# A Multiple Regression Technique in Data Mining

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

## ABSTRACT

The growing volume of data usually creates an interesting challenge for the need of data analysis tools that discover regularities in these data. Data mining has emerged as disciplines that contribute tools for data analysis, discovery of hidden knowledge, and autonomous decision making in many application domains. The Multiple regression generally explains the relationship between multiple independent or multiple predictor variables and one dependent or criterion variable. The 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 multiple regression technique, along with that multiple regression algorithm have been designed, further test data are taken to prove the multiple regression algorithm.

## Keywords

Multiple regression**,** dependent variable, independent variables, predictor variable, response variable

## 1. INTRODUCTION

Regression is a data mining (machine learning) technique which is used to fit an equation for the dataset. A Multiple regression technique is an extension of a linear regression technique which involves more than one predictor variable[1,2]. It allows response variable Y to be modeled as a linear function of multidimensional feature vector that is we have

$$Y= \alpha + \beta_1 X_{1+} \beta_2 X_2 \qquad \text{(eq 1)}$$

Where α, $\beta_1$ and $\beta_2$ are regression coefficients

The variable whose value is to be predicted is known as the **dependent variable** and the ones whose known values are used for prediction are known **independent (exploratory) variables**.

In this technique, a dependent variable is modeled as a function of several independent variables with corresponding multiple regression coefficients, along with the constant term[3,4]. Multiple regressions requires two or more predictor variables, and that is why it is called multiple regression

### 1.1 Multiple Regression Model

Multiple regression model maps a group of predictors x to a response variable y [3]. The multiple linear regressions is defined by the following relationship,

for i = 1, 2, n:

$$yi = a + b1xi1 + b2xi2 + \cdot\cdot\cdot +bkxik + ei \qquad \text{(eq 2)}$$

Equivalently, in more compact matrix terms:

$$Y = Xb + E \qquad \text{(eq3)}$$

For all the n considered observations,

Y is a column vector with n rows containing the values of the response variable; X is a matrix with n rows and k + 1 columns containing for each column the values of the explanatory variables for the n observations, plus a column (to refer to the intercept) containing n values equal to 1; b is a vector with k + 1 rows containing all the model parameters to be estimated on the basis of the data: the intercept and the k slope coefficients relative to each explanatory variable. Finally E is a column vector of length n containing the error terms[7,8].

In the bivariate case the regression model was represented by a line, now it corresponds to a (k + 1)-dimensional plane, called the regression plane[9,10]. This plane is defined by the equation

$$\hat{y}i= a + b1xi1 + b2xi2 + \cdot\cdot\cdot +bkxik+\mu i \qquad \text{(eq4)}$$

Where ŷi is dependent variable. Xi 's are independent variables, and μi is stochastic error term.

## 2. FORMULATION OF MULTIPLE REGRESSION TECHNIQUE

A Multiple regression technique is an extension of a linear regression technique which involves more than one predictor variable. It allows response variable Y to be modeled as a linear function of multidimensional feature vector. Multiple Regression model consist of random variable Y (called as a response variable) as a linear function of random variable X1 (called as a predictor variable) and $X_2$ and that is represented by the equation that is we have

$$Y= \alpha + \beta_1 X_{1+} \beta_2 X_2 \qquad \text{(eq 1)}$$

Where α , $\beta_1$ and $\beta_2$ are regression coefficients

The regression coefficient α , $\beta_1$ & $\beta_2$ are solved by the method of least squares, which minimize the error between the actual data & the estimate of the line. Basically multiple regression generally explains the relationship between multiple independent or multiple predictor variables and one dependent or criterion variable. In multiple regression, a dependent variable is modeled as a function of several independent variables with corresponding multiple regression coefficients, along with the constant term

### 2.1 Algorithm of Multiple Regression Technique

The Multiple regression technique works on the following algorithm

**Step 1**: Take the values of variable Xi, Xb and Yi

**Step 2**: Calculate the summation of the variable Xi,Xb,Yi

**Step 3**: Calculate the product of summation terms (∑ x1*x,∑x1*x2,∑ x2* y,∑x2*x2, ∑ x1*y)

**Step 4**: Solve the equations

$$\sum y = na0 + a1\sum x1 + a2\sum x2$$

$$\sum x1y = a0\ x1 + a1\sum x1*x1 + a2\sum x1x2$$

$$\sum x2\ y = a0\ x2 + a1\sum x1*x2 + a2\sum x2*x2$$

**Step 5**: Now calculate the value of a0, a1, a2 which is calculated by the inverse of a matrix

**Step 6**: Finally obtain the value of response variable Y by knowing the values of a0, a1, a2 in the equation Y= a0+a1X1+ a2X2 of β (calculated in step 4), average of $X_i$ and average of $Y_i$

**Step 7**: Finally substitute the value of regression coefficients α and β in the equation Y= α + βX

## 3. TEST DATA FOR MULTIPLE REGRESSION TECHNIQUE

In order to analyze the working and result of multiple 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 multiple regression**

| X1(Years of experience) | X2(Working hrs) | Y(Salary in K) |
|---|---|---|
| 4 | 6 | 10 |
| 6 | 8 | 14 |
| 9 | 12 | 16 |

