# ISE 3293/5013 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.

***Tasks***

All output made please copy and paste into **this word file**. Save and place in the dropbox when completed. Anything you are asked to make should be recorded under the question in this document. There will be two files you need to upload:

- a pdf of this document (pdf) or the word file (docx)
- a text file of all the code you used to create answers (txt)

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

- Task 1
  - Download from D2L 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()` and copy the output here.

```
"F:/Google Drive - Saied/Courses/02 OU/11 Fundamentals of Engi
neering Statistical Analysis/02 Labs/04 Lab 4"
```

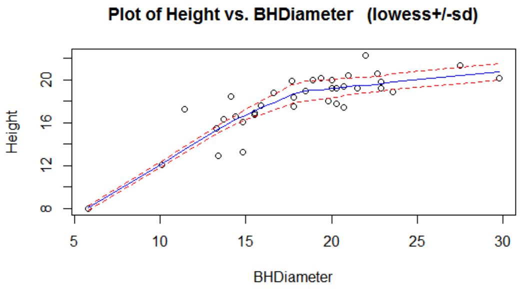
- Task 2
  - Find the file “SPRUCE.xls” inside LAB4

- 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()`” (use “`courier new`” font):

```
        BHDiameter Height
31         17.7   19.9
32         20.7   19.4
33         21.0   20.4
34         13.3   15.5
35         15.9   17.6
36         22.9   19.2
```

- 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))`**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.14684    1.12131   8.157 1.63e-09 ***
BHDiameter   0.48147    0.05967   8.069 2.09e-09 ***
```

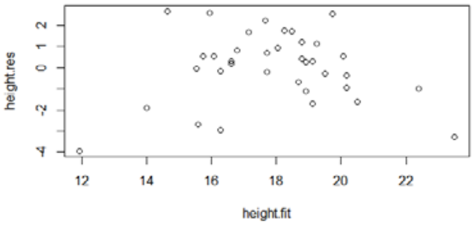
  - Find the residuals using **`residuals()`**, put them into an object called **`height.res`**

```
height.res = residuals(spruce.lm)
```

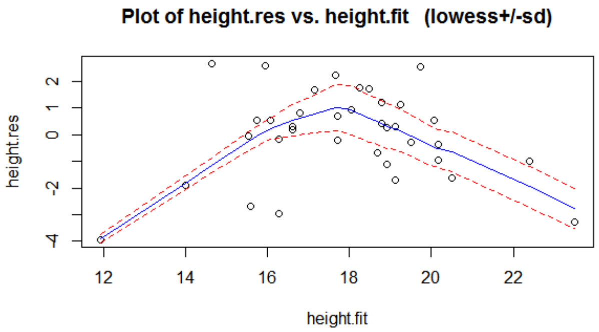
  - Find the fitted values using **`fitted()`** and place them in an object called **`height.fit`**.

```
height.fit = fitted(spruce.lm)
```

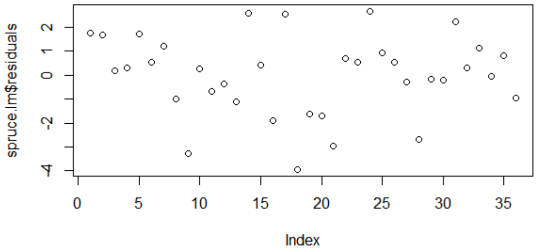
  - Plot the residuals vs fitted values.



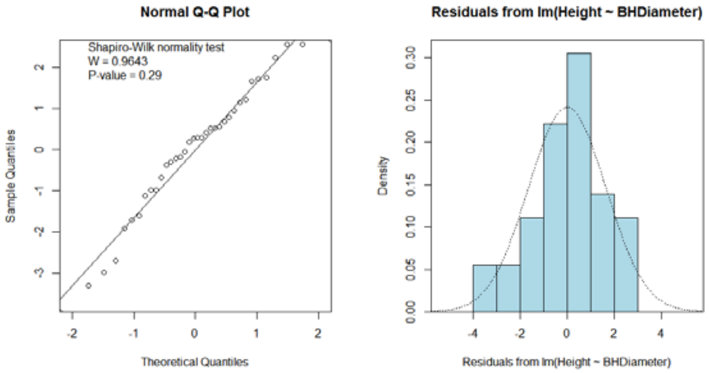
- Plot the residuals vs fitted values using **`trendscatter()`**



- 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?

