INTERMEDIATE ANALYTICS

ALY6015, FALL 2019

MODULE 1 ASSIGNMENT

DESCRIPTIVE STATISTICS AND REGRESSION ANALYSIS WITH R

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

CRN: 71933

SUBMITTED TO: PROF. LI, TENGLONG

DATE: 11/04/2019

Introduction

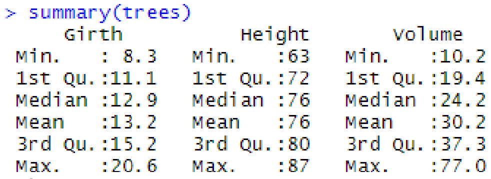
The assignment aims at describing the data numerically and graphically using multiple regression to predict influential variables. We have used R programming and functions to perform regressions for real world data.

Analysis

Part A: Using R functions to describe data numerically and graphically

We have used the in-built dataset “trees” provided by R for describing the data numerically and graphically. Using some of the functions present in R, we have performed the following:

1. The dataset “trees” has been used by viewing the “trees” dataset, by the View(trees) function which allows us to view the dataset.
2. We have utilized the summary(trees) function to display the 5 number summary which includes the minimum value, 1st quartile, median, mean, 3rd quartile and maximum values for the Height, Volume and Girth of the “trees” dataset.



```
> summary(trees)
     Girth           Height       Volume    
 Min.   : 8.3   Min.   :63   Min.   :10.2  
 1st Qu.:11.1   1st Qu.:72   1st Qu.:19.4  
 Median :12.9   Median :76   Median :24.2  
 Mean   :13.2   Mean   :76   Mean   :30.2  
 3rd Qu.:15.2   3rd Qu.:80   3rd Qu.:37.3  
 Max.   :20.6   Max.   :87   Max.   :77.0  
```

Fig.1: Output of the summary function

3. In order to plot the regression lines between two variables, we have used the plot() and the abline() and lm() functions as shown in the figures below. The regression lines shows the changes and predictions in the variable x with respect to variable y. The values for Girth vs Volume, are more near the slope whereas for Girth vs Height, they are more scattered.

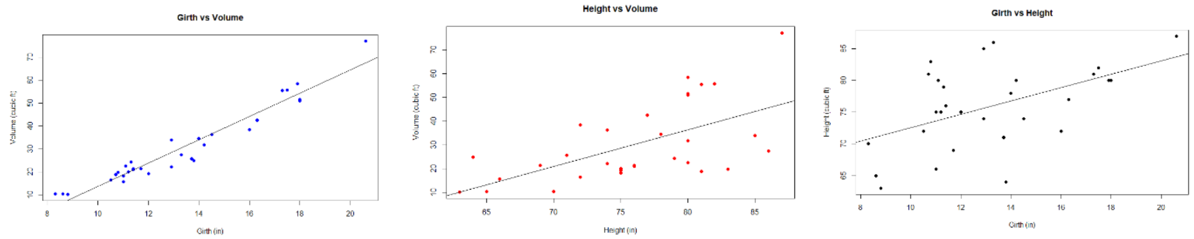


Fig.2: The plotting of regression lines

4. The histogram has been plotted using the hist() function. It can be observed that the histogram for height shows distribution equivalent to a normal distribution.

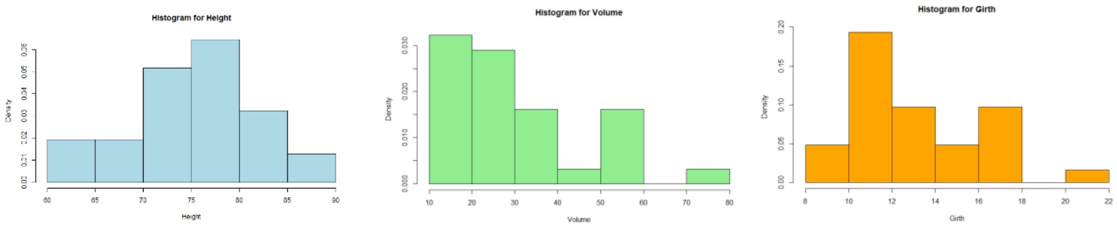


Fig.3: Histograms for Height, Volume and Girth

The density plots for Volume, Girth and Height have been plotted using the density() and plot() functions. A density plot shows the distribution of a numeric variable. As it is observed, the density for the height is highest and has a peak value at 78, whereas, the value is lowest for Girth.

