Insper

Ciência dos Dados

Aula 05

Análise Exploratória dos Dados

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 quanti;
- Descrever o comportamento médio entre duas variáveis quantitativas por meio de um ajuste linear.

Associação entre duas variáveis quantitativas

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

PréAula05
Indicadores sócio-economicos

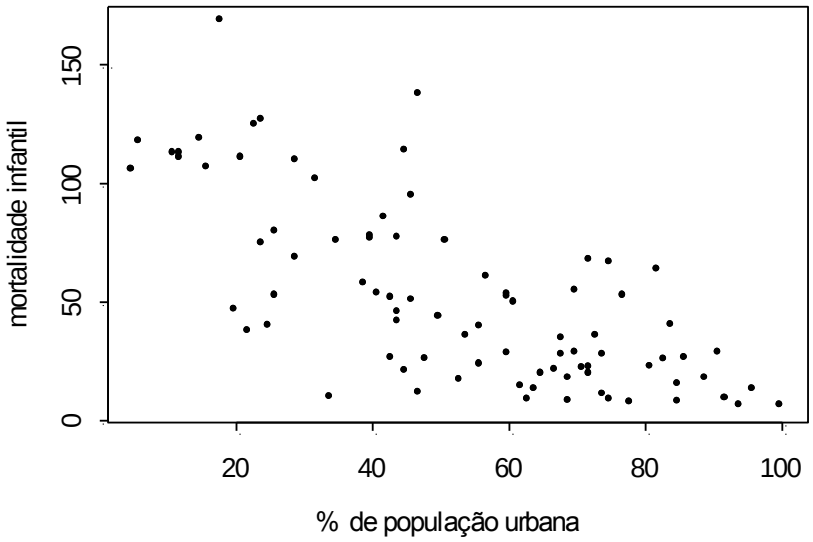
O arquivo **Mundo.xlsx** conta com uma amostra de **85 países**, para os quais levantou-se uma série de indicadores socioeconômicos.

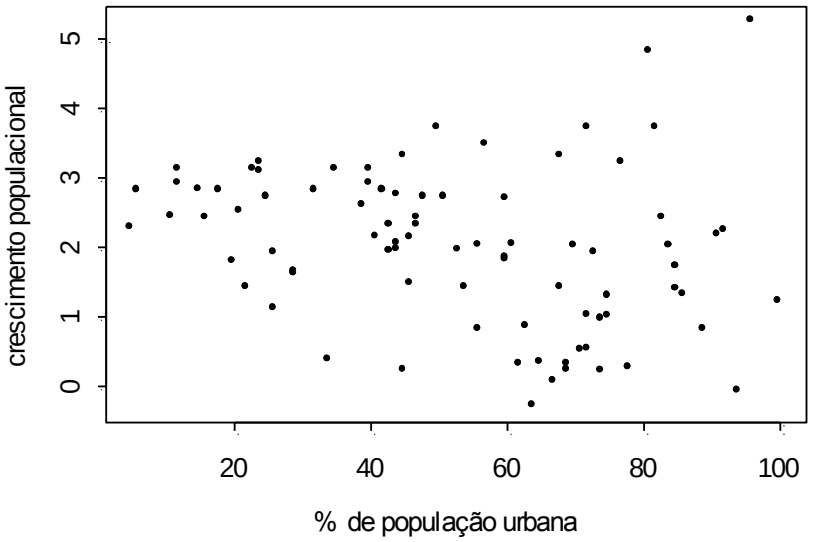
Variáveis:

