Mustafa Abdullayev

Descriptive Analysis With R

AY6015 – Intermediate Analytics

Term: Winter 2020 - A

Instructor: Joe Manseau

01/14/2020

Introduction

This is a Microsoft Word Report accompanying R Script. It contains my R code, outputs, my comments, and findings. In my analysis, I used 3 datasets: trees, Rubber, and oddbooks. These are built-in datasets in R. Data was cleaned, organized and ready for analysis. My main aim was to utilize R and its statistical methods to analyze data and find insights. I mainly used descriptive analysis, 5-number summary, and regression analysis. Also, I utilized powerful R built-in functions and graphs, such as box plots and histograms, to dive deeper to observe hidden patterns and visually communicate my findings to the audience. Since I also provided R script with all the codes and comments, I removed some of the codes and comments from my report (such as package loading). It is due to keep my report brief, succinct and to the point.

Part A Analysis

Firstly, I analyzed trees datasets. From their summary report, it seems that only Height values have a normal distribution since its mean is equal to the median. On the other hand, both Girth and Volume are positively skewed. I will check this assumption by using histograms later.

From Regression plots, the most obvious relationship is between Girth and Volume. They have a strong positive correlation between them. Also, we can see that its regression line has high accuracy in predicting Volume. Moreover, I do not think that there is a confounding variable and I would say that there is a causation between Girth and Volume.

# N2: 5-number summary numbers  
summary(trees)

## Girth Height Volume   
## Min. : 8.30 Min. :63 Min. :10.20   
## 1st Qu.:11.05 1st Qu.:72 1st Qu.:19.40   
## Median :12.90 Median :76 Median :24.20   
## Mean :13.25 Mean :76 Mean :30.17   
## 3rd Qu.:15.25 3rd Qu.:80 3rd Qu.:37.30   
## Max. :20.60 Max. :87 Max. :77.00

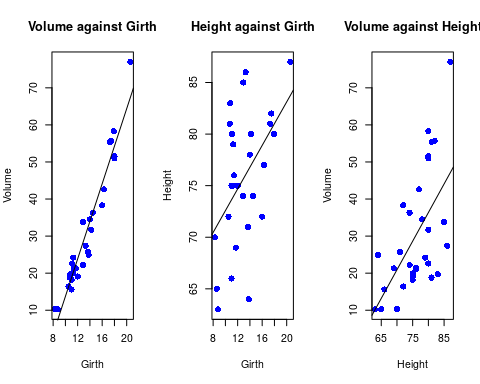
# N3 : Graphing straight line Regression for all pairs (3 pairs)  
plot(trees$Girth,trees$Volume,pch = 16,cex = 1.3, col = "blue",main = "Volume against Girth",xlab = "Girth",ylab = "Volume")  
abline(lm(trees$Volume ~ trees$Girth))

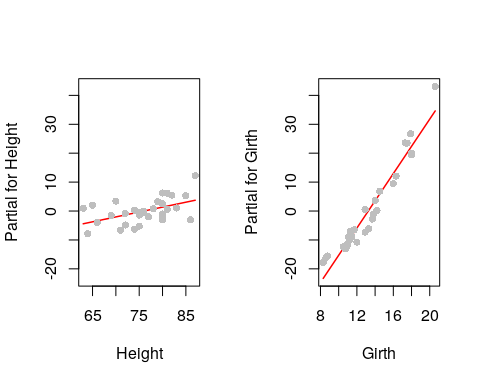
### Same code for other 2 with different variables ###

# Regression line for Volume versus Height and Girth

trees.mrl <- lm(Volume ~ Height+Girth,data = trees)  
termplot(trees.mrl,partial = TRUE)

In the second graph, I plotted volume versus both Height and Girth. Girth is the major factor in determining the Volume of a tree.

