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

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \]  
\[ (eq \ 1) \]

Where \( \alpha, \beta_1 \) 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,

\[ y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdot \cdot \cdot + b_k x_{ik} + e_i \]  
\[ (eq \ 2) \]

Equivalently, in more compact matrix terms:

\[ Y = XB + E \]  
\[ (eq \ 3) \]

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

\[ \hat{y_i} = a + b_1 x_{i1} + b_2 x_{i2} + \cdot \cdot \cdot + b_k x_{ik} + \mu_i \]  
\[ (eq \ 4) \]

Where \( \hat{y_i} \) is dependent variable. \( X_i 's \) are independent variables, and \( \mu_i \) 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 \( X_1 \) (called as a predictor variable) and \( X_2 \) and that is represented by the equation that is we have

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \]  
\[ (eq \ 1) \]

Where \( \alpha, \beta_1 \) and \( \beta_2 \) are regression coefficients

The regression coefficient \( \alpha, \beta_1 \), & \( \beta_2 \) 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 \( X_i, Xb \) and \( Y_i \)
Step 2: Calculate the summation of the variable $X_i, X_b, Y_i$

Step 3: Calculate the product of summation terms ($\sum x_1^*x, \sum x_1^*x_2, \sum x_2^*y, \sum x_2^*x_2, \sum x_1^*y$)

Step 4: Solve the equations

\[
\begin{align*}
\sum y &= a_0 + a_1 \sum x_1 + a_2 \sum x_2 \\
\sum x_1 y &= a_0 \sum x_1 + a_1 \sum x_1^*x_1 + a_2 \sum x_1^*x_2 \\
\sum x_2 y &= a_0 \sum x_2 + a_1 \sum x_1^*x_2 + a_2 \sum x_2^*x_2
\end{align*}
\]

Step 5: Now calculate the value of $a_0, a_1, a_2$ which is calculated by the inverse of a matrix

Step 6: Finally obtain the value of response variable $Y$ by knowing the values of $a_0, a_1, a_2$ in the equation $Y = a_0 + a_1 X_1 + a_2 X_2$ of $\beta$ (calculated in step 4), average of $X_i$ and average of $Y_i$

Step 7: Finally substitute the value of regression coefficients $\alpha$ and $\beta$ in the equation $Y = \alpha + \beta X$

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**

<table>
<thead>
<tr>
<th>$X_1$ (Years of experience)</th>
<th>$X_2$ (Working hrs)</th>
<th>$Y$ (Salary in K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

Here $X_1$ is the years of work experience, $X_2$ 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 $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$

We then calculate the summation of the variable $X_1$, $X_2$ and $Y$ the result is stored in a variable

\[
\sum x_1 = 19, \sum x_2 = 26 \text{ and } \sum y = 402
\]

We then calculate the product of summation terms

\[
(\sum x_1^*x_1 = 133, \sum x_1^*x_2 = 180, \sum x_2^*y = 364, \sum x_2^*x_2 = 244, \sum x_1^*y = 268)
\]

In order to solve the equations

\[
\begin{align*}
\sum y &= a_0 + a_1 \sum x_1 + a_2 \sum x_2 \\
\sum x_1 y &= a_0 \sum x_1 + a_1 \sum x_1^*x_1 + a_2 \sum x_1^*x_2 \\
\sum x_2 y &= a_0 \sum x_2 + a_1 \sum x_1^*x_2 + a_2 \sum x_2^*x_2
\end{align*}
\]

We now calculate the value of $a_0 = 10, a_1 = 6, a_2 = -4$ using inverse of a matrix

**USING THE LINEAR EQUATION $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$**

**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.
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 formulated 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 X2. \( \alpha \) and \( \beta \) are linear regression coefficients.

6. REFERENCES

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


