STA1002 Assignment 1__Lu Gan 1003119608

Dataset 1: Snow Gauge

In this dataset, gain is the predictor while density is the response variable. We need to build a proper model to describe their relationship. And use diagnostic plots and anova table to analyze the fitting result. Firstly, we will load and view data. Based on the observation count table, we have 9 levels of density totally and each level has ten observations. Then based on the figure of relationship between density and gain, we can describe that the snow density will decrease as the gain value becomes larger. However, it seems not to match a linear relation.

For the first simple linear model for gain and density, the fit result is not good in the first model plot. It has deviation with the data point trend without random residual pattern as well which indicates a nonlinear relation. Next, we transform gain to log(gain) and build the second model. Its line fits the data points well and the residual plot looks better.

Check three assumptions for valid model. Use the coincident result in QQ-plot to confirm normality of residuals. Then apply Bartlett's test to check homogeneity but it seems variances are different which leads a greater probability of falsely rejecting the null hypothesis and indicates difference between levels. Last, observations are independent. Fit data to null model and compare two models with anova. The result indicates lm2 is better with less sum of squares and significant p value of F test. We use two gain values in the context to predict the density with confidence interval.

In conclusion, the variables have negative relation. With the model between log(gain) and density, we can use a gain reading to predict density with a confidence interval. Note that there exists difference between levels.

Dataset 2: Crab shell size

This dataset includes information of 362 crabs with two variables. Size represents the size of the carapace and the other categorical variable shell shows molt classification. 1 stands for clean while 0 for fouled. They correspond to postmolt and premolt respectively. Our goal is to analyze whether differences about shell size exist between groups.

For accessing the relationship between a continuous and categorical variable, we can use boxplot that size against shell and statistics. We can observe that the mean size of premolt is larger than postmolt. But there are some outliers in premolt data leading larger deviation. It is reasonable based on the context since crabs will form new smaller shell and drop old in molting process.

The significant p-value in F test of anova table proves that there are differences between two molting groups. The assumptions of anova are normally distributed variables and equal variance. We use QQ-plot and histogram to check. From initial result, the distribution is right skewed. Then transform to use square according to boxcox curve. The new QQ-plot and distribution are better which can be used to make anova.

Then make analysis of variance for full linear model and reduced null model. In conclusion, the two variables have some relation. The mean size of shell in premolt is larger than postmolt. We can predict shell size based on molting state. And there exists significant difference between two groups.

STA1002 Assignment1

Lu Gan 1003119608

1. Snow-gauge dataset

a. View data and plot variables

```
## Observations: 93
## 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, ...
```

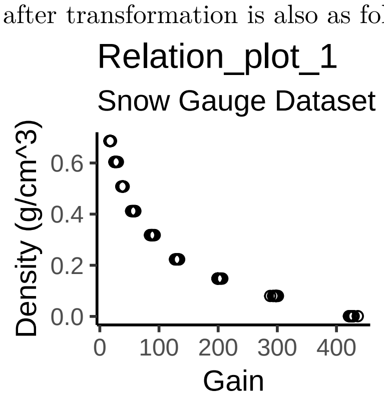
From above, we can observe the dataset includes two variables gain and density. Our goal is to estimate mean density based on a given gain. So gain is our predictor while density is the response variable. Next, we can summarize their statistics.

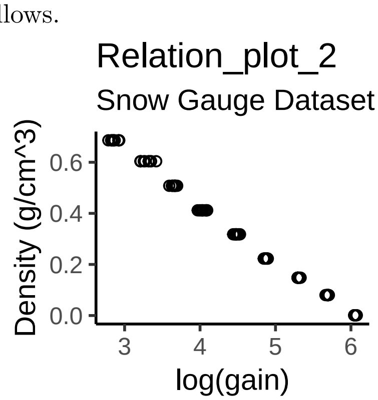
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0010  0.1480  0.3180  0.3311  0.5080  0.6860       3
```

From the summary result, there are three empty rows without both of the two variables. Since the number of NA values is small, we can delete these empty rows which have no significance to our analysis.

