Lab 3 Stat 135

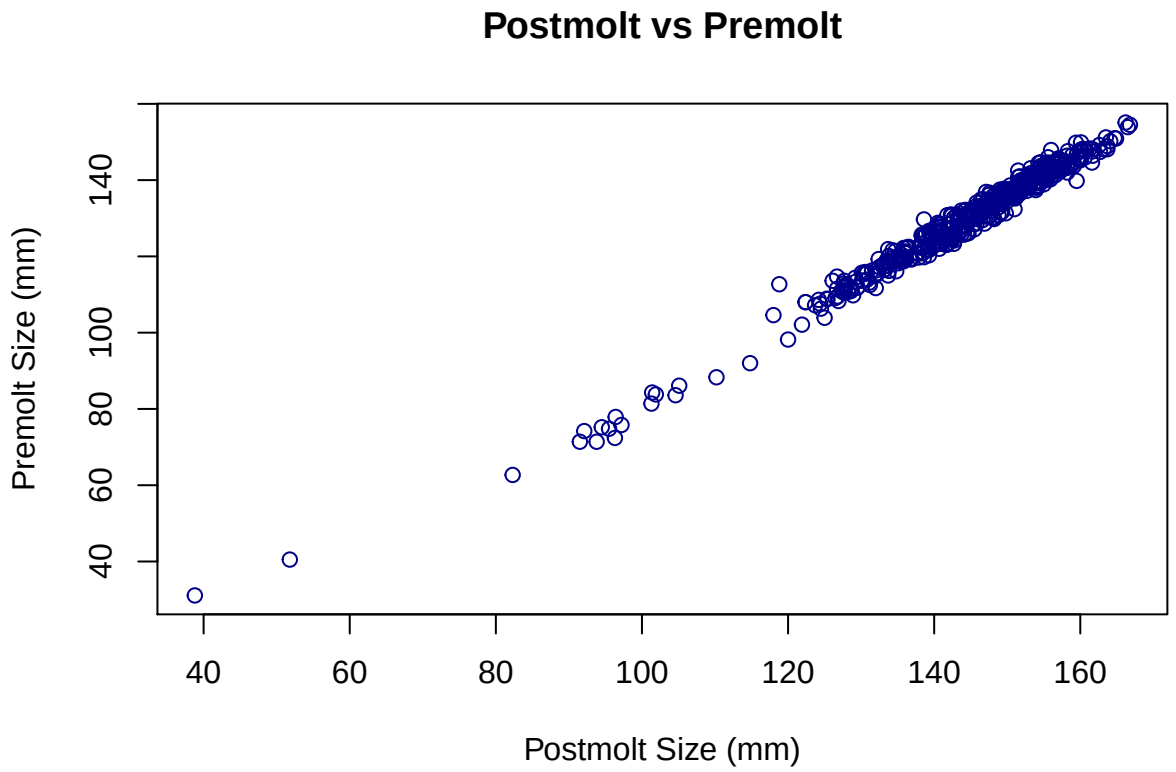
Leomart Crisostomo

4/22/2018

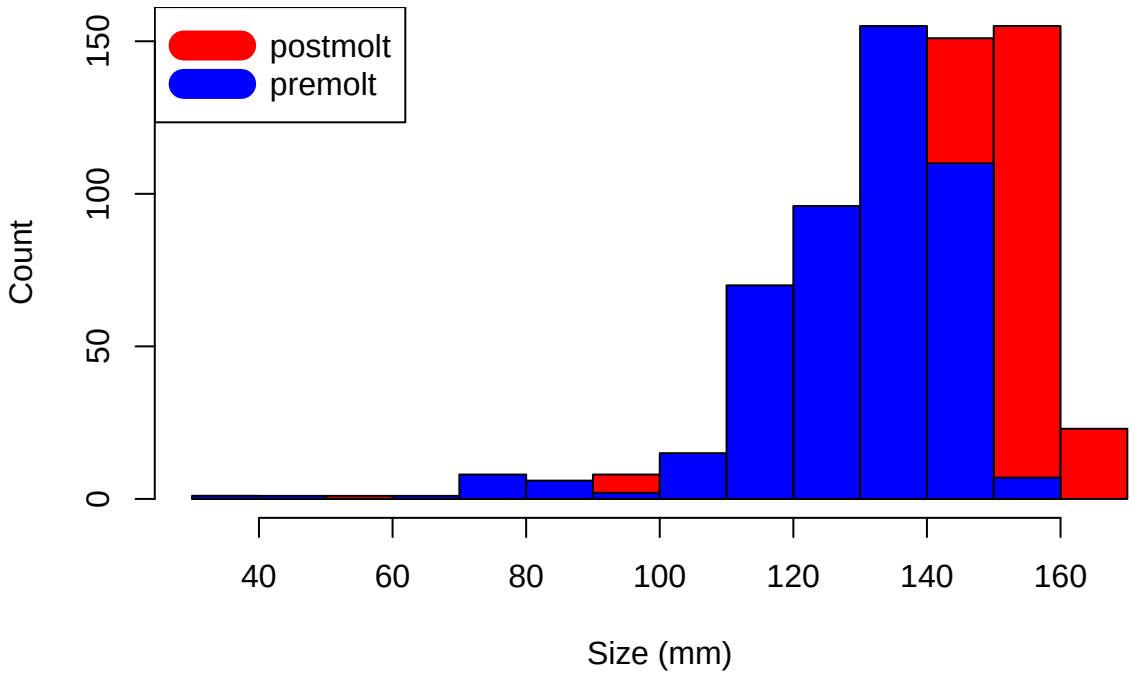
```
# 1
crabmolt <- read.csv("crabmolt.csv")

premolt <-  crabmolt$presz
postmolt <- crabmolt$postsz

plot(postmolt, premolt, ylab = "Premolt Size (mm)", xlab = "Postmolt Size (mm)",
     col = "dark blue", main = "Postmolt vs Premolt")
```




```
hist(postmolt, col = rgb(1,0,0), main = "" , xlab = "Size (mm)", ylab = "Count")
hist(premolt, add = TRUE, col = rgb(0,0,1))
legend("topleft", c("postmolt", "premolt"), col = c("red", "blue"), lwd = "15")
```




```
summary(premolt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    31.1   121.7   132.8   129.2   140.0   155.1
```

```
summary(postmolt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    38.8   138.0   147.4   143.9   153.4   166.8
```

```
pre_mean = mean(premolt)
post_mean =  mean(postmolt)
pre_sd = sd(premolt)
pre_sd
```

```
## [1] 15.86452
```

```
post_sd = sd(postmolt)
post_sd
```

```
## [1] 14.6406
```

```
size = length(postmolt)
```

```
# 2
# Y = premolt, X = postmolt
r = cor(postmolt, premolt)
r
```

