Predicting Crab Pre-Molt Size from Post-Molt Size with Simple Linear Regression

Issues

This study examines the application of a linear regression model to the pre- and post-molt sizes of crabs; the goal of a linear regression model is to predict the pre-molt size using the post-molt size. We also go over data interpretation to assess whether or not crab post-molt sizes may accurately predict crab pre-molt sizes. There was some mistake in the pre-molt size forecasts that we had to quantify and assess to see whether it would make the predictions risky to utilize owing to inaccuracy while developing the ability to estimate pre-molt sizes from post molt sizes. We also consider if we can only forecast the pre-molt sizes more precisely in specific places where we have more data, given the restricted quantity of data provided. Following the completion of these procedures, we are left with our conclusions regarding the effectiveness and dependability of applying post-molt sizing to generate or predict crab pre-molt sizes.

Findings

Considering the extremely high r-squared values, the data we have seems to function reasonably well in forecasting pre molt sizes from post molt sizes. Given that the inaccuracy is typically higher in the lower ranges of post molt sizes, care should be used if using this model to forecast pre molt sizes.

If wanting to improve this prediction model, more data on crabs molting sizes containing post molt sizes that are lower in the range could help improve the error of the prediction in the lower ranges, improving the model overall.

Discussion

We came to the conclusion that the model had more inaccuracy when forecasting in the lower ranges of post-molt sizes after completing general statistics on the crab molting data. This is related to our problem with the inadequate data and, thus, the poor accuracy across the post-molt sizing ranges. We can also confirm that the model is not dangerous due to these inaccuracies, but instead, we give caution about the higher error when using post molt sizes in the lower ranges. Since there is less accuracy in the lower ranges than particularly in the higher ranges of post molt sizing, between 125-160 roughly, this is less error and better predictions when giving post molt sizes in this range. The higher accuracy is due to most of the post molt sizes being in this range, therefore, the prediction has a better understanding of what the output pre molt sizes should be. Due to the better accuracy in the higher parts of the post molt range it proves that post molt size data can indeed predict pre molt sizes of crabs well within some bounds of tolerance or error.

Appendix A: Methods

This paper uses a dataset of crab molt sizes, two variables, pre-molt size, and post molt size with the goal of using post molt sizes of crabs to predict pre-molt sizing. The next steps were to evaluate the data by obtaining summaries of each variable, post molt and pre molt,

giving us minimum, 1st quartile, median, mean, 3rd quartile, and maximum. Next was to test whether the data for each variable followed a normal distribution; there are many ways to test this, including checking the kurtosis and skewness of each variable to see they are 3 and 0, respectively, or close to it. then by examining the density plots and histograms of each variable. The check for normality was finished after these three tests. We can see the difference or shift between the data by overlaying the density plots on top of each other and drawing two vertical lines for the mean of post-molt and pre-molt sizes. The linear model can then be constructed using a linear regression function in R, with post-molt sizing serving as our predictor for pre-molt sizes. The Pearson's r-squared value of the model, which contributes to explaining how well the post-molt sizing contributes to predicting the pre-molt size, can be obtained by looking at the plot of the line generated by the function from R on top of the data in order to determine the model's fit. We then compute the difference between the original pre-molt values in the data and the ones that our model predicted from the post-molt sizes in order to examine residuals or errors in the prediction. The next step is to determine whether the residuals or errors have a normal distribution; This is accomplished by running a quantile or q-q plot in R, which draws a line through a scatter plot of all the residuals. The residuals would not be normal if none of the points were on the line. We measure the residuals' kurtosis and skewness to fully confirm this, and if they are equal to 3 and 0, respectively, then the residuals are normal; Aside from that, they may or may not be out of the ordinary. We plotted the density plot over the histogram and the histogram of the crab residuals to comprehend the kurtosis of the data; First, look to see if the residuals quickly reach a steep peak or if the histogram has a long tail. The value of kurtosis in the residuals could be influenced by these. Last but not least, we determine whether the residuals exhibit heteroscedastic behavior by plotting them on the y-axis and making the x-axis the predictor variable—in this case, the post-molt sizes. There is heteroscedasticity in the residuals if there are visible clusters, conical shapes, which indicate that there are more points as you go right on the x-axis, or a discernible pattern in the plotted residuals. A conclusion regarding the model's accuracy and potential applications can be reached after these statistics have been completed.

