Rebecca Leu

Northeastern university

Descriptive statistics and Regression Analysis

The goal of the project is to demonstrate how to describe data numerically and graphically using R. In part B I will also be using multiple regression to predict influential variables.

Blue text is code from R

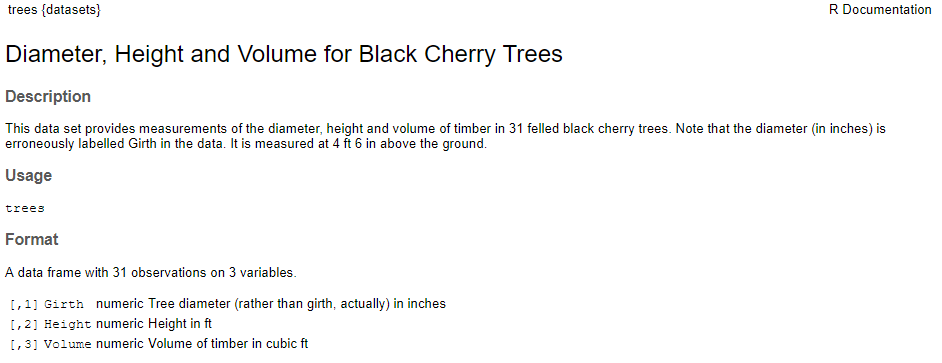
Yellow Highlight is numerical output from R

PART A

1. Describing data numerically

To begin I used a help command to read information about this dataset in the R help window.

> ?trees #HELP, INFORMATION ABOUT THIS DATASET



Although the help information tells us what the dataset holds I have listed out all of the data below.

> trees #LISTS FULL DATASET

Girth Height Volume

1 8.3 70 10.3

2 8.6 65 10.3

3 8.8 63 10.2

4 10.5 72 16.4

5 10.7 81 18.8

6 10.8 83 19.7

7 11.0 66 15.6

8 11.0 75 18.2

9 11.1 80 22.6

10 11.2 75 19.9

11 11.3 79 24.2

12 11.4 76 21.0

13 11.4 76 21.4

14 11.7 69 21.3

15 12.0 75 19.1

16 12.9 74 22.2

17 12.9 85 33.8

18 13.3 86 27.4

19 13.7 71 25.7

20 13.8 64 24.9

21 14.0 78 34.5

22 14.2 80 31.7

23 14.5 74 36.3

24 16.0 72 38.3

25 16.3 77 42.6

26 17.3 81 55.4

27 17.5 82 55.7

28 17.9 80 58.3

29 18.0 80 51.5

30 18.0 80 51.0

31 20.6 87 77.0

Next I listed out the full summary of each variable. This gives us the minimum, 1st quartile, median, mean, 3rd quartile and max for each of the 3 variables.

> data(trees)#LOAD DATA INTO WORKSPACE

> summary(trees) #FULL SUMMARY OF EACH VARIABLE

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

Tukeys 5 number summary gives the summary numbers without labels.

> fivenum(trees$Girth)#TUKEY 5 NUM SUMMARY FOR EACH VARIABLE - NO LABLES

[1] 8.30 11.05 12.90 15.25 20.60

> fivenum(trees$Height)

[1] 63 72 76 80 87

> fivenum(trees$Volume)

[1] 10.2 19.4 24.2 37.3 77.

> boxplot.stats(trees$Girth)#GIVES SAME NUM AS FIVENUM W/NUMBER OF OBSERVATIONS

$stats

[1] 8.30 11.05 12.90 15.25 20.60

$n

[1] 31

> #CONFIDENCE INTERVALS FOR THE MEDIAN, ALSO LISTS OUTLIERS IF ANY

$conf

[1] 11.70814 14.09186

$out

numeric(0)

1. Describing data graphically

Below are the regression plots showing correlation between the different variables. These plots give an easy to read representation of how each variable interacts with another. Height volume and girth of the trees all increase as the other increases, just as we would assume for these trees. The taller the tree, the bigger the girth, the heavier the volume.

