UNDERSTANDING SIMPLE LINEAR REGRESSION FROM POST-MOLT,PRE-MOLT OF CRABS

THE ISSUES :

In U.S Seas, only male crabs are fishing, as female crabs cannot be fished in order to maintain the feasible of crabs population. However, now it is observed that there is an abnormality in the sex ratio in crabs.

It has also been argued that the great imbalance in the sex ratio may have contributed to the decline in the crab population.

In the view of biologists, the imbalance may have caused an increase in the parasitic ribbon worm population to the level where the worm now destroys 50–80% of the crab eggs each year. In this project we will study the growth patterns of female dungeness crabs in order to assist biologists in developing recommendations for size restrictions on fishing female crabs.

THE FINDINGS :

Gathering the summary statistics, such as the minimum, maximum, median, mean, standard deviation, skewness, and kurtosis, is ahelpful approach to gain insights into the size distribution and variability of adult female Dungeness crabs. Probability density function histograms can also be utilized to detect any outliers and abnormalities, which can provide additional information for further analysis and modeling

To examine the correlation between pre-molt and post-molt sizesof Dungeness crabs, a scatter plot can be used to visualize their relationship, with a least squares linear regression line added to the plot. The slope of this line can provide the average increase

in pre-molt size for a given increase in post-molt size. Additionally, Pearson's r^2 can be computed to assess the degree of fit between the data and the linear regression line, which can help to predict pre-molt size based on post-molt size.

When studying the growth of Dungeness crabs, it is crucial to inspect for any systematic deviations from the regression line in aplot of residuals against the dependent variable and to examine visual heteroscedasticity patterns. These methods can help to detect any outliers and provide valuable information for modifying the regression line to obtain accurate predictions.

The Discussion :

In this report, the focus is on studying the correlation between the size of female Dungeness crabs before and after molting. Thegoal is to create a reliable method for predicting pre-molt size from post-molt size and generate a histogram showing the distribution of pre-molt sizes. The data for the study was obtained from a combination of laboratory experiments and capture-recapture methods over three fishing seasons. A sampleof 442 crabs was taken after the molting period. The report emphasizes the importance of statistical analysis in comprehending the relationships between the crab sizes and predicting pre-molt size based on post-molt size while also accounting for factors like the collection method, the conditions in which the crabs were kept, and the fishing period.

The focus is on studying the correlation between the size of female Dungeness crabs before and after molting. The goal is tocreate a reliable method for predicting pre-molt size from post- molt size and generate a histogram showing the distribution of pre-molt sizes. The data for the study was obtained from a combination of laboratory experiments and capture-recapture methods over three fishing seasons. A sample of 442 crabs wastaken after the molting period. The report emphasizes the

importance of statistical analysis in comprehending the relationships between the crab sizes and predicting pre-molt sizebased on post-molt size while also accounting for factors like thecollection method, the conditions in which the crabs were kept, and the fishing period.

APPENDIX A: METHODS

Data was downloaded as an excel file (.xsl) and imported into Rstudio. First we found out the minimum, maximum, mean, median, standard variation, skewness and kurtosis in Post-molt and Pre-molt sizes of the crab, respectively.

The Probability Density Function (PDF) histograms were createdin R studio of each variable. Post-molt or Pre-molt in the x-axis and Density in the y-axis.

Smooth approximations of the distributions of post-molt of crab, and of pre-molt, crab from the data were plotted together to give a visual appreciation for the different size distributions for the two situations.

Pre-molt size as a function of post-molt size had been plotted ina graph. Post-molt size in the x-axis as it is an independent variable and pre-molt size in the y-axis as it is a dependent variable.

A simple linear regression with “Post-molt” size as the predictor variable, and “Pre-molt” size as the predicted variable is carriedout and plotted in a graph.On the same plot least square linear regression line is plotted and from there Pearson’s for the regression is calculated.

To assess the distribution of a model's residuals, I utilized the residuals() method to calculate them and saved them in a new

variable. I then employed the summary() function to generate the descriptive statistics of the residuals, including their mean, median, minimum, maximum, and quartile values. In order to examine the normality of the residuals, I created a quantile- quantile (Q-Q) plot using the qqnorm() and qqline() functions. If the residuals were normally distributed, the Q-Q plot would exhibit a straight line connecting each point. This process enabled me to determine if the residuals were distributed normally or not.

