MATH 4753 Laboratory 3
Introduction to R and Regression

In this lab you will be introduced to simple linear regression (SLR) in R. We will do this in order to give you the basic tools needed to start your project. The theory will be developed in class (Wed, Fri) later in the semester, however we will be able to conduct an analysis and interpret the output (though not completely) after some rudimentary instruction. After 1 or 2 labs you should be able to conduct the basics of SLR.

Objectives

In this lab you will learn how to:

1. Create scatter plots.
2. Define a linear model in R.
3. Plot a least squares regression line.
4. Find the residuals and plot them.
5. Interpret regression summary output.
6. Make predictions.
7. Use a "shiny" server to make interactive plots

Tasks

- Use Rmarkdown and make clear and readable files. Do this:
    - Work on the Lab 3.R script file first and get all the correct code working first before using R markdown – only paste into Rmd after you have completed the lab.
    - Once you are happy with the final Rmd documents knit tasks 1-6 into html – then wait!
    - Use the separate rmd file for Task 7 – run in RStudio. There are three settings for the widget – make a screen print for each – include these in your Lab 3 rmd document under Task 7 – you can use { width=70% } where png is the name of your picture file.

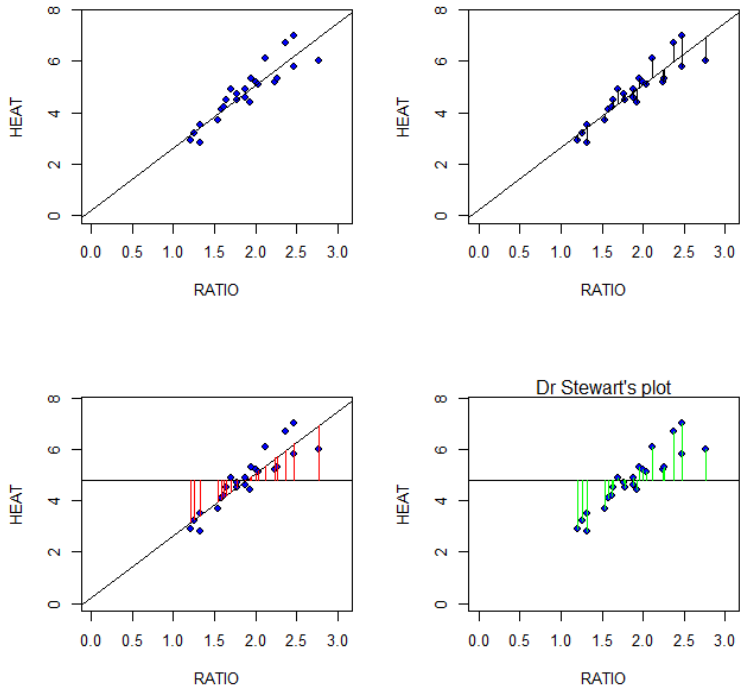
- Task 1
    - Download from CANVAS the zipped data files, "Dataxls"
    - Unzip the contents into a directory on your desktop (call it LAB3)
    - Download the file "lab3.r"
    - Place this file with the others in LAB3.
    - Start Rstudio
    - Open "lab3.r" from within Rstudio.
    - Go to the "session" menu within Rstudio and "set working directory" to where the source files are located.
    - Copy and paste the working directory by issuing the command `getwd()`:

- Task 2
    - Find the file "SPRUCE.xls" inside LAB3
    - Open it in Excel
    - Save As type CSV(comma delimited) "*.csv"

- Use `read.table()` to read the data into R (*or any other method available*), this function will already be available within the script lab3.r which you have opened in Rstudio.
- Obtain the first six lines of the data using "`head()`"
- Make a new file for your code in RStudio editor, call it "mylab3.R" and place in it all the code you need to answer the tasks of this lab (copy and paste from lab3.R).

