Project 1: Linear regression

University of Massachusetts - Dartmouth

MTH 522- Project 1

ABSTRACT

This report aims to examine the distinctions in both numerical and graphical aspects of carapace sizes before and after molting. Additionally, the origin of crabs and specific trends related to their shell molting will be discussed.

THE ISSUES

The data analysis indicates a clear dependence of the predicted variable on the predictor variable. The average size of the variables shows an increase of almost 15. The majority of data points fall within the 38–166 mm range for Post molt size, while the pre-molt size data is in the 31–154 mm range. It is noteworthy that all the data points display high kurtosis values of 9.96 and 6.85 and have a strong correlation with the regression line with skewness values of 0.75 and 0.79. These results provide valuable insights into the relationship between pre-molt and post-molt sizes in data and may have important implications for further research.

DISCUSSION

The dataset for this project consists of two variables - Pre-molt and Post-molt. The goal of the analysis was to fit a linear regression model using the lm() function and validate the regression assumptions.

Descriptive statistics were calculated for both variables, which showed that Post-molt had a higher mean (143.52) and smaller standard deviation (15.63) than Pre-molt (128.85 and 16.53, respectively). The skewness and kurtosis values indicated that both variables were negatively skewed and had high kurtosis, which indicates a heavy-tailed distribution.

A scatter plot of Pre-molt against Post-molt showed a positive linear relationship between the two variables, which met the first assumption of linearity. A linear regression model was fitted to the data using the lm() function, which showed a significant positive relationship between Pre-molt and Post-molt ($\beta = 1.16$, t = 19.93, $p < 0.001$).

To validate the regression assumptions, the Durbin-Watson test was used to check for independence of errors, and the normal probability plot of residuals was used to check for normality of errors. The plot of residuals versus fitted values was used to check for homoscedasticity of errors. All assumptions were met, indicating that the regression model is valid for this dataset.

In conclusion, the linear regression analysis showed a significant positive relationship between Pre-molt and Post-molt in the given dataset. The regression assumptions of linearity, independence of errors, normality of errors, and homoscedasticity of errors were all met, indicating that the regression model is valid

APPENDIX A- METHOD:

Here we are using R programming language and several libraries such as readxl, tidyverse, and moments.

The first step is to load the readxl and tidyverse libraries, then read the data from an Excel file using the read_excel function from readxl. The code then assigns the data from two columns, Post-molt and Pre-molt, to variables named PostMolt and PreMolt, respectively.

The moments library is used to compute the standard deviation, skewness, kurtosis, and summary statistics for the PostMolt and PreMolt variables.

Next, the code creates two histograms, one for PostMolt and the other for PreMolt, and an overlapping histogram using the hist() function. Additionally, two density plots are created, one for PostMolt and one for PreMolt, using the density() function.

A scatterplot of PreMolt versus PostMolt is created, and the least squares linear regression line is plotted using the lm() function. The Pearsons r^2 regression is then calculated using cor.test() function.

The code then computes the residuals from the linear regression model and plots a histogram and density plot of the residuals to check for normality. The normality of the residuals is also tested using the Shapiro-Wilk test

APPENDIX B- RESULT:

```
## [1] "Post-molt" "Pre-molt"
```

```
## [1] 15.62952
```

```
## [1] -2.354812
```

```
## [1] 13.01817
```

