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

Issues

The implementation of a linear regression model on pre- and post-molt crab sizes is the subject of this research. To put it simply, the pre- and post-molt sizes of crabs are predicted using the post-molt sizes. 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 examine if, given the limited data available, we can only forecast pre-molt sizes more precisely in those places where we have more data. We are then left with our conclusions regarding the effectiveness and dependability of employing post-molt sizing to generate or anticipate crab pre-molt sizes after completing these steps.

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.

The accuracy of the forecast in the lower ranges might be improved by adding more data on crabs molting sizes that include post-molt sizes that are lower in the range. This would improve the model as a whole.

Discussion

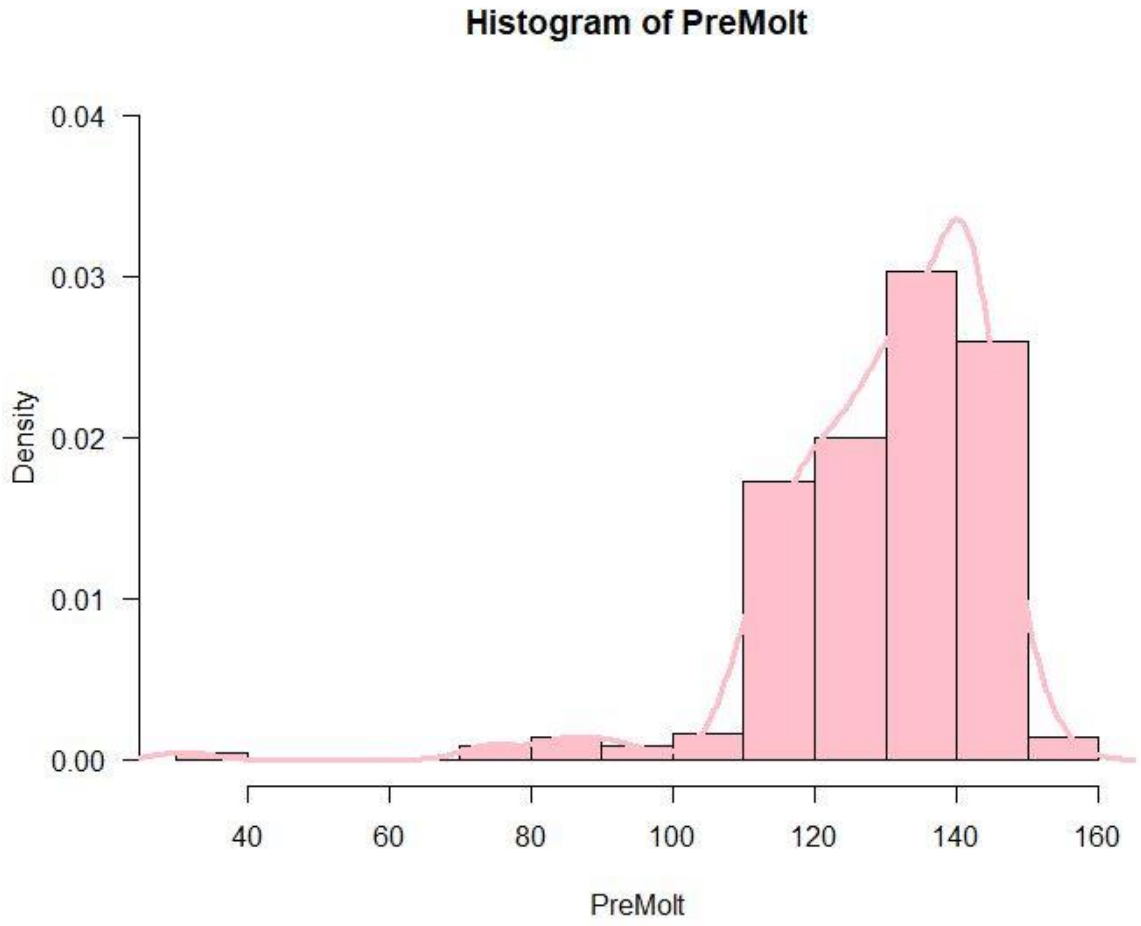
We came to the conclusion that the model had more inaccuracy when forecasting in the lower ranges of post-molt sizes after completing general statistics on the crab molting data. This is related to our problem with the inadequate data and, thus, the poor accuracy across the post-molt sizing ranges. We 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. Giving post molt sizes in this range results in less mistake and more accurate forecasts since the lower ranges of post molt sizing are less accurate than the upper ranges, roughly between 125 and 160. As the majority of post-molt sizes fall within this range, the prediction has a better idea of what the expected output pre-molt sizes should be, which accounts for the higher accuracy. It is demonstrated that post molt size data can predict pre molt sizes of crabs well within the limitations of tolerance or error due to the higher accuracy in the post molt range.

