Example

Let's take the previous example

$X \backslash Y$	5	7	9	11	13	$n_{i.}$	$\overline{Y}/x_i$	σ_{Y/x_i}
1				1	4	5	12.6	0.80
2			2	7	1	10	10.8	1.08
4			9	1		10	9.2	0.60
6	2	8	6	1		17	7.71	1.52
9	5	2	1			8	6	1.41
$n_{.j}$	7	10	18	10	5	50		
$\overline{X}/y_j$	8.14	6.6	4.72	2.5	1.2			
σ_{X/y_j}	1.36	1.20	1.48	1.32	1.33			

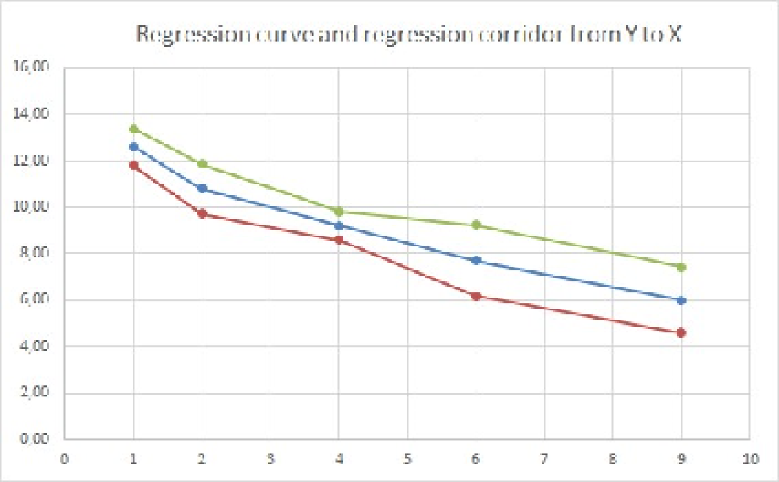
Example

To draw the regression corridor of Y in X we join the points

$$\left(x_i; \overline{Y}/x_i - \sigma_{Y/x_i}\right) = \left\{(1; 11.80), (2; 9.72), (4; 8.60), (6; 6.18), (9; 4.59)\right\}$$

then the points

$$\left(x_i; \overline{Y}/x_i + \sigma_{Y/x_i}\right) = \left\{(1; 13.40), (2; 11.88), (4; 9.80), (6; 9.23), (9; 7.41)\right\}.$$


Regression curve and regression corridor from Y to X
16,00
14,00
12,00
10,00
8,00
6,00
4,00
2,00
0,00
0
1
2
3
4
5
6
7
8
9
10

We have the following relationship called variance decomposition

$$\sigma_Y^2 = \overline{\sigma_{Y/X}^2} + \sigma_{\overline{Y/X}}^2.$$

where

$$\overline{\sigma_{Y/X}^2} = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\sigma_{Y/x_i}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(y_j - \overline{Y}/x_i\right)^2$$

is the average conditional variance of Y with respect to X (also called the variance around the regression corridor) is the average of the conditional variances, and

$$\sigma_{\overline{Y/X}}^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\left(\overline{Y}/x_i - \overline{Y}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\left(\overline{Y}/x_i\right)^2 - \overline{Y}^2$$

is the variance of conditional means.

Example

For the previous example we have $\overline{Y} = 8.84$,

$$\begin{aligned}\overline{\sigma^2_{Y/X}} &= (5 \cdot 0.64 + 10 \cdot 1.16 + 10 \cdot 0.36 + 17 \cdot 2.33 + 8 \cdot 2.00)/50 \\ &= 1.48\end{aligned}$$

and

$$\begin{aligned}\sigma^2_{\overline{Y/X}} &= \left(5 \cdot 12.60^2 + 10 \cdot 10.80^2 + 10 \cdot 9.20^2 + 17 \cdot 7.71^2 + 8 \cdot 6^2\right)/50 - \\ &= 3.96\end{aligned}$$

then $\overline{\sigma^2_{Y/X}} + \sigma^2_{\overline{Y/X}} = 5.44 \simeq \sigma^2_Y = 5.41$

Definition

We call the Pearson correlation ratio the real number

$$\eta^2_{Y/X} = \frac{\sigma^2_{\overline{Y/X}}}{\sigma^2_Y}.$$

It's the percentage of variability (of the variable Y) due to the differences between the modalities (of the variable X).

Remark

1. $0 \leq \eta^2_{Y/X} \leq 1$.
2. If $\eta^2_{Y/X} = 0 \iff \sigma^2_{\overline{Y/X}} = 0$ then the regression curve of Y in X is horizontal, this means that the variable Y is uncorrelated on average with the variable X.
3. If $\eta^2_{Y/X} = 1 \iff \overline{\sigma^2_{Y/X}} = 0$ then Y is totally linked to X.

