Experiment No.3

Title: Predicting missing data values using regression modeling

Batch: **Roll No.:** **Experiment No.: 2**

Aim: Predict missing data values using regression modeling.

Resources needed: Any programming language, any data source (RDBMS/Excel/CSV)

Theory:

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in data set and can have a significant effect on the conclusions that can be drawn from the data. There are various techniques proposed for handling missing values like deletion of records/attributes, filling with a random value or using some measures of central tendency, imputation using regression etc. Regression imputation is guessing missing variables using regression if we know there is a correlation between the missing value and other variables. Scatterplots can be used to identify correlation between variables.

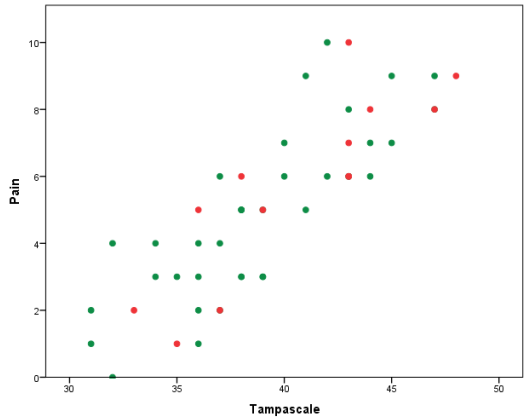

Figure 1: A scatter plot showing correlation between attributes pain and tampascale.

Once correlation is identified either linear regression or multiple regression can be used for imputation. Linear regression involves finding the "best" line as shown in fig. 1 to fit two attributes (or variables) so that one attribute can be used to predict the other.

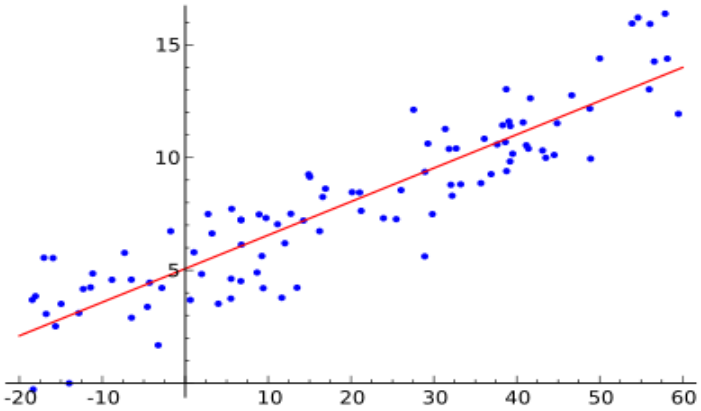

Figure 2: Example of simple linear regression, which has one independent variable

Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Prediction is predicting continuous or ordered values for a given input i.e. Numeric prediction, for example, predicting salary of employee with 10 years of experience.

Simple Linear Regression:

Straight line regression analysis involves a responsible variable *y* and a single predictor variable *x*. by modeling *y* as a linear function of *x* as given in equation 1,

$y=w0 + w1*x$ …….. (1)

where *w0* and *w1* are Regression co-efficient.

w0 = Y-intercept
w1 = Slope of the line

Calculate *w0* and *w1* by method of least squares, which estimates best fitting straight line.

Let *D* be a training set,
[*D*] = *{ (x1,y1),(x2,y2),(x3,y3),……,(x_n, y_n)}*

Regression co-efficient,

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i-\bar{x})\,(y_i-\bar{y})}{\sum_{i=1}^{|D|}(x_i-\bar{x})^2} \quad \text{……..(2)}$$

$$w_0 = \bar{y} - w_1\bar{x} \text{……..……(3)}$$

Where $\bar{x}$ is the mean value of *x1, x2, x3,….x_n*.

And $\bar{y}$ is the mean value of *y1, y2, y3, y4,…y_n*.

Multiple Linear Regression:

Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \ldots + b_nx_{in} + \varepsilon$$

Where , for i=1 to n obervetions:

y_i = dependent variable
x_{i1} , x_{i2} , ……= predictor variables
b_1, b_2, b_3…… = Regression coeffients
ε = Model's error term

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_k \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & . & . & . & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & . & . & . & x_{2,k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n,1} & x_{n,2} & . & . & . & x_{n,k} \end{bmatrix}$$

To handle the complications of multiple regression, we will use matrix algebra. The least squares normal equations can be expressed as: **Y=Xb**--------Multiply both sides with X^T

$\mathbf{X^TY = X^TXb}$ or $\mathbf{X^TXb = X^TY}$

Here, matrix $\mathbf{X^T}$ is the transpose of matrix **X**. To solve for regression coefficients, simply pre-multiply by the inverse of $\mathbf{X^TX}$:

$\mathbf{(X^TX)^{-1}X^TXb = (X^TX)^{-1}X^TY}$ ……since $\mathbf{(X^TX)^{-1}X^TX = I}$, the identity matrix, we get slope b as,

$\mathbf{b = (X^TX)^{-1}X^TY}$

Procedure / Approach /Algorithm / Activity Diagram:

1. Identify attributes suitable for applying Linear regression. Construct a linear regression model for your dataset and predict the missing values in your data set. Evaluate the accuracy of prediction.(usage of built in package for prediction is not expected)
2. Identify attributes suitable for applying Multiple Linear regression. Construct a linear regression model for your dataset and predict the missing values in your data set. Evaluate the accuracy of prediction. .(usage of built in package for prediction is not expected)

Results: (Program printout with output / Document printout as per the format)

Questions:

1. How will you choose between linear regression and non-linear regression?

2. Explain the nature or characteristics of a dataset where we can apply regression imputation.

Outcomes:

_____
_____
_____
_____

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

_____
_____
_____
_____
_____

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

_____

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition