Analysis of Carb Molt Data
Using
Linear Regression

The Issues:

The project uses Dungeness crabs, specifically the fishing of female crabs, to reduce variations in crab harvests throughout the year. In certain regions, like the central California coast, the imbalance in sex ratio has contributed to the collapse of the crab population. This may also have led to an increase in the population of parasitic ribbon worms, which consume about 90% of eggs annually. To create size limitations on fishing female crabs, we need to understand the growth processes of the Dungeness crabs. On the basis of the supplied database, we will address the following:

A) Based on the post-molt crab size, we will use linear regression to calculate the pre-molt crab size.

B) Determine the linear model's accuracy based on the linear model's heteroscedasticity and the data residuals.

Findings:

Based on the data obtained from crab molting, it was found that the size of the crabs before and after molting was primarily in the range of 120mm to 160mm, with the most common size being 140mm. Additionally, a residual analysis was conducted, which revealed the presence of some small residues and a low level of residual error. Furthermore, a visual examination of heteroskedasticity was performed, and it was determined that the regression model was mostly homoscedastic, indicating a consistent level of variability in the data.

Discussion

Appendix A: Method

The dataset containing two primary variables, pre-molt and post-molt, was downloaded as an .xls file and imported into R Studio. Descriptive statistics, including minimum and maximum values, first and third quartiles, median, mean, standard deviation, skewness, and kurtosis, were calculated for both variables using the summary function. The probability density function was applied to generate histograms for each variable and to identify the peak crab size values. An overlay histogram was also created to facilitate comparisons between the two variables. A scatter plot was generated using pre-molt size as a function of post-molt size, and least square regression and Pearson's (r^2) regression were performed on the data. Residual analysis was conducted using the Shapiro-Walks test to assess normality, and a visual examination was carried out to assess heteroscedasticity and evaluate the potential impact of residuals on the accuracy of the model.

