Descriptive Statistics and Regression analysis: Module 1 Project

G. Holt Williams

Northeastern University: ALY6015

Introduction

The purpose of this assignment is to practice coding in R and presenting findings both numerically and graphically to further our understanding of how these methods could be utilized for real world business problems. The following report is broken into two analysis sections with part A containing the bulk of visualizations of the “Tree” dataset, while part B looks at the “Rubber” and “oddbooks” datasets and the use of regression analysis.

Analysis: Part A

This portion of the assignment is a simple exploratory data analysis on the “Trees” dataset which has 31 observations of three variables: Girth (also known as diameter, in inches), Height, and Volume. Using the following R code\* the five summary numbers are shown in table 1.

Data(“trees”)

knitr::kable(summary(trees))

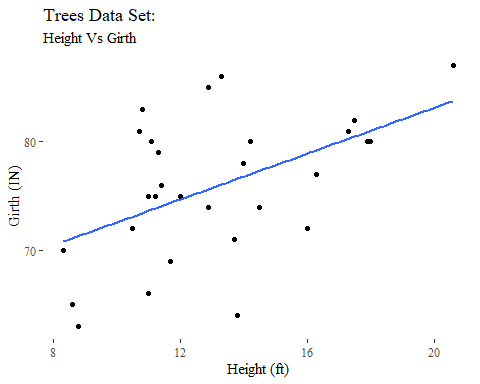
|  |  |  |  |
| --- | --- | --- | --- |
| Table 1: Five Summary Numbers (Should I add more relevant stuff? | | | |
|  | Girth | Height | Volume |
|  | Minimum: 8.30 | Minimum: 63 | Minimum: 10.20 |
|  | 1st Quartile:11.05 | 1st Quartile:72 | 1st Quartile: 19.40 |
|  | Median: 12.90 | Median: 76 | Median: 24.20 |
|  | 3rd Quartile: 15.25 | 3rd Quartile: 80 | 3rd Quartile: 37.30 |
|  | Maximum:20.60 | Maximum: 87 | Maximum: 77.00 |
|  | \*Note: Mean removed from output to maintain 5 summary number requirement | | |

Using the summary function in R will present the five summary numbers along with the mean. With these numbers we have an overview of what each variable looks like. Some potential areas of study we can draw from here is the potential for an outlier in the Volume variable as the third quantile is 37.30, but the max is 77.

Next we will graphically show straight line regressions for both Height vs Girth and Height vs Volume. Both graphs also include a scatter plot to better show the relationship between the points and the regression.

trees%>%  
 ggplot(aes(Girth,Height))+  
 geom\_point()+  
 geom\_smooth(method = "lm", formula = y~x, se=FALSE)+  
 labs( title = "Trees Data Set: ",  
 subtitle = "Height Vs Girth")+   
 xlab(label= "Height (ft)")+  
 ylab(label = "Girth (IN)")+  
 theme\_tufte()

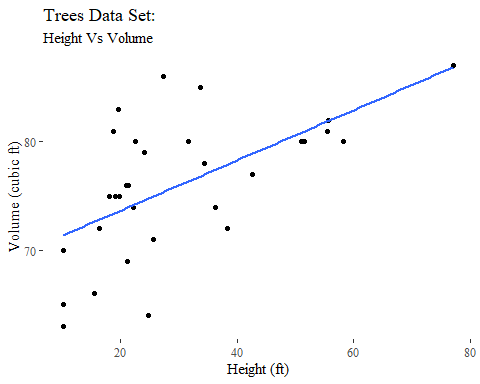
Figure



As shown above, the relationship between the two variables is positive which can be interrupted to mean that as the height of a tree increases the girth (diameter) will similarly increase. This is a pretty straightforward thought. If a tree is extremely tall then it will require a larger diameter to support the full weight. Note, the diameter is measured at the point 4ft 6in above ground level [1]. Below is the same graphic with a straight-line regression for Height and Volume:

trees%>%  
 ggplot(aes(Volume,Height))+  
 geom\_point()+  
 geom\_smooth(method = "lm", formula = y~x, se=FALSE)+  
 labs( title = "Trees Data Set: ",  
 subtitle = "Height Vs Volume")+   
 xlab(label= "Height (ft)")+  
 ylab(label = "Volume (cubic ft)")+  
 theme\_tufte()

