Additional Note on Linear Regression

2018年 前期講義資料 (作成者:大塚 淳)

Suppose we have measured two variables X and Y (say height and weight). We consider drawing a *regression line* that characterizes their relationship. Since the line is expressed as $Y = bX + a$, our job is to determine *coefficients* b and a.

To achieve this goal we calculate the deviance of a line from the actual data. For each data point i, we can measure the difference or "error" between actual y_i and the "prediction" $bx_i + a$ by

$$l_i = \{y_i - (bx_i + a)\}^2 \tag{1}$$

The square is required because we are interested just in the distance from the line to the data point, not a direction (whether the difference is positive or negative). If we have n data, we can sum up (1) to obtain

$$L = \sum_i^n \{y_i - (bx_i + a)\}^2 \tag{2}$$

The best line (i.e., a and b) minimizes this *sum of squared errors*. This is a quadratic function of a and b, so to minimize L we take its partial derivative with respect to a and b and set them to zero:

$$\begin{aligned}\frac{\partial L}{\partial b} &= \sum y_i - b\sum x_i - na = 0\\ \frac{\partial L}{\partial b} &= \sum x_i y_i - b\sum x_i^2 - a\sum x_i = 0\end{aligned}$$

By solving this system of equation yields

$$\begin{aligned}b &= \tfrac{\sum x_i y_i - n\bar{X}\bar{Y}}{\sum x_i^2 - n\bar{X}^2},\\ a &= \bar{Y} - b\bar{X}.\end{aligned}$$

Check by yourself that this b equals the regression coefficient $\mathrm{Cov}(X, Y)/\,\mathrm{Var}(Y)$ introduced in the note. This procedure is called *the method of least squares.*