STA 303 Assignment 1

Guanchen Zhang

July 14, 2018

Gauge Calibration

To monitor the water supply,people wants to analysize the snow-pack profile because show absorb the water up to a certain point, after which the water floods away. The snow gauge is used to determine a depth profile of snow density but does not directly measure snow density. The experimental data were acquired by varying density and measuring the gain. In the experiment, the polyethylene blocks are used to simulate snow whose density was recorded in dataset. The gauge measurement is called "gain".

In the context of the problem, a statistical model is necessay and ther is no alternative way. In general, linear regression has following assumptions: Linear relationship, Multivariate normality, No multicollinearity (no need in this case), No auto-correlation (no need in this case), Homoscedasticity. A simple linear regression model called fit1 was built to estimate mean density at a given gain. From the diagnostic plots of the fit1 model, there is a distinctive curvilinear pattern in the residual vs. fitted plot (figure 1.2). This could mean that we may get a better model is we try a model with a quadratic term included. The assumption of equal variance does not hold under the model of fit1.

When the variance is found to be nonconstant, we can consider to use transformation to stabilize variance. Also because the relationship between the response and the design variable is not linear, it is possible that a transormataion can put the relationship into a linear form. According to the boxcox function in R, a log transformation is recommended since the lamda is 0 which is acceptable for the log transformation. The resulting scatter plot has points forming a nearly perfect straightline, indicating an almost linear relationship. All assumptions for linear regression hold under the model of fit2.

Since the scatter plot is curved, a polynomial model may be appropriate. However, the physical model for the relationship between gain and density suggests proceeding with the log transformation.

Then we can conclude that to estimate mean density of snow pack at given gain we can use the model depicted as fit2 by taking log to "gain".

Appendix

We first need to coerce other objects to a data frame and glimpse at the data frame (table 1.1). There are 90 observations and 2 variables which are density and gain.

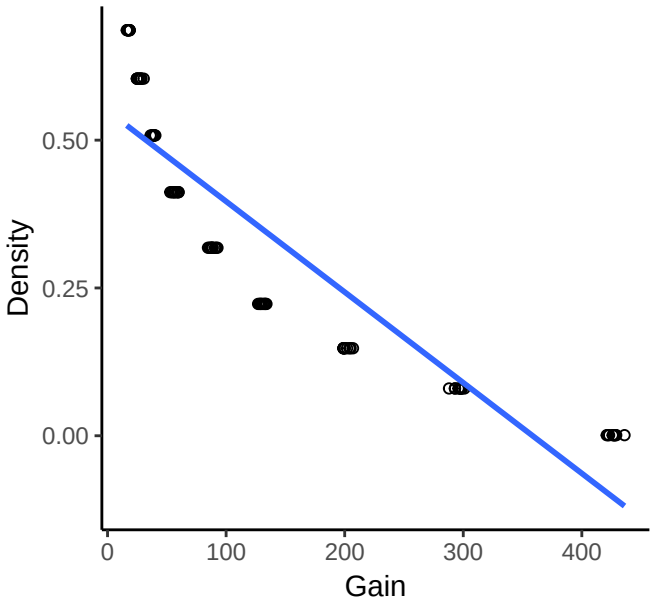
Then, a scatter plot of density by gain was plotted (figure 1.1), which indicates there is a curvilinear relationship instead of linear relationship where the density decrease exponentially when the gain increase.

Table 1.1 Glimpse

```
## 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, ...
```

