Yuyi Zhang

Apr 16, 2019

Spring 2019 CPS Quarter Term A

Instructor: Steward Huang

Assignment 1

ALY 6015\_Intermediate Analytics

# Instructions

Part (A): Use R functions to describe data numerically and graphically.

Part (B): Use R functions to build a multiple regression model for real world data.

# Analysis

## Part A

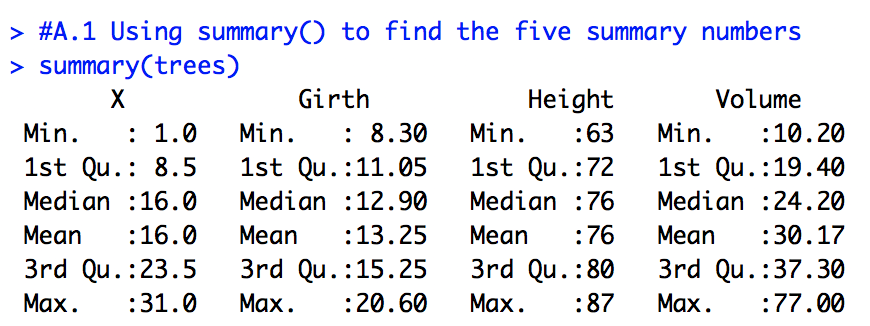
### Using summary() to find the five summary numbers

**Code:**

> trees <- read.csv("FCR/NEU/CPS/Analytics\_2018/ALY 6015\_Intermediate Analytics/Week 1/Assignment 1\_trees.csv", header = TRUE)

> summary(trees)

**Console:**



All summary numbers have been shown up in console.

### Using summary() and lm () functions for straight line regression function

**Code:**

# Independent Variable

> Girth <- trees[,c('Girth')]

# Dependent Variables

> Height <- trees[,c('Height')]

> Volume <- trees[,c('Volume')]

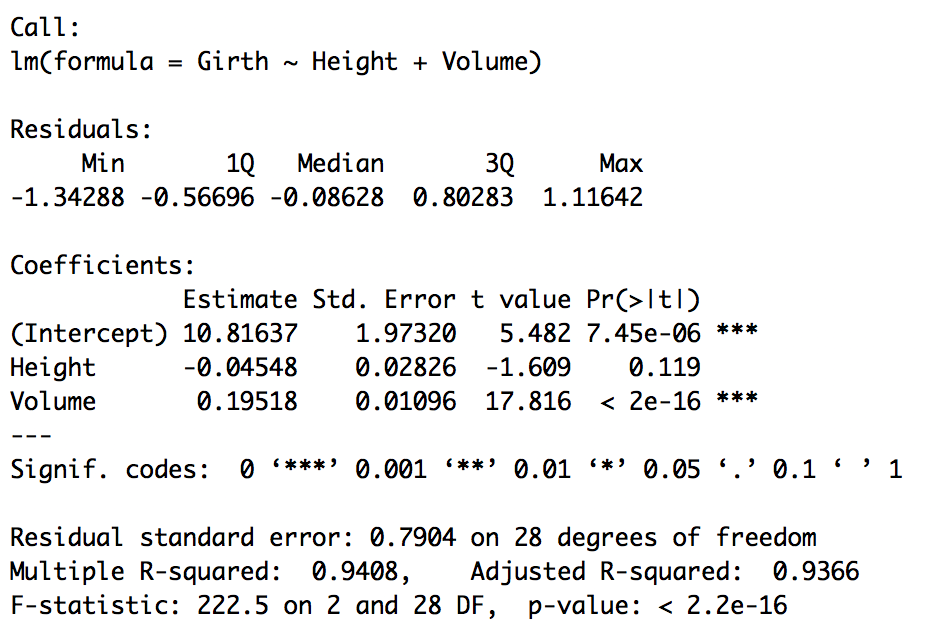
# Creating Regression Equation

> Regression <- lm(Girth ~ Height + Volume)