Theorem

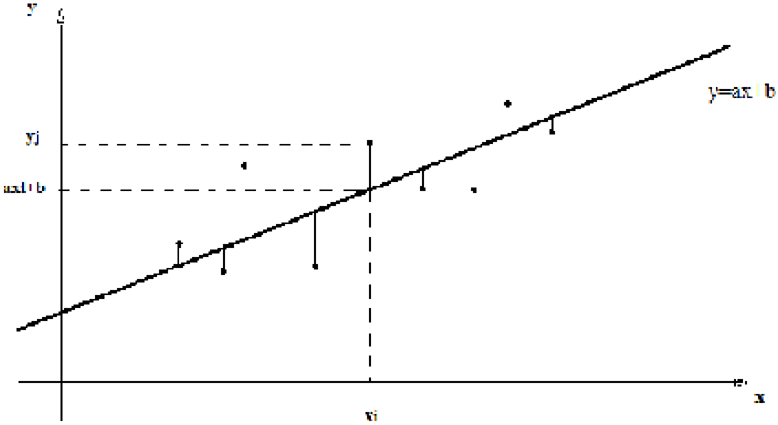
The linear correlation coefficient and the Pearson correlation ratio realize the following relationship $\rho^2(X, Y) \leq \eta^2_{Y/X}$.

Interpretation of results

1. If $\rho^2(X, Y) \leq \eta^2_{Y/X} \leq 0.1$ then Y and X are not correlated.
2. If $\rho^2(X, Y) \leq 0.1 < \eta^2_{Y/X} < 0.9$ then Y is partially linked to X but this link is not linear.
3. If $0.1 < \rho^2(X, Y) \leq \eta^2_{Y/X} < 0.9$ then Y is partially linked to X and this link is linear.
4. If $0.1 < \rho^2(X, Y) \leq 0.9 \leq \eta^2_{Y/X}$ then Y is linked to X and this link is functional but not linear.
5. If $0.9 < \rho^2(X, Y) \leq \eta^2_{Y/X}$ then Y is linearly related to X.

For the previous example we have $\eta^2_{Y/X} = \frac{3.96}{5.41} \simeq 0.7319$ **and** $\rho^2(X, Y) \simeq 0.7029$.

The linear fit consists in replacing the scatterplot by a line such that the estimated y-values along it, for the different x_i values are very close to the y_j values.



Definition

When $\rho^2(X, Y) > 0,9$ there exists a linear relationship between X and Y of the form $Y = aX + b$ which is called the regression line of Y in X and which minimize the sum

$$S = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} (y_j - ax_i - b)^2 .$$

The condition $\rho^2(X, Y) > 0,9$ is mandatory to confirm the linearity of the relationship between X and Y but we can still determine a regression line for values less than $0,9$.

Theorem

The regression line from Y to X is the line of the form $Y = aX + b$ with

$$a = \frac{Cov\,(X, Y)}{\sigma_X^2} \; et \; b = \overline{Y} - \frac{Cov\,(X, Y)}{\sigma_X^2}\overline{X}.$$

Remark

We can also determine the regression line from X to Y of the form $X = a'Y + b'$, où

$$a' = \frac{Cov\,(X, Y)}{\sigma_Y^2} \; et \; b' = \overline{X} - \frac{Cov\,(X, Y)}{\sigma_y^2}\overline{Y}.$$

Theorem

The regression line from Y to X is the line of the form $Y = aX + b$ with

$$a = \frac{Cov\left(X, Y\right)}{\sigma_X^2} \text{ et } b = \overline{Y} - \frac{Cov\left(X, Y\right)}{\sigma_X^2}\overline{X}.$$

Remark

We can also determine the regression line from X to Y of the form $X = a'Y + b'$, où

$$a' = \frac{Cov\left(X, Y\right)}{\sigma_Y^2} \text{ et } b' = \overline{X} - \frac{Cov\left(X, Y\right)}{\sigma_y^2}\overline{Y}.$$

The two lines pass through the mean point $\left(\overline{X}, \overline{Y}\right)$.