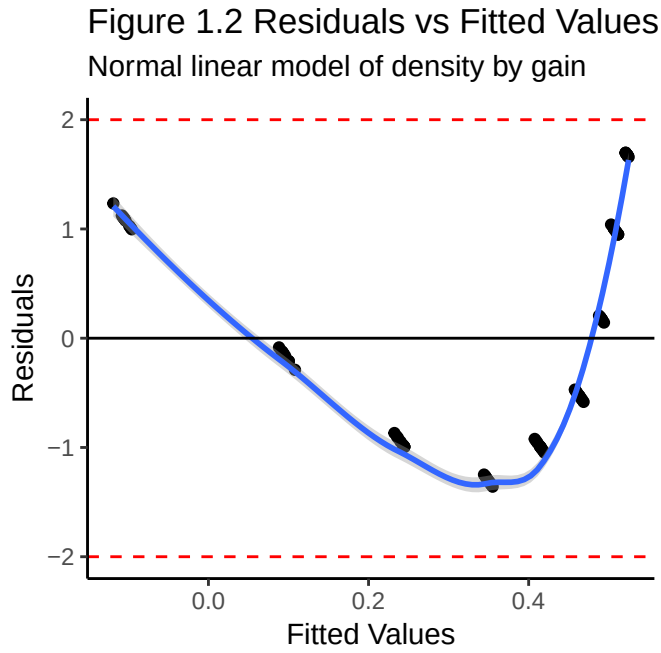
Figure 1.1 Linear Regression Model

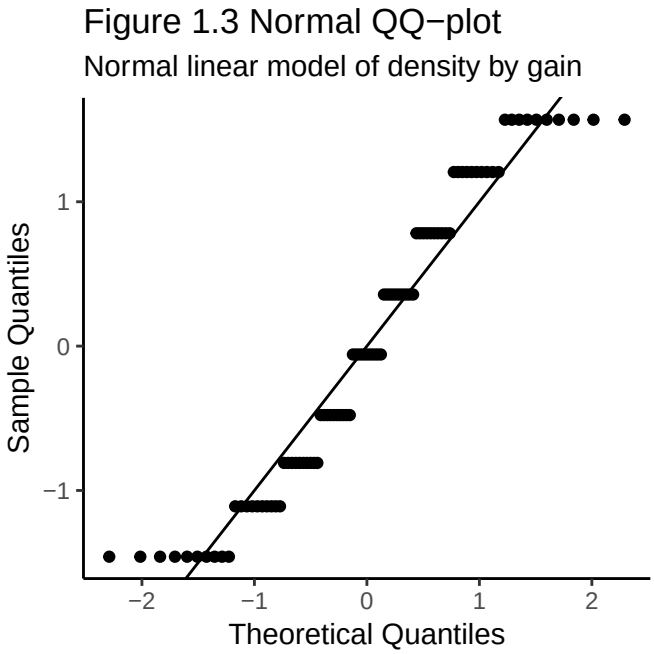


The assumption of normal errors is needed and can be checked by QQ plot (figure 1.3). In the normal QQ plot (figure 1.3) produces points close to a straight line so the data are said to be consistent with that from a normal distribution. Another assumption is that the errors have constant variance. In the residual vs. fitted plot (figure 1.2), the residuals are not spread equally around 0, indicating the assumption of equal variance is violated.

Table 1.2 ANOVA of fit1

```
## Analysis of Variance Table
##
## Response: density
##           Df Sum Sq Mean Sq F value    Pr(>F)
## gain       1 3.7169  3.7169  389.48 < 2.2e-16 ***
## Residuals 88 0.8398  0.0095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

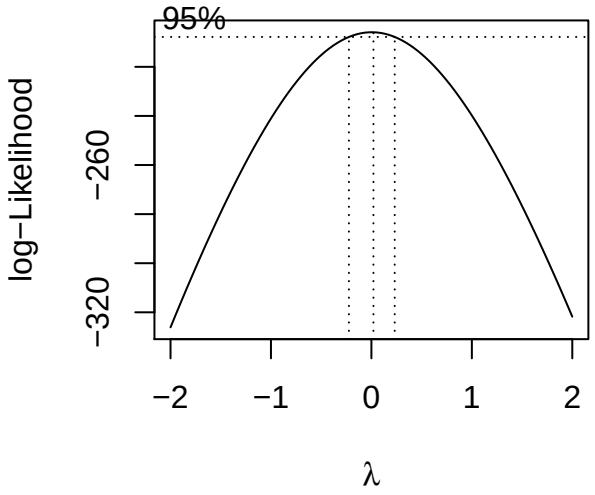




In the residual vs. fitted plot (figure 1.6), the residuals are spread equally around 0, indicating the assumption of equal variance is hold and the curvilinear pattern is not such obvious as the original model. Meanwhile, the assumption of normal errors is improved in the QQ plot (figure 1.5). Moreover, because the smaller the residual sum of squares, the better your model fits your data, comparing ANOVA of fit1 (table 1.2) with that of fit2 (table 1.3), the residual sum of squres of the fit2 is much smaller than that of fit1 and more closer to 0.

Table 1.3 ANOVA of fit2

```
## Analysis of Variance Table
##
## Response: density
##              Df Sum Sq Mean Sq F value    Pr(>F)
## I(log(gain))  1 4.5376  4.5376   20956 < 2.2e-16 ***
## Residuals    88 0.0191  0.0002
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