```
Because of high value for p, we can not reject the null hypoth
esis. Therefore; the distribution is normal.
```

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_{i,}\ \epsilon_i \sim N(0,\sigma^2)$ describes the model used above. Notice that the residuals $r_i$ estimate the model errors $\epsilon_i$. 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.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.14684    1.12131   8.157 1.63e-09 ***
BHDiameter   0.48147    0.05967   8.069 2.09e-09 ***
```

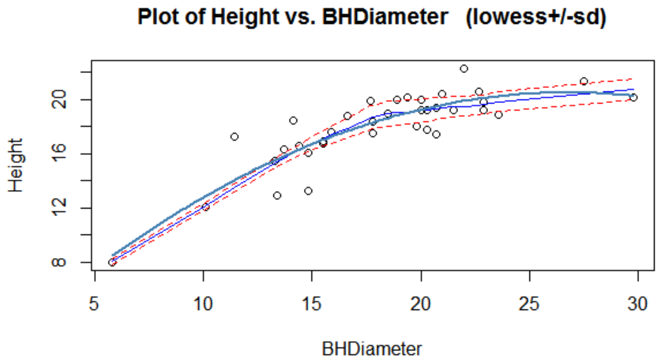
- Write a sentence outlining your conclusions concerning the validity of applying the straight line to this data set.

```
It says that there are not enough evidence to accept the null
hypothesis, therefore the estimated values are accepted. The m
odel is Height = 9.14684 + 0.48147 * Diameter.
```

