crab was measured in the field or in the lab. Additionally, Pearson's R-Squared values were calculated to determine how well the linear model fits the data for field measure crabs versus lab measure crabs.

Lastly, a cross validation method was employed to estimate the prediction accuracy of the linear model. This involved splitting the data into 5 groups, where ⅘ of the data is used for training, and ⅕ for testing each iteration. The process was repeated five times, then dividing the sum of the tests by the five to achieve the average mean square error, to see how accurate our model is.

Appendix B: Results

Table 1 provides basic descriptive statistics for the overall dataset, specifically focusing on two variables, pre molt and post molt sizes. This is disregarding whether the crab size comes

from the field or the lab. Given the skewness and kurtosis in the table, the data can already be identified as highly skewed to the left. The kurtosis indicates that the distribution has heavy tails, or some extreme values. To visually confirm the non-normality of the data, QQ plots were generated for both post molt and pre molt, which can be seen in Figures 1 and 2. These plots indicate the non-normality of the data, as evidenced by the pronounced discrepancies in the tails from the normality line. For quantitative assessment of normality, the Anderson Darling test was run in R studios and the resulting p value equaled 3.7e-24, which is an extremely small value. Being significantly below the chosen significance level of .05, the null hypothesis of the data being normally distributed was rejected, which aligns with the skewness values, kurtosis values and QQ plots.

Table 1:

	Min	Max	Mean	Median	SD	Skewness	Kurtosis
Pre Molt	31.1	155.1	121.29	132.8	15.85	-2.00	6.72
Post molt	38.3	166.8	143.90	147.4	14.63	-2.3	10.06

Figure 1:

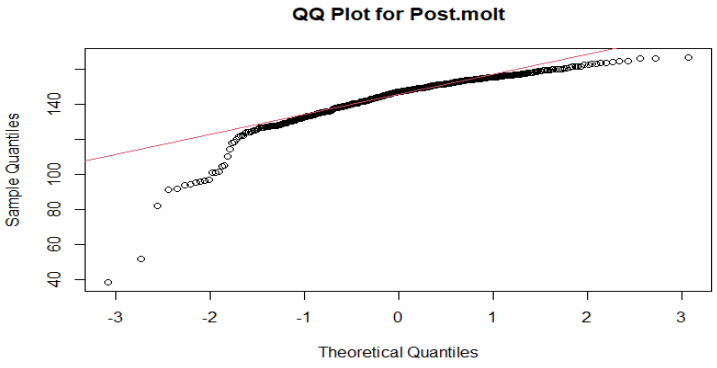
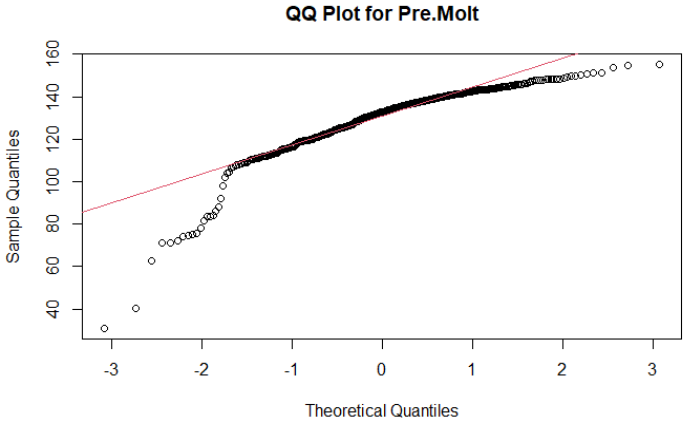


Figure 2:



The histograms presented below illustrate the distribution of crab sizes based on whether the crabs were post molt and pre molt. Visually, it is evident that the data contains outliers of smaller crabs, while the majority were in the 130-150 size region. The scatter plot in Figure 5 further illustrates the relationship between post molt (x-variable, predictor) and pre molt (y-variable, predicted). To predict the size of a crab pre molt based on the post molt data, a linear

regression equation was derived from the scatter plot: Y= 1.0732(X) -25.214. The Pearson R-squared value depicts how much of the variation in the predicted variable is accounted for by the predictor variable. This data set's R^2= .980833, which indicates that approximately 98% of the variation in premolt size can be accounted for by the variation in post molt data. This also indicates that our regression model is a good fit for the data given.