# Show the results

> summary(Regression)

**Console:**



### Using hist (), density() function for Histograms and Density Plots

**Code:**

# Create Histogram

> hist(Girth, breaks = 20, xlim = c(5,25), ylim = c(0,6), col = "#ffffe6", main = "Distribution of Black Cherry Trees by Girth", xlab = "Girth")

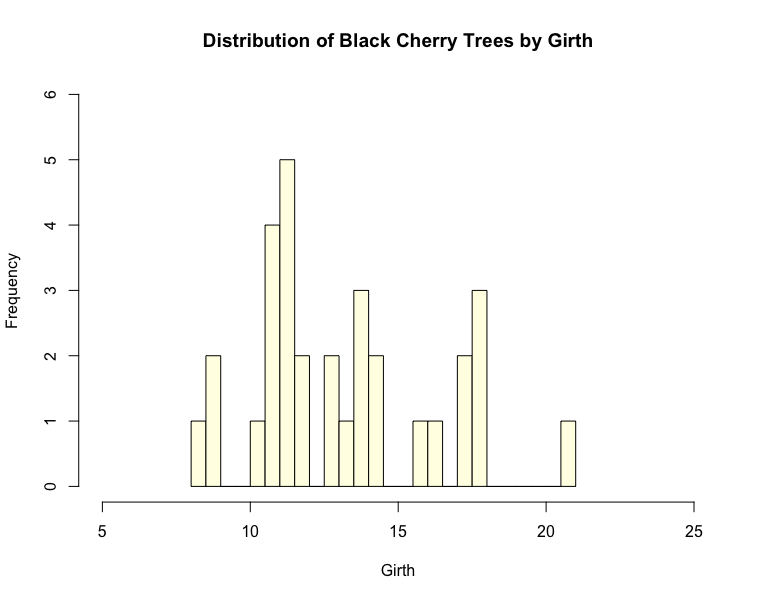
# Create Density Plots

> g <- density(Girth)

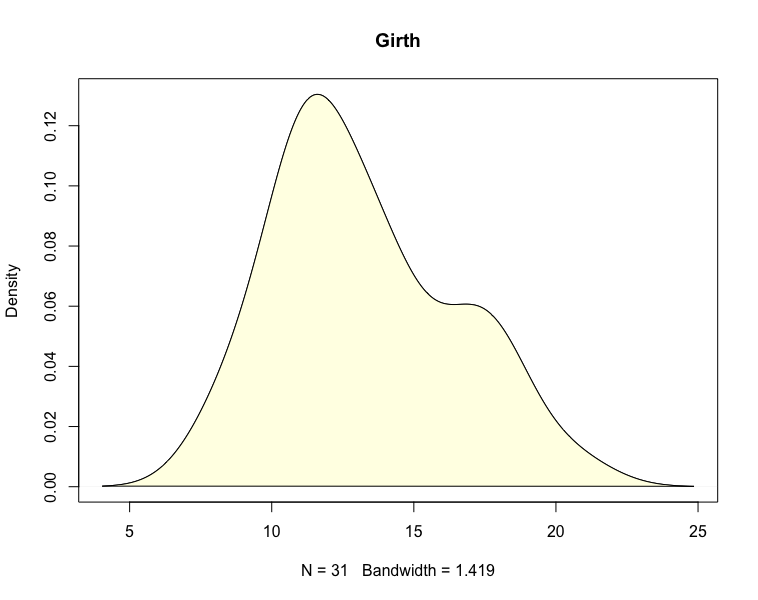
> plot(g, main = "Girth")

> polygon(g, col = "#ffffe6")

