

Ciência dos Dados

Análise Exploratória dos Dados

Análise Bidimensional

Objetivos de Aprendizagem

Os alunos devem ser capazes de:

- Estudar a **relação existente** entre duas variáveis quantitativas graficamente;
- Por meio de medidas adequadas, medir o grau de associação entre duas variáveis quantitativas;
- Descrever o comportamento médio entre duas variáveis quantitativas por meio de um ajuste linear.

**Acompanhe, previamente, o PLANO DE AULA
no BLACKBOARD!**

Associação entre duas variáveis quantitativas

- Gráfico de Dispersão
- Coeficiente de Covariância
- Coeficiente de Correlação Linear de Pearson

PréAula05

Indicadores sócio-econômicos

O arquivo **Mundo.xlsx** conta com uma amostra de **85 países**, para os quais levantou-se uma série de indicadores socioeconômicos.

Variáveis:

X_1 : população em milhares de habitantes

X_2 : densidade populacional

X_3 : % de população urbana

X_4 : expectativa de vida feminina

X_5 : expectativa de vida masculina

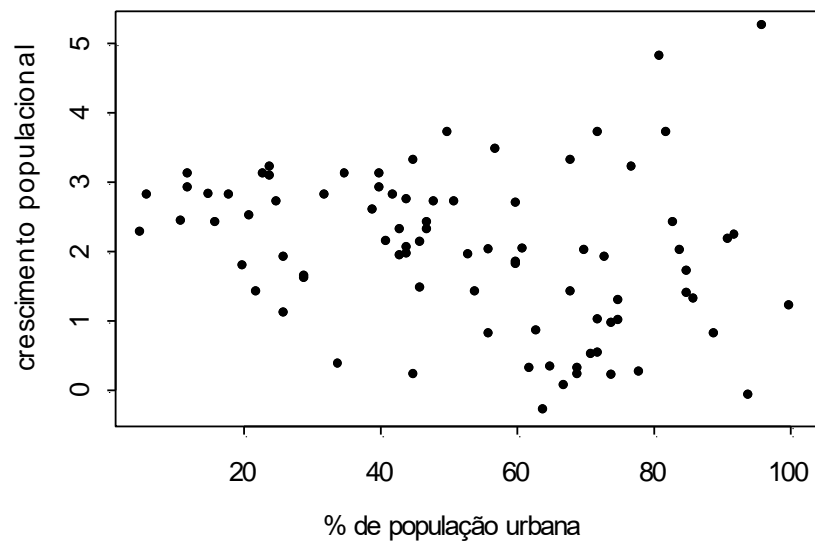
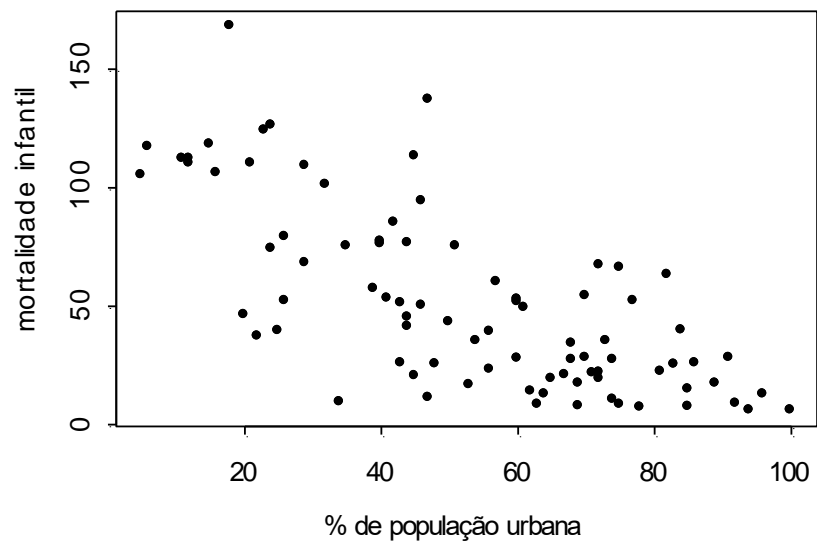
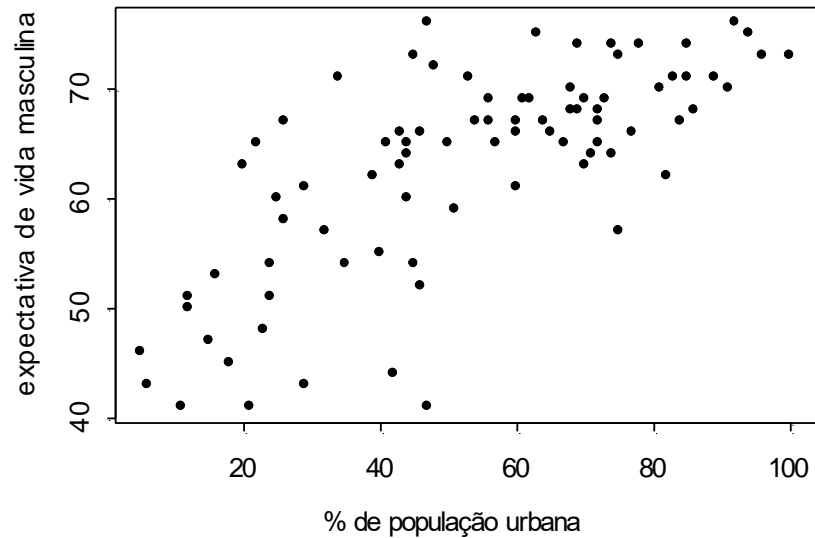
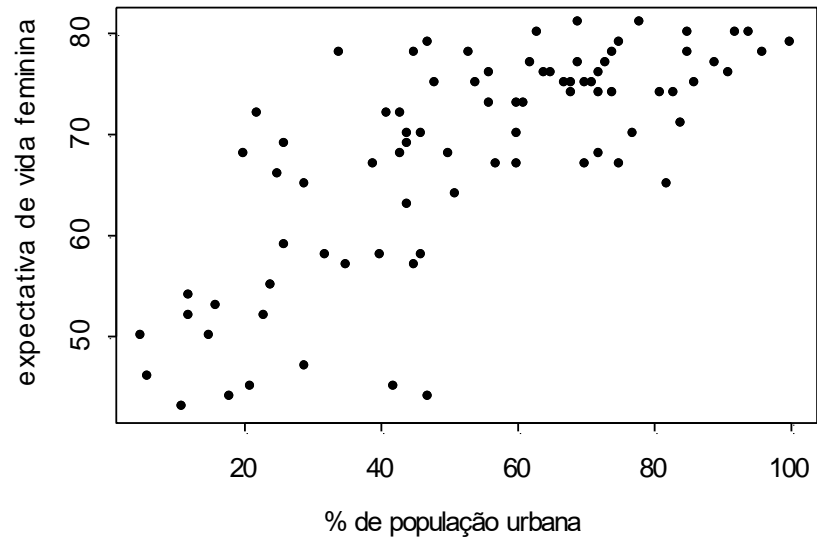
X_6 : crescimento populacional

X_7 : mortalidade infantil

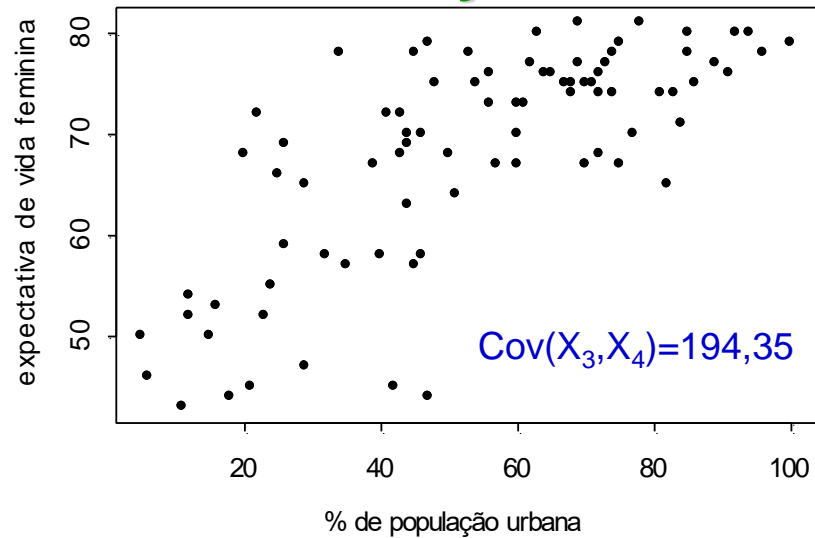
X_8 : PIB per capita

X_9 : % de mulheres alfabetizadas

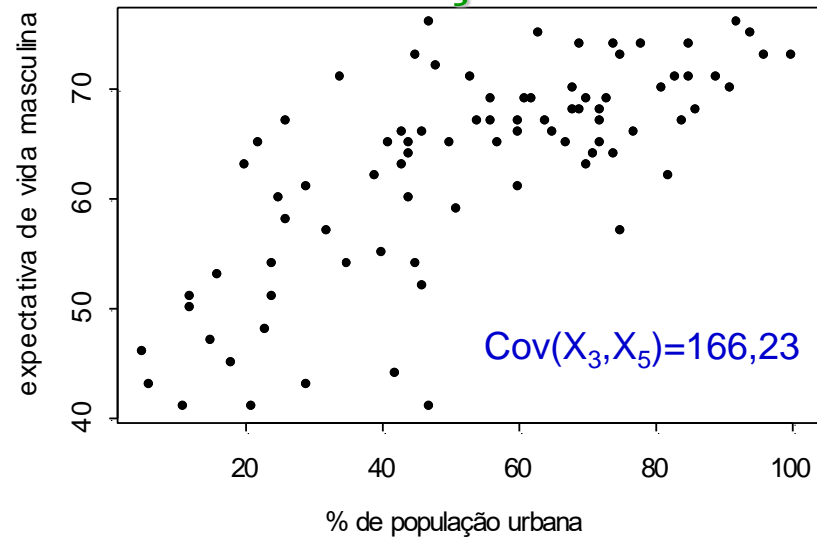
X_{10} : população em 100.000 habitantes