Figure 3

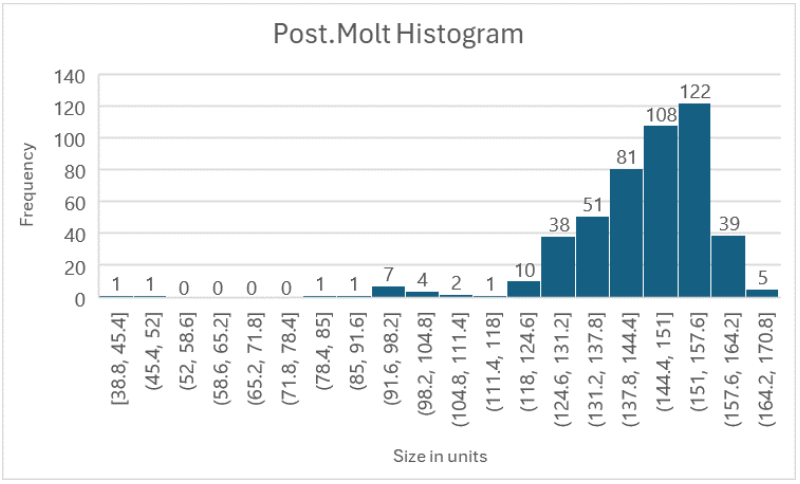
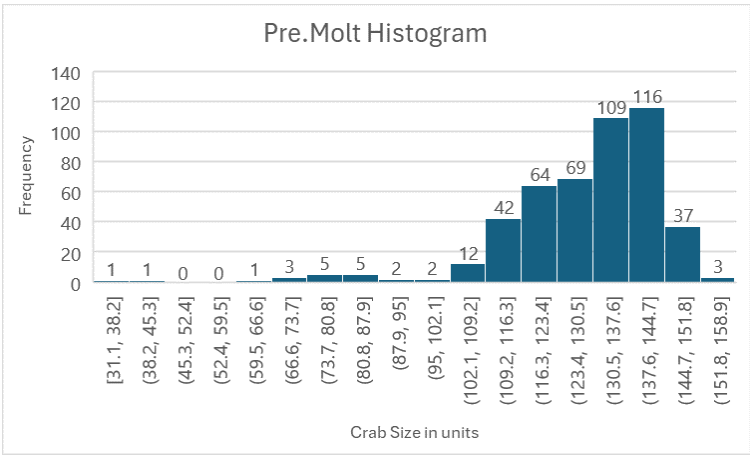
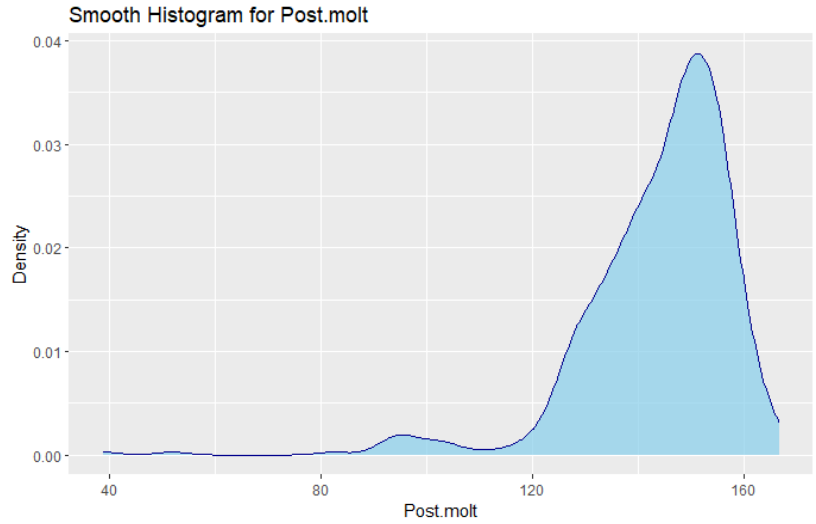