**Console:**



*Figure 1*. Histograms of Girth



*Figure 2*. Density Plots of Girth

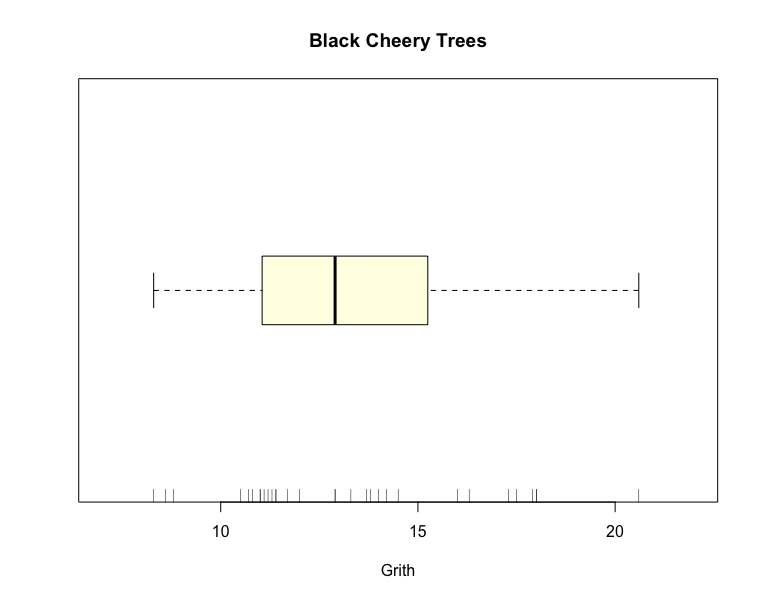
### Using boxplot() and rug() for Boxplots

**Code:**

> boxplot(Girth, col = "#ffffe6", main = "Black Cheery Trees", xlab = "Grith", ylim = c(7,22), frame.plot = TRUE, boxwex = 0.35, horizontal = TRUE)

> rug(Girth, side = 1)

**Console:**



*Figure 3*. Boxplots of Girth

### Using rnorm () and qqnorm () for Normal probability plots

**Code:**

#rnorm

> plot(rnorm(5000), main = "rnorm plots (mean=0, sd=1)")

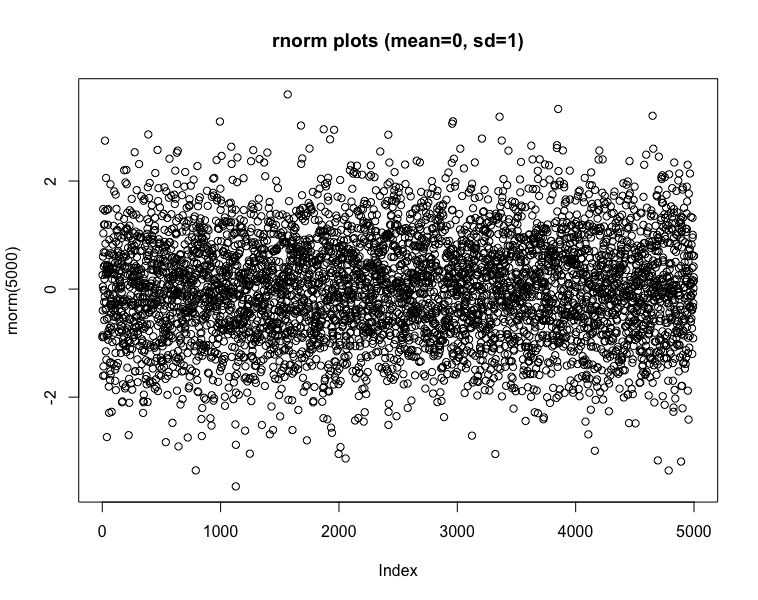
#qqnorm

> head(trees)

