Analysis on the pre-molt and post-molt carapace size of Dungeness crabs

The Issues:

The Dungeness crab, one of the largest and most abundant crabs, is a large crustacean found off the Pacific coast. The data-base for this analysis was collected as part of study of the adult female Dungeness crab. The data-base consists of pre-molt and post-molt carapace size of crabs of which some molted in laboratory and some molted in ocean.

We will examine the relationship between pre-molt and post-molt carapace size and summarize the results. This analysis will help to better understand the growth patterns of female Dungeness crabs.

Findings:

The results of the data analysis on pre-molt and post-molt carapace size of Dungeness crabs indicate that there is a strong linear relationship between the two variables. The analysis found a correlation coefficient of 0.99, which provides strong evidence against the null hypothesis. The p-value was found to be very small(less than 0.001) which further supports the strong statistical significance of the relationship between pre-molt size and post-molt size. Overall, these findings suggest that there is a clear and significant association between the two variables.

Discussion:

With the given data, we carry out a simple linear regression with post-molt size as the predictor variable and pre-molt size as the predicted variable. From the calculation we find that Pearson's r^2 coefficient is found to be 0.9915901, which signifies strong association between the two variables.

We do Shapiro-wilk test, which gives us values of w (=0.94589) and p-value (<0.001), which is strong evidence against null hypothesis implying that there is good statistical significance existing between pre-molt size and post-molt size.

Appendix A: Method

To conduct the analysis of the pre-molt and post-molt carapace size of Dungeness crabs, we first downloaded the data as a .csv file and imported it into R-studio. The data consists of two columns, one for post-molt size and the other for pre-molt carapace size. Pre-molt size, which is the size of the crab's shell before molting, is considered as a function of post-molt size and is analyzed accordingly.

We employed various statistical tools to analyze the relationship between pre-molt size and post-molt size, including histograms, scatter plots, and quantile plots. We used Pearson's method to evaluate the correlation coefficient between the two variables and performed a Shapiro-Wilk test to compute the associated p-value. By using these methods, we can gain a better understanding of the relationship between pre-molt and post-molt carapace size and determine the statistical significance of any observed associations.

Appendix B: Results

In our analysis, we utilized a data-set consisting of a total of 442 data points, which was imported into R-studio.

To compare the pre-molt size and post-molt size of Dungeness crabs, we generated histograms and smooth histograms for each variable.

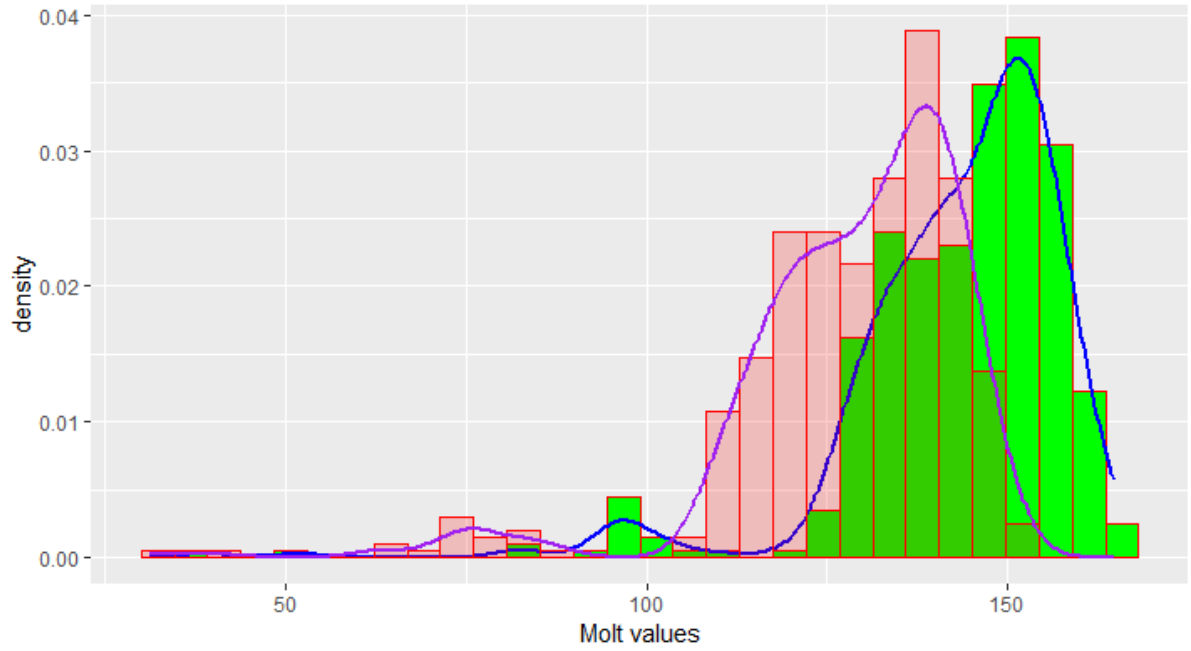

Figure 1: Distribution of post-molt size and pre-molt size

The histogram for post-molt size, represented by the green color, exhibited a significant shift to the right when compared to the histogram for pre-molt size, represented by the red color. This suggests that there is a noticeable difference between the two variables and indicates that post-molt size tends to be larger than pre-molt size in the sample.

