Pre-Molt vs Post-Molt Crab Sizes

Nick Collins

UMass Dartmouth

February 13, 2022

Abstract-

In recent years there have been efforts to raise and breed seafood in laboratories. It raises the question of whether these laboratory crabs are the same as the wild crabs. The data from the Stats Lab Data Site gave us three major columns to look at. Pre-molt size, post-molt size and the measurement sight. For the analysis we need to split the data based on the measurement site and then plot a regression line using the post-molt size (independent variable) and pre-molt size (dependent variable). Using this regression line and scatter plots we want to predict the post-molt sizes and see if there are any differences in the laboratory and wild caught crab sizes.

Introduction-

The data used to perform this analysis was crab shell sizes of pre-molt versus post-molt sizes in millimeters. Through some analysis we want to see if there is a statistically significant difference in the sizes of the crabs based on where they were caught and measured. People say that seafood is usually better when caught in the wild, but recently efforts have been made to make the laboratory grow crabs bigger and better. In this analysis we will see if there is any difference between the two groups and see some predictions using the regression line.

Methods-

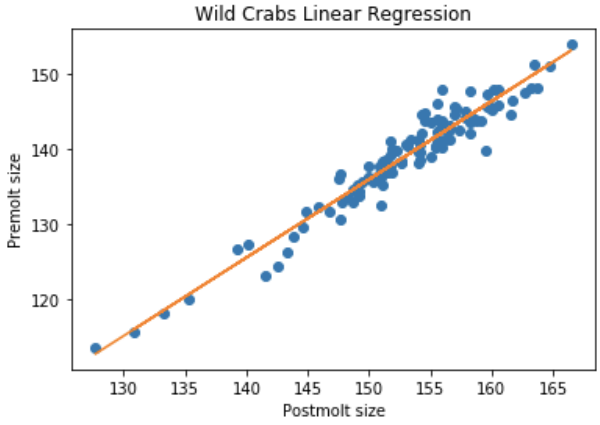
The data was found on the Stats Lab Data Site and consisted of five columns and 472 rows. When I first downloaded this data, it was condensed into one column with 472 rows. After downloading the data, I imported it into excel and separated all the values into their own columns. After this I had five columns and 472 rows. I dropped two rows where the values were very low compared to the other data values. I feel this will give us a truer representation of the data and the plots produced later will be more focused on the data we want to see and the axis ranges won't be as large.

Although the data has five columns, we only need to look at three of the five. The main two columns that have the sizes of the crabs are "presz" and "postsz". These two columns have the sizes of the pre-molt and post-molt crab shells. The last column that we will look at is the column called "lf". This is the measurement site and where it took place. The values for this column are either one or zero, where 0 = field, and 1 = lab. This column lets us sort the data by where the measurement took place. To group the data, I used the get group method in pandas. Pandas was mainly used to sort and organize the data into two data frames and then made into one large data frame to get a plot for the whole dataset.

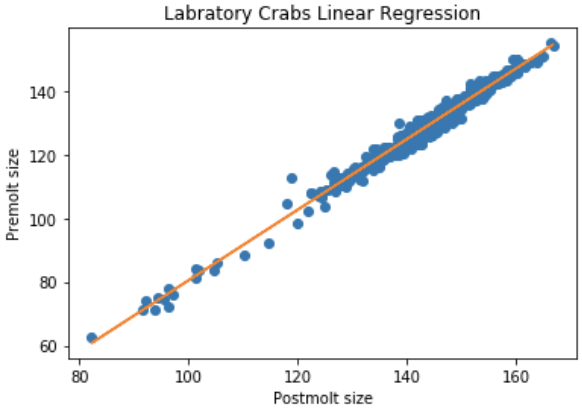
After splitting the data into two groups it was easy getting the plots I needed. The only other step I had to take to clean the data was to turn the values in "presz" and "postsz" into floats because they were strings when first imported but this was easily solved with two lines of code. In regard to plotting, I used the two libraries matplotlib and seaborn to plot the scatterplot, regression line and residuals.

Results-

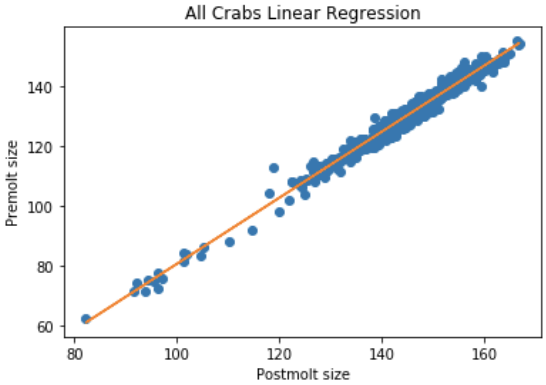
The main objective of this assignment was to plot linear regression for the lab measured, field measured and all the crabs measurements in both sites. The first linear regression model I had to output was for the wild crab measurements. In this plot there were 111 points out of the 470 total



crab measurements. The next plot was the remaining 359 points of the 470 which was the plot for the laboratory crabs.