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

Figure 1.4 Log Regression Model

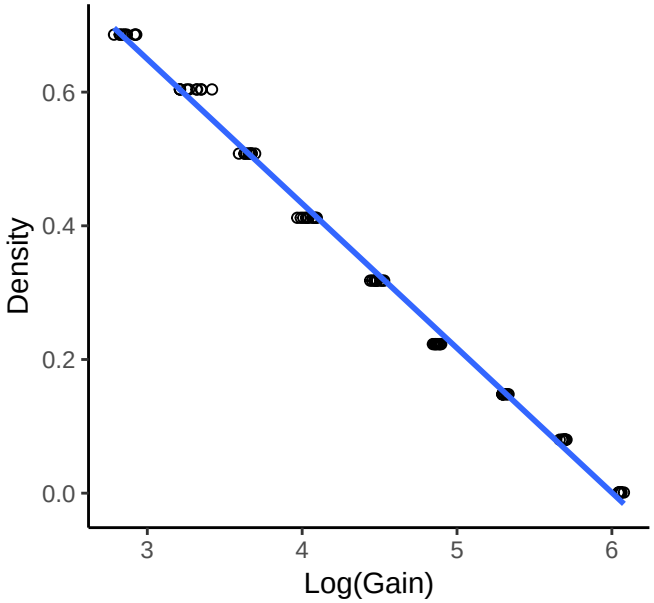
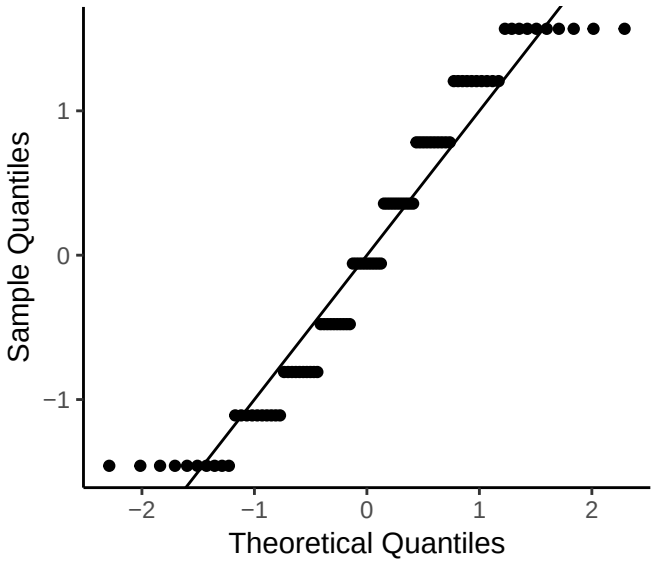
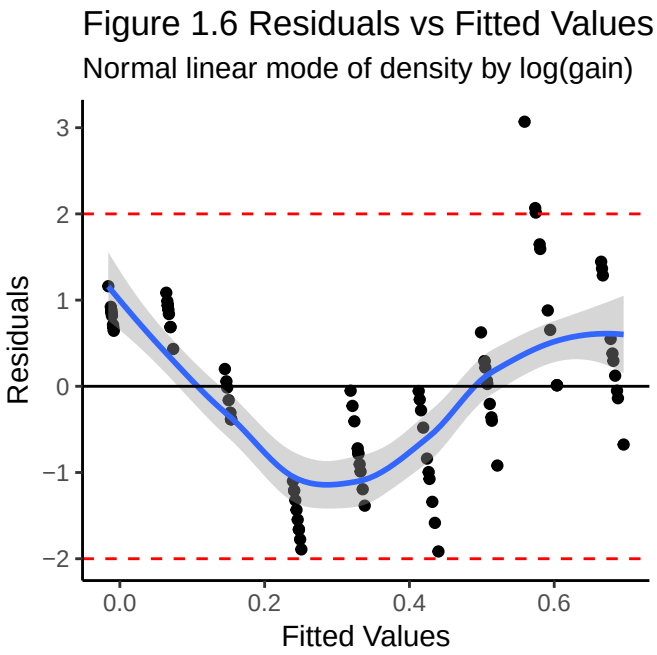


Figure 1.5 Normal QQ−plot

Normal linear model of density by log(gain)





Dugeness crabs

Dugeness crabs are fished extensively and fisihing female crabs is a possible means of controlling hte fluctuation in yearly catches of crabs. To determine size restrictions for female crabs, we are trying to estimate mean carapace size for crabs who have recently molted, for those who have not, and whether the difference is significant. The shell width was recorded as "size" in the dataset and information on whether the crab had molted in the most recent molting season or not was recorded as dummy variable called "shell". According to the background reading, "shell" is 0 means the crab has a fouled shell indicating it had molted in the most recent molting season, whereas, 1 means the crab has a clean shell, indicating it had not molted in the most recent molting season.

A box plot (Figure 2.1) is drawn as well which enable us to study the distributional characteristics of the group of the variable and provide some indication of the data's symmetry and skewness. Unlike other methods of data display, boxplots show outliers. By using a boxplot for each categorical variable side-by-side, it is easy to compare data set.

Regarding the problem of the crab growth, ANOVA is used to assess the relationship between a continuous and categorical variable, unlike the linear regression model used to solve the problem of gauge calibration where both the dependent variable and the independent variable are continuous.

Then we can conclude that estimated mean carapace size for crabs who have recently molted is 142cm, for those who have not is 149cm, and there is significant evidence that whether the female crab had molted in the most recent molting season or not impact on the width of the female crab.