In R, I performed a formal test to check if the residuals from a linear model were normally distributed using the shapiro.test() function. To do this, I extracted the residuals using the resid() tooland input them into the shapiro.test() function to obtain the test results, which included the test statistic and p-value. I set the significance level to 0.05 and rejected the null hypothesis that theresiduals were normally distributed if the p-value was less than 0.05.
To check for heteroscedasticity, I utilized the resid() and predict() methods in R to extract the residuals and predicted values and saved them as separate variables. After that, I plotted the dependent variable against the predicted values using the plot() function. To visually inspect for heteroscedasticity, I added the residuals to the plot using the points() method, passing the residuals variable as the y-axis values. A funnel shape in the plotwould indicate the presence of heteroscedasticity. Specifically, ifthe spread of the residuals changed as the dependent variable increased or decreased.

APPENDIX B: RESULTS

In the dateset, there are 442 female crab samples, and the size ofeach sample before and after molting is considered as independent variables. In the comparison of the two variables, the

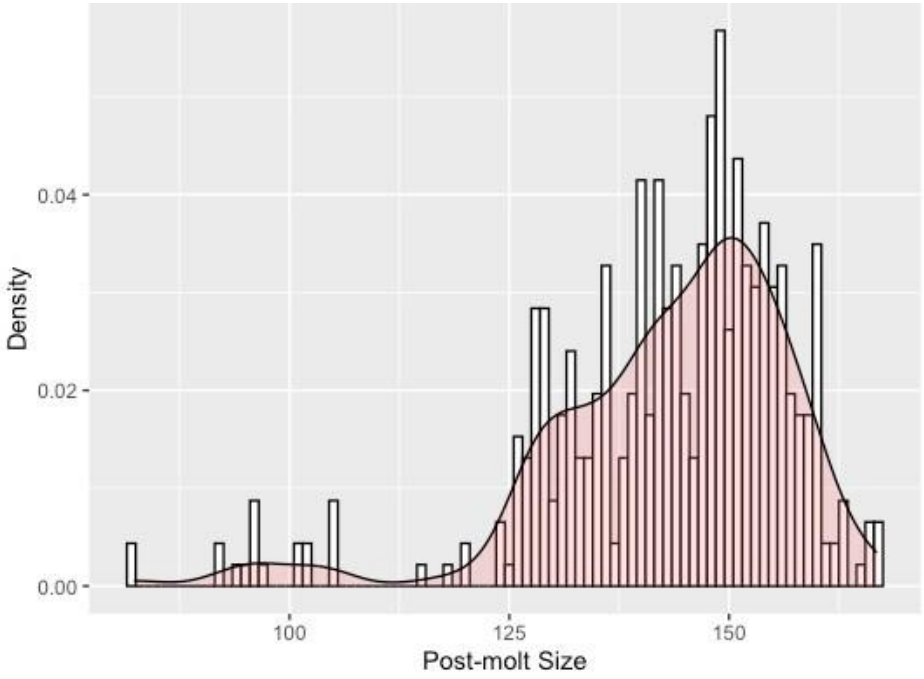
post-molt sizes of female crabs were found to have a higher mean of 143.2986, as compared to the pre-molt sizes which had a mean of 131.65. The standard deviation of post-molt sizes was lower at 15.07034 compared to the pre-molt sizes at 16.14941, which suggests that the pre-molt sizes are more variable than the post-molt sizes. The post-molt sizes had a slightly greater negative skewness of -1.46 than the pre-molt sizes, which had a skewness of -1.42, indicating that both variables are negatively skewed. Both pre-molt and post-molt sizes showed positive kurtosis values, suggesting that their distributions have more peaks than a normal distribution. Although the kurtosis value for post-molt size (6.16) was greaterthan that for pre-molt size (6.01), it can be inferred that the distribution of post-molt sizes is more distinct than that of pre- molt sizes.

