Shawn Koohy

13 February 2022

Dr. Davis

MTH332

Prediction of Crabs Carapace Using Linear Regression

Abstract

Being able to predict the sizes of a crab's carapace is helpful for both the fishers and the crabs themselves. Considering the severe de-escalation in the Dungeness crab population, the disappearance of them by the northern Californian coast would prove to be a major economic hit. Fishermen would lose their jobs and any that still would be working would not have much to catch, not to mention the obvious detriments that the ocean ecosystem would suffer from the absence of the Dungeness crabs. To find the optimal carapace size for the crabs to be caught at will solve this problem. We will be successful in doing this by using various statistical, analytical, and computational methods with an end goal of finding a linear model to predict the pre-molt sizes from the measured post-molt size. We will also be proving why the linear model works and is accurate. How our model differs from the actual data will be examined to confirm our predications and confirmations of the assumed accurate regression lines. The results from this will help the fishing economy on the California coast, restore the Dungeness crab population and overall, better the aquatic ecosystem in the Pacific Ocean.

Introduction & Background

Unregulated or overfishing can lead to a major disruption in a sea creatures' population and can disrupt their appropriate environment along with their living conditions. In specific, Dungeness crabs are commercially fished along the Pacific coast of North America mainly between December and June. The vast majority of what is caught by the fishermen are male crabs with the intent to keep the female crabs to maintain the sustainability of the overall crab population. The overfishing of these crabs has led to an imbalance in the sex ratio and possibly contributed to the decline in crab population near the Californian central coast. Due to this imbalance, there has been a large surge in the parasitic ribbon worm population. This increase in the parasitic population has led to worms destroying anywhere from 50-90% of the Dungeness crab eggs every year.

To try to reverse this effect, size restriction has been placed on the male and female crabs for them to have more opportunities to mate before being caught. A carapace is essential the crabs shell which is measured in millimeters along it's center from side-to-side. Male crabs tend to have growth marks on their shells where female crabs do not, so it is more difficult to determine their age and if they can be caught or not. The marks appear when a crab has molted or removed its old shell for a new one. When it comes to female crabs they do not molt very often and sometimes not even yearly. So there needs to be some way in specifying their age, last time they molted and their growth pattern. We can do this by looking at female crabs who have recently molted and viewing their change in carapace size to predict growth patterns and extrapolate the data with the intent to develop recommendations on size restriction of the female crabs.

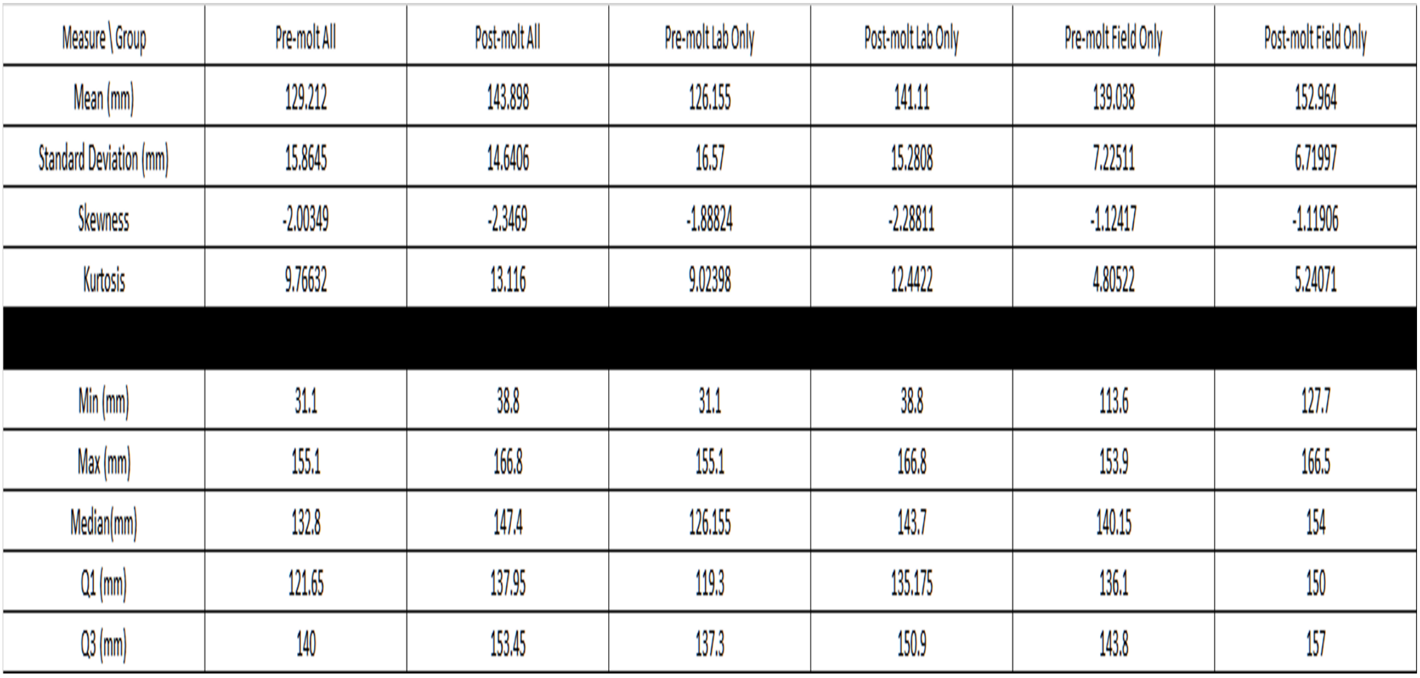
The data for this observation was conducted by David Hankin, Nancy Diamond, Michael Mohr, James Ianelli, with help from the California Department of Fish and Game and commercial fishers from northern California and southern Oregon. There are two sets of data, one including 472 carapace sizes of crabs from laboratory data (This set will sometimes be denoted by a 1) and another 472 carapace sizes from crabs collected in a capture-recapture manner (This set will sometimes be denoted by a 0). Each data set also included the pre-molt and post-molt sizes of the carapace for the lab (1) and the Field/capture-recapture (0).

Methods

To analyze the given data Mathematica was used. From Mathematica I imported the data and organized it so the post-molt size was the independent variable, and the pre-molt size was the dependent variable. I created about 9 different .csv files to organize all my data correctly in the way I wanted. Those files were organized into the following, “Pre-molt All vs Post-molt All”, “Pre-molt Only”, “Post-molt Only”, “Pre-molt (1) Only”, “Post-molt (1) Only”, “Pre-molt (0) Only”, “Post-molt (0) Only”, “Pre-molt (0) vs Post-molt (0)”, and “Pre-molt (1) vs Post-molt (1)”. The majority of these were made so I could find the Mean, Standard Deviation, Skewness, Kurtosis, Minimum, Maximum, Median, Quartile 1 and Quartile 3 of very specific groups of crabs to get a better understand of how result vary from group to group (Note that the names derive from classification of lab data (1), field data (0) or All/Both). After having all the data separated correctly, I then created a scatter plot along with a plotted regression line for the “Pre-molt All vs Post-molt All”. Once I had my regression line I found and plotted a corresponding scatter plot of residual values along with a histogram, smooth histogram, and theoretical normal distribution with the same Mean, Standard Deviation, Minimum and Maximum of the residuals. I then plotted a Quantile Plot for my residuals, and all sets of data groups. I repeated the same steps and created similar graphs for the remaining two data sets “Pre-molt (1) vs Post-molt (1)” and “Pre-molt (0) vs Post-molt (0)”. Lastly, I plotted the smooth histograms of the 3 sets of residuals on top of each other.

These statistical measurements helped me understand the distribution and spread of all data groups. The quantile plots were created to compare the data sets distribution to the theoretical normal distribution. The histograms give a great visual to many of the measurement made such as skewness, kurtosis and mean. Plotting the residual smooth histograms on top of each other makes it easy to compare the relative difference between them. The generated regression line is what will be used to make our prediction, before, during and after the extent of our data.