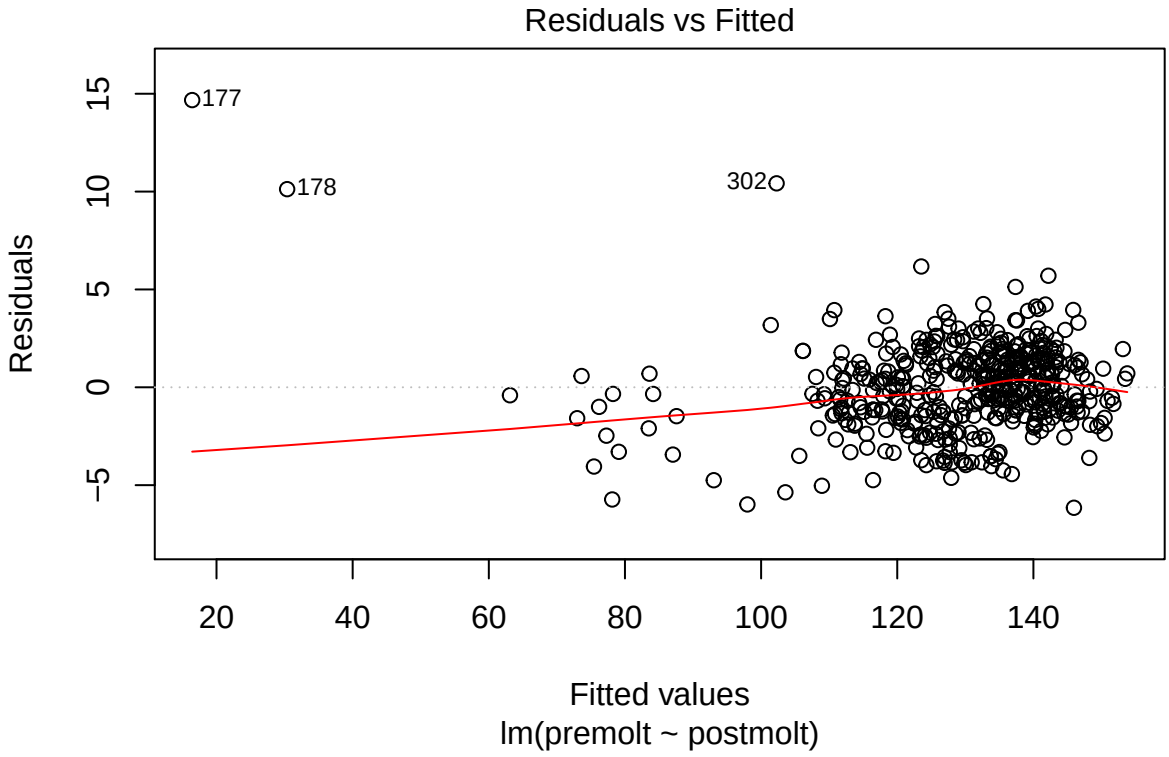
```
## [1] 0.9903699
```

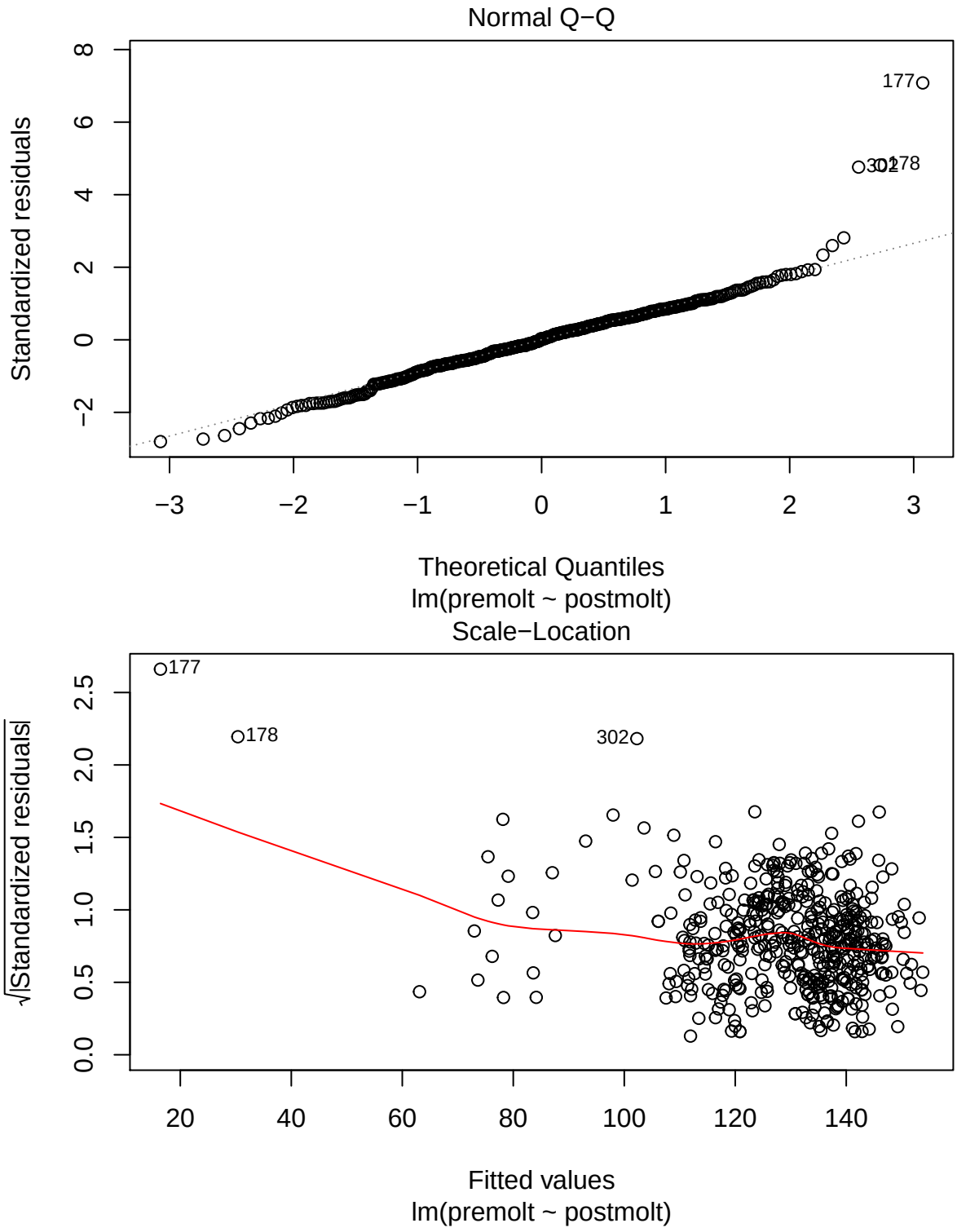
```
slope = r * pre_sd / post_sd
intercept = pre_mean - (slope * post_mean)
lm(premolt ~ postmolt)
```

