*Name: Hagar Haytham*

*Sec: 2*

*B.N: 28*

***Data Mining, Big Data and Analytics***

***Lab 5 – Linear Regression***

***Part (1):***

***For the following questions I’m going to compare sd =7 to sd =7 to illustrate the differences and how sd affects the problem.***

***1.Try changing the value of standard deviation (sd). How do the data points change for different values of standard deviation?***

The more the standard deviation increases the more scattered the data points get (noise is getting added to the data)

|  |  |
| --- | --- |
| Sd=2 | Sd=7 |
|  |  |

***2.How are the coefficients of the linear model affected by changing the value of standard deviation in Q1?***

We are trying to fit a line (5+6x) and added to it some noise, the more the sd increases the more the noise increases the more the coefficients gets far from the coefficients of the original line.

|  |  |
| --- | --- |
| Sd=2 | Sd=7 |
| Coefficients:  (Intercept) x  5.075 5.908 | Coefficients:  (Intercept) x  5.447 5.888 |

***3.How is the value of R-squared affected by changing the value of standard deviation in Q1?***

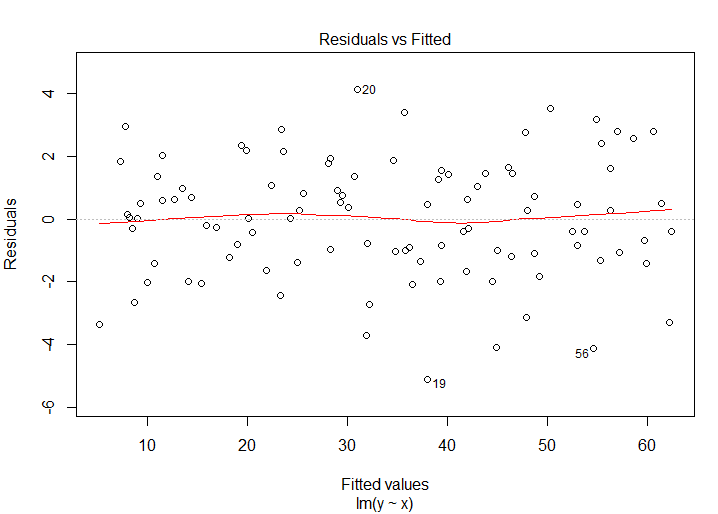
The R-squared value is equivalent to the correlation coefficient, it measures the correlation between the data points and the fitted line, when the sd was small it was easier to fit the model on the data points so they were highly (positively of course) correlated. When the sd increased it became harder to fit the data points so the R-squared value decreased a little bit.

|  |  |
| --- | --- |
| Sd=2 | Sd=7 |
| R-sqr = 0.985385 | R-sqr = 0.8372801 |

***4.What do you conclude about the residual plot? Is it a good residual plot?***

We conclude from the residual plot that the model fits the data in a good manner.

Yes, the residual plot is good as points are scattered and show no pattern so that means that linear regression model is good to fit this problem.

******

***Part (2):***

***5.What do you conclude about the residual plot? Is it a good residual plot?***

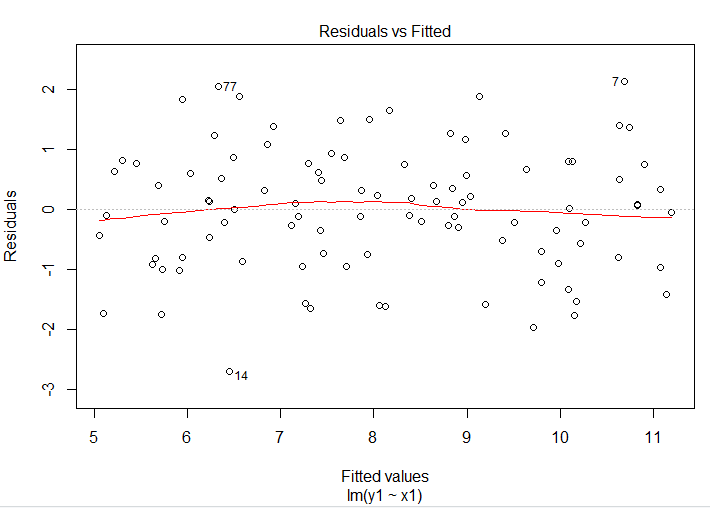
We conclude from the residual plot that the model fits the data in a good manner.

Yes, the residual plot is good as points are scattered and show no pattern so that means that linear regression model is good to fit this problem.