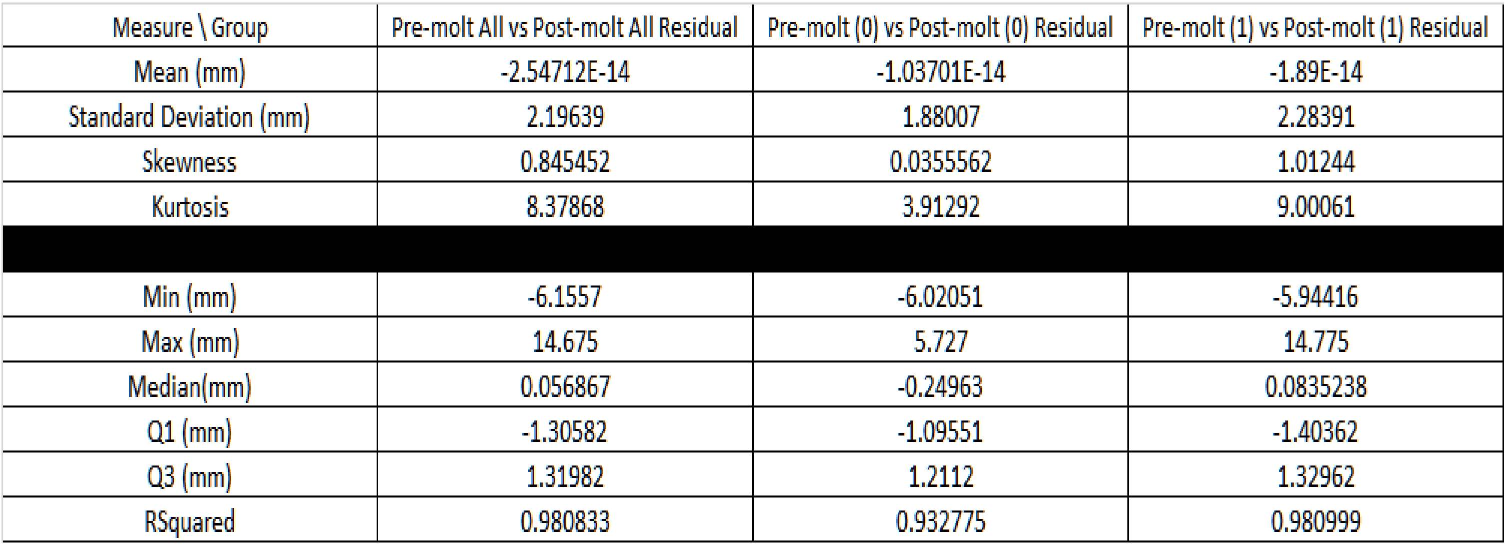
Results



Measure \ Group	Pre-molt All	Post-molt All	Pre-molt Lab Only	Post-molt Lab Only	Pre-molt Field Only	Post-molt Field Only
Mean (mm)	129.212	143.898	126.155	141.11	139.038	152.964
Standard Deviation (mm)	15.8645	14.6406	16.57	15.2808	7.22511	6.71997
Skewness	-2.00349	-2.3469	-1.88824	-2.28811	-1.12417	-1.11906
Kurtosis	9.76632	13.116	9.02398	12.4422	4.80522	5.24071
Min (mm)	31.1	38.8	31.1	38.8	113.6	127.7
Max (mm)	155.1	166.8	155.1	166.8	153.9	166.5
Median(mm)	132.8	147.4	126.155	143.7	140.15	154
Q1 (mm)	121.65	137.95	119.3	135.175	136.1	150
Q3 (mm)	140	153.45	137.3	150.9	143.8	157

Figure 1. Table of Descriptive Statistics for the 6 divided data groups

We can see in Figure 1. That the kurtosis for all groups is high and the skewness is decently negative. This can be interpreted as telling us that our data is not normally distributed.



Measure \ Group	Pre-molt All vs Post-molt All Residual	Pre-molt (0) vs Post-molt (0) Residual	Pre-molt (1) vs Post-molt (1) Residual
Mean (mm)	-2.54712E-14	-1.03701E-14	-1.89E-14
Standard Deviation (mm)	2.19639	1.88007	2.28391
Skewness	0.845452	0.0355562	1.01244
Kurtosis	8.37868	3.91292	9.00061
Min (mm)	-6.1557	-6.02051	-5.94416
Max (mm)	14.675	5.727	14.775
Median(mm)	0.056867	-0.24963	0.0835238
Q1 (mm)	-1.30582	-1.09551	-1.40362
Q3 (mm)	1.31982	1.2112	1.32962
RSquared	0.980833	0.932775	0.980999

Figure 2. Residual Descriptive Statistics

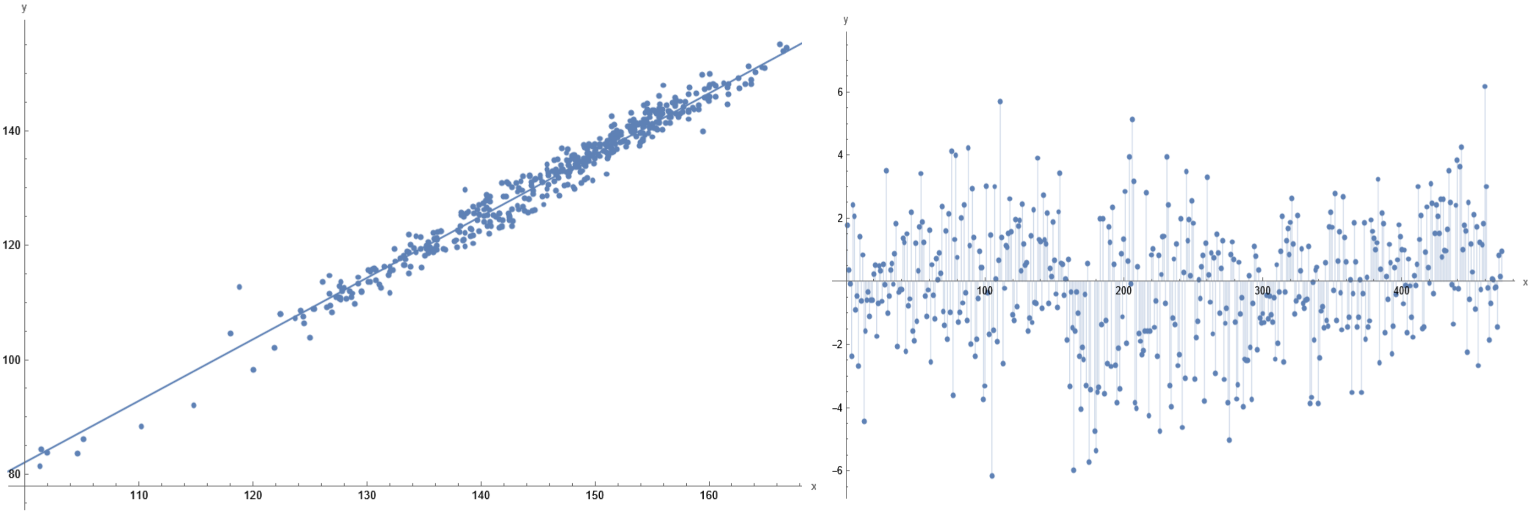


Figure 3. Scatter Plot with Regression Line (Left) and Plot of Residuals (Right) for Pre-molt All vs Post-molt All