Appendix B: Results

First, we will start off with the summary of both variables in the data set,

Post Molt Sizes:

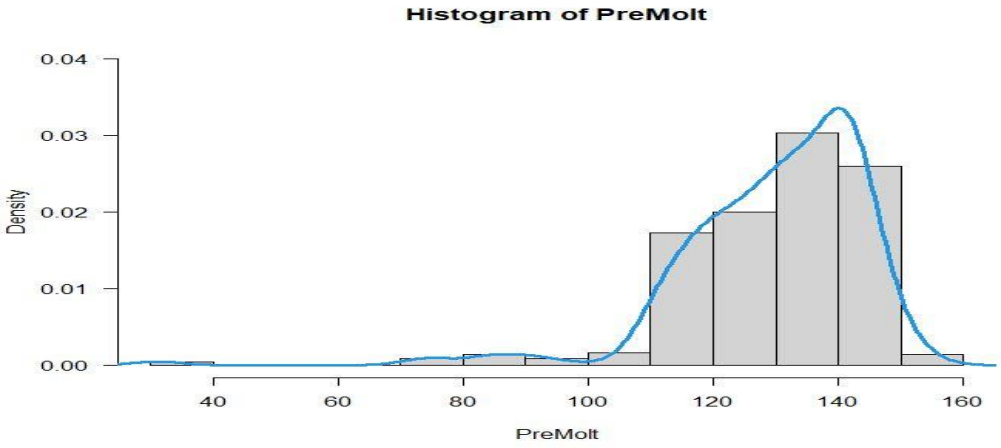
Min.	1st Qu	Median	Mean	3rd Qu	Max.
38.8	136.3	146.8	143.3	152.5	166.8

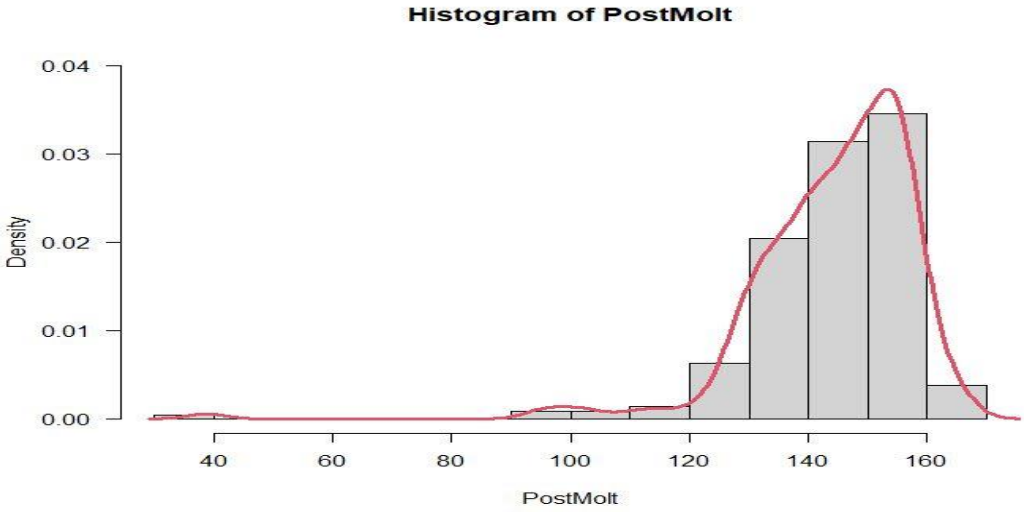
Pre-Molt Sizes:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
31.1	120.1	132.3	128.5	139.2	155.1

Kurtosis:	16.32761	10.87427
Skewness:	-2.589865	-2.016654

The fact that the Kurtosis and Skewness do not correspond to 3 and 0 immediately indicates that neither of these variables follows a normal distribution.
The following are two histograms alongside two thickness plots those on the left compare to the post shed variable and on the right relate to pre shed. These density and histogram plots, in addition to the Kurtosis and Skewness, demonstrate that the data do not follow a normal distribution. For comparison, a typical normal distribution histogram and density are provided below.)

