Multiple linear regression

Descriptive data وصف البيانات

Data typically consist of n observations on the dependent variable y and m independent variables $(X_1, X_2, \dots, X_m)$. The data is arranged as in the following table:

رقم المشاهدة	y_i	X_1	X_2	…	X_m
1	y_1	x_{11}	x_{12}	⋯	x_{1m}
2	y_2	x_{21}	x_{22}	⋯	x_{2m}
3	y_3	x_{31}	x_{32}	⋯	x_{3m}
⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	⋯	x_{nm}

Graphical representation

The graph of a multiple linear regression equation is a dimensional surface (m+1) Where m is the number of independent variables. If there are two independent variables (m=2). The appropriate surface for the data is a three-dimensional surface that is best represented by points:

$$(x_{11}, x_{12}, y_1), \left(x_{21,}x_{22}, y_2\right), \dots, (x_{n1}, x_{n2}, y_n)$$

Whereas (x_{i1}, x_{i2}, y_i) represent values X_1, X_2, y_i For observation i of the sample.

Therefore, the multiple regression equation in this case is a surface that represents the average of the y values for different values $X_{1,}X_2$, that is, the model

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$

It can be represented by a three-dimensional surface, as shown in Figure (1).

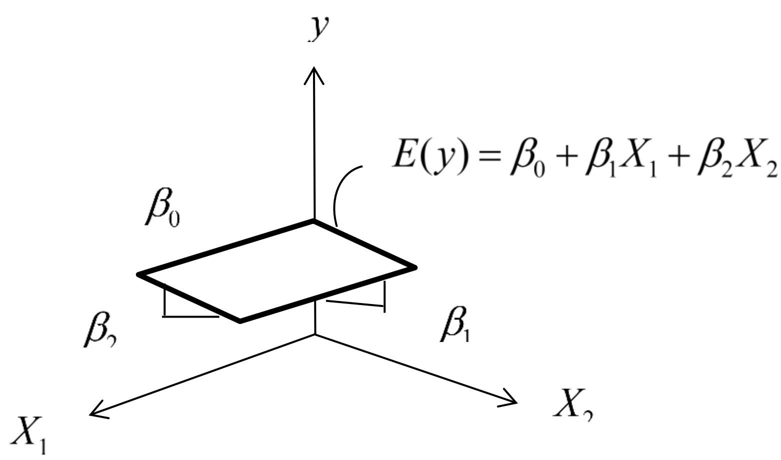

Figure (1): Graphical representation of the three−dimensional regression equation

Mathematical model

The significant relationship between the variable y and the independent variables x_{is} in multiple regression analysis can be expressed as a linear function as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + e_i, i = 1,2, \ldots, n$$

That is:

$$y = X\beta + e$$

whereas:

y_i : is the value of the dependent variable or response variable.

$x_{i1}, x_{i2}, \ldots, x_{im}$: they are fixed values of m of independent variables.

$\beta_0, \beta_1, \ldots, \beta_m$: they are constants or parameters of the regression equation.

whereas:

β_0 : It is the location of the intersection of the slope plane with the y-axis, and β_1 gives the average response when the values of x_{is} are equal to zero.

β_i It is the partial regression coefficient of y on X_i. When the rest of the independent variables are held constant. It also represents the amount of change in y for a one-unit increase in X_i when the rest of the independent variables are constant.

e_i it is a random or residual error.

The previous equation is called multiple because it contains more than one independent variable, and it is linear because each of the parameters $\beta_0, \beta_1, \ldots, \beta_m$ as well as the independent variables $X_1, X_2, \ldots, X_m$ It is of first order, meaning it has a power equal to one.

Analysis assumptions

1. The dependent variable y is a random variable and its values are statistically independent from one another and distributed normally with an arithmetic mean of $\mu_{y/x_1,\ldots,x_m}$ and variance $\sigma^2_{y/x_1,\ldots,x_m} = \sigma^2$ that is, the average y is a linear function

$$\mu_{y/X_1,\ldots,X_m} = E(y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

and the variance of y is

$$\sigma^2_{y/X_1,\ldots,X_m} = \sigma^2_{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_m X_m} = \sigma^2_e = \sigma^2_I$$

This property is called homoscedasticity, i.e. homogeneity of error

2. The error term e_i is a random error distributed normally with an arithmetic mean of zero, that is $E(e) = 0$, and a variance of the amount $\sigma^2_e = \sigma^2_I$ that is $E(e'e) = \sigma^2_I$.

The covariance between is $\mathrm{cov}(e_i e_i^{'}) = 0$ because we assume that there is no correlation between the values e_i.

3. There is no specific or complete linear relationship between the independent variables.