Figure 4





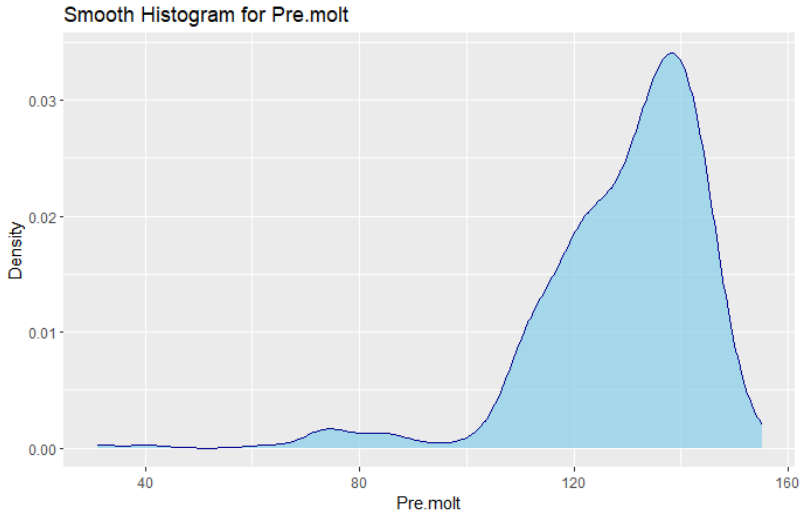
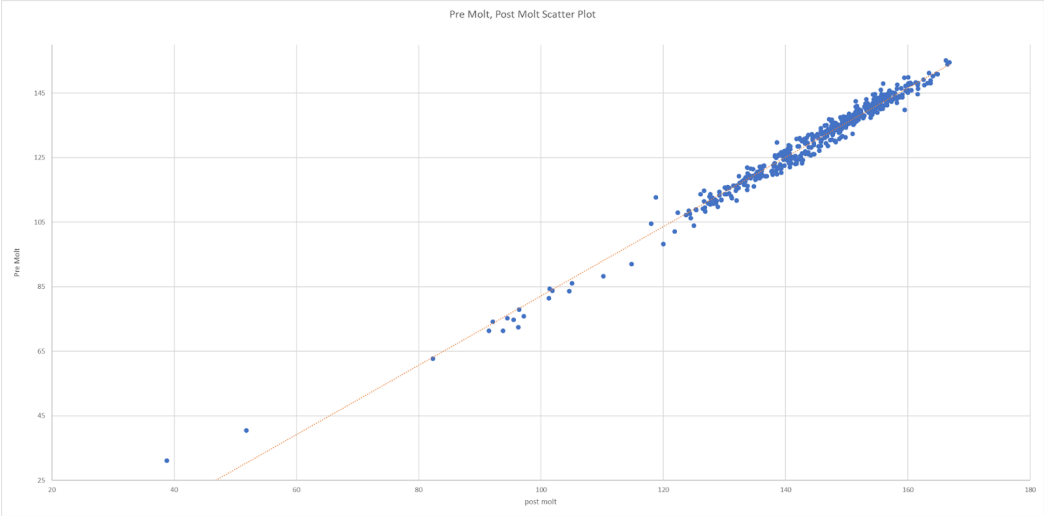


Figure 5



In Figure 6, a residual plot was created to assess the goodness of fit for the regression model. Residuals represent the difference, or error, between the observed values and the values predicted by the model, or linear regression equation. The accompanying smooth histogram (Figure 7) reveals a right skewness, with some residuals exceeding 10+. To validate this observation, I also calculated the skewness and kurtosis of residuals. The skewness of the residuals equaled .8427671, which indicates moderate positive, or right, skewness, which aligns with the histogram. The kurtosis of residuals equaled 5.34218, which indicates a higher distribution of residual than a normal distribution of kurtosis = 3. This indicates heavier tails and more extreme values, which also aligns with the histogram. Again, to test the null hypothesis that the residuals are normally distributed, the Anderson Darling test was run again, and came out to a p value of 1.271e-06, which is very small and obviously lower than the standard .05, so we can reject the null and say that the residuals do not follow a normal distribution. This result aligns with the skewness, kurtosis values and histogram observations.