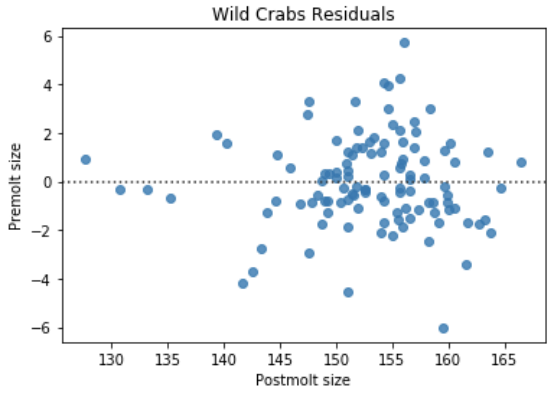


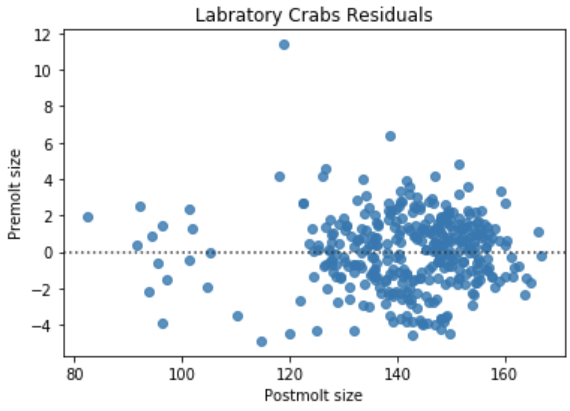
Lastly, we have all 470 points in the same scatter plot with a regression line.

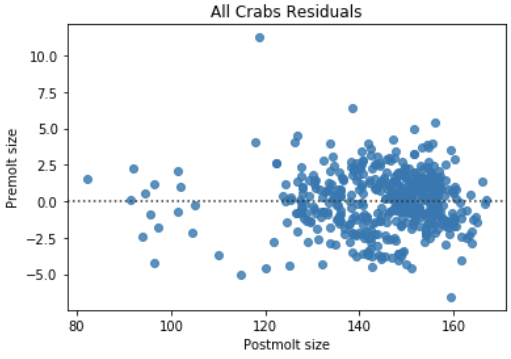


These three plots show that there is a strong relationship between the post-molt on the x-axis and the y-axis being the pre-molt size. The wild crabs seem to have a higher size on average. The range where wild crabs

are found the most are (145,130) to (160,150) and the laboratory crabs range the most from (120,110) to (160,140).

The next objective of the assignment was to plot the residuals for all three plots. The residuals for the wild crabs were 

and the residuals for the laboratory crabs were 

. Finally, the last residual plot was the residuals of all the data. 

The residuals for wild crabs appear to be more spread out than the residuals for the laboratory crabs. This could be the result of more data for the laboratory crabs because there are more than three times as many points.

Conclusion-

After performing the analysis on the crab data, we got some interesting outputs. Like I said earlier, the range where wild crabs are found the most are (145,130) to (160,150) and the laboratory crabs range the most from (120,110) to (160,140). This was interesting to see that the wild crabs had a much higher range and averages for post-molt and pre-molt sizes. The regression line for these plots are also different. The wild crabs' regression line has a slope of 1.042 and a y intercept of -20.40. The laboratory crabs' regression line has a slope of 1.11 and a y intercept of -30.69.

The residual plots give us some understanding of what exactly these crab shell sizes are doing. Residuals measure how far our actual values are from the expected value based off the regression line. For the regression line to be a "good fit" we want the residuals to be close to zero. In these three plots we see the majority of the residuals are pretty much around zero. There are some residuals that aren't as close to very as we'd like, but that is expected with any dataset. The wild crabs residuals are more spread out than the laboratory residuals and I think this is a sign that we might need some more data to make a better regression line. Since there are only a little over 110 data points for wild caught crabs compared to 360 data points for the laboratory crabs. If the wild crabs had as many points as the laboratory, I think we would have gotten some better residuals and a better regression line.

Overall, we did get good regression lines and residuals for the data we were given. We see the wild crabs being a little larger on average than the laboratory crabs. With the laboratory crabs we see around 15-20 crabs in the post-molt size range of 80-120, while we see the wild crabs lowest post-molt size around 125. If we got more data for wild crabs, we might expect to see some more outliers. However, our regression lines would do a good job predicting pre-molt sizes based off the post-molt sizes of the crabs.

References-

"Dungeness Crab Growth:" *Stat Labs Data Page*, https://www.stat.berkeley.edu/~statlabs/labs.html#crabs

Appendix-

Problem 9, page 154

$$\sum (y_i - (ax_i + b))^2$$

After taking the derivative we get:

$$2\sum y_i - (ax_i + b)$$

Simplifying the equation:

$(\sum x_i)a + (\sum 1)b = 2\sum y_i$