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

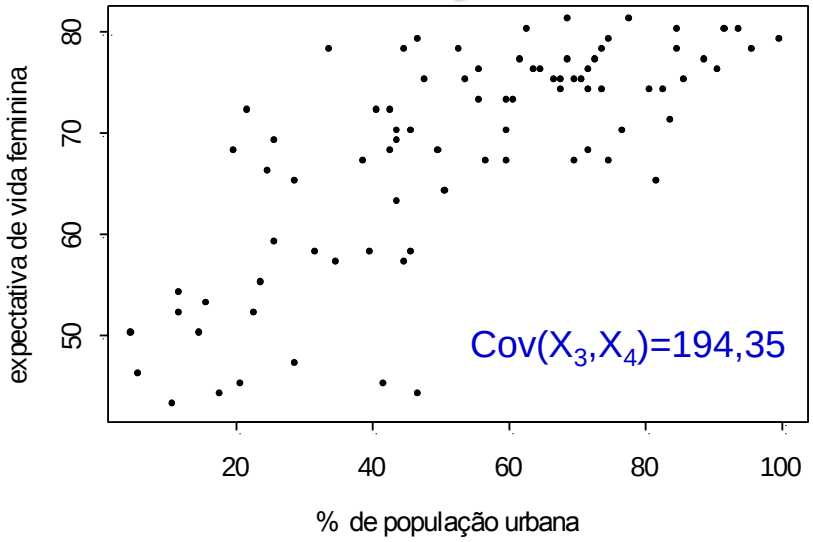

expectativa de vida feminina
50
60
70
80
20
40
60
80
100
% de população urbana


expectativa de vida masculina
40
50
60
70
20
40
60
80
100
% de população urbana

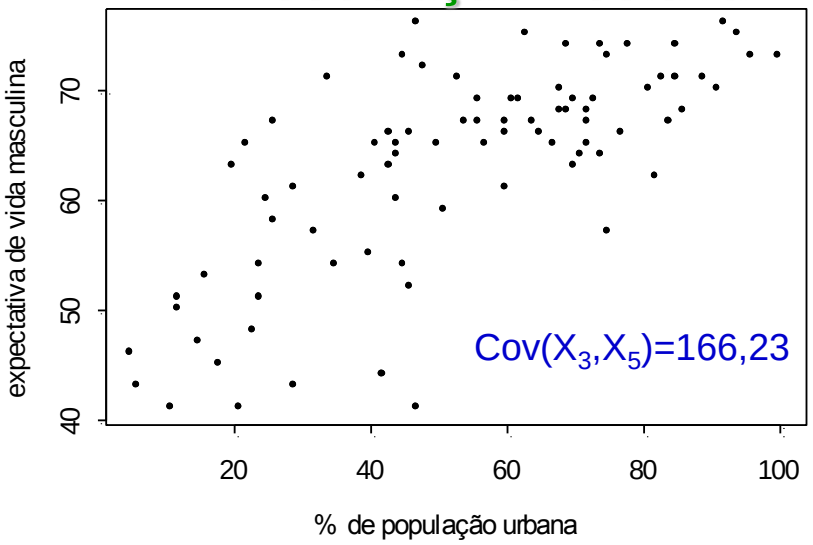

mortalidade infantil
0
50
100
150
20
40
60
80
100
% de população urbana


crescimento populacional
0
1
2
3
4
5
20
40
60
80
100
% de população urbana

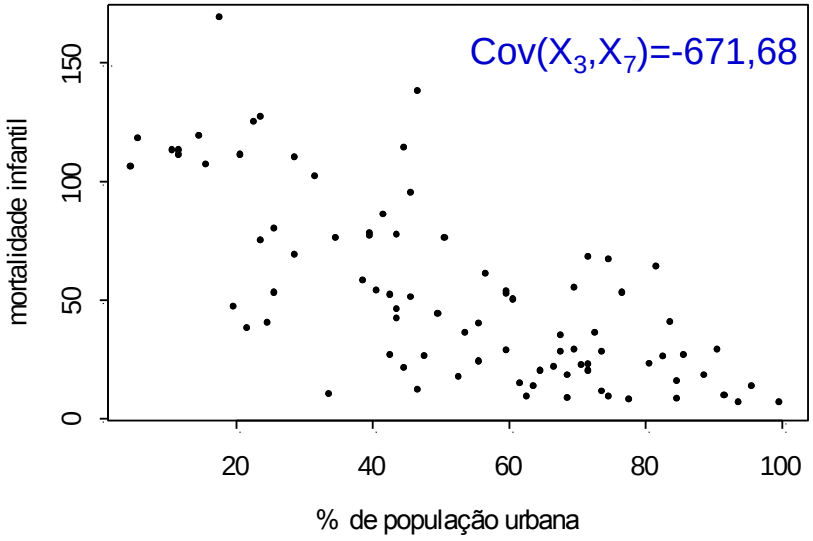
Associação Positiva



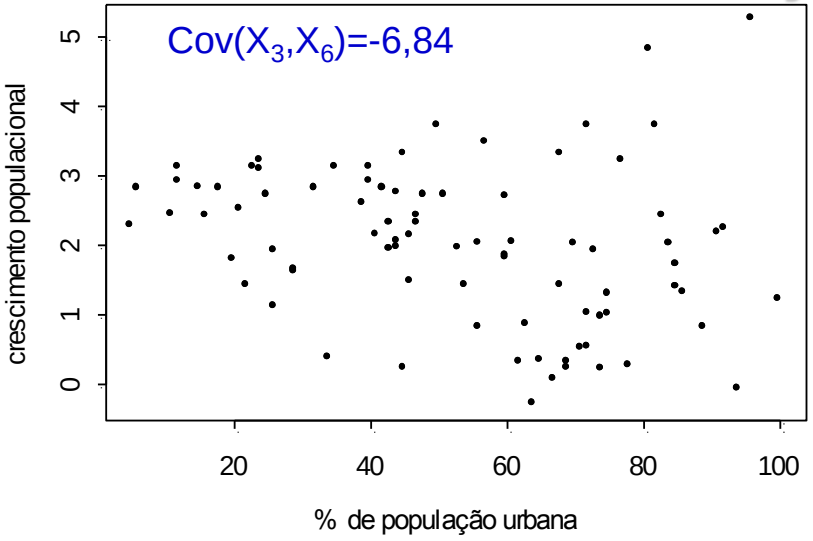