The Q-Q plot in Figure 8 provides further evidence that the residuals are not normally distributed, as the tails deviate from the normal line, especially on the right tail, there are some major value differences. To check for heteroskedasticity or homoscedasticity, we can look at the residual plots of the Breusch-Pagan test. If it is heteroskedastic, then the variance of residuals is not constant across all the levels of the independent variable, the Post molt data. After running this test, the test returned a p-value of 3.247e-21, which indicates to reject the null that residuals are uniformly scattered, and suggests the presence of heteroskedasticity.

Figure 6:

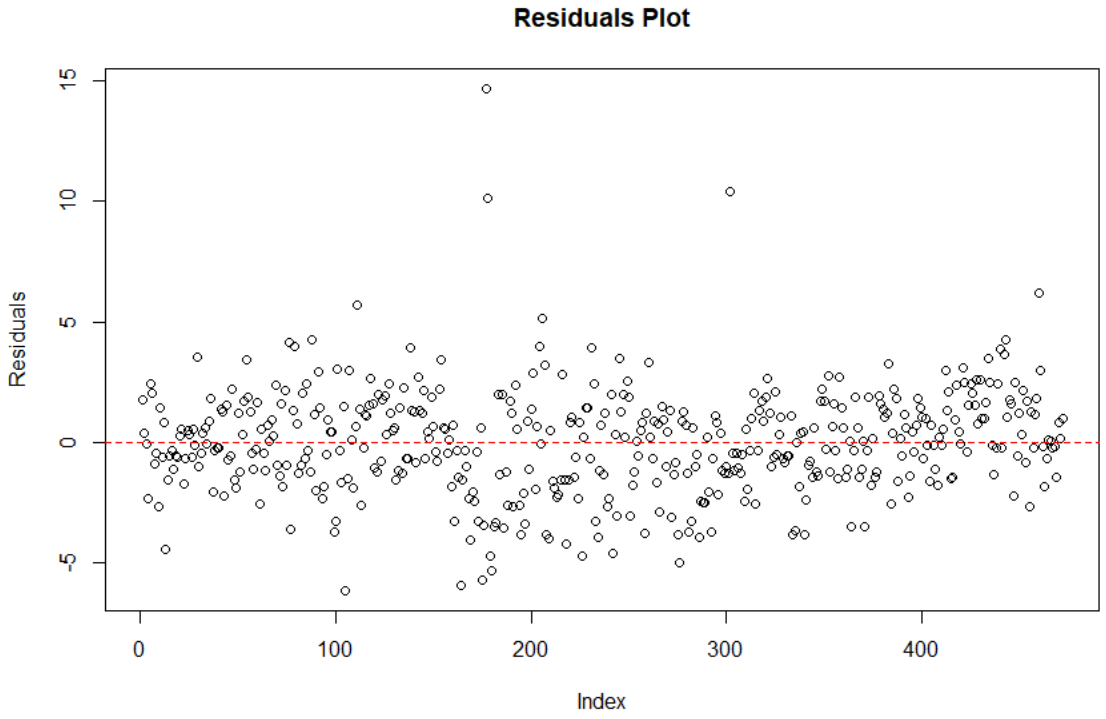


Figure 7

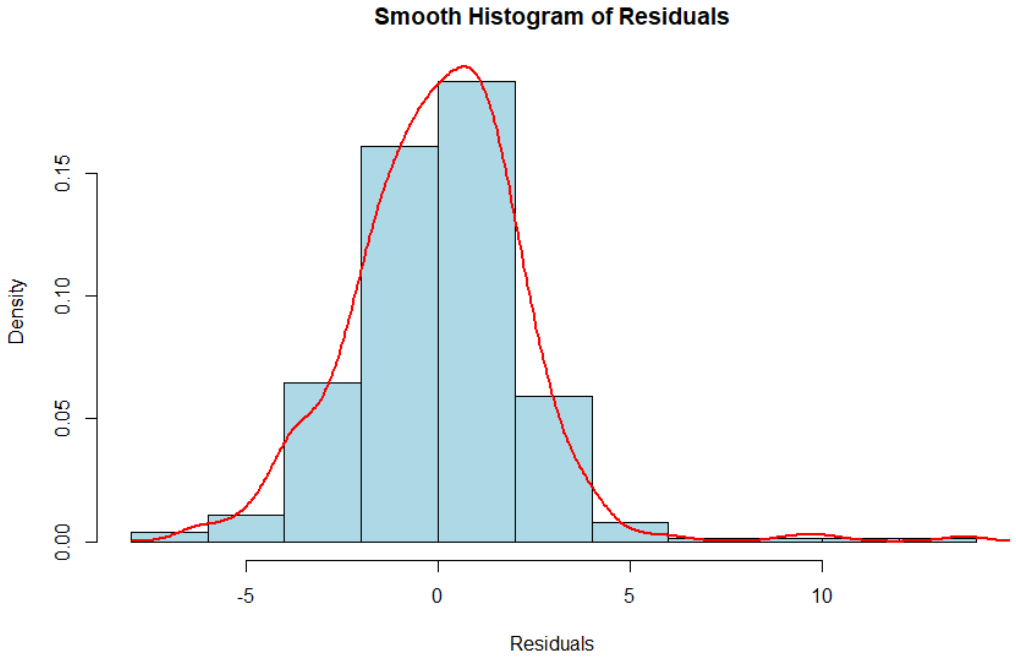
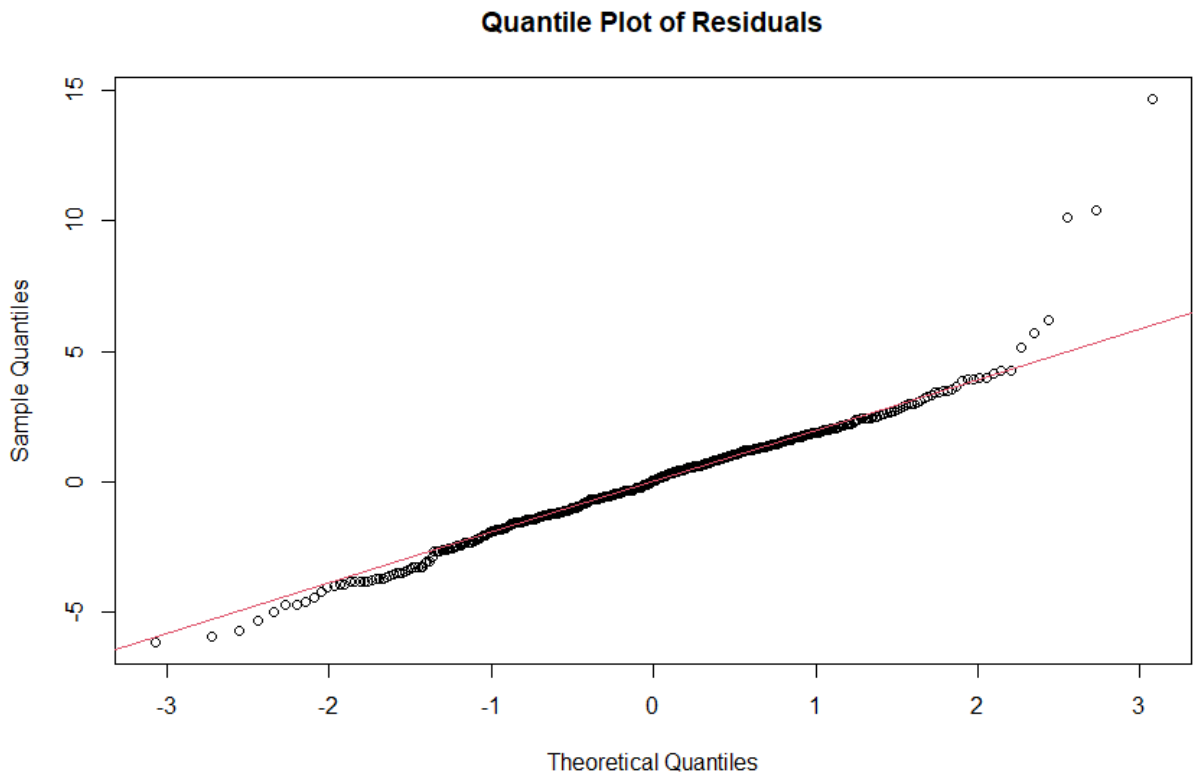