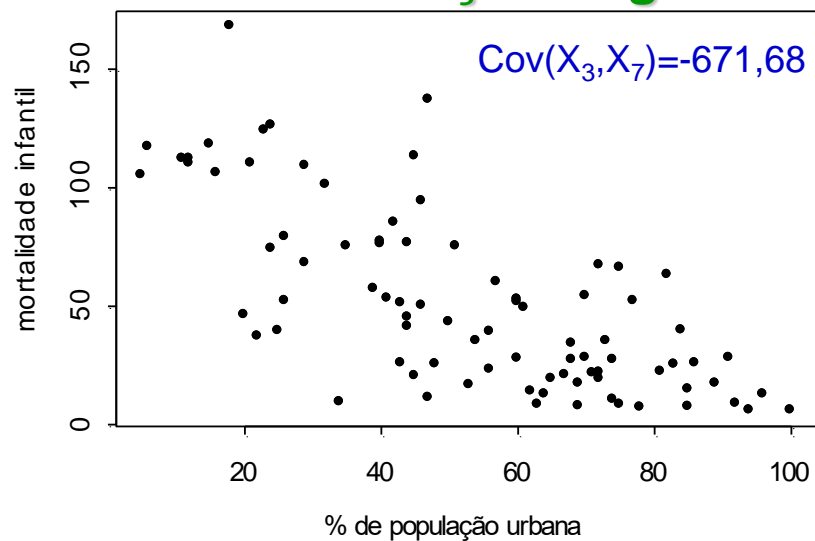
Associação Positiva



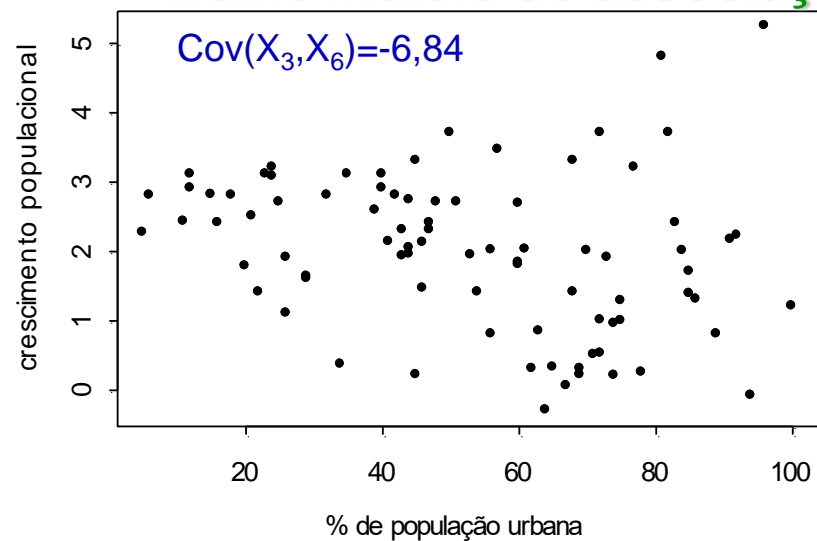
Associação Positiva



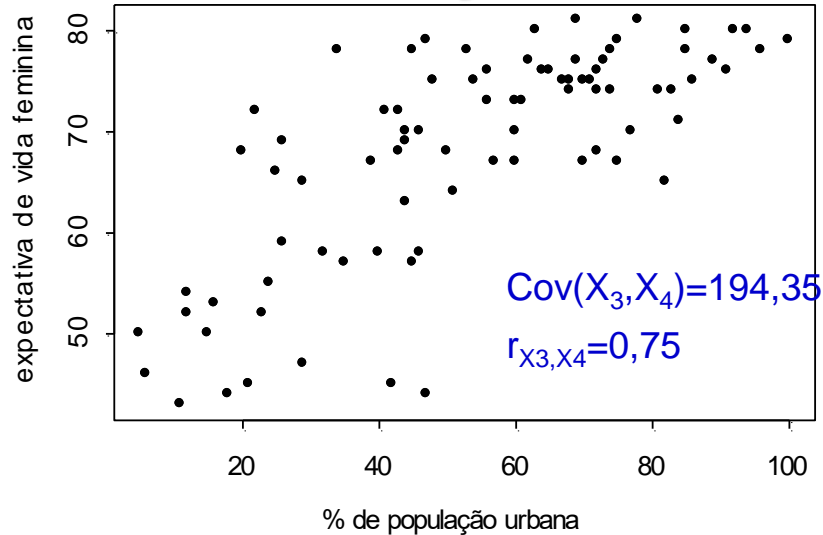
Associação Negativa



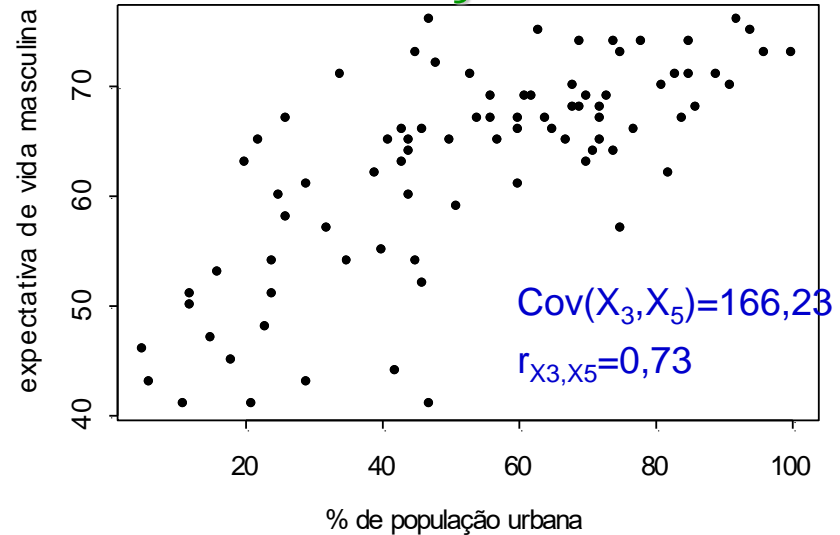
Baixo índice de associação



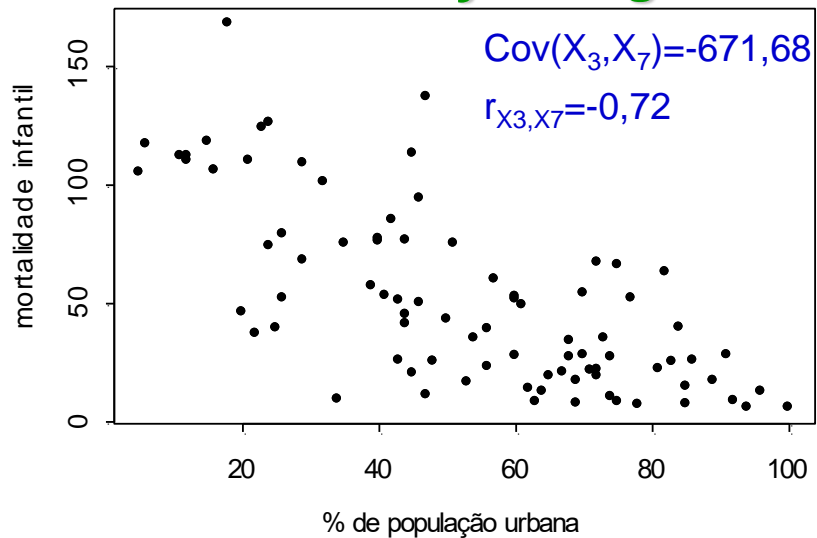
Associação Positiva



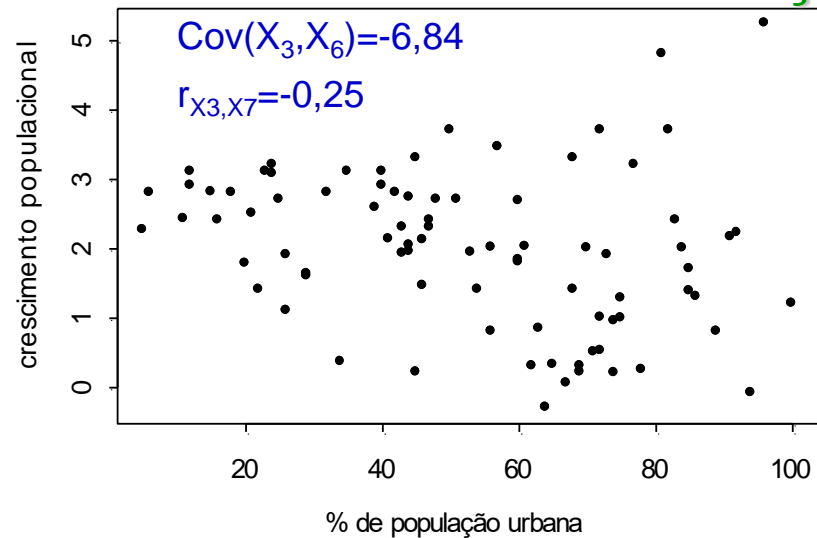
Associação Positiva



Associação Negativa



Baixo índice de associação



Exemplo: Gráfico de dispersão e medidas de associação

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Taxa de analfabetismo: Percentual de pessoas com 15 ou mais anos de idade que não sabem ler e escrever pelo menos um bilhete simples, em determinado espaço geográfico, no ano considerado.

Taxa de mortalidade infantil: Número de óbitos de menores de um ano de idade, por mil nascidos vivos, em determinado espaço geográfico, no ano considerado.

Exemplo: Gráfico de dispersão e medidas de associação

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Considere :