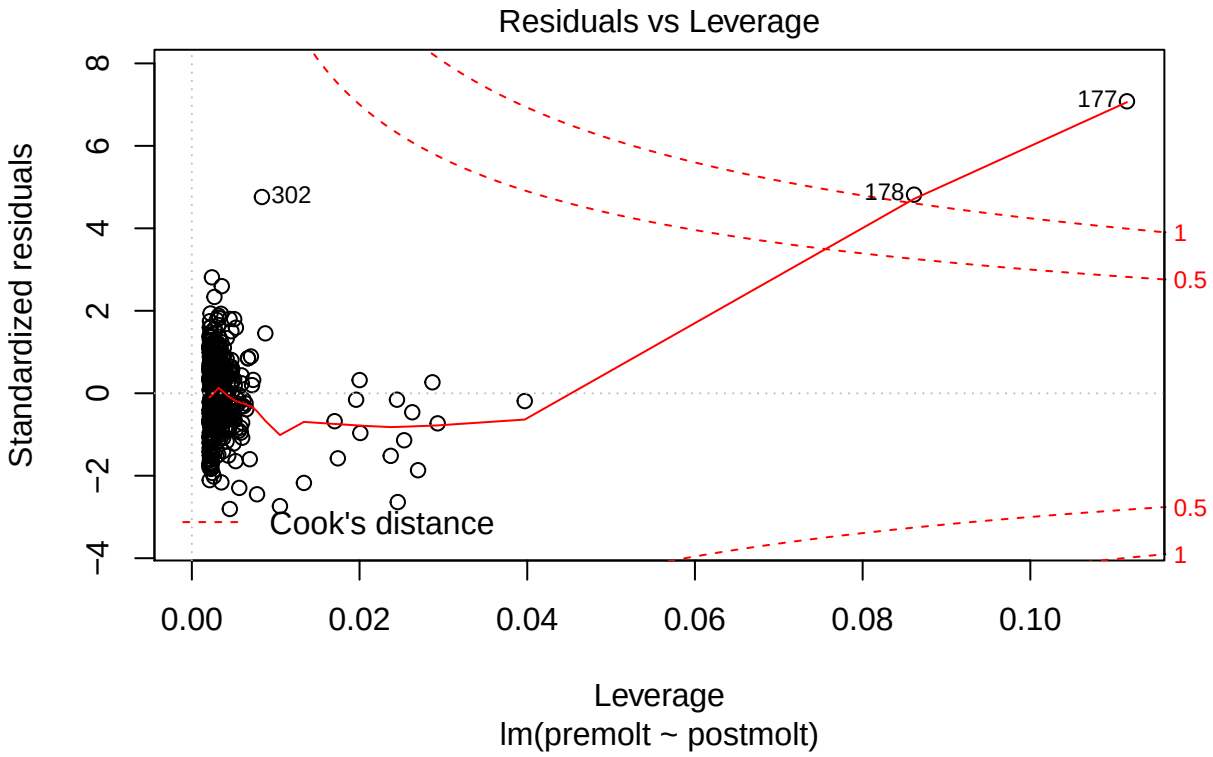
```
##
## Call:
```

```
## lm(formula = premolt ~ postmolt)
##
## Coefficients:
## (Intercept)     postmolt
##     -25.214        1.073
```

```
regr = lm(premolt ~ postmolt)
plot(regr)
```