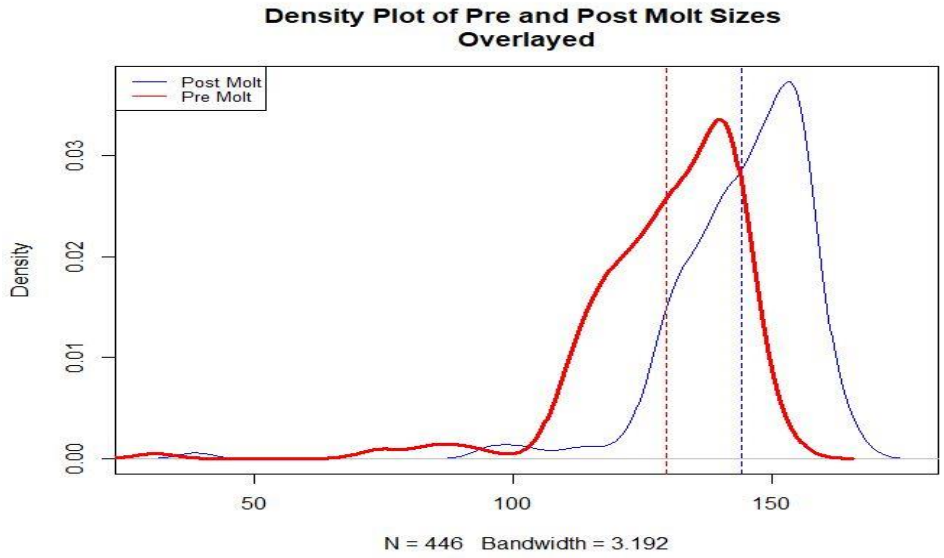




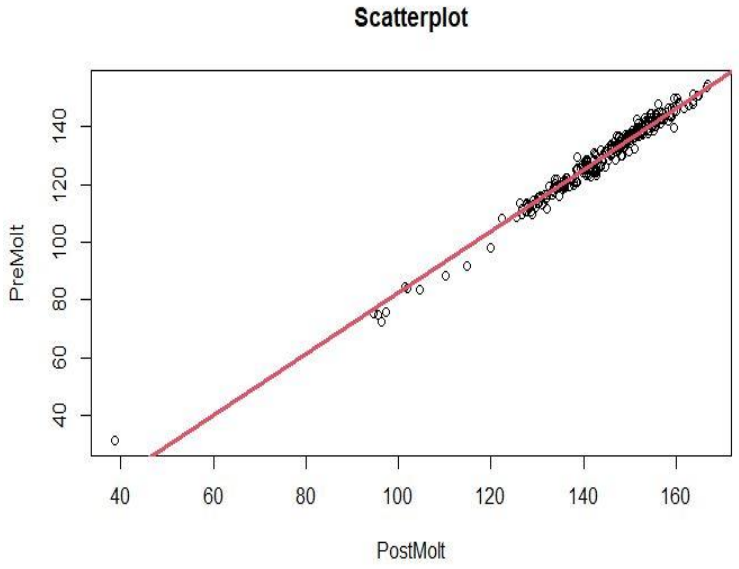
After looking at the example normal distribution you can see that the variables do not follow a normal distribution.

Example Normal Distribution

The next plot shows the difference between the data by overlaying the density plot for post-molt and pre-molt sizes and plotting a line for each variable's mean. On the molt change, the differences between the means are 143.3 - 128.5 = 14.8. The mean of the corresponding variables is represented by the colored dotted lines, with the red line representing the pre-molt mean.)



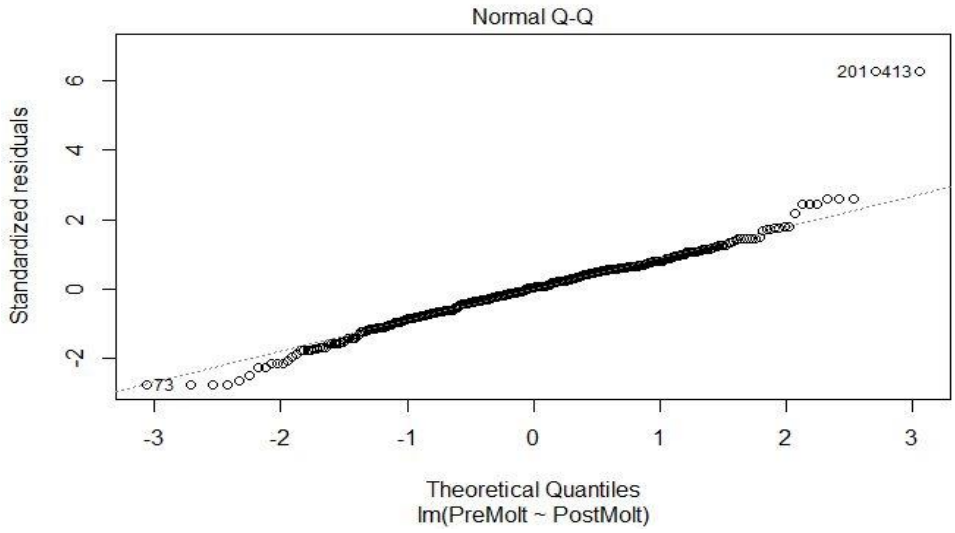
The post-molt data are plotted as the x-axis, and the pre-molt data are plotted as the y-axis. The linear regression function's line is then placed below the plot. A summary of significant values from the linear regression summary can be found on the right side.