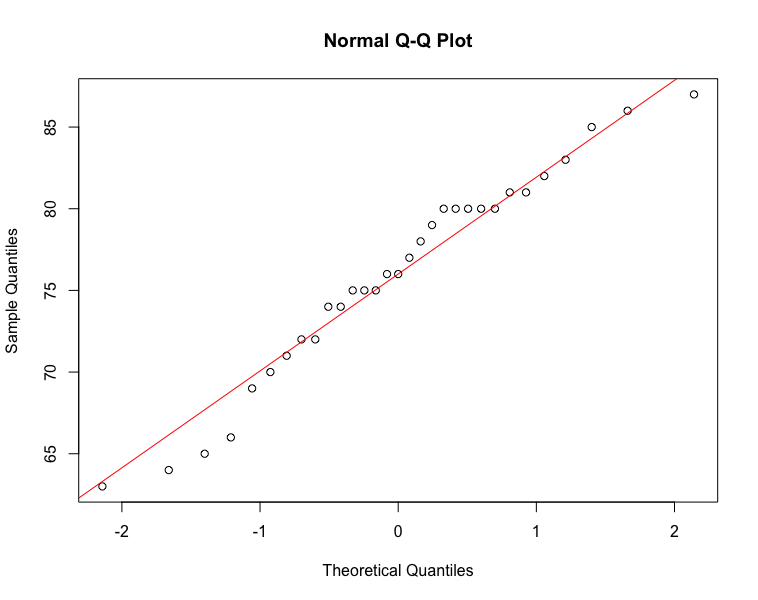
> qqnorm(trees$Height)

> qqline(trees$Height, col = "red")

**Console:**



*Figure 4*. Normal Probability Plots



*Figure 5*. QQ Plots of Height

## Part B

**Code:**

> library(MASS)

> library(ggplot2)

> library(ggcorrplot)

> library(DAAG)

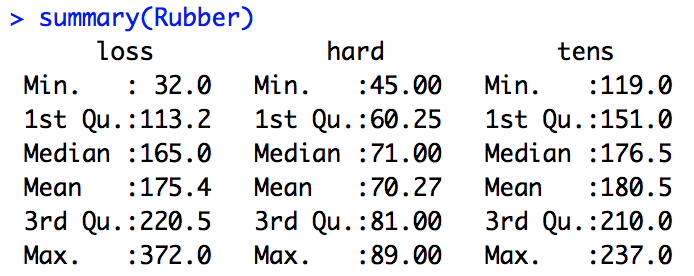
### Rubber

**Code:**

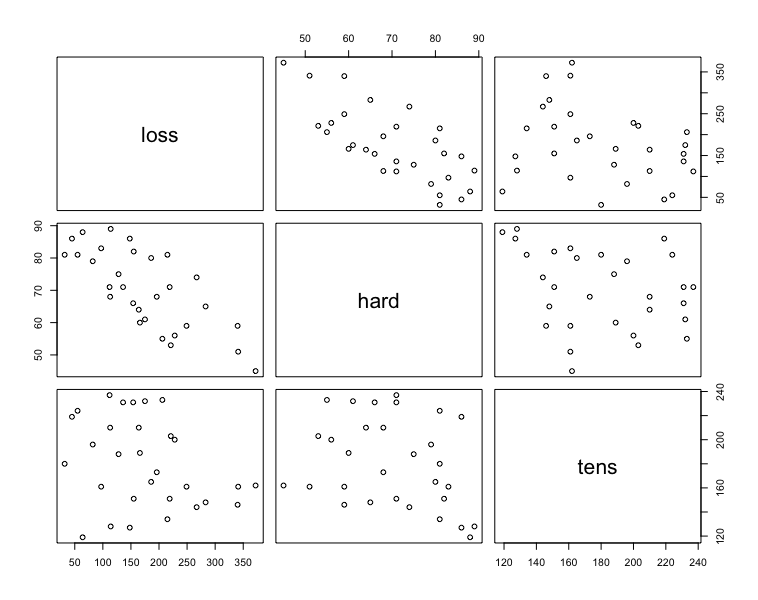
> summary(Rubber)

> pairs(Rubber)