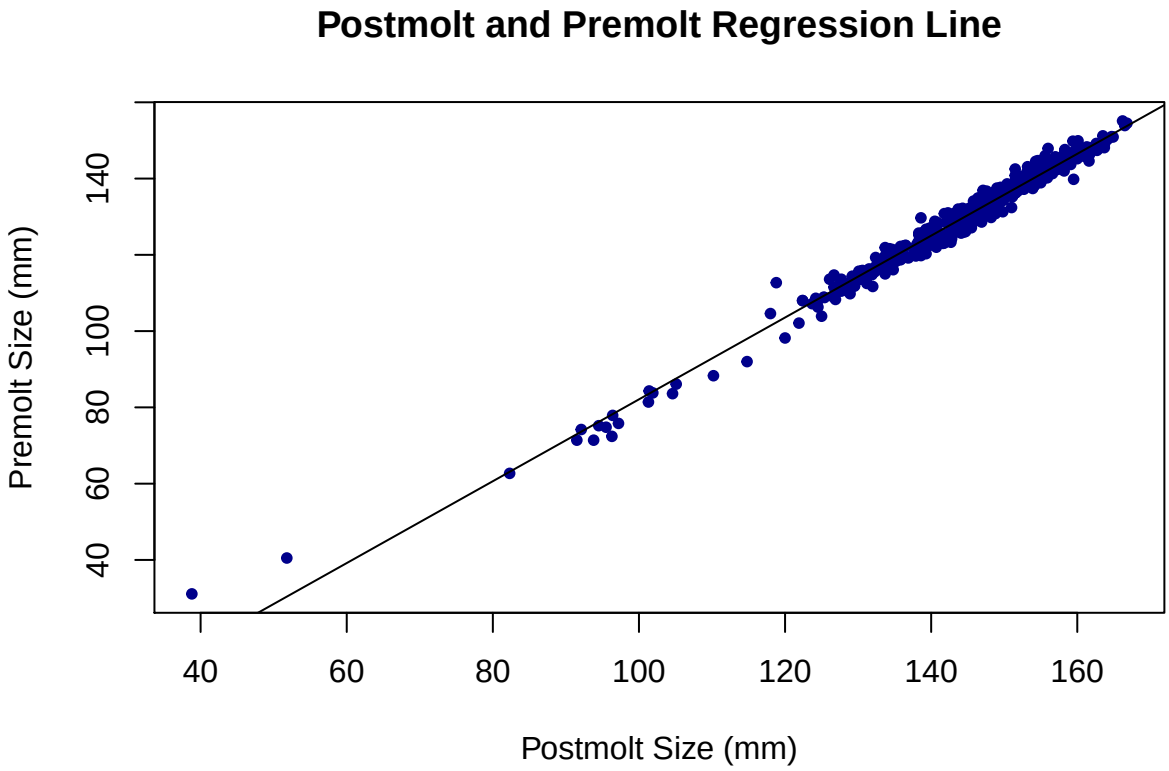

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(premolt ~ postmolt)
177
178
302
Scale−Location
√|Standardized residuals|
Fitted values
lm(premolt ~ postmolt)




```
predicted_premolt = postmolt * slope + intercept
# First 20 samples of the the predicted and observed premolt size comparison
table = data.frame(premolt[0:20], round(predicted_premolt,2)[0:20])
kable(table,  col.names = c("Observed Premolt Size", "Predicted Premolt Size"))
```

Observed Premolt Size	Predicted Premolt Size
113.6	111.83
118.1	117.73
119.9	119.99
126.2	128.57
126.7	124.28
127.3	125.24
128.2	129.11
129.5	129.97
130.5	133.19
131.6	130.18
131.7	132.33
132.2	131.36
132.4	136.83
132.8	134.37
132.8	133.40
133.6	133.94
133.8	134.90
134.1	134.69
134.3	134.90
134.6	134.37

```r
# 3
plot(postmolt, premolt, pch = 20,ylab = "Premolt Size (mm)", xlab = "Postmolt Size (mm)",
     col = "dark blue", main = "Postmolt and Premolt Regression Line")
abline(lm(premolt ~ postmolt))
```



3

As we can see on the plot, most of the data fit the regression line and only few of the data are far away from the line. It also follows a trend starting from the bottom left and increases as it goes to the right. This tells us that postmolt size and premolt size of the crab may be highly correlated. Thus, we invetigate further between these two variables.

```r
# 4
r_sq = r ** 2
r_sq
```