Associação Positiva



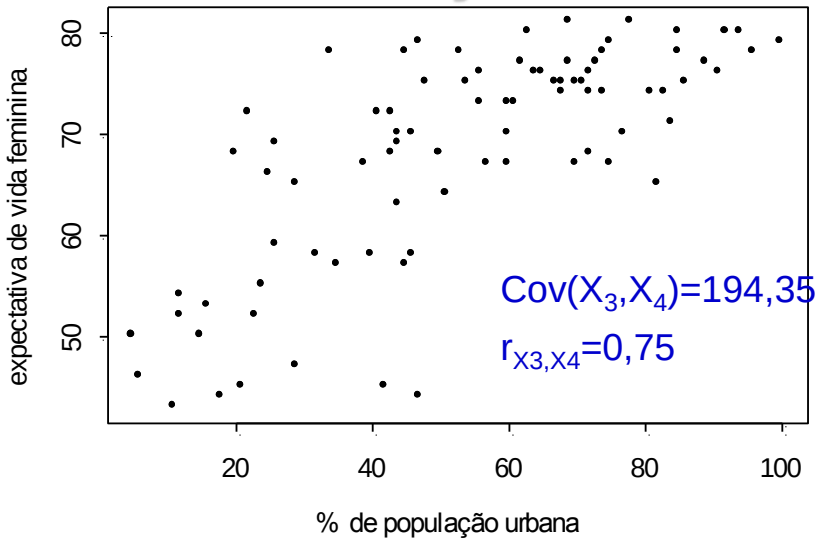
Associação Negativa



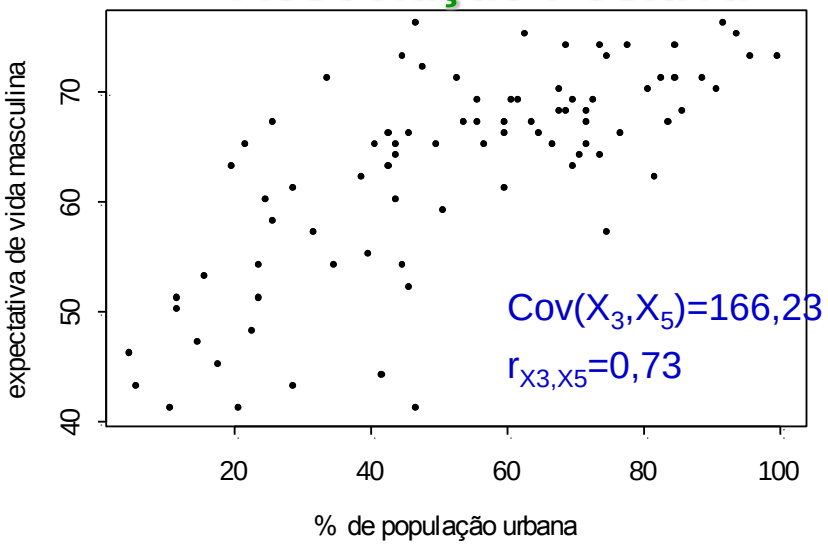
Baixo indício de associação



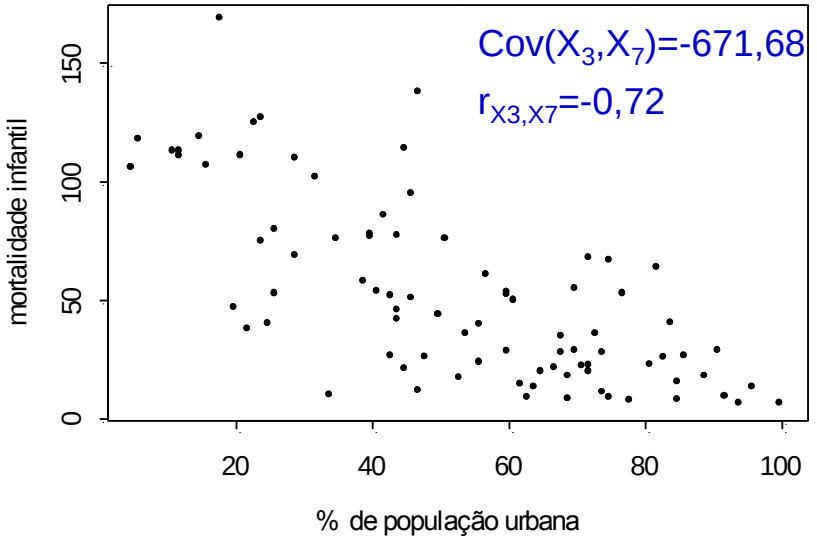
Associação Positiva



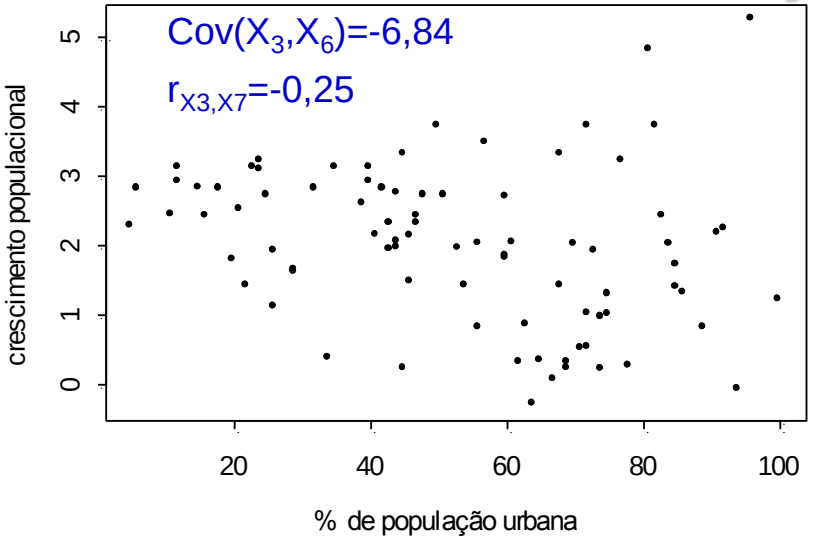
Associação Positiva



Associação Negativa



Baixo indício de associação



Insper

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

Fonte: IBGE.

Taxa de analfabetismo: **Percentual de pessoas com 15 e 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.**