**Console:**

****

*Figure 6*. Summary of Rubber



*Figure 7*. Matrix Scatterplots of Rubber

**Code:**

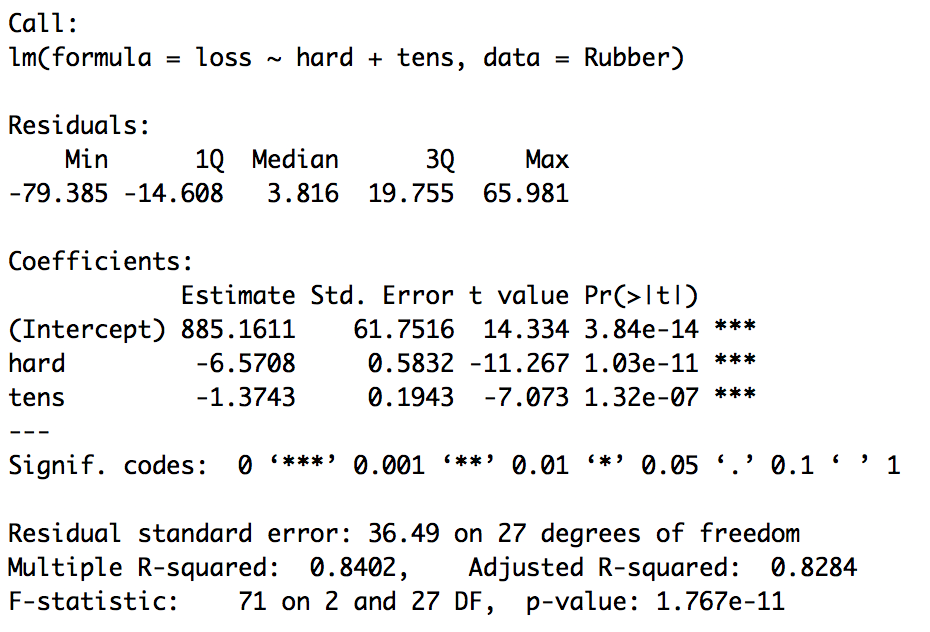
> Rubber.lm <- lm(loss~hard+tens, data = Rubber)

> summary(Rubber.lm)

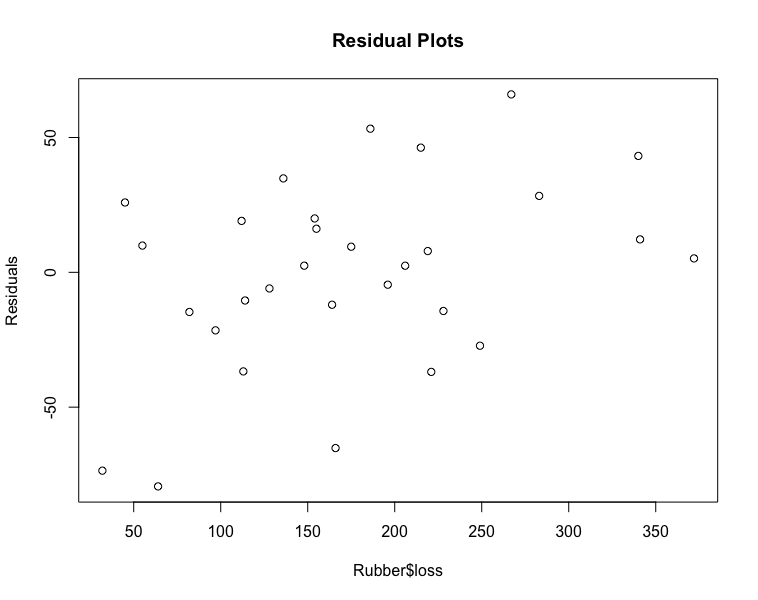
> Residuals <- resid(Rubber.lm)

> plot(Rubber$loss,Residuals, main = 'Residual Plots')

**Console:**



*Figure 8*. Residual Summary of Rubber



*Figure 9*. Residual Plots of Rubber

To illustrate, the value of multiple coefficient of determination, R-squared is 0.84. This high value of R-squared implies that using the independent variables hard and tens explains 84% of the total sample variation in loss.

