Assignment 1

Kevin Dang

July 18, 2018

Important Note: Footers are used in this report in order to save space. Italicized text containing a footer notation indicates that there is a plot or code output that is to be referred to in the appendix.

Calibrating a Snow Gauge

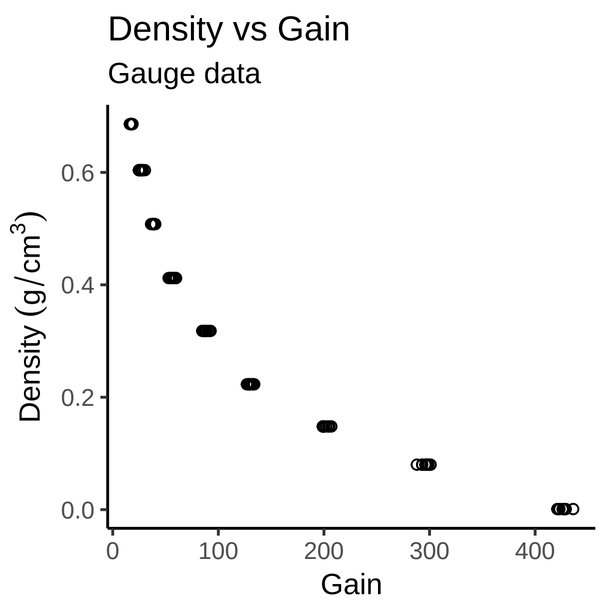
Snow gauges are used to indirectly measure the density of snow; a high snow density leads to less absorption of water. Analyzing this information is important because we want to monitor water levels and prevent floods from occurring. My analysis involves specifying the relationship between `density` of polyethylene blocks (a substitute for snow) and `gain` – an amplified version of gamma photon count. From the *Density vs Gain*[1] plot, it appears as though there is an inverse exponential relationship between the variables. A linear model was initially created, however the *standardized residuals*[2] appear to follow a distinct pattern, so a standard linear model cannot directly be fit to the data. A *box-cox transformation*[3] was done on the `gain` variable, and the plot shows that a value of $\lambda = 0.02020202$ is the best power transformation; in this case, a log transformation is appropriate. After completing a log transformation on the `gain` variable, a valid linear model for *Density vs log(Gain)*[4] was produced since the new *Residuals vs Fitted Values*[5] plot does not show a distinct pattern. Also, the *Normal QQ plot*[6] on the transformed data does not show evidence of skew – the normality condition is met. The *regression output*[7] shows a significant relationship between `log(Gain)` and `density`, as the p-value is extremely small. In addition, the multiple R-squared value of 0.9958 provides further evidence that this model is appropriate. The linear model is: mean `density` = 1.298013 g/cm^3 - (0.216203 g/cm^3 * `log(gain)`). This model can be used to estimate the mean `density` of snow at a particular value of `gain` since the snow gauge has been calibrated, but we must proceed with caution because polyethylene blocks were used in place of snow blocks for the model.

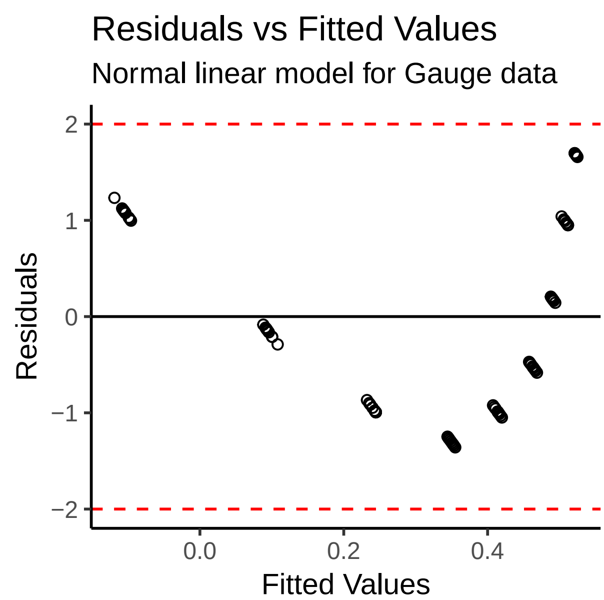
[1] Appendix A, Density vs Gain (Gauge data)
[2] Appendix A, Residuals vs Fitted Values (Normal linear model for Gauge data)
[3] Appendix A, Box-Cox Transformation
[4] Appendix A, Density vs log(Gain) (Transformed log model for Gauge data)
[5] Appendix A, Residuals vs Fitted Values (Transformed log model for Gauge data)
[6] Appendix A, Normal QQ plot (Transformed log model for Gauge data)
[7] Appendix A, Gauge Regression

