Crab Growth Data Regression Analysis

By Edward Huynh

Feb 16, 2020

1. Abstraction: In this statistical report, I will examine the relationship between premolt and postmolt carapace size in both numerically and graphically. The goal of this lab is to create histogram for the size distribution of the crabs before the molting season, with the shaded region representing the molted crabs. The data where both premolt and postmolt sizes are available can be used to determine the relationship between a crab's premolt and postmolt size, and this relationship can be used to develop a method for predicting a crab's premolt size from its postmolt size.

2. Introduction: The data in this report was obtained from a study of adult female Dungeness crabs. It contains premolt and postmolt width sizes of the shells of 472 crabs.I use the R Studio Integrated Development Environment programing language to analyse this report both in numerically and graphically. This R Script is created that allowed to perform all statistical operations as well as perform work with linear regression.

3. Methods and Results

- *Data Collecton:* The data for this lab were collected as part of a study of the adult female Dungeness crab. Two sets of data are provided in "crabs.data" from the Stat Lab. The first consists of premolt and post molt widths of the shells of 472 female Dungeness crabs in northern California and Southern Oregon. The data were obtained over three fishing seasons. The first two were in 1981 and 1982; the third, in 1992.
- *Analsis" I use R studio to import the dataset "crabs.data", then analyze the investigation from the Stat Lab textbook as steps by steps below:

Part1: Begin by considering the problem of predicting the premolt size of a crab given only its postmolt size. Develop a procedure for doing this, and derive an expression for the average squared error you expect in such a prediction.

Import data

```
data<-read.table("C:/Users/Sang/Documents/crabs.data", header=TRUE,sep="")
str(data)
```

```
## 'data.frame':    472 obs. of  5 variables:
##  $ presz : num  114 118 120 126 127 ...
##  $ postsz: num  128 133 135 143 139 ...
##  $ inc   : num  14.1 15.1 15.4 17.1 12.6 12.9 15.6 15.1 17.1 13.2 ...
##  $ year  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ lf    : int  0 0 0 0 0 0 0 0 0 0 ...
```

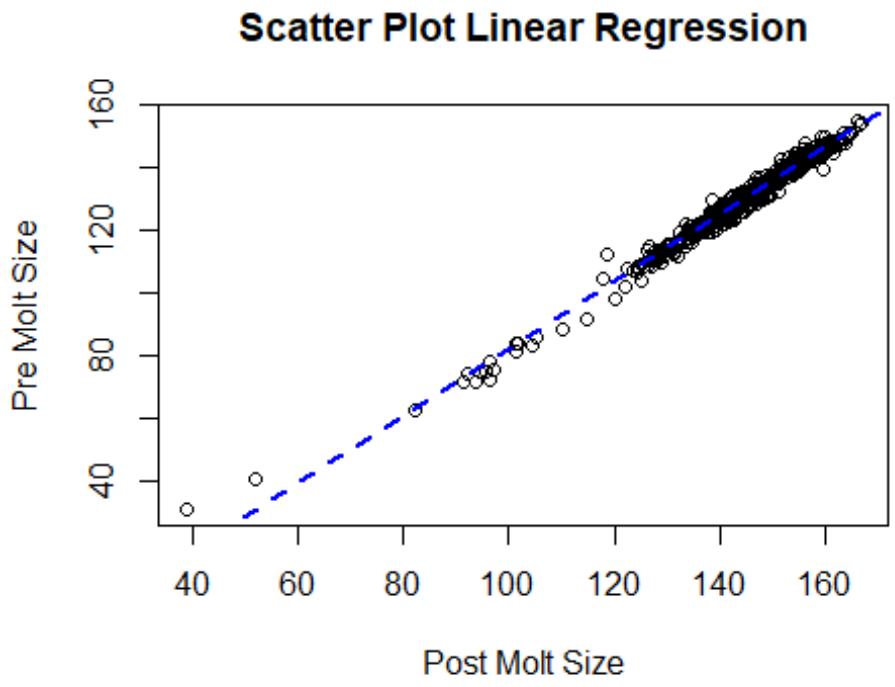
Extract Pre-molt and Post-molt data

```
moltdata <- subset(data, select = c("postsz","presz"))
summary(moltdata)
```

```
##      postsz          presz
##  Min.   : 38.8   Min.   : 31.1
##  1st Qu.:138.0   1st Qu.:121.7
##  Median :147.4   Median :132.8
##  Mean   :143.9   Mean   :129.2
##  3rd Qu.:153.4   3rd Qu.:140.0
##  Max.   :166.8   Max.   :155.1
```

