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 . 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 and 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 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() and copy the output here.

"C:/Users/cglen/Documents/Stat Methods/Labs/LAB4"

* 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

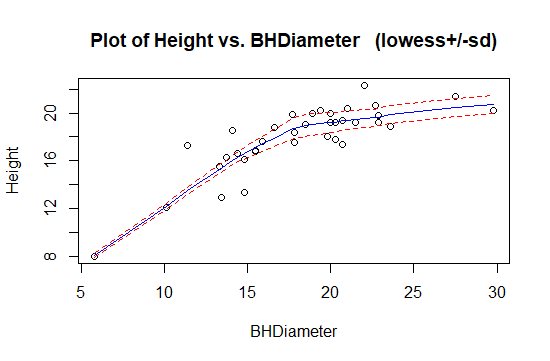
33 21.0 20.4

34 13.3 15.5

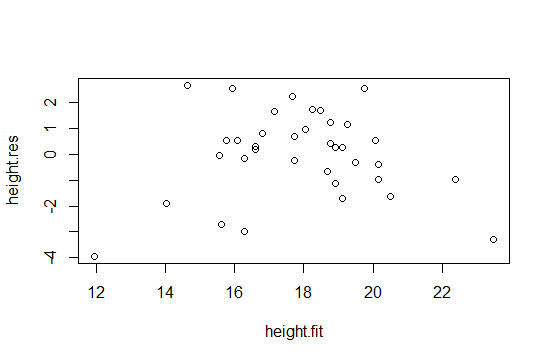
35 15.9 17.6

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.



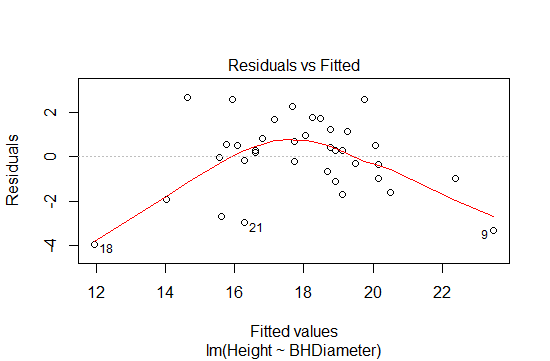
* + Make a linear model object, **spruce.lm=with(spruce.df,lm(Height~BHDiameter))**
  + Find the residuals using **residuals()**, put them into an object called **height.res**
  + Find the fitted values using **fitted()** and place them in an object called **height.fit**.
  + Plot the residuals vs fitted values.



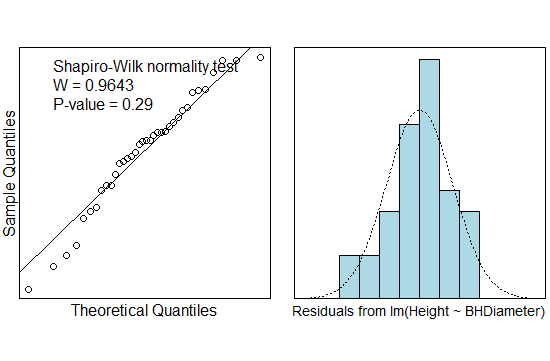
* + Plot the residuals vs fitted values using **trendscatter()**



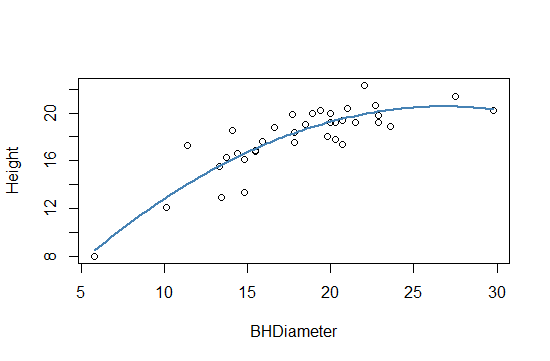
* + What shape is seen in the plot? Compare it with the curve made with the trendscatter function (second line after Task3).
    - Looks like a mountain with a Steep Positive slope, Plane out, then Steep Decline.
    - This looks different from the trendscatter as in the plot looks very skewed. The decline line segment and plots look pulled down to bottom of the plot.
  + Using the plot() function and spruce.lm, make the residual plot.



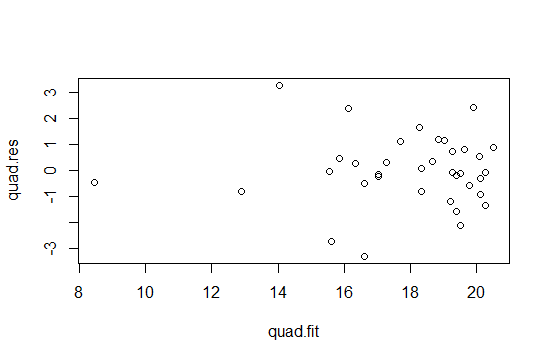
* + Check normality using the s20x function **normcheck()**. Please note that you may need to add an additional option to show the Shapiro-Wilk test (use **?normcheck** )



