Lab 03 - Dungeness Crabs

Nicholas Lai

April 30, 2018

1. Data Exploration

All sizes are in millimeters.

crabmolt.csv

presz - Premolt size of the carapace before molting.

postsz - Postmolt size of the carapace after molting.

inc - Increment postmolt-premolt.

year - Collection year (not provided for recaptured crabs).

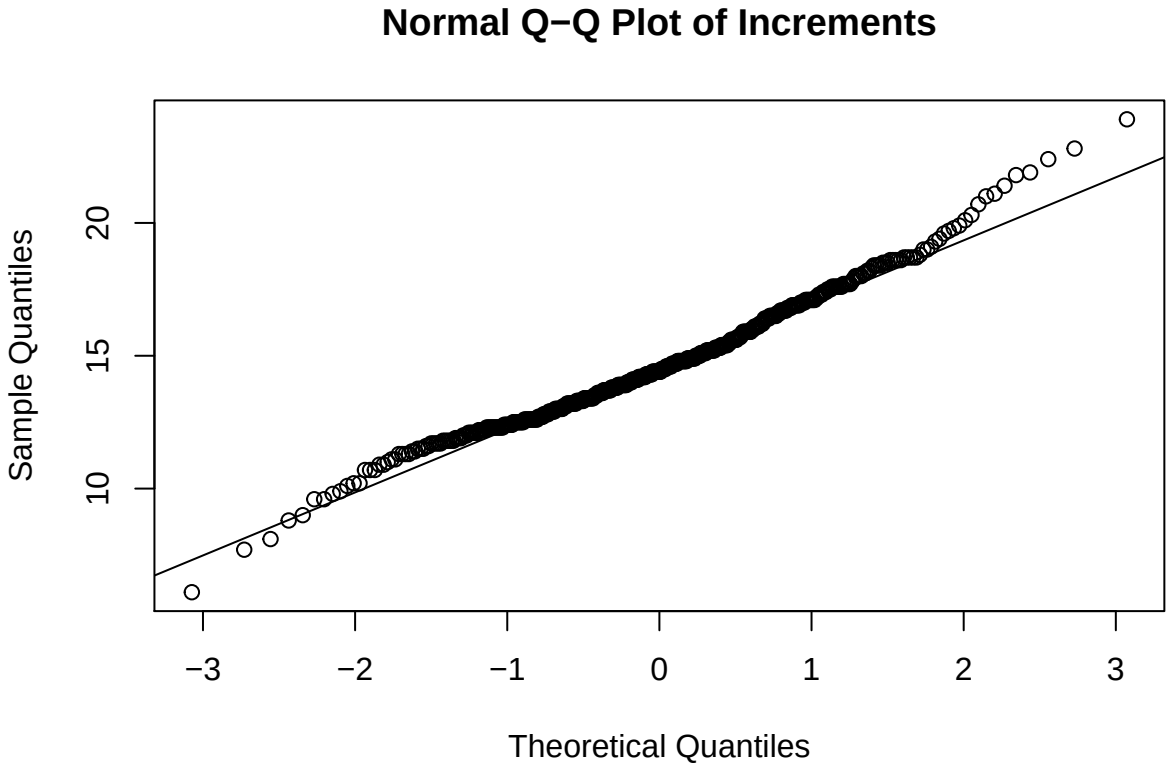
lf - Source. 1=molted in laboratory; 0=capture-recapture.

crabpop.csv

size - Postmolt size of the carapace after molting.

shell - Molt classification. 1=clean carapace; 0=fouled carapace.

a). An Interesting Plot - Quantile Plot of Increments



One might expect the difference of pre-molt and post-molt sizes of Dungeness crab to be normally distributed, as a natural assumption might be that the pre-molt and post-molt sizes of crab are normally distributed. But this quantile plot shows that the distribution is not a difference of normal populations, but rather more heavily tailed.