Scattter Plot Pre-molt and Post-molt size data

```
plot(presz ~ postsz, data = moltdata, ylab = "Pre Molt Size", xlab ="Post
Molt Size", main = "Scatter Plot Linear Regression")
mod<-lm(presz~postsz, data=moltdata)
abline(mod, col = "blue", lty = 2, lwd = 2)
```



Summary mod and Pearson's method for the line of least squares to show average square error

```
summary(mod)
```

```
##
## Call:
## lm(formula = presz ~ postsz, data = moltdata)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -6.1557 -1.3052  0.0564  1.3174 14.6750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.21370    1.00089  -25.19   <2e-16 ***
## postsz       1.07316    0.00692  155.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.199 on 470 degrees of freedom
## Multiple R-squared:  0.9808, Adjusted R-squared:  0.9808
## F-statistic: 2.405e+04 on 1 and 470 DF,  p-value: < 2.2e-16
```

```
attributes(mod)
```

```
## $names
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
##
## $class
## [1] "lm"
```

```
mod$coefficients
```

```
## (Intercept)      postsz
##  -25.213703    1.073162
```

```
confint(mod)
```

```
##                   2.5 %    97.5 %
## (Intercept) -27.180474 -23.24693
## postsz        1.059565   1.08676
```

```
r=with(data=moltdata,cor(presz,postsz))
rsq=r^2
```

Since R^2 = 1 means no error, my average square error shows R^2 = 0.981 which means the error is extremely low for the linear regression line fairly accurate model.

Part 2: Examine a subset of the data collected, say those crabs with postmolt carapace width between 147.5 and 152.5 mm. Compare the predictions of premolt size for this subset with the actual premolt size distribution of the subset. Do this for one or two other small groups of crabs.

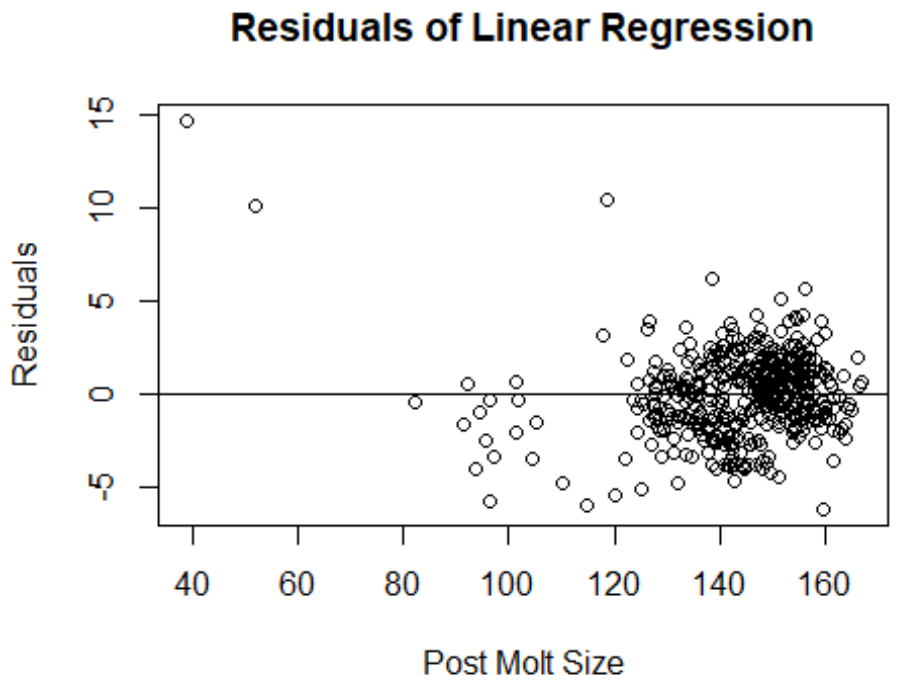
Import Postmolt Data

```
postmolt<-read.table("C:/Users/Sang/Documents/crabpop.data",
header=TRUE,sep="")
str(postmolt)
```

```
## 'data.frame':    362 obs. of  2 variables:
##  $ size : num  117 117 118 120 120 ...
##  $ shell: int  1 1 1 1 1 1 1 1 1 ...

postmoltsize<-subset(postmolt, select=c("size"))
str(postmoltsize)

## 'data.frame':    362 obs. of  1 variable:
##  $ size: num  117 117 118 120 120 ...
```

