Statistical Natural Language Processing

**assignment**

Course Name

Introduction to Data Science with Python

Course Code PM-ASDS04

submitted to

Dr. Ajit Kumar Majumder

Professor, Department of Statistics

jahangirnagor Univarsity

sumitted From

MD Jahidul alam

PM-ASDS-Batch a

Roll-201900101051

Introduction: Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. We will discuss about some statistical calculation which considered a major part in Data Science platform. The discussion is related with Random Number Generation and Regression Analysis from data.

Objectives:

Objectives of Random Number Generation: To generate a sequence of numbers or symbols that cannot be reasonably predicted well than by a random chance.

Objectives of Regression Analysis: To understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

Methodology:

Random Number Generation: A random number generator (RNG) is a mathematical construct, either computational or as a hardware device, that is designed to generate a random set of numbers that should not display any distinguishable patterns in their appearance or generation, hence the word random.

Random Number Generation working procedure in excel:

1. To generate random numbers, first click the Data tab’s Data Analysis command button.
2. In the Data Analysis dialog box, select the Random Number Generation entry from the list and then click OK. Excel displays the Random Number Generation dialog box.
3. Then we describe how many columns and rows of values that we want.
4. Select the distribution method.
5. (Optional) Provide any parameters needed for the distribution method.
6. (Optional) Select a starting point for the random number generation.
7. Identify the output range.
8. After we describe how we want Excel to generate random numbers and where those numbers should be placed, click OK. Excel generates the random numbers.

Regression Analysis: Regression Analysis includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. A regression model is a mathematical equation that describes the relationship between two or more variables. A regression model that includes two or more independent variables is called a multiple regression model. It is written as

y = A + B1x1 + B2x2 + B3x3+ … + Bkxk + ε

where y is the dependent variable, x1, x2, x3, …, xk are the k independent variables, and ε is the random error term.

Regression Analysis working procedure in excel:

1. On the Data tab, in the Analysis group, click the Data Analysis button.
2. Select Regression and click OK.
3. In the Regression dialog box, configure the following settings: Select the Input Y Range, which is your dependent variable. Select the Input X Range, i.e. our independent variable. If we are building a multiple regression model, select two or more adjacent columns with different independent variables.

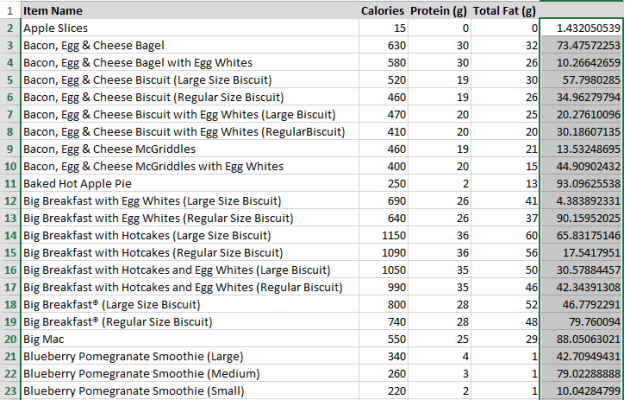
• Check the Labels box if there are headers at the top of your X and Y ranges.

• Choose your preferred Output option, a new worksheet in our case.

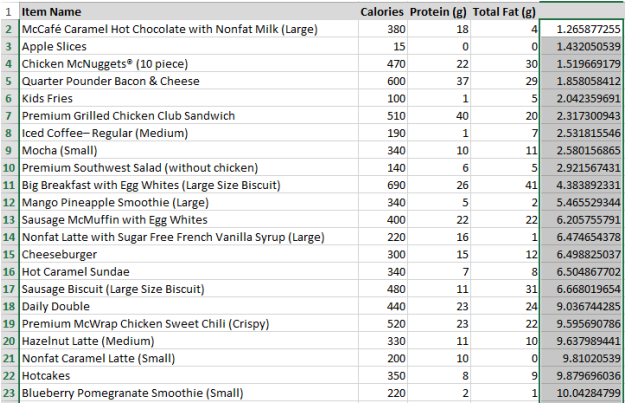
• Optionally, select the Residuals checkbox to get the difference between the predicted and actual values. IV. Click OK and observe the regression analysis output created by Excel.

Results Analysis:

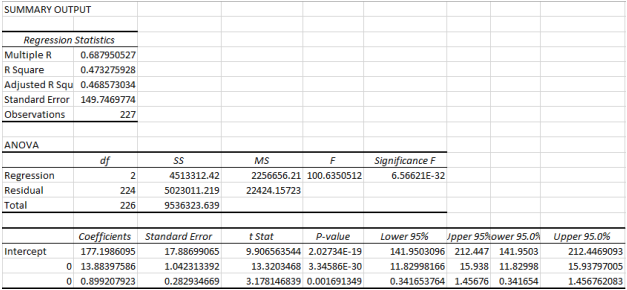
1. We generate unique random numbers from the given data (which data is provided in our classroom) in Figure-01.



1. We also sort the generated random numbers from smallest to largest in Figure-02.



1. Regression Analysis: Calories versus Protein, Total Fat from given data (which data is provided in our classroom) in Figure-03.



Discussion:

Let, y = Calories x1 = Portion x2 = Total Fat

We are to estimate the regression model

y = A + B1x1 + B2x2 + ε

From the output given in Screen, the estimated regression equation is:

y = 177.19 + 13.88x1 + 0.89x2

The value of a = 177.19 in the estimated regression equation gives the value of y for x1 = 0 and x2 = 0. The value of b1 = 13.88 in the estimated regression model gives the change in y for a one-unit change in x1 when x2 is held constant.

The value of b2 = 0.89 in the estimated regression model gives the change in y for a one-unit change in x2 when x1 is held constant. Here P-value is approximately ‘0’, so one variable has linear influence on calories. H0: B1 = 0 H1: B1 < 0 Portion has linear influence on calories.

I wanted to do Random Number Generation and Regression Analysis from the given data (which data is provided in our classroom) with Python but I could not find out our desired output for my programming lacking, so I try my best in excel to solve the problem.

Summary: We can find out our desired output by a statistical process for estimating the relationships among variables which is called Regression Analysis. By this analysis we know that portion has linear influence on calories. We also generate random numbers to generate a sequence that does not have any pattern, therefore appear to be random.