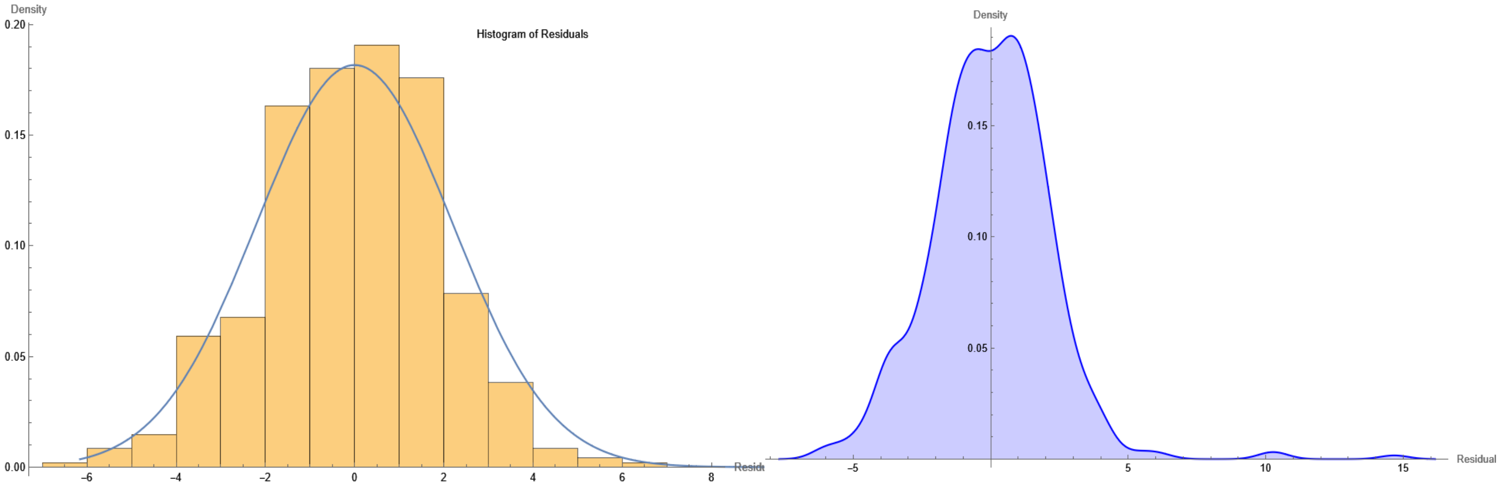


Figure 4. Histogram with Normal Distribution (Left) and Smooth Histogram of Residuals for Pre-molt All vs Post-molt All (Right)

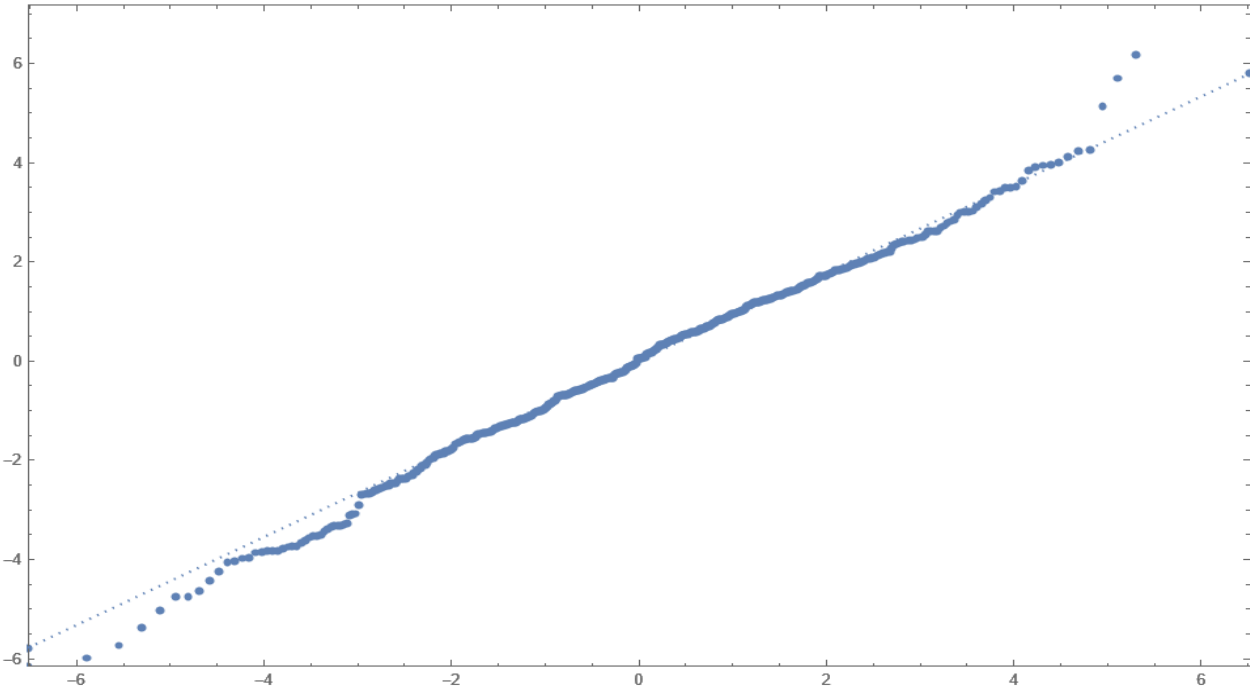


Figure 5. Q-Q Plot of Residuals for Pre-molt All vs Post-molt All

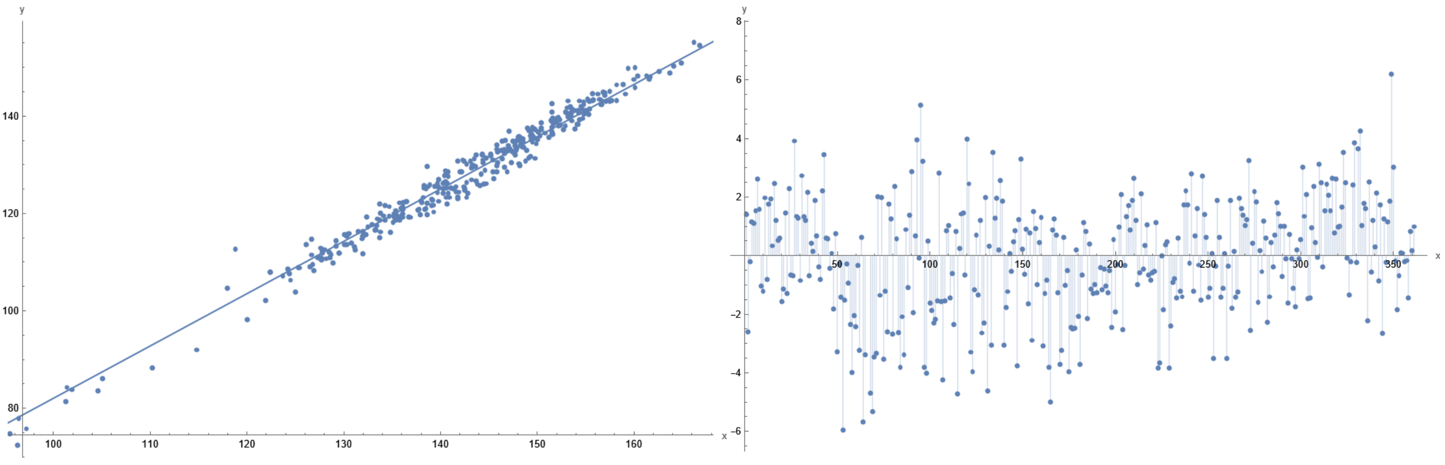


Figure 6. Scatter Plot with Regression Line (Left) and Plot of Residuals (Right) for Pre-molt (Lab-1) vs Post-molt (Lab-1)