**Code:**

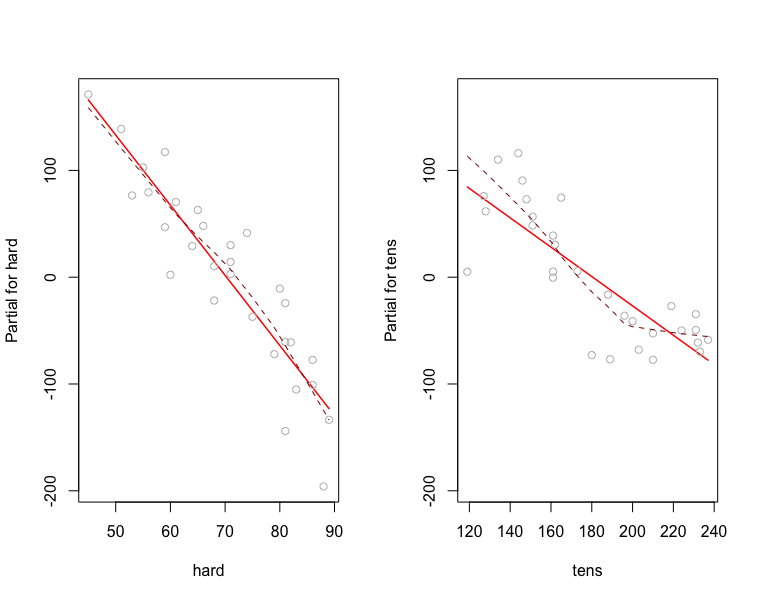
> par(mfrow=c(1,2))

> termplot(Rubber.lm, partial = TRUE, smooth = panel.smooth)

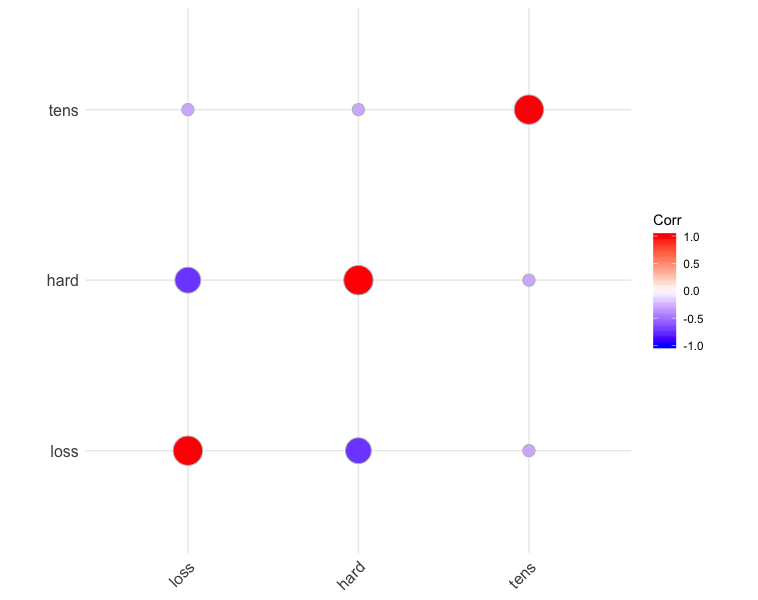
> par(mfrow=c(1,1))

> ggcorrplot(cor(Rubber), method = "circle")

**Console:**



*Figure 10*. Plots Regression Terms of Rubber



*Figure 11*. Correlation Plots of Rubber

According to correlation plots, hard and loss are quite negative related.

