Correlation and Regression

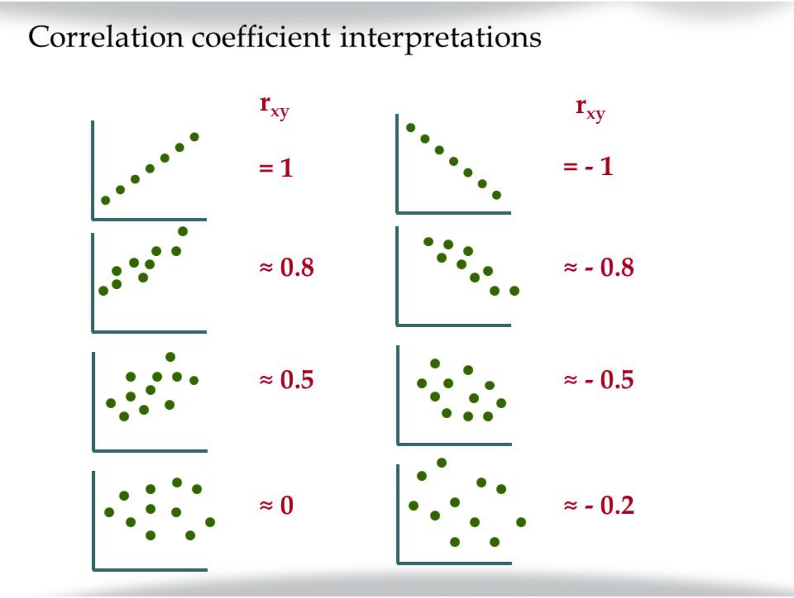
Correlation

Correlation deals with the measure of strength of the linear relationship between variables.

Correlation

Correlation deals with the measure of strength of the linear relationship between variables.

- Graphical - Scatter plot
- Correlation coefficient (due to Karl Pearson)
- Rank correlation


Correlation coefficient interpretations
r_{xy}
= 1
≈ 0.8
≈ 0.5
≈ 0
r_{xy}
= - 1
≈ - 0.8
≈ - 0.5
≈ - 0.2

Karl Pearson Correlation coefficient (Product moment coefficient)

$$r_{XY} = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}$$

Correlation

Karl Pearson Correlation coefficient (Product moment coefficient)

$$r_{XY} = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}$$

$$r_{XY} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{(N\sum X^2 - [\sum X]^2)(N\sum Y^2 - [\sum Y]^2)}}$$

Properties

1. $-1 \leq r_{XY} \leq 1$
   If $r_{XY} = -1 \implies$ perfect negative correlation
   If $r_{XY} = 1 \implies$ perfect positive correlation
   If $r_{XY} = 0 \implies$ Uncorrelated (no linear relationship bet X and Y)

Properties

1. $-1 \leq r_{XY} \leq 1$
   If $r_{XY} = -1 \implies$ perfect negative correlation
   If $r_{XY} = 1 \implies$ perfect positive correlation
   If $r_{XY} = 0 \implies$ Uncorrelated (no linear relationship bet X and Y)

2. $r_{XY} = r_{UV} = \frac{N\sum UV - \sum U \sum V}{\sqrt{(N\sum U^2 - [\sum U]^2)(N\sum V^2 - [\sum V]^2)}}$
   where $U = \frac{X-a}{h}$ and $V = \frac{Y-b}{k}$

Regression

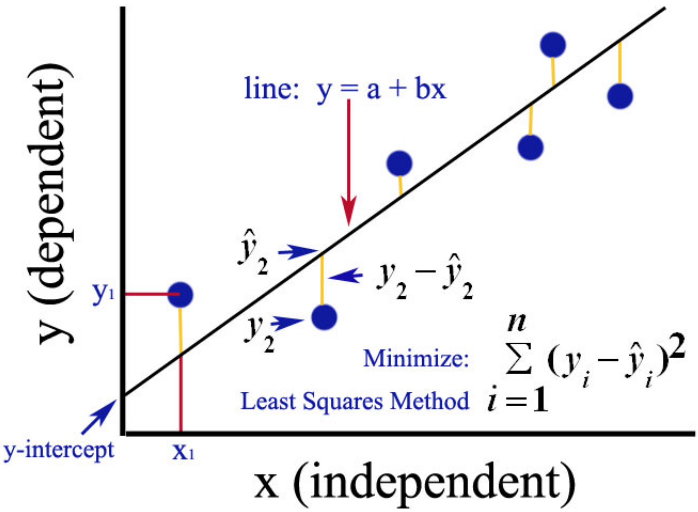
It is mathematical measure of average relationship between two or more variables

It is mathematical measure of average relationship between two or more variables

Regression line

Line which gives the best estimate to the value of one variable for any specific value of the other variable.

