Analysis to Determine the Growth Pattern of Female Dungeness Crab using Pre-molt and Post-molt sizes

---

1 The Issues

In U.S. waters, nearly the entire adult male Dungeness crab population is fished each year. Female crabs are not fished in order to maintain the viability of the crab population. The great imbalance in the sex ratio of crabs have contributed to the decline in the crab population. Size restrictions on male crabs are set to ensure that they have at least one opportunity to mate before being fished. To help determine similar size restrictions for female crabs, more needs to be known about the female crab's growth. We address the following questions: 1. What is the relation between "Post-molt" size and "Pre-molt" size? 2. Is there any heteroscedasticity in the residuals and how that might affect the accuracy of linear model of prediction?

2 Findings

Firstly, when I calculated the descriptive statistics and plotted the Histograms of both Pre-molt size and Post-molt size data, I found that there is a considerable difference in the sizes of Pre molt and Post molt.

The mean post-molt size of crabs is 145.26 mm, while the mean pre-molt size is 130.63 mm. This suggests that crabs grow by about 14.63 mm on average between molts. The standard deviation of post-molt size is 13.21 mm, which indicates that there is some variation in the size of crabs after molting. The standard deviation of pre-molt size is 14.45 mm, which suggests that there is even more variation in crab size before molting. The median post-molt size is 148.05 mm, and the median pre-molt size is 133.4 mm. The skewness values of -2.47 and -2.21 for the post-molt and pre-molt columns, respectively, indicate that the distribution is negatively skewed. What that means is, there are more crabs with smaller sizes than larger sizes.

The R-squared value is 0.9896098, it means that 98.96 percentage of the variability in the response variable (crab sizes in this case) can be explained by the predictor variable (post-molt/pre-molt). This indicates a strong correlation between the two variables, and as

a result, any predictions made using the regression model will likely have a high degree of accuracy. Even though this model has a 0.98 R- squared value, we cannot directly trust these results. On performing Shapiro Wilk test on residuals, the p-value was obtained to be 6.059e-12. This is a very small value and as it is less than 0.05 I found that the data is not normally distributed and therefore cannot be statistically significant. After plotting the residuals, I observed that the residual plots are displaying problematic patterns leading to a biased model and are not normally distributed. Hence, it's a biased model, I cannot trust the results. That means linear regression model is not the right fit for this crab data.

3 Discussions

The accuracy of the predictions can be affected by other factors that influence the size of the crabs. Going forward with the Simple linear Regression model with the dependent variable as Pre molt size and Independent variable as Post molt size will not be statistically significant.

4 Appendix A: Method

Data was downloaded as a comma-separated (.csv) file and imported into R Studio. The CSV file consists of the Dungeness crabs' size data Post-molt and Pre-molt. Firstly, plotted the Post-molt vs Pre-molt data and found the summary of the data such as minimum, maximum, mean, and median, standard deviation, skewness, and kurtosis. By applying the probability density function, I plotted the histogram of each variable individually and overlaying both. A scatter plot has been plotted for post and pre-molt data to analyze the relationship between dependent and independent variables. Plotted the least square linear regression on the same plot with "Post-molt" size and "Pre-molt size". Calculated the Pearson r^2 correlation. Performed the descriptive statistics for the residuals which were found in the least square regression and plotted the residuals with help of histogram plot and density lines and checked for their normality by using Quantile plot test and Shapiros Walks test. Then, checked for heteroskedasticity by plotting the residuals against the Pre-molt size (dependent variable).

5 Appendix B: Results

There are a total of 436 observations for both post-molt and pre-molt sizes.

On applying the descriptive statistics on the both the variables, I found the insights in Figure 1

	PreMolt	PostMolt
Minimum	31.1	38.8
Maximum	155.1	166.8
Median	133.4	148.05
Mean	130.63	145.26
Standard Deviation	14.45	13.21
Skewness	-2.21	-2.47
Kurtosis	11.48	15.11

Figure 1: Descriptive Statistics of the data

A smooth histograms for each variable, overlaid, shows the comparision clearly in figure 2