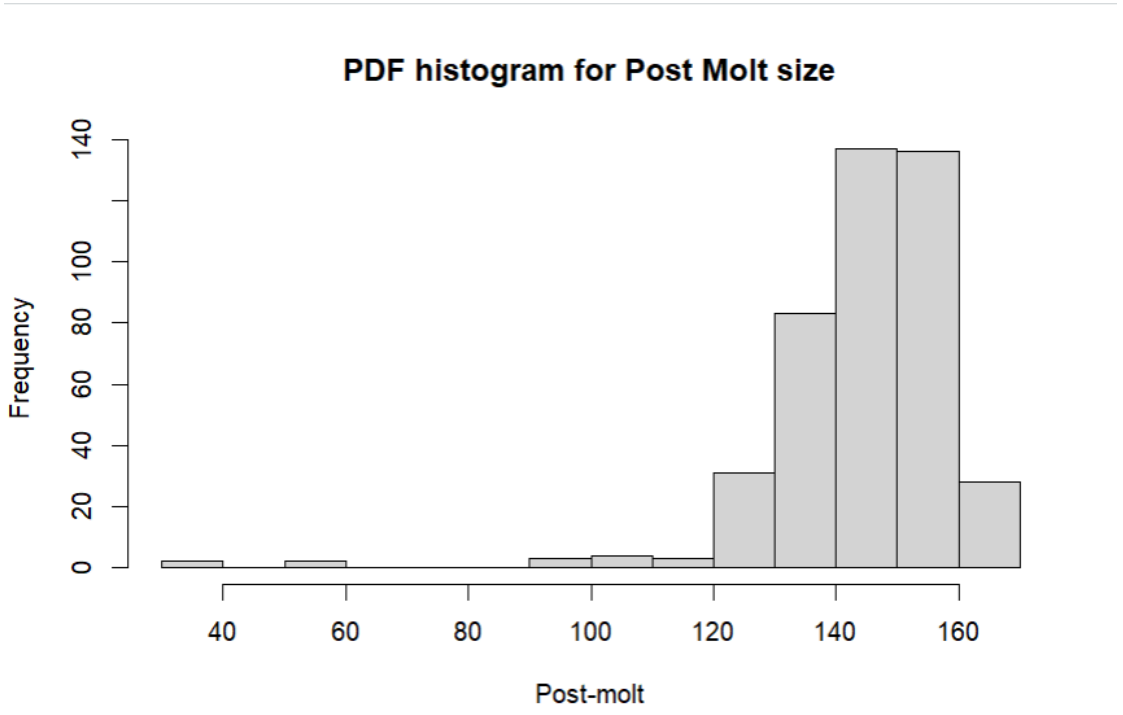
Appendix B: Results

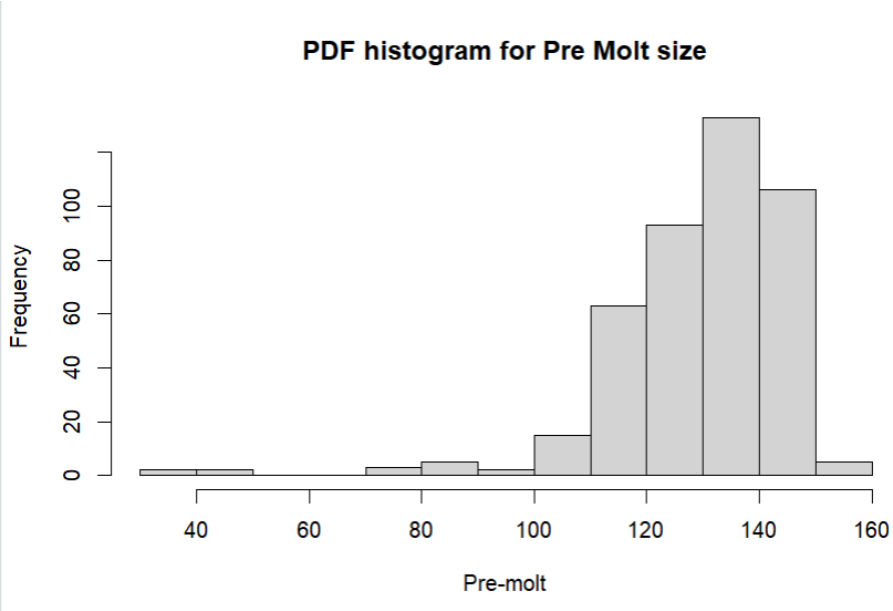
The data consisted of 401 points which had Pre-molt and Post-molt sizes.

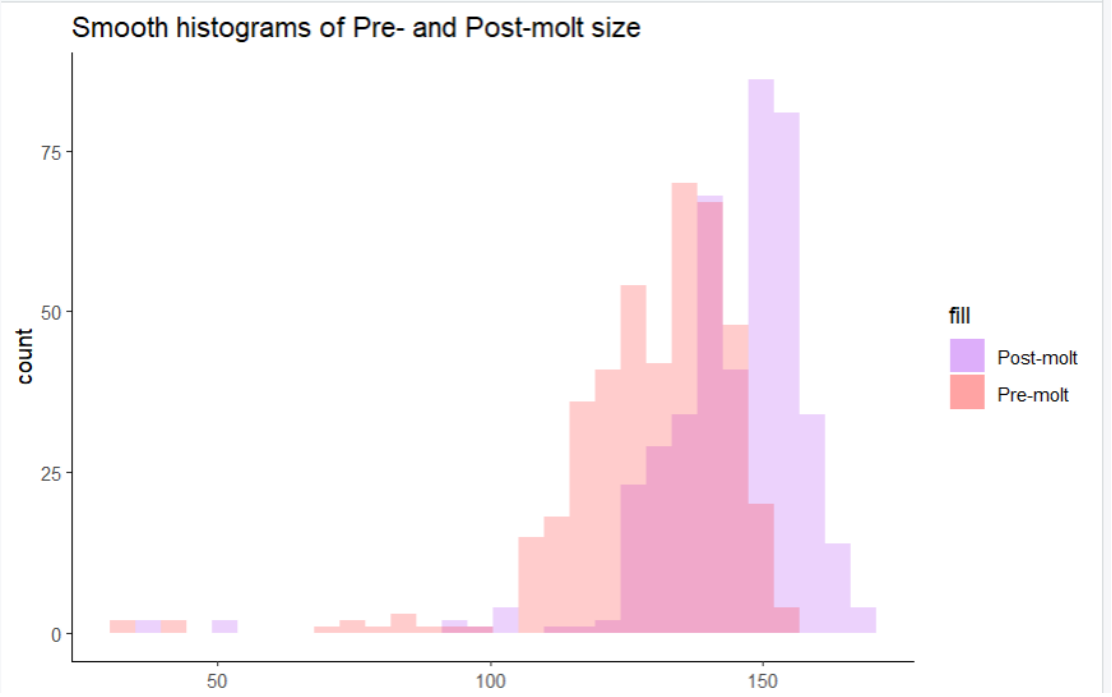
The summary function was applied to fetch the descriptive statistics of both the variables:

```
> summary(data$'Post-molt')
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   38.8   138.6   147.7   144.2   153.6   166.5
> summary(data$'Pre-molt')
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   31.1   122.0   133.6   129.6   140.7   155.1
```

