**Experiment No.3**

**Title:**Predicting missing data using regression modeling

|  |
| --- |
|  |

**Batch: A4 Roll No.:1914078 Experiment No.:3**

**Aim:**Predict missing data 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.

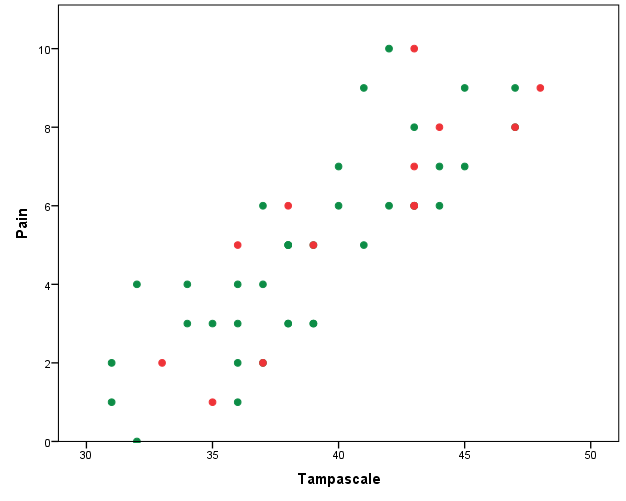


Figure 1: A scatter plot showing correlation between attributes pain and tampascale.

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.

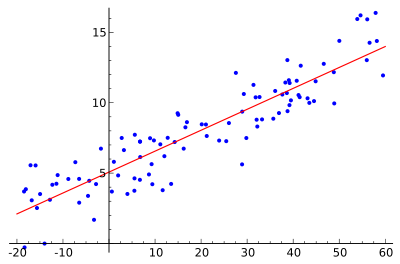


Figure 2:Example of simple linear regression, which has one independent variable

*Multiple linear regression* is an extension of linear regression, where more than twoattributes 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),…….,(xn,yn)}*

Regression co-efficient,

………………………………………………………………..(2)

…………………………………………………..…………………….(3)

Where is the mean value of *x1,x2,x3,….xn*.

And is the mean value of *y1,y2,y3,y4,…yn.*

**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.​

= + ++…+ +ε

Where , for i=1 to n obervetions:

= dependent variable

= predictor variables

= Regression coeffients

ε = Model’s error term

A picture containing diagram

Description automatically generated

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 XT​

**XTY** = **XTXb**    or    **XTXb** = **XTY**​

Here, matrix **XT** is the transpose of matrix **X**. To solve for regression coefficients, simply pre-multiply by the inverse of **XTX**:​