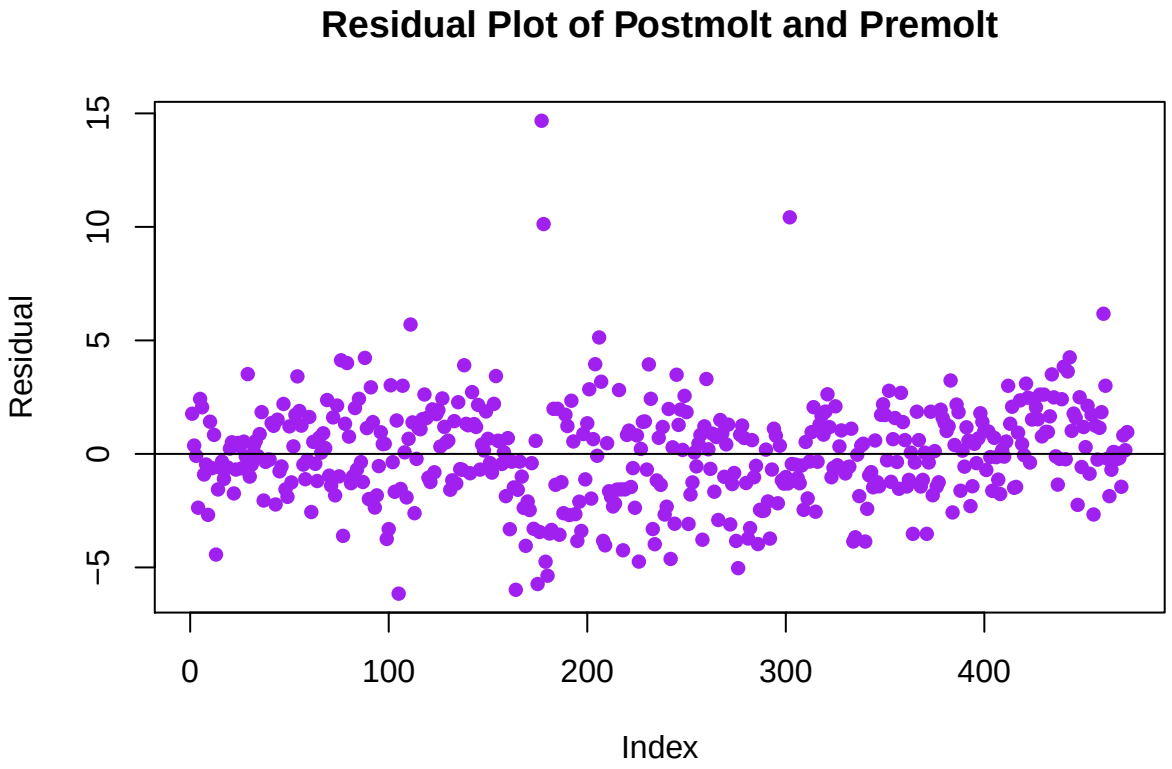
```
## [1] 0.9808326
```

4

For every 1 mm increase in postmolt size, there is a 1.073162 (the solpe of the regression line) mm increase in the premolt size. However, the y-intercept does not really make sense in this context since it turns out to be a negative value, and size can't be a negative value, and this is also because our data do not incluse where the postmolt is at 0.

The percentage variation in premolt size explained by the postmolt size is 98.08%, which is the coefficient of determination.

```r
#5
plot(resid(lm(premolt ~ postmolt)), ylab = "Residual",
     main = "Residual Plot of Postmolt and Premolt", pch = 16, col = "purple")
abline(0,0)
```



5

As we can see in the residual plot, the residuals are mostly centered and symmetric around the horizontal line that is equal to 0. There is no upward or downward trend and aside from the few outliers that are above 8, the residuals are fairly even cross the plot. This means that the accuracy of the regression line appears to be the same across the postmolt size predicting the premolt size. Therefore, the regression line obtained is a reasonable predictor of premolt size given a postmolt size

```r
# 6
# Hypothesis Testing
mean_slope = slope
est_variance_post = sum(resid(lm(premolt ~ postmolt)) ** 2) / (size - 2)
est_variance_post
```

```
## [1] 4.834374
```

```r
se_slope = ((size * (est_variance_post ** 2)) / ( (size * sum(postmolt **2)) - (sum(postmolt)) ** 2 )) =
se_slope
```

```
## [1] 0.01521496
```

```r
# Since 472 is a large sample, the t distribtion becomes a z distibution and the 95% Confidence Interva
lower_bound = mean_slope - (1.96 * se_slope)
upper_bound = mean_slope + (1.96 * se_slope)
c(lower_bound, upper_bound)
```

```
## [1] 1.043341 1.102984
```

```
test_stat = mean_slope / se_slope
test_stat
```

```
## [1] 70.53336
```

```
p_value = 2 * pnorm(-abs(test_stat))
p_value
```

```
## [1] 0
```

```
""
```

```
## [1] ""
```

6

Since the test_stat, 70.53, is not inside the 95% Confidence Interval (1.043, 1.10), we reject the null hypothesis that the slope is 0, and we have a strong evidence to conclude that the slope is non zero. Therefore, there is a significant linear relationship between postmolt and premolt size of the crabs. Also, the p-value turns out to be 0 which is a strong evidence against the null; therefore, the slope of the regression line is not 0.