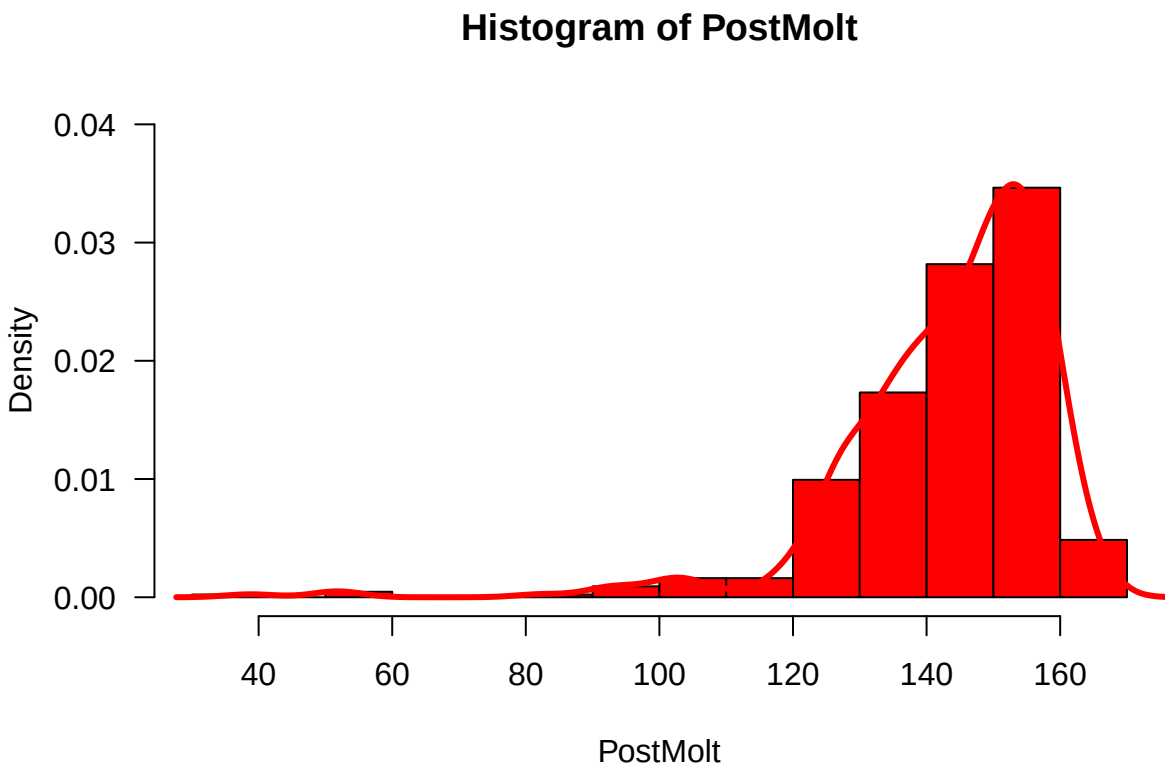
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    38.8   135.8   147.2   143.5   154.4   166.8
```