## Appendix A: Methods

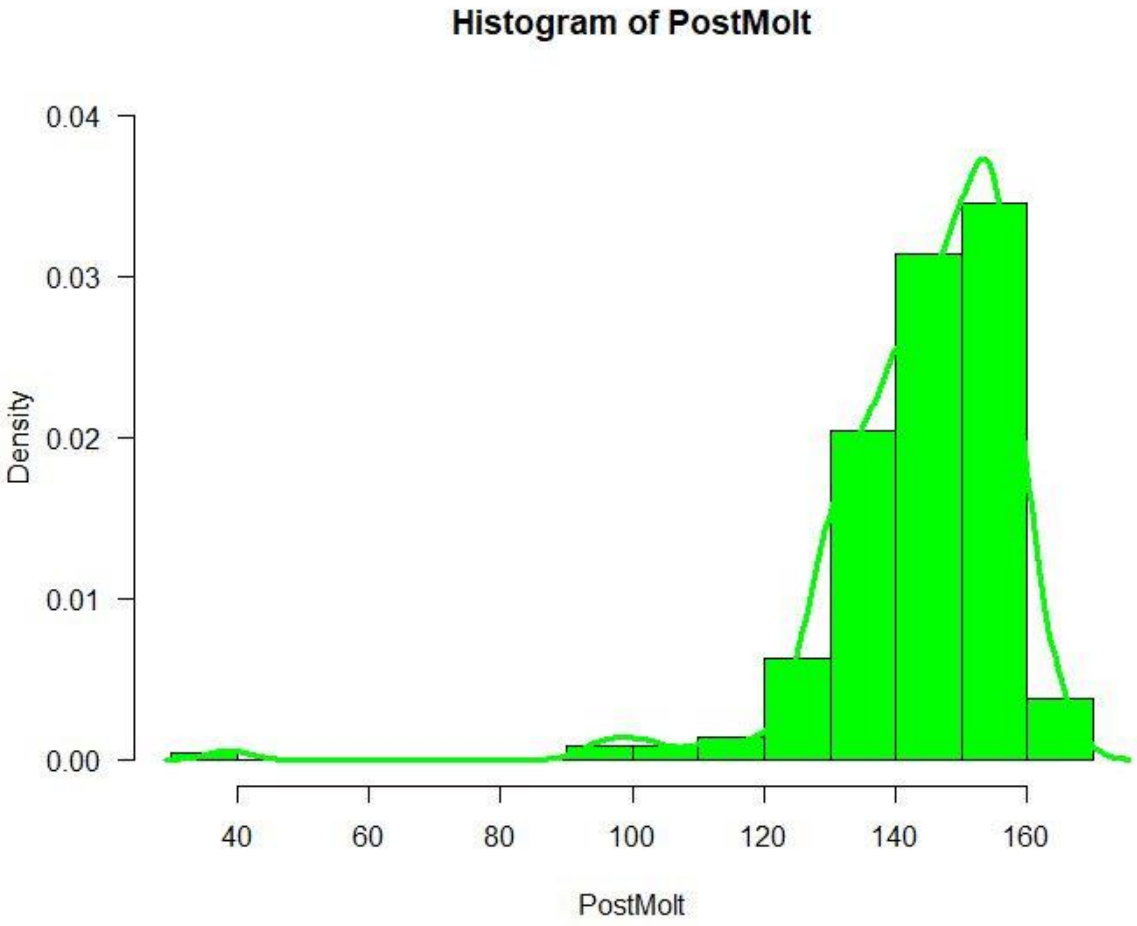
In order to forecast pre-molt sizes of crabs, this study uses a dataset of crab molt sizes and the variables pre-molt size and post-molt size. The following steps involved evaluating the data by collecting post- and pre-molt summaries for each variable, which gave us the minimum, first quartile, median, mean, third quartile, and maximum. The next step was to determine whether the data for each variable followed a normal distribution. There are several ways to do this, such as by looking at each variable's kurtosis and skewness to see if they are 3 and 0, respectively, or very close to them. Then by examining the density plots and histograms for each variable. The three tests were followed by the completion of the normalcy check. To illustrate the difference/shift in the data after plotting the density plots, we may superimpose them on top of one another and draw two vertical lines representing the means of the post- and pre-molt sizes. We can then develop the linear model in R using a linear regression function, forecasting pre-molt sizes based on post-molt values. We may also obtain the model's Pearson's r-squared value, which helps to show how effectively post-molt sizing contributes to forecasting pre-molt size, in order to assess the model's fit by examining the plot of the line produced by the R function over the data. Following the prediction model's computation, we compute the difference between the data's initial pre-molt values and the values our model predicted from the post-molt sizes to examine residuals/errors of the prediction. Next, by using a quantile or q-q plot in R, which generates a scatter plot of all the residuals and draws a line across it, we examine if the residuals or errors follow a normal distribution. The residuals would not be typical if some of the spots were not on the line. To thoroughly verify this, we calculate the residuals' kurtosis and skewness. If these values are 3 and 0, respectively, then the residuals are normal; otherwise, they could be extremely close to or very far from normal. We plot the density plot over the histogram and the histogram of the crab residuals in order to understand the kurtosis of the data. We next check to see if the histogram has a long tail or if the residuals reach a steep peak rapidly. Several factors may have an impact on the residuals' value of kurtosis. The residuals are then plotted on the y-axis while the post molt sizes are used as the predictor variable on the x-axis to determine whether the residuals exhibit heteroscedastic behavior. Heteroscedasticity in the residuals is present if there are visual clusters, conical forms, which indicate that there are more points as you move further right on the x-axis, or a discernible pattern to the residuals plotted. These statistics will enable a conclusion to be drawn regarding the model's accuracy and potential applications.