Regression Line



Lines of regression

Regression line of y on x

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Lines of regression

Regression line of y on x

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Regression line of x on y

$$x - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Lines of regression

Regression line of y on x

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Regression line of x on y

$$x - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Regression coefficients

$$b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y}$$

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$
2. r_{xy}, b_{xy}, b_{yx} will have same sign

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$
2. r_{xy}, b_{xy}, b_{yx} will have same sign
3. Both the lines of regression pass through $(\bar{X}, \bar{Y})$

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$
2. r_{xy}, b_{xy}, b_{yx} will have same sign
3. Both the lines of regression pass through $(\bar{X}, \bar{Y})$
4. If there is a perfect correlation between two variables, then there is only one regression line.

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$
2. r_{xy}, b_{xy}, b_{yx} will have same sign
3. Both the lines of regression pass through $(\bar{X}, \bar{Y})$
4. If there is a perfect correlation between two variables, then there is only one regression line.
5. If $r = 0$, then the two lines of regression are perpendicular. If $r = 1$ or -1, then the two lines coincide.

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$
2. r_{xy}, b_{xy}, b_{yx} will have same sign
3. Both the lines of regression pass through $(\bar{X}, \bar{Y})$
4. If there is a perfect correlation between two variables, then there is only one regression line.
5. If $r = 0$, then the two lines of regression are perpendicular. If $r = 1$ or -1, then the two lines coincide.
6. $$b_{XY} = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - [\sum Y]^2}$$

Properties

1. $r_{xy}^2 = b_{xy} * b_{yx}$
2. r_{xy}, b_{xy}, b_{yx} will have same sign
3. Both the lines of regression pass through $(\bar{X}, \bar{Y})$
4. If there is a perfect correlation between two variables, then there is only one regression line.
5. If $r = 0$, then the two lines of regression are perpendicular. If $r = 1$ or -1, then the two lines coincide.
6. $$b_{XY} = \frac{N \sum XY - \sum X \sum Y}{N \sum Y^2 - [\sum Y]^2}$$
7. $$b_{YX} = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - [\sum X]^2}$$

1. Calculate the correlation coefficient for the following heights (in inches) of fathers' (x) and their sons' (y):

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Obtain the lines of regression for the above data and find the estimate of x for $y = 70$

Problems

1. Calculate the correlation coefficient for the following heights (in inches) of fathers' (x) and their sons' (y):

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Obtain the lines of regression for the above data and find the estimate of x for $y = 70$

$r_{xy} = 0.603, \bar{x} = 68, \bar{y} = 69, \sigma_x = 4.5, \sigma_y = 5.5$

Problems

1. Calculate the correlation coefficient for the following heights (in inches) of fathers' (x) and their sons' (y):

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Obtain the lines of regression for the above data and find the estimate of x for $y = 70$

$r_{xy} = 0.603, \bar{x} = 68, \bar{y} = 69, \sigma_x = 4.5, \sigma_y = 5.5$

Regression equation of x on y: $x = 0.5454y + 30.3674$

Regression equation of y on x: $y = 0.6666x + 23.6712$

Try!!!!

2. Calculate the coefficient of correlation between X and Y by Karl Pearson's method:

X	25	30	28	29	32	24	36	28	27	21
Y	18	20	21	16	14	13	22	15	19	12

Also, obtain the regression equations.

Try!!!!

2. Calculate the coefficient of correlation between X and Y by Karl Pearson's method:

X	25	30	28	29	32	24	36	28	27	21
Y	18	20	21	16	14	13	22	15	19	12

Also, obtain the regression equations. $r_{XY} = 0.5955$, positive correlation

Problems

3. A computer while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following: $n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508$. It was later discovered that they had copied two pairs as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient.

Problems

3. A computer while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following:
$n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508.$ It was later discovered that they had copied two pairs as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient.
$r_{xy} = 0.667$
4. Can $y = 5 + 2.8x$ and $x = 3 - 0.5y$ be the estimated regression equations of y on x and x on y respectively?