In figure 2, we see plotting of pre-molt size as a function of post-molt size and in figure 3, we see the result of carrying out a simple linear regression with post-molt size as the predictor variable and pre-molt size as the predicted variable.

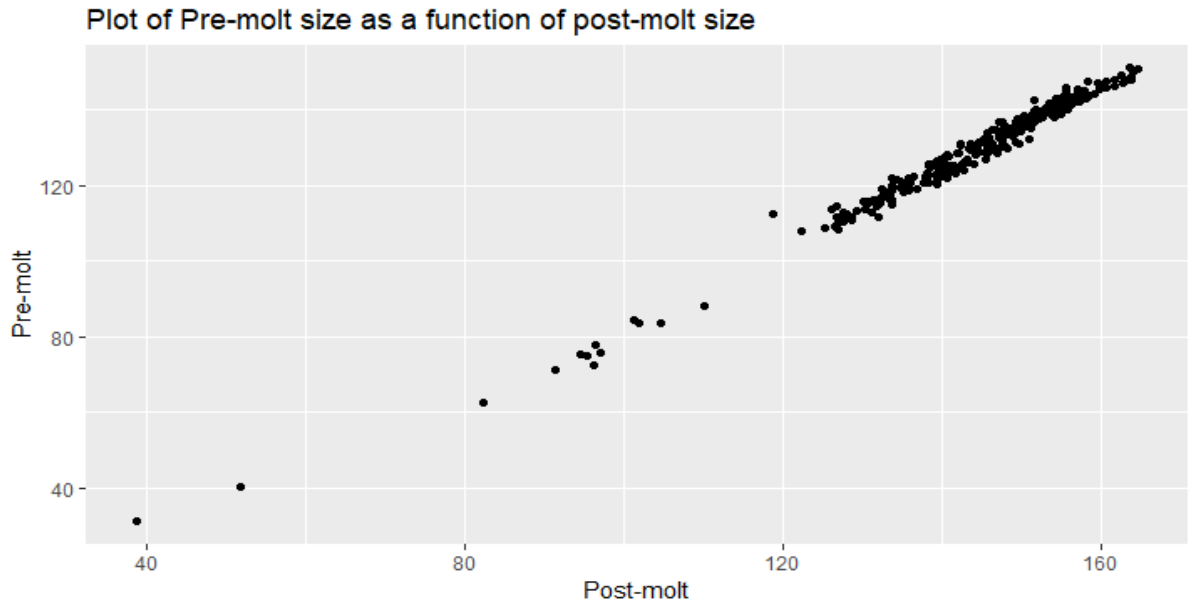

Figure 2: Scatter plot between pre-molt size and post-molt size

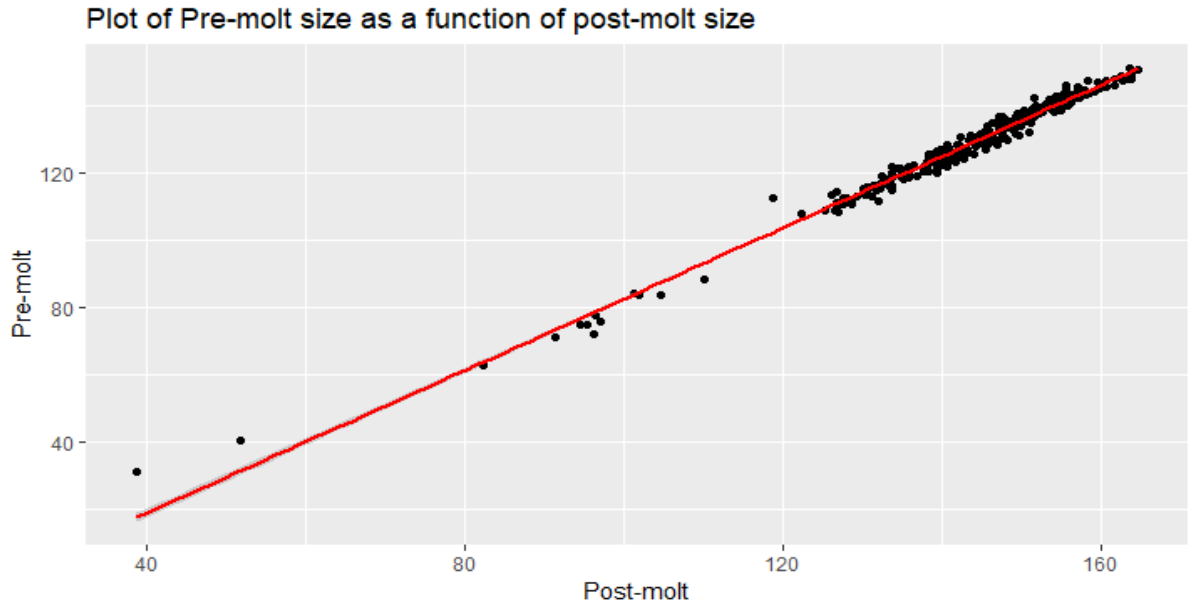

Figure 3: Linear regression relation between post-molt size and pre-molt size

The red line passing through the scattered points in figure 3, shows that the relation between pre-molt size and post-molt size is linear.