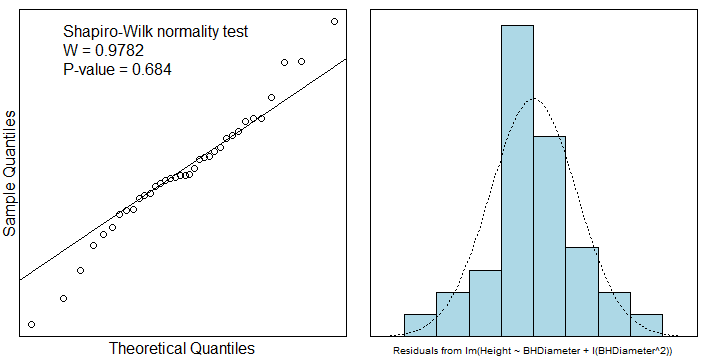
* + What is the pvalue for the Shapiro-Wilk test? What is the NULL hypothesis in this case?
    - The pvalue is 0.29. As the Height of the Tree is larger, the Breadth Width is larger.
  + describes the model used above. Notice that the residuals estimate the model errors . If the model works well with 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.
    - Applying a straight line to this data set is a bad conclusion, due to the data not being linear. A quadratic would be a better fit for this data set. The probability of the linear model being true is very low.
* 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?
    - Pvalue is 0.684
    - The probability is better the population to fit the line when the line is a quadratic. The Quadratic Model is a better option.
* Task 5\*
  + Summarize quad.lm paste it here.

Call:

lm(formula = Height ~ BHDiameter + I(BHDiameter^2), data = spruce.df)

Residuals:

Min 1Q Median 3Q Max

-3.2966 -0.6245 -0.0707 0.7442 3.2541

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 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.382 on 33 degrees of freedom

Multiple R-squared: 0.7741, Adjusted R-squared: 0.7604

F-statistic: 56.55 on 2 and 33 DF, p-value: 2.182e-11

* + What is the value of ?
    - 0.860896
  + What is the value of ?
    - 1.469592
  + What is the value of ?
    - -0.027457
  + Make interval estimates for .
    - .9, 1.5, -.03
  + Write down the equation of the fitted line.
  + Predict the Height of spruce when the Diameter is 15, 18 and 20cm (use predict())
    - 15 :: 16.72690
    - 18 :: 18.41740
    - 20 :: 19.26984
  + Compare with the previous predictions.
    - The previous predictions were smaller heights all around for all trees 15, 18, and 20.
  + What is the value of multiple ? Compare it with the previous model.
  + Quad
    - 0.7741266
  + Line
    - 0.6569146
  + 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 ( mean in this case?
    - Percent of total Variation within Equation
  + Which model explains the most variability in the Height?
    - The Quadratic Model has a better percentage of total variation.
  + 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

Analysis of Variance Table

Response: Height

Df Sum Sq Mean Sq F value Pr(>F)

BHDiameter 1 183.245 183.245 95.975 2.701e-11 \*\*\*

I(BHDiameter^2) 1 32.696 32.696 17.125 0.0002269 \*\*\*

Residuals 33 63.007 1.909

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table

Response: Height

Df Sum Sq Mean Sq F value Pr(>F)

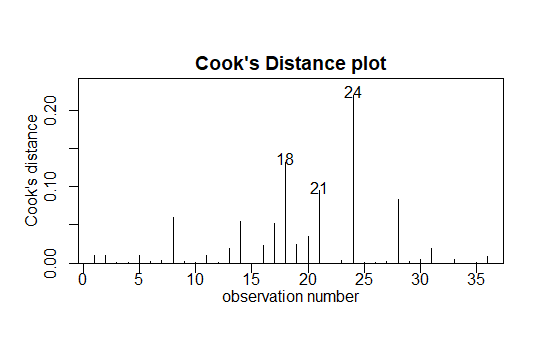
BHDiameter 1 183.245 183.245 65.101 2.089e-09 \*\*\*

Residuals 34 95.703 2.815

---

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.
    - Cooks Distance is a way of finding outliers. If the point is 3 times the mean, then it is most likely an outlier. These are based on leverage and the residuals.
  + What does cooks distance for the quadratic model and data tell you?
    - Anything above .5 should be investigated, which no data is at or above 0.5.
  + 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.

Call:

lm(formula = Height ~ BHDiameter + I(BHDiameter^2), data = spruce.df[-24,

])

Residuals:

Min 1Q Median 3Q Max

-3.11233 -0.48227 0.01253 0.71727 2.59146

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.341500 2.068479 -0.165 0.87

BHDiameter 1.564793 0.226102 6.921 7.78e-08 \*\*\*

I(BHDiameter^2) -0.029242 0.006114 -4.782 3.74e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.266 on 32 degrees of freedom

Multiple R-squared: 0.8159, Adjusted R-squared: 0.8044

F-statistic: 70.91 on 2 and 32 DF, p-value: 1.74e-12

* + Compare with the summary information from quad.lm
  + What do you conclude?
    - The two quadratic lines are almost parallel, but if 24 is taken out, then the y intercept estimate decreases drastically.

###################### 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())**

