Multiple linear regression

Recall that a multiple linear regression model is as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

For calculation purposes, it is better to use matrix notation. Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{1p} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Then

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

It can be shown that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Example

The data in the following table relate the suicide rate to the population size and the divorce rate at eight different locations.

Location	Population in Thousands	Divorce Rate per 100,000	Suicide Rate per 100,000
Akron, OH	679	30.4	11.6
Anaheim, CA	1,420	34.1	16.1
Buffalo, NY	1,349	17.2	9.3
Austin, TX	296	26.8	9.1
Chicago, IL	6,975	29.1	8.4
Columbia, SC	323	18.7	7.7
Detroit, MI	4,200	32.6	11.3
Gary, IN	633	32.5	8.4

Let us fit a model of the form

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

where Y is the suicide rate, x_1 is the population and x_2 is the divorce rate.

The estimated regression line is

$$Y = 3.5073 - 0.0002\, x_1 + 0.2609\, x_2 + e$$

The population does not play a major role in predicting the suicide rate (at least when the divorce rate is also given).

Polynomial regression

Sometimes we try to fit to the data set a functional relationship of the form

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + \epsilon$$

The parameters are estimated by the method of Least Squares. Solution of Least Squares equation is easier if a multiple linear regression set-up (discussed above) is used after treating x, x^2 etc. as different variables.

In fitting a polynomial to a set of data pairs, it is often possible to determine the necessary degree of the polynomial by a study of the scatter diagram. We Should always use the lowest possible degree that appears to adequately describe the data.

Logistic regression model for binary output data

When the response (dependent variable) is a categorical variable with two categories called "success" and "failure", we try to explain the probability of success $p(x)$ with the help of the covariate x. When $p(x)$ is of the form

$$p(x) = \frac{\exp(a + bx)}{1 + \exp(a + bx)} = \frac{1}{1 + \exp\left(-(a + bx)\right)}$$

then the model is called a logistic regression model. If $b > 0$, $p(x)$ is an increasing function that converges to 1 as $x \to \infty$. If $b < 0$, $p(x)$ is a decreasing function that converges to 0 as $x \to \infty$. The curve is shaped like an S.

If we let $o(x)$ be the odds for success when the value of the covariate is x, then

$$o(x) = \frac{p(x)}{1 - p(x)} = \exp\,(a + bx)$$

so that the log-odds or "logit" of success is given by

$$\log\left(o(x)\right) = a + bx$$

Iterative algorithm is required to estimate the parameters a and b.

Example

The table below shows the number of hours each of 20 students spent studying weekly, and whether they passed (1) or failed (0). (Source: Wikipedia)

Hours	Whether passed
0.50	0
0.75	0
1.00	0
1.25	0
1.50	0
1.75	0
1.75	1
2.00	0
2.25	1
2.50	0
2.75	1
3.00	0
3.25	1
3.50	0
4.00	1
4.25	1
4.50	1
4.75	1
5.00	1
5.50	1

The results are: $\hat{a} = -4.08$ and $\hat{b} = 1.51$. For a student who studies 2 hours per week, the odds of passing is

$$o(2) = \frac{p(2)}{1 - p(2)} \approx \exp(-4.08 + 1.51 \times 2) = 0.3465$$

For 1 hour increase in study time, how much does the odds increase?

The probability of passing is

$$p(2) \approx \frac{1}{1 + \exp\big(-(-4.08 + 1.51 \times 2)\big)} = 0.2573$$