X: Taxa de analfabetismo

Y: Taxa de mortalidade infantil

$$\bar{x} = 14,2$$

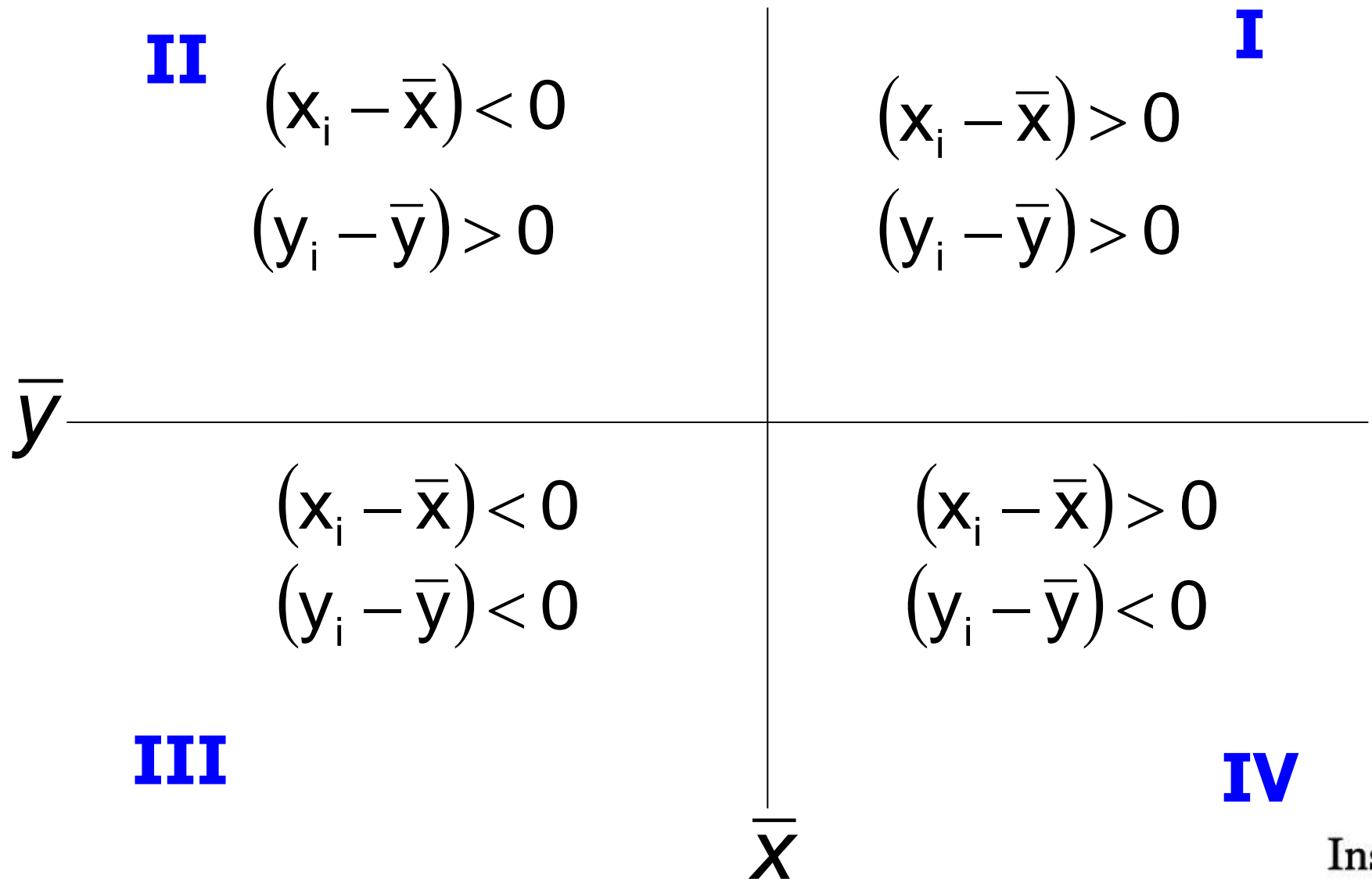
$$\bar{y} = 33,4$$

Coefficiente de Covariância

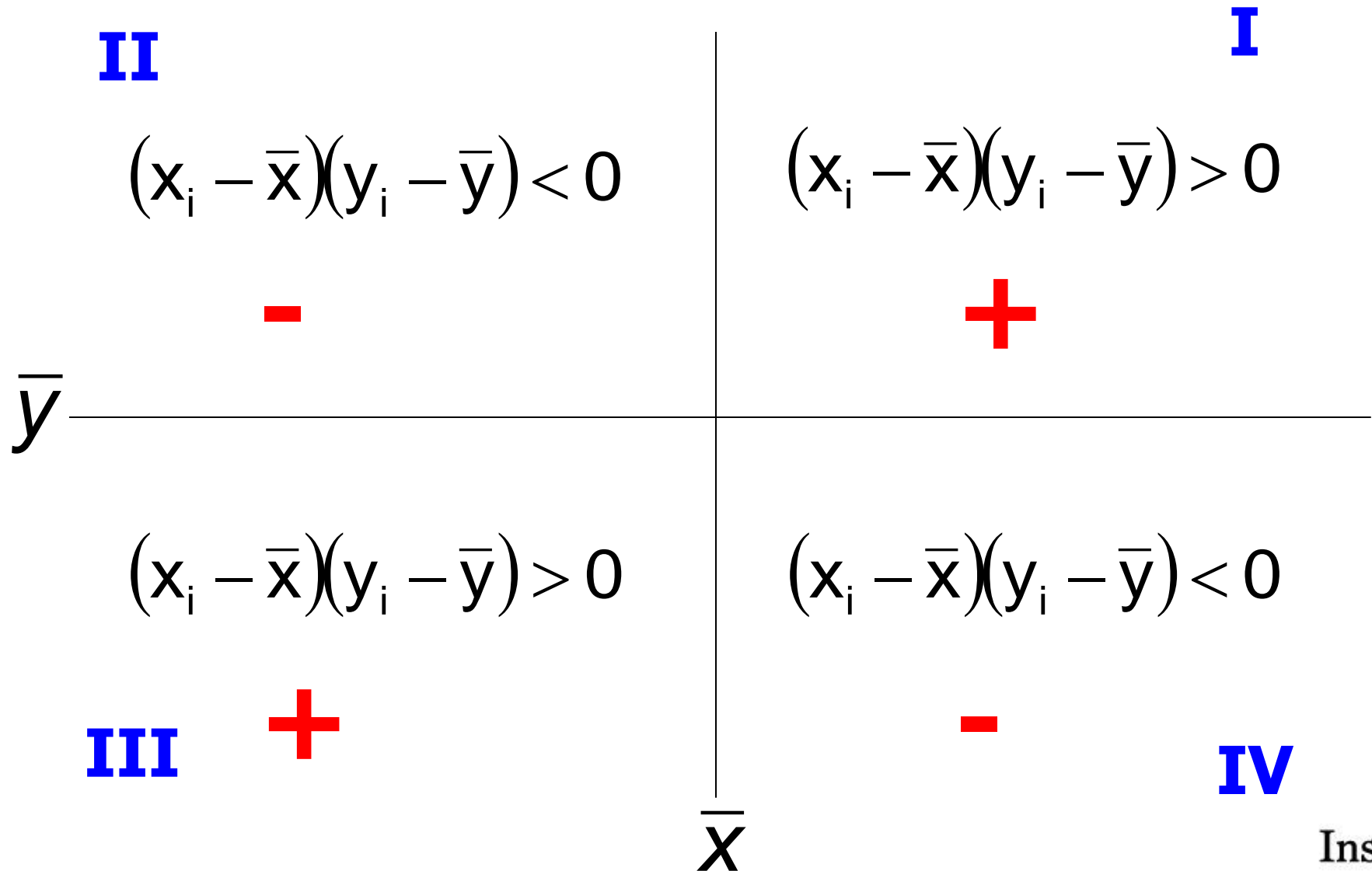
$$\text{Cov}(X, Y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- ✓ $\text{Cov}(X, Y) > 0$ se a associação linear for positiva.
- ✓ $\text{Cov}(X, Y) < 0$ se a associação linear for negativa.
- ✓ $\text{Cov}(X, Y) = 0$ indica que não existe associação linear positiva, nem negativa, mas pode existir outro tipo de associação.

