IV-3-3. Simple linear regression and correlation

When a data has multiple variables fluctuate having relation in a dataset, we want to know the relation among them. The work for expressing the relation by mathematical formula is regression. When the variances are only two and the relation can be expressed by primary linear expression, the work is simple linear regression.

When we watch following scatter diagram, we feel relation between x and y. Simple linear regression is to draw a line expressing the relation between x and y in the scatter diagram.

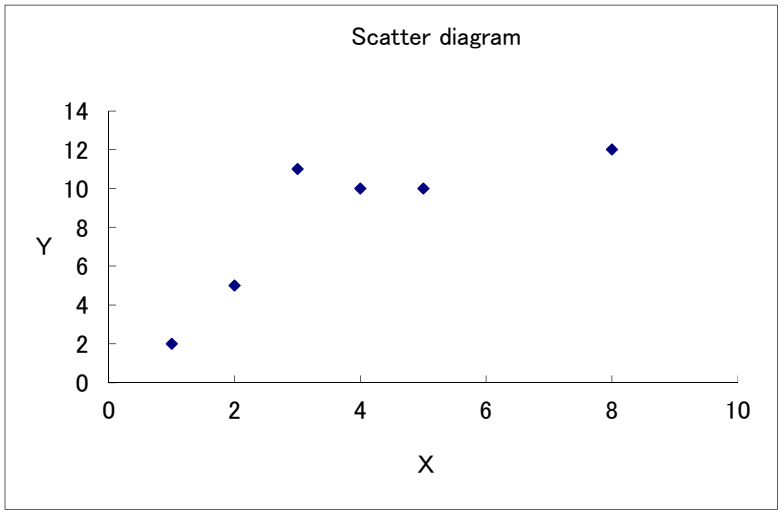


Fig. 36. Scatter diagram of sample dataset.

Table 35 is the dataset of the diagram. No is the number for identification of data pair.

Table 35. Dataset of the diagram

No	X	Y
1	1	2
2	2	5
3	3	11
4	5	10
5	8	12
6	4	10

We discuss the method to confirm adequacy of our feeling that there is a relation between x and y and how we can draw a line expressing the relation.

In previous chapters, we discussed variance of difference and sum. This is a discussion of variance of products (covariance). However, discussion of covariance is

abstract and complicate. When we discuss the method of regression, we can easily accept the concept of covariance.

We assume the regression line as follow.

$$y = bx + a$$

This equation means that we can estimate y when we give the value of x. We express the estimated y from x_i as $\bar{y}$

$$\bar{y}_i = bx_i + a$$

$$y_i - \bar{y}_i = r_i$$

$$y_i = \bar{y}_i + r_i$$

This means that y_i can be separated to two parts. One is a portion explained by x_i. That is $\bar{y}_i$, the other portion is residual which is not explained by x_i. That is r_i.

When we express x_i and y_i by average and deviation

$$x_i = M_x + e_{x_i}$$

$$y_i = M_y + e_{y_i}$$

M: average

e: distance from average

We put these equations to the relation of $y_i - \bar{y}_i = r_i$

$$M_y + e_{y_i} - a - b(M_x + e_{x_i}) = r_i$$

$$M_y - bM_x - a + e_{y_i} - be_{x_i} = r_i$$

When we assume $y = bx + a$ at mean

$$M_y - bM_x - a = 0$$

$$e_{y_i} - be_{x_i} = r_i$$

We consider SS of r

$$SS_r = \sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} \left(e_{y_i} - be_{x_i}\right)^2$$

$$= \sum_{i=1}^{n} {e_{y_i}}^2 - 2b \sum_{i=1}^{n} e_{x_i} e_{y_i} + b^2 \sum_{i=1}^{n} {e_{x_i}}^2$$

First and third term are SS of each variables and coefficient. Second term is new for us. The value obtained by division of $\sum_{i=1}^{n} e_{x_i} e_{y_i}$ by degree of freedom is covariance. Source of covariance is relation between x and y.