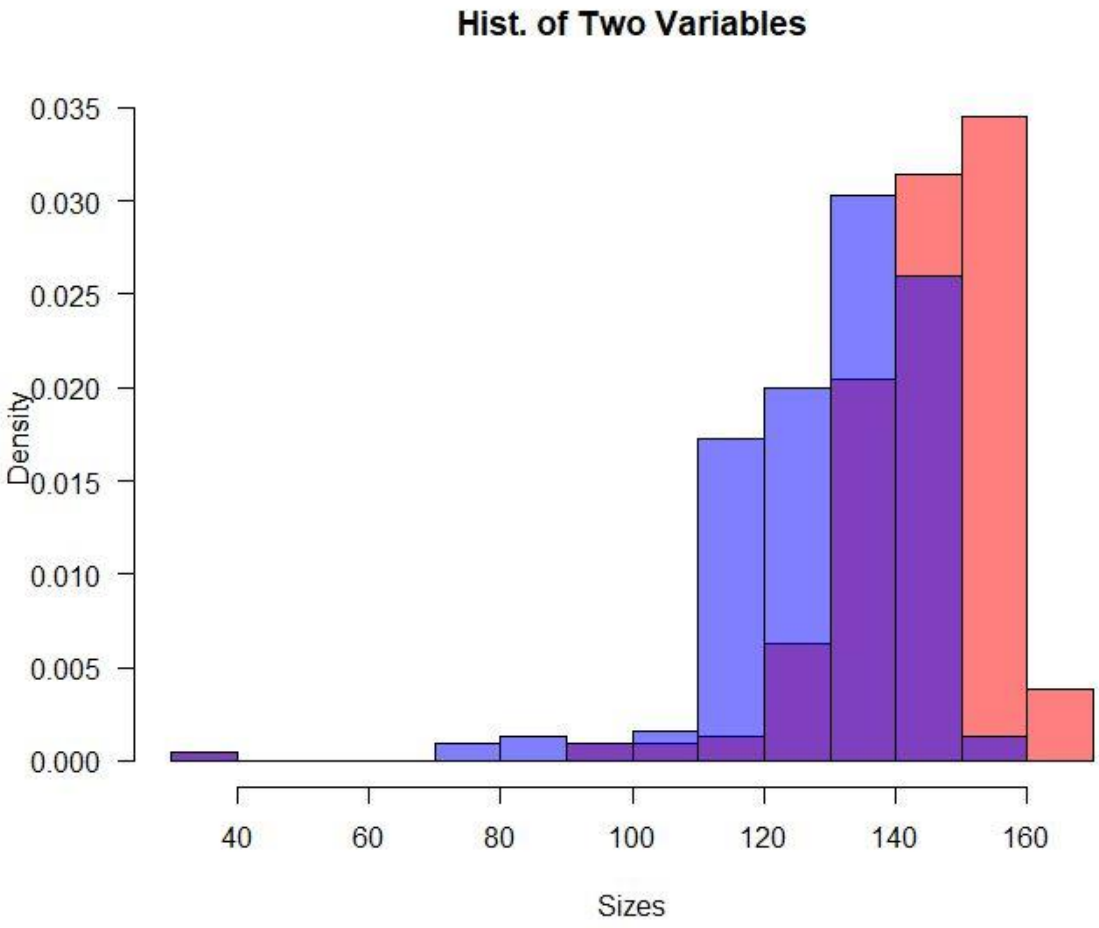
## Appendix B: Results

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




As the Kurtosis and Skewness are not 3 and 0, respectively, we can immediately conclude that neither variable has a normal distribution. Below are two histograms and two density plots; the left histogram represents the post-molt variable, and the right histogram