Estudo de Sinal

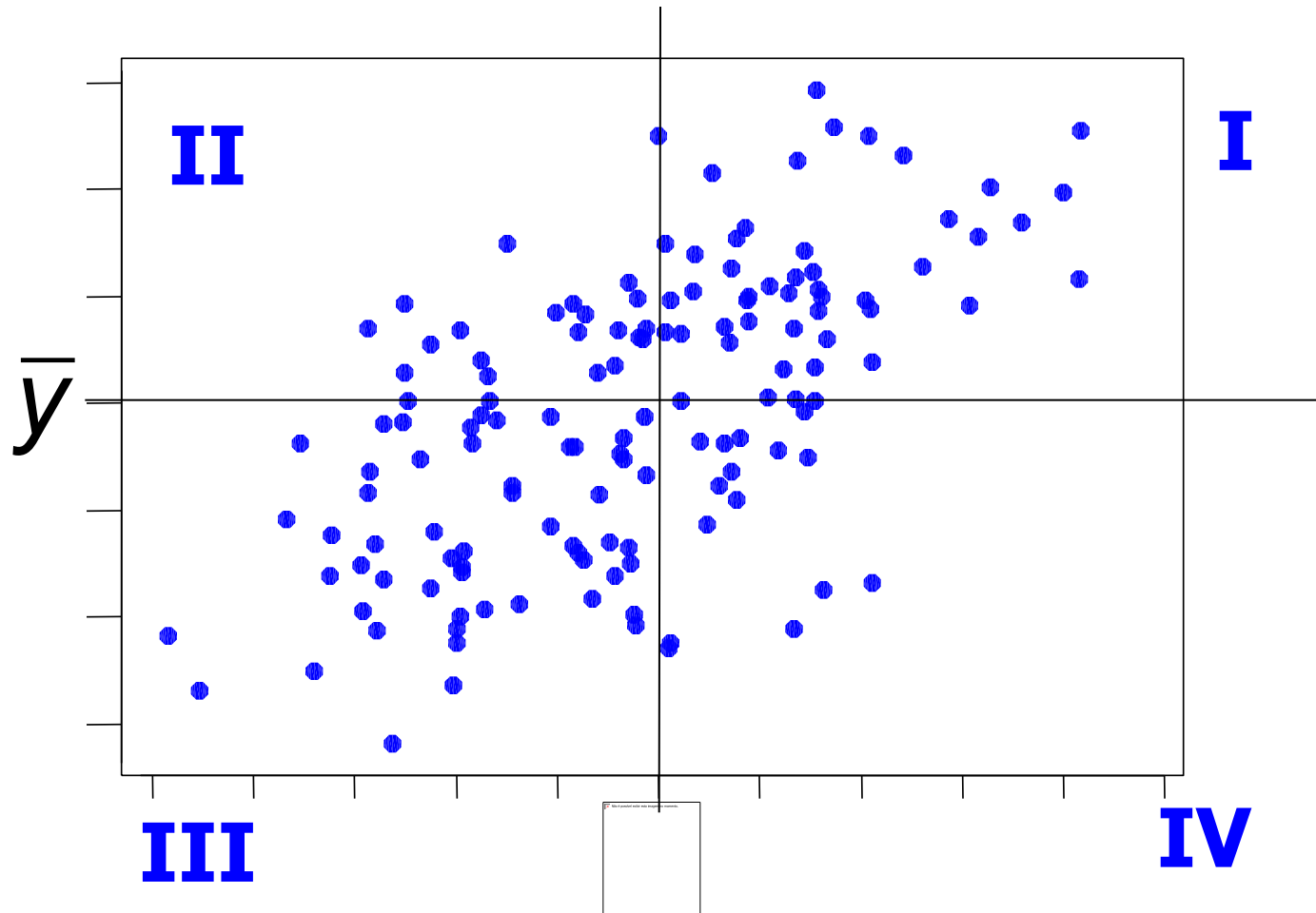


Estudo de Sinal



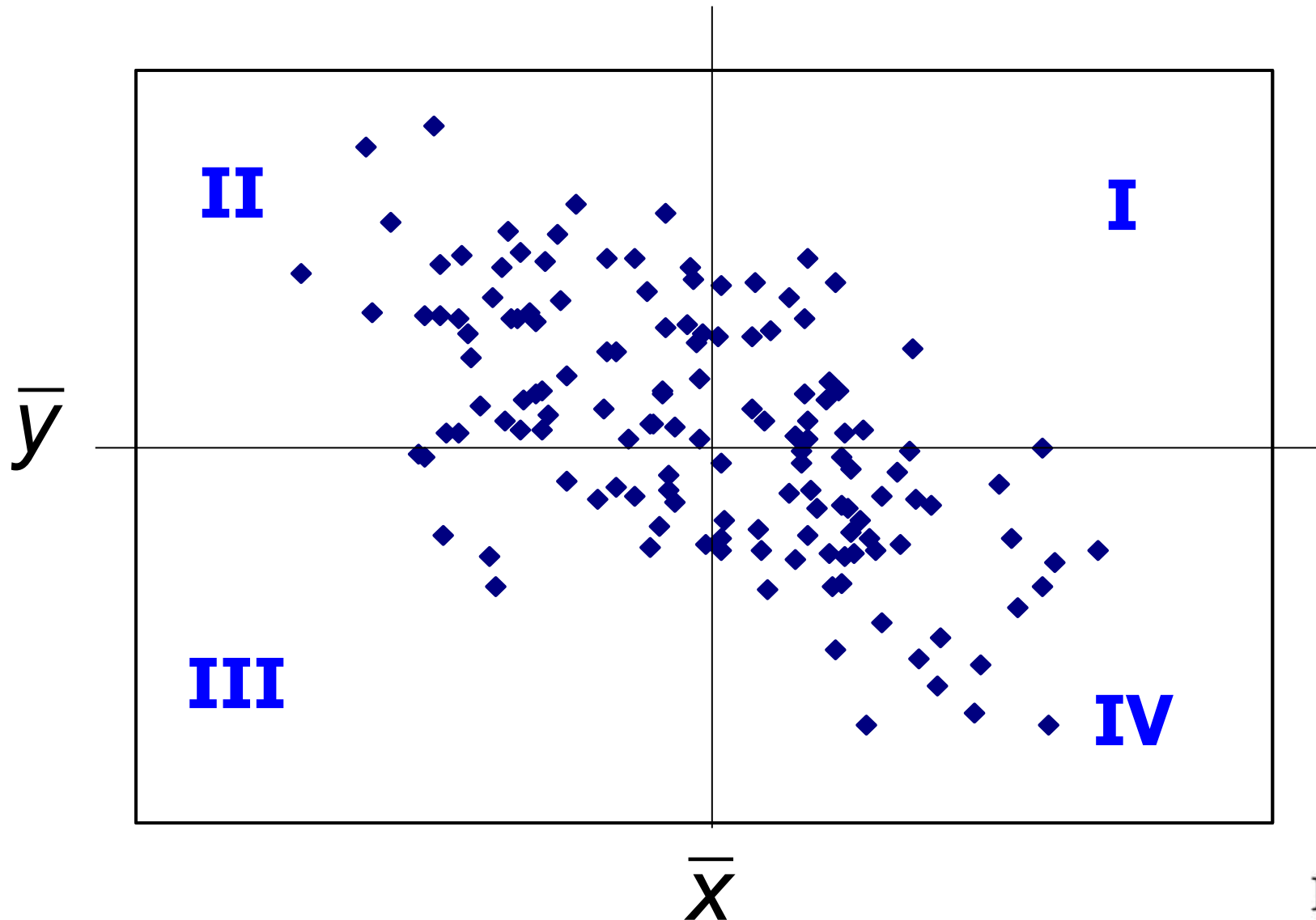
Associação Positiva

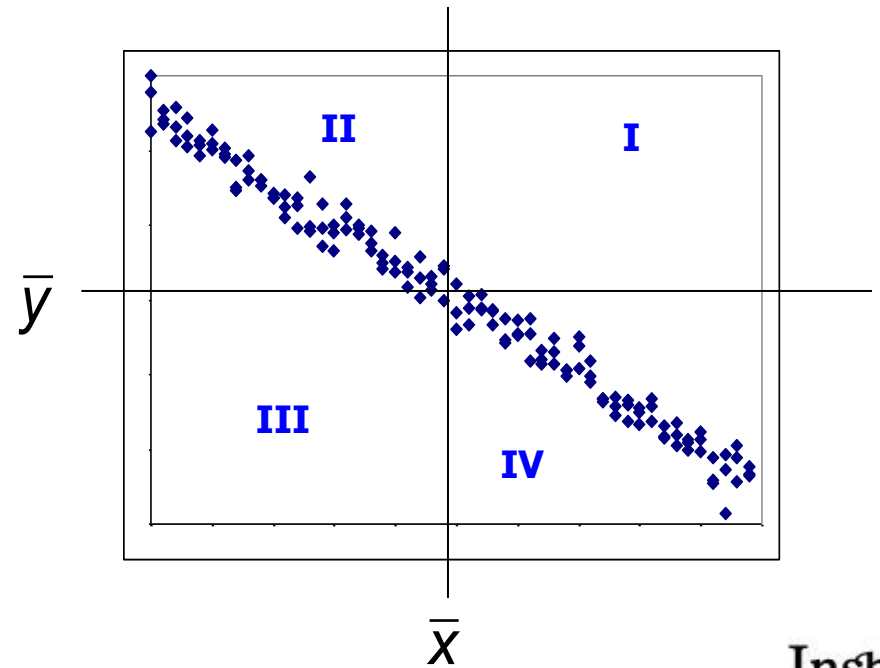
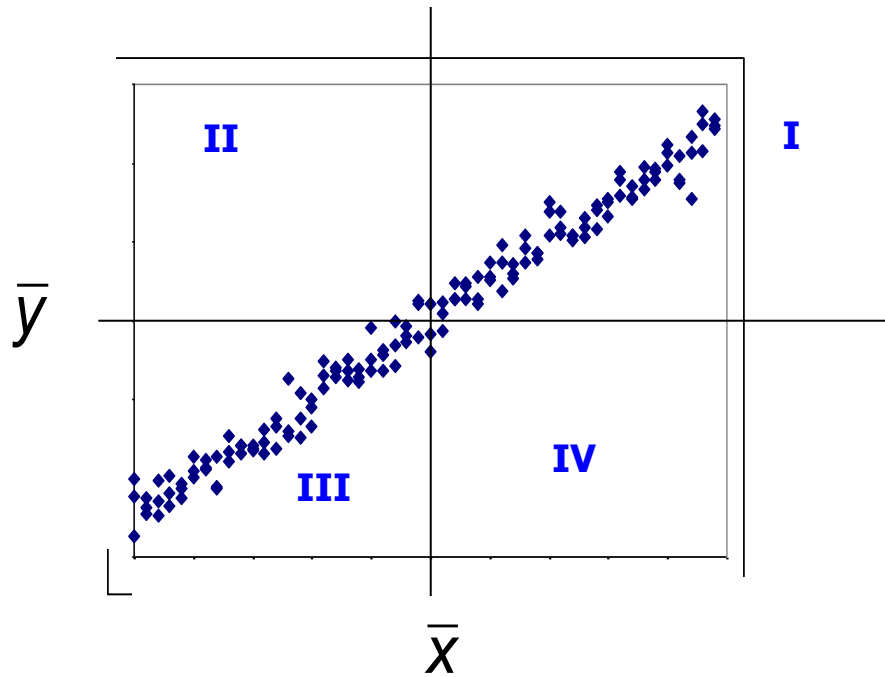
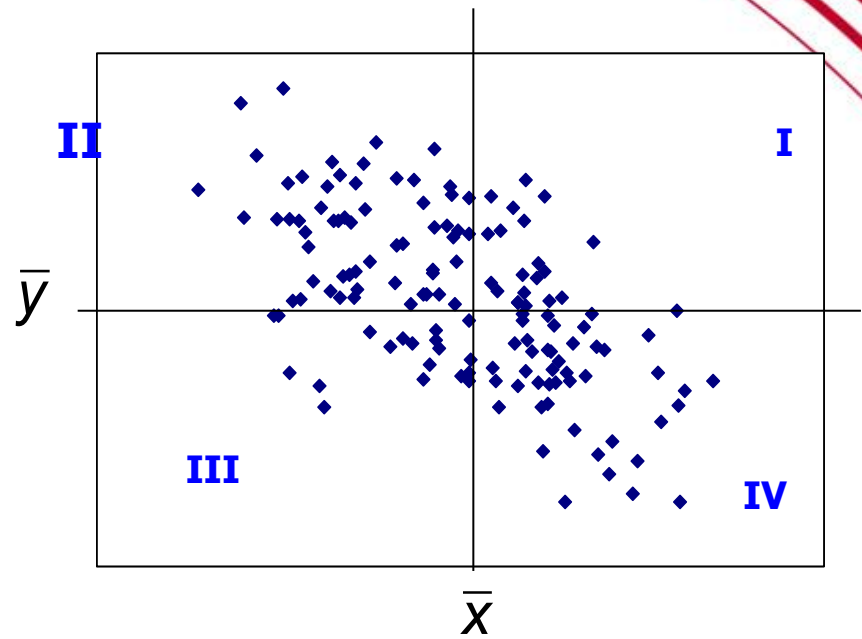
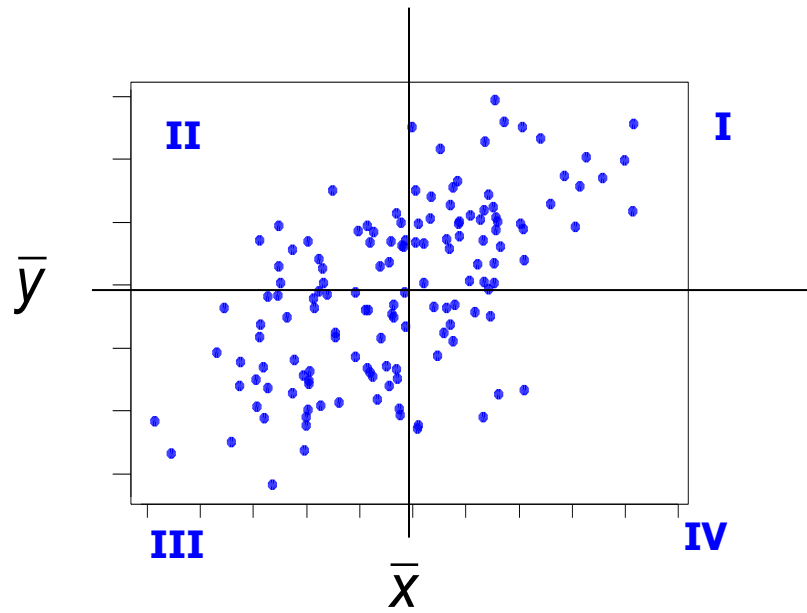
Percebe-se um acúmulo de pontos nos quadrantes ímpares.



Associação Negativa

Percebe-se um acúmulo de pontos nos quadrantes pares.





Comportamento Geral

- ✓ Quando existe uma associação positiva (crescente) entre as variáveis, há um predomínio de pontos nos quadrantes ímpares.
- ✓ Quando existe uma associação negativa (decrescente) entre as variáveis, há um predomínio de pontos nos quadrantes pares.
- ✓ Quanto mais próxima de uma reta estiverem os pontos, maior é o predomínio nos quadrantes ímpares (se crescente) ou pares (se decrescente).