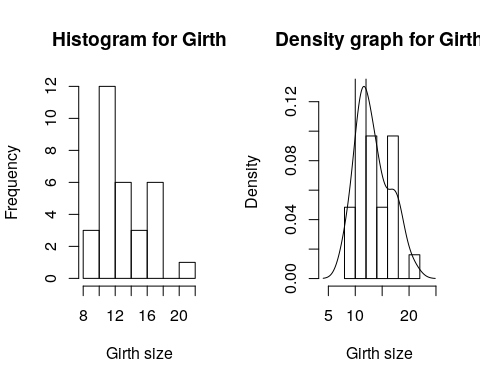


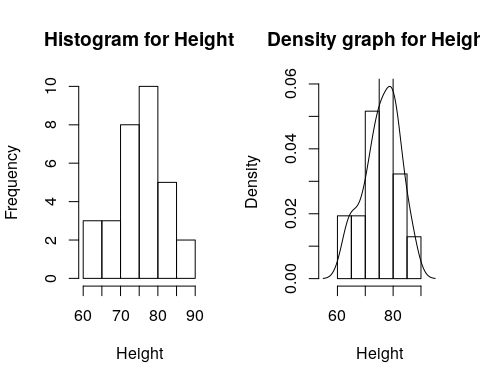


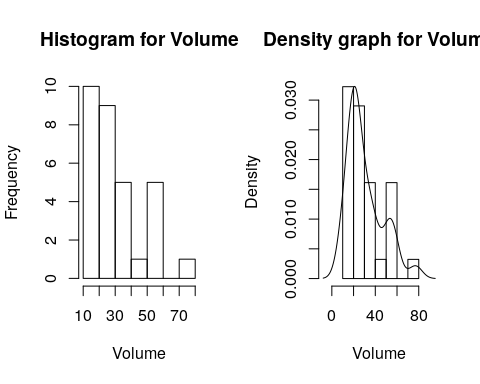
# N4: Histograms and Density Plots  
hist(trees$Girth,main = "Histogram for Girth",xlab ="Girth size" )  
dens <- density(trees$Girth)  
xlim <- range(dens$x)  
ylim <- range(dens$y)  
hist(trees$Girth,probability = TRUE,xlim = xlim,ylim = ylim,main = "Density graph for Girth",xlab = "Girth size")  
lines(dens)

### Same code for other 2 with different variables ###

Afterward, I dived deeper to check their distribution. From their summary statistics, I supposed that Volume does not have a normal distribution. We can prove this claim by looking at the third graph. It demonstrates that its distribution is positively skewed. Also, its graph suggests that there can be outliers. When I check other distribution, their histograms, and density plots, Height values are the closest to normal distribution. Indeed, its mean and median values are the same. Girth values are also almost symmetric. But anyone should keep in mind that, verifying distribution is should not be done with only histograms since it is highly dependent on selecting bin values.



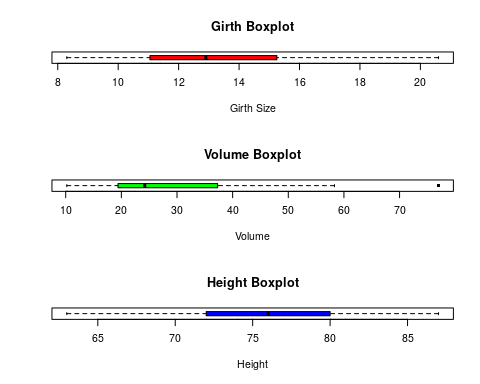




Next, I utilized boxplots to visualize 5 number summaries. Once more, I observed that Volume values and Girth values are positively skewed. Also, Height has an almost normal distribution. Moreover, one more thing boxplot added to my analysis is that I verified that there is an outlier point in Volume. The definition of an outlier can change from person to person. I define an outlier to be 1.5 times the IR above the third quartile or below the first quartile (mathworld.wolfram.com).

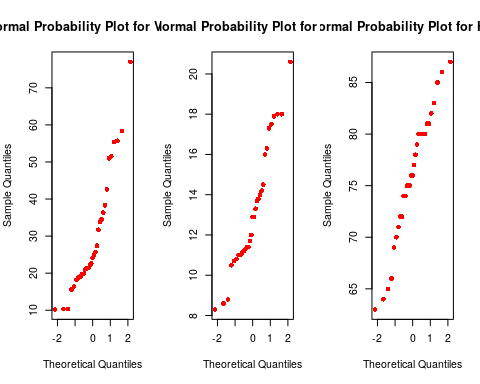
# N5: Boxplots

boxplot(trees$Girth,horizontal = TRUE, main = "Girth Boxplot", xlab = "Girth Size", outline = TRUE, col = "red")  
rug(trees$Girth,side = 1)  
### Same code for other 2 with different variables ###



Finally, I used QQ plots to check the normality of my data points. Again, I observed that only Height values lie along the perfect normal line (the majority of them). And since my dataset is small, I can assume that Height values are normally distributed. On the other hand, Volume values away from Perfect normal line.