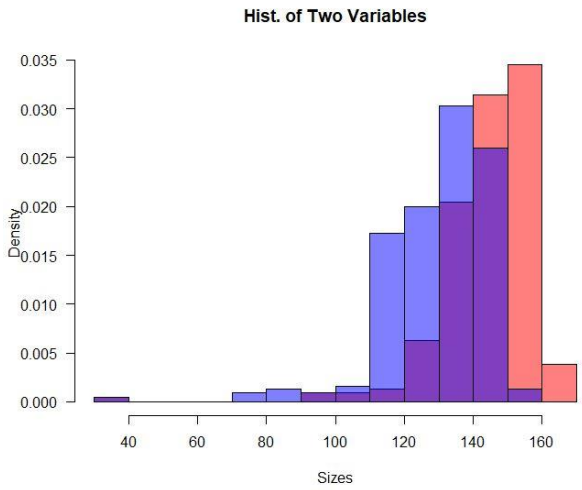
Slope:	1.071465
Intercept:	-25.084758
R-Squared:	0.9796 or 97.96%
Pearson's R-Squared:	0.9897423 Almost 99%
P-Value:	< (less than) 2.2e-16
F-Statistic:	2.093e+04

Summary

A Q-Q or quantile plot, which is a test for normality among the residuals, is created from the residuals from the created linear model as shown below. The residuals have a normal distribution if all points are on the line.



A histogram with a density plot overlayed and a summary of the residuals on the right side are provided to confirm this non-normality. The plot with the long tail may be affecting the residuals' kurtosis, which is otherwise close to normal but not normal.

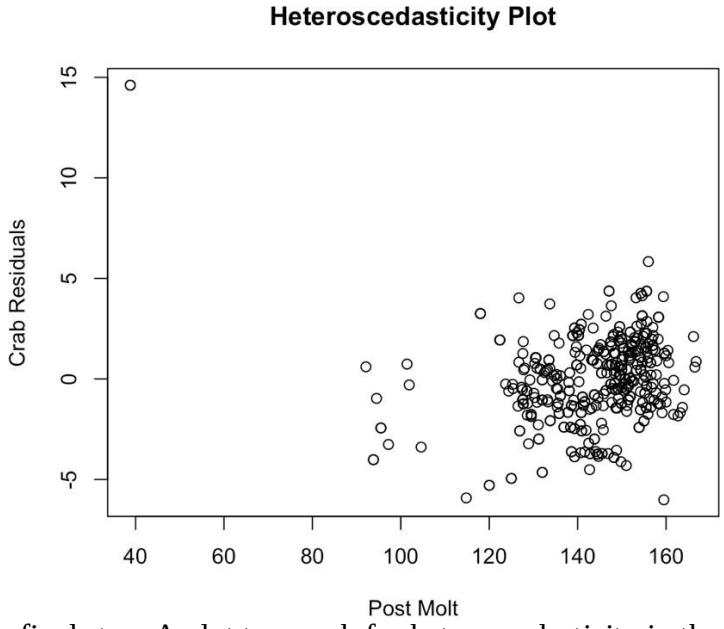


Summary

Min.	1st Qu	Median	Mean	3rd Qu	Max.
31.1	120.1	132.3	128.5	139.2	155.1

Kurtosis:	10.73385
Skewness:	1.075962

Plotting the residuals against the predictor variable, post molt size—with post molt as the x-axis and residuals



as the y-axis—is the final step. A plot to search for heteroscedasticity in the prediction model is shown here. This demonstrates that the model performs better at the higher end of the post-molt range.

Appendix C: Data & Code

Data: Pre-molt Sizes and Post-molt Sizes. 2 columns, 416 rows.

Background for the data used can be found here: Stat Labs: Mathematical Statistics Through Applications Chapter 7, pgs. 139 – 150

R-Studio Code:

```
library(readxl)
crab_molt_data <- read_excel("C:/Users/Sai Krishna chikkala /Desktop/MTH /crab_molt_data.xls")
View(crab_molt_data)
attach(crab_molt_data)

library(moments) # To import the skewness and kurtosis function.
#Now we have two variables in the Data Set  i.e. PostMolt and PreMolt and we have to describe these two variables
#Let's start with Post-molt
min (Post-molt)
max (Post-molt)
median(Post-molt)
mean(Post-molt)
sd(Post-molt)
skewness (Post-molt)
kurtosis (Post-molt)

#Pre-molt

min (Pre-molt)
max (Pre-molt)
median(Pre-molt)
mean(Pre-molt)
sd (Pre-molt)
skewness (Pre-molt)
kurtosis(Pre-molt)

#Now we have to make a Probability Density Function(PDF) histogram for each variable #In the histogram plot , the Y axis will be represented by the frequency and we want the density function, So we will replace F with density function by typing "freq=F" #Lets begin with PostMolt

hist(PostMolt, freq=F, las=1,ylim=c(0,0.040),col="red")

#Now the histogram plot of Pre-molt

hist (Pre-molt, freq=F,las=1,ylim=c(0,0.040),col ='blue')
```