Exemplo: Gráfico de dispersão e medidas de associação

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

Ano: 1997

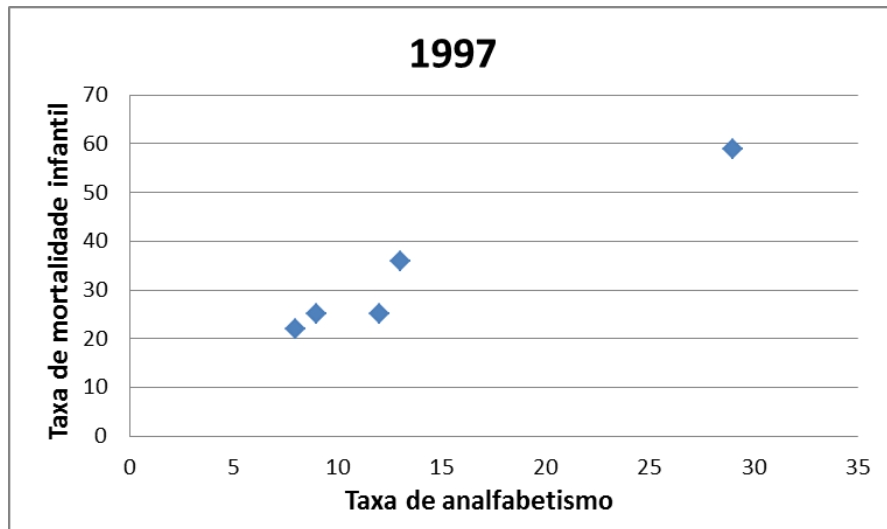
Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Ano: 2009

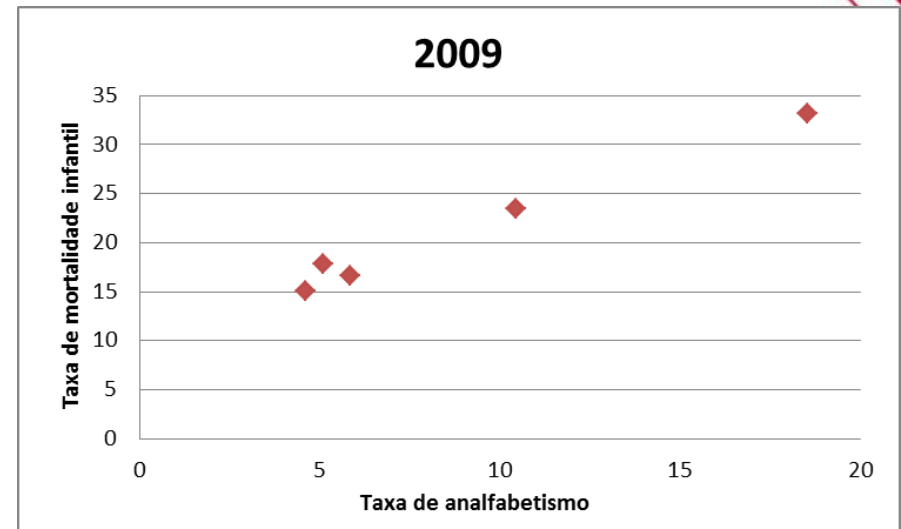
Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	10,45	23,5
Nordeste	18,53	33,2
Sudeste	5,84	16,6
Sul	4,62	15,1
Centro Oeste	5,09	17,8

Fonte: IBGE.

$$\text{Cov}(X,Y)= 101,72$$



$$\text{Cov}(X,Y)= 34,45$$



O gráfico azul possui um coeficiente de covariância maior, mas a associação não parece ser mais forte do que a observada no gráfico vermelho.

Como resolver isso?

Padronização da covariância

Resultado teórico $|\text{Cov}(X, Y)| \leq \text{dp}(X) \text{dp}(Y)$

Consequência $-1 \leq \frac{\text{Cov}(X, Y)}{\text{dp}(X) \text{dp}(Y)} \leq 1$

$$r = \frac{\text{Cov}(X, Y)}{\text{dp}(x) \text{dp}(y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Coeficiente de Correlação Linear

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{DP}(X)\text{DP}(Y)}$$

Vantagem em relação à covariância:

$$-1 \leq r \leq 1$$

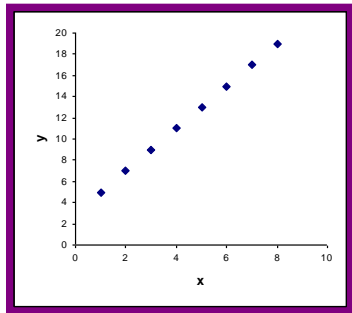
Quando $r = 1$ ou $r = -1$ os pontos estarão perfeitamente alinhados sobre uma reta.

Propriedades do coeficiente de correlação

- ✓ Medida de associação linear entre duas variáveis quantitativas (varia entre -1 e $+1$).
- ✓ Valores próximos a $+1$: indicam forte relação linear positiva
- ✓ Valores próximos a -1 : indicam forte relação linear negativa
- ✓ Valores próximos a zero: indicam ausência de relação linear.

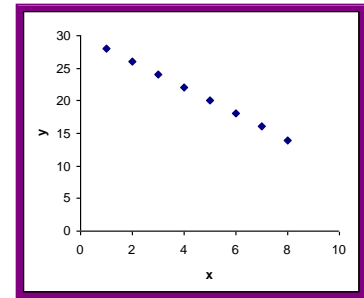
Interpretação do Coeficiente de Correlação

Relação perfeita

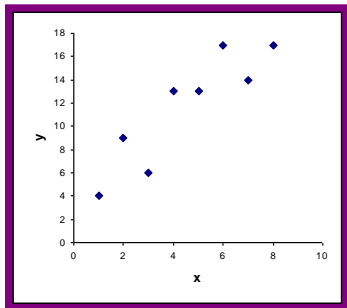


$$r = +1$$

Relação perfeita

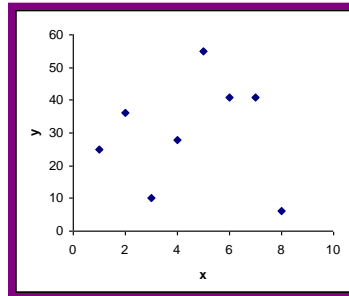


$$r = -1$$

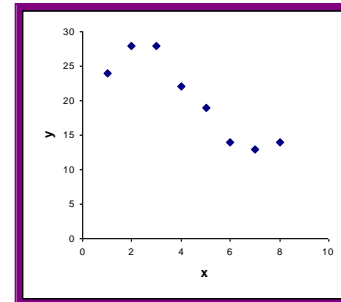


$$r \approx 0,80$$

**Ausência
de relação**



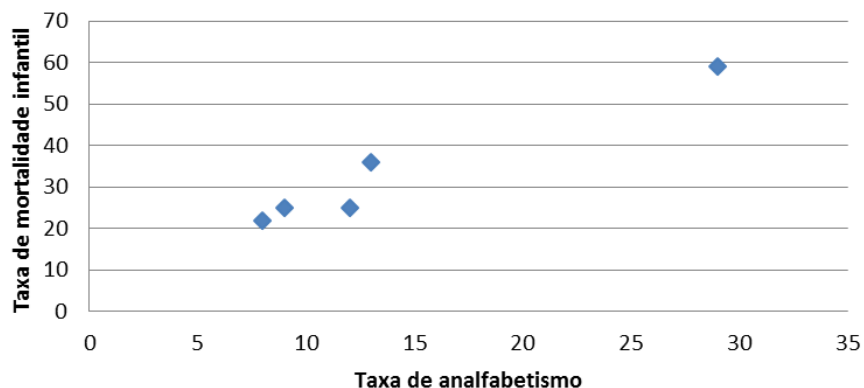
$$r \approx 0$$



$$r \approx -0,80$$

Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região.

1997



Covariância

101,72

Correlação

0,976

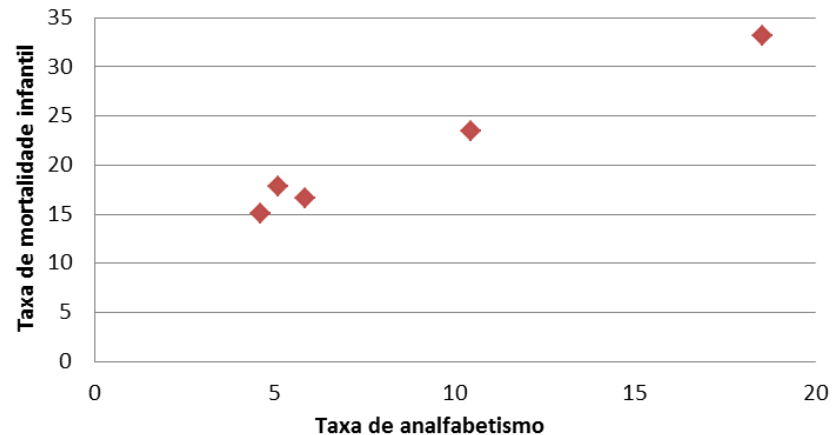
Covariância

34,45

Correlação

0,993

2009



Associação entre duas variáveis quantitativas

Ajuste de Curvas

Outro Problema

De acordo com os valores mensurados de **Taxa de mortalidade infantil e taxa de analfabetismo no Brasil, segundo região** para cada região brasileira, em 2009:

Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	10,45	23,5
Nordeste	18,53	33,2
Sudeste	5,84	16,6
Sul	4,62	15,1
Centro Oeste	5,09	17,8

Qual deve ser a taxa de mortalidade infantil prevista (ou esperada) (ou média) se uma região passar a ter taxa de analfabetismo igual a 8?

Análise de Regressão

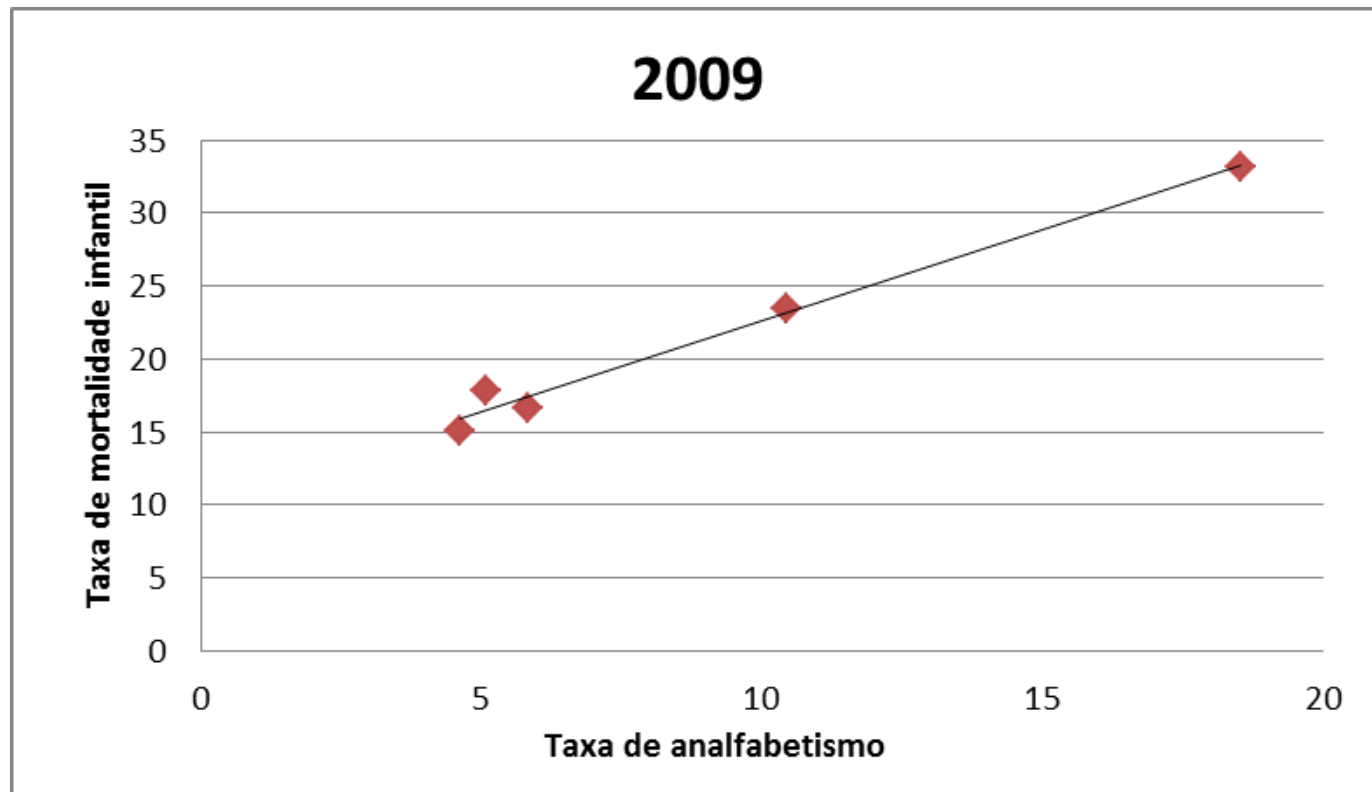
Objetivo: Explicar como uma variável se comporta em função de outra.