Fonte: RIPSA

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

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

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

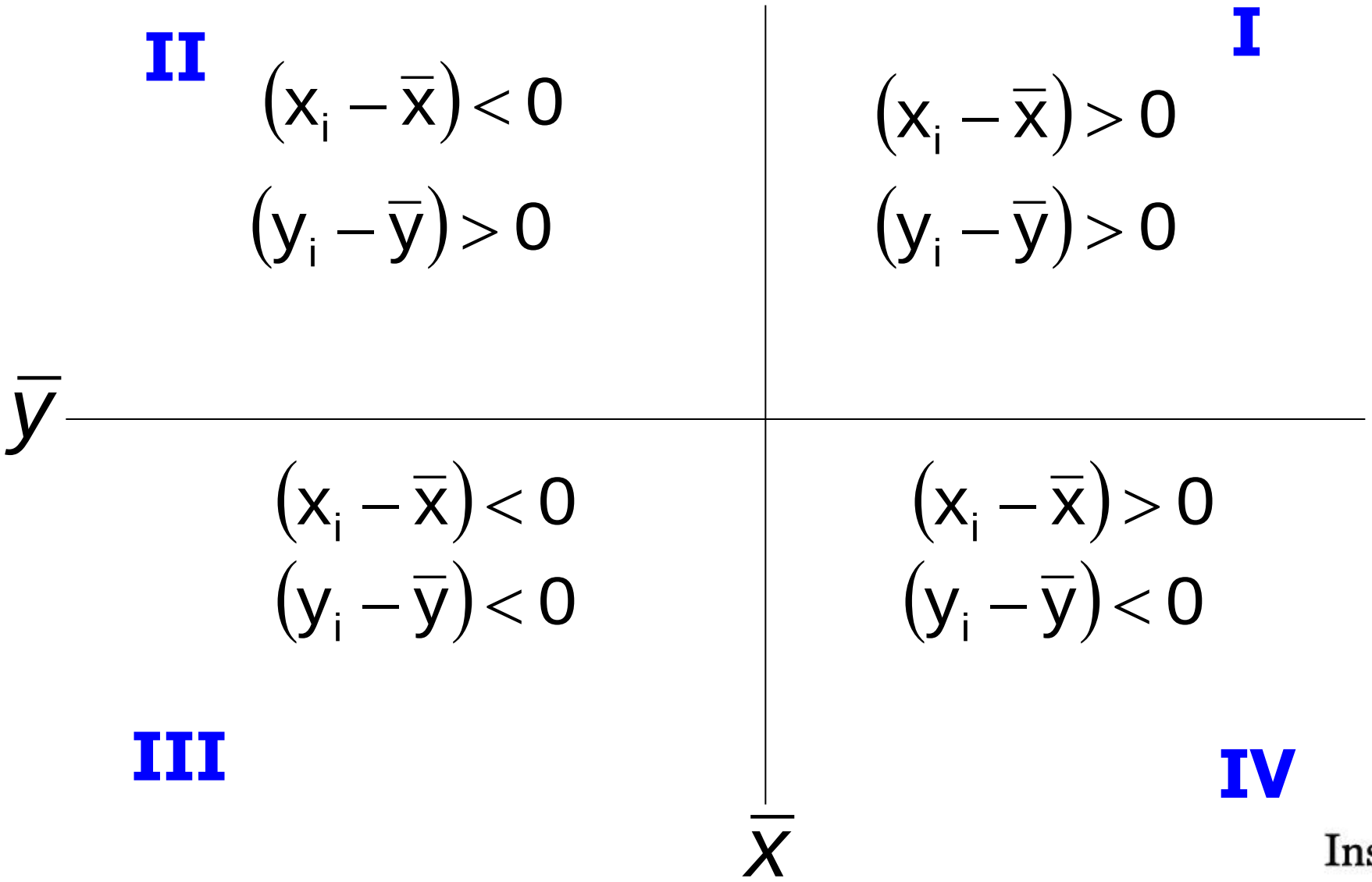
Fonte: IBGE.

Coeficiente de Covariância

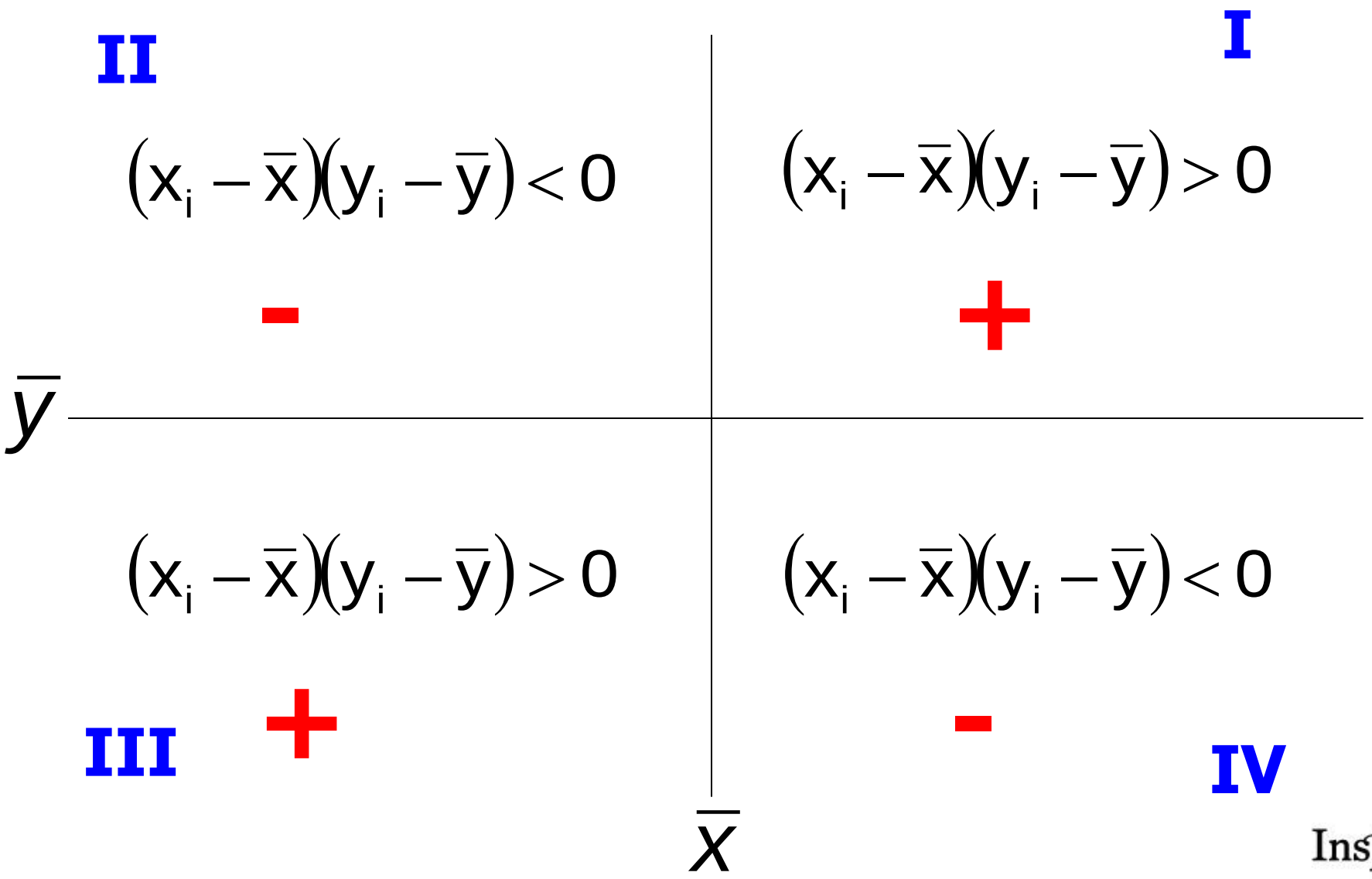
$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