The minimum, maximum, median, mode, standard deviation, skewness, and kurtosis results of the two variables Post-molt and Pre-molt crab

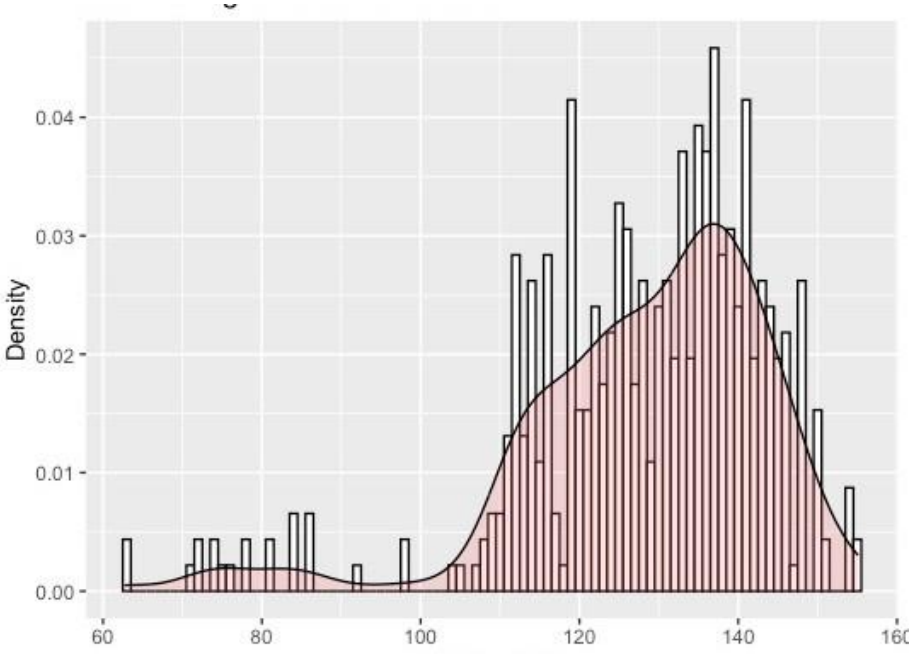
	Post-molt	**Pre-molt**
Minimum	38.8	31.1
Maximum	164.7	151.2
Median	146.5	131.65
Mean	143.2986	128.5676
Standard Deviation	15.07034	16.14941
Skewness	-1.457535	-1.421444
Kurtosis	6.162931	6.012548

In both the graph (Pre-molt and Post-molt) the shape of the PDFcurve gives us information about the distribution of the data.
While it is a skewed curve, it indicates a non-normal distribution.

Outliers also can be identified by looking at the tails of the PDF.And it is asymmetric graph thus we can say the mean and median are different which is satisfying the above table.
The Pre-molt and Post-molt sizes of the crabs by using Smooth Distribution of the variables.The blue area in this Smooth