>

> #HEIGHT AND GIRTH REGRESSION PLOT RED

> par(mfrow=c(2,3))

> plot(Height, Girth, pch=16, col="red")

> abline(lm(Girth~Height))

>

> #HEIGHT AND VOLUME REGRESSION PLOT BLUE

> plot(Height, Volume, pch=16, col="blue")

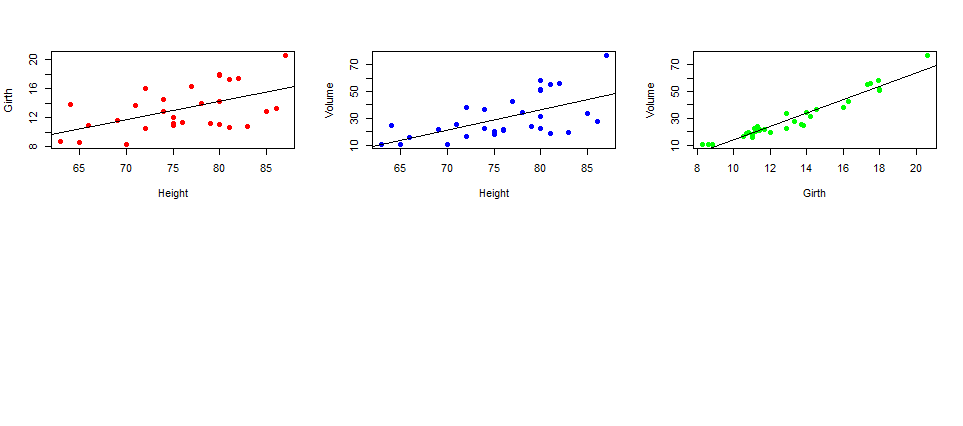
> abline(lm(Volume~Height))

>

> #GIRTH AND VOLUME REGRESSION PLOT GREEN

> plot(Girth, Volume, pch=16, col="green")

> abline(lm(Volume~Girth))

> 

Histograms give us a visual representation of the range of our dataset. These histograms tell us how frequently each of the values are measured. The histogram of height has almost a normal probability curve while the volume histogram has a negative slope.

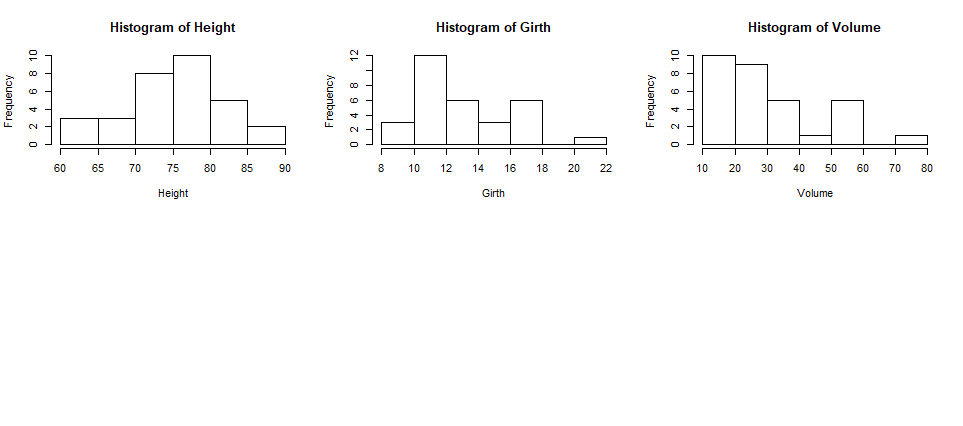
> #HISTOGRAMS

> par(mfrow=c(2,3))

> hist(Height)

> hist(Girth)

> hist(Volume)



Density plots give us a similar visual representation as a histogram does. As we can see the density plot correlates nicely with the histogram.