Figure 8



In this data set, while the primary focus has been on analyzing the relationship between post-molt and pre molt data in crabs, it's important to acknowledge the presence of another significant variable, the location of measurements, whether they were taken in the field or in the lab. This new addition could potentially add a new influence in the data patterns. Tables 2 and 3 reveal a clear difference in average size of crabs. While there are some very small crabs measured in the lab, the median size is 10+ units bigger for the field measured crabs. Isolating the post and pre molt data into field measured and lab measured crabs, two different scatter plots were exported, along with their linear regressions. For the field measured crabs, the linear regression equation is Y = 1.0421-20.402. For the lab crabs, Y = 1.0739x-25.344. These equations offer predictive models specific to the two locations, reflecting the observed relationships between post and pre molt sizes in the field and lab.

Table 2:

Field	#	Min	Max	Mean	Median	SD
Post Molt	111	127.7	166.5	152.96	154	6.66
Pre Molt	111	113.6	153.9	139.01	140.5	7.19

Figure 9:

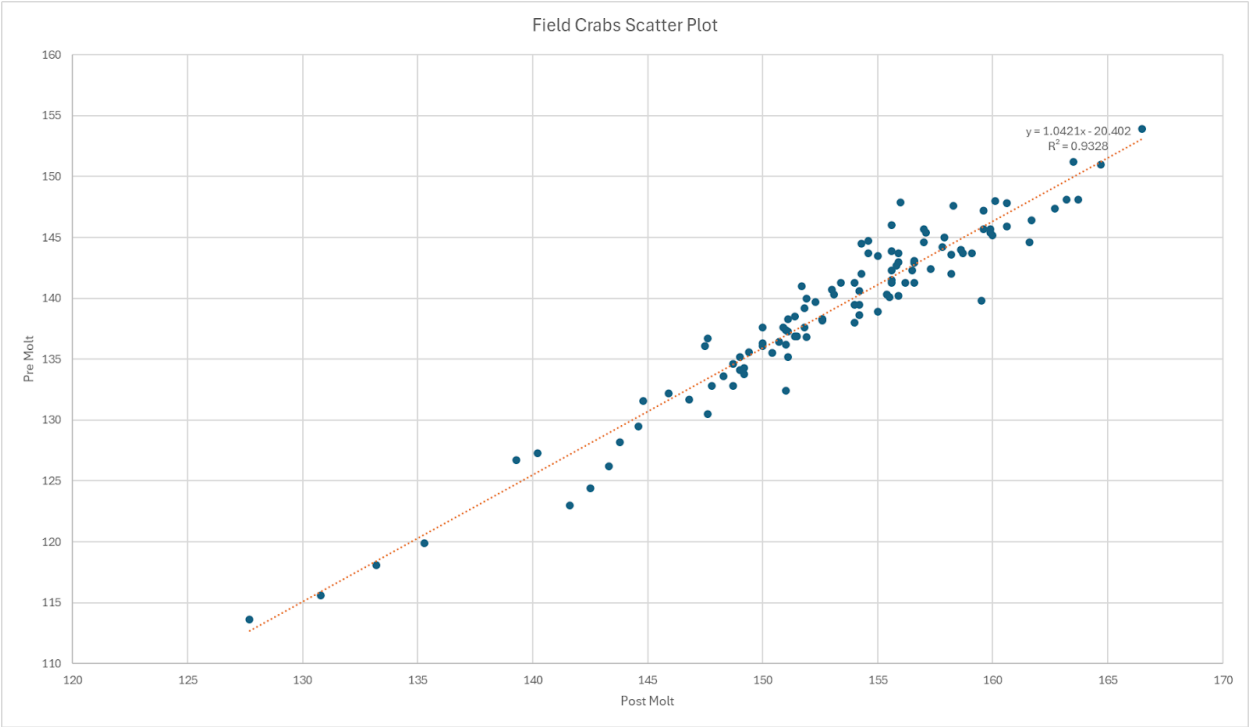
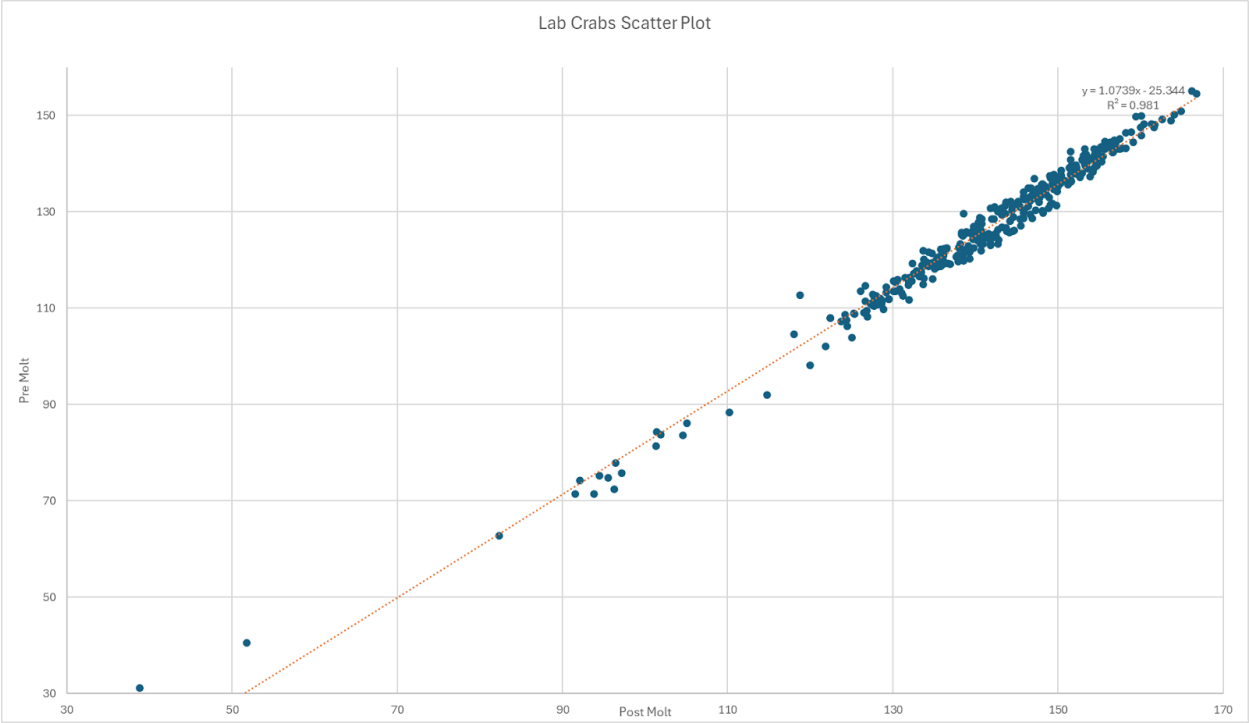