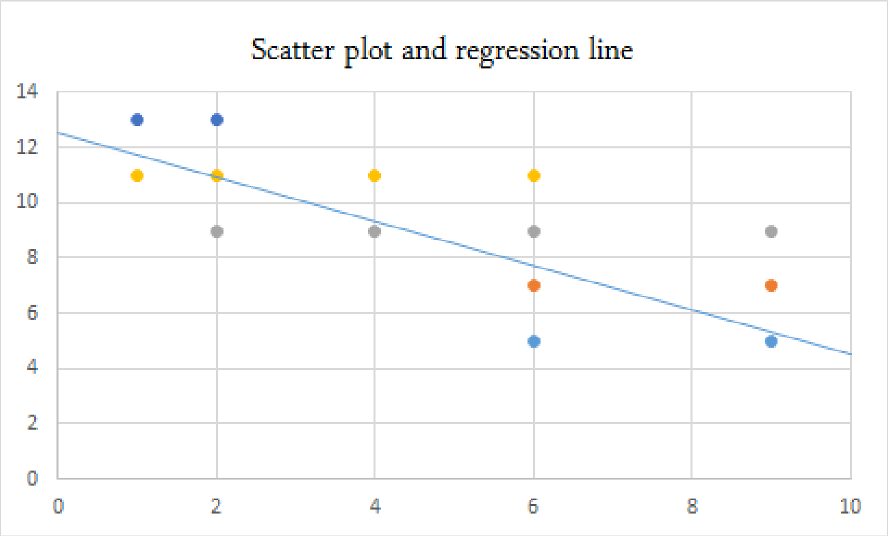
Example

We consider the previous example. The equation of the regression line from Y to X is determined.
We have $\rho^2(X,Y) \approx 0,7029 < 0,9$ the fit is not actually linear but there is a strong enough correlation that we can still fit it by a line of the form $(D_{Y/X}) : y = ax + b$ where $a = \frac{Cov(X,Y)}{\sigma_X^2}$ and $b = \overline{Y} - \frac{Cov(X,Y)}{\sigma_X^2}\overline{X}$, such that

$$\begin{aligned} a &= \frac{Cov(X,Y)}{\sigma_X^2} = \frac{-4,9552}{2,54^2} \approx -0,7681 \\ b &= \overline{Y} - \frac{Cov(X,Y)}{\sigma_X^2}\overline{X} = 8,84 + 0,7681 \cdot 4,78 = 12,5115 \end{aligned}$$

hence the equation of the regression line from Y to X is

$$(D_{Y/X}) : y = -0,7681x + 12,5115.$$

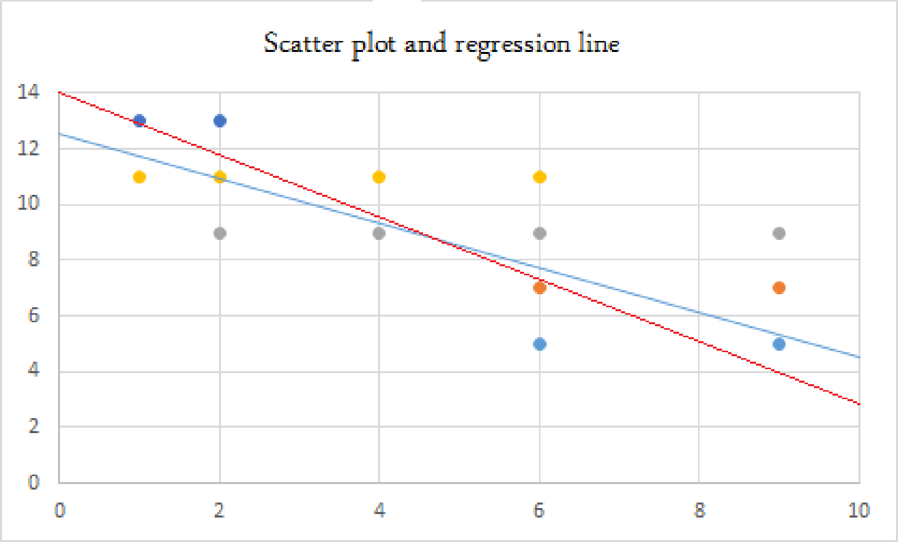

Scatter plot and regression line
14
12
10
8
6
4
2
0
0
2
4
6
8
10

We determine the equation of the regression line from X to Y.
$(D_{X/Y}) : x = a'y + b'$ such that:

$$\begin{aligned} a' &= \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{-4,9552}{2,3269^2} \approx -0,9152 \\ b' &= \overline{X} - \frac{Cov(X,Y)}{\sigma_Y^2}\overline{Y} = 4,78 + 0,9152 \times 8,84 \\ &= 12,8704 \end{aligned}$$

hence the equation of the regression line from X to Y is

$$(D_{X/Y}) : x = -0,9152y + 12,8704.$$


Scatter plot and regression line
14
12
10
8
6
4
2
0
0
2
4
6
8
10