Variável dependente (resposta): variável de interesse, cujo comportamento se deseja explicar.

Variável independente (explicativa): variável que é utilizada para explicar a variável dependente.

Modelo de regressão: equação (reta) que associa y e x .

Ajuste no Exemplo do IBGE



y (var. dep.) = Taxa de mortalidade infantil

x (var. indep.) = taxa de analfabetismo

Como é o comportamento de **y** em função de **x**?

Ajuste de Reta

y: variável de interesse (também conhecida como dependente ou resposta)

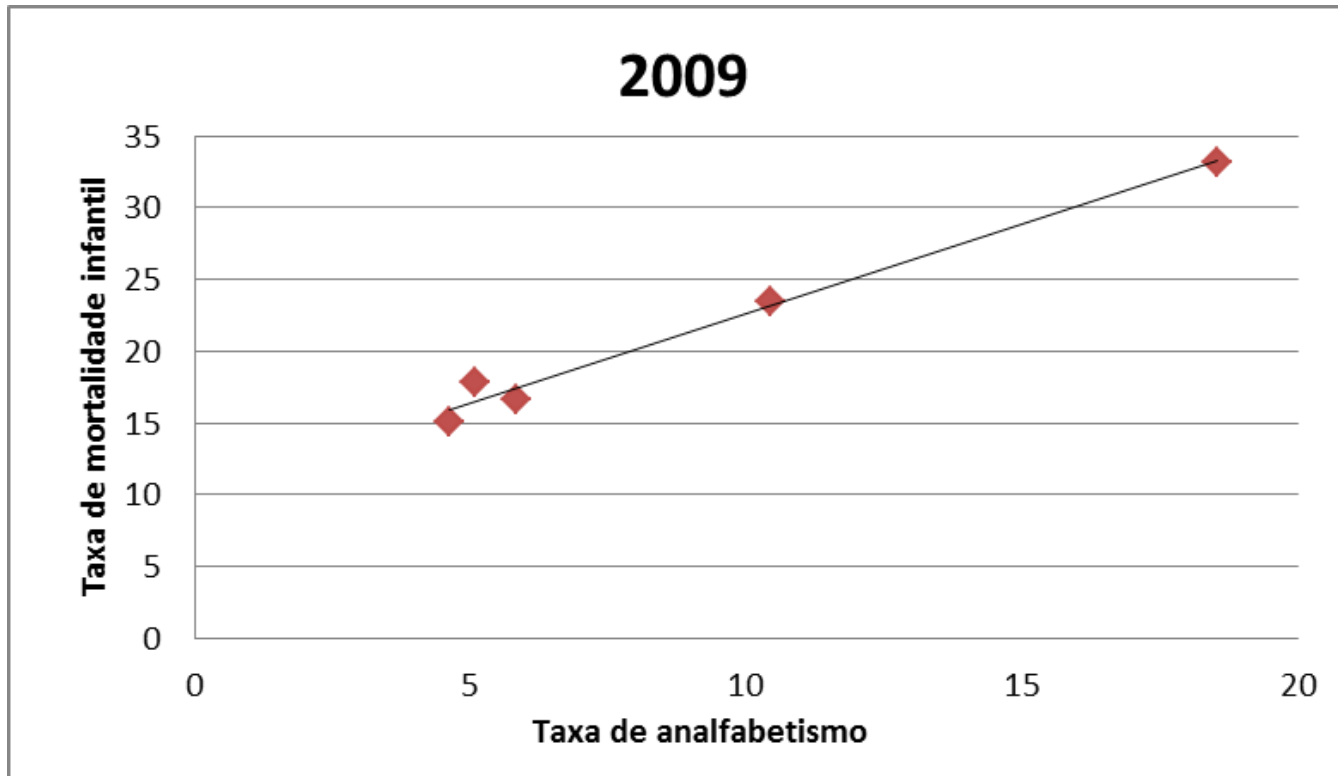
x: variável explicativa

\hat{y} : reta ajustada

$$\hat{y} = ax + b$$

Como obter a e b?

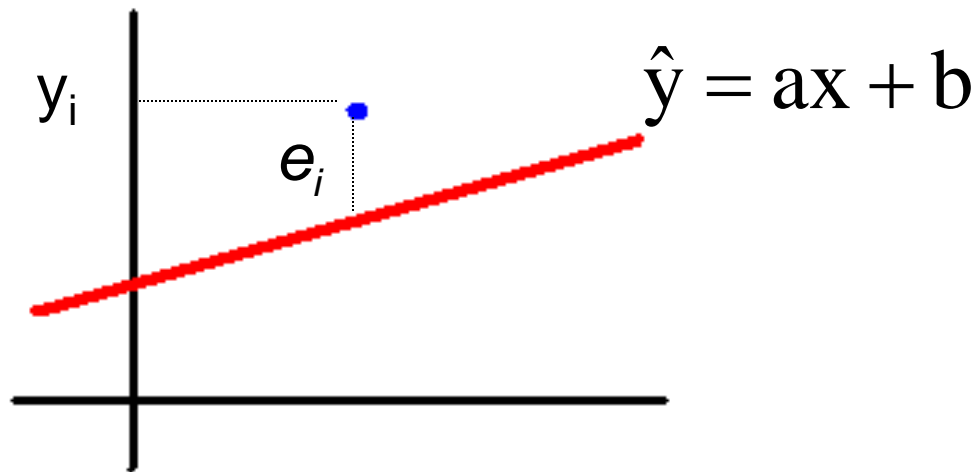
Diagrama de Dispersão



Como
estimar
essa reta?

Regressão Linear Simples

Dados: $(x_1, y_1), \dots, (x_n, y_n)$, com y_1, \dots, y_n não correlacionados.



Modelo

$$y_i = ax_i + b + e_i$$

Característica
populacional

Característica
individual

(erro)

Método dos Mínimos Quadrados

Critério de qualidade:

Encontrar a e b que minimizem:

$$S(a; b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - ax_i - b)^2$$

Como encontrar a e b ?

Método dos Mínimos Quadrados

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} -2 \sum_{i=1}^n x_i (y_i - a x_i - b) = 0 \\ -2 \sum_{i=1}^n (y_i - a x_i - b) = 0 \end{cases}$$

$$a = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

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

Método dos Mínimos Quadrados

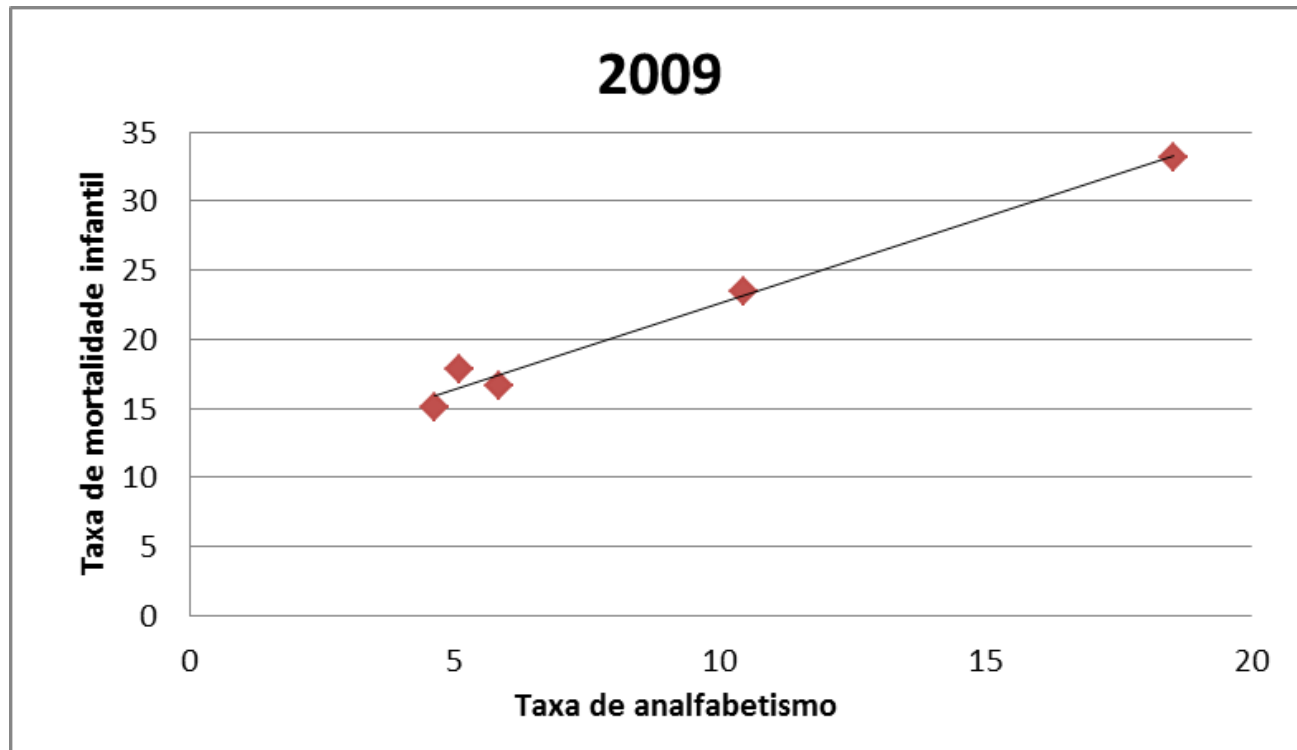
A melhor reta que descreve a relação linear entre **y** e **x** ocorre quando **a** e **b** assumem as seguintes expressões:

$$a = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \text{Corr}(X, Y) \frac{\text{DP}(Y)}{\text{DP}(X)}$$

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

sendo $\text{DP}(Y)$ e $\text{DP}(X)$ os desvios padrões de Y e X , respectivamente; $\text{Cov}(X, Y)$ e $\text{Corr}(X, Y)$ a covariância e a correlação, respectivamente, entre X e Y .