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

Estudo de Sinal

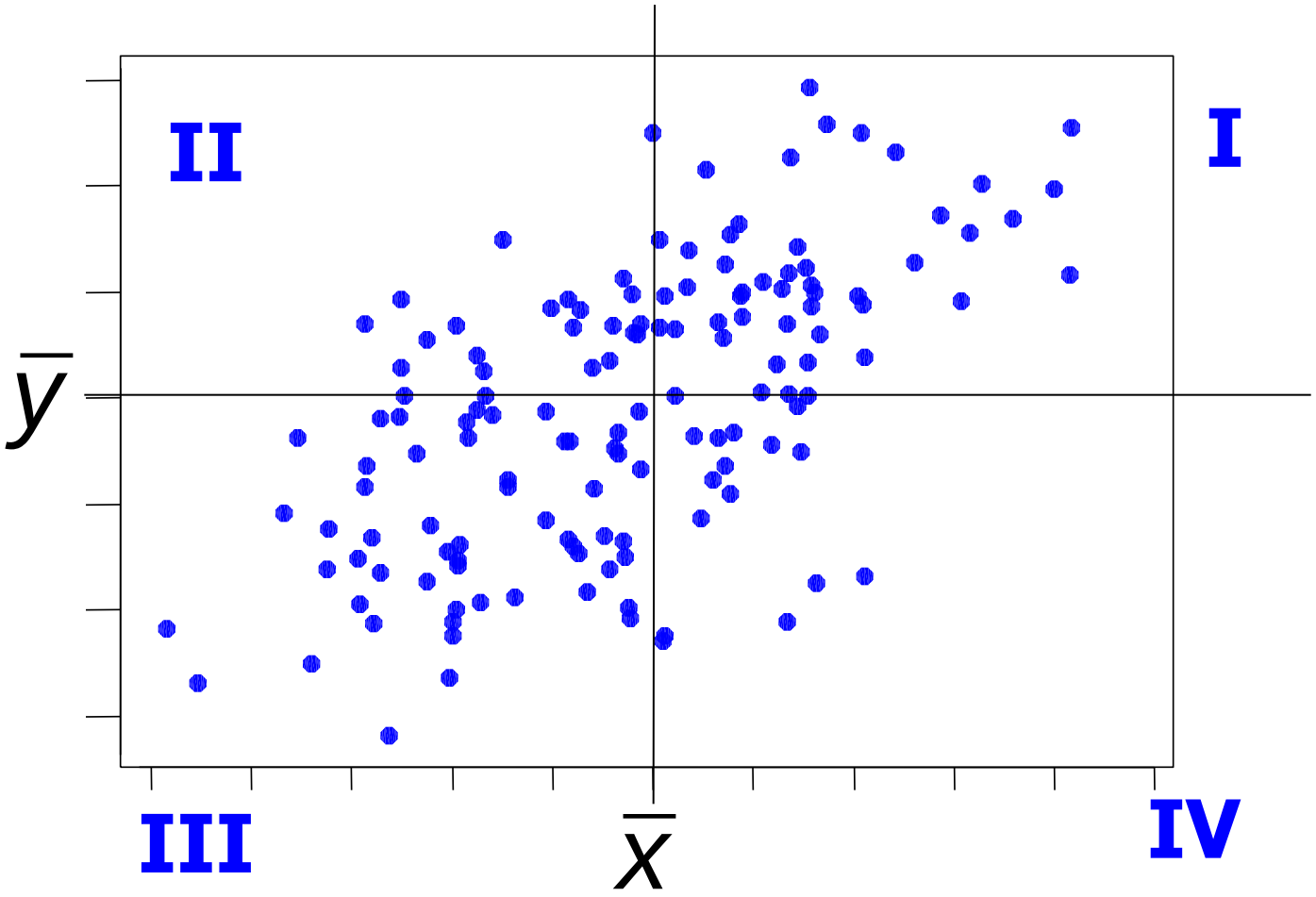


Estudo de Sinal



