A Regression Modeling Technique on Data Mining

Swati Gupta
Assistant Professor, Department of Computer Science
Amity University Haryana, Gurgaon, India

ABSTRACT
A regression algorithm estimates the value of the target (response) as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

In this paper, we have discussed the formulation of linear regression technique, along with that linear regression algorithm have been designed, further test data are taken to prove the linear regression algorithm.

Keywords
Linear regression, dependent variable, independent variables, predictor variable, response variable

1. INTRODUCTION
Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression [2], uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x[1,6]. Basically a Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation *solves for*) is called the **dependent variable**. The factors that are used to predict the value of the dependent variable are called the **independent variables**.

1.1 Simple Linear Regression Model
The simple linear regression model is represented by the equation:

$$y = \alpha + \beta X$$

By mathematical convention, the two factors that are involved in a simple linear regression analysis are designated *X* and *y*. The equation that describes how *y* is related to *x* is known as the **regression model** [2, 3]. Here in the equation α is the *y* intercept of the regression line and *β* is the slope.

A regression line can show a positive linear relationship, a negative linear relationship, or no relationship [5, 6]. If the graphed line in a simple linear regression is flat (not sloped), there is no relationship between the two variables. If the regression line slopes upward with the lower end of the line at the *y* intercept (axis) of the graph, and the upper end of line extending upward into the graph field, away from the *x* intercept (axis) a positive linear relationship exists. If the regression line slopes downward with the upper end of the line at the *y* intercept (axis) of the graph, and the lower end of line extending downward into the graph field, toward the *x* intercept (axis) a negative linear relationship exists[7,8].

2. FORMULATION OF LINEAR REGRESSION TECHNIQUE
Linear Regression model consist of random variable Y (called as a response variable) as a linear function of another random variable X (called as a predictor variable) that is represented by the equation

$$Y = \alpha + \beta X \qquad \text{(eq 1)}$$

α & β are regression coefficients specifying the Y intercept and slope of the line respectively.

The regression coefficient α & β are solved by the method of least squares, which minimize the error between the actual data & the estimate of the line[9] .Given s sample of data or data points of the form (x1,y1),(x2,y2)…….(xs, ys) than the regression coefficients α & β are given by

$$\beta = \sum(xi - x)\ (yi - y) / \sum(xi - x) \qquad \text{(eq 2)}$$

$$\alpha = y - \beta x \qquad \text{(eq 3)}$$

These values of regression coefficients α and β calculated in equation (2) & (3) are substituted in equation (1) so as to obtain the relationship between the response variable X and the target variable Y.

2.1 Algorithm of Linear Regression Technique
The linear regression technique works on the following algorithm

Step 1: Take the values of variable Xi and Yi

Step 2: Calculate the average for variable Xi such that average is x= (X1 +X2 +……..+ Xi)/ Xi

Step 3: Calculate the average for variable Yi such that average is y= (Y1 +Y2 +……..+ Yi)/ Yi

Step4: Calculate the value of regression coefficient β by substituting the values of Xi , Yi average of Xi and average of Yi in the equation 2

Step 5: Calculate the value of another regression coefficients α by substituting the values of β (calculated in step 4), average of Xi and average of Yi in the equation 3

Step 6: Finally substitute the value of regression coefficients α and β in the equation Y= α + βX

3. TEST DATA FOR LINEAR REGRESSION
In order to analyze the working and result of linear regression technique we have taken a different test data. We put these data values in the regression equations and then analyze the result that has been obtained.

Table 1: The test data for linear regression

X(years of experience)	Y(Salary)(in K)
3	**30**
8	**57**

9	**64**
13	**72**
3	**36**
6	**43**
11	**59**
21	**90**
1	**20**
16	**83**

Here X is the number of years of work experience and Y is the corresponding salary. We model a relationship that the salary must be related to the number of years of experience with the equation $Y= \alpha + \beta X$

1) Given the above test data we first compute the average of x and y such that average of x is 9.1 and average of y is 55.4

2) Next we compute the value of regression coefficients α and β

Such that $\beta = \sum(xi - x) (yi -y)/ \sum(xi -x)$ and

$\alpha = y - \beta x$

We now have

$\beta=(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\ldots\ldots+(16-9.1)(83-55.4)/(3-9.1)+(8-9.1)+\ldots.+(16-9.1)=3.5$

$\alpha =55.4-(3.5)(9.1)=23.6$

and now finally the equation of the least square line is estimated by

$$Y=23.6+3.5X$$

using this equation we can predict the salary of college graduate with say 10 years of experience