Figure 4 shows quantile plot for the distribution of post-molt size and pre-molt size. The quantile plot holds a proof that the data which is analyzed is proper.

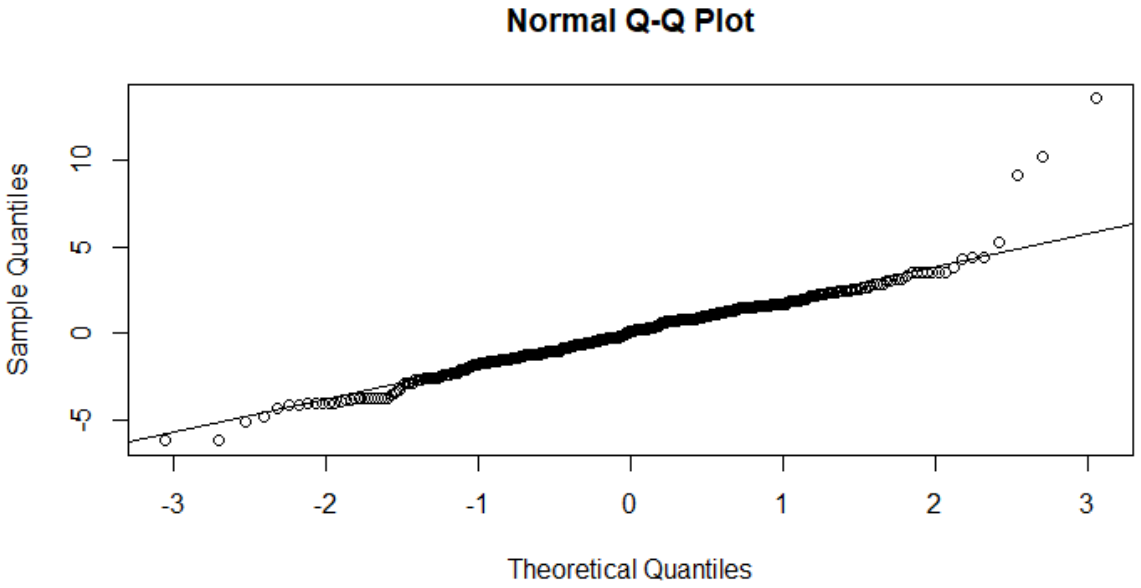

Figure 4: Quantile plot for the distribution of post-molt and pre-molt size

From figure 4, we can see that most of the points are around the reference line and very less points are deviating from the reference line indicating that only for few points errors is present.

Below Figure 5, shows plot between residuals and dependent variable. After the Shapiro-wilk test we get a p-value of 1.271e-11, which is very much less than 0.01. This indicates

that the relation between post-molt size and pre-molt size is significant.