Appendix A

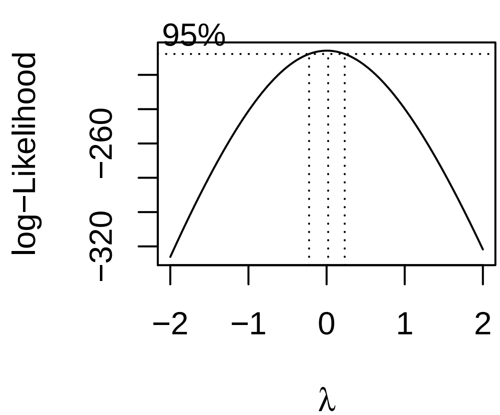
Snow Gauge Data

```
## Observations: 90
## Variables: 2
## $ density <dbl> 0.686, 0.686, 0.686, 0.686, 0.686, 0.686, 0.686, 0.686...
## $ gain    <dbl> 17.6, 17.3, 16.9, 16.2, 17.1, 18.5, 18.7, 17.4, 18.6, ...
```

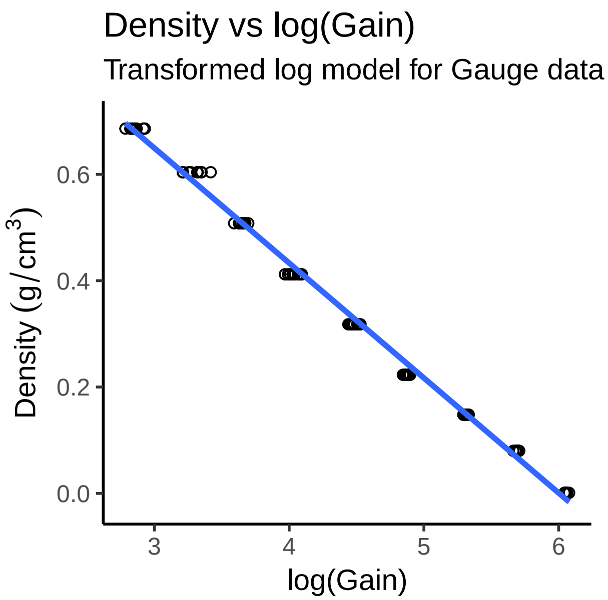


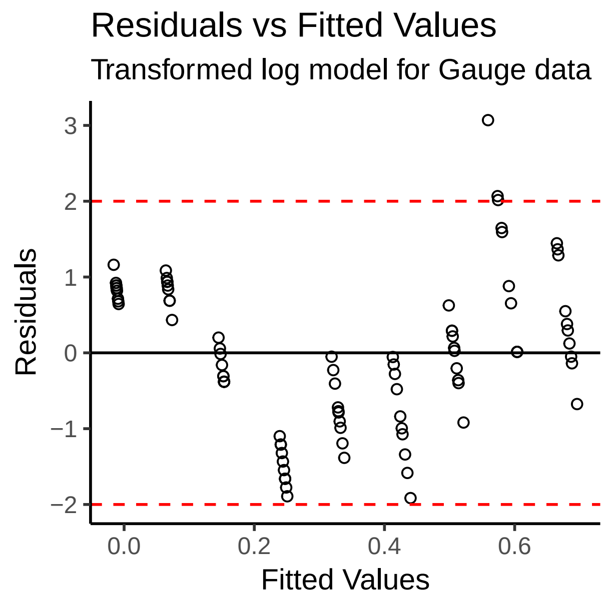


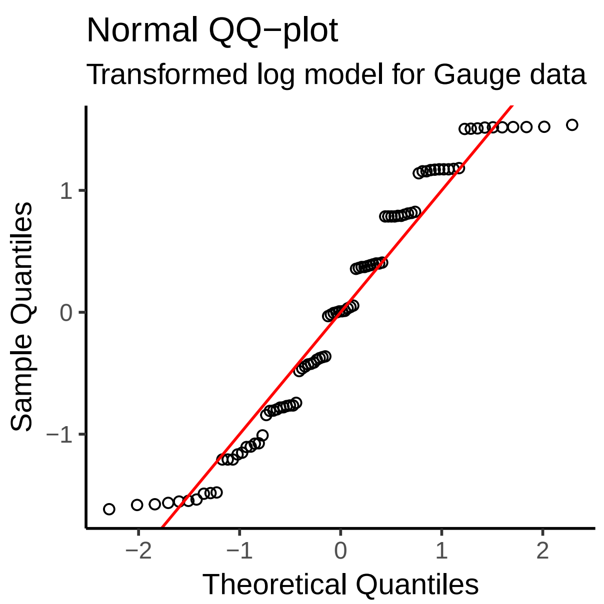
Box-Cox Transformation




```
## [1] 0.02020202
```


Density vs log(Gain)
Transformed log model for Gauge data
Density (g/cm^3)
0.0
0.2
0.4
0.6
3
4
5
6
log(Gain)


Residuals vs Fitted Values
Transformed log model for Gauge data
Residuals
−2
−1
0
1
2
3
0.0
0.2
0.4
0.6
Fitted Values


Normal QQ−plot
Transformed log model for Gauge data
Sample Quantiles
−1
0
1
−2
−1
0
1
2
Theoretical Quantiles

Gauge Regression

```
##
## Call:
## lm(formula = density ~ log_gain, data = gauge_transform)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028031 -0.011079 -0.000018  0.011595  0.044911
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.298013   0.006857   189.3   <2e-16 ***
## log_gain    -0.216203   0.001494  -144.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01471 on 88 degrees of freedom
## Multiple R-squared:  0.9958, Adjusted R-squared:  0.9958
## F-statistic: 2.096e+04 on 1 and 88 DF,  p-value: < 2.2e-16
```

Dungeness Crab Growth

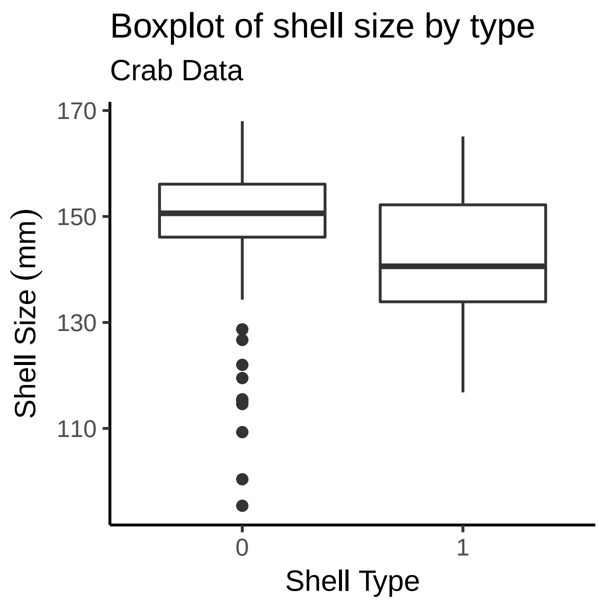