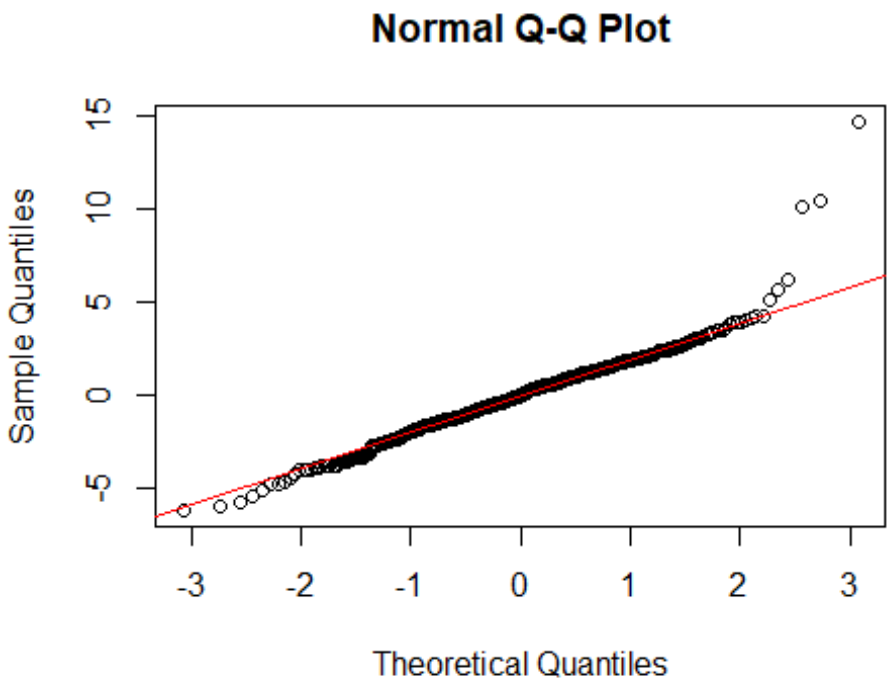
Residual Linear Regression Plot

```
m<-mod$coefficients[2]
b<-mod$coefficients[1]
premoltsize<-postmoltsize
for(i in 1:length(postmoltsize)){premoltsize[i]<-m*postmoltsize[i]+b}
mod_res<-resid(mod)
summary(mod_res)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.15570 -1.30517  0.05639  0.00000  1.31741 14.67500

plot(moltdata$postsz, mod_res, ylab="Residuals", xlab="Post Molt Size",
main="Residuals of Linear Regression")
abline(0,0)
```



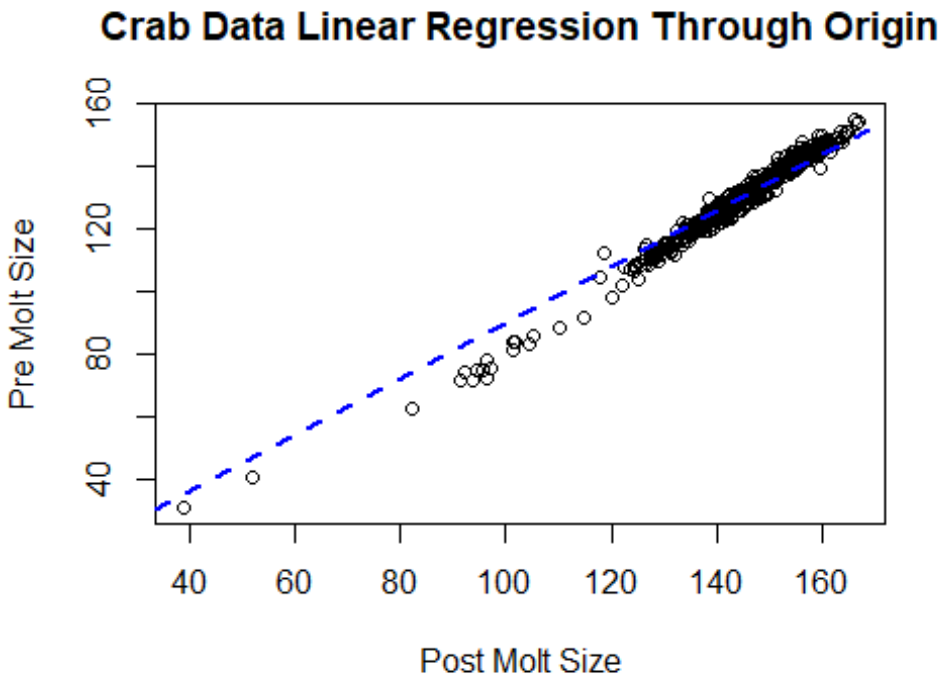
Normality Q-Q Plot

```r
qqnorm(mod_res)
qqline(mod_res, col=2)
```