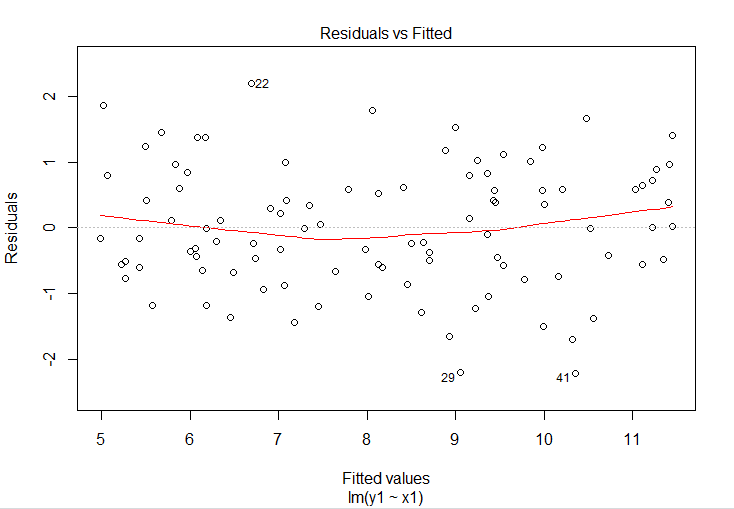
***6.Now, change the coefficient of the non-linear term in the original model for (A) training and (B) testing to a large value instead. What do you notice about the residual plot?***

When we increase the value of the non-linear term, the residual plot tends to have a quadratic pattern which shows that linear regression won’t be suitable for that case.

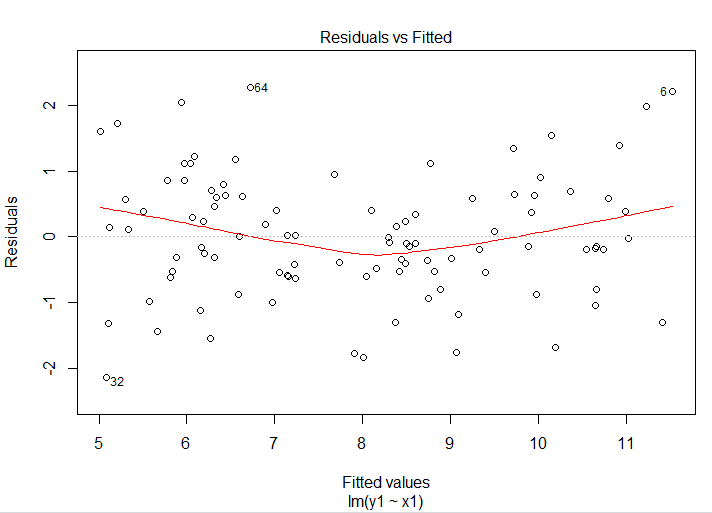
The residual plot for the model when the non linear coefficient = 0.1

******

The residual plot for the model when the non linear coefficient = 0.5

******

The residual plot for the model when the non linear coefficient = 0.8

******

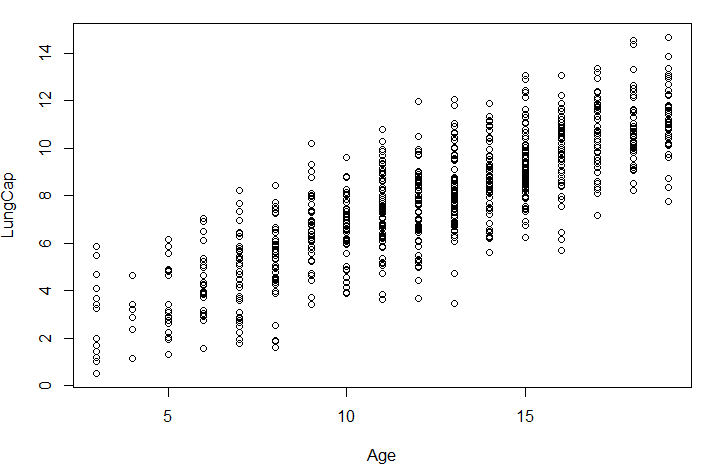
***Part (3):***

***The coding parts are int R file attached.***

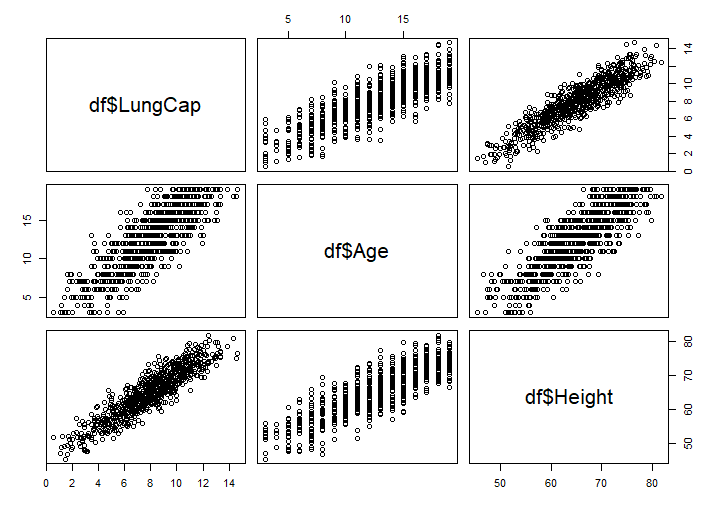
***7.Import the dataset LungCapData.tsv. What are the variables in this dataset?***

The variable names are "LungCap" "Age" "Height" "Smoke" "Gender" "Caesarean"

***8.Draw a scatter plot of Age (x-axis) vs. LungCap (y-axis). Label x-axis "Age" and y-axis "LungCap"***

******

***9.Draw a pair-wise scatter plot between Lung Capacity, Age and Height. Hint: Check the tutorial slides for how to plot a pair-wise scatterplot***

******

***10.Calculate the correlation between Age and LungCap, and between Height and LungCap. Hint: You can use the function cor.***

Correlation between Age and LungCap is 0.8196749

Correlation between Height and LungCap is 0.9121873

***11.Which of the two input variables Age and Height are more correlated to the dependent variable LungCap?***

The Height is more correlated.