This may suggest that Dungeness crabs have a higher tendency to either grow much faster or slower after molting than would be expected normally.

b). Summary of one relevant variable - crabmolt$lf

```
## [1] 111
```

```
## [1] 361
```

The number of wild-caught crabs studied in this dataset number 111, while the number of lab-raised crabs number 361. This means that the data disproportionately represents laboratory-studied crabs by a margin of more than $3:1$. If any irregularities that could effect the measurements of the crabs arose during the lab gathering procedure, it could have had an effect on the data and its representativeness of wild crabs, the group of interest.

2. Building a Linear Model

We want to predict the premolt size of Dungeness crabs, given postmolt size. Let us consider the population model:

$y_i = \beta_0 + \beta_1 x_i + e_i$

where the x_is are the postmolt sizes of the Dungeness crabs, and the y_is are the premolt sizes of the Dungeness crabs, and e_i are *iid* standard normal error terms.

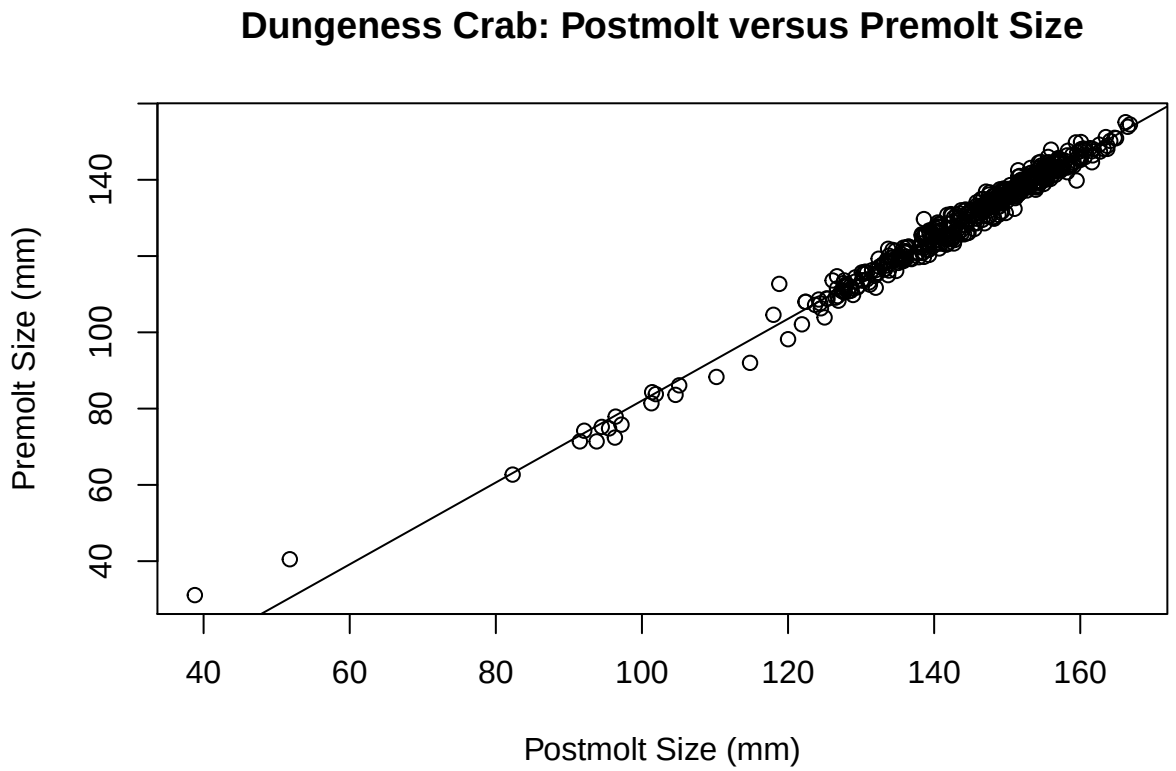
This gives rise to a linear regression model for prediction, with the idea of minimizing the squared error of the predicted variable, premolt size:

$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$

$\hat{\beta_0}$ can be interpreted as the y-intercept of the linear regression model, and $\hat{\beta_1}$ can be interpreted as the slope. We can calculate the estimators for these parameters from the data:

```
##
## Call:
## lm(formula = presz ~ postsz, data = crabmolt)
##
## Coefficients:
## (Intercept)       postsz
##     -25.214        1.073
```

3. The Linear Model, Visualized



4. Interpretation of the Linear Model

For every 1mm increase of postmolt size of a Dungeness crab, the linear model predicts a 1.073mm increase in the premolt size of the crab, minus a 25.214 constant negative adjustment.

