Multiple Linear Regression

Multiple regression analysis is an extension of simple regression analysis, as described previously, to applications involving the use of two or more ***independent variables*** (predictors) to estimate the value of the ***dependent variable*** (response variable). In the case of two independent variables, denoted by X1 and X2, the linear algebraic model is

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_I$$

The definitions of the above terms are equivalent to the definitions in previous classes for simple regression analysis, except that more than one independent variable is involved in the present case.

Based on sample data, the linear regression equation for the case of two independent variables is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

The multiple regression equation identifies the best-fitting line based on the method of least squares. In the case of multiple regression analysis, the best-fitting line is a line through n-dimensional space (3-dimensional in the case of two independent variables).

Determining Regression Estimates

The calculations required for determining the values of the parameter estimates in a multiple regression equation and the associated standard error values are quite complex and generally involve matrix algebra. However, computer software, such as `R`, is widely available for carrying out such calculations.

Assumptions

The assumptions of multiple linear regression analysis are similar to those of the simple case involving only one independent variable. For point estimation, the principal assumptions are that

(1) the dependent variable is a random variable,

(2) the relationship between the several independent variables and the one dependent variable is linear. Additional assumptions for statistical inference (estimation or hypothesis testing) are that

(3) the variances of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal,

(4) the conditional distributions of the dependent variable are normally distributed, and

(5) the observed values of the dependent variable are independent of each other. Violation of this assumption is called autocorrelation.