# N6: Normal Probability plots  
qqnorm(trees$Volume,col = "red",main = "Normal Probability Plot for Volume")  
qqnorm(trees$Girth,col = "red",main ="Normal Probability Plot for Girth")  
qqnorm(trees$Height,col = "red",main = "Normal Probability Plot for Height")



Part B Analysis

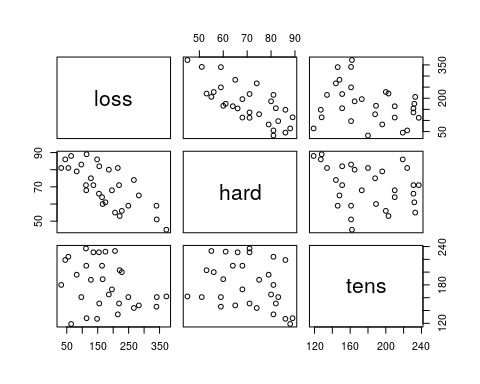
In this part of my analysis, I used Rubber and oddbooks datasets. These datasets are provided by R, in the packages of MASS and DAAG. Rubber dataset is from accelerated testing of tire rubbers. Here loss, hard, tens means the abrasion loss in gm/hr, the hardness in Shore units and tensile strength kg/sq m, respectively. Oddbooks data set is about thickness, height, breadth, and weight of selected books.

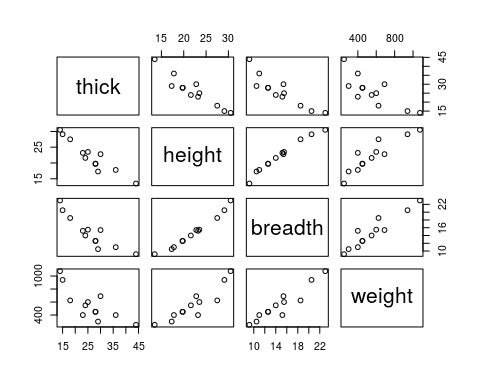
At first, I used scatterplot to observe their correlation between each other. Tensile strength in the Rubber dataset is not in correlation with loss or hard (or correlation is not impressive). On the other hand, hardness and loss are in negative correlation.

But for oddbooks, it seems all of them are correlated with each other in some way. Especially breadth and height are strongly correlated. To test these claims I used R correlation plot.

pairs(Rubber)

pairs(oddbooks)

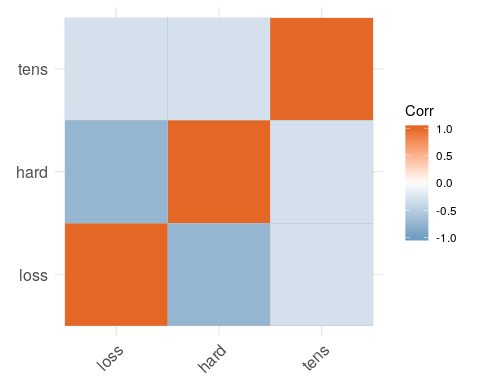
a

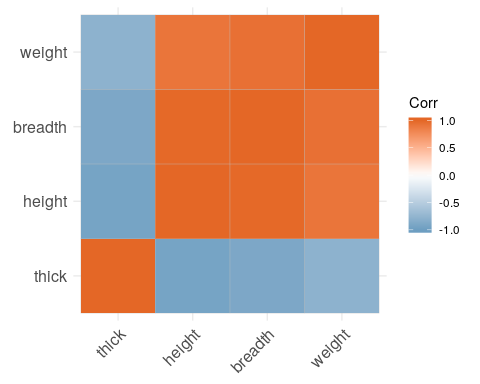


####### correlation plots  
ggcorrplot(cor(Rubber),method = "square", show.diag = TRUE,colors = c("#6D9EC1", "white", "#E46726"))

####### oddbooks   
ggcorrplot(cor(oddbooks),method = "square", colors = c("#6D9EC1", "white", "#E46726"))

Indeed, we can see that all the above claims are true. There are strong correlations between oddbooks dataset variables and are weak between Rubber dataset values.