Figure

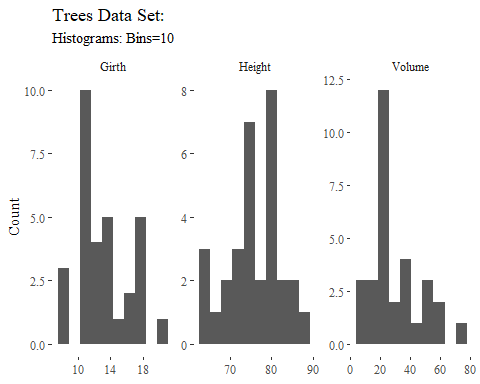


A similar relationship is shown, as all three variables are dependent on each other.

Next we look at each variable individually through both histograms and density plots. In a histogram the variable is placed into bins by its value and is then graphically shown by putting the count of each bin over the values it contains.

#histogram  
trees%>%gather()%>%  
 ggplot(aes(value))+  
 geom\_histogram(bins=10)+  
 facet\_wrap(~key, scales='free')+  
 labs( title = "Trees Data Set: ",  
 subtitle = "Histograms: Bins=10")+   
 xlab(label= NULL)+  
 ylab(label = "Count")+  
 theme\_tufte()

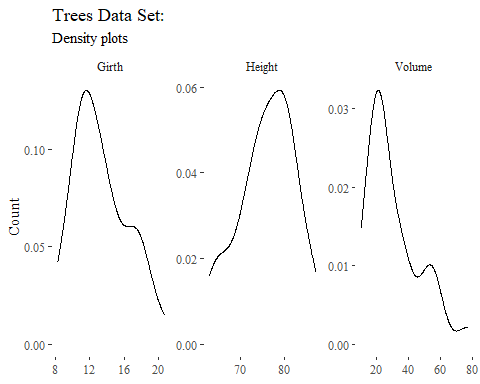
Figure



Above, we have histograms for all three variables. Looking at Girth, we find that a larger quantity of values is centered around 10-15 inches, with another smaller peak around 15-18 inches. Height has only one peak around 75 to 80 ft, which may be interpreted that as black cherry trees grow, they tend to stop growing higher in that range, but can still grow girthier. The volume histogram has a peak around 20-25 cubic ft, but has a long tail out to 80 cubic ft. Below are density plots, which is another way to show the same information.

#density Plot  
trees%>%gather()%>%  
 ggplot(aes(value))+  
 geom\_density()+  
 facet\_wrap(~key, scales='free')+  
 labs( title = "Trees Data Set: ",  
 subtitle = "Density plots")+   
 xlab(label= NULL)+  
 ylab(label = "Count")+  
 theme\_tufte()

Figure

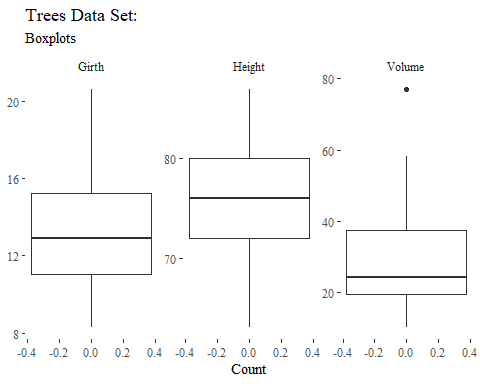


In a density plot it is easier to see the tails of the data. This is useful as we can say the volume is positively skewed along with the girth variable. However, the height variable would likely be categorized as negatively skewed.