So now if we want to know the salary of a person whose experience is 10 years we can get the result

$$Y=23.6+3.5(10)$$

$$=58.6$$

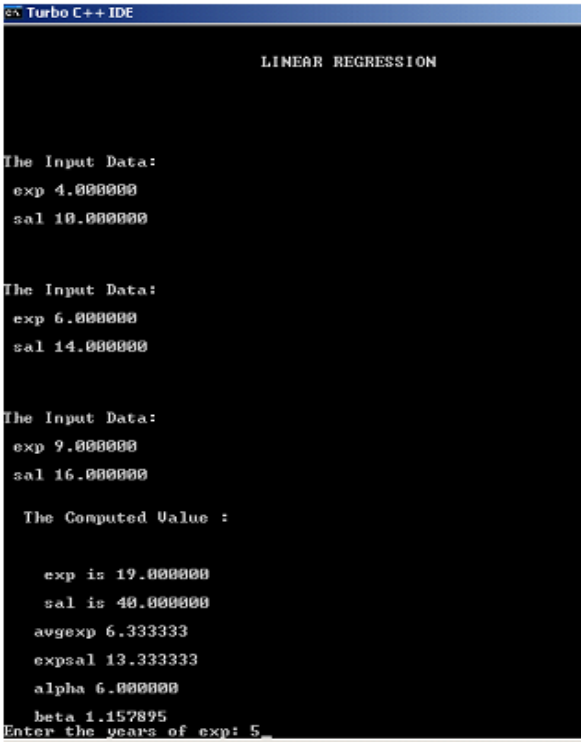
So, a person with 10 years of experience has a salary of 58.6K

Thus the above test data can be used to verify that the linear relationship exist between the variables X and Y.

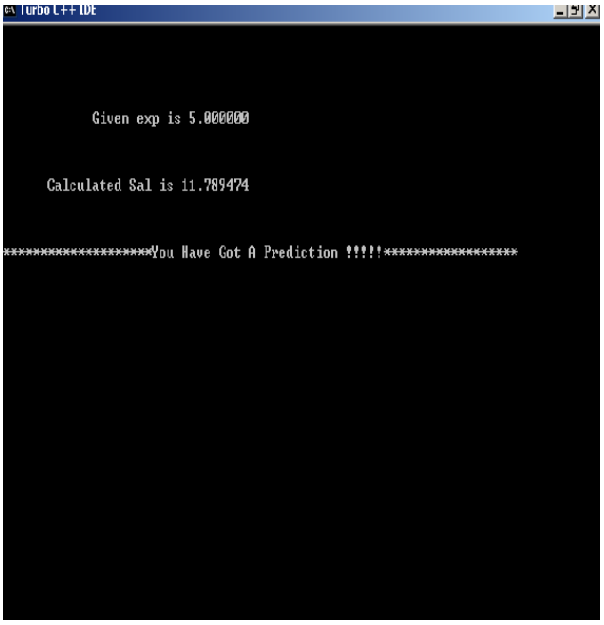
4. IMPLEMENTATION OF LINEAR REGRESSION TECHNIQUE (SNAP SHOTS)

The linear regression technique has been implemented in C. The following snapshot are taken

1) Input Data Screen:



2) Output Data Screen



5. CONCLUSIONS

The Linear Regression technique predicts a numerical value. Regression performs operations on a dataset where the target values have been defined already. And the result can be extended by adding new information. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other datasets where the target values are not known. In this paper we have formulate a linear regression technique, further we have designed the linear regression algorithm. The test data are taken to prove the relationship between predictor and target variable which is being represented by the linear regression equation

$$Y= \alpha + \beta X$$

where variable Y (called as a response variable) as a linear function of another random variable X (called as a predictor variable) , α and β are linear regression coefficients.

6. REFERENCES

[1] Manisha rathi Regression modeling technique on data mining for prediction of CRM CCIS 101, pp.195-200,2010Springer–Verlag Heidelberg 2010.

[2] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[3] Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of UAI-1998: The Fourteenth Conference on Uncertainty in Artificial Intelligence.

[4] Burges, C. (1998). A tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):955–974.

[5] CiteSeer (2002). CiteSeer Scientific Digital Library. http://www.citeseer.com.

[6] Duda, R. O. and Hart, P. E. (1973). Pattern Classification and Scene Analysis. John Wiley & Sons.

[7] GLIM (2004). Generalised Linear Interactive Modelingpackage. http://www.nag.co.uk/stats/GDGEsoft.asp, http://lib.stat.cmu.edu/glim/.

[8] Greenbaum, A. (1997). Iterative Methods for Solving Linear Systems, volume 17 of Frontiers in Applied Mathematics. SIAM..

[9] Kubica, J., Goldenberg, A., Komarek, P., Moore, A., and Schneider, J. (2003). A comparison of statistical and machine learning algorithms on the task of link completion. In KDD Workshop on Link Analysis for Detecting Complex Behavior, page 8.