MATH 4753 Laboratory 4
SLR assumptions

The last lab introduced you to SLR with a data set that had a non-linear trend. This meant that a straight line was an inappropriate choice for a model. However, this model was applied and some skills developed like plotting points, segments, adding the fitted line and determining estimates of parameters from summary output and interpreting multiple R^2. Today we will begin where the last lab left off and examine the assumptions of the linear model. If the assumptions hold we say that the analysis performed is valid.

Objectives

In this lab you will learn how to:

1. Create a linear model with x^2 and x variables.
2. Create residual plots for two models and be able to compare and interpret them.
3. Create QQ plots and interpret them.
4. Create and interpret the Shapiro Wilk test.
5. Interpret regression summary output (similar to last lab).
6. Make predictions for the new model.
7. Learn about piecemeal regression.
8. Learn how to make an R package using the package roxygen2

Tasks

Make an RMD document and then knit to HTML
Upload both to canvas

Note: All plots you are asked to make should be recorded in this document.

- Task 1
  - Download from CANVAS the zipped data files, "Dataxls"
  - Unzip the contents into a directory on your desktop (call it LAB4)
  - Download the file "lab4.r"
  - Place this file with the others in LAB4.
  - Start Rstudio
  - Open "lab4.r" from within Rstudio.
  - Go to the "session" menu within Rstudio and "set working directory" to where the source files are located.
  - Issue the function `getwd()`

- Task 2
  - Find the file "SPRUCE.xls" inside LAB4
  - Open it in Excel
  - Save As type CSV(comma delimited) "*.csv"

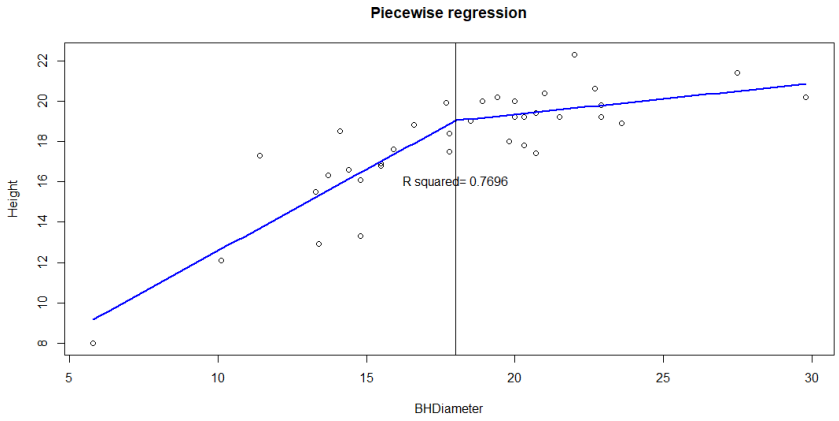
- Use `read.table(file.choose(), header=TRUE,sep=",")` to read the data into R (*or any other method available*), this function will already be available within the script lab4.r which you have opened in Rstudio.
- Copy and paste the last six lines of the data using "`tail()`":
- Make a new file for your code in RStudio editor, call it "mylab4.R" and place in it all the code you need to answer the tasks of this lab (copy and paste from lab4.R).
- Use the hash # symbol and write your own comments in the code file explaining what the code does.

- 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.
  - Load the library s20x and make a lowess smoother scatter plot (Height Vs BHDiameter) using **`trendscatter()`** (use f=0.5) record the plot.
  - Make a linear model object, **`spruce.lm=with(spruce.df,lm(Height~BHDiameter))`**
  - Find the residuals using **`residuals()`**, put them into an object called **`height.res`**
  - Find the fitted values using **`fitted()`** and place them in an object called **`height.fit`**.
  - Plot the residuals vs fitted values.
  - Plot the residuals vs fitted values using **`trendscatter()`**
  - What shape is seen in the plot? Compare it with the curve made with the trendscatter function (second line after Task3).
  - Using the plot() function and spruce.lm, make the residual plot.
  - Check normality using the s20x function **`normcheck()`**. Please note that you may need to add an additional option to show the Shapiro-Wilk test (use **`?normcheck`**)
  - What is the pvalue for the Shapiro-Wilk test? What is the NULL hypothesis in this case?
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_{i,}\ \epsilon_i \sim N(0, \sigma^2)$ describes the model used above. Notice that the residuals r_i estimate the model errors ϵ_i. If the model works well with the data we should expect that the residuals are approximately Normal in distribution with mean 0 and constant variance.
  - Write a sentence outlining your conclusions concerning the validity of applying the straight line to this data set.
- Task 4
  - Fit a quadratic to the points using the appropriate formula inside the lm() function and placing the output in the object **`quad.lm`**.
  - Make a fresh scatter plot of Height Vs BHDiameter and add the quadratic curve to it.
  - Make **`quad.fit`**, a vector of fitted values.
  - Make a plot of the residuals vs fitted values, use **`plot()`** and quad.lm
  - Construct a QQ plot using **`normcheck()`**
  - What is the value of the p-value in the Shapiro-Wilk test? What do you conclude?

- Task 5
  - Summarize `quad.lm` paste it here.
  - What is the value of $\widehat{\beta_0}$?

- What is the value of $\widehat{\beta_1}$
- What is the value of $\widehat{\beta_2}$
- Make interval estimates for $\beta_0, \beta_1, \beta_2$.
- Write down the equation of the fitted line.
- Predict the Height of spruce when the Diameter is 15, 18 and 20cm (use `predict()`)
- Compare with the previous predictions.
- What is the value of multiple R^2? Compare it with the previous model.
- Make use of adjusted R squared to compare models to determine which is "better". Use the web to learn about adjusted R squared.
- What does $(multiple\ R^2)$ mean in this case?
- Which model explains the most variability in the Height?
- Use anova() and compare the two models. Paste anova output here and give your conclusion underneath.
- Find TSS, record it here
- Find MSS, record it here
- Find RSS, record it here
- What is the value of MSS/TSS?

- Task 6
  - Investigate unusual points by making a cooks plot using cooks20x(). Place the plot here.
  - Use the web to find out what cooks distance is and how it is used – write a couple of sentences here.
  - What does cooks distance for the quadratic model and data tell you?
  - Make a new object called quad2.lm which is made from the same quadratic model using the data with the datum which has highest cooks distance removed.
  - Summarize the new object here.
  - Compare with the summary information from quad.lm
  - What do you conclude?
- Task 7



  - Prove using latex that $y = \beta_0 + \beta_1 x + \beta_2 (x - x_k) I(x > x_k)$ where I() is 1 when $x > x_k$ and 0 else.
  - Reproduce the above plot using the code included in the R script $(x_k = 18)$, you may need to change some of the parameter values.

- Task 8
  - Please install the following packages. You may use the single function:
  - install.packages(c("devtools", "roxygen2", "testthat", "knitr"))
  - Follow the demonstration here: https://youtu.be/DWkIbk_HE9o or that given by me.
  - Add one function to your package from the labs we have done so far (Labs 1-4). Make sure it is documented. This is VERY important – learn about how to do this with roxygen – see http://r-pkgs.had.co.nz/man.html#man
  - Every lab you will add a function to this package.
  - In an RMD chunk, load your package using library()
  - Use your function so that I can see the output.
  - Explain in a couple of sentences what the function does.

###################### LAB 4 comes to here – the rest is extra if you finish early ###############

Extra for experts: Produce the plot below (you will need, segments(), text(), arrows())