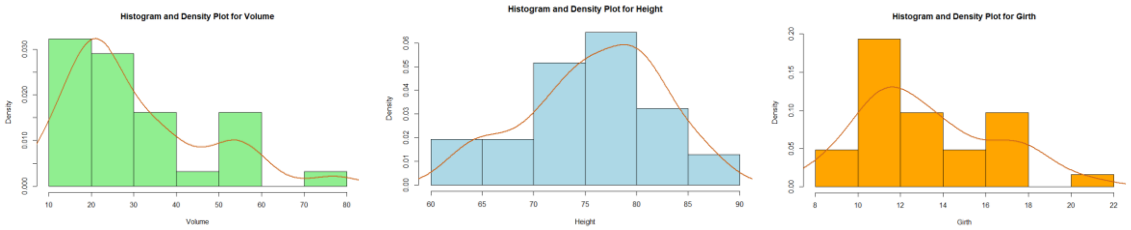


Fig.4: Density Plots and Histograms for Height, Volume and Girth

5. The box plot has been plotted for Height, Volume and Girth using the boxplot() function. Boxplots give a visual representation about the Quartiles, IQR, mean, median and outliers.

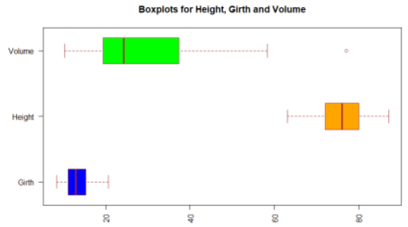


Fig.5: Boxplot for Volume, Girth and Height

It can be observed, the values for volume are more dispersed, with median of 22 with a greater range and also has an outlier. The values for Girth are less dispersed with median of 10 and Height having median of 78.

6. Normal probability plots have been plotted using the qqnorm() and qqline() functions for Height, Volume and Girth. It can be interpreted, that the points have a linear pattern which indicates that they are normally distributed.

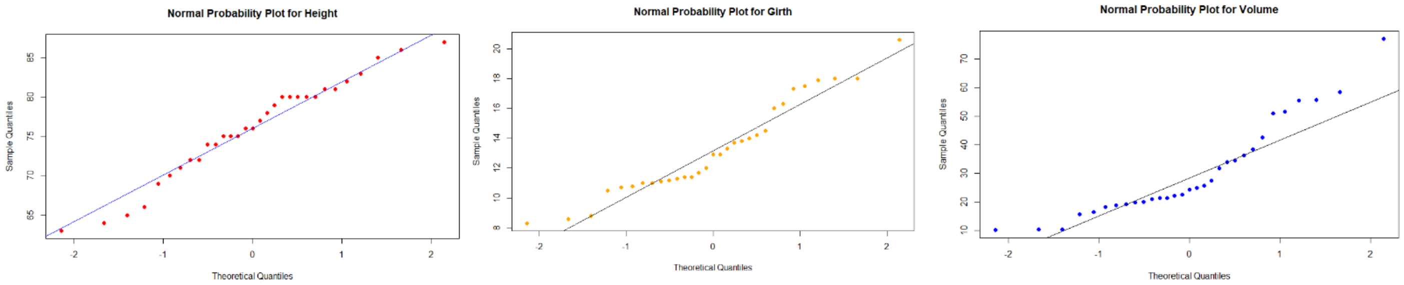
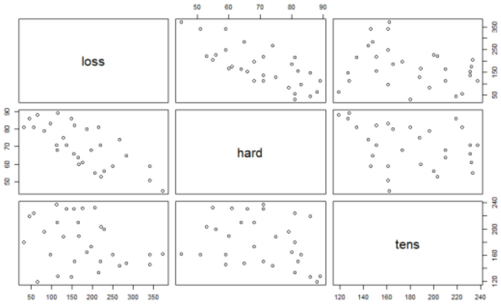
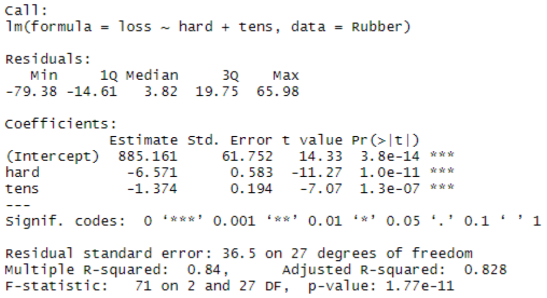


Fig.6: Normal Probability Plot for Volume, Girth and Height

Part B: Using R functions to build a multiple regression model for real world data

We have used the Rubber dataset from the MASS and DAAG package containing variables like loss, hard and tens. Using this data, we have obtained the scatterplot matrix for the variables. It can be interpreted from the scatterplot matrix for the Rubber dataset, that there exists a negative correlation between the loss and hardness. So, we regress the loss on hard and tens by using Rubber.lm <- lm(loss~hard+tens, data=Rubber) function.