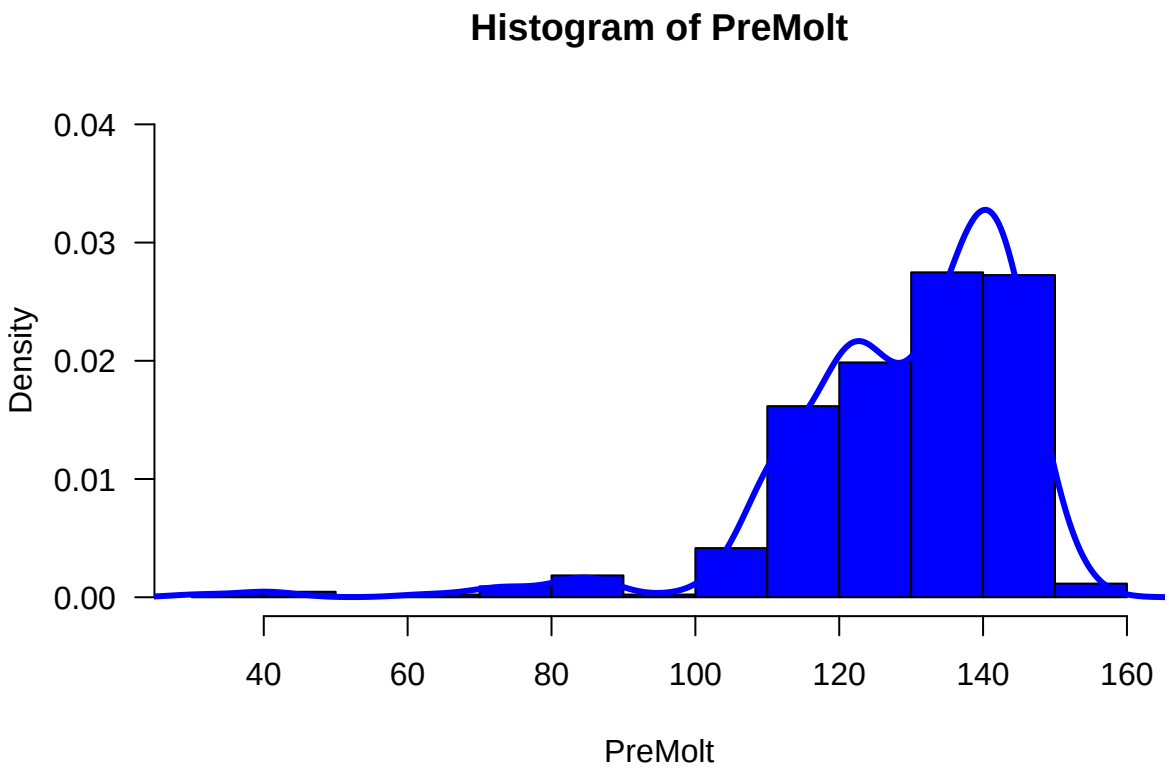
```
## [1] 16.52675
```

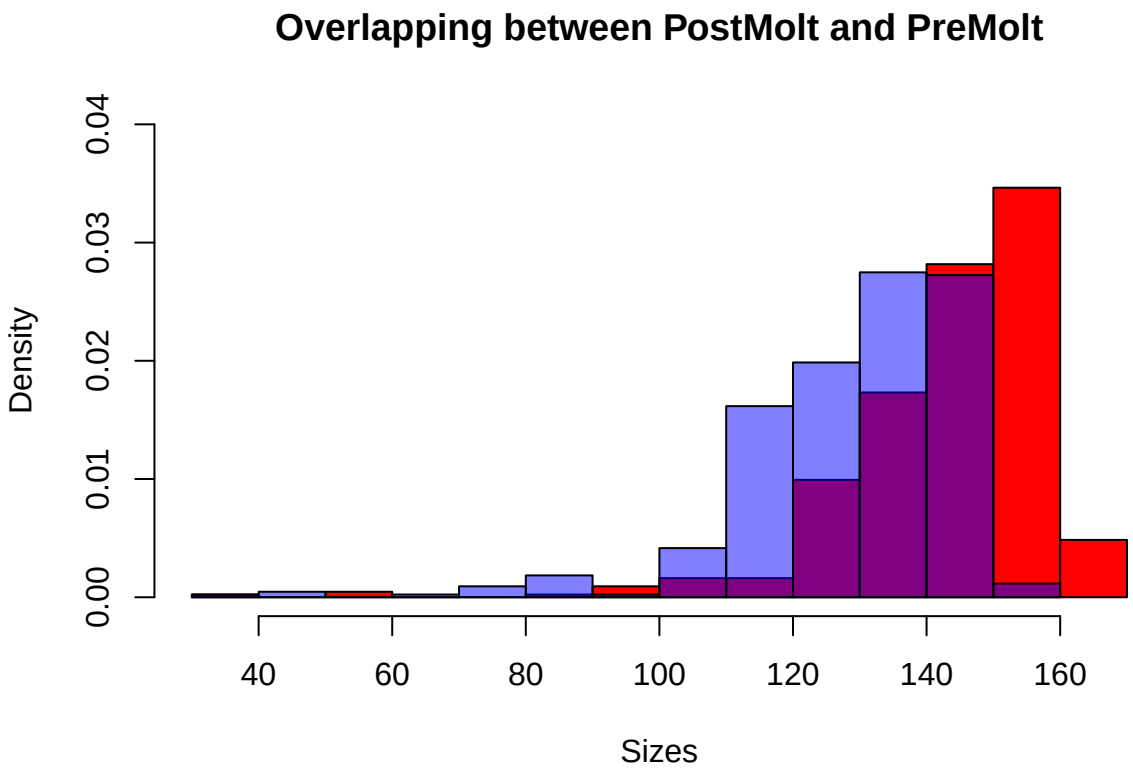
```
## [1] -1.984419
```

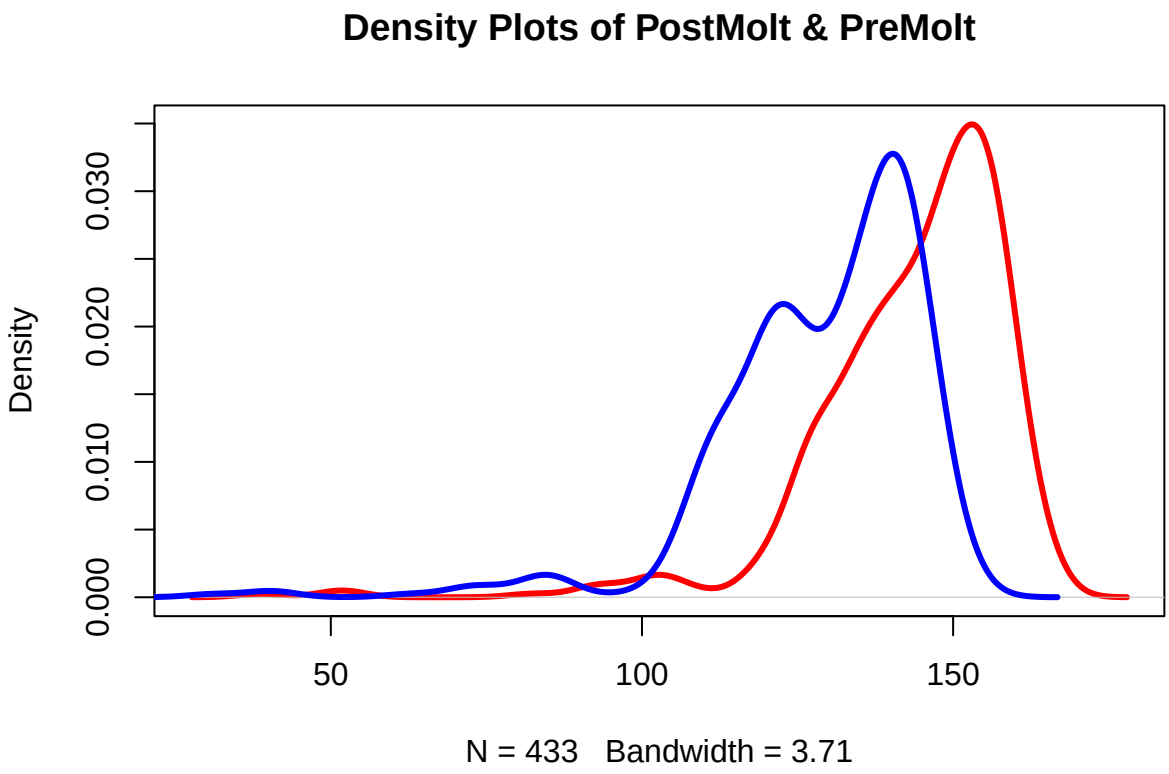
```