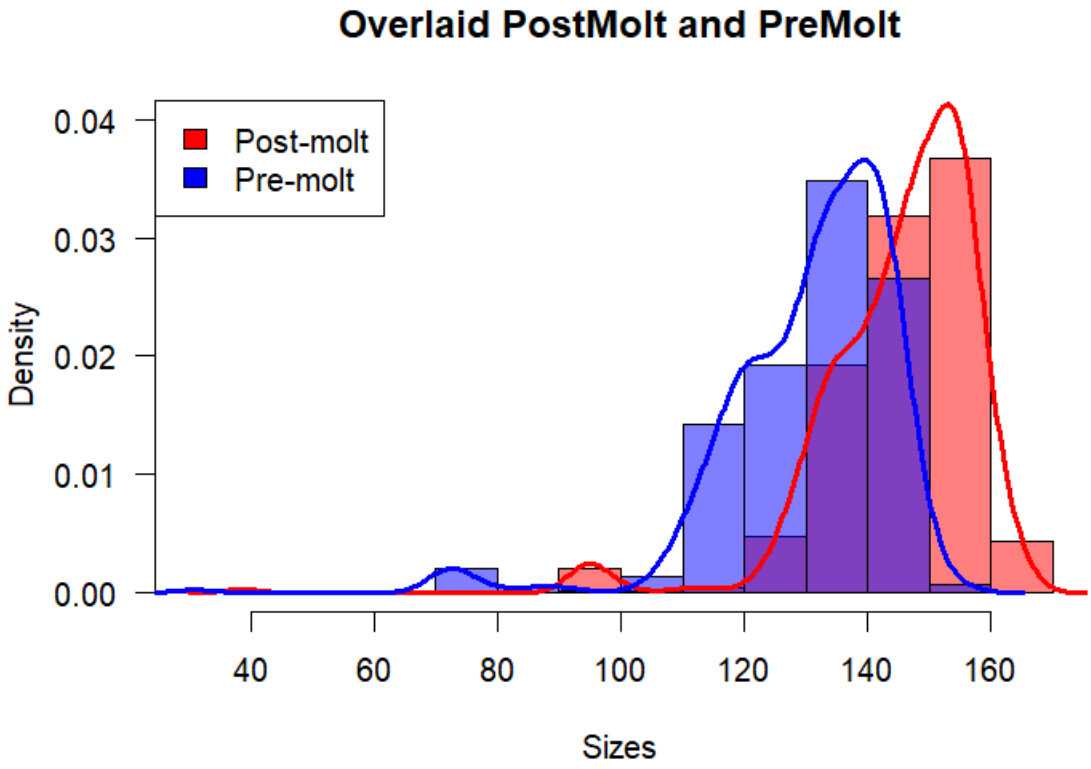

Figure 2: Overlaid Histogram representation of Pre molt and Post molt data

After this I plotted the Scatter plot with Pre-Molt as a function of Post-Molt and applied the least square linear regression on the same plot. Refer Figure 3

