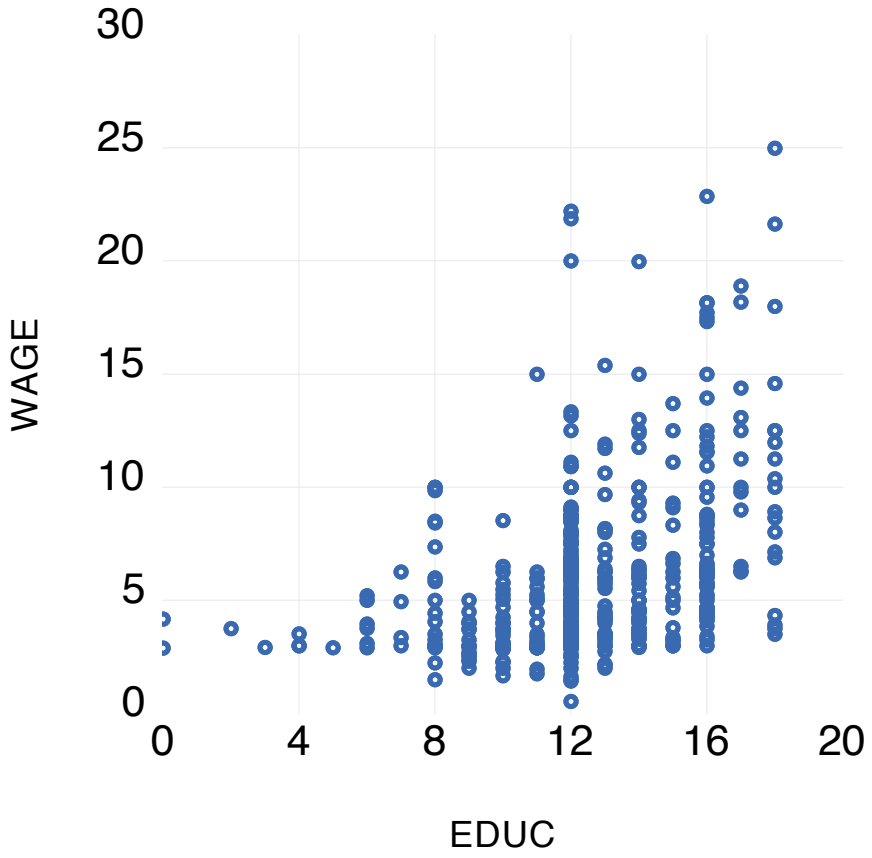
Correlação e Regressão

Associação entre variáveis

- O objetivo de estabelecer a distribuição conjunta de duas variáveis é o de compreender a existência de alguma associação entre elas, ou o grau de dependência entre elas

Exemplo wage1.wf1

- Gráfico de Dispersão



- Dados hipotéticos

Covariância

- Dados n pares de valores $(x_1, y_1)..., (x_n, y_n)$, chamaremos de covariância entre as variáveis X e Y, consideradas como população:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$$

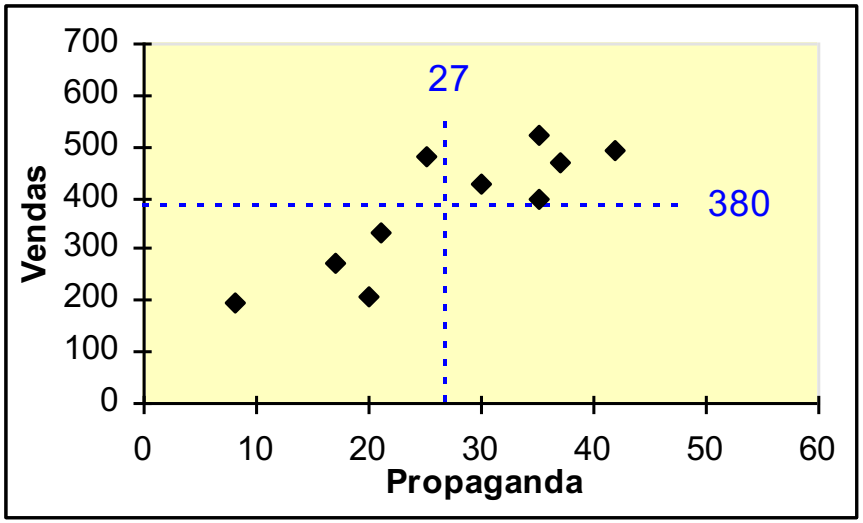
- É a média dos produtos dos valores centrados das variáveis