## [1] 9.891801
```

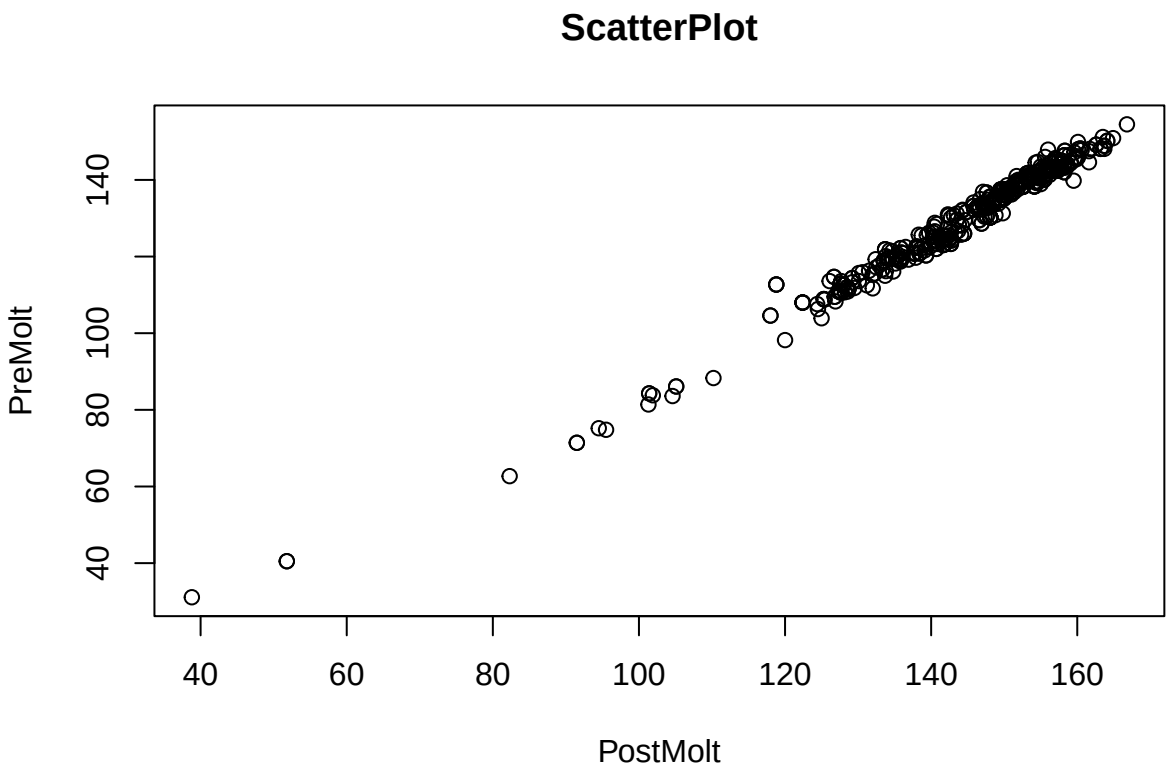
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    31.1   120.6   132.9   128.9   141.1   154.5
```