The percentage of variation of premolt sizes of the Dungeness crab explained by the linear model applied to postmolt sizes is equal to the coefficient of determination, R^2. The value is calculated to equal:

$R^2 = \frac{s_y^2 - s_{\hat{e}}^2}{s_y^2}$

For simplicity, the expression can be rewritten as:

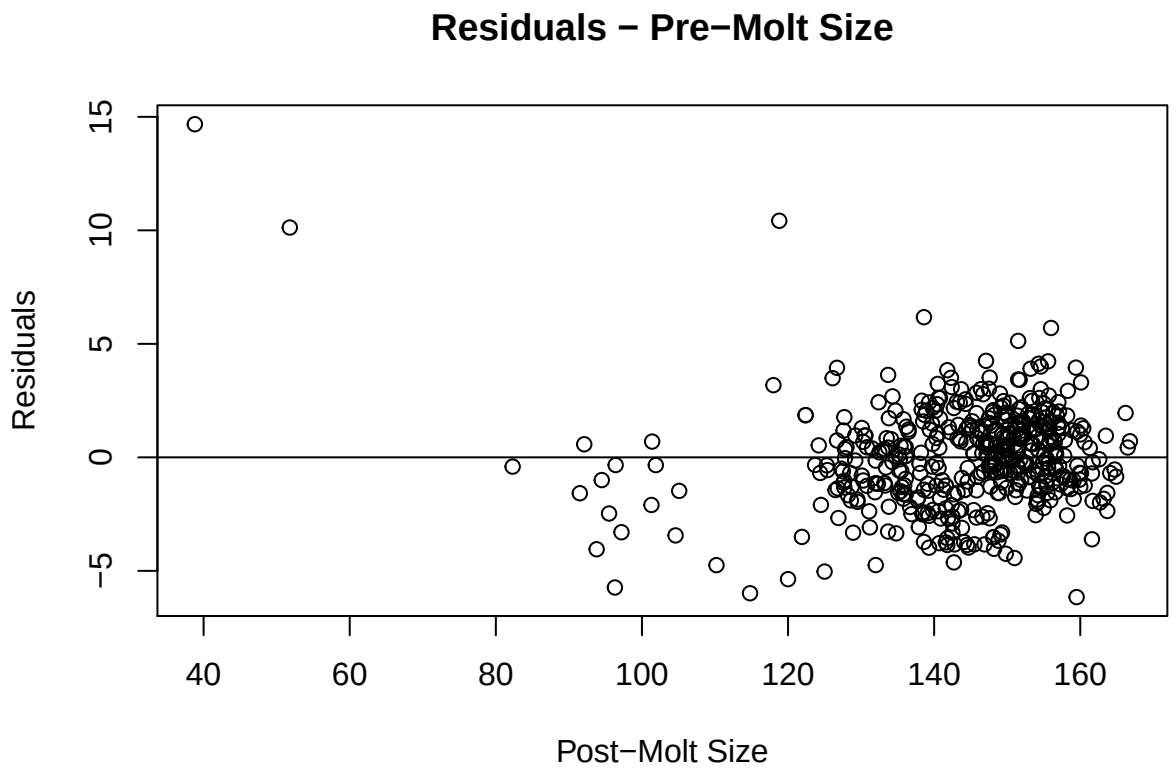
$R^2 = 1 - \frac{s_{\hat{e}}^2}{s_y^2}$

```
## [1] 0.9808303
```

The calculated R^2 value is about 0.9808, so about 98% of the variation in premolt sizes is explained by the model.