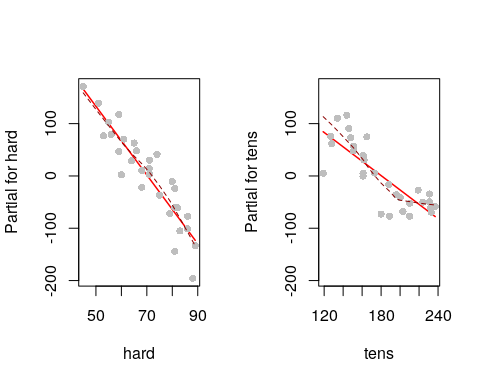


I continued my analysis with regression analysis. Firstly, we can see that, R-squared for the Rubber.lm model, which is trying to predict loss versus hardness and tensile, is almost 83. This indicates that our model is working nicely and can explain almost 83% of variance around its mean. Also, from the second figure I can see that, hardness factor is more important for predicting a loss.

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

##   
## Call:  
## lm(formula = loss ~ hard + tens, data = Rubber)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.385 -14.608 3.816 19.755 65.981   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 885.1611 61.7516 14.334 3.84e-14 \*\*\*  
## hard -6.5708 0.5832 -11.267 1.03e-11 \*\*\*  
## tens -1.3743 0.1943 -7.073 1.32e-07 \*\*\*  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Residual standard error: 36.49 on 27 degrees of freedom  
## Multiple R-squared: 0.8402, Adjusted R-squared: 0.8284

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



In the second part, I used regression analysis in order to analyze oddbooks data. First of all, I took the log values of data. This way I can make them more normally distributed. My main aim was to predict weight with the help of the other 3 variables. I started with only thickness and my model has the 0.68 R-Squared value. This is small considering our model can only explain 67% of variability around the mean.

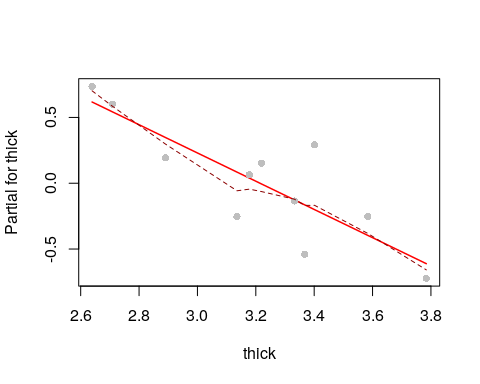
summary(oddbooks)

## thick height breadth weight   
## Min. :14.00 Min. :13.50 Min. : 9.20 Min. : 250.0   
## 1st Qu.:21.75 1st Qu.:19.23 1st Qu.:12.20 1st Qu.: 400.0   
## Median :26.50 Median :22.20 Median :14.60 Median : 500.0   
## Mean :26.17 Mean :22.19 Mean :14.83 Mean : 560.8   
## 3rd Qu.:29.25 3rd Qu.:24.50 3rd Qu.:16.25 3rd Qu.: 641.2   
## Max. :44.00 Max. :30.50 Max. :23.00 Max. :1075.0

logbooks <- log(oddbooks)  
logbooks.lm1 <- lm(weight ~ thick,data = logbooks)  
summary(logbooks.lm1)

##   
## Call:  
## lm(formula = weight ~ thick, data = logbooks)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.37651 -0.12228 0.00898 0.12479 0.49276   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.6920 0.7076 13.697 8.35e-08 \*\*\*  
## thick -1.0726 0.2190 -4.897 0.000626 \*\*\*  
  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
## Residual standard error: 0.2446 on 10 degrees of freedom  
## Multiple R-squared: 0.7057, Adjusted R-squared: 0.6762

termplot(logbooks.lm1,partial = TRUE,smooth = panel.smooth)