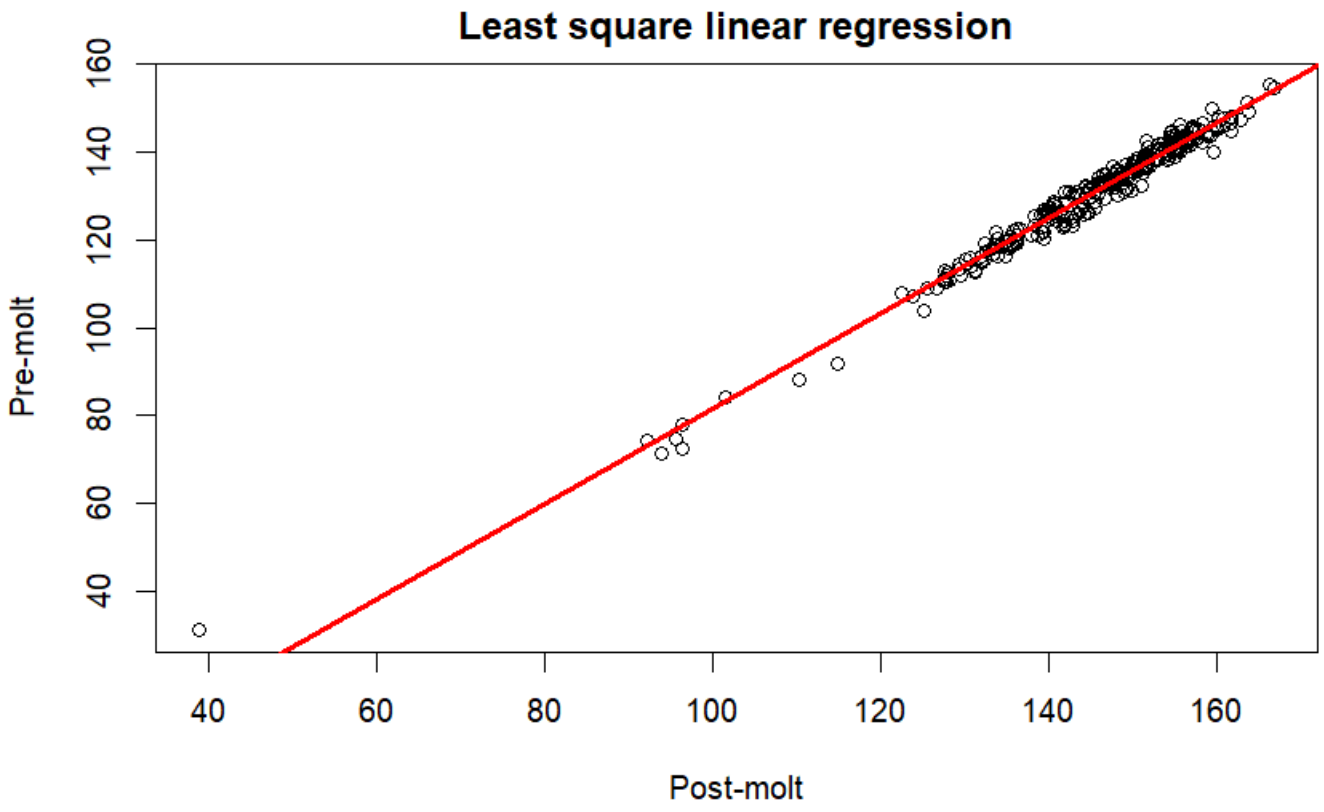

Figure 3: Scatter Plot along with the least square linear regression of PreMolt (dependent variable) as a function of Post Molt (independent variable)

To check the normality of these residuals I plotted Quantile plots. Refer Figure 4