Table 3:

Lab	#	Min	Max	Mean	Median	SD
Post Molt	361	38.8	166.8	141.11	143.65	15.24
Pre molt	361	31.1	155.1	126.12	128.85	16.52

Figure 9:



Finally, cross validation was employed as a technique to estimate the prediction accuracy of our model. This involved dividing the crab molt dataset into five groups, using one of the groups in training the model, then using the other remaining groups to train the data. This resulted in Root Mean Square Error, RMSE, and Mean absolute Error,MAE, values of 2.003 and 1.606. These metrics provide insight into the accuracy of our model's predictions, with lower values indicating a better fit.

Appendix C: Code

Code for Skewness

```
> data <- Crab.Molt.Data$Post.molt
>
> skew <- skewness(data)
>
> cat("Skewness for the 'post.molt' column:", skew, "\n")
```

*changed to Pre.Molt for premolt skewness.

Kurtosis Code:

```
> data <- Crab.Molt.Data$Post.molt

> kurt <- kurtosis(data)

> cat("Kurtosis for the 'post.molt' column:", kurt, "\n")
```

*change to pre.molt for premolt kurtosis

QQ Plot Code

```
> data <- Crab.Molt.Data$Post.molt
>
> # Creating QQ plot
> qqnorm(data, main = "QQ Plot for Post.molt")
> qqline(data, col = 2)
```

*change to pre.molt for premolt QQ plot

Creating a smooth histogram

```
> # create smooth histogram
> ggplot(data = Crab.Molt.Data, aes(x = Post.molt)) +
+    geom_density(fill = "skyblue", color = "darkblue", alpha = 0.7) +
+    labs(title = "Smooth Histogram for Post.molt", x = "Post.molt", y =
"Density")
```

*change to pre.molt for premolt Smooth histogram

Plot of residuals vs fitted values

```
> plot(fitted(model), resid(model))
> abline(h = 0)
```

R-squared and plot of residuals

```
> pearson_r_squared <- cor(Crab.Molt.Data$Pre.molt, predict(model))^2
> cat("Pearson's R-squared value:", pearson_r_squared, "\n")
Pearson's R-squared value: 0.9808326
>
> # Residuals plot
> plot(residuals(model), main = "Residuals Plot", xlab = "Index", ylab =
"Residuals")
> abline(h = 0, col = "red", lty = 2)
```

Anderson Darling:

```
> # Post molt AD test
> post_molt_ad_test <- ad.test(Crab.Molt.Data$Post.molt)
> cat("Anderson-Darling test for Post.molt:", post_molt_ad_test$statistic,
"\n")
Anderson-Darling test for Post.molt: 13.97019
> cat("p-value:", post_molt_ad_test$p.value, "\n")
p-value: 3.7e-24
>
> # Pre molt AD test
> pre_molt_ad_test <- ad.test(Crab.Molt.Data$Pre.molt)
> cat("Anderson-Darling test for Pre.molt:", pre_molt_ad_test$statistic, "\n")
Anderson-Darling test for Pre.molt: 12.94778
> cat("p-value:", pre_molt_ad_test$p.value, "\n")
p-value: 3.7e-24
```

Skewness of residuals:

```
> set.seed(123)
> X <- rnorm(100)
> Y <- 1.0732 * X - 25.214 + rnorm(100)

# Fit a linear regression model
> model <- lm(Y ~ X)

> residuals <- residuals(model)

# Skewness of residuals
> residual_skewness <- skewness(residuals)
> cat("Skewness of residuals:", residual_skewness, "\n")
```

Kurtosis of residuals

```
> set.seed(123)
> X <- rnorm(100)
> Y <- 1.0732 * X - 25.214 + rnorm(100)

# Fit a linear regression model
> model <- lm(Y ~ X)
> residuals <- residuals(model)
> residual_kurtosis <- kurtosis(residuals)
> cat("Kurtosis of residuals:", residual_kurtosis, "\n")
```

Q-Qplot for residuals

```
> model <- lm(Y ~ X)
> residuals <- residuals(model)
> qqnorm(residuals)
> qqline(residuals)
```

Anderson darling residual test

```
> ad_test_res <- ad.test(residuals(model))
> cat("Anderson-Darling test for residuals:", ad_test_res$statistic, "\n")
Anderson-Darling test for residuals: Inf
> cat("p-value:", ad_test_res$p.value, "\n")
p-value: 1.271186e-06
```

Bresuch pagan test:

```
> bp_test_res <- bptest(model)
> cat("Breusch-Pagan test for residuals:", bp_test_res$statistic, "\n")
Breusch-Pagan test for residuals: 89.38639
> cat("p-value:", bp_test_res$p.value, "\n")
p-value: 3.247625e-21
```

Cross Validation:

```
> set.seed(123)

> random_sample <- createDataPartition(Crab.Molt.Data$Pre.molt, p = 0.8, list
= FALSE)
> training_dataset <- Crab.Molt.Data[random_sample, ]
> testing_dataset <- Crab.Molt.Data[-random_sample, ]
>
> #build model
> model <- lm(Pre.molt ~ Post.molt, data = training_dataset)
>
> # Predicting
> predictions <- predict(model, testing_dataset)
>
> # computation
> metrics <- data.frame(
+    R2 = R2(predictions, testing_dataset$Pre.molt),
+    RMSE = RMSE(predictions, testing_dataset$Pre.molt),
+    MAE = MAE(predictions, testing_dataset$Pre.molt)
+ )
>
> print(metrics)
```