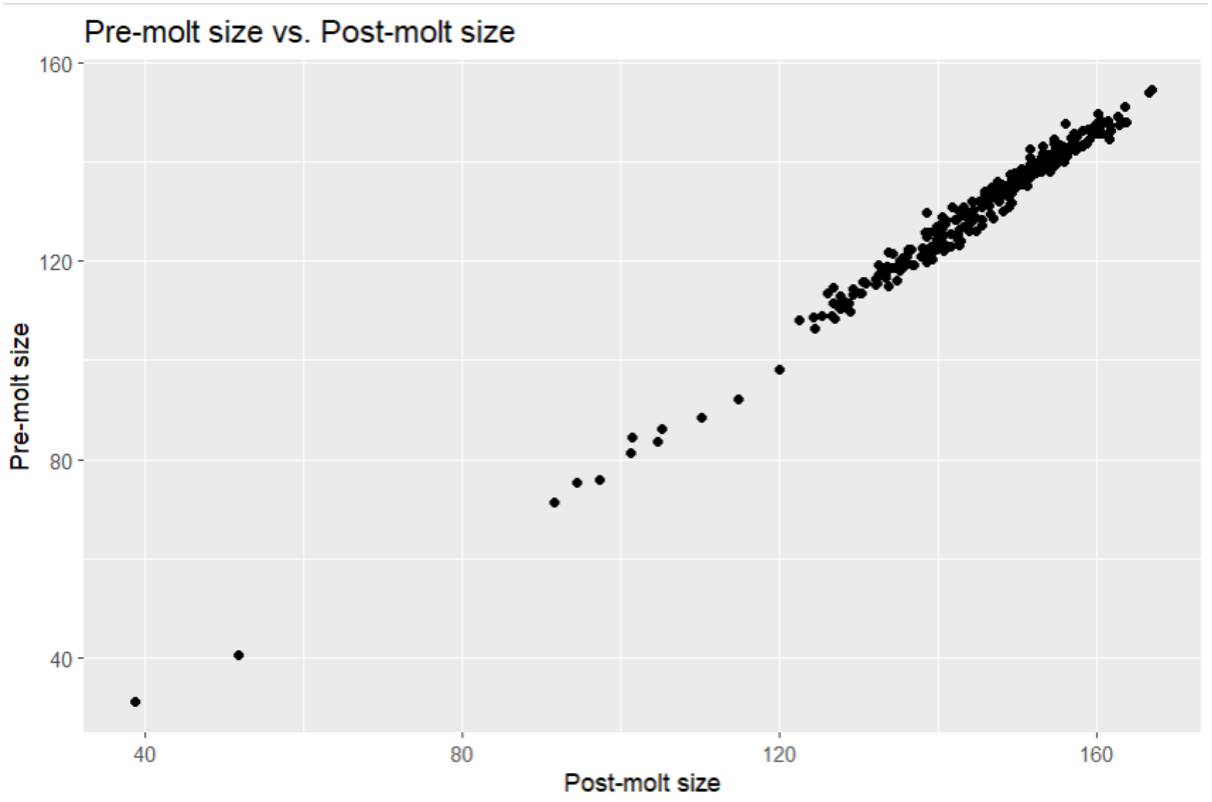
A probability function histogram was plotted for both Post Molt size and Pre Molt size as well as an overlay histogram to compare and observe the similarities in size.