The next set of graphs are boxplots, which are useful to graphically show the 5 summary statistics.

#boxplots  
trees%>%gather()%>%  
 ggplot(aes(value))+  
 geom\_boxplot()+coord\_flip()+  
 facet\_wrap(~key, scales='free')+  
 labs( title = "Trees Data Set: ",  
 subtitle = "Boxplots")+   
 xlab(label= NULL)+  
 ylab(label = "Count")+  
 theme\_tufte()

Figure

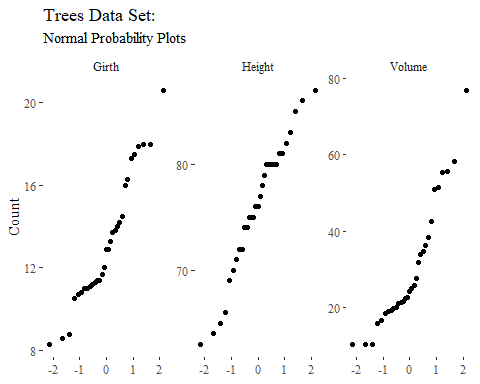


Looking at the Volume variable, we can see the single outlier mentioned earlier in the report. We can also see the positive skewness here with most of the graph/image only including half the data points. In the girth graph we can see some smaller positive skewness for the same reason. The Height graph also has some negative skewness here. Each of the horizontal lines correspond to the five summary values shown in table 1.

Next are the normal probability plots:

trees%>%gather()%>%  
 ggplot(aes(sample=value))+  
 geom\_qq()+  
 facet\_wrap(~key, scales='free')+  
 labs( title = "Trees Data Set: ",  
 subtitle = "Normal Probability Plots")+   
 xlab(label= NULL)+  
 ylab(label = "Count")+  
 theme\_tufte()

Figure



Analysis: Part B

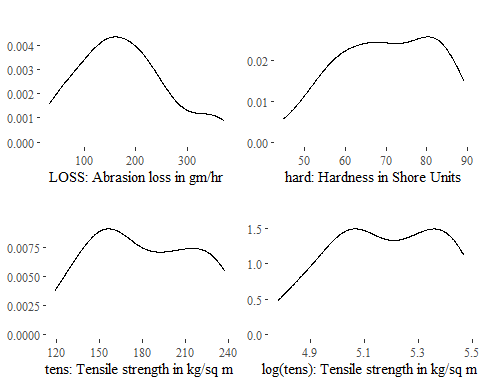
This portion of the paper looks at two datasets “Rubber” and “oddbooks.” Using this code:

#Load Relevant Libraries and Datasets  
library(MASS)  
library(DAAG)

library(ggcorrplot)  
data("Rubber")  
data("oddbooks")

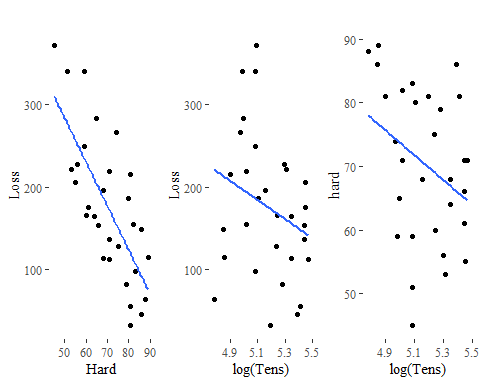
the requisite libraries and datasets are loaded. Although not required a set of histograms and scatter plots with linear models was included for each dataset to give a visual of what the datasets contain before performing regressions. The code can be found in the supporting documentation. Looking first at the “Rubber” dataset we see it has three variables: loss, hard, and tens with the following histograms in figure 7:

Figure



As seen in figure 7, a fourth variable has been included: log(tens). In the model further in the report, the decision to log tensile was made due to the longer tail it has. Below we have included graphic depictions of linear models for each of the possible variable pairs.