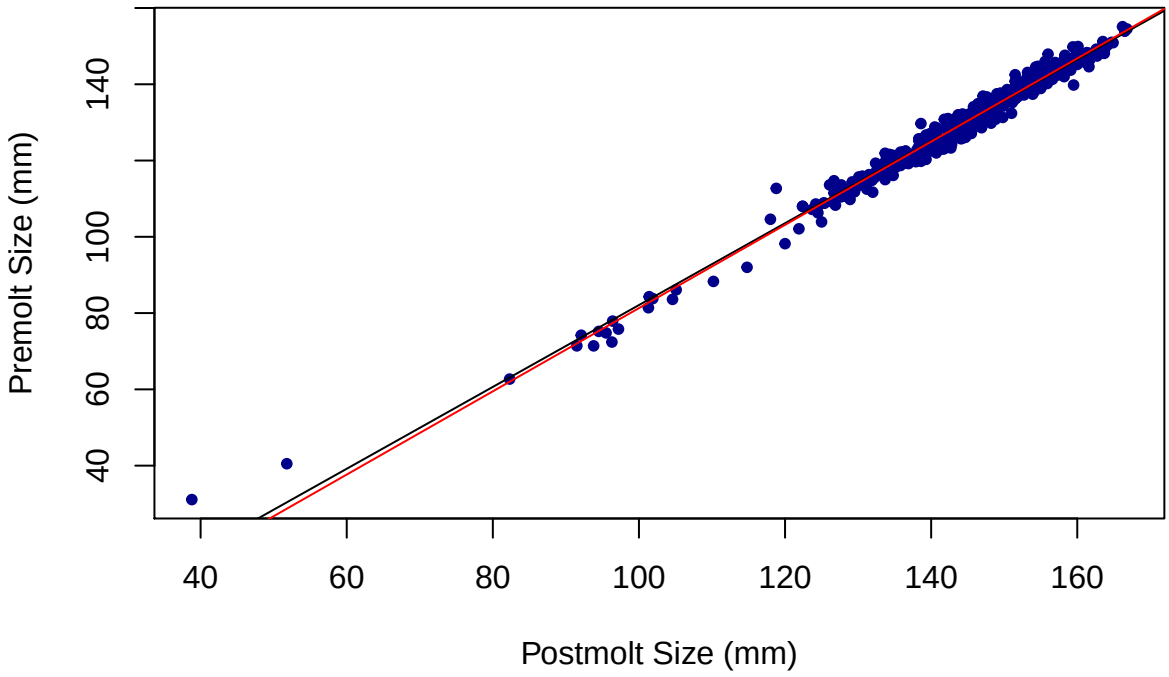
```
# 7
crabmolt1 <- crabmolt[crabmolt$presz >= 100,]
premolt1 = crabmolt1$presz
postmolt1 = crabmolt1$postsz
lm(premolt1 ~ postmolt1)
```

```
##
## Call:
## lm(formula = premolt1 ~ postmolt1)
##
## Coefficients:
## (Intercept)    postmolt1
##     -27.819        1.091
```

```
cor(postmolt1, premolt1)
```

```
## [1] 0.9840395
```

```
plot(postmolt, premolt, pch = 20,ylab = "Premolt Size (mm)",
     xlab = "Postmolt Size (mm)", col = "dark blue")
abline(lm(premolt ~ postmolt))
abline(lm(premolt1 ~ postmolt1), col = 'red')
```



7

Based on the plot, there is a small difference between the new regression line (red line) and the previous regression line (black line). The two lines are different when the size is less than 100 and almost the same line for sizes higher than 100, and it's because we only get rid of the sizes that are less than 100 mm. I expect them to be mostly the same since we are not getting rid of the outliers. If we get rid of the outliers, then the lines would be more different. The new regression line is slightly steeper than the previous line since its slope is higher and has a smaller y intercept, which would result to smaller correlation.

```
# 8
crabpop <-  read.csv("crabpop.csv")
molted <- crabpop[crabpop$shell == 1,]$size
not_molted <- crabpop[crabpop$shell == 0,]$size
predict_molted = molted * slope + intercept
predict_not_molted = not_molted * slope + intercept

summary(crabpop$size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    95.4   138.7   147.1   145.2   154.2   168.0
```

```
summary(predict_molted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   100.1   118.5   125.7   127.3   138.1   152.0
```