For the confirmation, we assume there is no relation between x and y.

$$\sum_{i=1}^{n} e_{x_i} e_{y_i} = 0$$

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} {e_{y_i}}^2 - 2b \sum_{i=1}^{n} e_{x_i} e_{y_i} + b^2 \sum_{i=1}^{n} {e_{x_i}}^2$$

$$= \sum_{i=1}^{n} {e_{y_i}}^2 + \sum_{i=1}^{n} \left(be_{x_i}\right)^2$$

$$\mathrm{SS}_r = SS_y + SS_{bx}$$

We can explain the variance of residual only by variance of x and y.

Conversely, when y can be explained only by the relation $y = bx + a$, in other word, when all plots in scatter diagram are on a line

$$e_{y_i} = be_{x_i}$$

$$SS_r = \sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} \left(e_{y_i} - be_{x_i}\right)^2 = 0$$

This is very important theoretically, and we will discuss the meaning of this fact in later part, though we discuss the method to obtain optimum value of b at first.

There several methods. Today most likelihood method becoming common. This the method to maximize the probability. However, this is basic text book for explanation analysis of covariance. We select least square method.

As a question in the test of math class of high school, the question is "Obtain the optimal value of b to minimize left side following equation"

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} {e_{y_i}}^2 - 2b \sum_{i=1}^{n} e_{x_i} e_{y_i} + b^2 \sum_{i=1}^{n} {e_{x_i}}^2$$

This is question of minimization of quadric function. There various possible methods in minimization and maximization of quadric function. Easiest way is to use "Excel solver". Generally, calculation of extremum using differential is used. Philosophy of the text book is "understandable to a lot of people and minimize necessary back ground knowledge". Here the author uses the method for minimization of quadric function which he learned in junior high school.

For simplification,

$$SS_{xy} = \sum_{i=1}^{n} e_{x_i} e_{y_i}$$

Then

$$\sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} {e_{y_i}}^2 - 2b\sum_{i=1}^{n} e_{x_i}e_{y_i} + b^2\sum_{i=1}^{n} {e_{x_i}}^2$$

$$= SS_y \pm 2bSS_{xy} + b^2SS_x$$

$$= SS_x\left(b - \frac{SS_{xy}}{SS_x}\right)^2 + SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

$$\left(b - \frac{SS_{xy}}{SS_x}\right)^2 \geq 0$$

When $b - \frac{SS_{xy}}{SS_x} = 0$

$$\sum_{i=1}^{n} {r_i}^2 = SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

When

$$b = \frac{SS_{xy}}{SS_x}$$

We can obtain minimum value of

$$SS_y - \frac{{SS_{xy}}^2}{SS_x}$$

Value of a is obtainable from b using the relation between mean x and y.
Then we discuss the statistical fluctuation of the estimated values. Sum of square of y is originally SS_y, and $SS_y - \frac{{SS_{xy}}^2}{SS_x}$ is sum of value obtained by optimizing value of b. So, the difference of both SS means obtainable effect by optimizing of b.

$$SS_y - \left(SS_y - \frac{{SS_{xy}}^2}{SS_x}\right) = \frac{{SS_{xy}}^2}{SS_x}$$

When we divide this value by SS_y, we can obtain ration of effect of optimizing of b in total variance. This is contribution ratio of b. Generally, the symbol of contribution ratio is r^2.

$$r^2 = \frac{{SS_{xy}}^2}{SS_xSS_y}$$

r^2: contribution ratio

Formula 44

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

r: coeffcient of correlation

From other perspective, $SS_y - \frac{{SS_{xy}}^2}{SS_x}$ is a SS. We can obtain a variance dividing the value by degree of freedom. The value is obtained two variables and mean of dataset which has n pairs of data. Degree of freedom is $n-2$.

$$\sigma_y{}^2 = \frac{1}{n-2}\left(SS_y - \frac{{SS_{xy}}^2}{SS_x}\right) = \frac{1-r^2}{n-2}SS_y$$

This variance is secondary moment of y around estimated line (regression line). If we discuss the possibility of mean of parent population of y, the standard error is $\frac{\sigma_y}{\sqrt{n}}$ and we can obtain observed t value as follow.

$$\mathrm{t}=\frac{M_y}{\frac{\sigma_y}{\sqrt{n}}} = \sqrt{n}\frac{\frac{\sum_{i=1}^{n} y_j}{n}}{\sigma_y} = \frac{\sum_{i=1}^{n} y_j}{\sqrt{n}\sigma_y}$$

Then we can discuss the possibility comparing observed t value with threshold at df $= n-2$ in table of T distribution.