### Oddbooks

**Code:**

> logbooks <- log(oddbooks)

> logbooks.lm1 <- lm(weight ~ thick, data = logbooks)

> summary(logbooks.lm1)

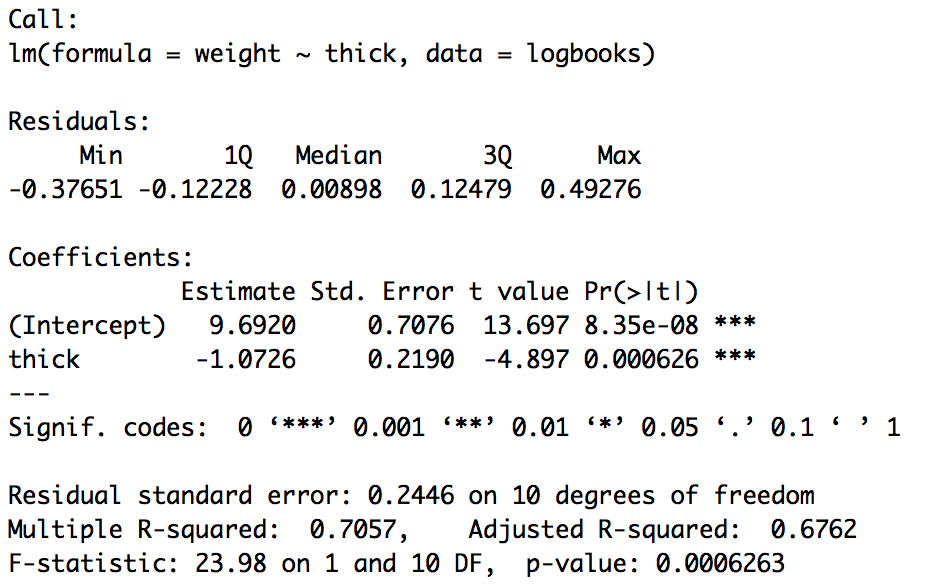
> logbooks.lm2 <- lm(weight ~ thick + height, data = logbooks)

> summary(logbooks.lm2)

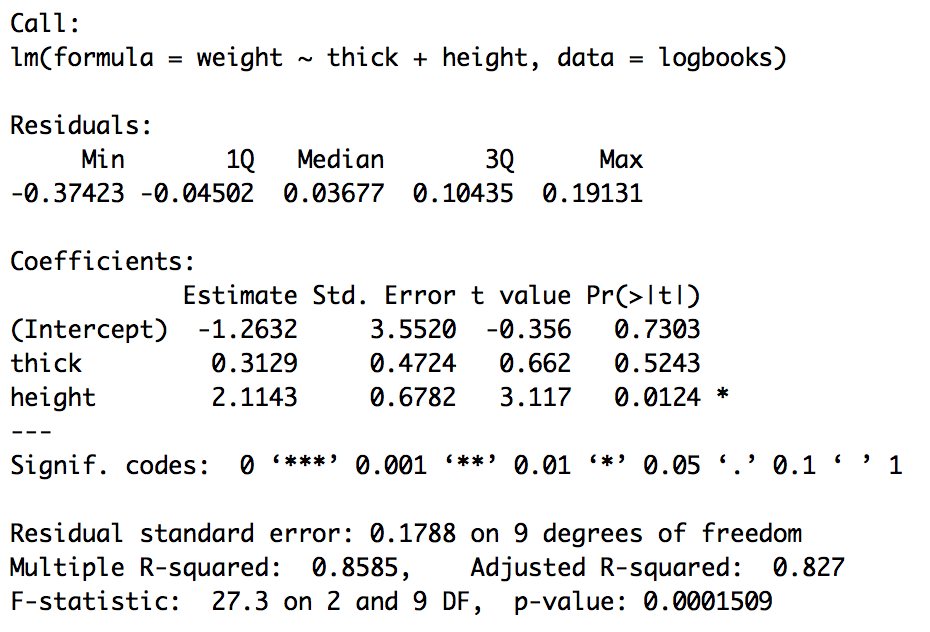
> logbooks.lm3 <- lm(weight ~ thick + height + breadth, data = logbooks)

> summary(logbooks.lm3)