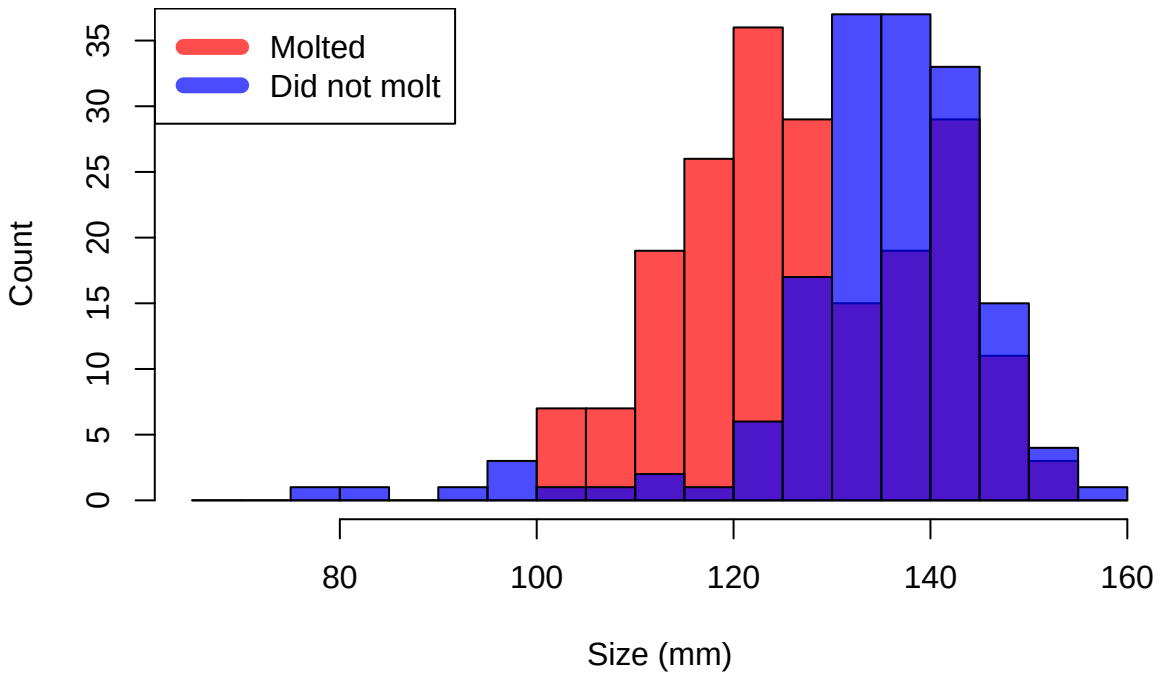
```
summary(predict_not_molted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   77.17  131.58  136.40  134.81  142.31  155.08
```

```
breaks = c()
for (i in c(1:20)){
```

```
  breaks = c(breaks, 60 + (5 * i))
}


hist(predict_molted, col = rgb(1,0,0, 0.7), main = "" , xlab = "Size (mm)",
     ylab = "Count", breaks = breaks )
hist(predict_not_molted, add = TRUE, col = rgb(0,0,1, 0.7), breaks = breaks)
legend("topleft", c("Molted", "Did not molt"), col = c(rgb(1,0,0, 0.7),
    rgb(0,0,1, 0.7)), lwd = "8")
```



The distribution of the molted crabs and non-molted crabs have similar distribution, but the distribution of the non-molted crabs is shifted more to the right, which suggests that they have larger premolt size than the premolt size of the crabs that molted.

Summary

The correlation coefficient between postmolt and premolt size of the crabs is .990, very close to 1, which means that postmolt and premolt size of crabs are highly correlated. We reinforce this idea by testing the hypothesis of slope being equal to 0, but we rejected this and concluded that the true slope is not 0. Therefore, there is a significant linear relationship between postmolt and premolt size of the crab. Since the sign of the correlation coefficient is positive, this implies that if the postmolt size increases, the premolt size of the crabe also has to increase. We used this relationship to predict the premolt size of the Female Dungesness crab after the 1983 molting season given only its postmolt size by using the regression line we obtained.