Figure



This graphic can be useful as it shows the general relationship between each variables. Next, we look at creating a multiple regression model. The model chosen uses loss (Abrasion loss in gm/hr) as the dependent variable and hardness and the log of tensile as the independent variables.

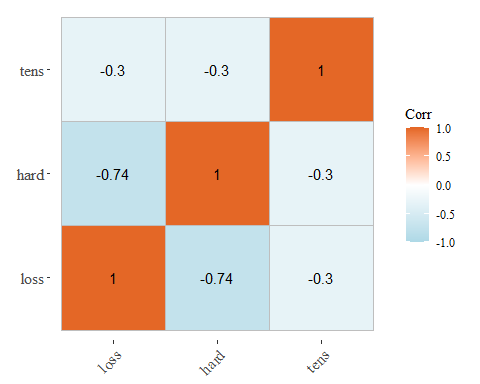
#Model Creation  
RubberModel<- lm(loss~hard + log(tens), data=Rubber)  
summary(RubberModel)

##   
## Call:  
## lm(formula = loss ~ hard + log(tens), data = Rubber)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -89.746 -14.567 2.037 18.552 66.132   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1919.9258 196.7864 9.756 2.40e-10 \*\*\*  
## hard -6.7153 0.5878 -11.424 7.58e-12 \*\*\*  
## log(tens) -245.9099 34.6141 -7.104 1.22e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.38 on 27 degrees of freedom  
## Multiple R-squared: 0.8412, Adjusted R-squared: 0.8294   
## F-statistic: 71.49 on 2 and 27 DF, p-value: 1.634e-11

Looking at the output, the model accounts for 84% (R-squared) of the variation in loss, with each of the independent variables having statistically significant (alpha=0.001) affects (large t-values). Next, a correlation plot is shown to show the interactions between the variables:

ggcorrplot(round(cor(Rubber),2), lab=TRUE, ggtheme=theme\_tufte(),colors = c("Light Blue", "white", "#E46726"))

Figure

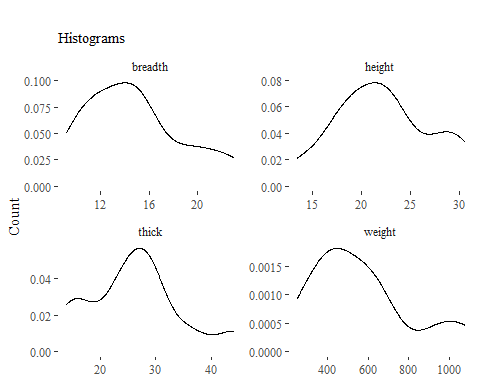


Here it is shown that Hardness and Abrasion loss are highly correlated at -0.74, whereas the other variable interactions both have a correlation value of -0.3. Overall it is clear that Hardness has a larger effect on Abrasion loss compared to tensile strength.

Next is the same analysis repeated utilizing the “oddbooks” dataset which has four variables Weight, Breadth, Height, and Thickness which is seen in the following histogram visualizations. Included is one visualization with just the variables, while the second visualization shows what happens if you log each variable.

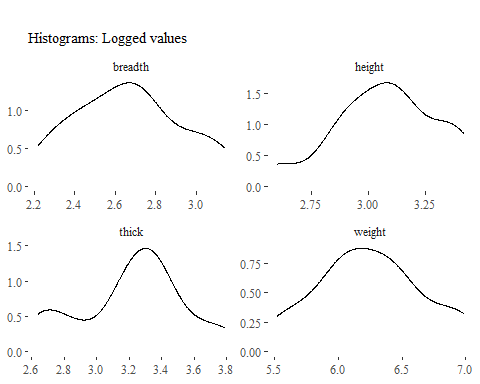
#Oddbooks EDA  
oddbooks%>%gather()%>%  
 ggplot(aes(x=value))+  
 geom\_density()+  
 facet\_wrap(~key, scales='free')+  
 labs( title = " ",  
 subtitle = "Histograms")+   
 xlab(label= NULL)+  
 ylab(label = "Count")+  
 theme\_tufte()