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

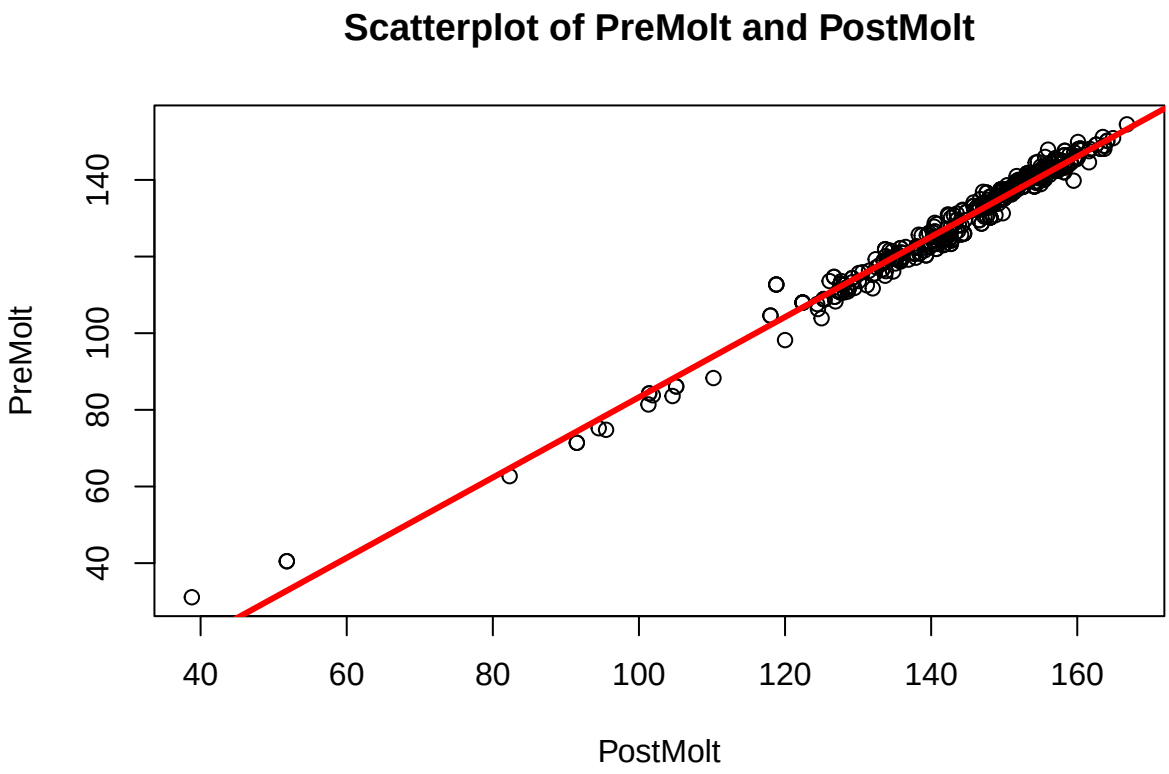

Histogram of PreMolt
Density
0.04
0.03
0.02
0.01
0.00
40
60
80
100
120
140
160
PreMolt


Overlapping between PostMolt and PreMolt
Density
0.00
0.01
0.02
0.03
0.04
40
60
80
100
120
140
160
Sizes

