

Statistics and Dot Products

Nobuyuki TOSE

November 29, 2016

Bivariate Data

We are given a set of bivariate data:

x	y
x_1	y_1
$\vdots$	$\vdots$
x_n	y_n

In this situation, first consider the arithmetic mean of x and y :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Next consider the variance of x and y

$$V(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad V(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Bivariate Data (2)

We also consider the covariance of x and y

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Now we associate two vectors to the data:

$$\vec{x} = \frac{1}{\sqrt{n}} \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}, \quad \vec{y} = \frac{1}{\sqrt{n}} \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$$

to express the variance and the covariance by

$$V(x) = \|\vec{x}\|^2, \quad V(y) = \|\vec{y}\|^2, \quad C_{xy} = (\vec{x}, \vec{y})$$

Moreover the correlation coefficient of x and y is expressed by

$$\rho_{xy} := \frac{C_{xy}}{\sqrt{V(x)}\sqrt{V(y)}} = \frac{(\vec{x}, \vec{y})}{\|\vec{x}\| \cdot \|\vec{y}\|}$$

Bivariate Data (3)—Cauchy-Schwartz inequality

Recall the Cauchy-Schwartz inequality:

$$|(\vec{a}, \vec{b})| \leq \|\vec{a}\| \cdot \|\vec{b}\| \quad (\vec{a}, \vec{b} \in \mathbf{R}^n)$$

It follows that

$$-1 \leq \rho_{xy} \leq 1$$

We introduce a model to figure out what happens when $\rho_{xy} \rightarrow \pm 1$.

$$y = ax + b$$

We try to fit the model to the given data in the following way.

Bivariate Data (4)—Regression Line

First introduce a variable

$$\varepsilon = y - (ax + b), \quad \text{namely} \quad \varepsilon_j = y_j - (ax_j + b) \quad (j = 1, \dots, n)$$

The variable ε means the error of the data based on the model and we set up the coefficients a and b so that

- **(i)** $\bar{\varepsilon} = 0$,
- **(ii)** $V(\varepsilon)$ is minimized.

Bivariate Data (5)—Regression Line

The condition (i) is equivalent to

$$\bar{\varepsilon} = \bar{y} - a\bar{x} - b = 0$$

To look into the condition (ii), we introduce a vector $\vec{\varepsilon}$ by

$$\vec{\varepsilon} = \frac{1}{\sqrt{n}} \begin{pmatrix} \varepsilon_1 - \bar{\varepsilon} \\ \vdots \\ \varepsilon_n - \bar{\varepsilon} \end{pmatrix}$$

Moreover we have

$$\varepsilon_j - \bar{\varepsilon} = (y_j - ax_j - b) - (\bar{y} - a\bar{x} - b) = (y_j - \bar{y}) - a(x_j - \bar{x})$$

Then it follows that

$$\vec{\varepsilon} = \vec{y} - a\vec{x}$$

Bivariate Data (6)—Regression Line

Now we can minimize $V(\varepsilon)$ by

$$\begin{aligned} V(\varepsilon) &= \|\vec{y} - a\vec{x}\|^2 \\ &= \|\vec{y}\|^2 - 2a(\vec{x}, \vec{y}) + a^2\|\vec{x}\|^2 \\ &= \|\vec{x}\|^2 \left(a - \frac{(\vec{x}, \vec{y})}{\|\vec{x}\|^2} \right)^2 + \|\vec{y}\|^2 - \frac{(\vec{x}, \vec{y})^2}{\|\vec{x}\|^2} \\ &\geq \|\vec{y}\|^2 - \frac{(\vec{x}, \vec{y})^2}{\|\vec{x}\|^2} \end{aligned}$$

The equality holds at the end of the line when

$$a = \frac{(\vec{x}, \vec{y})}{\|\vec{x}\|^2}$$

Bivariate Data (7)—Regression Line

Regression line

The line $y = ax + b$ is called the regression line when

$$a = \frac{(\vec{x}, \vec{y})}{\|\vec{x}\|^2} = \frac{C_{xy}}{V(x)}, \quad b = \bar{y} - a\bar{x}$$

In the case of regression line we have

$$\begin{aligned} V(\varepsilon) &= \|\vec{y}\|^2 \left(1 - \frac{(\vec{x}, \vec{y})^2}{\|\vec{y}\|^2 \cdot \|\vec{x}\|^2} \right) \\ &= V(y) (1 - \rho_{xy}^2) \end{aligned}$$

Moreover since $\bar{\varepsilon} = 0$, we have

$$V(\varepsilon) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$$

Bivariate Data (7)—Regression Line

The sum $\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2$ is called the residual square and it approaches to 0 when $\rho_{xy} \rightarrow \pm 1$