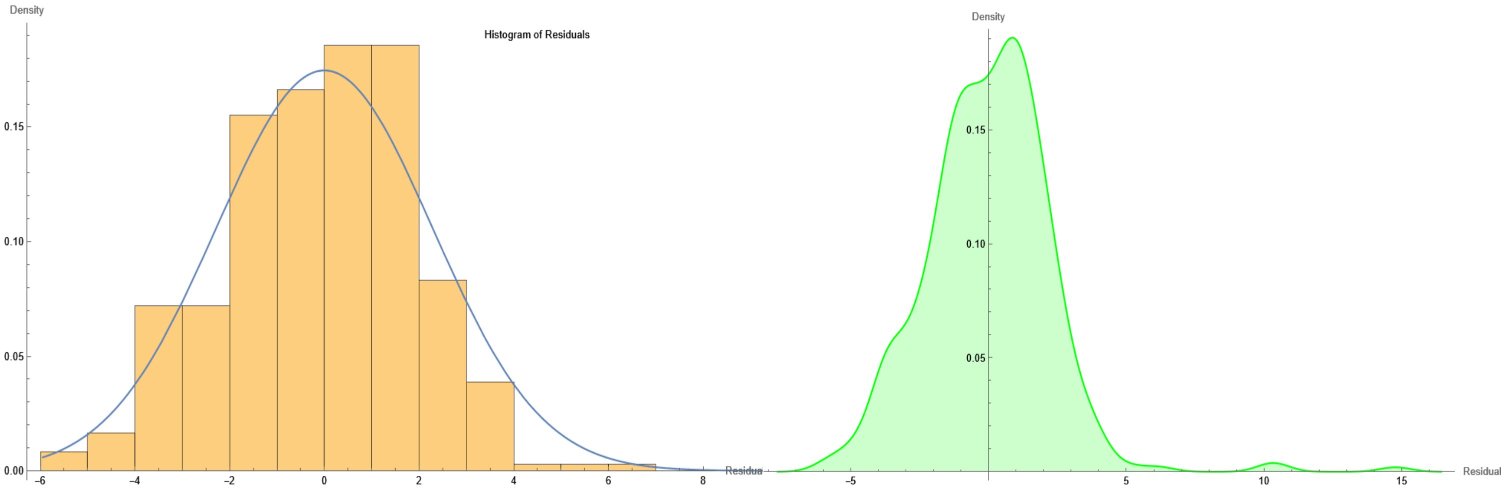


Figure 7. Histogram with Normal Distribution (Left) and Smooth Histogram of Residuals for Pre-molt (Lab-1) vs Post-molt (Lab-1) (Right)

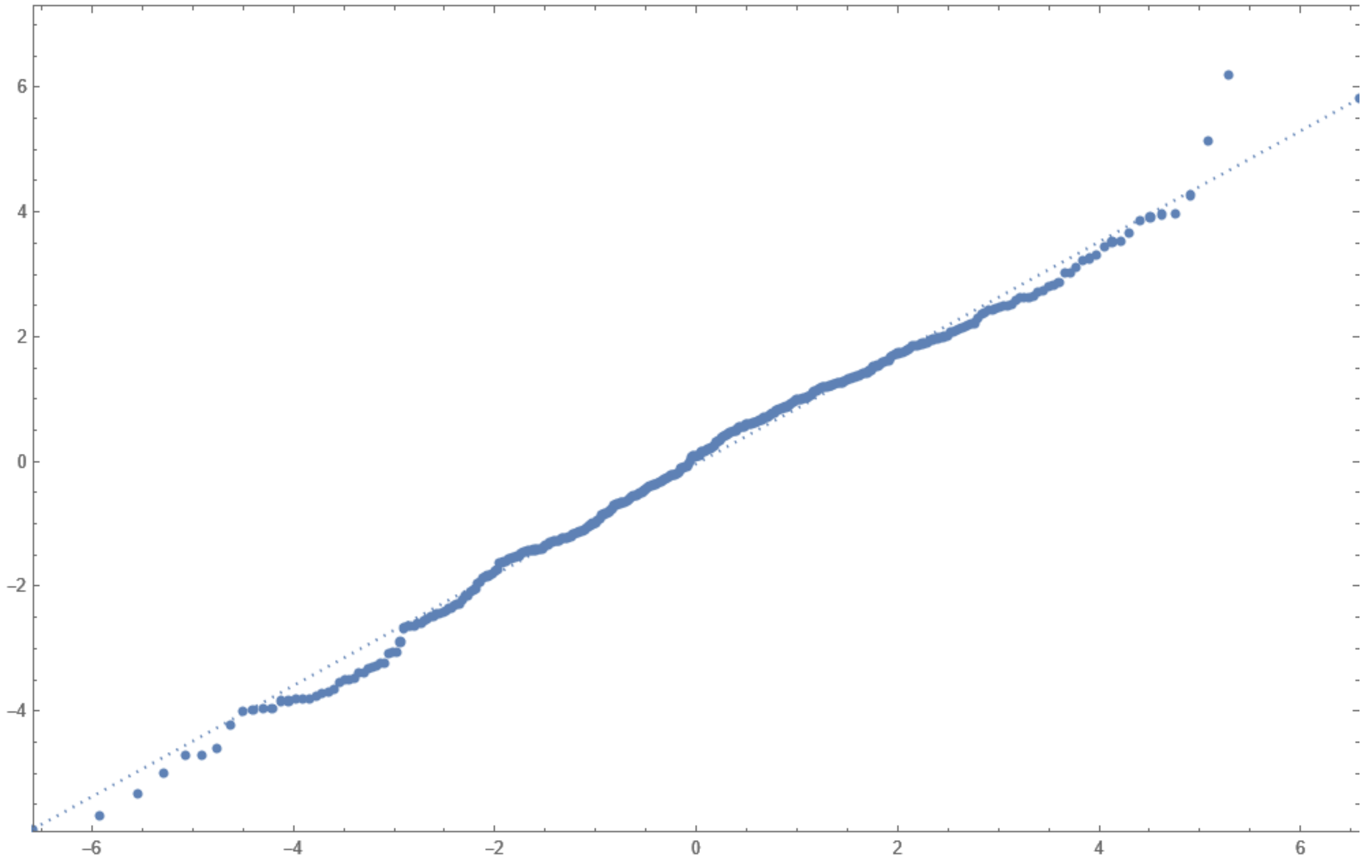


Figure 8. Q-Q Plot of Residuals for Pre-molt (Lab-1) vs Post-molt (Lab-1)

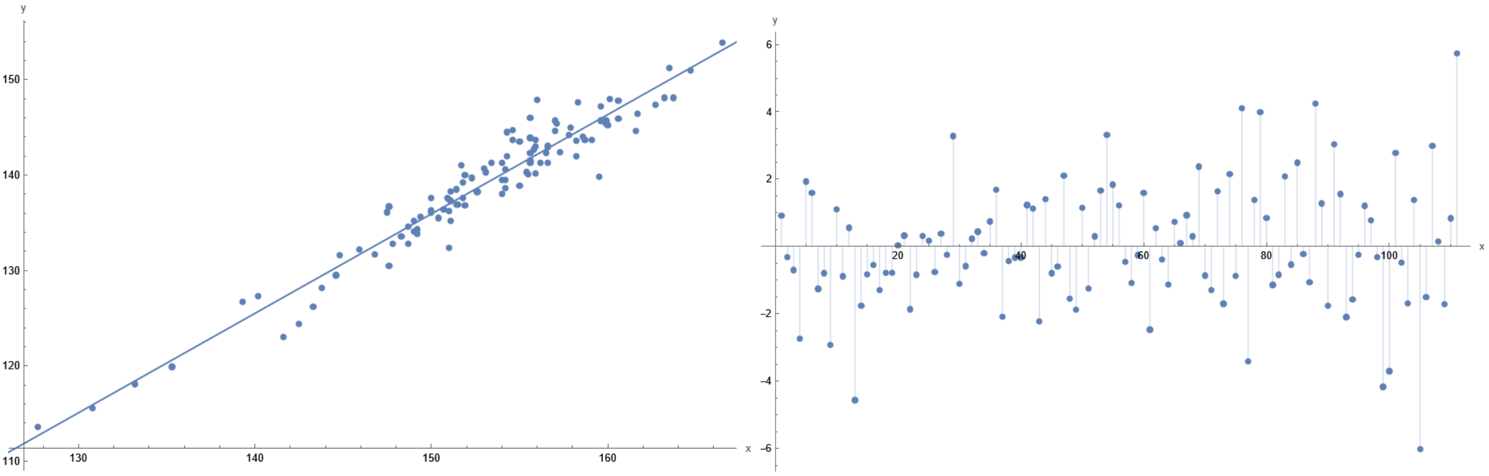


Figure 9. Scatter Plot with Regression Line (Left) and Plot of Residuals (Right) for Pre-molt (Field-0) vs Post-molt (Field-0)