Figure



oddbooks%>%gather()%>%  
 ggplot(aes(x=log(value)))+  
 geom\_density()+  
 facet\_wrap(~key, scales='free')+  
 labs( title = " ",  
 subtitle = "Histograms: Logged values")+   
 xlab(label= NULL)+  
 ylab(label = NULL)+  
 theme\_tufte()

Figure



Comparing the two visualizations one might conclude to log the weight and breadth variables as the outcome looks to be more normally distributed. For the model creation the weight variable was chosen as the dependent variable to see how the Height, Breadth, and Thickness effect the weight of a book.

# Model Creation  
oddbooksModel<- lm(oddbooks$weight~ oddbooks$height+oddbooks$breadth+oddbooks$thick)  
summary(oddbooksModel)

##   
## Call:  
## lm(formula = oddbooks$weight ~ oddbooks$height + oddbooks$breadth +   
## oddbooks$thick)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -156.60 -12.22 22.54 49.87 71.56   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -441.651 720.770 -0.613 0.5571   
## oddbooks$height -21.947 45.898 -0.478 0.6453   
## oddbooks$breadth 91.666 43.534 2.106 0.0683 .  
## oddbooks$thick 4.961 10.308 0.481 0.6432   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 87.47 on 8 degrees of freedom  
## Multiple R-squared: 0.9089, Adjusted R-squared: 0.8747   
## F-statistic: 26.59 on 3 and 8 DF, p-value: 0.0001635

Using this formula the outcome has no significant variables at the 95% confident level. This is most likely due to all the variables being highly correlated with each other as shown in Figure 12, the correlation plot. Trying different variations, the final model chosen used the log of weight, thickness, and the log of breadth.

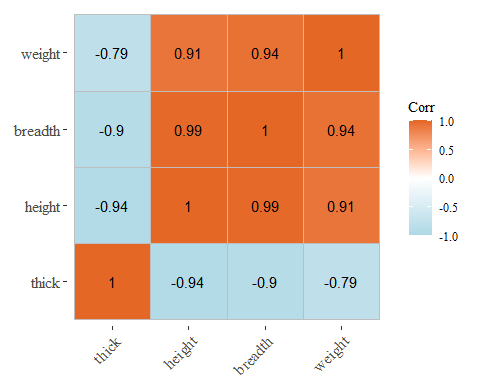
oddbooksModel<- lm(log(oddbooks$weight)~oddbooks$thick+log(oddbooks$breadth))  
summary(oddbooksModel)

##   
## Call:  
## lm(formula = log(oddbooks$weight) ~ oddbooks$thick + log(oddbooks$breadth))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.31947 -0.03533 0.05037 0.09131 0.13883   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.89855 1.47930 0.607 0.5586   
## oddbooks$thick 0.01399 0.01379 1.015 0.3368   
## log(oddbooks$breadth) 1.87066 0.42834 4.367 0.0018 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1541 on 9 degrees of freedom  
## Multiple R-squared: 0.8949, Adjusted R-squared: 0.8715   
## F-statistic: 38.3 on 2 and 9 DF, p-value: 3.963e-05

This model finds the log of breadth to have a statistically significant effect at the 99% confidence interval. Interpreting the value we can say that a 10% increase in breadth will have approximately a 20% increase in the weight. Next we look at the correlation values between all the variables.

#Correlation plot  
ggcorrplot(round(cor(oddbooks),2), lab=TRUE, ggtheme=theme\_tufte(),colors = c("Light Blue", "white", "#E46726"))

Figure



As is seen here, all the values are highly correlated, which is the likely the reasoning behind the low t-values of the model. By having highly correlated predictors it is hard to determine exactly how much effect each variable has on the dependent variable.

References

[1] (Trees) Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) The Minitab Student Handbook. Duxbury Press.

[2] Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition. Springer.