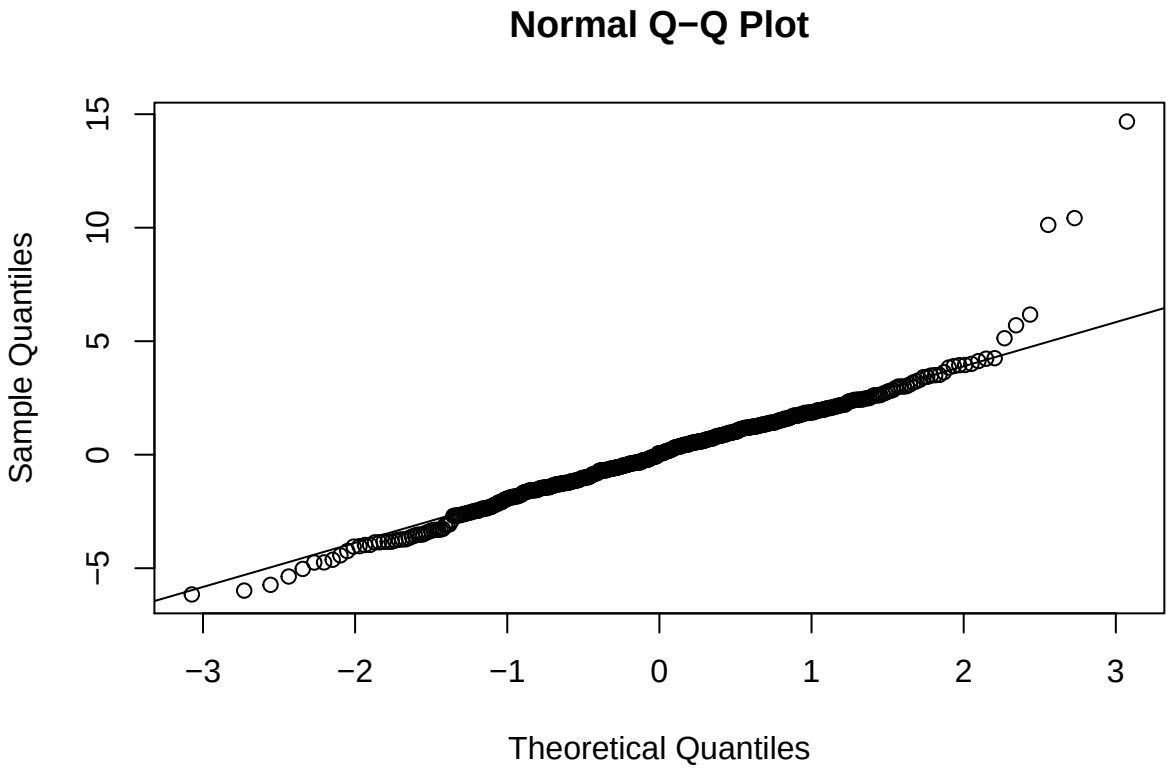
5. Residuals

One consequence of the model we defined in part 2 being true is the homoskedasticity of error. One way to verify this is to plot the residuals:



The residuals on the left of the graph seem to vary differently than the residuals on the right, which taken alone seem homoskedastic over their range. This suggest that there are some problems in our model that need to be addressed.

Another way to verify our assumption that error is standard-normally distributed is to view the normal quantile plot of the residuals.



It is immediately apparent that more values than expected have large residuals, which is evidence that our linear model is problematic without some further refinement.

6. Testing for Predictive Value

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$\alpha = 0.05$

Test Statistic: $\frac{\hat{\beta_1}}{s_{\hat{\beta_1}}}$

Given the null hypothesis, the test statistic is distributed as a t-statistic, with $df = n - 2 = 470$. As $df \to \infty$, $t_{df} \to$ Standard Normal. Because n is large, we can approximate the t-test with a z-test with the same test statistic.

The value of the test statistic is:

```
## [1] 155.0506
```

The critical values for the test are ± 1.96, which our test statistic is well above, so we reject the null hypothesis that $\beta_1 = 0$. Thus, we reject the implication of the null hypothesis, that the linear model we've created has no predictive value. Note that this conclusion would not change even if a one-tailed test were conducted.