#Let’s find the density of the Pre-molt and PostMolt variables

lines(density (PostMolt),col=“red”,lwd=3) lines(density(Pre-molt),col=“blue”,lwd=3)

#Now we will overlap the two histograms in such a way that the difference in the distribution would be visible by naked eye

hist (PostMolt, freq=F,ylim=c(0,0.040),main=“Overlapping between PostMolt and Pre-molt”, xlabel=“Sizes”, Col=rgb(1,0,0,0.5),las=1) hist(Pre-molt, freq=F,add=TRUE, col=rgb(0,0,1,0.5))

#Now we do the density plot for the overlapping of two variables

plot(density (PostMolt),col=“red”,lwd=3,main=“Density Plots of PostMolt& Pre-molt”) lines(density(Pre-molt),col=“blue”,lwd=3)

#In this step we will plot the dependent variable(Pre-molt) as a function of independent variable(PostMolt) with the help of Scatter Plot

plot (PostMolt, Pre-molt, main= “ScatterPlot”)

#Now we must plot the least square linear regression on the same plot as the data

model <- lm (Pre-molt ~ PostMolt) summary(model) abline (model,col=“darkorange”, lwd =3)

#Now we calculate find the Pearsons r^2 regression

results <- cor.test (Pre-molt, PostMolt, method= “pearson”) results

#Let’s do the descriptive statistics for the residuals

residuals <- model$residuals
sapply (residuals, sum)

#Plotting the residuals in the histogram plot

hist (residuals, freq=F,las=1,col=“green” ,ylim=c(0,0.20))

#Plotting the density line for the residuals

```
plot(density(residuals), col= “green” ,lwd=3,ylim =c(0,0.20),main=“Density Plot of Residuals”)
lines(density(residuals),col=“green”, lwd=3)

#Quantile Plot of residuals to check the normality

qqnorm (residuals, pch=1,frame=FALSE, main=“Quantile Plot of residuals”)
qqline (residuals, col= “steelblue”, lwd=2)

#Performing Shapiro-Walks Test

shapiro.test((residuals))

#Plot the residuals against the dependent variable (Pre-molt)

plot (residuals, Pre-molt, main = “ScatterPlot”)
r_model <- lm (Pre-molt ~residuals) summary(r_model) abline(r_model, col=“brown”,lwd=3)
plot(r_model)

# After plotting residuals we can note that most points do indeed lie on the line
```

```
# but, there are some points that do not lie on the line telling us that it is not
# normal.


# Looking at the Kurtosis and Skewness we can see the kurtosis is very high
# for the residuals and for a normal distribution it would be 3 and the skewness
# would be 0. So we can see that we definitely dont have a normal distribution on
# our residuals.
kurtosis(crab_residuals) # 10.73385
skewness(crab_residuals) # 1.075962

# We can see we have a pretty long tail on the residuals and the data is peaky
# around the mean in the histogram which are probably the causes of such a high
# kurtosis.
hist(crab_residuals, col="steelblue", main="Crab Molting Data Residuals", probability=TRUE)
lines(density(crab_residuals), lwd=2, col="black")



# We are going to perform a Breusch-Pagan test to check for heteroscedastic
# behavior in the Residuals.
bptest(crabMolt_LinearModel)

# Plotting the residuals against the dependent variable or predictor to look at
# heteroscedasticity of our model. Even here in this plot we can see the heteroscedastic
# behavior telling showing us the model will perform better around the the higher
# ranges and worse for the lower ranges with the Post Molt data.
plot(crabMoltData$Post.molt, crab_residuals, xlab="Post Molt", ylab="Crab Residuals",
    main="Heteroscedasticity Plot")

# Due to the p-value is less than the alpha value (0.05) which means that the
# residuals shows heteroscedastic behavior and reject homoscedasticity among
# the residuals. The possible cause could be because at the lower end of the
# model our prediction is worse given we have very minimal data there.
```

References

Nolan, D., & Speed, T. (n.d.). *Mathematical Statistics Through Applications*. Retrieved February 16, 2023, from https://mth332.files.wordpress.com/2022/12/stat-labs-book.pdf