Ajuste no Exemplo do IBGE



Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Média	8,91	21,24
Desvio padrão	5,24	6,62
Covariância	34,45	

Encontre a e b:

$$\hat{y} = ax + b$$

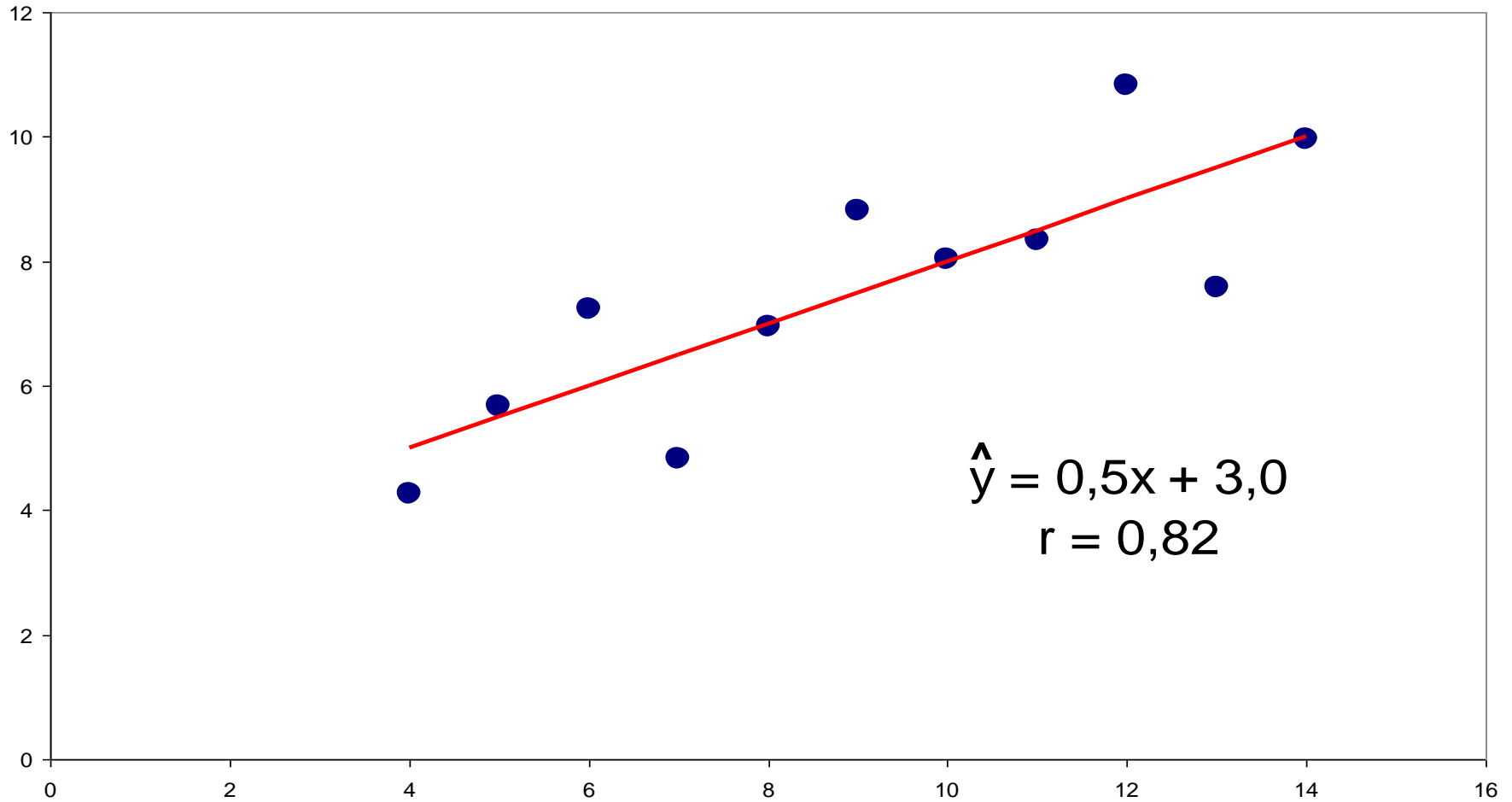
Interpretação dos coeficientes

- 1,255: variação média (esperada) na taxa de mortalidade infantil quando aumenta 1 unidade na taxa de analfabetismo.
- 10,058: taxa esperada de mortalidade infantil quando a região tem valor zero para taxa de analfabetismo.

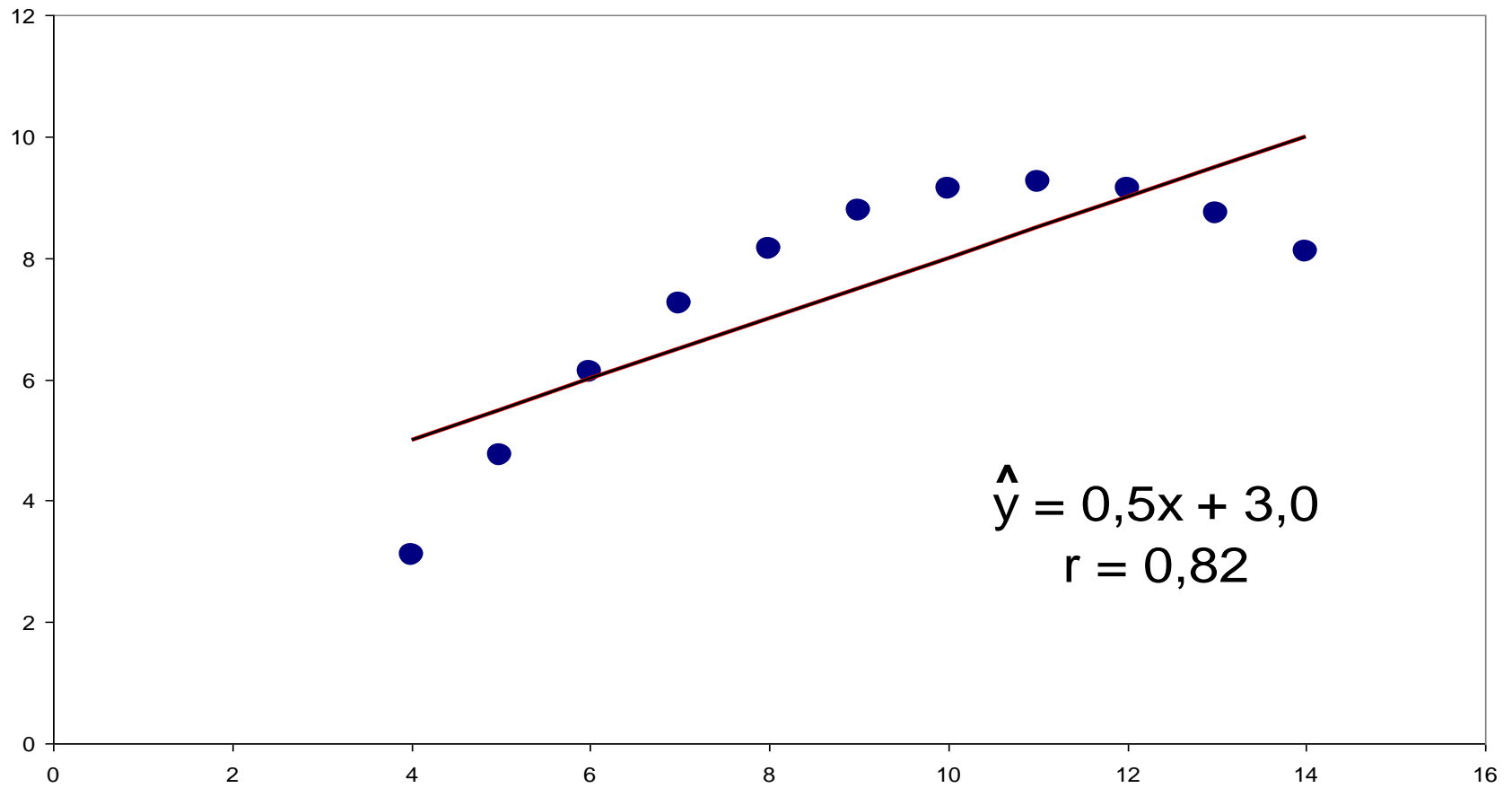
Na prática, nem sempre o intercepto tem uma interpretação que faz sentido no problema.

Limitações no uso do coeficiente de correlação (r)

