Chapter 06.07

Multivariate Least Squares Fitting

Up until this point, we have considered single predictor variables in the specification of linear regression prediction equation. However, in most practical engineering problems, the independent variables or factors that determine or affect the dependent or the response variable are not often single predictor variables. If multiple independent variables affect the response variable, then the analysis calls for a model different from that used for the single predictor variable. In a situation where more than one independent factor (variable) affects the outcome of a process, a multiple regression model is used. This is referred to as multiple linear regression model or multivariate least squares fitting. Although flexibility is introduced into the regression analysis by the existence of multiple predictor variables, the complexity added by the use of multiple predictor variables makes this approach most suited for computer usage. A simple example problem for which multiple predictor variables may be required is the consideration of factors on which the total miles traveled per gallon of gas by a car depends. Some of the factors that determine gas usage by a car include its speed, its weight and the wind conditions etc. Thus, for its analysis, a multiple regression model is used which is often referred to as multiple linear regression model or multivariate least squares fitting.

Unlike the single-variable analysis, the interpretation of the output of a multivariate least squares fitting is made difficult by the involvement of several predictor variables. Hence, even if the data base is sound and correct model specified, it is not sufficient and correct to merely examine the magnitudes of the estimated coefficients in order to determine which predictor variables most affect the response. In the same vein, it is not sound to ignore the interactions of the predictor variables when considering the influence of any of the parameters. It is obvious from the fore going that modern day computer tools might have solved the computational aspect of the multivariate least squares method, but discerning the implications of the computational result remains a challenge.

The multivariate least squares discussion will be very brief. Consider N observations on a response y, with m regressors x_j, $j = 1,2,3,\ldots,m,$, the multiple linear regression model is written as

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \qquad i = 1,2,\ldots,N \tag{1}$$

In matrix form, we can arrange the data in the following form

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \qquad \hat{\hat{a}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \tag{2}$$

where $\hat{\beta}_j$ are the estimates of the regression coefficients, β_j which can be obtained from the solution of the matrix equation:

$$\hat{\hat{\alpha}} = (X'X)^{-1} X'Y \tag{3}$$

Equation 3 is obtained by setting up the sum of squares of the residuals and differentiating with respect to each of the unknown coefficients. Similar to the single variable regression, the adequacy of the multiple least square regression model can be checked by computing the residuals and checking if they are normally distributed.

1.1 Example

For the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, determine β_j for the data in Table 1

y	x_1	x_2
144	18	52
142	24	40
124	12	40
64	30	48
96	30	32
92	22	16

Table 1: Data for multiple least square regression

1.1.1. Solution

he single variable regression, setting up sum of squares of the residuals,

$$SR_m = \sum_{i=1}^{6} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2 \tag{4}$$

and differentiating with respect to each unknown coefficient arid equating each partial derivative to zero,

$$\frac{\partial SR_m}{\partial \beta_0} = -2\sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0 \tag{5}$$

$$\frac{\partial SR_m}{\partial \beta_1} = -2\sum x_{1i}\left(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}\right) = 0 \tag{6}$$

$$\frac{\partial SR_m}{\partial \beta_2} = -2\sum x_{2i}\left(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}\right) = 0 \tag{7}$$

we obtain the following matrix expression:

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{Bmatrix} \tag{8}$$

i	y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
1	144	18	52	324	2704	936	2592	7488
2	142	24	40	576	1600	960	3408	5680
3	124	12	40	144	1600	480	1488	4960
4	64	30	48	900	2304	1440	1920	3072
5	96	30	32	900	1024	960	2880	3072
6	92	22	16	484	256	352	2024	1472
Σ	662	136	228	3328	9488	5128	14312	25744

Table 2: Computations for example problem

Using the computed data in Table 2 in eqn. 8 we obtain

$$\begin{bmatrix} 6 & 136 & 228 \\ 136 & 3328 & 5128 \\ 228 & 5128 & 9488 \end{bmatrix} \begin{Bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{Bmatrix} = \begin{Bmatrix} 662 \\ 14312 \\ 25744 \end{Bmatrix} \tag{9}$$

Equation 9 is a system of linear algebraic equations and can be solved by any of method suitable for solving simultaneous equations including Gauss elimination or matrix inversion methods, etc. Using matrix inversion method, the solution to eqn. 9 gives $\beta_0 = 150.166, \beta_1 = -2.731, \beta_2 = 0.581$. As the number of predictor variables increase, solving eqn8 becomes more challenging, hence the use of computational software for the multivariate modeling.