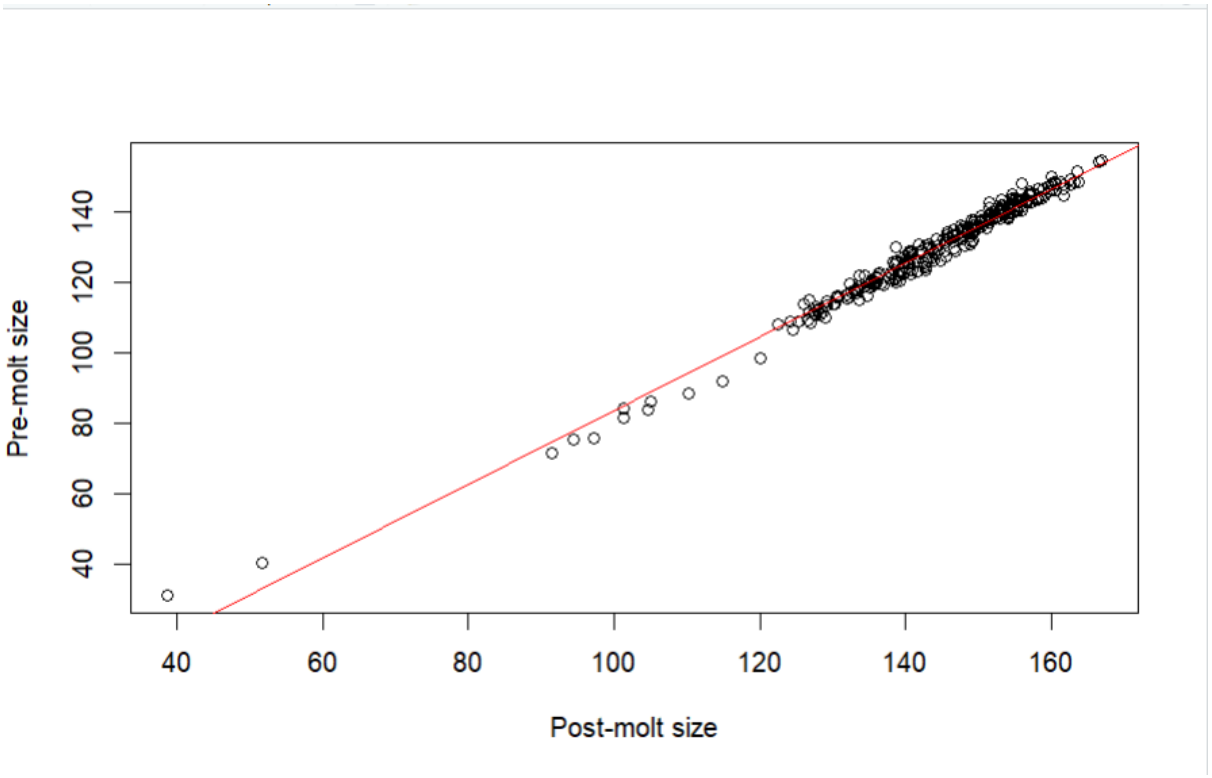


Scatter plot

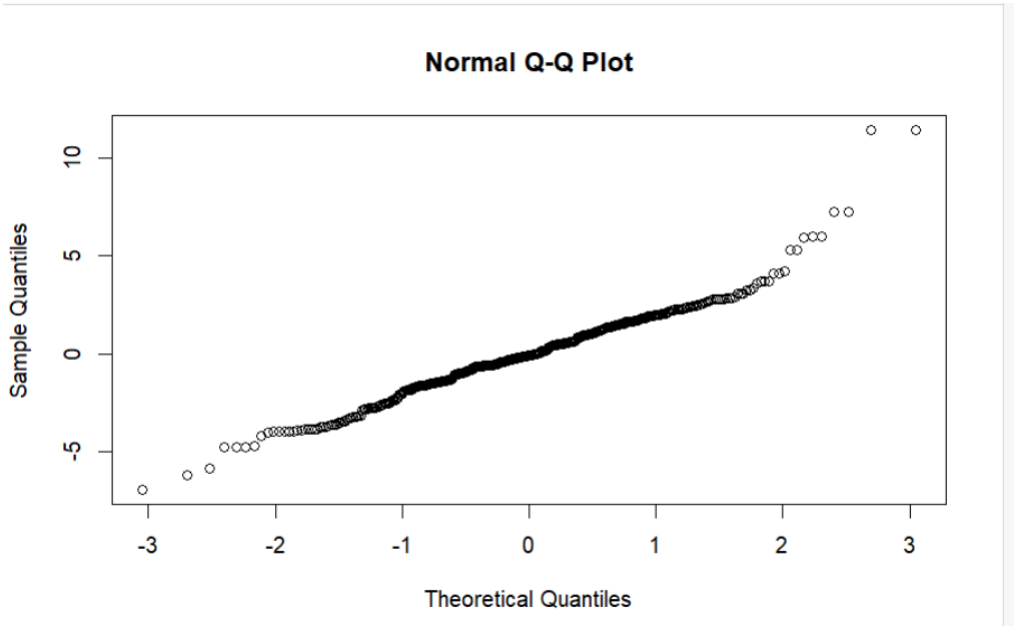


Using Post-molt size as the predictor variable and Pre-molt size as the predicted variable, a linear regression analysis was conducted. The resulting least squares linear regression was plotted on the same line as the data.