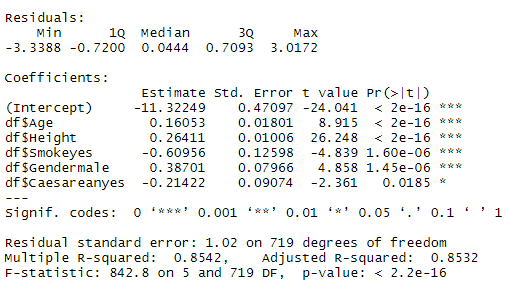
***12.Do you think the two variables Height and LungCap are correlated? Why?***

Yes, they ‘re very correlated because the correlation coefficient of them is a high positive value.

But from another perspective, the more the person is tall the more he has capacity for his lungs which makes sense.

***13.Fit a liner regression model where the dependent variable is LungCap and use all other variables as the independent variables.***

***14.Show a summary of this model.***



***15.What is the R-squared value of this model? What does R-squared indicate?***

The R-squared value is 0.8542478, and this value indicates that the data points are highly correlated with the fitted model.

***16.Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense? What should you do?***



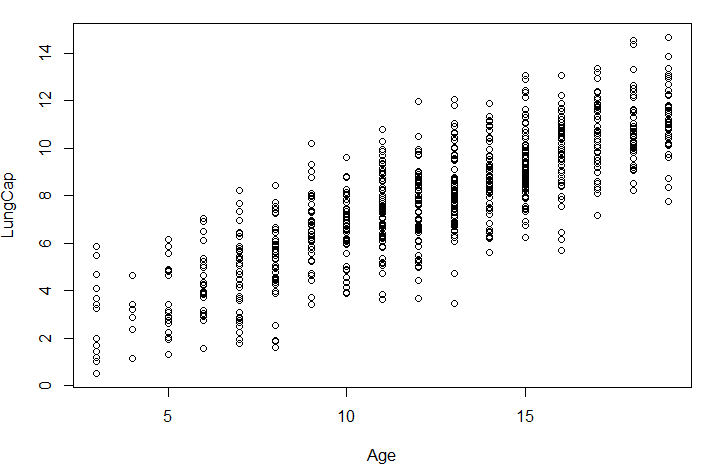
The coefficients don’t make sense, Age and Height have a high positive correlation with the LungCap, but in the model they were given very weak contribution (small coefficients)

We should remove correlated input variables.

***17.Redraw a scatter plot between Age and LungCap. Display/Overlay the linear model (a line) over it. Hint: Use the function abline (model, col="red").***

***Note (1): A warning will be displayed that this function will display only the first two coefficients in the model. It's OK.***

***Note (2): If you are working correctly, the line will not be displayed on the plot. Why?***



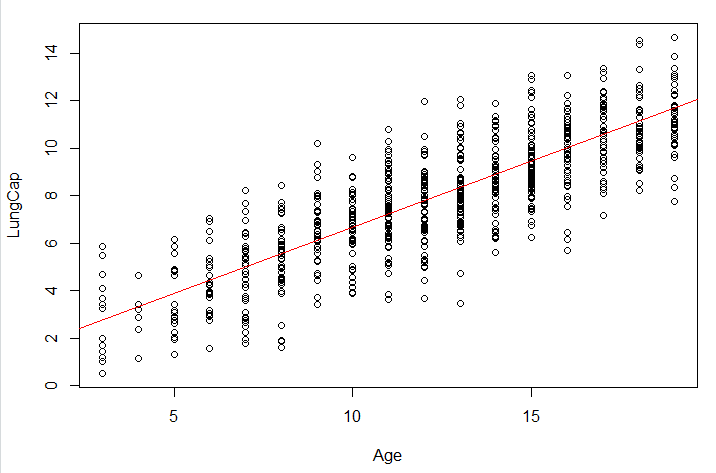
Because the line is not really affected by the age (have a very small coefficient) and it is more affected by the other input variables.

***18.Repeat Q13 but with these variables Age, Smoke and Cesarean as the only independent variables.***

***19.Repeat Q16, Q17 for the new model. What happened?***



The coefficients now make more sense as the age gained more contribution to the model as it is highly correlated with the dependent variable.



***20.Predict results for this regression line on the training data.***

***21.Calculate the mean squared error (MSE) of the training data.***