Covariância - Exemplo

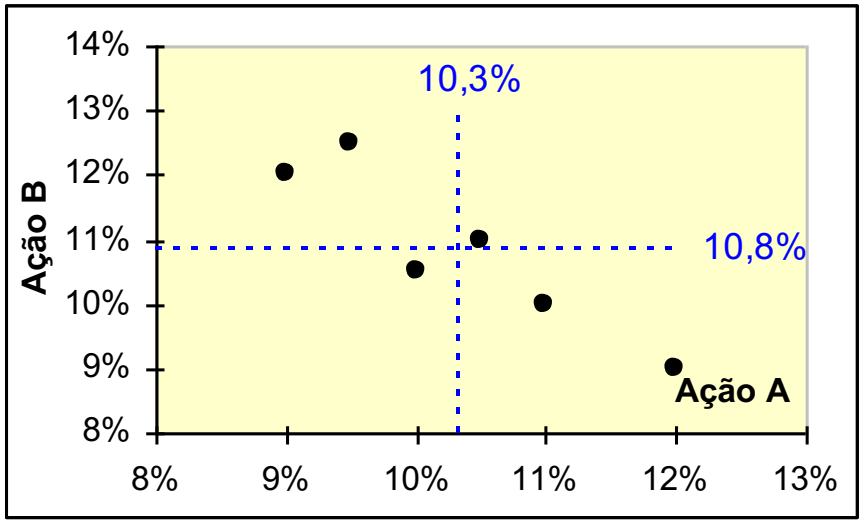
Matriz de Variância-Covariância

	WAGE	EDUC
WAGE	13.61295	4.142973
EDUC	4.142973	7.652908

- A covariância pode ser nula, negativa ou positiva.
- A covariância é a medida do afastamento simultâneo das respectivas médias.
- Se as ambas variáveis aleatórias tendem a estar simultaneamente acima, ou abaixo, de suas respectivas médias, então a covariância tenderá a ser positiva e nos outros casos poderá ser negativa, como mostram os gráficos abaixo.



A maioria dos pares de valores tem os dois valores acima de sua média correspondente, provocando covariância positiva.



A maioria dos pares de valores tem um valor acima de sua média e outro abaixo da média correspondente, provocando covariância negativa.

Características da covariância

- A covariância de uma variável e ela mesma é a própria variância da variável, seja no caso de população ou amostra. Como *Y* = *X*,

$$\sigma_{XX} = \frac{\sum_{i=1}^{N}(X_i - \mu_X) \times (X_i - \mu_X)}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N} = \sigma_X^2$$

- A permutação das variáveis não altera o resultado da covariância, se os pares de valores não forem alterados

$$\sigma_{XY} = \sigma_{YX}$$

Características da covariância

- Da mesma forma que a variância, a covariância é afetada pelos valores extremos da variável, ela não é uma medida resistente.
- A unidade de medida é o resultado do produto das unidades dos valores das variáveis.

Coeficiente de correlação

- Para facilitar o entendimento da relação entre duas variáveis e evitar a unidade de medida da covariância, foi definido o coeficiente de correlação r_{XY}.
- Os valores de r_{XY} estão limitados entre os valores –1 e +1, e sem nenhuma unidade de medida

Coeficiente de correlação

- Dados n pares de valores $(x_1, y_1)..., (x_n, y_n)$, chamaremos de covariância entre as variáveis X e Y, consideradas como população:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$$

- É a média dos produtos dos valores centrados das variáveis
- Tendo esta definição, podemos escrever o coeficiente de correlação como:

$$corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{dp(X).dp(Y)}$$

Coeficiente de correlação

- O coeficiente de correlação busca auferir a direção da relação entre as variáveis, dentro de um intervalo determinado entre -1 e 1
- O objetivo do intervalo é discriminar a direção e a intensidade da relação:
  - valores próximos de zero indicam ausência de relação entre as variáveis
  - valores próximos de 1 indicam forte relação positiva
  - valores próximos de -1 indicam forte relação negativa