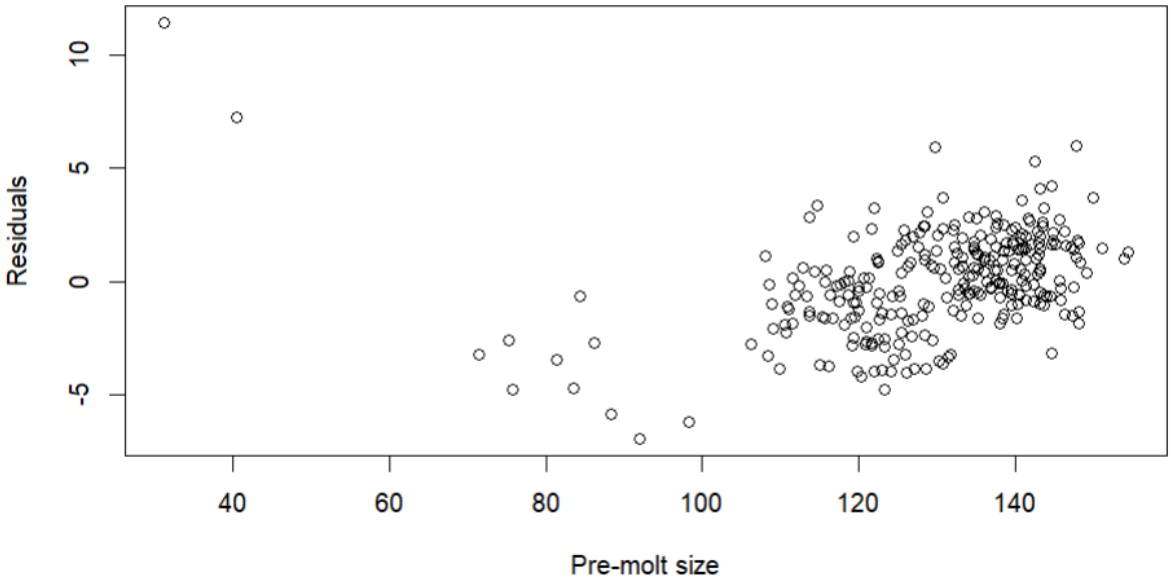
Linear regression of Post-molt size vs Pre-molt size.



Normality for the residuals:



A scatter plot was plotted for Residues against Pre-molt size:



Appendix C: Code

```
library("readxl")
library(tidyverse)
```

```
data <-
read_excel("D:/MSDS/MTH522/assigment1/crab_molt_data_vislavath_likh
il.xls")
PostMolt <- data$`Post-molt`
PreMolt<- data$`Pre-molt`

View(data)
attach(data)
names(data)
library(moments)
#ProMolt
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)

#histogram plot of PostMolt
hist(PostMolt, freq=F, las=1,ylim=c(0,0.040),col="red")
lines(density (PostMolt),col="red",lwd=3)
#histogram plot of PreMolt
hist (PreMolt, freq=F,las=1,ylim=c(0,0.040),col ="blue")
lines(density(PreMolt),col="blue",lwd=3)
#overlap the two histograms
```

```
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))
# density plot
plot(density (PostMolt),col="red",lwd=3,main="Density Plots of
PostMolt&PreMo
lt")
lines(density(PreMolt),col="blue",lwd=3)
#plot the dependent variable(PreMolt) as a function of i
plot (PostMolt, PreMolt, main= "ScatterPlot")
#plot the least square linear regression
data
model <- lm (PreMolt ~ PostMolt)
summary(model)
abline (model,col="red", lwd =3)
#Now we calculate find the Pearsons r^2 regression
results <- cor.test (PreMolt, PostMolt, method = "pearson" )
results
# residuals
residuals <- model$residuals
sapply (residuals, sum)
#Plotting the residuals
hist (residuals, freq=F,las=1,col = "red",ylim=c(0,0.20))
#Plotting the density line for the residuals
plot(density(residuals), col= "red" ,lwd=3,ylim =c(0,0.20),main="Density
Plot of Residuals")
lines(density(residuals),col= "red", 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
```

```
par (mfrow = c(2,2))
r_model <- lm (PreMolt~residuals)
summary(r_model)
plot(r_model)
```