However, this test has little practical meaning. In many case, it is obvious whether value of y is 0 or not 0. In the case when we need to discuss the possibility, we can discuss by getting standard error of y directly from original data. No one want to do such a complicated calculation. For argument, it may have meaning, when we use σ_y for discussion of fluctuation of estimated y at specified x. For example, we consider fluctuation of y at $x=0$. This is estimation of y intercept.

Regression line is $y = bx + a$

Intercept is $a = y - bx$

So instinctually

$$a = M_y - bM_x$$

For confirmation,

$$\overline{y_i} = a + bx_i$$
$$y_i - \overline{y_i} = r_i$$
$$y_i - (a + bx_i) = r_i$$
$$y_i = r_i + (a + bx_i)$$
$$y_i = e_{y_i} - be_{x_{i\,i}} + b\left(e_{x_i} + M_x\right) + a$$
$$y_i = e_{y_i} + bM_x + a$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} e_{y_i} + n(bM_x + a)$$

$$\sum_{i=1}^{n} y_i = n(bM_x + a)$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = bM_x + a$$

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - bM_x = M_y - bM_x$$

Q.E.D

The value of a is estimated value, and it fluctuates around the true value of a having secondary moment. The secondary moment is standard error of y. Standard error of y is $\frac{\sigma_y}{\sqrt{n}}$

We can calculate observed t as follow.

$$t = \frac{M_y - bM_x}{\frac{\sigma_y}{\sqrt{n}}}$$

$$\mathrm{df} = n - 2$$

However, this is still incomplete discussion, because we did not consider the fluctuation of b. The value of b also fluctuates independently with a.

We also need to discuss fluctuation of b. This has practical meaning. Many people want to know whether there is relation between x and y. This question is the same with the question whether $r = 0$ or $r \neq 0$. This is test of correlation. The null hypothesis is $r = 0$. The meaning of this null hypothesis is same as $SS_{xy} = 0$ and $b = 0$.

Because,

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}}$$

$$b = \frac{SS_{xy}}{SS_x}$$

$$SS_x \neq 0$$

$$SS_y \neq 0$$

Expected value of b is $\frac{SS_{xy}}{SS_x}$. This means that the difference between $b = 0$ and $b = \frac{SS_{xy}}{SS_x}$ is $\frac{SS_{xy}}{SS_x}$. We can obtain observed t value dividing the difference by secondary moment of expected value.

We should consider the calculation of secondary moment around b. We obtained σ^2 as secondary moment around the mean of parent population by summing up square of difference between data value and estimated mean in the discussion of variance. We can use similar method for calculation of secondary moment around true b. Observed b_i obtained from x_i and y_i is

$$b_i = \frac{e_{y_i}}{e_{x_i}}$$

Estimated b from data set is

$$b = \frac{SS_{xy}}{SS_x}$$

We discuss the difference between b_i and b.

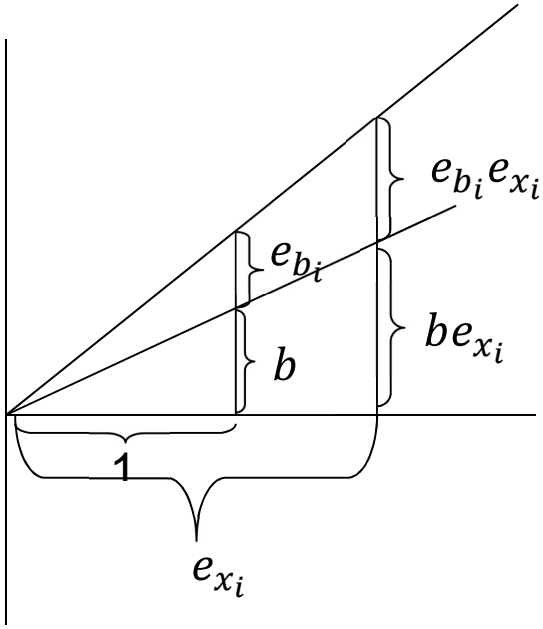


Fig. 36. Relation between b, e_{x_i} and e_{y_i}

The slope of regression formula is increase of y, when x increase 1. The value of e_{b_i} is deviation of b. The slope $b_i = \frac{e_{y_i}}{e_{x_i}}$ is magnification of e_{b_i} when e_{x_i} increase from1 to e_{x_i}. So, deviation of y_i from estimated from y using $y_i = bx_i$ is

$$e_{y_i} = e_{b_i}e_{x_i}$$

$$e_{b_i} = b_i - b = \frac{e_{y_i}}{e_{x_i}} - \frac{SS_{xy}}{SS_x} = \frac{1}{e_{x_i}}\left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)$$

$$e_{b_i}{}^2 = \frac{1}{e_{x_i}{}^2}\left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2$$