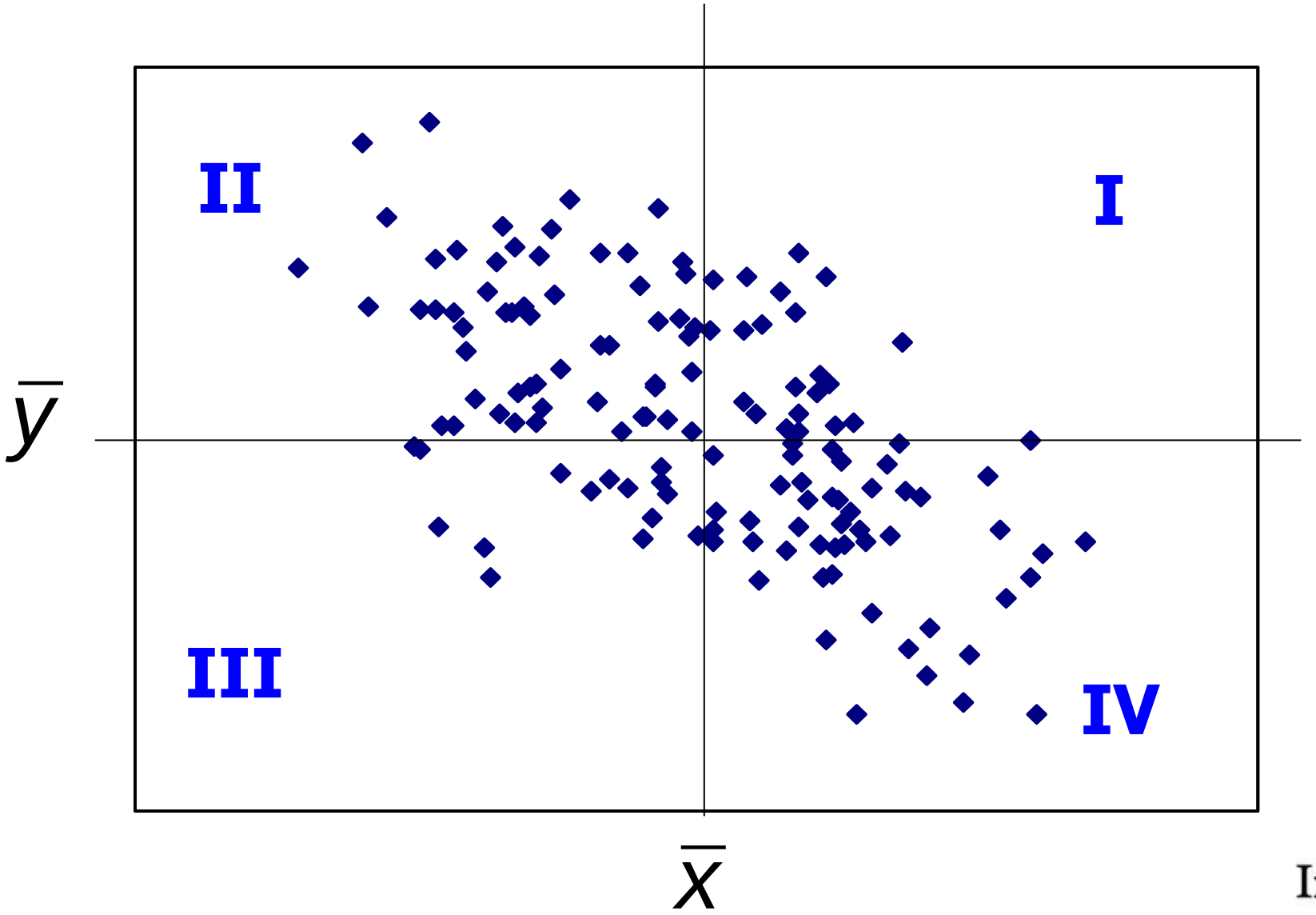
Associação Positiva

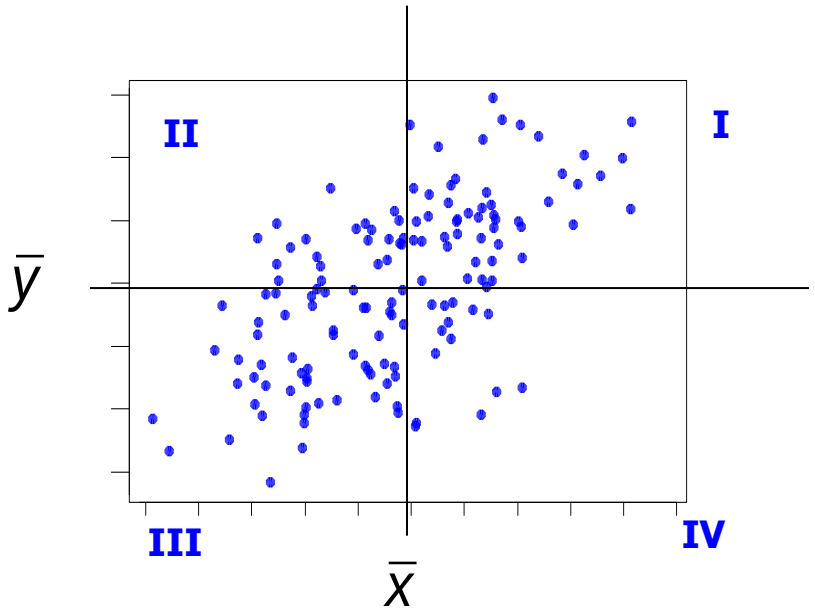
Percebe-se um acúmulo de pontos nos quadrantes ímpares.