Problems

3. A computer while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following:
$n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508$. It was later discovered that they had copied two pairs as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient.
$r_{xy} = 0.667$
4. Can $y = 5 + 2.8x$ and $x = 3 - 0.5y$ be the estimated regression equations of y on x and x on y respectively? No

Problems

3. A computer while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following:
$n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508$. It was later discovered that they had copied two pairs as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient.
$r_{xy} = 0.667$
4. Can $y = 5 + 2.8x$ and $x = 3 - 0.5y$ be the estimated regression equations of y on x and x on y respectively? No
5. Out of two lines of regression, which is the regression line of X on Y.

$$X + 2Y - 5 = 0, 2X + 3Y - 8 = 0$$

Also, obtain (i) the value of correlation coefficient, (ii) mean values of X and Y, (iii) if the variance of X is 12, find σ_Y.

Problems

3. A computer while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following: $n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508$. It was later discovered that they had copied two pairs as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient.

$r_{xy} = 0.667$

4. Can $y = 5 + 2.8x$ and $x = 3 - 0.5y$ be the estimated regression equations of y on x and x on y respectively? No

5. Out of two lines of regression, which is the regression line of X on Y.

$$X + 2Y - 5 = 0, 2X + 3Y - 8 = 0$$

Also, obtain (i) the value of correlation coefficient, (ii) mean values of X and Y, (iii) if the variance of X is 12, find σ_Y.

$r_{XY} = -0.866, b_{XY} = -1.5, b_{YX} = -0.5$

$\bar{X} = 1, \bar{Y} = 2, \sigma_Y = 2$

Partial correlation

A partial correlation measures the relationship between two variables while controlling the influence of the third variable by holding it constant.

Partial correlation

A partial correlation measures the relationship between two variables while controlling the influence of the third variable by holding it constant.
Zero-order partial correlation coefficient - r_{xy}, r_{xz}, r_{yz}

Partial correlation

A partial correlation measures the relationship between two variables while controlling the influence of the third variable by holding it constant.
Zero-order partial correlation coefficient - r_{xy}, r_{xz}, r_{yz}
First-order partial correlation coefficient :

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

Problems

1. The simple correlation coefficients between temperature (X_1), corn yield (X_2), and rainfall (X_3) are

$$r_{12} = 0.59, r_{13} = 0.46, \quad \text{and} \quad r_{23} = 0.77$$

Find the partial correlation coefficients $r_{12.3}, r_{23.1},$ and $r_{31.2}$.

Problems

1. The simple correlation coefficients between temperature (X_1), corn yield (X_2), and rainfall (X_3) are

$$r_{12} = 0.59, r_{13} = 0.46, \quad \text{and} \quad r_{23} = 0.77$$

Find the partial correlation coefficients $r_{12.3}, r_{23.1},$ and $r_{31.2}$.

2. If all the correlation coefficients of zero order in a set of p-variates are equal to r, show that every partial correlation of first order is $\frac{r}{1+r}$

Problems

1. The simple correlation coefficients between temperature (X_1), corn yield (X_2), and rainfall (X_3) are

$$r_{12} = 0.59, r_{13} = 0.46, \quad \text{and} \quad r_{23} = 0.77$$

Find the partial correlation coefficients $r_{12.3}, r_{23.1},$ and $r_{31.2}$.

2. If all the correlation coefficients of zero order in a set of p-variates are equal to r, show that every partial correlation of first order is $\frac{r}{1+r}$

3. The correlation between a general intelligence test and school achievement in a group of children from 6 to 15 years is 0.8. The correlation between the general intelligence test and age in the same group is 0.7 and the correlation between school achievement and age is 0.6. What is the correlation between general intelligence and school achievement in children of the same age?
(Hint: $X_1 =$ General intelligence, $X_2 =$ School achievement, $X_3 =$ Age. Given $r_{12} = 0.8, r_{13} = 0.7, r_{23} = 0.6$. Calculate $r_{12.3}$)

Multiple correlation

We study the effects of all the independent variables simultaneously on a dependent variable. For example, to study the correlation coefficient between the yied of paddy (X_1) and the other independent variables namely, manure (X_2), humidity (X_3), type of seedlings (X_4), rainfall (X_5), we use multiple correlation, denoted by $R_{1.2345}$