This is square of difference of between observed value and estimated value. In the discussion of secondary moment of average. We multiply the probability to the value and sum up the products to obtain secondary moment. We need to consider the probability of each $e_{b_i}{}^2$. One possible idea is to consider that the probabilities are the same among data. This is adequate. However, we are discussing standardized data dividing by e_{X_i}. There is an inverse relation between e_{x_i} and e_{b_i}. Fluctuations of data closer to mean makes wider fluctuation of b, and fluctuation of data further to mean makes smaller fluctuation of b. We need weighting by the distance. In this case the data is square. We weight the data by $\frac{e_{x_i}{}^2}{\sum_{i=1}^n e_{x_i}{}^2}$ to obtain SS_b.

$$SS_b = \sum_{i=1}^n \frac{e_{x_i}{}^2}{\sum_{i=1}^n e_{x_i}{}^2} e_{b_i}{}^2 = \frac{1}{\sum_{i=1}^n e_{x_i}{}^2} \sum_{i=1}^n e_{x_i}{}^2 e_{b_i}{}^2 = \frac{1}{SS_x} \sum_{i=1}^n \left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2$$

$$\because \sum_{i=1}^n e_{x_i}{}^2 = SS_x$$

$$e_{b_i}{}^2 = \frac{1}{e_{x_i}{}^2}\left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2$$

Expansion of $\sum_{i=1}^n \left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2$。

$$\sum_{i=1}^n \left(e_{y_i} - \frac{SS_{xy}e_{x_i}}{SS_x}\right)^2 = \sum_{i=1}^n e_{y_i}{}^2 - 2\frac{SS_{xy}}{SS_x}\sum_{i=1}^n e_{x_i}e_{y_i} + \frac{SS_{xy}{}^2}{SS_x{}^2}\sum_{i=1}^n e_{x_i}{}^2$$

$$= SS_y - \frac{SS_{xy}{}^2}{SS_x}$$

$$\because \sum_{i=1}^n e_{y_i}{}^2 = SS_y\ ,\ \sum_{i=1}^n e_{x_i}e_{y_i} = SS_{xy}\ ,\ \sum_{i=1}^n e_{x_i}{}^2 = SS_x$$

$$= (n-2)\sigma_y{}^2 = (1-r^2)SS_y$$

$$\because \quad \sigma_y{}^2 = \frac{1}{n-2}\left(SS_y - \frac{SS_{xy}{}^2}{SS_x}\right) = \frac{1-r^2}{n-2}SS_y$$

$$SS_b = \frac{SS_y}{SS_x}(1-r^2)$$

$$\sigma_b{}^2 = \frac{(1-r^2)SS_y}{(n-2)SS_x}$$

$$\sigma_b = \sqrt{\frac{(1-r^2)SS_y}{(n-2)SS_x}}$$

$\sigma_b{}^2$is secondary moment around true value and standard error of b. The difference between b=0 and $b_{estimated} = \frac{SS_{xy}}{SS_x}$ is $\frac{SS_{xy}}{SS_x}$

We can obtain observed t by

$$\mathrm{t} = \frac{|0 - b_{estimated}|}{\sigma_b} = \frac{SS_{xy}}{SS_x}\sqrt{\frac{(n-2)SS_x}{(1-r^2)SS_y}} = \frac{SS_{xy}}{\sqrt{SS_xSS_y}}\sqrt{\frac{n-2}{1-r^2}} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

$$\mathrm{df} = n-2$$

This is a test of existence of correlation. This test has practical meaning.

Here, we go back to the discussion of secondary moment of y intercept.

Ratio of fluctuation between fluctuation in y-axis and x-axis is $\frac{\sigma_y}{\sqrt{SS_x}}$.