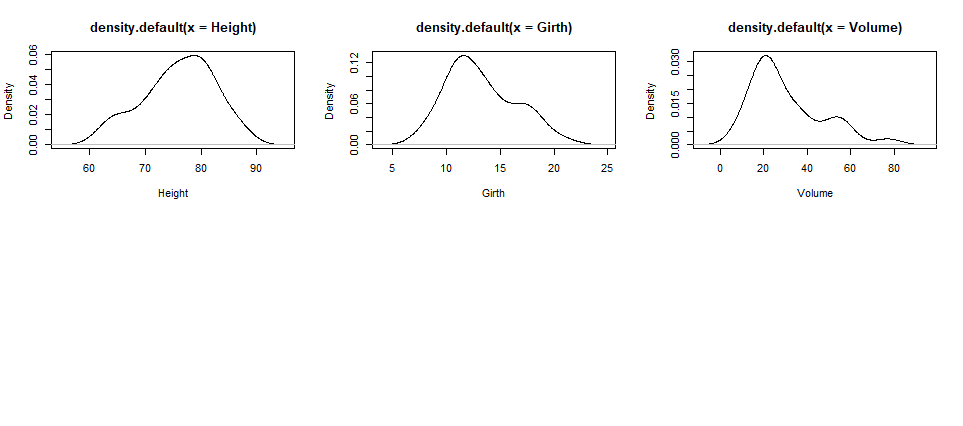
>

> #DENSITY PLOTS

> par(mfrow=c(2,3))

> plot(density(Height), xlab="Height")

> plot(density(Girth), xlab="Girth")

> plot(density(Volume), xlab="Volume")  


Boxplots give us a visual representation of the summary information numerically given before.

>

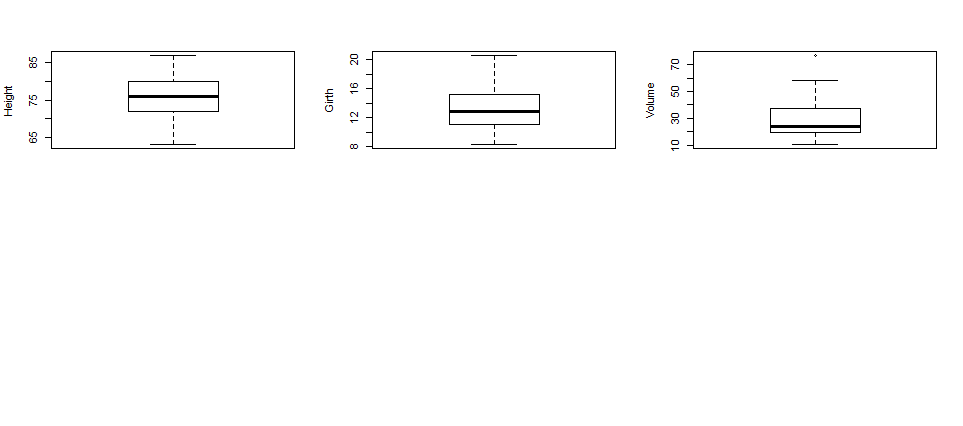
> #BOXPLOTS

> par(mfrow=c(2,3))

> boxplot(Height, ylab="Height")

> boxplot(Girth, ylab="Girth")

> boxplot(Volume, ylab="Volume")

> 

Normal probability pots help us see outliers in the data. All the dots should follow a pretty straight line. If we see a dot that is far off from the normal probability line, it may be skewing are data one way or another. It appears that all of the data in the trees dataset is pretty normal and there does not appear to be any outliers that could skew the data.