**( XTX)-1 XTXb** = **( XTX)-1 XTY …........**since **( XTX)-1 XTX** = **I**, the identity matrix, we get slope b as,​

**b** = **(XTX)-1 XTY**

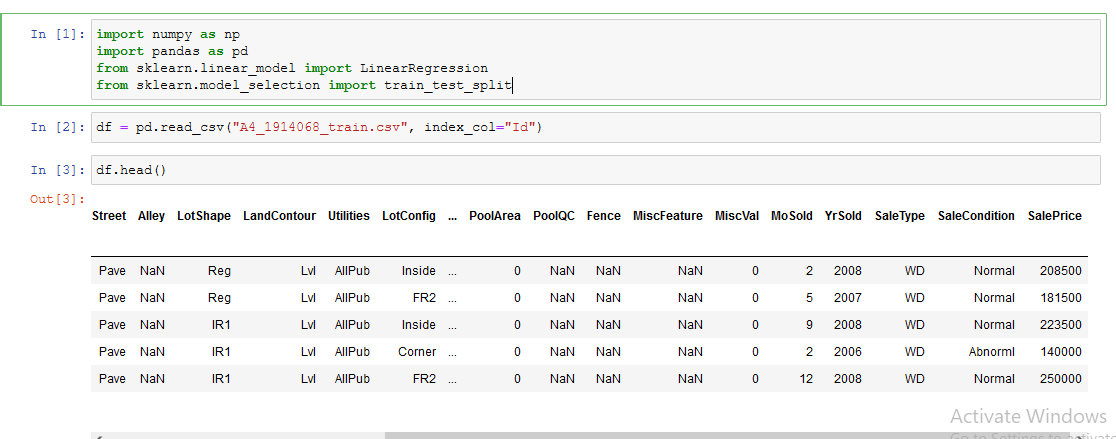
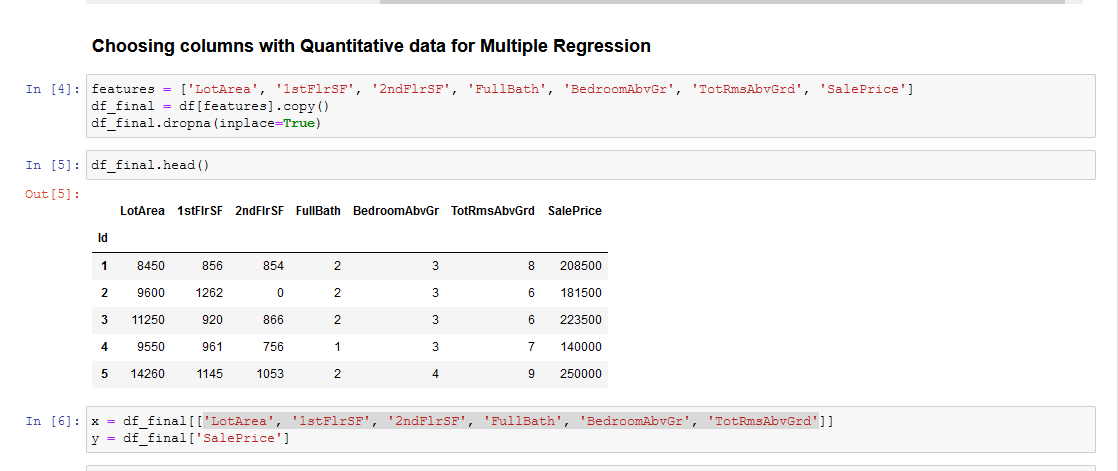
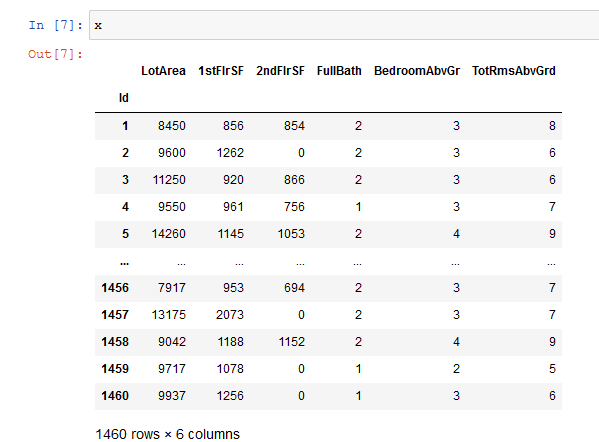
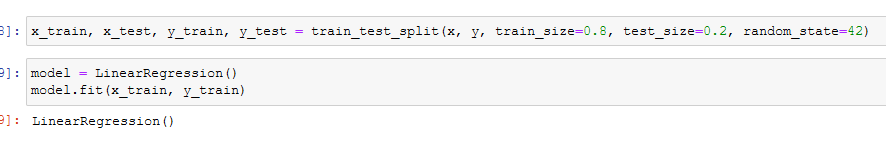
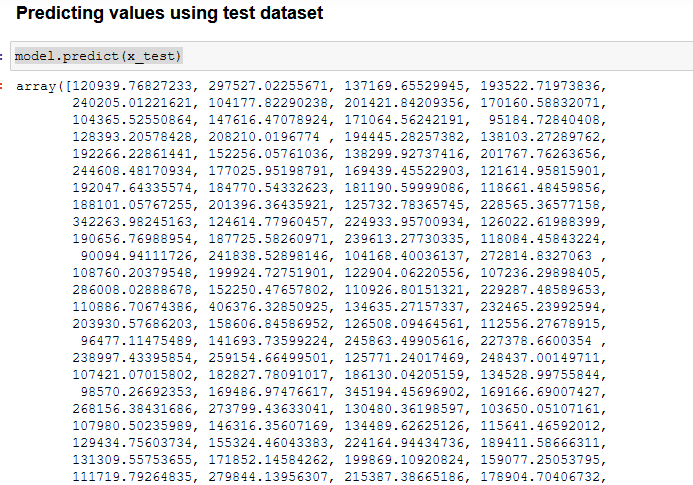
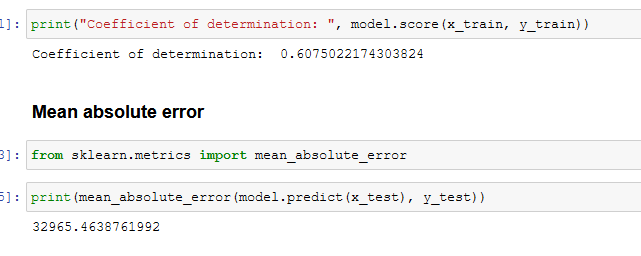
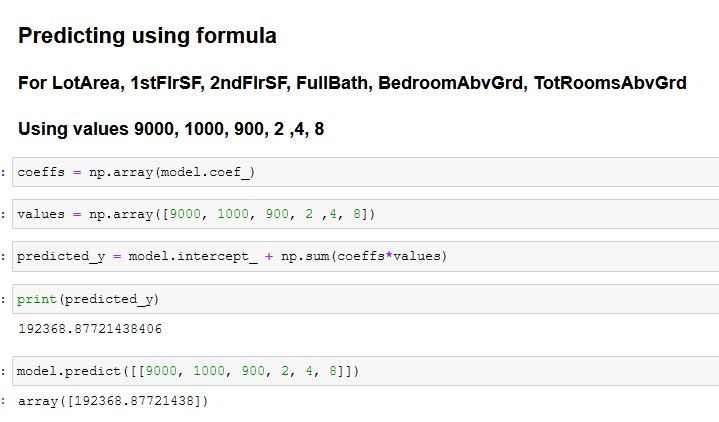
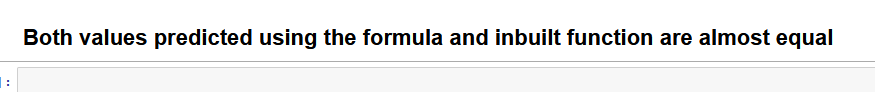
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Procedure / Approach /Algorithm / Activity Diagram:**

1. Identify attributes suitable for applying Linear regression. Constructa linear regressionmodel 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. Constructa linear regressionmodel 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?

* Linear or Non-linear, both models are trying to fit to the curvature of your data. Linear models have to follow a particular form of straight line
  + + …..
* Any other model is a non linear regression model

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

* A regression model is estimated to predict observed values of a variable based on other variables, and that model is then used to impute values in cases where the value of that variable is missing. In other words, available information for complete and incomplete cases is used to predict the value of a specific variable. Fitted values from the regression model are then used to impute the missing values. The problem is that the imputed data do not have an error term included in their estimation, thus the estimates fit perfectly along the regression line without any residual variance. This causes relationships to be over identified and suggest greater precision in the imputed values than is warranted. The regression model predicts the most likely value of missing data but does not supply uncertainty about that value.

**Outcomes:**

Comprehend descriptive and proximity measures of data

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

Used Linear and Multiple regression on paper, Excel and my dataset.

**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