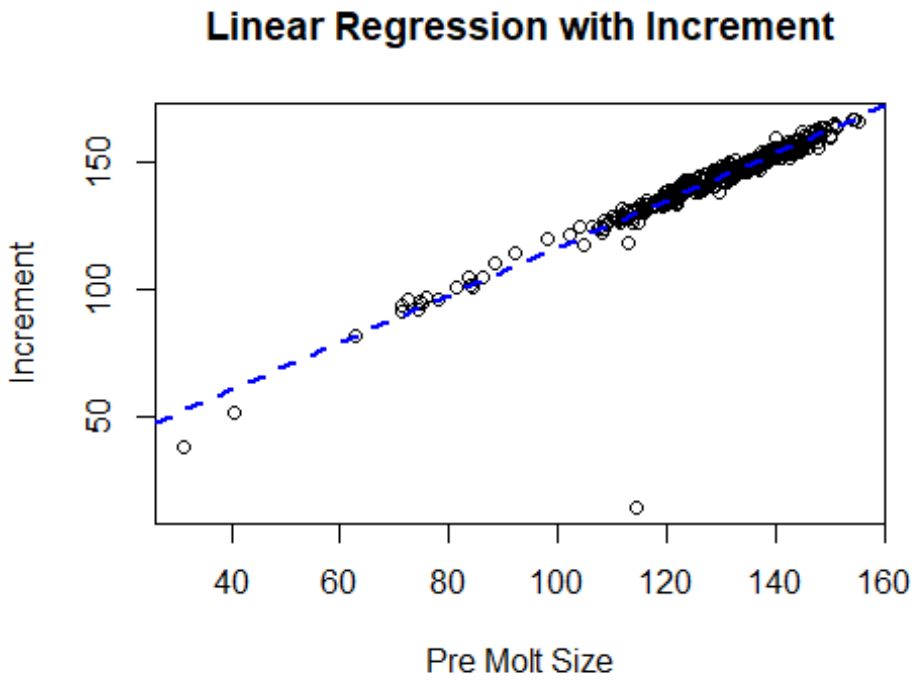
The Regression line go through origin

```r
plot(presz ~ postsz, data = moltdata, ylab = "Pre Molt Size", xlab = "Post
Molt Size", main = "Crab Data Linear Regression Through Origin")
mod.origin = lm(presz ~ postsz - 1, data = moltdata)
abline(mod.origin, col = "blue", lty = 2, lwd = 2)
```



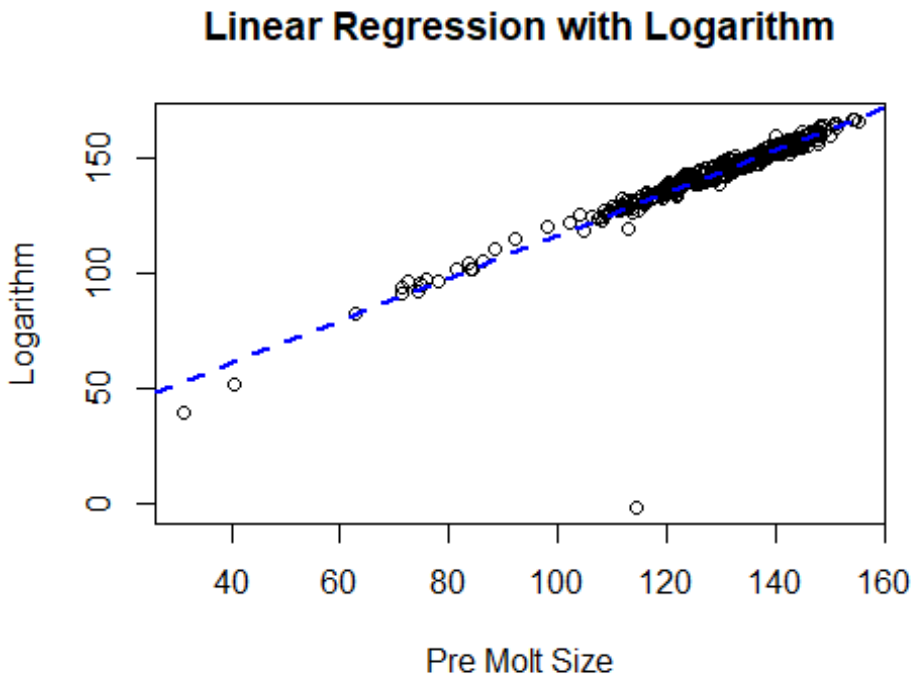
The regression line by using increment (postsz~presz)

```
increment<-moltdata$postsz
for (i in length(increment)){increment[i]<-moltdata$postsz[i]-
moltdata$presz[i]}
plot(increment ~ presz, data = moltdata, ylab = "Increment", xlab ="Pre Molt
Size", main = "Linear Regression with Increment")
mod.increment <- lm(increment ~ presz, data = moltdata)
abline(mod.increment, col = "blue", lty = 2, lwd = 2)
```



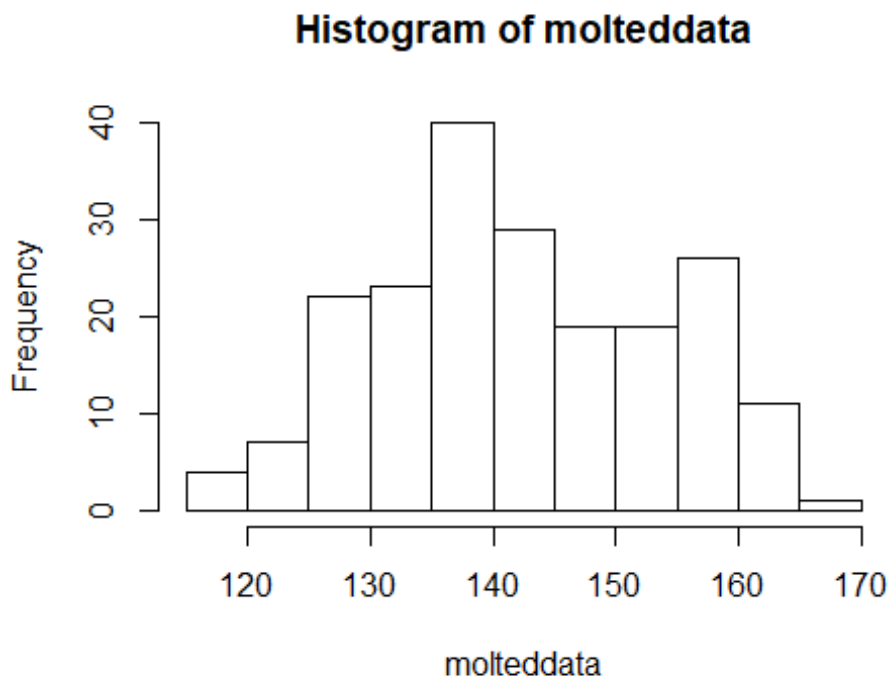
#The Regression line with logarithm

```
logarithm <- moltdata$postsz
for (i in length(logarithm)){logarithm[i]<-
log(increment[i]/moltdata$presz[i])}
plot(logarithm ~ presz, data = moltdata, ylab="Logarithm", xlab ="Pre Molt
Size", main = "Linear Regression with Logarithm")
mod.log <- lm(logarithm ~ presz, data = moltdata)
abline(mod.log, col = "blue", lty = 2, lwd = 2)
```