Coeficiente de correlação

- O coeficiente de correlação é a medida do grau de associação linear entre duas variáveis
- Fórmula do coeficiente de correlação:

$$corr(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

Coeficiente de correlação - Exemplo

Matriz de Correlação

	WAGE	EDUC
WAGE	1	0.405903329
EDUC	0.405903329	1

Características de r

- Se a variável *Y* é a mesma variável *X*, então o coeficiente de correlação é igual a 1:

$$r_{XX} = \frac{\sigma_{XX}}{\sigma_X \times \sigma_X} = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

- A permutação das variáveis não altera o resultado do coeficiente de correlação, se os mesmos pares de valores forem mantidos.

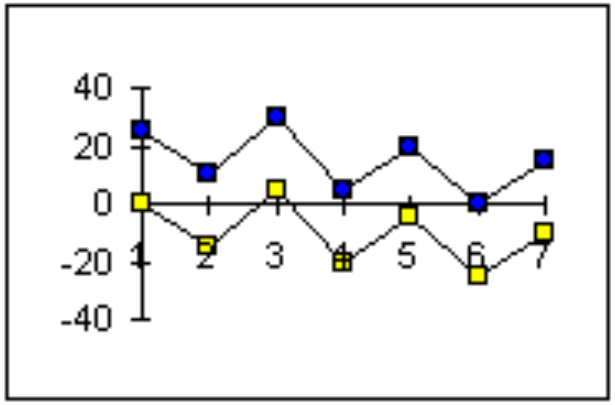
$$r_{XY} = r_{YX}$$

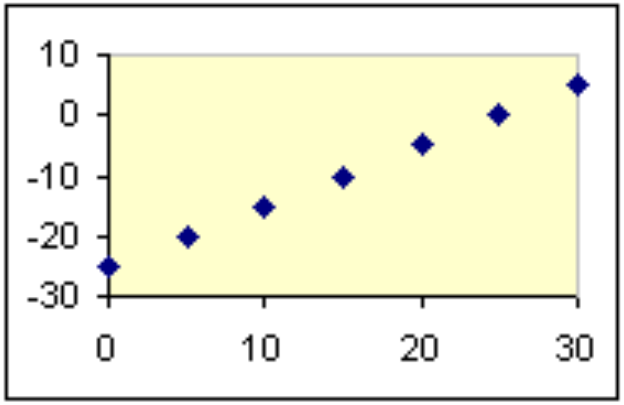
$r = +1$

Amostras perfeitamente correlacionadas em sentido positivo

X	Y
25	0
10	-15
30	5
5	-20
20	-5
0	-25
15	-10

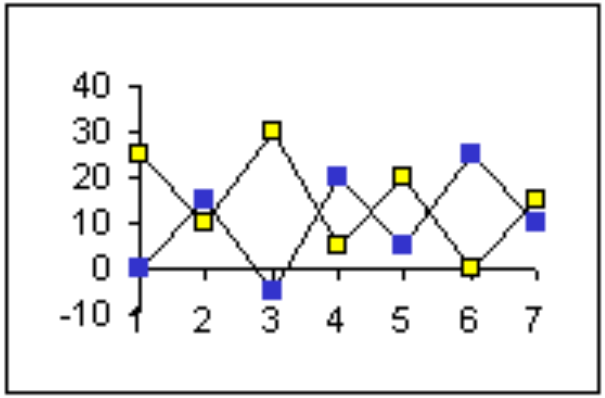
r	1,00

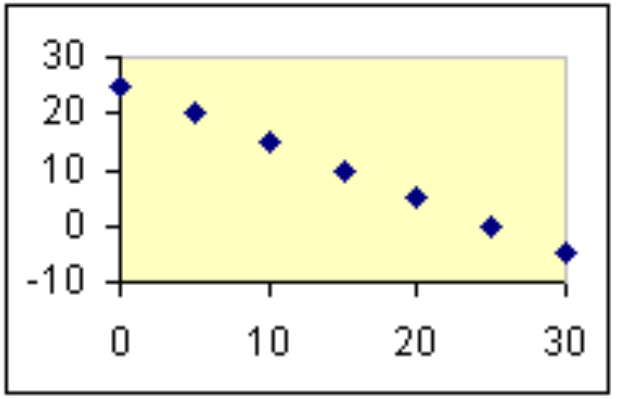




$r = -1$

	A	B	C	D	E	F	G	H	I
13	**Amostras perfeitamente correlacionadas em sentido negativo**								
14		**X**	**Y**		***r***	**-1,00**			
15		25	0						
16		10	15						
17		30	-5						
18		5	20						
19		20	5						
20		0	25						
21		15	10						
22									