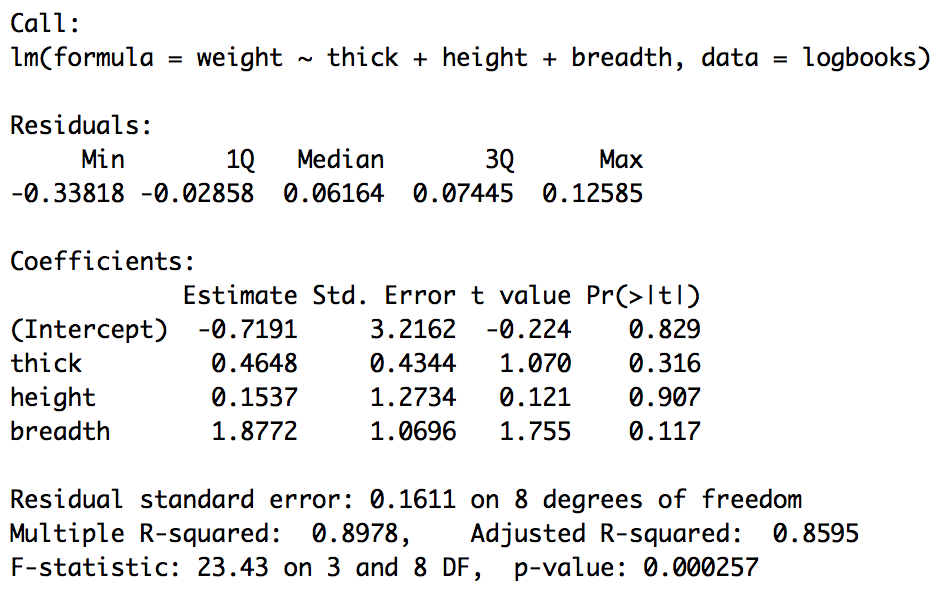
**Console:**



*Figure 12*. Residual Summary of weight ~ thick



*Figure 13*. Residual Summary of weight ~ thick & height



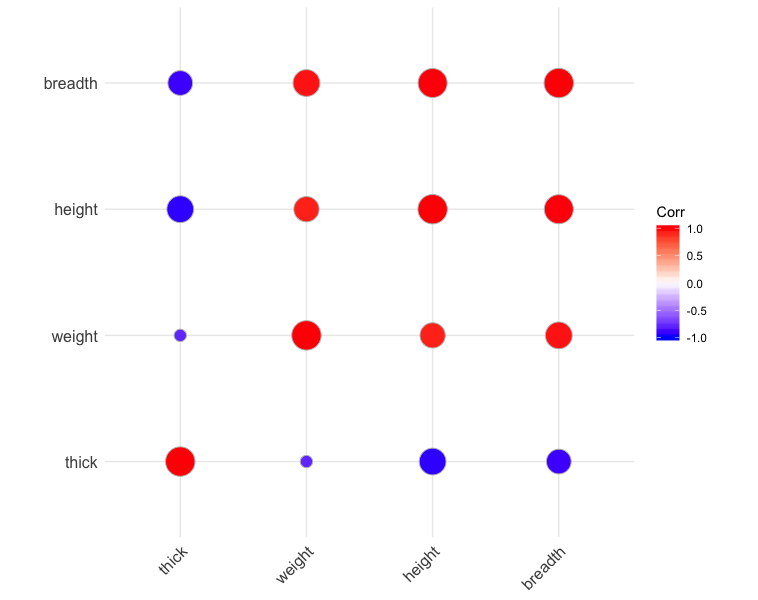
*Figure 14*. Residual Summary of weight ~ thick, height & breadth

The value of R-squared is gradually increased since more independent factors have been added one by one. Therefore in logbooks.lm3, independent variables thick, height and width could explain 89.8% of the total book weight.

**Code:**

> ggcorrplot(cor(oddbooks), hc.order = TRUE, method = "circle")

**Console:**



*Figure 15*. Correlation Plots of Oddbooks

According to correlation plot, breadth and height are positive related. Thick is negative related with breadth and height.

Reference:

1. Maindonald, J. H. (2008). *Using R for Data Analysis and Graphics.* Retrieved from https://cran.r-project.org/doc/contrib/usingR.pdf