represents the pre-molt variable. These density and histogram plots demonstrate that the data do not follow a normal distribution along with the Kurtosis and Skewness. (For comparison, see the example normal distribution histogram and density below.)

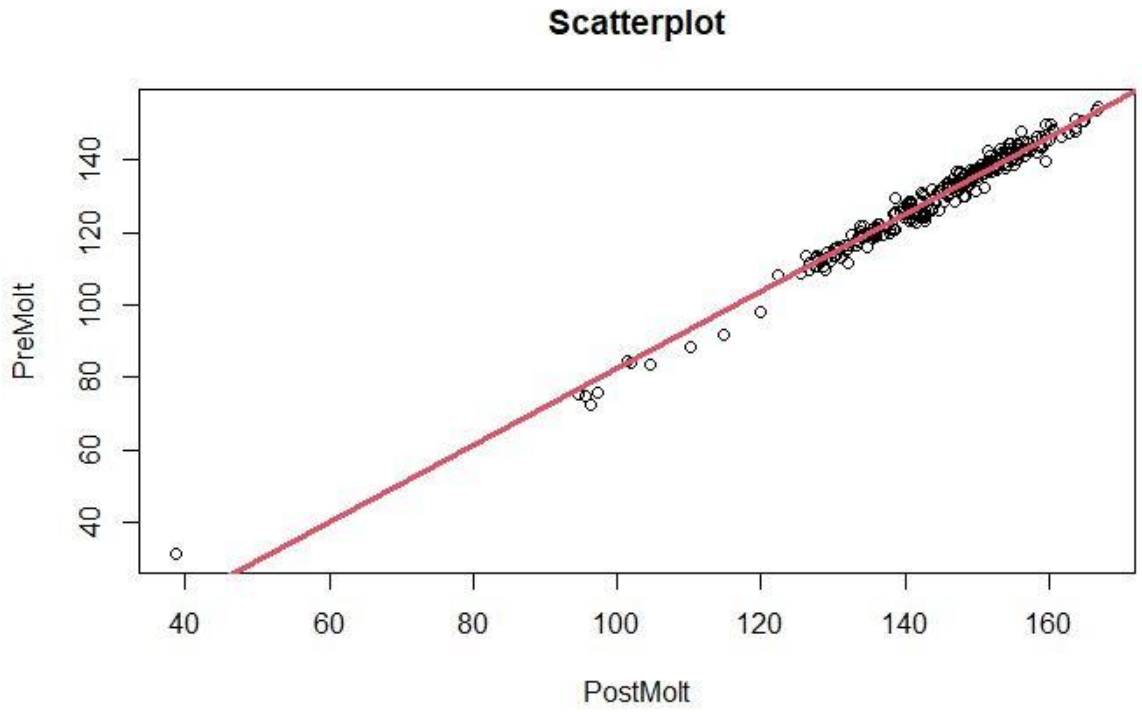

Histogram of PostMolt
Density
0.04
0.03
0.02
0.01
0.00
40
60
80
100
120
140
160
PostMolt