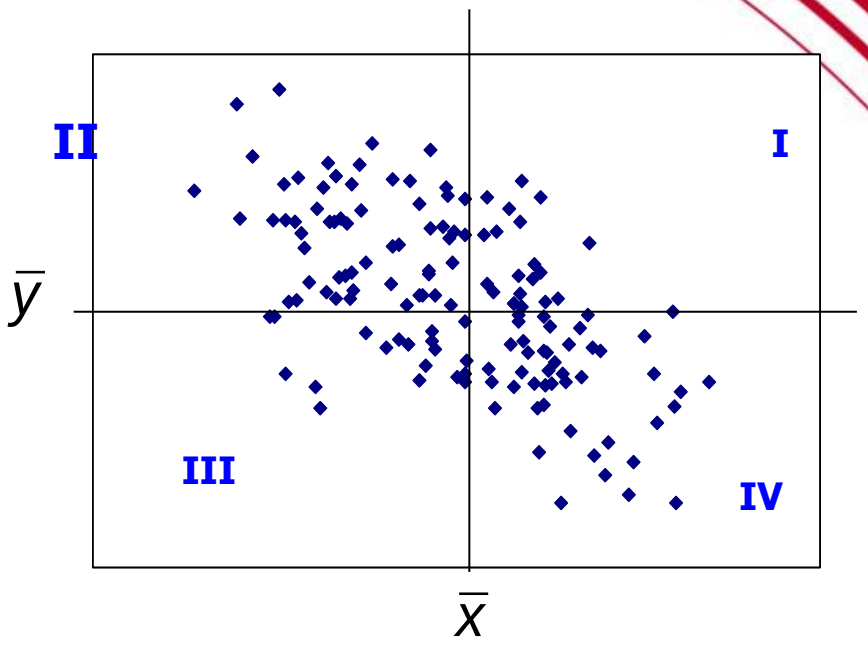


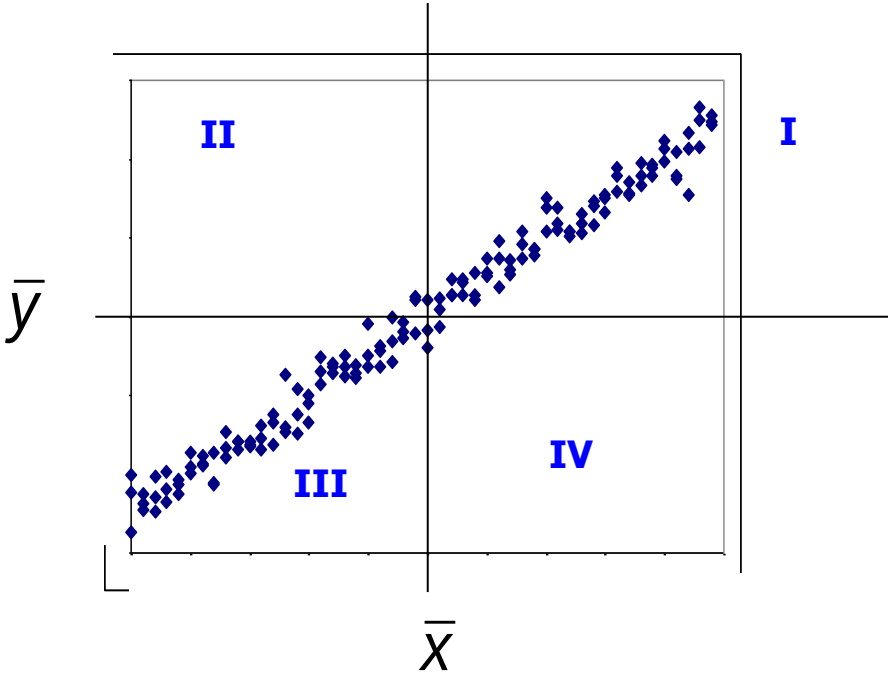
Associação Negativa

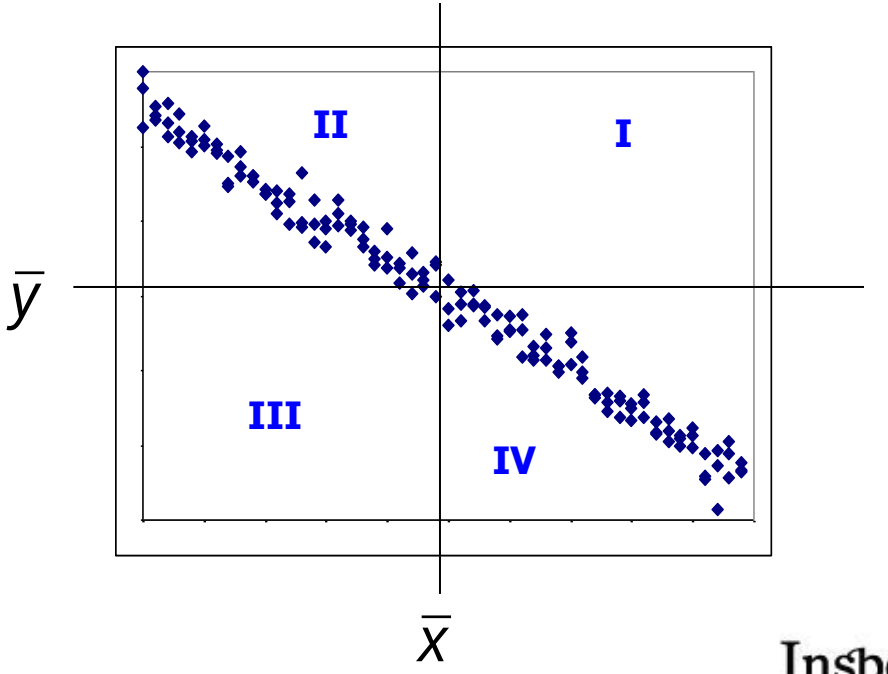
Percebe-se um acúmulo de pontos nos quadrantes pares.




II
I
III
IV
$\bar{y}$
$\bar{x}$


II
I
III
IV
$\bar{y}$
$\bar{x}$


II
I
III
IV
$\bar{y}$
$\bar{x}$


II
I
III
IV
$\bar{y}$
$\bar{x}$

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

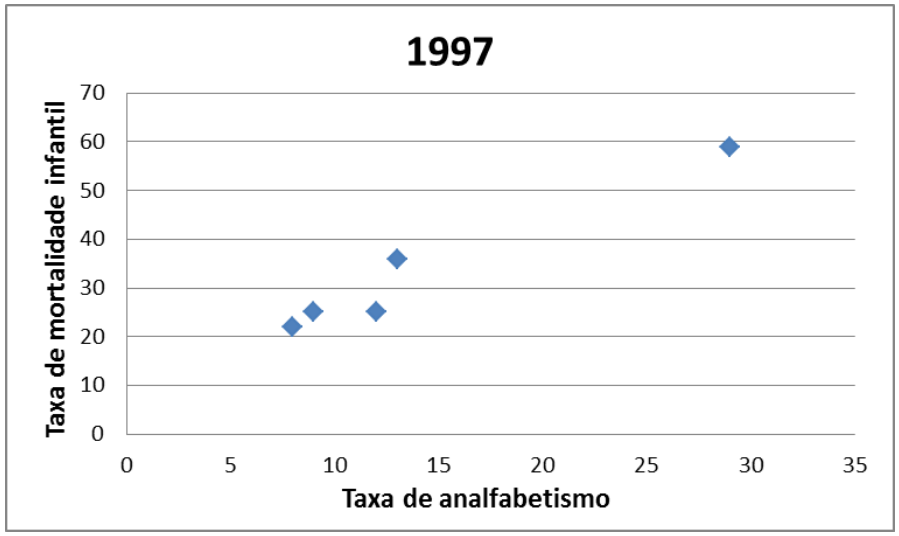
Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	13	36
Nordeste	29	59
Sudeste	9	25
Sul	8	22
Centro Oeste	12	25

