Cairo University
Faculty of Engineering
Computer Engineering Department

CMPN451
Spring 2020

Data Mining, Big Data and Analytics.

Lab 5 – Linear Regression

Note 1: At the end of the lab, you should submit a document containing **all plots** as well as full and clear answers to **non-coding questions.** The answers and plots should be in the same order as questions.

Note 2: The grade of the lab will be based on the submitted document at the end of the lab.

Part (1):

1.	Try changing the value of standard deviation (sd). How do the data points change for different values of standard deviation?
2.	How are the coefficients of the linear model affected by changing the value of standard deviation in Q1?
3.	How is the value of R-squared affected by changing the value of standard deviation in Q1?
4.	What do you conclude about the residual plot? Is it a good residual plot?

Part (2):

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

Part (3):

7.	Import the dataset **LungCapData.tsv.** What are the variables in this dataset?
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`.
11.	Which of the two input variables **Age and Height** are more correlated to the dependent variable **LungCap**?
12.	Do you think the two variables **Height** and **LungCap** are correlated? Why?
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?
16.	Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense? What should you do?
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?
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?
20.	Predict results for this regression line on the training data.
21.	Calculate the mean squared error (MSE) of the training data.