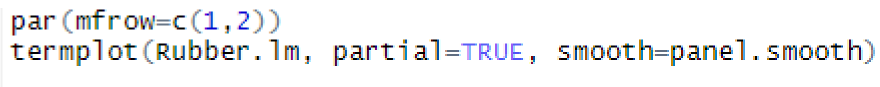


```
Call:
lm(formula = loss ~ hard + tens, data = Rubber)

Residuals:
   Min     1Q Median     3Q    Max 
-79.38 -14.61   3.82  19.75  65.98 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  885.161     61.752   14.33  3.8e-14 ***
hard          -6.571      0.583  -11.27  1.0e-11 ***
tens          -1.374      0.194   -7.07  1.3e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.5 on 27 degrees of freedom
Multiple R-squared:  0.84,	Adjusted R-squared:  0.828 
F-statistic:    71 on 2 and 27 DF,  p-value: 1.77e-11
```

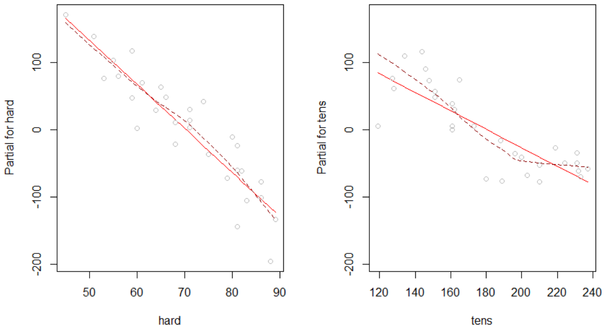
Fig.7: Scatterplot matrix for Rubber dataset

Fig.8: Output showing summary for Rubber

We have used the termplot() function to perform the linear regression for rubber dataset. The code used for the same is:



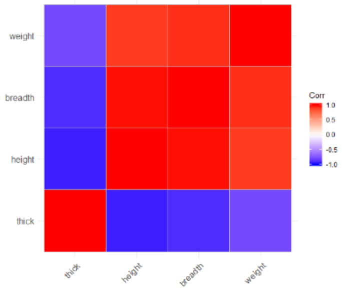
```
par(mfrow=c(1,2))
termplot(Rubber.lm, partial=TRUE, smooth=panel.smooth)
```

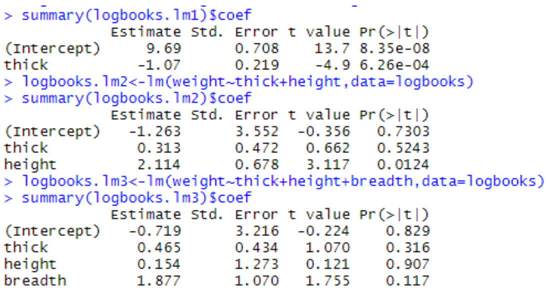


The plot shows the mean contributions for each term in the model. It can be observed that the response is not linear with the tensile strength. While, the response appears to be linear for hardness.

Fig.9: Plots obtained using termplot() function

We have used the oddbooks dataset which covers a wide range of weight to height ratios. It can be observed from the dataset, that, as the thickness increases, the weight decreases.





```
> summary(logbooks.lm1)$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.69      0.708    13.7 8.35e-08
thick          -1.07      0.219    -4.9 6.26e-04
> logbooks.lm2<-lm(weight~thick+height,data=logbooks)
> summary(logbooks.lm2)$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.263      3.552  -0.356   0.7303
thick          0.313      0.472   0.662   0.5243
height         2.114      0.678   3.117   0.0124
> logbooks.lm3<-lm(weight~thick+height+breadth,data=logbooks)
> summary(logbooks.lm3)$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.719      3.216  -0.224    0.829
thick          0.465      0.434   1.070    0.316
height         0.154      1.273   0.121    0.907
breadth        1.877      1.070   1.755    0.117
```

Observations: The correlations between height, thickness and weight are very strong. We have calculated the error and test statistic to compare the height, width and thickness parameters and perform analysis.

Fig.10: Correlation between parameters

Fig.11: Output of the logbooks summary function

Reference

1. Maindonald, J. H., & Braun, J. (2010). Data analysis and graphics using R: an example-based approach. Cambridge: Cambridge University Press.
2. Kabacoff, R. (n.d.). R Tutorial. Retrieved from https://www.statmethods.net/r-tutorial/index.html.