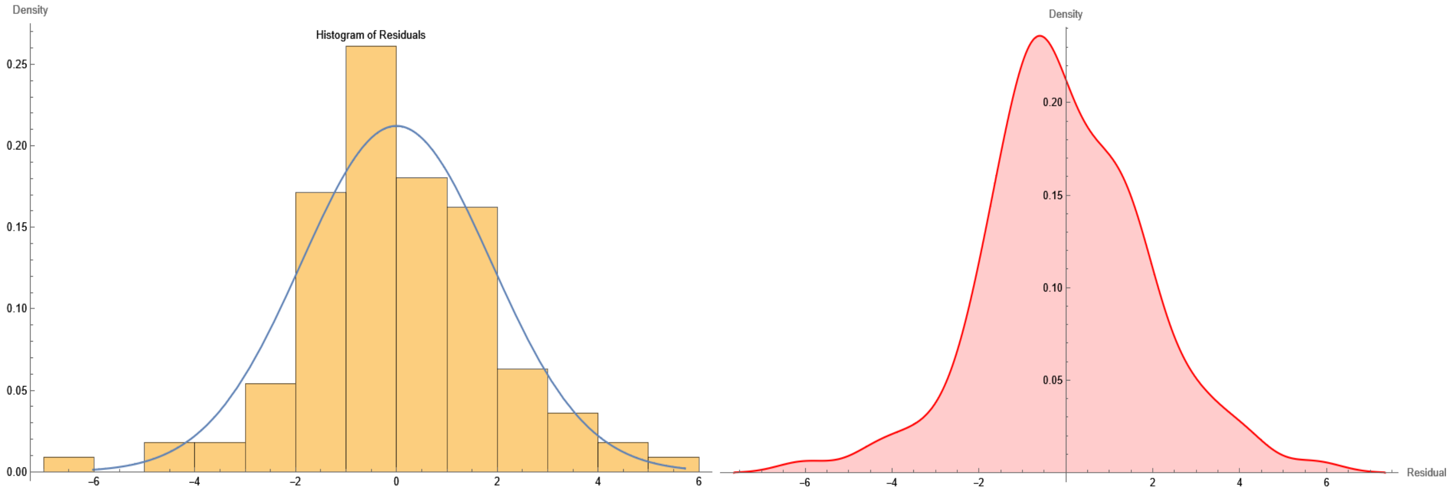


Figure 10. Histogram with Normal Distribution (Left) and Smooth Histogram of Residuals for Pre-molt (Field-0) vs Post-molt (Field-0) (Right)

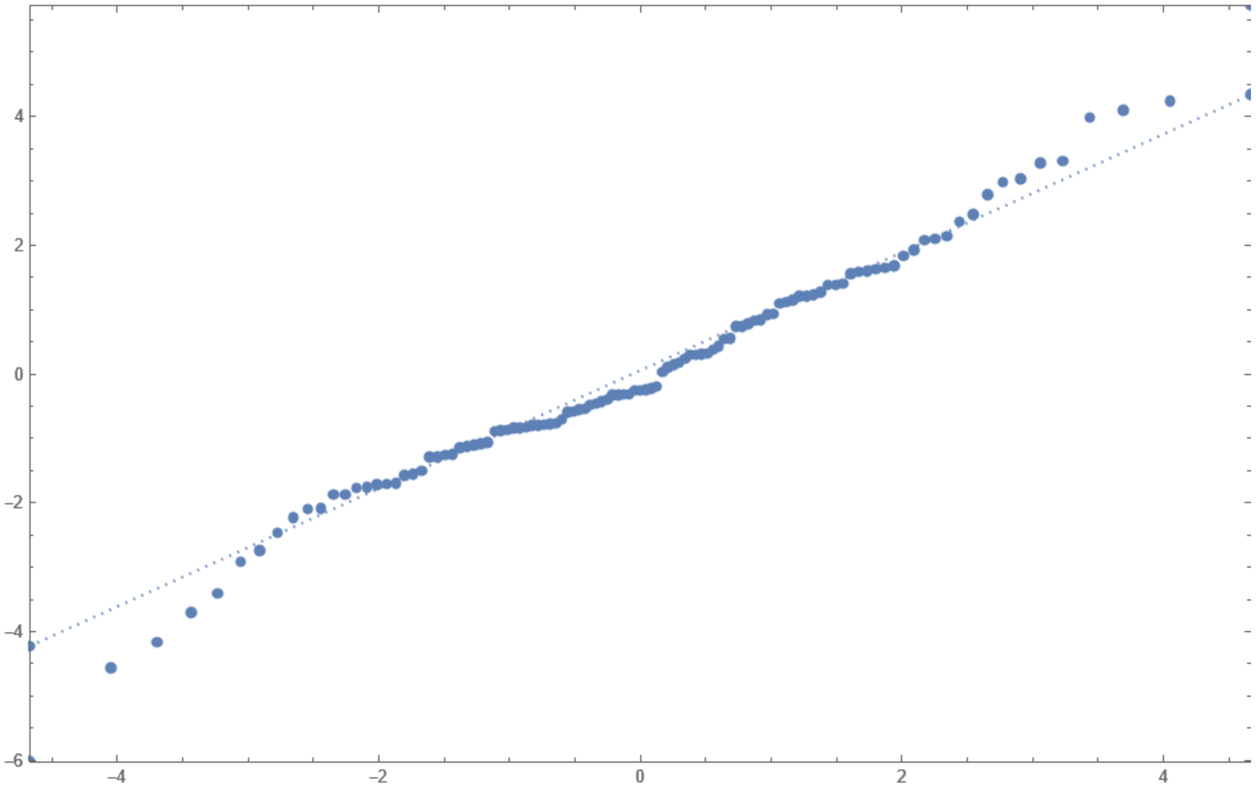


Figure 11. Q-Q Plot of Residual Pre-molt (0-Field) vs Post-molt (0-Field)

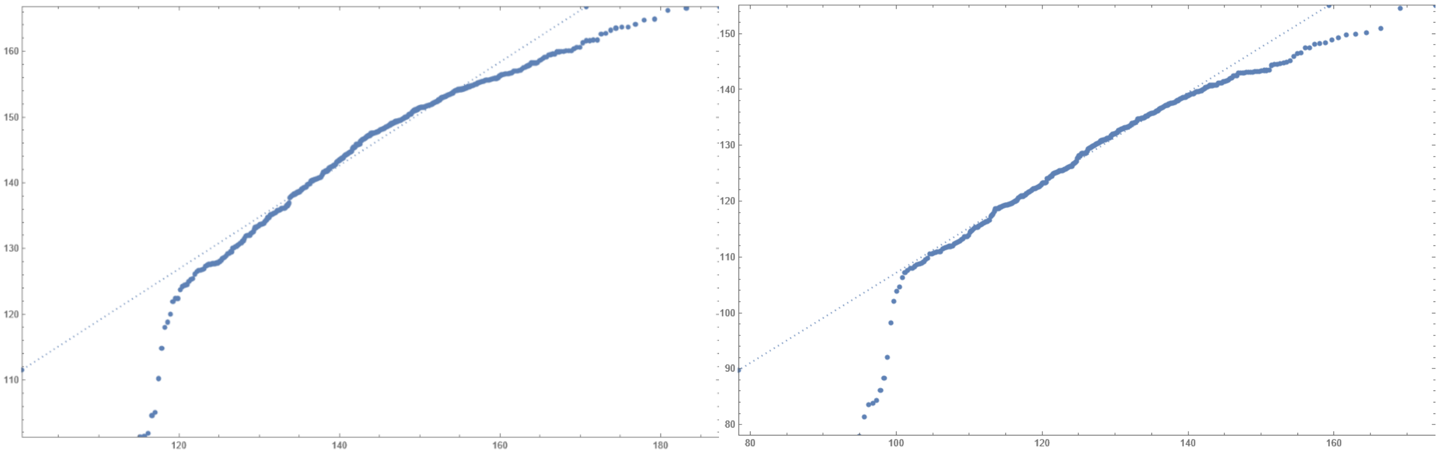


Figure 12. Q-Q Plot for Post-molt All (Left) and Pre-molt All (Right)