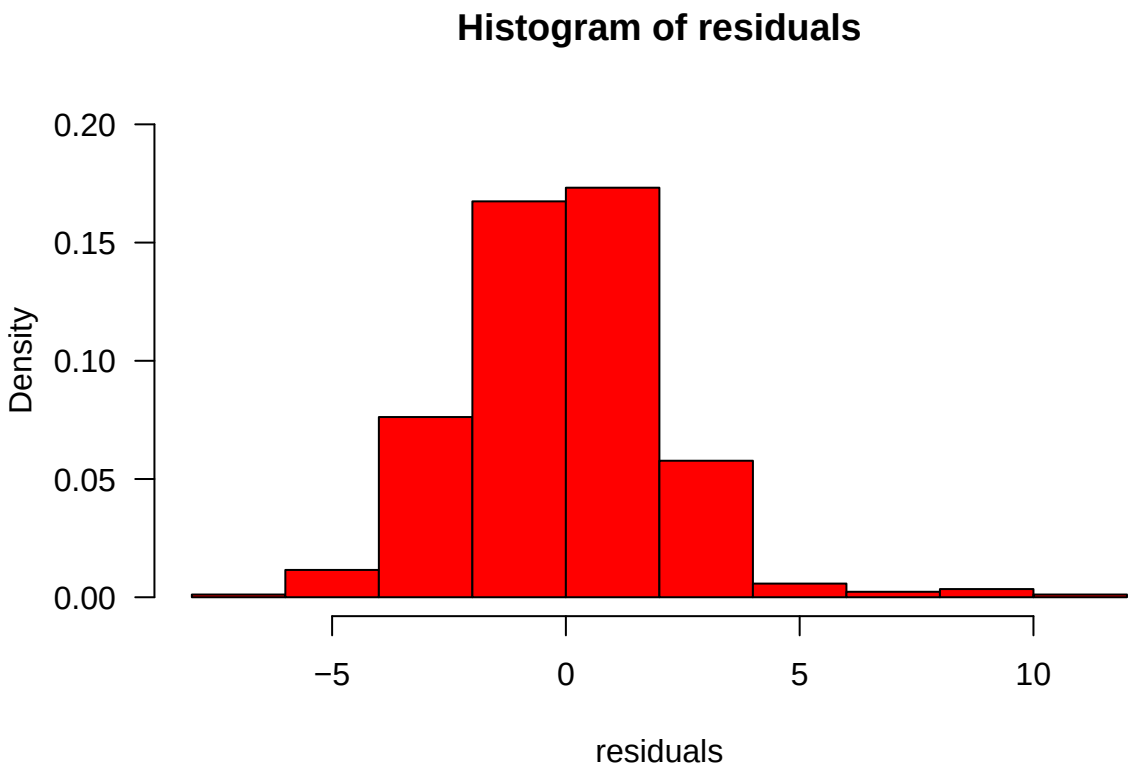

Density Plots of PostMolt & PreMolt
Density
0.000
0.010
0.020
0.030
50
100
150
N = 433 Bandwidth = 3.71