As Dungeness crabs grow, they need to replace their carapace; a process referred to as molting. My analysis involves grouping the adult female Dungeness crabs by whether they recently molted or not, estimating the mean carapace size of both groups, then determining whether there is a significant difference between the groups. First, a *boxplot*[8] of the shell size (`size`) by the shell type (`shell`) was created. Shell type 0 represents a fouled carapace which can be interpreted as an old shell, while shell type 1 represents a clean carapace – a recently molted shell. The boxplot shows that older shells contain some outliers, while recent shells have no outliers. From the *summary statistics*[9], type 0 shells are larger than type 1 shells by 7mm on average. A *two sample t-test*[10] yielded a statistically significant p-value; this indicates that the means of the 2 groups are not equivalent. The two sample groups are independent,since the traps that were used were designed to catch adult female Dungeness crabs of all sizes, meaning that this sample is representative of the population. An *F-test to compare two variances*[11] shows that the two sample group variances are similar – the constant variance condition is met. Both the *Normal QQ Plot*[12] and *Histogram of Shell Size*[13] show skew in the data, which may be a problem. Fortunately, the sample size of 362 (161 type 0, 201 type 1) is sufficiently large. By the Central Limit Theorem, means of samples from a population approach a normal distribution as sample size increases – regardless of the population distribution. Thus, the normality condition for the t-test is met. Given the strong supporting evidence, adult female Dungeness crabs with older carapaces (shell type 0) on average have larger shells than those with recently molted carapaces (shell type 1).