As explained in the discussion of secondary moment of b, impact of fluctuation of b increase with the distance from the mean. Intercept of y is value of y at $x = 0$. Distance from M_x to is M_x

The standard deviation of y sourcing fluctuation in x-axis at $x = 0$ is

$$M_x\frac{\sigma_y}{\sqrt{SS_x}}$$

Variance is

$$M_x{}^2\frac{\sigma_y{}^2}{SS_x}$$

Originally, the fluctuation at y intercept has fluctuation sourcing fluctuation in y-axis.

Total fluctuation around y-intercept is sum of both fluctuations.

The secondary moment of y intercept is

$$\sigma_{intercept}{}^2 = \frac{\sigma_y{}^2}{n} + M_x{}^2\frac{\sigma_y{}^2}{SS_x} = \sigma_y{}^2\left(\frac{1}{n} + \frac{M_x{}^2}{SS_x}\right)$$

Standard error of y intercept is

$$\sigma_{intercept} = \sigma_y\sqrt{\frac{1}{n} + \frac{M_x{}^2}{SS_x}}$$

We can obtain observed t value of intercept by dividing estimated intercept by $\sigma_{intercept}$.

Similarly, we can obtain observed t at given x, and we can consider various test including the significance of difference of slopes between tow regression lines.

Weakness of regression and correlation analysis

1. Impact of outlier

Sometimes dataset include outlier, the data extremely far from other data. Following is an example of dataset including outlier.

Table 37. Dataset including outlier

X	Y
2	1
3	5
5	5
1	3
5	1
20	22

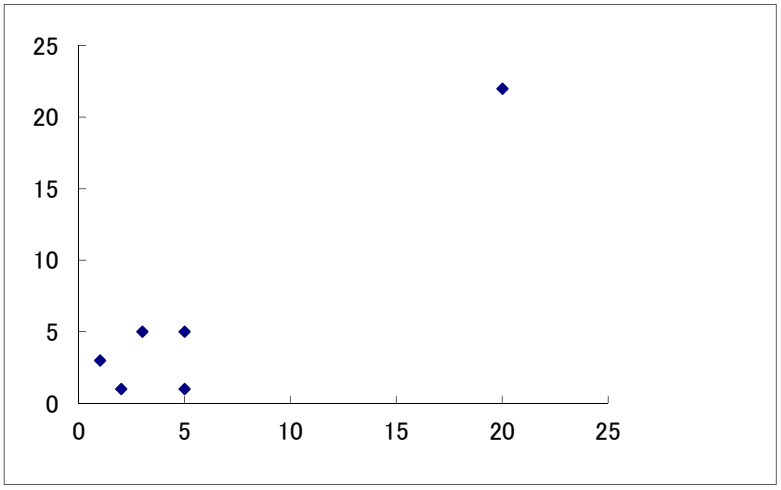


Fig. 37. Scatter diagram of the dataset including outlier

It looks as if there is a correlation between x and y. When we implement analysis of correlation, $r = 0.956$, and the correlation is significant ($p \leq 0.05$). However, this dataset includes an outlier in upper right of scatter diagram. When we remove the outlier, $r = 0.140$. The significance is produced by the existence of the outlier. We should consider, whether we have to remove the outlier or have to include outlier. There are several methods concerning removal of outlier. However, the analyzers

should consider the question by themselves finally. Because analyzer can obtain the background of existence of outlier. Another reason of strong impact of outlier is least square method. When we use methods to minimize the error, outliers have strong impact on the result of analysis. Recently, many people use maximum likelihood method. Maximum likelihood method is better than least square method concerning this issue, though it is not complete solution of this issue and the procedure is complicate.

2. Correlation and causal link.

Regression is used to clarify the relation by regression line, when the existence of correlation is recognized beforehand. Analysis of correlation is used for the confirmation of existence of relation. Even when we confirm the existence of correlation, we cannot say there is a causal link. For example, laundries are well dried, and many people go outside in fine day. There is correlation between dryness of laundries and number of people out of house, though there is no causal link between dryness of laundries and number of people out of house. In order to say about causal links, we have to clarify the mechanism using scientific method in each discipline.

Geometric meaning of correlation coefficient.