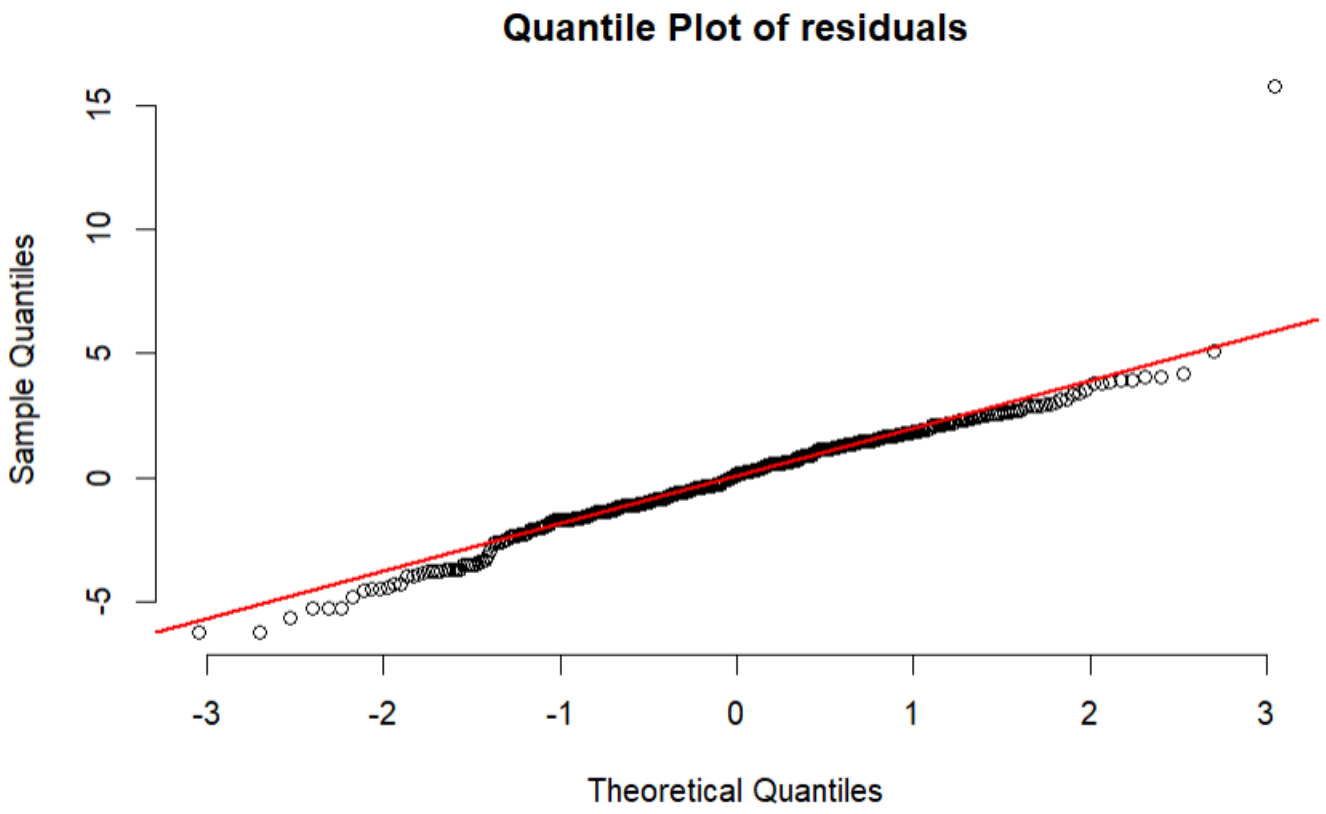

Figure 4: Quantile Plots for the residuals to check the normality

Plotted residuals against the dependent variable (Pre-molt) and visually checked for heteroskedasticity. Refer figure 5