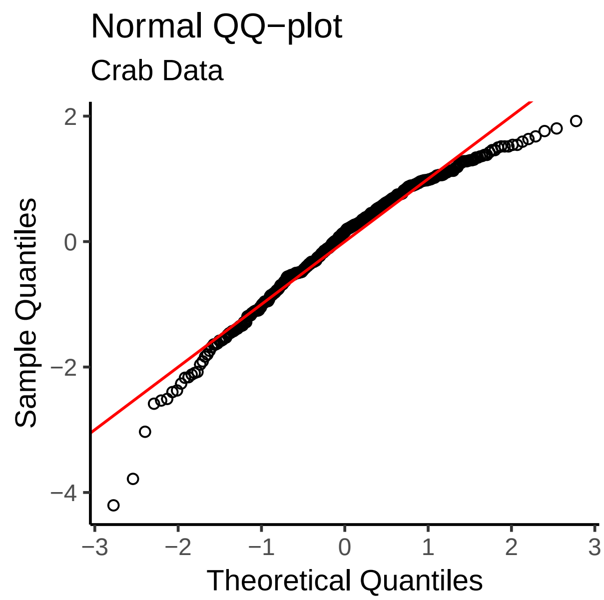
[8] Appendix B, Boxplot of shell size by type (Crab data)
[9] Appendix B, Summary Statistics
[10] Appendix B, Two Sample t-test
[11] Appendix B, F test to compare two variances
[12] Appendix B, Normal QQ plot (Crab data)
[13] Appendix B, Histogram of Shell Size (Crab data)

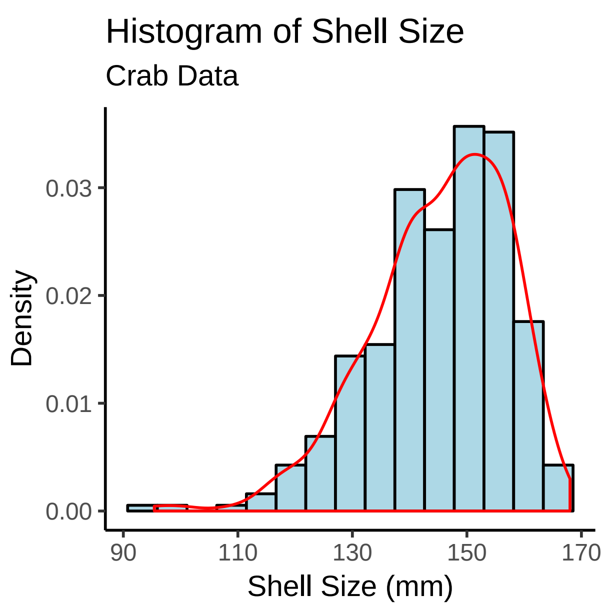
Appendix B

Crab Growth Data

```
## Observations: 362
## Variables: 2
## $ size  <dbl> 116.8, 117.1, 118.4, 119.6, 120.1, 120.4, 120.6, 122.6, ...
## $ shell <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "...
```







Summary Statistics

```
## # A tibble: 2 x 5
##   shell group_mean group_median group_sd group_size
##   <chr>      <dbl>        <dbl>    <dbl>      <int>
## 1 0           149.         151.     11.3        161
## 2 1           142.         141.     11.4        201
```

Two Sample t-test

```
##
##  Two Sample t-test
##
## data:  size by shell
## t = 5.8328, df = 360, p-value = 1.215e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.637563 9.355447
## sample estimates:
## mean in group 0 mean in group 1
##        149.1099        142.1134
```

F test to compare two variances

```
##
##  F test to compare two variances
##
## data:  size by shell
## F = 0.97771, num df = 160, denom df = 200, p-value = 0.8851
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.729754 1.316331
## sample estimates:
## ratio of variances
##          0.9777051
```