Purpose of simple linear regression and analysis of correlation is different. However, both analysis use coefficient of correlation as key element of analysis. For the expansion of our discussion to multiple variance analysis. It is important to understand geometric meaning of correlation coefficient.

Our spatial cognition of data scatter in simple liner regression is planar map. The map is called scatter diagram as shown in figure 38.

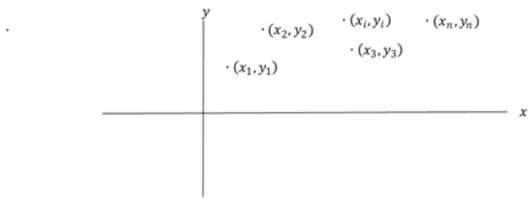


Fig. 38. Scatter diagram on two axes plane

When we shift our perspective, we can recognize the scatter in diagram in n dimension space. We cannot draw such scatter diagram in two-dimension plane. However, in the case when number of data pair is only 3 it can be draw as follow.

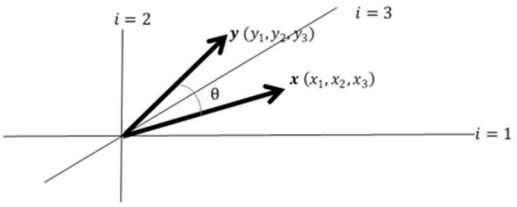


Fig.39.Expression or dataset by vector space.

The angle of vector $\boldsymbol{x}$ and vector $\boldsymbol{y}$ is θ

Definition of inner product $\boldsymbol{x} \cdot \boldsymbol{y}$ in multiple space is as follow.

$$\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i = SS_{xy}$$

The definition of inner product $\boldsymbol{x} \cdot \boldsymbol{y}$ in vector space is

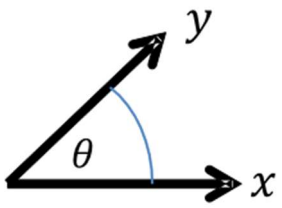
$$\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}| \cdot |\boldsymbol{y}| \cos\theta = \sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2} \cos\theta = \sqrt{SS_x} \sqrt{SS_y} \cos\theta$$

$$SS_{xy} = \sqrt{SS_x} \sqrt{SS_y} \cos\theta$$

$$\cos\theta = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}} = r$$

.

θ: angle of two vectors



We can say the geometric meaning of coefficent of correlation is $\cos\theta$. This is important in multiple variance analysis.

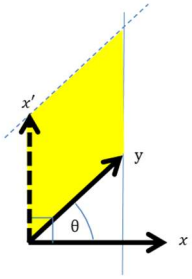


Fig. 40. Geometric meaning of inner product

In the author's perception, inner product is area of yellow parallelogram in figure 40. Coefficient of correlation is obtained by dividing the area by product of length of two vectors. The meaning of the value is degree of squash of the yellow parallelogram. When the parallelogram is completely squashed, the space is 0.

$$r = 0$$

And the vectors are orthogonal and independent each other.

This is also geometrical proof of Cauchy-Schwarz inequation.

$$(\alpha^2 + \beta^2 + \gamma^2)(\delta^2 + \varepsilon^2 + \zeta^2) \geq (\alpha\delta + \beta\varepsilon + \gamma\,\zeta)^2$$

Left side of inequation is product of squares of length of tow vector and right side is square of inner product. We express the vectors by unit vectors as follow.

$$\boldsymbol{a} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

$$\boldsymbol{b} = \begin{pmatrix} \delta \\ \varepsilon \\ \zeta \end{pmatrix}$$

$$|\boldsymbol{a}||\boldsymbol{b}| \cos\theta = \boldsymbol{a} \cdot \boldsymbol{b}$$

We are considering domain of $0 \leq \theta \leq \frac{\pi}{2}$

$$0 \leq \cos\theta \leq 1$$

$$|\boldsymbol{a}||\boldsymbol{b}| \geq \boldsymbol{a} \cdot \boldsymbol{b}$$

$$\sqrt{\alpha^2 + \beta^2 + \gamma^2}\sqrt{\delta^2 + \varepsilon^2 + \zeta^2} \geq \alpha\delta + \beta\varepsilon + \gamma\,\zeta$$

Equation 46