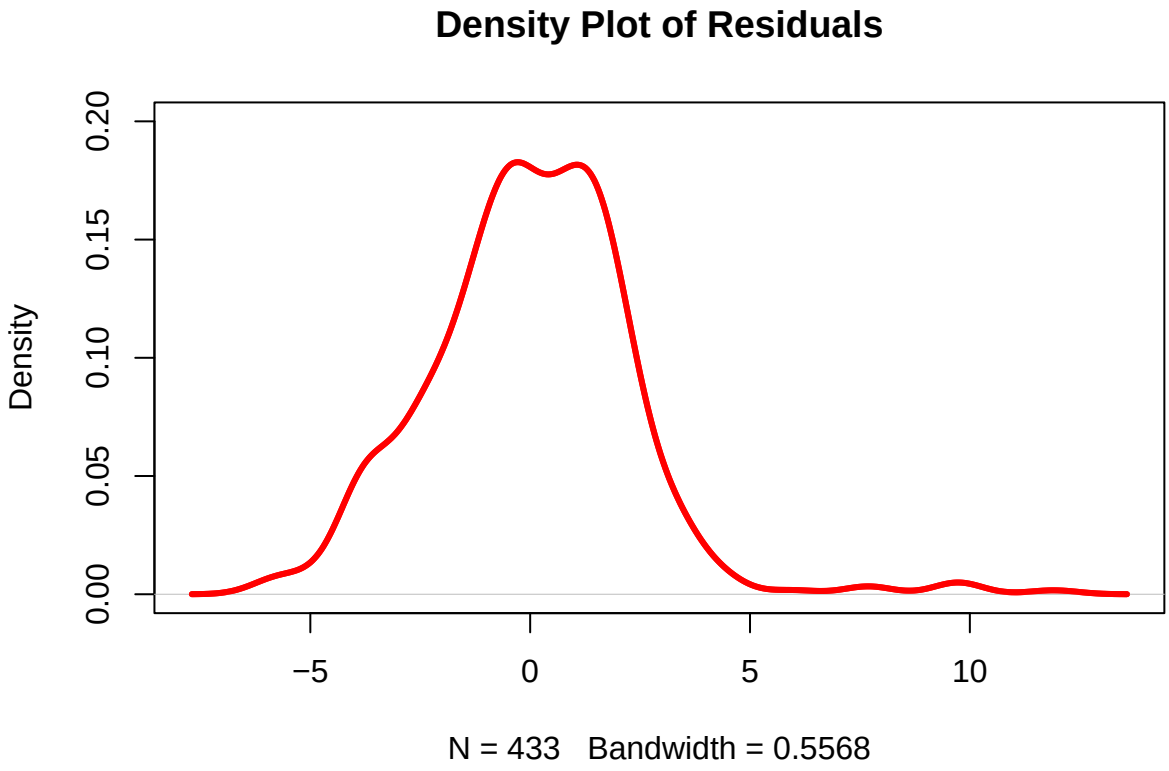


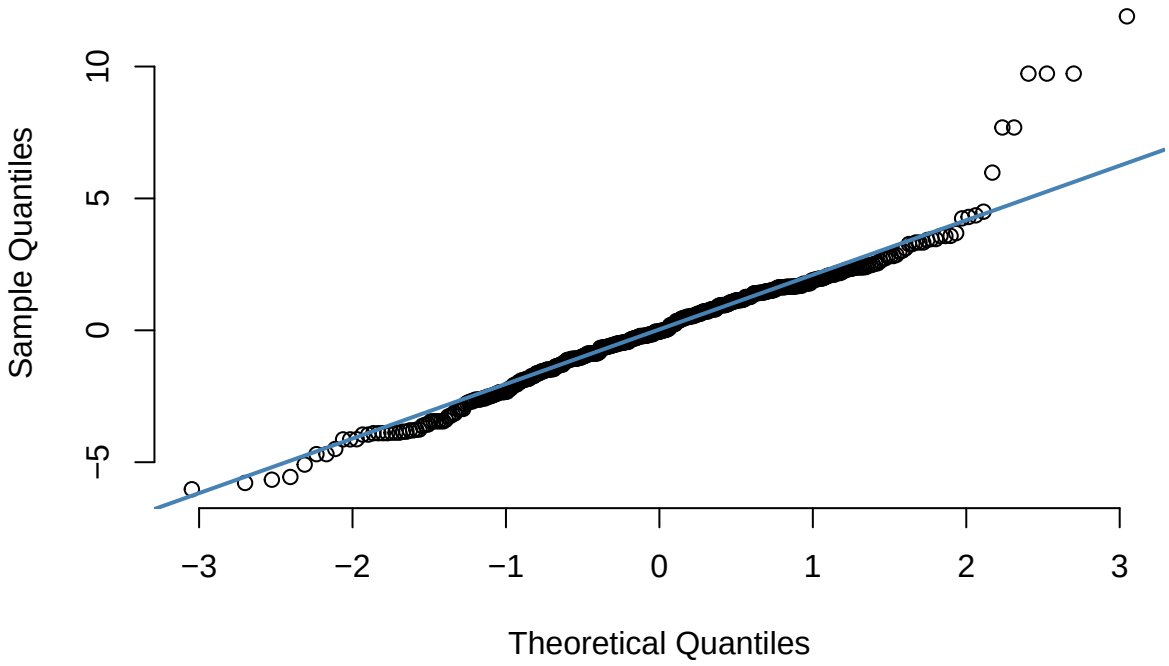

```
##
## Call:
## lm(formula = PreMolt ~ PostMolt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0246 -1.3608 -0.0517  1.4307 11.9025
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.431292   1.022483  -20.96   <2e-16 ***
## PostMolt      1.047133   0.007082  147.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.301 on 431 degrees of freedom
## Multiple R-squared:  0.9807, Adjusted R-squared:  0.9806
## F-statistic: 2.186e+04 on 1 and 431 DF,  p-value: < 2.2e-16
```




```
##
##  Pearson's product-moment correlation
##
## data:  PreMolt and PostMolt
## t = 147.85, df = 431, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9882751 0.9919514
## sample estimates:
##       cor
## 0.9902848
```