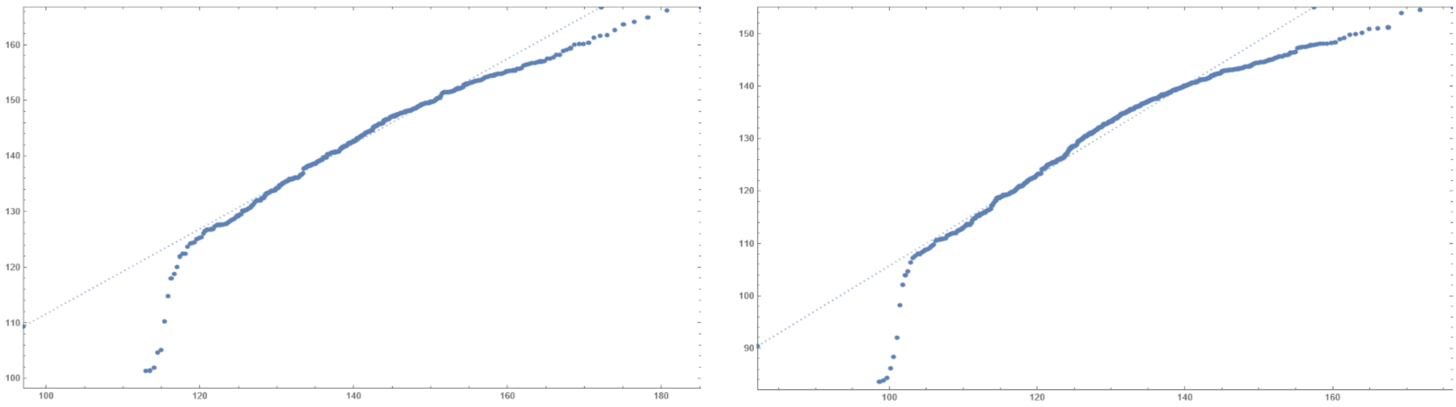


Figure 13. Q-Q Plot for Post-molt Lab-1 (Left) and Pre-molt Lab-1 (Right)

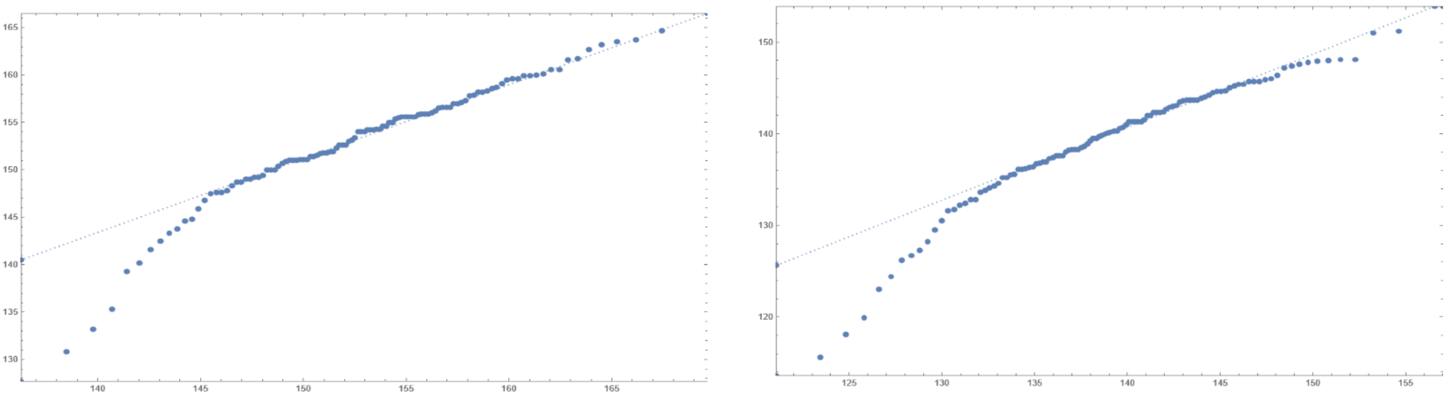


Figure 14. Q-Q Plot for Post-molt Field-0 (Left) and Pre-molt Field-0 (Right)