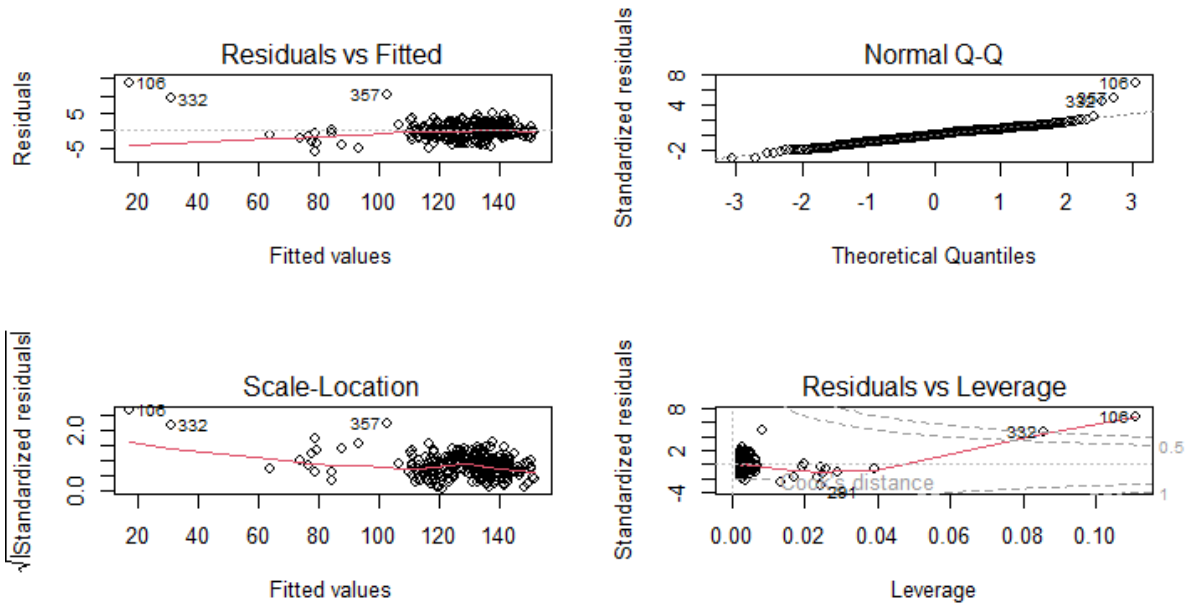

Figure 5: Residuals against the dependent variable

The residual vs. dependent variable plot in Figure 5 does not show any discernible pattern or shape. The residuals are randomly scattered without any clear trend or relationship between the residual values and the dependent variable. This indicates that there is no evidence of heteroscedasticity in the model, as the variance of the residuals is roughly constant across the range of the dependent variable. Therefore, the model appears to be a good fit for the data, and there is no need to explore alternative modeling approaches to handle heteroscedasticity.

Appendix C: Code

In this appendix we will document the code written in R studio to plot histograms, scatter plot, quantile plot and code used to get the Pearson's r^2 and p-value from the Shapiro-wilk test.

```
install.packages("readxl")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("moments")
library(moments)
library(ggplot2)
library(readxl)
library(dplyr)
library(DAAG)
 #Importing data-set
file <- "C:/Users/DELL/Downloads/crab_molt_data_payili_ramana.xls"
mydata <-  read_excel(file)
View(mydata)
nrow(mydata)
ncol(mydata)
colnames(mydata)
```

```
str(mydata)

summary(mydata)

skewness(mydata)

kurtosis(mydata)

class(mydata)

sd(mydata$`Posr-molt`)

sd(mydata$`Pre-molt`)

#Histogram for each variable

hist(mydata$`Posr-molt`, breaks = 50, freq = FALSE, main = "PDF Histogram
for post-mold", xlab = "Data Values", ylab = "Probability Density")

hist(mydata$`Pre-molt`, breaks = 30, freq = FALSE, main = "PDF Histogram
for pre-mold", xlab = "Data Values", ylab = "Probability Density" )

#Overlaying of histograms for each variable

mydata %>% ggplot(aes(x=`Posr-molt`)) +

  geom_histogram(aes(y=..density..),color='red',fill='green')+

  geom_density(aes(y=..density..),color='blue', lwd=1) +

  geom_histogram(aes(x= `Pre-molt`, y=..density..),color='red'
,fill='red',alpha=0.2)+

  geom_density(aes(x=`Pre-molt` ,y=..density..),color='purple', lty=1,lwd=1) +

  labs(x="Molt values")

  ggtitle("Overlaying of Smooth Histograms for Each Variable")

#Plotting the pre-molt size as a function of post-molt size

scatter <- ggplot(mydata,aes(x=`Posr-molt`,y=`Pre-molt`))+

  geom_point()+

  labs(x="Post-molt" , y="Pre-molt")+
```

```
  ggtitle("Plot of Pre-molt size as a function of post-molt size")
#Simple linear regression
model <- lm(`Pre-molt`~`Posr-molt`, data = mydata)
summary(model)


linear <- scatter+ geom_smooth(method="lm", col='red')
linear
correlation <- cor(mydata$`Posr-molt`, mydata$`Pre-molt`, method = 'pearson')
correlation
#Stats of residuals and quantile plot and the Shapiro-Walks test
residuals <- resid(model)
summary(residuals)
qqnorm(residuals)
qqline(residuals)
shapiro.test(residuals)
#Plot the residuals against the dependent variable
plot(mydata$`Pre-molt`, residuals, xlab="Pre-molt values" , ylab="residuals" ,
main ="Residual analysis")
abline(0,0)
# Heteroscedasticity
par(mfrow=c(2,2))
plot(model)
```