7. Excluding Juveniles

To address the problems with our previous model, we can omit the juveniles by excluding data points with premolt sizes of less than 100mm.

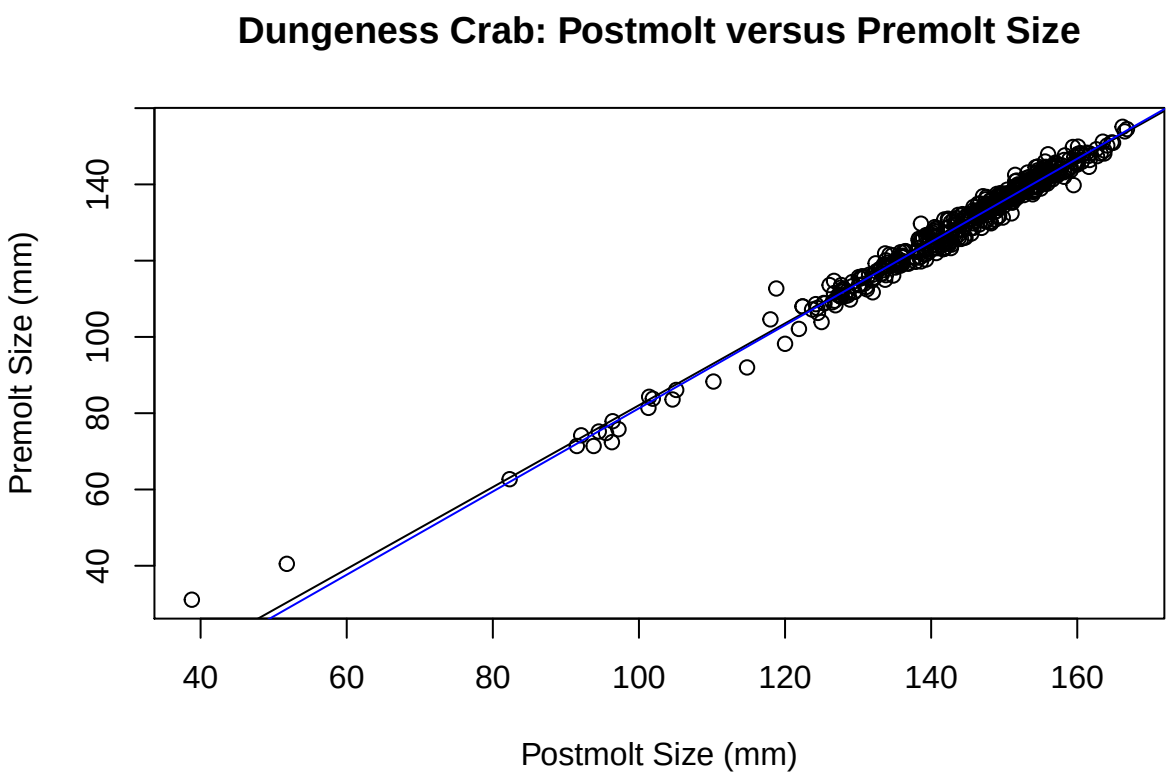
Our new linear regression constants are as follows:

```
##
## Call:
## lm(formula = presz ~ postsz, data = crabmolt_adj)
##
## Coefficients:
## (Intercept)       postsz
##     -27.819        1.091
```

Note that these coefficients are very similar to the ones found in Part 2,$\hat{\beta}_0 = -25.214$, $\hat{\beta}_1 = 1.073$.

The new model predicts that for every 1mm increase in the post-molt size of non-juvenile Dungeness crabs a 1.091mm increase in the pre-molt size of the crabs, minus a 27.819mm constant negative adjustment.

We can use these coefficients to plot a new regression line next to the old one:



We see that this line is very close to the regression line that we had before. This makes sense, as from an examination of the residual plot in Part 5, there were very few juveniles in the data to begin with as a proportion of total crabs.