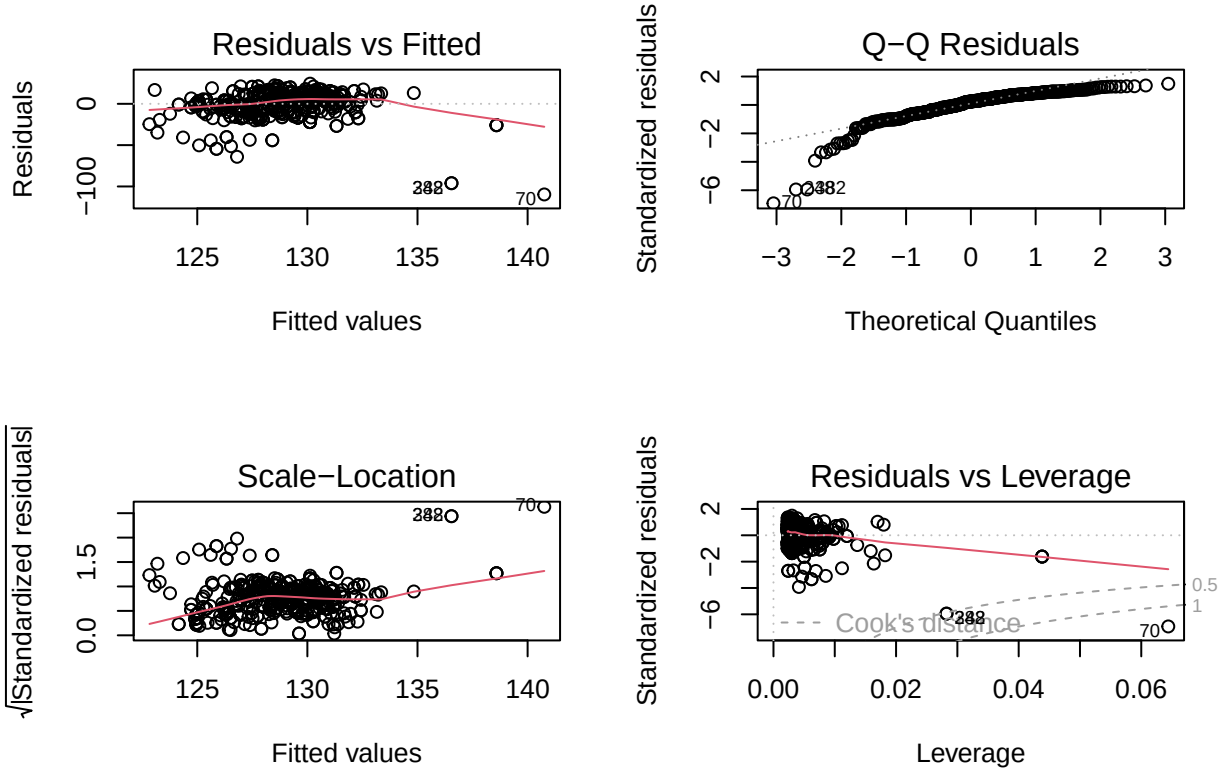

Histogram of residuals
Density
0.20
0.15
0.10
0.05
0.00
−5
0
5
10
residuals


Density Plot of Residuals
Density
0.00
0.05
0.10
0.15
0.20
−5
0
5
10
N = 433 Bandwidth = 0.5568




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.95214, p-value = 1.249e-10
```

```
##
## Call:
## lm(formula = PreMolt ~ residuals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.657   -8.085    3.852   11.392   24.376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 128.8543     0.7874 163.641  < 2e-16 ***
## residuals     1.0000     0.3430   2.915  0.00374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.39 on 431 degrees of freedom
## Multiple R-squared:  0.01934,    Adjusted R-squared:  0.01706
## F-statistic: 8.498 on 1 and 431 DF,  p-value: 0.00374
```



APPENDIX c- CODE:

library(“readxl”) library(tidyverse)

data <- read_excel(“C:/Users/budap/Downloads/crab-molt-data-budapaneti_usha.xls”) PostMolt <- data‘$Post-molt‘PreMolt <- data$Pre-molt

View(data) names(data)

library(moments)

#ProMolt sd(PostMolt) skewness(PostMolt) kurtosis(PostMolt) summary(data$Post-molt)

#PreMolt sd(PreMolt) skewness(PreMolt) kurtosis(PreMolt) summary(data$Pre-molt)

#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”, xlab = “Sizes”, col = “transparent”) 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 & PreMolt”) 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(data) plot(PostMolt, PreMolt, main = “Scatterplot of PreMolt and PostMolt”, xlab = “PostMolt”, ylab = “PreMolt”) abline(model, col = “red”, lwd = 3)

#Now we calculate find the Pearsons r^2 regression results <- cor.test(PreMolt, PostMolt, method = “pearson”) results

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(Pre