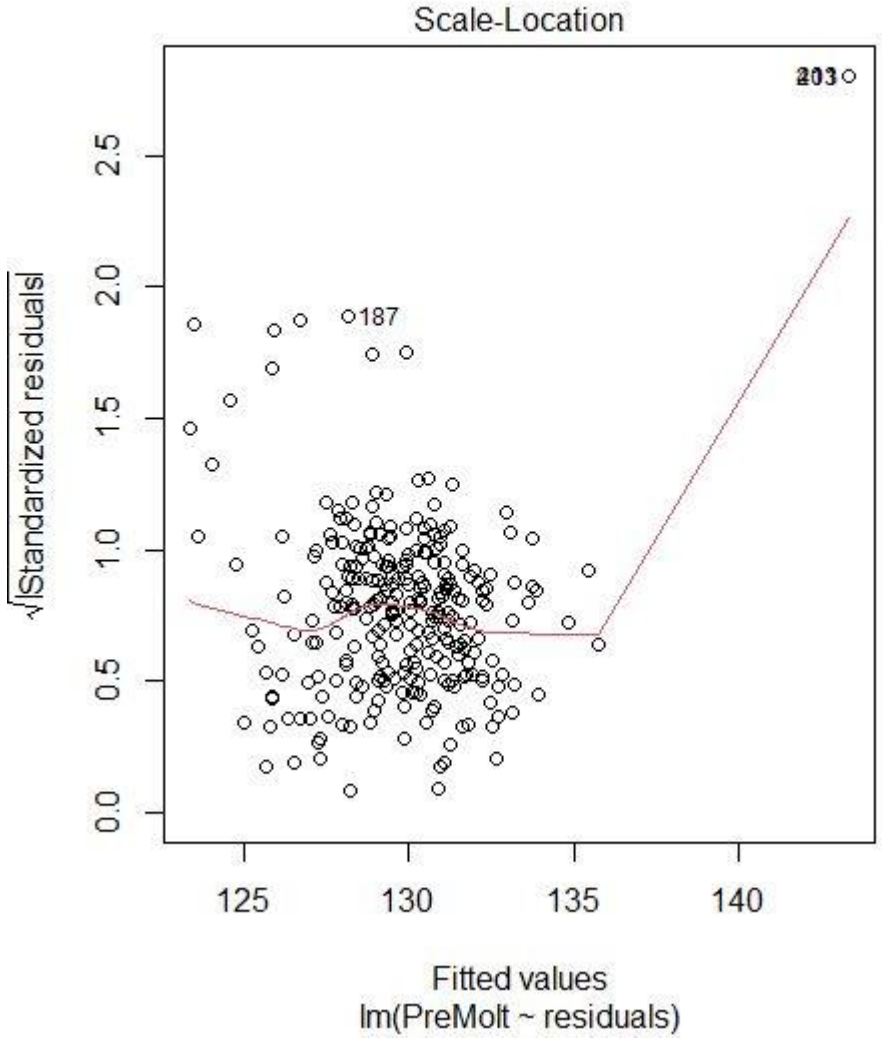

You can observe that the variables do not follow a normal distribution by examining the example normal distribution.

The density plot for pre- and post-molt sizes is overlaid in the following graphic, which also shows a line for each variable's mean to illustrate how the data differ. On the molt change, the mean differences total 143.3 - 128.5 = 14.8. (Each color dotted line represents the mean of the corresponding variable; for example, the mean for pre-molt is represented by the red dotted line.)