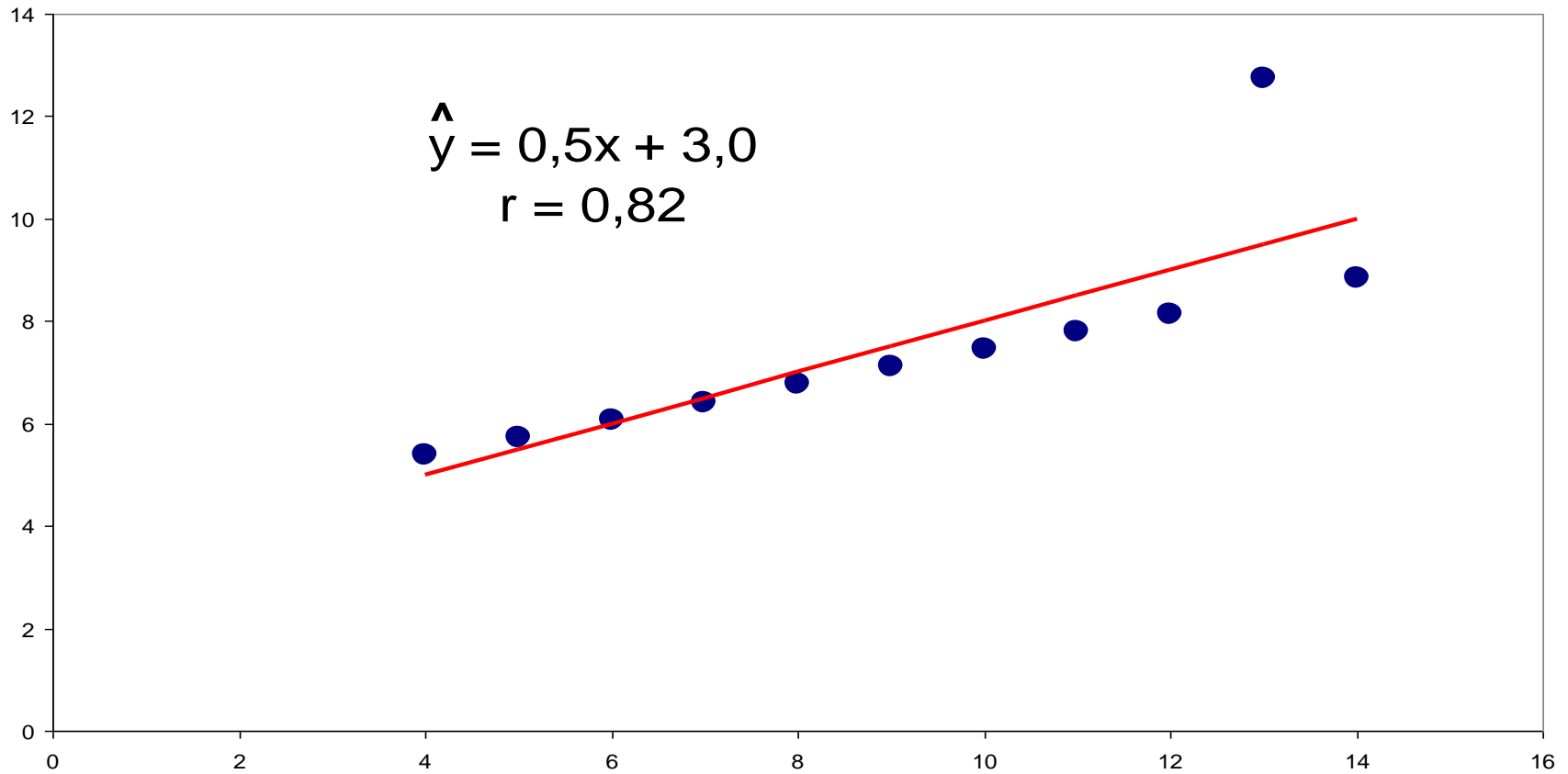
y1 vs x1



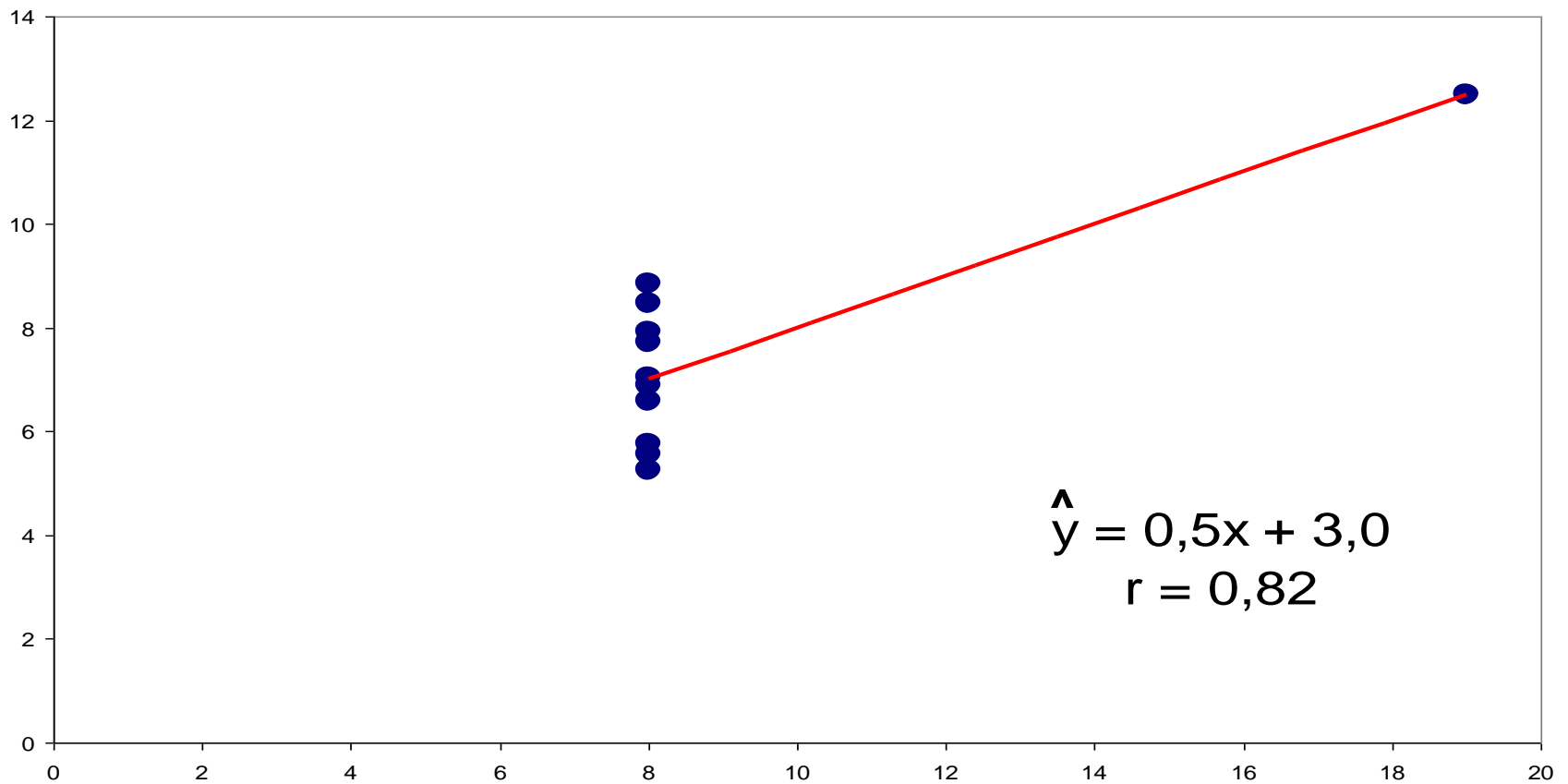
y2 vs x1



y3 vs x1

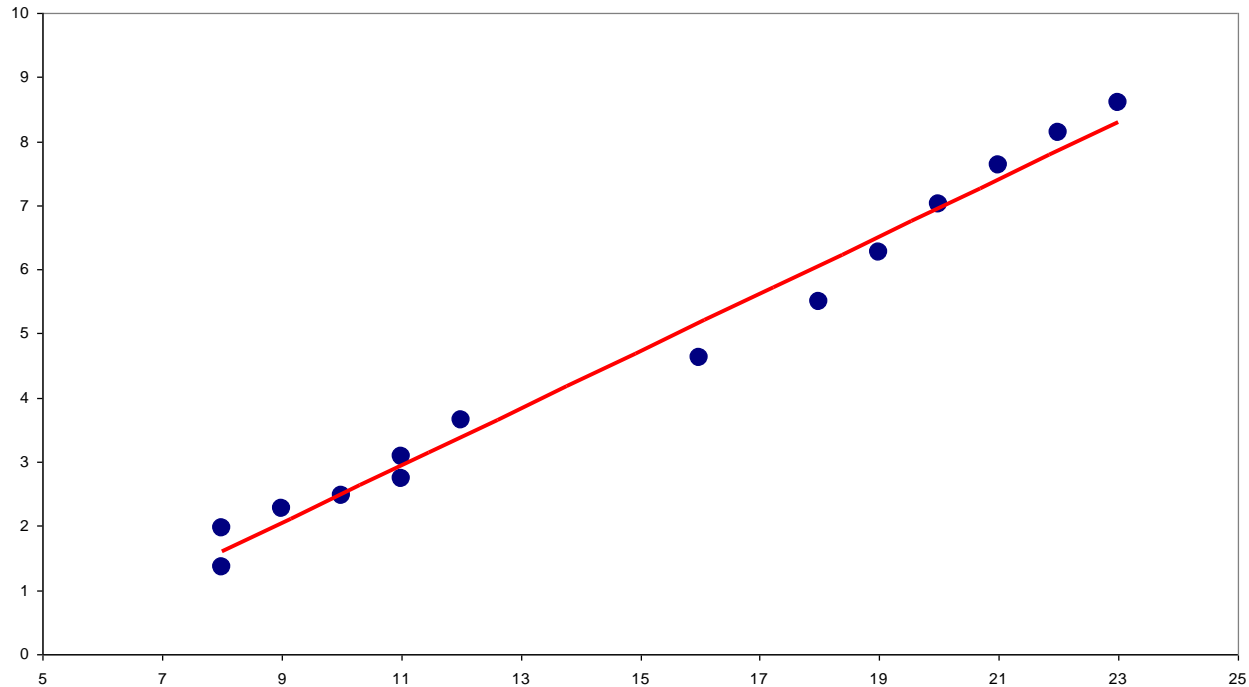


y4 vs x2



Mau uso da regressão

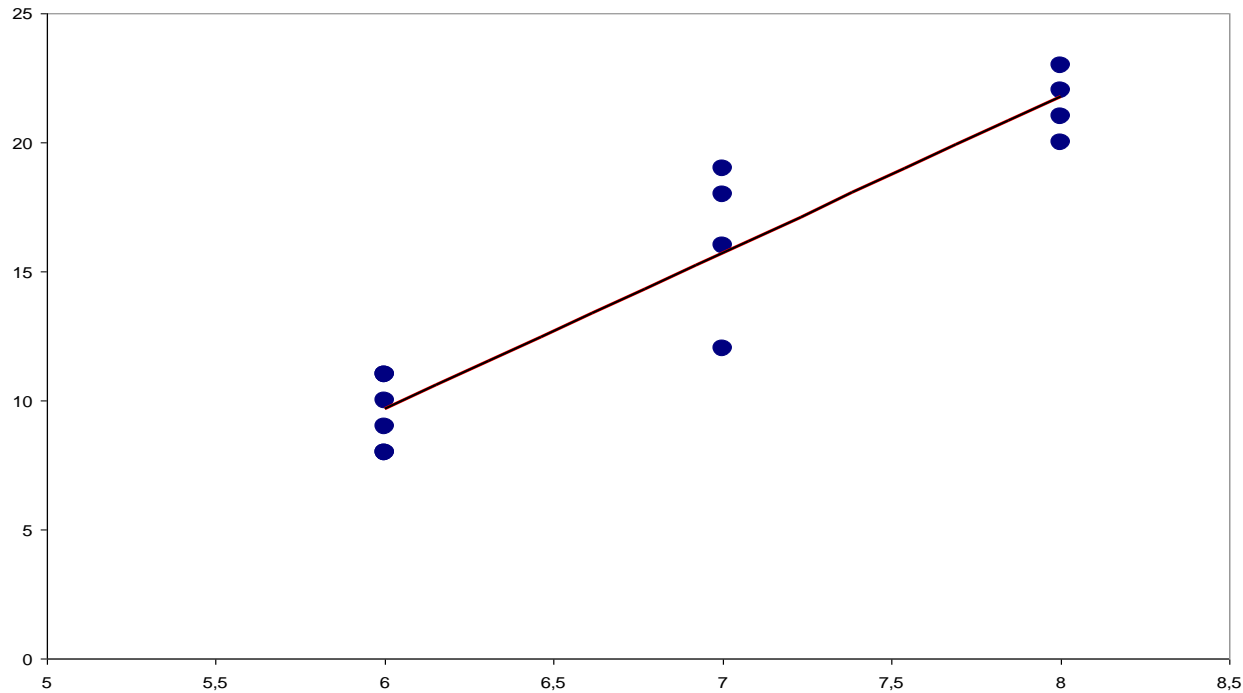
Regressão de y vs x_1



$$\hat{y} = 0,4465x - 1,9722$$

$$r = 0,99$$

Regressão de y vs x_2



$$\hat{y} = 6,0441x - 26,588$$

$$r = 0,94$$

Associação não é causalidade

Suponha que encontremos alta correlação entre duas variáveis A e B. Podem existir diversas explicações do porque elas variam conjuntamente, incluindo:

- Mudanças em outras variáveis causam mudanças tanto em A quanto em B.
- Mudanças em A causam mudanças em B.
- Mudanças em B causam mudanças em A.
- A relação observada é somente uma coincidência (correlação espúria).

A primeira explicação é frequentemente a mais apropriada. Isto indica que existe algum processo de conexão atuando.

Fonte: <http://leg.ufpr.br/~silvia/CE003/node77.html>

Atividade para analisar MUNDO...

15 minutos:

Estudo entre as relações de algumas variáveis quantitativas que descrevem comportamento de alguns países de acordo com o % de população que vivem em área urbana.

Arquivo:

Aula06_Atividade1_ExplorandoDuas
VariáveisQuantitativas_Mundo.ipynb

Atividade para analisar Discriminação Salarial...

40 minutos:

A sua empresa está sendo acusada de pagar um salário maior para os homens do que para as mulheres. Para justificar a acusação, apresentou-se uma lista de salários de uma amostra de funcionários.

Você tem motivos para se preocupar?

Arquivo:

Aula06_Atividade2_ExplorandoDuasVariáveisQuantitativas_Discriminacao.ipynb

CRONOGRAMA PROJETO 1

DATA	Finalização:
14/02	APS1: Check descrito na Aula 02 (COM NINJAS ATÉ 20/02)
21/02	APS2: Check descrito na Aula 04 (COM NINJAS ATÉ 25/02)
?/03	PROJETO 1 FINALIZADO (COMBINAR COM ALUNOS) Fazer git push em seu Github até 23:59 do Projeto 1 finalizado.