Department of Computer Science and Engineering

CSE-454: Data Warehousing and Data Mining Sessional

Assignment-3 (Regression Analysis)

In our third lab, we have seen basics of regression analysis, mathematical background and hands-on on linear regression analysis and multiple linear regression analysis respectively with some practical datasets.

In this assignment:

1. You have to do Polynomial Regression Analysis on *world Covid-19 dataset.*

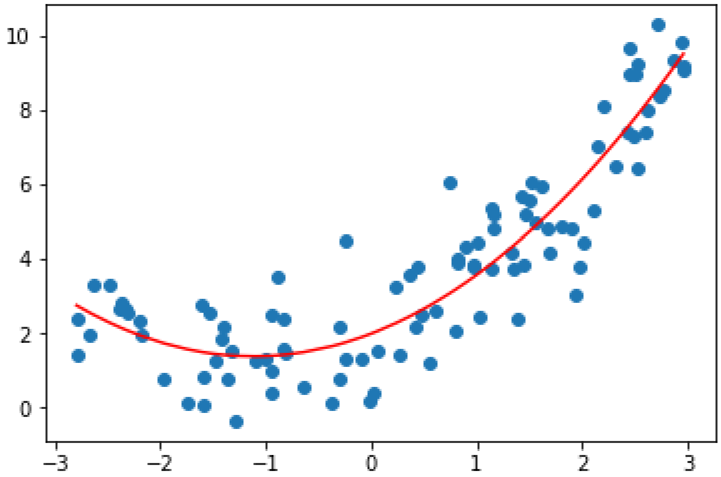
2. In linear and multiple linear regression, we fit data points in a straight line. The mathematical model for linear regression is

$$Y = \beta_0 + \beta_1 X$$

And the mathematical model for multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N$$

Polynomial linear regression is necessary where datapoints cannot be fitted into a straight line. Such case is shown in the following diagram.



The mathematical model for polynomial regression is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_N X^N$$

Here ***N*** is the *order/degree* of the model.

3. You have to analyze only Bangladesh’s data for the months April and May. Export *date*, *total_cases*, *new_cases*, and *total_deaths* from the dataset for Bangladesh from April 01, 2020 to May 31, 2020. Handle missing data appropriately.

4. Create a regression model keeping *date* as independent variable (X) and *total_case* as dependent variable (Y). Similarly create another 2 models keeping *new_cases* and *total_deaths* as dependent variable respectively. Find R^2 score for each of the models. Set appropriate value to ***N*** for each of the models which minimize R^2 scores.

5. As dates cannot be used as independent variable, you might need to convert it to appropriate format so that it can be used in the regression models.

6. Create a user defined function which will take a date as parameter and print the predicted and actual values of total cases, new cases and total deaths.

7. Use Jupyter notebook to create your project. Rename the notebook file with your student ID. File name must be your student ID (<std id>.ipynb). Do not add your name or your section in the file name.

8. Remove unnecessary code blocks before submitting your assignment.

9. Answer the following questions.

   i. What would be consequence of taking lower value of ***N*** in the regression model?

   ii. What would be consequence of taking higher value of ***N*** in the regression model?

10. Submit only one Jupyter notebook (<std id>.ipynb). file. Do not compress it nor include any other file with your submission. Answer the questions in a Markdown block at the end of the notebook.

11. DO NOT COPY FROM ANYWHERE

Marks Distribution

Ser	**Description**	**Marks**
1	Loading dataset and data preprocessing	5
2	Creating and executing regression models	10
3	Creating a user function	3
4	Answering questions	2
Total		**20**

* Deadline for submission is **Tuesday 29 September, 2020 11:55pm**

* This assignment will carry 10% weight in final grading.

* Don’t do copy-and-paste programming. Severe actions will be taken against any sort of plagiarism.

* Please leave a comment if find any difficulties.