Multiple Linear Regression

Dr. Kiah Wah Ong

The Need for Several Predictor Variables

- How's a person's waist size and height relate to his/her % body fat.
- How's a student's SAT verbal and math scores relate to the graduation rate among the students who took the exam.
- How's the concentration of acetic acid, hydrogen sulfide, and lactic acid relate to the taste of cheese.

Multiple Linear Regression Model

Let x_1, x_2 and y be the students' SAT verbal score, math score and graduation rate, respectively. Then the relationship between x_1, x_2 and y might be describe by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

This is a **multiple linear regression model** with two regressors (predictors).

The term linear is used because the equation above is a linear function of the unknown parameters β_0, β_1 and β_2.

Multiple Linear Regression Model

Example

Suppose we have

$$y = 15 + 3x_1 + 4x_2 + \epsilon$$

then because of $E(\epsilon) = 0$. We obtain $E(y) = 15 + 3x_1 + 4x_2$.

- The parameter $\beta_0 = 15$ is the "y" intercept of the regression plane.
- If the scope of the data includes $x_1 = x_2 = 0$, then $\beta_0 = 15$ represents the mean response $E(y)$ at $x_1 = x_2 = 0$.
- The parameters β_1 indicates the change of in the mean response $E(y)$ per unit increase in x_1 when x_2 is held constant.
- Likewise, β_2 indicates the change in mean response per unit increase in x_2 when x_1 is held constant.

Multiple Linear Regression Model

To see this, suppose x_2 is held constant at $x_2 = 2$. The regression function is then given by

$$E(y) = 15 + 3x_1 + 4(2) = 23 + 3x_1$$

This is a straight line with slope $\beta_1 = 3$. Hence $\beta_1 = 3$ indicates that the mean response $E(y)$ increases by 3 unit when x_1 increases by 1 unit, while x_2 being held constant.

Multiple Linear Regression Model

Suppose you collected n data

$$(x_{11}, x_{12}, y_1), \cdots, (x_{i1}, x_{i2}, y_i), \cdots, (x_{n1}, x_{n2}, y_n).$$

then we can write the regression model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \cdots, n$$

where we assume

- x_{i1} and x_{i2} are known constants,
- ϵ_i are independent $\mathrm{N}(0, \sigma^2)$.

Multiple Linear Regression Model

In general, the response y maybe related to k predictor variables, that is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

The regression coefficients β_j represents the expected change in the response y per unit change in x_j when all of the remaining regressor variables x_i, $i \neq j$ are held constant.

Suppose you have n data

$$(x_{11}, x_{12}, \cdots, x_{1k}, y_1), \cdots, (x_{i1}, x_{i2}, \cdots, x_{ik}, y_i), \cdots, (x_{n1}, x_{n2}, \cdots, x_{nk}, y_n)$$

then we write

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \cdots, n$$

Multiple Linear Regression Model in Matrix Form

Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

then

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \cdots, n$$

can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The Least Squares Method

From

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \cdots, n$$

The least square function is

$$S(\beta_0, \beta_1, \cdots, \beta_k) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

We want to minimize S with respect to $\beta_0, \beta_1, \cdots, \beta_k$.