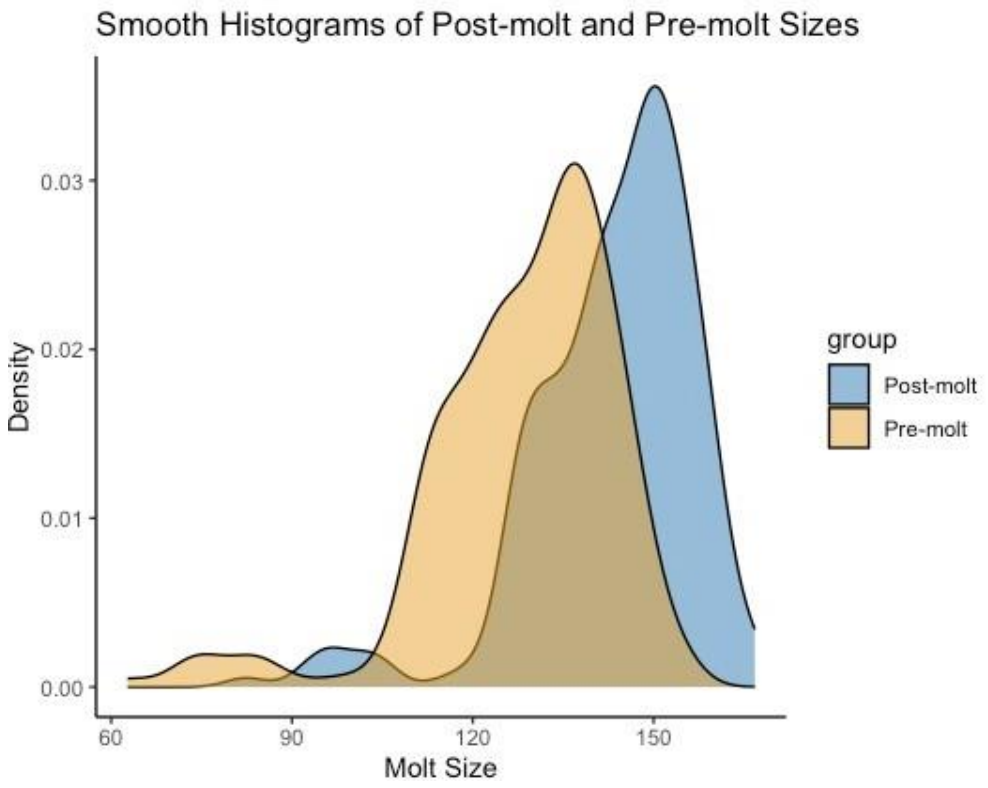


PDF of Post-molt size



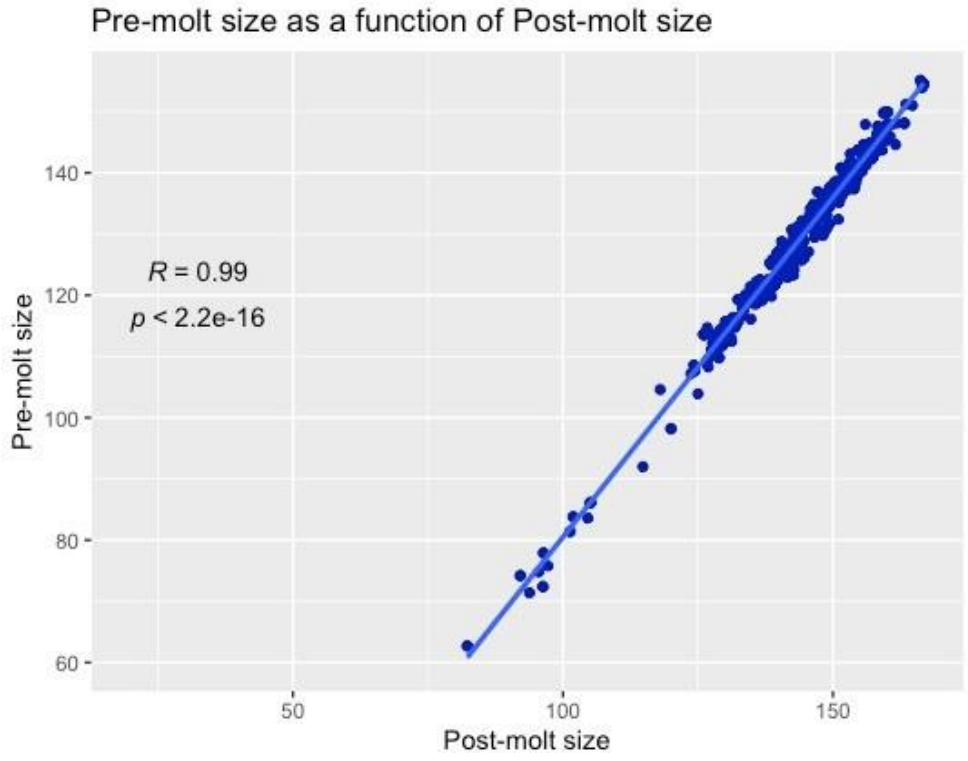
PDF of Pre-molt Size

Distribution Graph indicates the Post-Molt sizes of the crab,while the yellow area reflects the Pre-Molt sizes.
A comparison of smooth histogram approximations to the size distributions for post-molt, and for pre-molt, size of crabs showsa significant shift to the right(higher average size) for post-molt crabs:



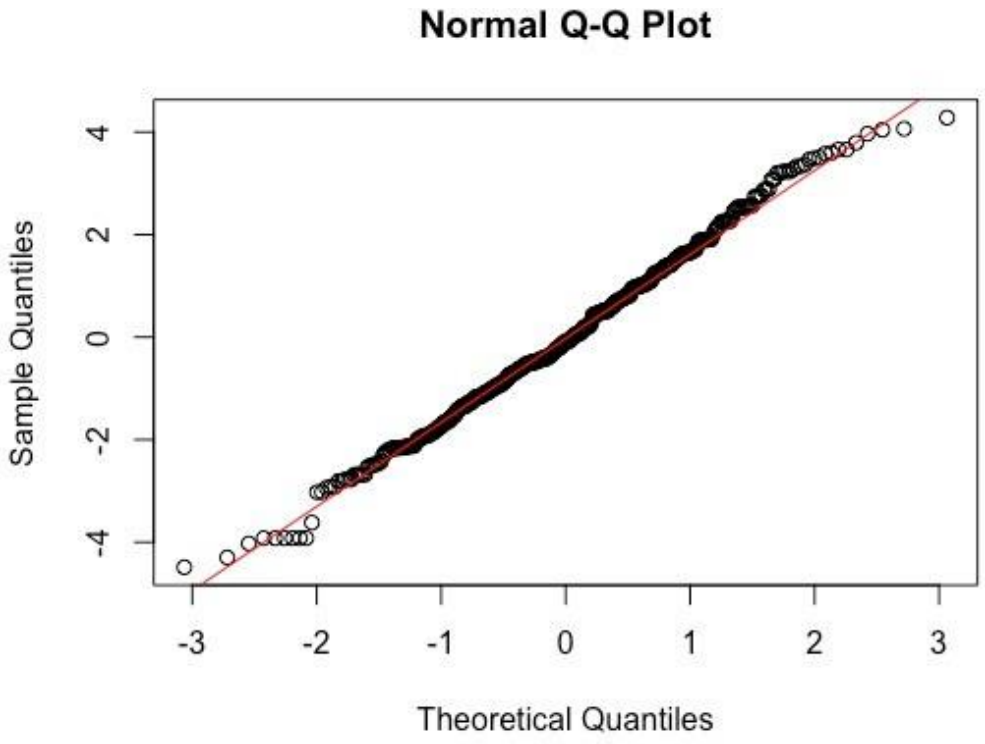
Smooth Histograms of Post-molt and Pre-molt Sizes

From the scatter plot, we see that the pre- and post-molt sizes of the crab shells are highly linearly associated. Pre-molt size on the Y-axis and Post-molt size on the X-axis .
We can observe that the Pre-molt size plots are densely packedbetween 100 and 160 with regard to the Post-molt size,
That is, the points on the scatter plot in figure below are closelybunched around a line.
A least squares linear regression line is represented in blue, which is a straight line that is fitted to the provided data to minimize the sum of squared residuals. As the regression line hasa positive slope, it can be inferred that the size of female crabs before molting increases as the size of the crabs after molting increases. There is a strong positive correlation between the
post-molt and pre-molt sizes of female crabs, which is indicated by the R value of 0.99. This suggests that the line is a good fit forthe data, and the model can be used with some degree of certainty to estimate the pre-molt size of female crabs based on their post-molt size.



Pre-molt size as a function of Post-molt size

In a regression model, the residuals are the discrepancies between the predicted and actual values of the response variable, representing the extra variability in the data that cannotbe explained by the predictor variable. The expected values of the residuals, which follow a normal distribution, are depicted bythe red line.



Normal Q-Q Plot

In the case of the post- and pre-molt sizes of female crabs, the residuals can be used to evaluate the goodness of fit of the linearregression model to the data. The scatter plot of the residuals shows an approximately straight line, indicating that the residualshave a normal distribution and the linear regression model is a good fit for the data.

Shapiro-wilk Normality Test
data: residuals
W =0.99495 p-value<2.2e-16

However, the Shapiro-Wilk test results indicate that the null hypothesis of normality is rejected, as the p-value is lower thanthe significance level of 0.05, and the W statistic is 0.91067.
Thus, it can be concluded that the residuals are not normally distributed. Nevertheless, the scatter plot of the residuals versusthe dependent variable does not show any clear pattern, suggesting that there is no significant heteroscedasticity in the model.

APPENDIX C : Code

Code for importing data from excel

```
library(readxl)
crab_molt_data_payili_ramana <- read_excel("Downloads/
crab_molt_data_payili_ramana.xlsx")
View(crab_molt_data_payili_ramana)
```

Code for minimum, maximum, median, mean, standard deviation, skewness, and kurtosis in Post-molt and Pre-molt variable respectively.