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4362 -1.2997  0.0863  1.2920 10.8779
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.819060   1.359695  -20.46   <2e-16 ***
## postsz        1.091255   0.009292  117.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.977 on 451 degrees of freedom
## Multiple R-squared:  0.9683, Adjusted R-squared:  0.9683
## F-statistic: 1.379e+04 on 1 and 451 DF,  p-value: < 2.2e-16
```

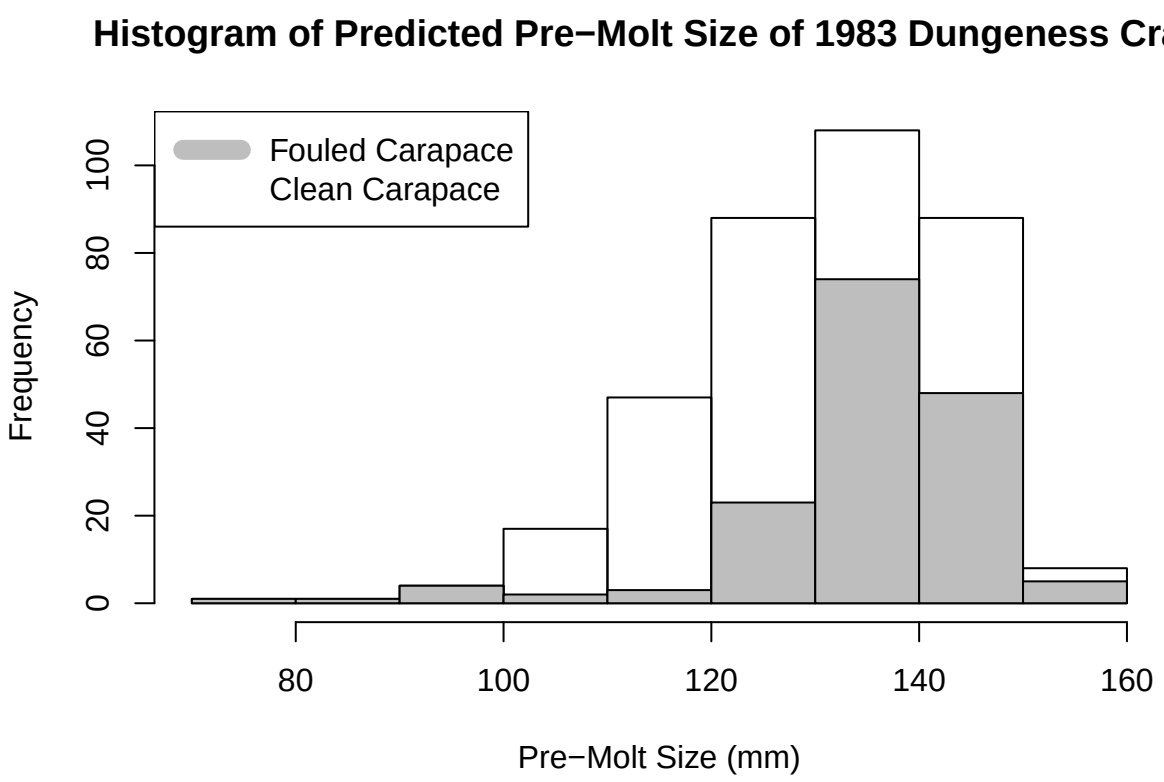
From the summary of the regression excluding the juveniles, we see that the R^2 value of the new model is 0.9683, no improvement on the old model. ~96% of the variation in y can be explained by the model on x.

8. The Model in Practice - 1983 Molting Season

We use the linear model developed in Part 2 to predict the pre-molt shell size of Dungeness crab prior to the 1983 molting season. Below is a numerical summary of the predicted pre-molt shell sizes:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   77.15  123.61  132.57  130.61  140.24  155.05
```

This is a histogram of the previous data, seperated by colour into molted and non-molted crabs. Crabs that have molted are observed to have clean, new shells, while the crabs that have not molted are observed to have dirty, fouled shells:


Histogram of Predicted Pre−Molt Size of 1983 Dungeness Crabs
Fouled Carapace
Clean Carapace
Frequency
0
20
40
60
80
100
80
100
120
140
160
Pre−Molt Size (mm)