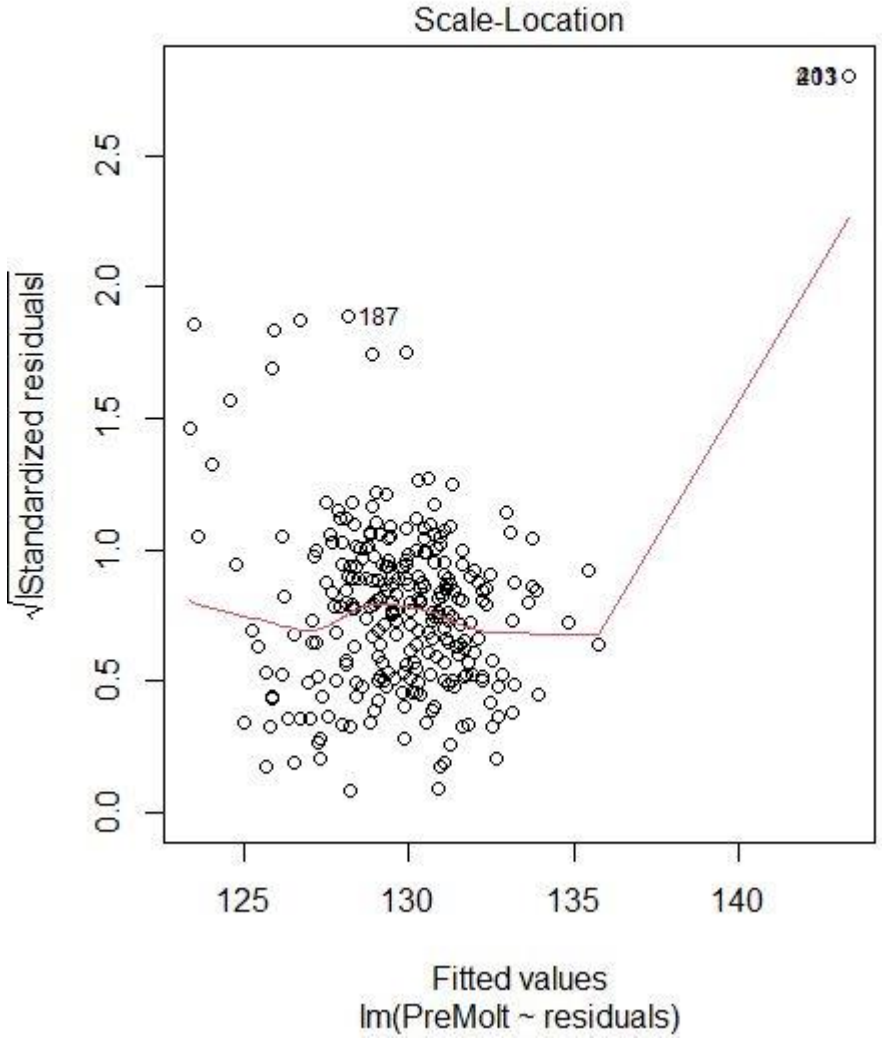
Below is a plot with the post molt data as the x-axis and pre-molt along the y-axis, then placed is the line generated from linear regression function. On the right-hand side is a summary of important values from the linear regression summary.




As seen below, a Q-Q or quantile plot is constructed using the residuals from the created linear model as a test for residual normality. The residuals have a normal distribution if all points lie on the line.




A histogram with a density plot overlay and a list of the residuals on the right side are provided to demonstrate this non-normality. The kurtosis of the residuals, which are otherwise close to normal but not normal, may be impacted by the plot with the long tail in this case.




The residuals are then plotted against the predictor variable, post molt size, with post molt acting as the x-axis and residuals as the y-axis. This graphic is used to check the prediction model for heteroscedasticity. This demonstrates that the model performs better at the higher end of the post molt range.

Appendix C: Data & Code

```
library(readxl)

crab_molt_data <- read_excel("C:/Users/Mahesh Varma/Desktop/MTH 522/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 PostMolt

min (PostMolt)

max (PostMolt)

median(PostMolt)

mean(PostMolt)

sd(PostMolt)

skewness (PostMolt)

kurtosis (PostMolt)


#PreMolt


min (PreMolt)

max (PreMolt)

median(PreMolt)

mean(PreMolt)

sd (PreMolt)

skewness (PreMolt)

kurtosis(PreMolt)
```

#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 PreMolt

hist (PreMolt, freq=F,las=1,ylim=c(0,0.040),col =‘blue’)

#Let’s find the density of the PreMolt and PostMolt variables

lines(density (PostMolt),col=“red”,lwd=3) lines(density(PreMolt),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 PreMolt”, xlabel=“Sizes”, Col=rgb(1,0,0,0.5),las=1) hist(PreMolt, 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&PreMolt”) lines(density(PreMolt),col=“blue”,lwd=3)

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

plot (PostMolt, PreMolt, main= “ScatterPlot”)

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

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

```
#Now we calculate find the Pearsons r^2 regression

results <- cor.test (PreMolt, 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 (PreMolt)

plot (residuals, PreMolt, main = “ScatterPlot”)
```

r_model <- lm (PreMolt~residuals) summary(r_model) abline(r_model, col="brown",lwd=3)

plot(r_model)

Refrence:


1. *Andrew Raposo - Data by Professor*. MTH 522 (Advanced Mathematical Statistics, section02B). (2023, January 21). Retrieved February 16, 2023, from https://mth522.wordpress.com/about/12-individual-grades/raposo-andrew/