Post-molt

```
> min(mydata$`Post-molt`)
[1] 38.8
> max(mydata$`Post-molt`)
[1] 164.7
> median(mydata$`Post-molt`) [1]
146.5
> mean(mydata$`Post-molt`) [1]
143.2986
> sd(mydata$`Post-molt`)
[1] 15.07034
> skewness(mydata$`Post-molt`) [1]
-1.457535
> kurtosis(mydata$`Post-molt`)
[1] 6.162931
```

Pre-molt

```
> min(mydata$`Pre-molt`)
[1] 31.1
> max(mydata$`Pre-molt`)
```

[1] 151.2
> median(mydata$`Pre-molt`)
[1] 131.65
>mean(mydata$`Pre-molt`) [1]
128.56767
> sd(mydata$`Pre-molt`)
[1] 16.14941
> skewness(mydata$`Pre-molt`) [1]
-1.421444
> kurtosis(mydata$`Pre-molt`)
[1] 6.012548

```
library(ggplot2)
ggplot(mydata, aes(x = `Postmolt`, y = `Premolt`)) +
geom_point() +
geom_smooth(method = "lm") +
xlab("Post-molt size") + ylab("Pre-
molt size") +
ggtitle("Pre-molt size as a function of Post-molt size”)
```

PDF of Pre-molt and Post-molt

Pre-molt

```
ggplot(mydata = df, aes(x = Premolar)) + geom_histogram(aes(y =
..density..), binwidth = 1, color = "black", fill = "white") +
geom_density(alpha = 0.2, fill = "#FF6666") + ggtitle("PDF
Histogram of Pre-molt Size") +
xlab("Pre-molt Size") +
ylab(“Density")
```

Post-molt

```
ggplot(mydata = df, aes(x = Postmolt)) + geom_histogram(aes(y =
..density..), binwidth = 1, color = "black", fill = "white") +
geom_density(alpha = 0.2, fill = "#FF6666") + ggtitle("PDF
Histogram of Post-molt Size") +
xlab("Post-molt Size") +
ylab(“Density")
```

Smooth histogram for each variable overlaid

```
Data_smooth <-
data.frame(value = c(postmolt, premolt), group =c(rep("Post-molt",
length(postmolt )), rep("Pre-molt", length(premolt))))
ggplot(Data_smooth, aes(x = value, fill = group)) +
geom_density(alpha = 0.5) +
labs(title = "Smooth Histograms of Post-molt and Pre-molt Sizes", x =
"Molt Size", y = "Density") + scale_fill_manual(values
= c("#0072B2", "#E69F00")) +
theme_classic()
plot(x=postmolt-y=premolt)
```

Pre-molt size (dependent variable) as a function ofPost-molt size (independent variable)

```
ggplot(Data, aes(x = postmolt, y = premolt)) + geom_point(color
= "#001eb2") +
geom_smooth(method = "lm")
xlab("Post-molt size") +
ylab("Pre-molt size") +
ggtitle("Pre-molt size as a function of Post-molt size")
```

Linear regression line on the same plot as the data

```
ggplot(mydata, aes(x = postmolt, y = premolt)) + geom_point(color
= "#001eb2") +
geom_smooth(method = "lm") +
xlab("Post-molt size") + ylab("Pre-
molt size") +
ggtitle("Pre-molt size as a function of Post-molt size") +
stat_cor(label.x = 20, label.y = 120, method = "pearson",
label.sep = "\n")
```

Residuals

```
residuals<-resid(lm(mydata))
```

For getting the line

```
qqnorm(residuals) qqline(residuals, col = “red")
summary(residuals)
```

Shapiro-Walks test

```
shapiro.test(residuals)
>Shapiro-Wilk normality test data:
>residuals W = 0.99495, p-value = 0.1406
```

The residuals against the dependent variable

```
ggplot(Data, aes(x
= premolt, y = residuals)) + geom_point() +ggtitle("Residuals vs
Pre-molt  Size")  + geom_smooth(method = "lm", se = FALSE)
xlab("Pre-molt Size") +
ylab("Residuals")
```

References:

1)Stat Labs: Mathematical Statistics Through Applications.

2) https://bookdown.org.

3) An Introduction to Statistical Learning with Applications in R.