Here X1 is the years of work experience, X2 is the working hrs and Y is the corresponding salary. We model a relationship that the salary must be related to years of work experience and working hrs with the equation Y= α + β1 X1+ β2 X2

We than calculate the summation of the variable X1, X2 and Y the result is stored in a variable

∑ x1 =19 , ∑x2 =26 and ∑y=402

We than calculate the product of summation terms

(∑x1*x1=133, ∑x1*x2=180, ∑ x2* y=364, ∑x2*x2=244, ∑ x1*y=268)

In order to solve the equations

$$\sum y = a0 + a1\sum x1 + a2\sum x2$$

$$\sum x1*y = a0\ x1 + a1\sum x1*x1 + a2\sum x1*x2$$

$$\sum x2*y = a0\ x2 + a1\sum x1\ *x2 + a2\sum x2*x2$$

We now calculate the value of a0=10,a1=6, a2=-4 using inverse of a matrix

USING THE LINEAR EQUATION Y= α + β1 X1+ β2 X2

WE CAN PREDICT THE SALARY OF A PERSON WITH SAY 5 YEARS OF WORK EXPERIENCE,7 WORKING HRS BY SUBSTITUTING THE COMPUTED VALUE IN THE EQUATION

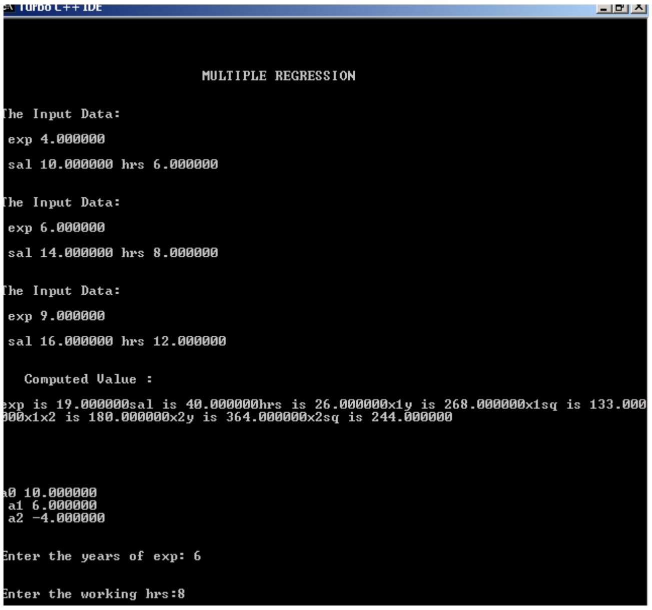
$$Y = 10+6*5+ (-4)\ (7) = 12$$

So the salary of a person is 12 having 5 year of experience and 7 working hrs.

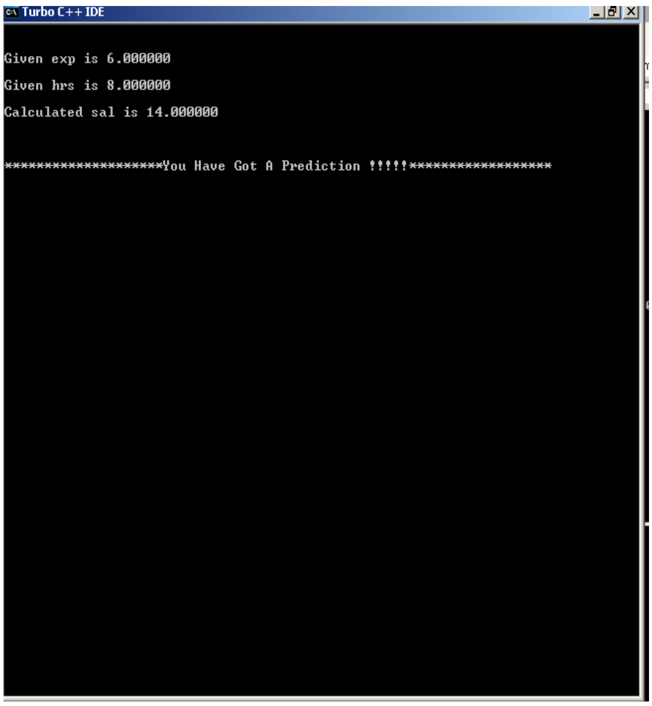
## 4. IMPLEMENTATION OF LINEAR REGRESSION TECHNIQUE (SNAP SHOTS)

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

1) Input Data Screen:




2) Output Data Screen



# 5. CONCLUSION

The Multiple 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 multiple regression technique, further we have designed the multiple 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

$Y= \alpha + \beta_1 X_{1+} \beta_2 X_2$ where random variable Y (called as a response variable) as a linear function of random variable X1 (called as a predictor variable) and $X_2$ $\alpha$ and $\beta$ 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] Giudici Paolo, "Applied Data Mining-Statistical methods for business and industry" wiley, (2003) [5] Dash, M., and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis. 1:3 (1997) pp. 131-156. [6] Rencher C. Alvin, "Methods of Multivariate Analysis" 2nd Edition, Wiley Interscience, (2002).

[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/GDGE soft.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.

[10] Lay, D. C. (1994). Linear Algebra and Its Applications. Addison-Wesley.