```
## # A tibble: 9 x 2
##   density count
##     <dbl> <int>
## 1   0.001    10
## 2   0.08     10
## 3   0.148    10
## 4   0.223    10
## 5   0.318    10
## 6   0.412    10
## 7   0.508    10
## 8   0.604    10
## 9   0.686    10
```

From the table above, we have 9 levels of density in this dataset. Each level has ten observations. We can make a scatter plot to show the relationship between gain and density. Relation of log(gain) and density after transformation is also as follows.

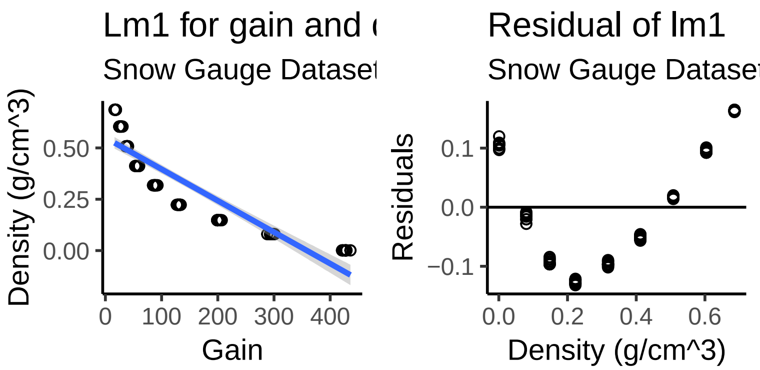




From their relation curve, we can describe that the snow density will decrease as the gain value becomes larger. However, it seems not to match a linear relation.

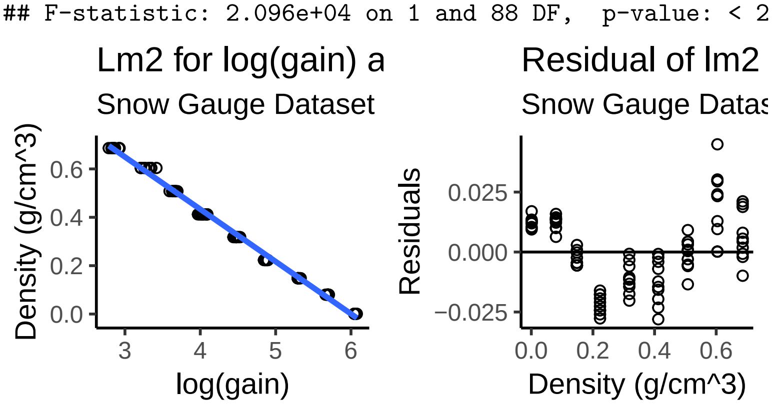
b. Build models and make transformation

We can firstly try to build a simple linear model for gain and density.



From the linear model figure, we can observe the fit result of simple linear model is not good. It has deviation with the relation trend. The residual points do not show a random scatter pattern which indicates the nonlinear relation. Next use log(gain) to build a second model.

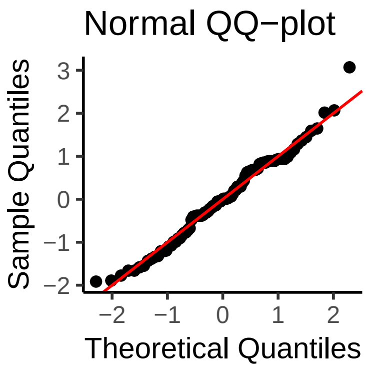
```
##
## Call:
## lm(formula = density ~ log_gain, data = snow_gauge)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028031 -0.011079 -0.000018  0.011595  0.044911
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.298013   0.006857   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
```



The model line fits the data points well and the residual plot looks better. Thus after transformation, the lm2 model is more appropriate than lm1.

c. Limitations and assumptions

Normality of residuals. The residual errors are assumed to be normally distributed. Use QQ plot to check this assumption.



Homogeneity of variance. The assumption of homogeneity of variance means that the level of variance for a particular variable is constant across the sample. The assumption of homogeneity is important for ANOVA testing and in regression models. In ANOVA, when homogeneity of variance is violated, there is a greater probability of falsely rejecting the null hypothesis. In regression models, the assumption needs to be checked with regards to residuals.

```
##
##  Bartlett test of homogeneity of variances
##
## data:  log_gain by density
## Bartlett's K-squared = 55.194, df = 8, p-value = 4.048e-09
```

The p-values is less than 0.05 suggesting variances are significantly different and the homogeneity of variance assumption has been violated. It indicates difference between levels. This may influence the anova result as well.

d. Compare models

```
## Analysis of Variance Table
##
## Model 1: density ~ 1
## Model 2: density ~ log_gain
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     89 4.5567
## 2     88 0.0191  1    4.5376 20956 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fit data to null model and compare two models with anova. The result indicates lm2 is better with less sum of squares and significant p value of F test.

e. Prediction

With the model between log(gain) and density, we can use a gain reading to predict density with a confidence interval.

```
##         fit       lwr       upr
## 1 0.5081678 0.5042423 0.5120933
```

```
##           fit          lwr          upr
## 1 -0.01133153 -0.01695305 -0.005710022
```