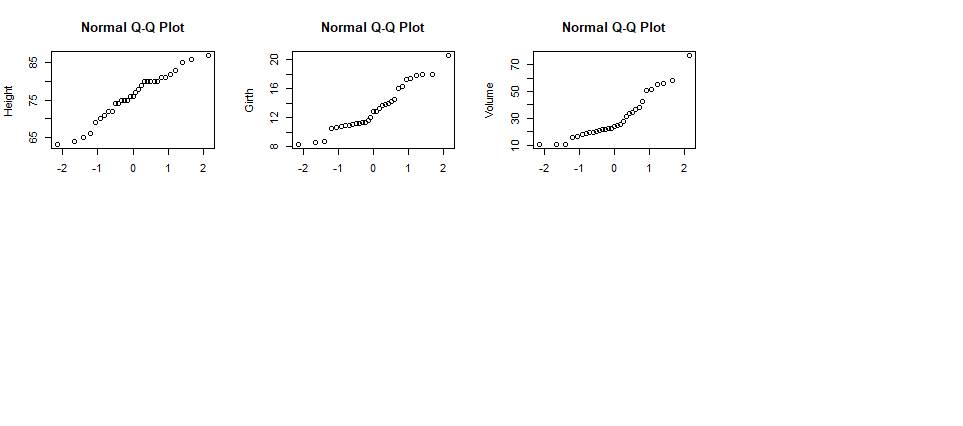
> #NORMAL POROBABILITY PLOTS

> par(mfrow=c(2,4)) #COMBINES PLOTS ON TO ONE PAGE

> qqnorm(Height, xlab="", ylab="Height") #LABLES Y AXIS

> qqnorm(Girth, xlab="", ylab="Girth")

> qqnorm(Volume, xlab="", ylab="Volume")



From part A we have found that listing out all the data numerically does not easily show us patterns in the data. The numeric summary of the data was slightly more helpful, but overall, most people are visual. Seeing the data visually in graphs and plots helps the reader see trends in the data as well as correlations and potential outliers that could be skewing data.

PART B

In part B we will be looking into two other datasets and using multiple regression to predict influential variables. The first dataset is in the MASS package. We will be using the Rubber dataset. Again there are 3 variables and we are going to see how each of these correlates with each other and what variables are more influential than the others.