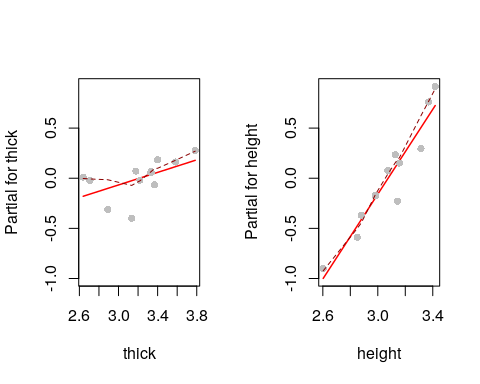


Secondly, I used thickness and height of a book in order to predict its weight. At this scenario, the coefficient of determination increased to 0.83. This is a great increment and it suggested that the height of a book is a better predictor of weight. Also, we should keep in mind that, the correlation between height and thickness was small, thus, using these 2 variables together was a good idea. Finally, I plotted their impact on the final model separately and observed that, indeed, the height of a book is a better predictor.

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

##   
## Call:  
## lm(formula = weight ~ thick + height, data = logbooks)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.37423 -0.04502 0.03677 0.10435 0.19131   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.2632 3.5520 -0.356 0.7303   
## thick 0.3129 0.4724 0.662 0.5243   
## height 2.1143 0.6782 3.117 0.0124   
  
## Residual standard error: 0.1788 on 9 degrees of freedom  
## Multiple R-squared: 0.8585, Adjusted R-squared: 0.827

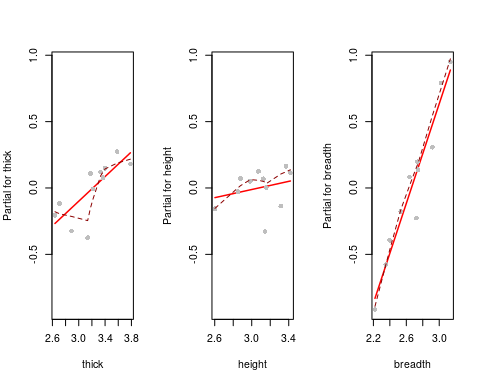
termplot(logbooks.lm2,partial = TRUE,smooth = panel.smooth)



To conclude, I decided to use all 3 variables to predict weight. Before that, I assumed that since the correlation between breadth and height is high, using breadth alongside height will not increase my R-squared value too much. Indeed, my model, logbooks.lm3, had 0.86 R-squared value which is an only 0.03-point increment. (While describing R-squared values, I rounded them to 2 decimal points). But interestingly when I plotted all three of them separately, breadth was the best predictor for the weight of a book.

logbooks.lm3 <- lm(weight ~ thick+height+breadth,data = logbooks)  
summary(logbooks.lm3)  
## Call:  
## lm(formula = weight ~ thick + height + breadth, data = logbooks)  
   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33818 -0.02858 0.06164 0.07445 0.12585   
  
## Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.7191 3.2162 -0.224 0.829  
## thick 0.4648 0.4344 1.070 0.316  
## height 0.1537 1.2734 0.121 0.907  
## breadth 1.8772 1.0696 1.755 0.117  
  
## Residual standard error: 0.1611 on 8 degrees of freedom  
## Multiple R-squared: 0.8978, Adjusted R-squared: 0.8595

termplot(logbooks.lm3,partial = TRUE,smooth = panel.smooth)



Conclusion

To conclude, I analyzed 3 datasets namely trees, oddbooks, and Rubber. These were built-in datasets provided by R. My main aim was to analyze these datasets to find insights and patterns by using R and its functions. I also utilized regression analysis to dive deeper. I found that, for trees dataset, only height values are normally distributed. Also, there is an outlier in Volume values. It is the only outlier I have found. For the second part, I mainly used regression analysis and correlation analysis. The most interesting fact was that there is a high correlation between oddbooks dataset variables.

References

Maindonald, J. H., & Braun, W. J. (2019, May 2). oddbooks: Measurements on 12 books in DAAG: Data Analysis and Graphics Data and Functions. Retrieved from  [https://rdrr.io/cran/DAAG/man/oddbooks.html](%09https://rdrr.io/cran/DAAG/man/oddbooks.html)

Outlier. (n.d.). Retrieved from <http://mathworld.wolfram.com/Outlier.html>.

pmagunia. (2018, March 9). Retrieved from [https://app.quadstat.net/dataset/r-dataset- package-mass-rubber](https://app.quadstat.net/dataset/r-dataset-%09package-mass-rubber).