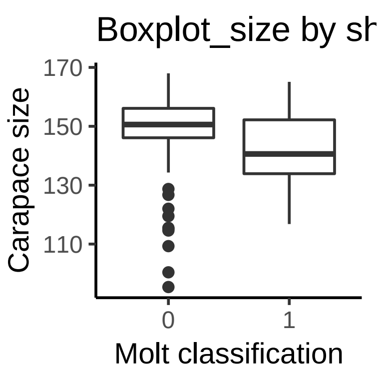
2. Crabpop dataset

a. View data and plot variables

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

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

```
# Pairwise bloxplot
crab_boxplot <- crab_tbl %>%
  ggplot(aes(x = shell,y = size)) +
  theme_classic() +
  geom_boxplot() +
  labs(title = "Boxplot_size by shell",
       x = "Molt classification",
       y = "Carapace size")
cowplot::plot_grid(crab_boxplot)
```



b. Summary statistics

```
## # A tibble: 1 x 2
##   size_mean size_sd
##       <dbl>   <dbl>
## 1      145.    11.8
```

```
## # A tibble: 2 x 5
##   shell shell_mean shell_median shell_sd shell_size
##   <chr>      <dbl>        <dbl>    <dbl>      <int>
## 1 0           149.         151.     11.3        161
## 2 1           142.         141.     11.4        201
```

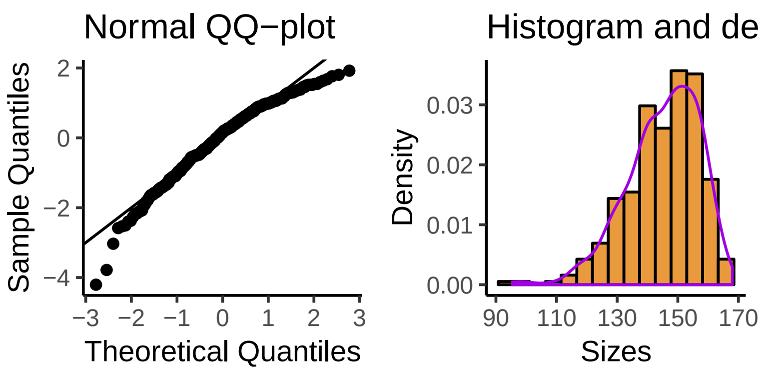
```
## # A tibble: 3 x 2
##   type      SS
##   <chr>  <dbl>
## 1 total 50681.
```

```
## 2 error 46305.
## 3 model  4376.
```

```
size_anova <- aov(size ~ shell,data = crab_tbl)
summary(size_anova)
```

```
##              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
```

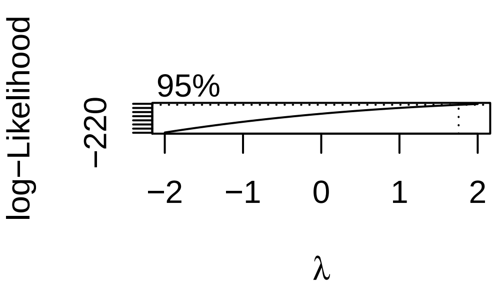
c. Assumptions




```
# Courtesy Alex Stringer.
library(MASS)
```

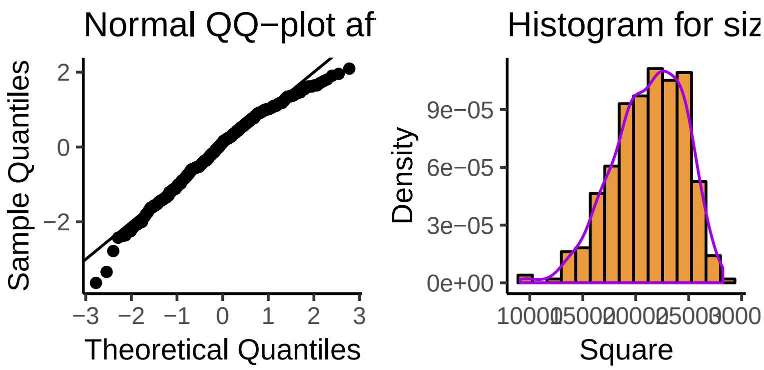
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
crab_boxcox <- boxcox(size ~ 1,data=crab_tbl)
```




```
crab_boxcox$x[which(crab_boxcox$y == max(crab_boxcox$y))]
```

```
## [1] 2
```



d. Models and anova

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)    
## shell         1 3.700e+08 370025501    36.4 3.98e-09 ***
## Residuals   360 3.659e+09  10164326                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## Residuals   361 4.029e+09 11161172
```

```
## Analysis of Variance Table
## 
## Model 1: square ~ 1
## Model 2: square ~ shell
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)    
## 1    361 4029182949                                  
## 2    360 3659157447  1 370025501 36.404 3.984e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```