Appendix

In general, the shape of the box are different between fouled shell and clean shell crab. Comparing the box for clean shell, the box for fouled shell is much shorter, suggesting that

size of fouled shell carbs have a high level of agreement with each other, meanwhile, it is much higher, suggesting a difference between groups. Regarding the boxplot for clean shell carbs, the box are uneven in size, showing that many carbs have similar sizes at certain parts of the scale, but in other parts of the parts of scale sizes of clean carbs vary much. Outliers are very noticable for the carbs with fouled shell. ### Table 2.1 Glimpse

```
## 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", "...
```

Figure 2.1 Boxplot of size by sh

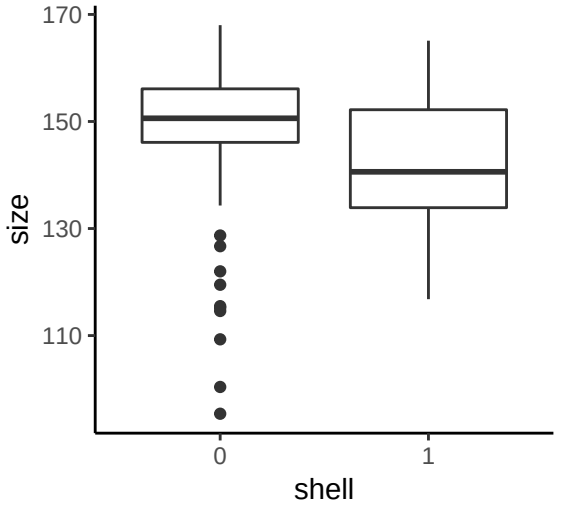


Table 2.2 Grand statistics

```
## # A tibble: 1 x 2
##   grand_mean grand_sd
##        <dbl>    <dbl>
## 1       145.     11.8
```

Table 2.3 Group statistics

```
##
## Call:
## lm(formula = size ~ shell, data = crabs_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.710  -5.986  -0.260   7.789  22.987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 149.1099     0.8938 166.823  < 2e-16 ***
## shell1       -6.9965     1.1995  -5.833 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 11.34 on 360 degrees of freedom
## Multiple R-squared:  0.08634,    Adjusted R-squared:  0.08381
## F-statistic: 34.02 on 1 and 360 DF,  p-value: 1.215e-08
```

```
## # A tibble: 2 x 5
##   shell count median  mean    sd
##   <chr> <int>  <dbl> <dbl> <dbl>
## 1 0       161   151.  149.  11.3
## 2 1       201   141.  142.  11.4
```

In the ANOVA table(table 2.4), as the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the group of fouled shell carbs and the group of clean shell carbs. Since there are only 2 groups, it is not necessary to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of group are statistically significant.

The ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. From the output (table 2.5) we can see that the p-value is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, the assumption of homogeneity of variances in the different treatment groups holds. Anova assumes that the data in each group are distributed normally. This assumption is equivalent saying that the residuals of the best-fitting model are distributed normally. In the normal QQ plot (Figure 2.2), as almost all the points fall approximately along this reference line, the assumption holds.

Table 2.4 ANOVA table

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## shell         1   4376    4376   34.02 1.21e-08 ***
## Residuals   360  46305     129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

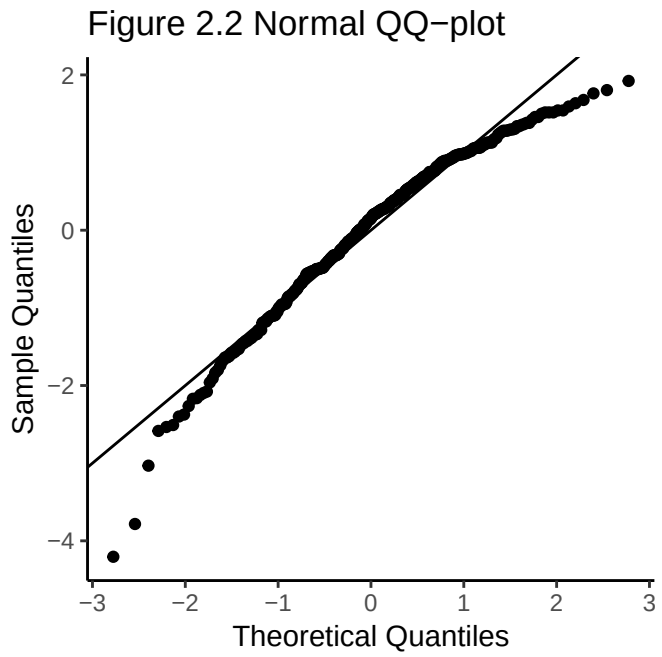


Table 2.5 Equal variance test

```
##
##  Bartlett test of homogeneity of variances
##
## data:  size by shell
## Bartlett's K-squared = 0.022512, df = 1, p-value = 0.8807
```