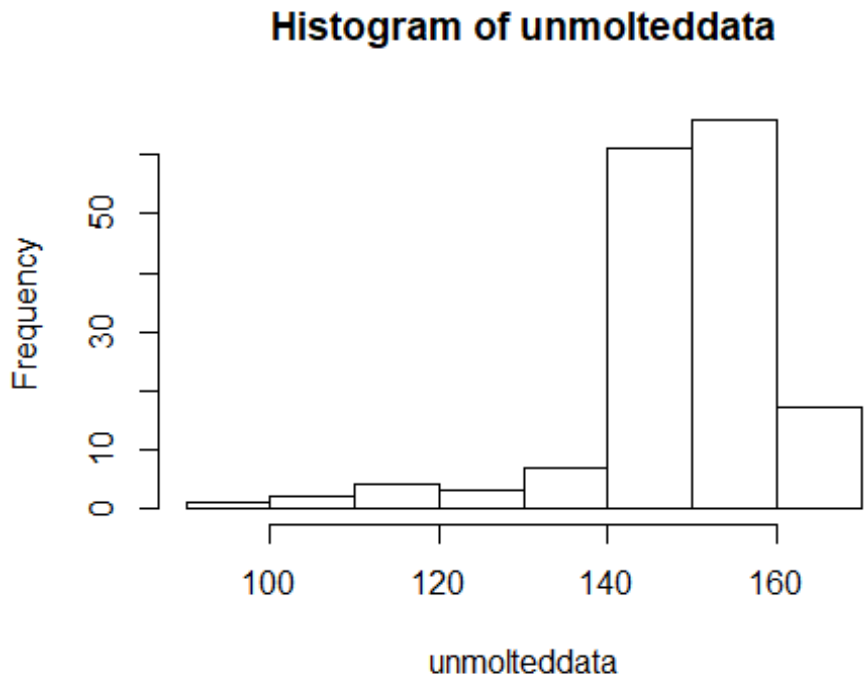


Part 3: Use your procedure to describe the premolt size distribution of the molted crabs collected immediately following the 1983 molting season. Make a histogram for the size distribution prior to the molting season of the crabs caught in 1983. Use shading to distinguish the crabs that molted from those that did not molt

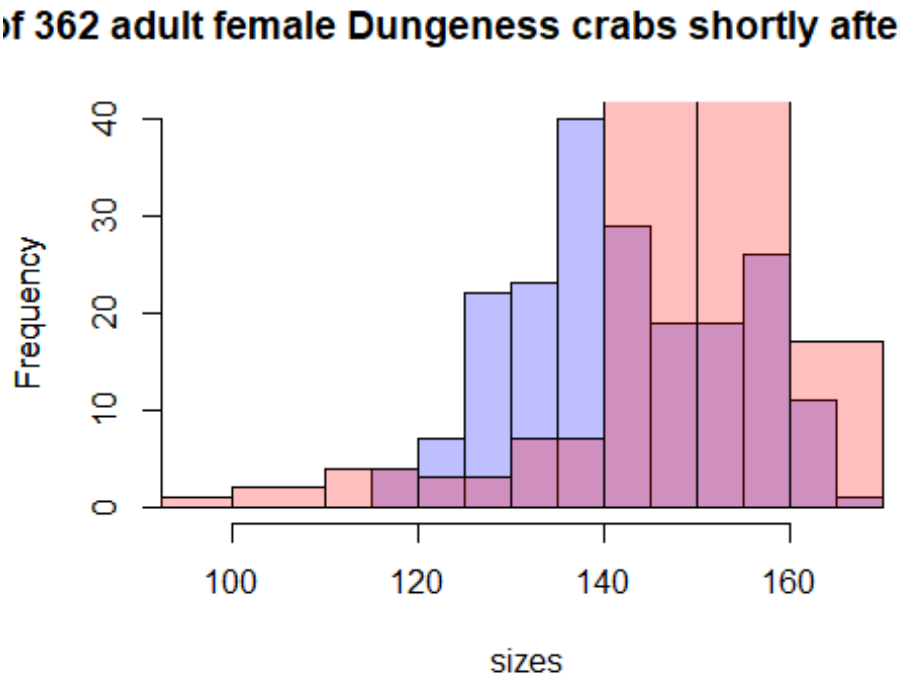
```
molted<-subset(postmolt,shell==1)
unmolted<-subset(postmolt,shell==0)
molteddata<-molted[['size']]
unmolteddata<-unmolted[['size']]
p1<-hist(molteddata)
```




```
p2<-hist(unmolteddata)
```



```r
a<-min(min(molteddata),min(unmolteddata))
b<-max(max(molteddata),max(unmolteddata))
plot(p1,col=rgb(0,0,1,1/4),main="",sub="",xlab="",xlim=c(a,b))#first
histogram
plot(p2,col=rgb(1,0,0,1/4),main="",sub="",xlab="",xlim=c(a,b),add=T)#second
histogram
title(main="Size distribution of 362 adult female Dungeness crabs shortly
after the 1983 molting season",xlab="sizes",ylab="Frequency")
```



4. Conclusion & Discussion: From the statistical analysis above, the results show me that the residuals are more variable on the frst half of data set, but more condense and less variable at the second half of dataset. In addition, the regression line fit the data well which mean that the distribution of residual data is mostly normal. The linear regression model better fits the data which give better predictions of premolt size ftom it's postmolt size.