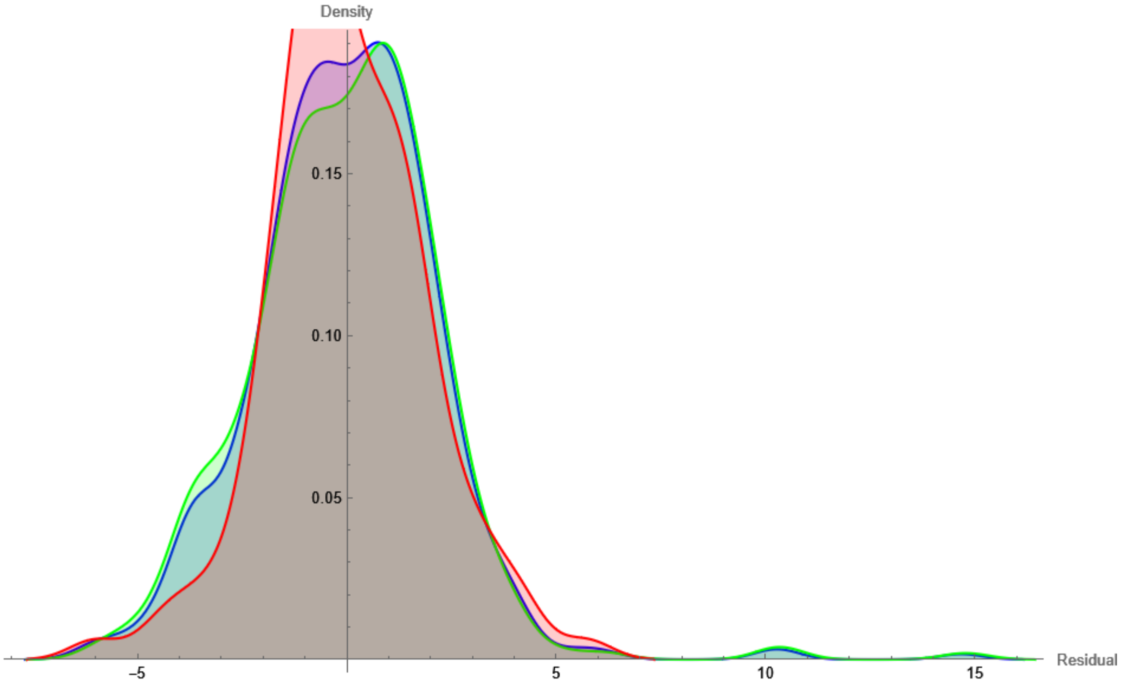


Figure 15. Smooth Histogram of All 3 Group Comparison Residuals

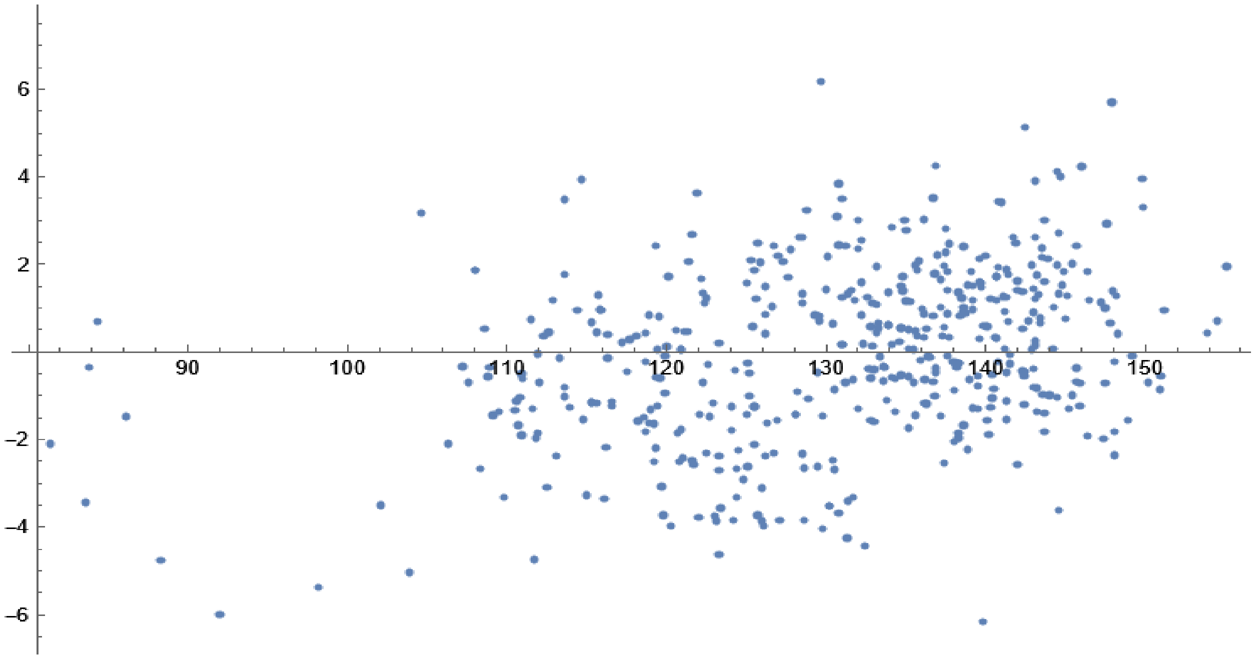


Figure 16. Pre-molt All vs Residuals

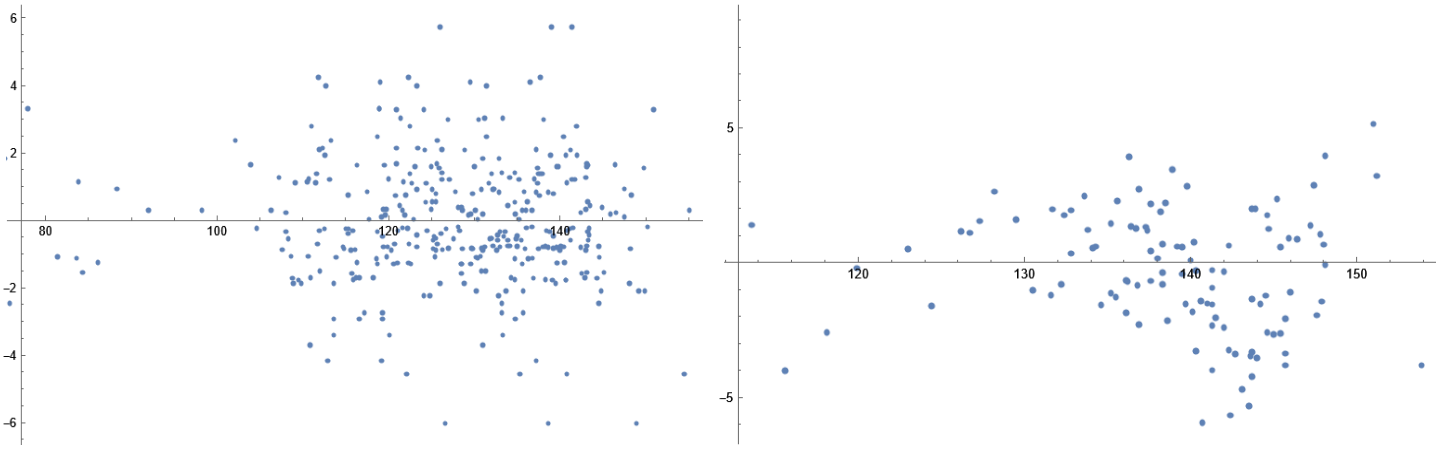
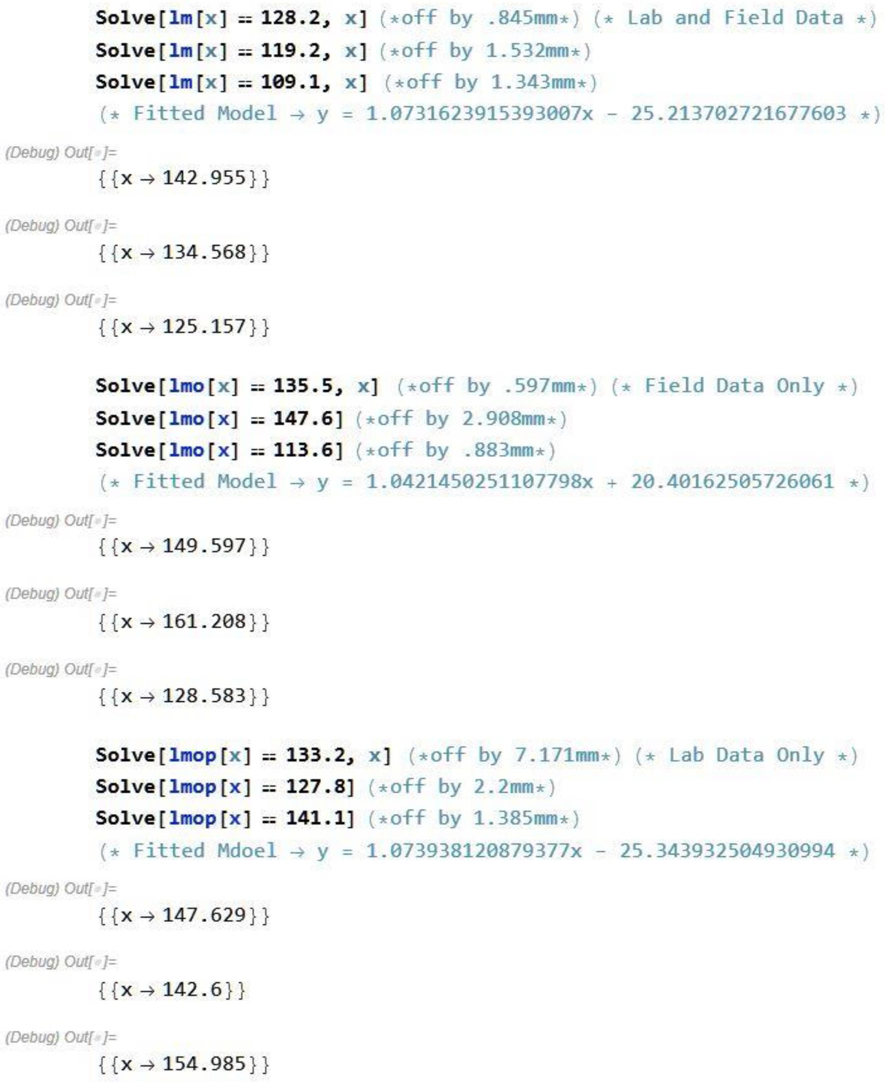