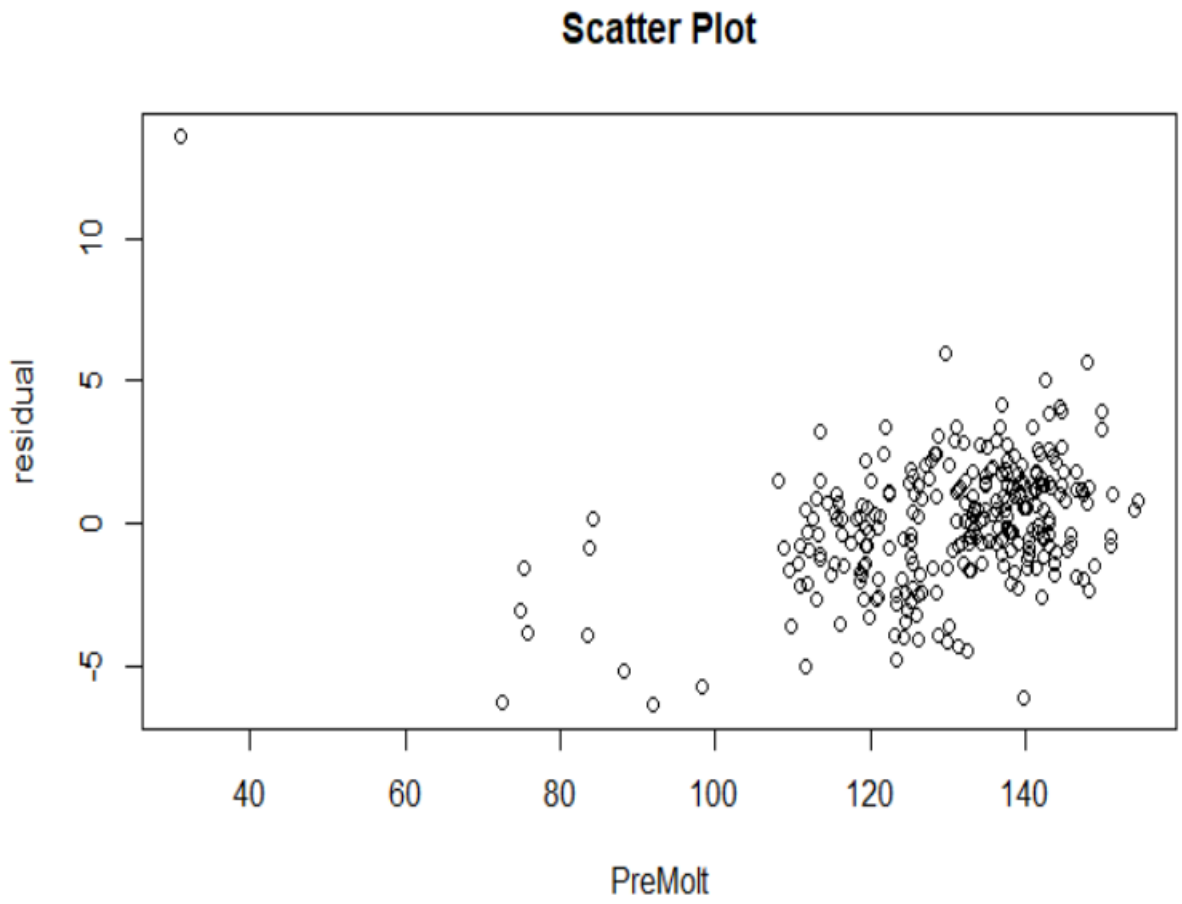

Figure 5: scatter plot of the residuals and PreMolt

6 Appendix C: Code

```
#Finding Descriptive statistics for the Data

#Summary of Post molt data
library(moments)
min('Post-molt')
max('Post-molt')
median('Post-molt')
mean('Post-molt')
sd('Post-molt')
skewness('Post-molt')
kurtosis('Post-molt')

#Summary of Pre molt data
library(moments)
```

```
min(`Pre-molt`)
max(`Pre-molt`)
median(`Pre-molt`)
mean(`Pre-molt`)
sd(`Pre-molt`)
skewness(`Pre-molt`)
kurtosis(`Pre-molt`)


#Making Probability Density Function (PDF) histogram for each variable

# For Post molt
hist(`Post-molt`, freq=F, las=1,ylim=c(0,0.040),col="red")
lines(density (`Post-molt`),col="red",lwd=3)

# For Pre molt
hist(`Pre-molt`, freq=F,las=1,ylim=c(0,0.040),col ="blue")
lines(density(`Pre-molt`),col="blue",lwd=3)


#Overlaid PDF histograms so that the difference is visible to nbaked eye
hist(`Post-molt`,freq=F,ylim = c(0,0.040),main ="Overlaid PostMolt and
PreMolt", xlabel ="Sizes",col=rgb(1,0,0,0.5),las=1)
hist(`Pre-molt`,freq = F,add=TRUE,col = rgb(0,0,1,0.5))


#Density plot for Overlaid histograms
plot(density(`Post-molt`),col="red",lwd=3,main="Density Plots of PostMolt & P
lines(density(`Pre-molt`),col="blue",lwd=3)

#Combined plot for above
par(mar = c(5, 4, 4, 8) + 0.1)
hist(`Post-molt`, freq = F, ylim = c(0, 0.040), main = "Overlaid PostMolt and
hist(`Pre-molt`, freq = F, add = TRUE, col = rgb(0, 0, 1, 0.5))
lines(density(`Post-molt`), col = "red", lwd = 3)
lines(density(`Pre-molt`), col = "blue", lwd = 3)
legend("topleft", c("Post-molt", "Pre-molt"), fill = c("red", "blue"))


#Scatter Plot of dependent variable(PreMolt) as a function of independent var
plot (`Post-molt`, `Pre-molt`, main= "ScatterPlot")

# Ploting least square linear regression model on the above
model <- lm (`Pre-molt` ~ `Post-molt`)
summary(model)
```

```
plot (`Post-molt`, `Pre-molt`, main= "Least square linear regression")
abline(model,col="red", lwd =3)


#Finding Pearsons r   (Corelation Coefficient)
results <- cor.test (`Pre-molt`, `Post-molt`, method = "pearson" )
results


#descriptive statistics of the residuals
residuals <- model$residuals
sapply (residuals, sum)


#Histogram plot for the residuals
hist (residuals, freq=F,las=1,col = "yellow",ylim=c(0,0.20))

#Plotting Density Line of residuals
plot(density(residuals), col= "green" ,lwd=3,ylim =c(0,0.20),main="Density Pl
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= "red", lwd=2)


#Performing Shapiro Wilks test
shapiro.test((residuals))


#Ploting residuals against the dependent variable (PreMolt)
par (mfrow = c(2,2))
r_model <- lm (`Pre-molt`~residuals)
summary(r_model)
```