|  |
| --- |
| 1. “Rubber” Regression Model   > install.packages("MASS")  package ‘MASS’ successfully unpacked and MD5 sums checked  The downloaded binary packages are in  C:\Users\rebec\AppData\Local\Temp\RtmpQf8Vpd\downloaded\_packages  > install.packages("ggplot2")  package ‘ggplot2’ successfully unpacked and MD5 sums checked  The downloaded binary packages are in  C:\Users\rebec\AppData\Local\Temp\RtmpQf8Vpd\downloaded\_packages  > library(MASS)  > Rubber  loss hard tens  1 372 45 162  2 206 55 233  3 175 61 232  4 154 66 231  5 136 71 231  6 112 71 237  7 55 81 224  8 45 86 219  9 221 53 203  10 166 60 189  11 164 64 210  12 113 68 210  13 82 79 196  14 32 81 180  15 228 56 200  16 196 68 173  17 128 75 188  18 97 83 161  19 64 88 119  20 249 59 161  21 219 71 151  22 186 80 165  23 155 82 151  24 114 89 128  25 341 51 161  26 340 59 146  27 283 65 148  28 267 74 144  29 215 81 134  30 148 86 127  > data(Rubber)  > pairs(Rubber)  The below model gives us a visual representation of each variable plotted against each other in pairs. From this visual I would conclude that loss and hard have a negative correlation but loss and tens as well as hard and tens do not appear to have much of a correlation.    >  > #LABLE REGRESSION MODEL, LOSS IS OUTCOME VARIABLE, PREDICTORS ARE HARD AND TENS  > #, DATA=RUBBER TO NOTE ALL THESE VARIABLES ARE FROM THE RUBBER DATASET  > #NO NEED TO PUT RUBBER$ BEFORE EACH VARIABLE  > Rubber.lm <- lm(loss~hard+tens, data=Rubber)  > options(digits=3)  > #TO SEE RESULTS MUST REUN SUMMARY OF REGRESSION MODEL CREATED  > summary(Rubber.lm)  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  After seeing some correlations in our initial plots, a regression model was built with loss as the outcome variable and the predictors being hard and tens. Additionally, a color-coded correlation plot was produced that confirms what we saw in the previous plots. Tensile strength has almost no correlation with hard and loss.  >  > install.packages("ggcorrplot")  package ‘ggcorrplot’ successfully unpacked and MD5 sums checked  The downloaded binary packages are in  C:\Users\rebec\AppData\Local\Temp\RtmpQf8Vpd\downloaded\_packages  > library(ggcorrplot)  >  > data(Rubber)  > corr<-round(cor(Rubber),1)  > head(corr[, 1:3])  loss hard tens  loss 1.0 -0.7 -0.3  hard -0.7 1.0 -0.3  tens -0.3 -0.3 1.0  > p.mat<-cor\_pmat(Rubber)  > head(p.mat[,1:3])  loss hard tens  loss 0.00e+00 3.29e-06 0.109  hard 3.29e-06 0.00e+00 0.108  tens 1.09e-01 1.08e-01 0.000  > ggcorrplot(corr)     1. “Oddbooks” Regression Model   > install.packages("DAAG")  package ‘DAAG’ successfully unpacked and MD5 sums checked  The downloaded binary packages are in  C:\Users\rebec\AppData\Local\Temp\RtmpQf8Vpd\downloaded\_packages  >  > library(DAAG)  > #LIST ODDBOOKS DATASET  > oddbooks  thick height breadth weight  1 14 30.5 23.0 1075  2 15 29.1 20.5 940  3 18 27.5 18.5 625  4 23 23.2 15.2 400  5 24 21.6 14.0 550  6 25 23.5 15.5 600  7 28 19.7 12.6 450  8 28 19.8 12.6 450  9 29 17.3 10.5 300  10 30 22.8 15.4 690  11 36 17.8 11.0 400  12 44 13.5 9.2 250  > #LABLE THE LOG ON THE ODDBOOKS DATASET LOGBOOKS  > logbooks <-log(oddbooks)  > #LABLE THE FIRST REGRESSION MODEL, FIRST MODEL HAS WEIGHT AS THE OUTCOME VARIABLE  > #ADDING ONLY THICKNESS AS A PREDICTER  > logbooks.lm1<-(lm(weight~thick,data=logbooks))  > #TO SEE RESULTS MUST RUN SUMMARY OF REGRESSION MODEL CREATED  > summary(logbooks.lm1)$coef  Using the lm() function we are completing a stepwise regression starting with only thickness as a predictor.  Estimate Std. Error t value Pr(>|t|)  (Intercept) 9.69 0.708 13.7 8.35e-08  thick -1.07 0.219 -4.9 6.26e-04  >  Our second regression includes thickness and height as predictors. In this model it appears that the height is a somewhat significant predictor.  > logbooks.lm2<-lm(weight~thick+height,data=logbooks)  > summary(logbooks.lm2)$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  >  Finally, we add breadth into our model. This shows that breadth is our most significant predictor in this model, but mostly this data tells us that thickness, height, and breadth are all very similar information and as we use more than one variable the coefficients are less significant.  > 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  >  > install.packages("ggcorrplot")  Error in install.packages : Updating loaded packages  > library(ggcorrplot)  >  > data(oddbooks)  > corr<-round(cor(oddbooks),1)  > head(corr[, 1:4])  thick height breadth weight  thick 1.0 -0.9 -0.9 -0.8  height -0.9 1.0 1.0 0.9  breadth -0.9 1.0 1.0 0.9  weight -0.8 0.9 0.9 1.0  > p.mat<-cor\_pmat(oddbooks)  > head(p.mat[,1:4])  thick height breadth weight  thick 0.00e+00 5.90e-06 7.28e-05 2.24e-03  height 5.90e-06 0.00e+00 4.26e-09 4.43e-05  breadth 7.28e-05 4.26e-09 0.00e+00 4.28e-06  weight 2.24e-03 4.43e-05 4.28e-06 0.00e+00  > ggcorrplot(corr) |
| The gg correlation plot that was used with the Rubber dataset it used again. This again shows that all the variables correlate with each other strongly whether they be a negative or positive correlation. In this model, each variable can be a predictor of the other. |
|  |

References

Kabacoff, R. (n.d.). R Tutorial. Retrieved April 12, 2020, from https://www.statmethods.net/r-tutorial/index.html

Maindonald, J. H. (n.d.). Using R for Data Analysis and Graphics. doi: https://cran.r-project.org/doc/contrib/usingR.pdf