Figure 17. Pre-molt (Lab) vs Residuals (Left) and Pre-molt (Field) vs Residuals (Right)



```
          Solve[lm[x] == 128.2, x] (*off by .845mm*) (* Lab and Field Data *)
          Solve[lm[x] == 119.2, x] (*off by 1.532mm*)
          Solve[lm[x] == 109.1, x]  (*off by 1.343mm*)
          (* Fitted Model → y = 1.0731623915393007x - 25.213702721677603 *)
(Debug) Out[ ]=
          {{x → 142.955}}

(Debug) Out[ ]=
          {{x → 134.568}}

(Debug) Out[ ]=
          {{x → 125.157}}

          Solve[lmo[x] == 135.5, x]  (*off by .597mm*) (* Field Data Only *)
          Solve[lmo[x] == 147.6] (*off by 2.908mm*)
          Solve[lmo[x] == 113.6] (*off by .883mm*)
          (* Fitted Model → y = 1.0421450251107798x + 20.40162505726061 *)
(Debug) Out[ ]=
          {{x → 149.597}}

(Debug) Out[ ]=
          {{x → 161.208}}

(Debug) Out[ ]=
          {{x → 128.583}}

          Solve[lmop[x] == 133.2, x] (*off by 7.171mm*) (* Lab Data Only *)
          Solve[lmop[x] == 127.8] (*off by 2.2mm*)
          Solve[lmop[x] == 141.1] (*off by 1.385mm*)
          (* Fitted Mdoel → y = 1.073938120879377x - 25.343932504930994 *)
(Debug) Out[ ]=
          {{x → 147.629}}

(Debug) Out[ ]=
          {{x → 142.6}}

(Debug) Out[ ]=
          {{x → 154.985}}
```

Figure 18. Estimating with Acquired Linear Model

Discussion & Conclusion

One of the biggest things to notice is that in every Q-Q Plot at the beginning and end of the plotted data points goes off in a non-liner fashion rather than following the theoretical uniform distribution. This tells us that if we want to accurately predict pre-molt sizes either before/near the beginning or at the end/beyond our data it is going to become less accurate than if we predict values between the end points.

When looking to identify heteroscedasticity or homoscedasticity we could look at Figures 3, 6 and 9 and focus on the variance of the residuals but it is quite difficult to examine this from just the residual values themselves. What we can do is look at Figure 16. where we have plotted pre-molt (All) vs the residuals. In Figure 16. there are homoscedastic properties. The reason for this is because most of the residual lies in-between +/-4 and consistently stay there. In Figure 17. when we look at the pre-molt (Lab) and pre-molt (Field) data there tends to have more heteroscedastic characteristics. Though in pre-molt (Lab) it is quite difficult to exactly determine if it would be considered hetro or homoscedastic than in pre-molt (Field) where we can defiantly define that as being heteroscedastic.

An important aspect of how accurate our regression line is by looking at the kurtosis of the residuals. We can see in Figure 2. that the kurtosis values of Residuals (All) and Residuals (Lab) are very large, greater than 8. This is a good thing because it means that most residual values are small, meaning less error between them and the regression line, which is what we are looking for. Though the kurtosis for the Residuals (Field) are not as large as 8, it is still above the normal kurtosis value 3, ours here is 3.91. This is still good as the residual is peakier so more of the residuals are closer to zero.

Looking at figure 18. we can see that most of the randomly selected data points are predicted accurately. As said before when we try to predict pre-molt sizes at the end points of the data set our answers will be more inaccurate which and be seen by the larger differences found such as 7.171mm, 2.908mm and 1.532mm.

The last important measurement we should look at is the RSqaured value for the three groups. For both the All group and the Lab group we have a RSquared of ~.98 and for the Field group ~.93. These tell us that we have a good relationship between our two variables, pre-molt and post-molt.

From what was discovered about our groups through statistical, visual, and analytical methods/measurements it can be concluded that we are able to accurately determine the size of the pre-molt carapaces. However, predicating them from the Lab will give us a more accurate estimation than from the Field. Linear models for our groups were found and proved why they are good fits and how they compare to each other. These models and the process that was used to assess and create them can be used to help restore the Dungeness crab population and halt overfishing by placing restrictions on the specific sizes at which they can be caught at.

References

Nolan, D. A., & Speed, T. (2001). *Stat labs: Mathematical statistics through applications*. New York, NY: Springer.

David G. Hankin, Nancy Diamond, Michael S. Mohr, James Ianelli, Growth and reproductive dynamics of adult female Dungeness crabs (*Cancer magister*) in northern California, *ICES Journal of Marine Science*, Volume 46, Issue 1, 1989, Pages 94-108, https://doi.org/10.1093/icesjms/46.1.94

Appendix

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

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

Proof.

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

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

Let

$$\frac{\partial S(a,b)}{\partial a} = 0$$

$$\frac{\partial S(a,b)}{\partial b} = 0$$

$$0 = 2\sum_{i=1}^{n} (y_i - (ax_i + b))(-1)$$

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

$$= \sum_{i=1}^{n} (-2y_i + 2ax_i) + 2nb$$

$$-2nb = \sum_{i=1}^{n} (-2y_i + 2ax_i)$$

$$nb = \sum_{i=1}^{n} (y_i - ax_i)$$

$$b = \frac{\sum_{i=1}^{n} (y_i)}{n} - a\frac{\sum_{i=1}^{n} (x_i)}{n}$$

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

Next

$$0 = 2\sum_{i=1}^{n}(y_i - (ax_i + b))(-x_i)$$

$$= 2\sum_{i=1}^{n}(-x_iy_i + ax_i^2 + bx_i)$$

$$= -2\sum_{i=1}^{n}(x_iy_i) + 2a\sum_{i=1}^{n}(x_i^2) + 2b\sum_{i=1}^{n}(x_i)$$

$$= -2\sum_{i=1}^{n}(x_iy_i) + 2(\overline{y} - a\overline{x})\sum_{i=1}^{n}(x_i) + 2a\sum_{i=1}^{n}(x_i^2)$$

$$= -2\sum_{i=1}^{n}(x_iy_i) + 2\overline{y}\sum_{i=1}^{n}(x_i) - 2a\overline{x}\sum_{i=1}^{n}(x_i) + 2a\sum_{i=1}^{n}(x_i^2)$$

$$= -2\sum_{i=1}^{n}(x_iy_i) + 2\frac{\sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i)}{n} - 2a\frac{\sum_{i=1}^{n}(x_i)\sum_{i=1}^{n}(x_i)}{n} + 2a\sum_{i=1}^{n}(x_i^2)$$

$$= -\sum_{i=1}^{n}(x_iy_i) + \frac{\sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i)}{n} - a\frac{\sum_{i=1}^{n}(x_i)\sum_{i=1}^{n}(x_i)}{n} + a\sum_{i=1}^{n}(x_i^2)$$

$$= -n\sum_{i=1}^{n}(x_iy_i) + \sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i) - a\sum_{i=1}^{n}(x_i)\sum_{i=1}^{n}(x_i) + an\sum_{i=1}^{n}(x_i^2)$$

$$an\sum_{i=1}^{n}(x_i^2) - a(\sum_{i=1}^{n}(x_i))^2 = n\sum_{i=1}^{n}(x_iy_i) - \sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i)$$

$$a(n\sum_{i=1}^{n}(x_i^2) - (\sum_{i=1}^{n}(x_i))^2) = n\sum_{i=1}^{n}(x_iy_i) - \sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i)$$

$$a = \frac{n\sum_{i=1}^{n}(x_iy_i) - \sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i)}{n\sum_{i=1}^{n}(x_i^2) - (\sum_{i=1}^{n}(x_i))^2}$$

$$a = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\sum_{i=1}^{n}(y_i)\sum_{i=1}^{n}(x_i)}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n}(x_i))^2}$$

□