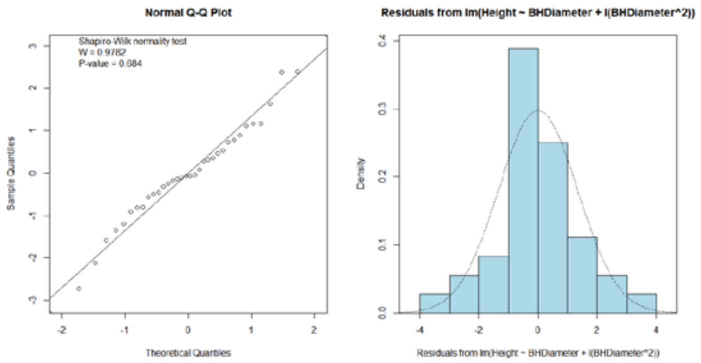
- Task 4
  - Fit a quadratic to the points using the appropriate formula inside the lm() function and placing the output in the object **`quad.lm`**.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.860896   2.205022   0.390 0.698731
BHDiameter       1.469592   0.243786   6.028 8.88e-07 ***
I(BHDiameter^2) -0.027457   0.006635  -4.138 0.000227 ***
```

  - 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()`**

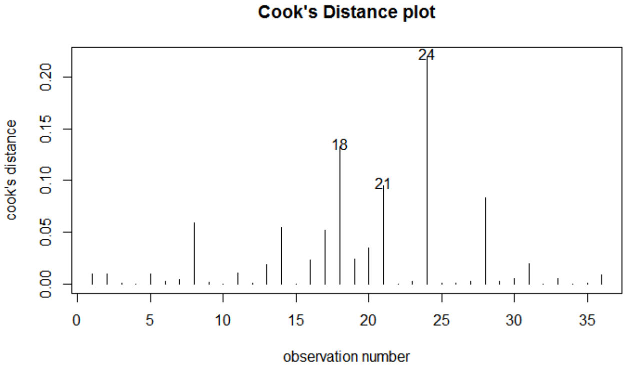


  - What is the value of the p-value in the Shapiro-Wilk test? What do you conclude?

- Task 5
  - Summarize `quad.lm` paste it here.
  - What is the value of $\widehat{\beta_0}$?
    ```
    0.860896
    ```
  - What is the value of $\widehat{\beta_1}$
    ```
    1.469592
    ```
  - What is the value of $\widehat{\beta_2}$
    ```
    -0.027457
    ```
  - Make interval estimates for $\beta_0, \beta_1, \beta_2$.
  - Write down the equation of the fitted line.
    ```
    Y = 0.86089580 +1.46959217*x  -0.02745726*x^2
    ```
  - Predict the Height of spruce when the Diameter is 15, 18 and 20cm (use `predict()`)
    ```
    16.72690 18.41740 19.26984
    ```
  - Compare with the previous predictions.
    ```
    16.36895 17.81338 18.77632
    ```
  - What is the value of multiple $R^2$? Compare it with the previous model.
    ```
    Multiple R-squared:  0.7741,
    ```
  - Make use of adjusted R squared to compare models to determine which is "better". Use the web to learn about adjusted R squared.
    ```
    It's used to compare two different models.
    ```
  - What does $(multiple\ R^2)$ mean in this case?
    ```
    How much the model fits data.
    ```
  - 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.
    ```
    Analysis of Variance Table

    Model 1: Height ~ BHDiameter
    Model 2: Height ~ BHDiameter + I(BHDiameter^2)
      Res.Df    RSS Df Sum of Sq      F    Pr(>F)
    1     34 95.703
    2     33 63.007  1    32.696 17.125 0.0002269 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    ```
  - Find TSS, record it here
    ```
    278.9475
    ```
  - Find MSS, record it here
    ```
    215.9407
    ```
  - Find RSS, record it here
    ```
    63.00683
    ```
  - What is the value of MSS/TSS?
    ```
    0.7741266
    ```

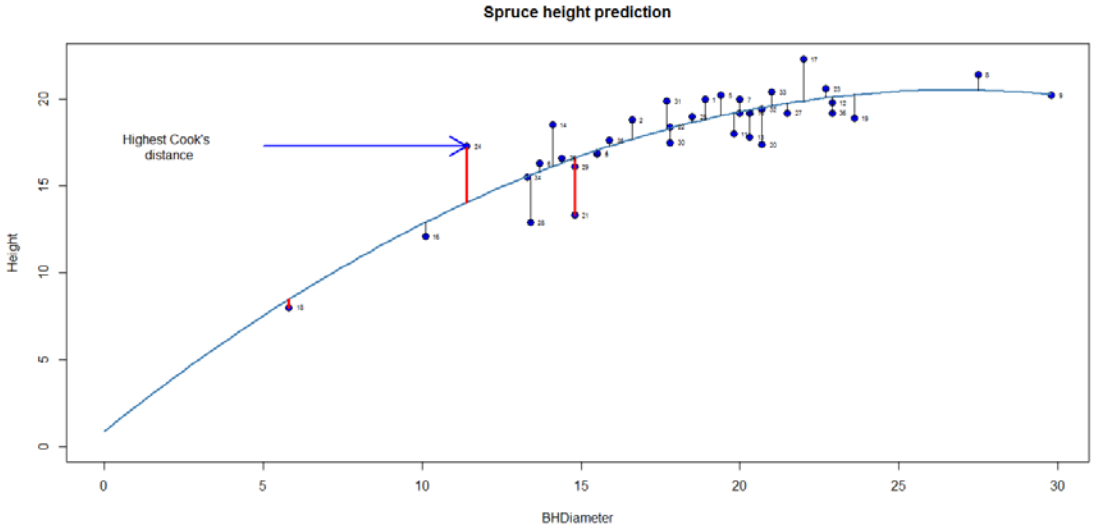
- Task 6
  - Investigate unusual points by making a cooks plot using cooks20x(). Place the plot here.



- 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?

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

**<u>Extra for experts: Produce the plot below (you will need, segments(), text(), arrows())</u>**


Spruce height prediction
Highest Cook's distance
Height
BHDiameter