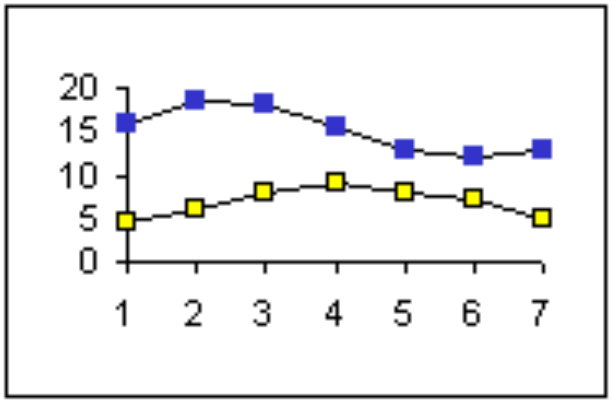


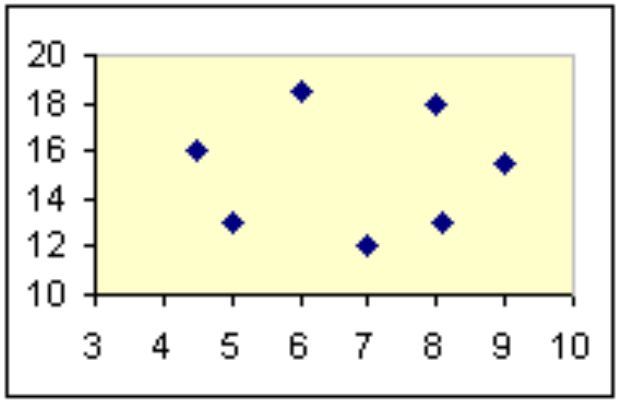
$r = 0$

Amostras não correlacionadas

X	Y
4,5	16
6	18,5
8	18
9	15,5
8,1	13
7	12
5	13

r	0,00





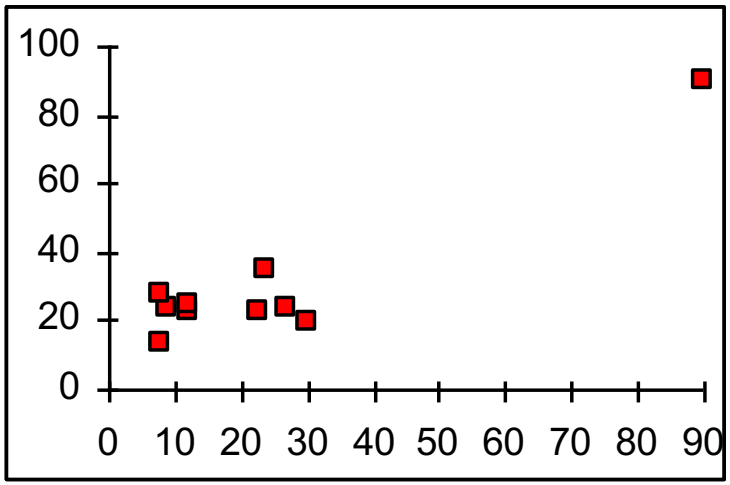
Alguns cuidados

- O coeficiente de correlação **não mede a relação causa-efeito entre as variáveis, apesar de que essa relação possa estar presente**.

- Por exemplo, uma correlação fortemente positiva entre as variáveis *X* e *Y* não autoriza afirmar que variações da variável *X* provocam variações na variável *Y*, ou vice-versa.

- O coeficiente de correlação **sozinho não identifica a relação causa-efeito entre as duas variáveis**

Exemplo de anomalia com *r* próximo de +1



Exemplo de anomalias com *r* próximo de 0