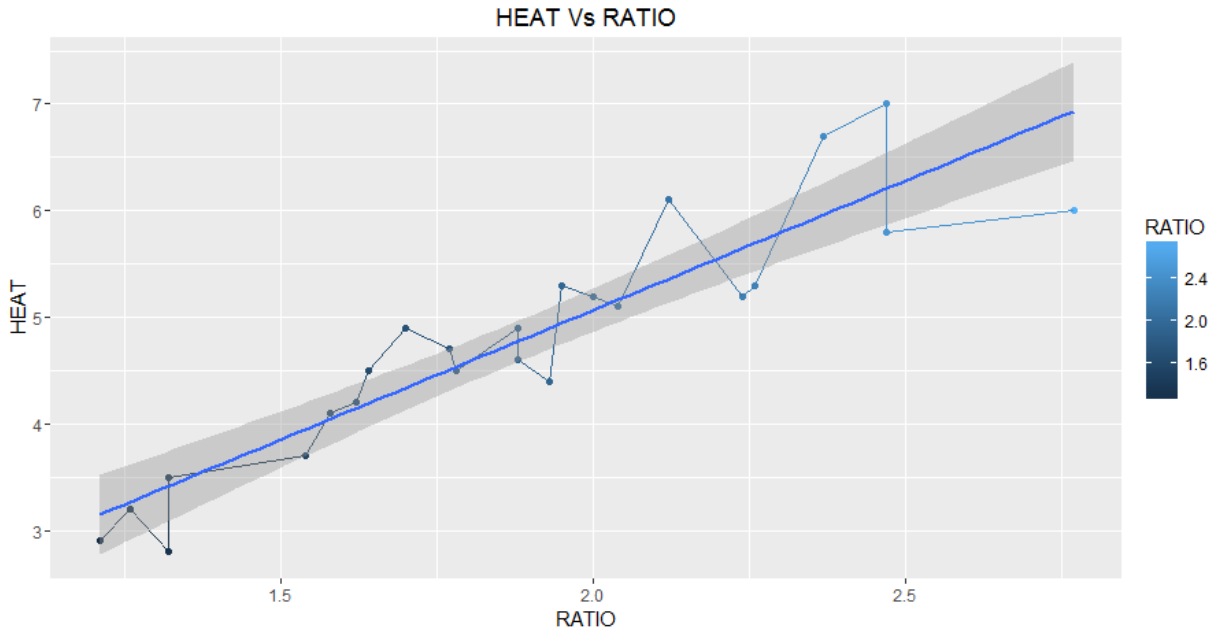
- Task 3
  - The SPRUCE data set is described in MS 10.52, pages 478 and 479. This data set has two variables, Height = Height of Spruce trees in m (this is what we want to predict) and BHDiameter = Breast height Diameter in cm. The idea is that breast height diameter is an easy measurement to make whereas the height of the trees is much more difficult. We want to see if there is a relationship between the two variables that enables us to predict Height from Diameter.
  - Make a scatter plot of the data (y axis will be the Height).
    - Make sure it has a heading
    - Y axis is labelled
    - X axis is labelled
    - Points are filled blue, use pch=21 and bg="Blue"
    - The points are 1.2Xs the default character size (use cex=1.2)
    - The y and x axes are adjusted to include 0 and **`1.1*max(y), and 0`** and **`1.1*max(x)`** respectively
  - Does there appear to be a straight line relationship?
  - Load the library s20x and make a lowess smoother scatter plot using **`trendscatter()`** (try a few values of f, f=0.5,0.6,0.7) record all three plots, use **`layout().`**
  - We will *assume* (this may in fact be a bad assumption) a straight line relationship, use lm and make a linear model object, call it **`spruce.lm`**
  - Make a new scatter plot and add the least squares regression line to the points - **`abline(spruce.lm) –`** record the plot.
  - Comment on the graph, is a straight line appropriate? Consider the smoother curve also.
- Task 4
  - Divide the graphical interface into 4 equal areas, use **`layout.show(4)`** and record the picture.
    - In the first square, plot the scatter plot and fitted line.
    - In the second square plot the same with the residual line segments (deviations about the fitted line). (RSS=residual sum of squares)
    - In the third square plot the mean of Y versus X i.e. mean of Height vs BHDiameter, with the fitted line and deviations of the fitted line from the mean height added. (MSS=model sum of squares)
    - In the fourth square plot the mean of Height versus BHDiameter and show the total deviation line segments $\hat{y} = \bar{y}$. (TSS=total sum of squares)

In brief reproduce the plot below for the SPRUCE data set.



- Calculate TSS, MSS and RSS
- Calculate $\frac{MSS}{TSS}$, and interpret it!
- Does TSS=MSS+RSS?

- Task 5
  - Summarize `spruce.lm` paste it here.
  - What is the value of the slope?
  - What is the value of the intercept?
  - Write down the equation of the fitted line.
  - Predict the Height of spruce when the Diameter is 15, 18 and 20cm (use `predict()`)
- Task 6
  - Use appropriate code using the ggplot2 package to make a plot of Height Vs Diameter with shading lines and legend as shown below:



- Task 7
  - Now that you have made the ggplot of Task 6 we will take our presentation one step further by making an interactive document.
  - Open the file Task 7.rmd in R markdown.
  - There are a number of controls that will create input for your dynamic plot.
  - Using the example given in Task 7.rmd create a shiny interactive document that has a selectInput() widget function with choices
    - One straight simple linear regression line that estimates the trend (see blue line above)
    - Points alone
    - Points joined with line segments
  - Run your document and for each selection of your control (widget) take a picture using PrtScr – save in your working directory.
  - Place in your rmd document using { width=70% }