Multiple correlation

We study the effects of all the independent variables simultaneously on a dependent variable. For example, to study the correlation coefficient between the yied of paddy (X_1) and the other independent variables namely, manure (X_2), humidity (X_3), type of seedlings (X_4), rainfall (X_5), we use multiple correlation, denoted by $R_{1.2345}$
The multiple correlation coefficient of X_1 on X_2 and X_3 is denoted by $R_{1.23}$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Multiple correlation

We study the effects of all the independent variables simultaneously on a dependent variable. For example, to study the correlation coefficient between the yied of paddy (X_1) and the other independent variables namely, manure (X_2), humidity (X_3), type of seedlings (X_4), rainfall (X_5), we use multiple correlation, denoted by $R_{1.2345}$

The multiple correlation coefficient of X_1 on X_2 and X_3 is denoted by $R_{1.23}$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Properties

- $0 \leq R_{1.23} \leq 1$
- $R_{1.23} \geq r_{12}, r_{13}, r_{23}$

Problems

1. The simple correlation coefficients between temperature (X_1), corn yield (X_2), and rainfall (X_3) are

$$r_{12} = 0.59, r_{13} = 0.46, \quad \text{and} \quad r_{23} = 0.77$$

Find the multiple correlation coefficients $R_{1.23}, R_{2.31}$, and $R_{3.12}$.

Problems

1. The simple correlation coefficients between temperature (X_1), corn yield (X_2), and rainfall (X_3) are

$$r_{12} = 0.59, r_{13} = 0.46, \quad \text{and} \quad r_{23} = 0.77$$

Find the multiple correlation coefficients $R_{1.23}, R_{2.31},$ and $R_{3.12}$.

2. The following zero order correlation coefficients are given $r_{12} = 0.98, r_{13} = 0.44, r_{23} = 0.54.$ Calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Problems

1. The simple correlation coefficients between temperature (X_1), corn yield (X_2), and rainfall (X_3) are

$$r_{12} = 0.59, r_{13} = 0.46, \quad \text{and} \quad r_{23} = 0.77$$

Find the multiple correlation coefficients $R_{1.23}, R_{2.31},$ and $R_{3.12}$.

2. The following zero order correlation coefficients are given $r_{12} = 0.98, r_{13} = 0.44, r_{23} = 0.54$. Calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

3. If all the correlation coefficients of zero order in a set of p-variates are equal to r, show that every multiple correlation $R_{1.23} = R_{2.13} = R_{3.12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$

Multiple Regression

If X, Y, and Z are three variables, then the regression equation of X on Y and Z is

$$X = aY + bZ + c$$

Multiple Regression

If X, Y, and Z are three variables, then the regression equation of X on Y and Z is

$$X = aY + bZ + c$$

Problem

Find the multiple linear regression of X_1 on X_2 and X_3 from the data relating to three variables

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

If X, Y, and Z are three variables, then the regression equation of X on Y and Z is

$$X = aY + bZ + c$$

Problem

Find the multiple linear regression of X_1 on X_2 and X_3 from the data relating to three variables

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

$X_1 = 0.3899X_2 - 0.6233X_3 + 16.4776$

Multiple Regression

For a multivariate data, the regression equation of X on Y and Z is

$$(X-\bar{X})\frac{\omega_{11}}{\sigma_1}+(Y-\bar{Y})\frac{\omega_{12}}{\sigma_2}+(Z-\bar{Z})\frac{\omega_{13}}{\sigma_3}=0$$

where

$$\omega = det\begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

$$\omega_{11} = det\begin{bmatrix} 1 & r_{23} \\ r_{23} & 1 \end{bmatrix}$$

$$\omega_{12} = -det\begin{bmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{bmatrix}$$

$$\omega_{13} = det\begin{bmatrix} r_{12} & 1 \\ r_{13} & r_{23} \end{bmatrix}$$

Problems

1. Find the regression equation of X on Y and Z given the following results:

Variables	Mean	SD	r_{12}	r_{23}	r_{31}
X	35.8	4.2	0.6	-	-
Y	52.4	5.3	-	0.7	-
Z	48.8	6.1	-	-	0.8

Problems

1. Find the regression equation of X on Y and Z given the following results:

Variables	Mean	SD	r_{12}	r_{23}	r_{31}
X	35.8	4.2	0.6	-	-
Y	52.4	5.3	-	0.7	-
Z	48.8	6.1	-	-	0.8

$$\omega = det \begin{bmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.7 \\ 0.8 & 0.7 & 1 \end{bmatrix}$$

$$X = 0.062Y + 0.513Z + 7.6$$