Ano: 2009

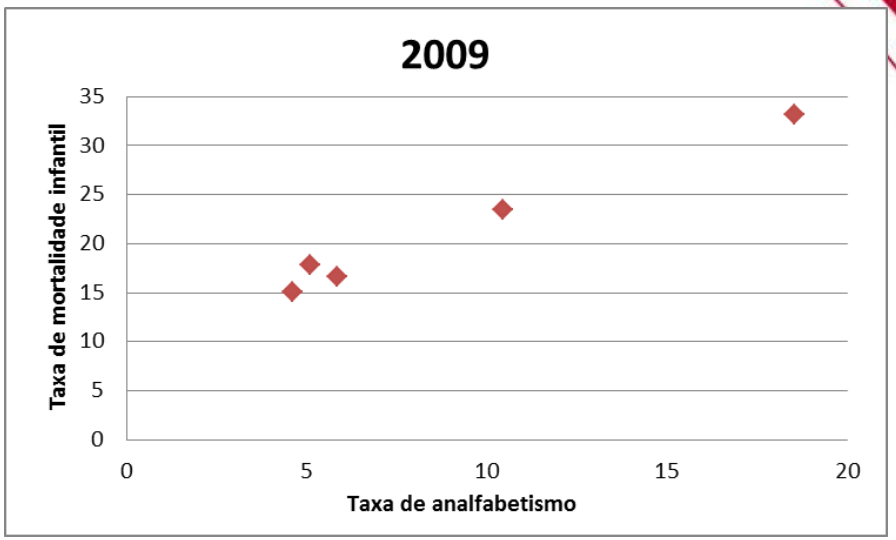
Região	Taxa de analfabetismo	Taxa de mortalidade infantil
Norte	10,45	23,5
Nordeste	18,53	33,2
Sudeste	5,84	16,6
Sul	4,62	15,1
Centro Oeste	5,09	17,8

Fonte: IBGE.

Cov(X,Y)= 101,72



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 $\left|\mathrm{Cov}(X,Y)\right| \leq \mathrm{dp}(X)\,\mathrm{dp}(Y)$

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

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

Coeficiente de Correlação Linear

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

Vantagem em relação à covariância:

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

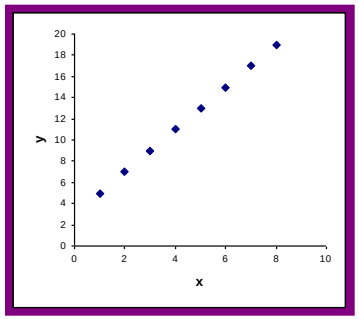
Quando r = 1 ou r = -1 os pontos estarão perfeitamente alinhados sobre uma reta.

Propriedades do coeficiente de correlação

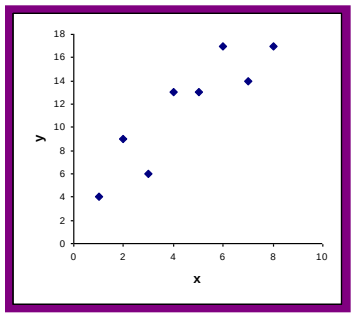
- ✓ Medida de associação linear entre duas variáveis quantitativas (varia entre –1 e +1).
- ✓ Valores próximos a +1 : indicam forte relação linear positiva
- ✓ Valores próximos a -1 : indicam forte relação linear negativa
- ✓ Valores próximos a zero: indicam ausência de relação linear.

Interpretação do Coeficiente de Correlação

Relação perfeita

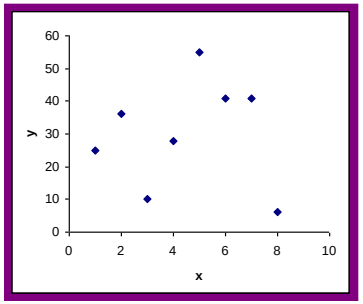


r = +1

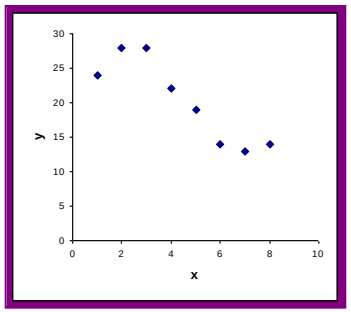


r ≈ 0,80

Ausência de relação

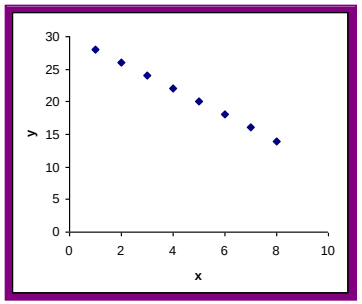


r ≈ 0



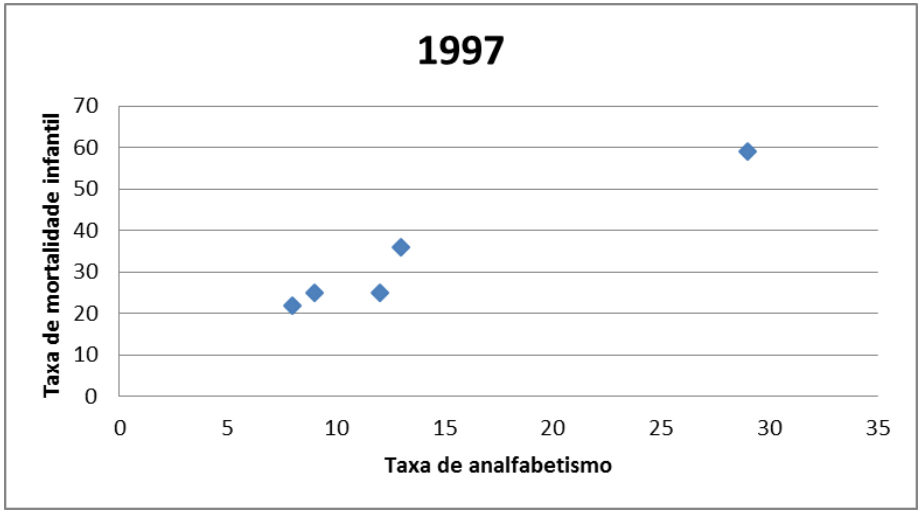
r ≈ -0,80

Relação perfeita



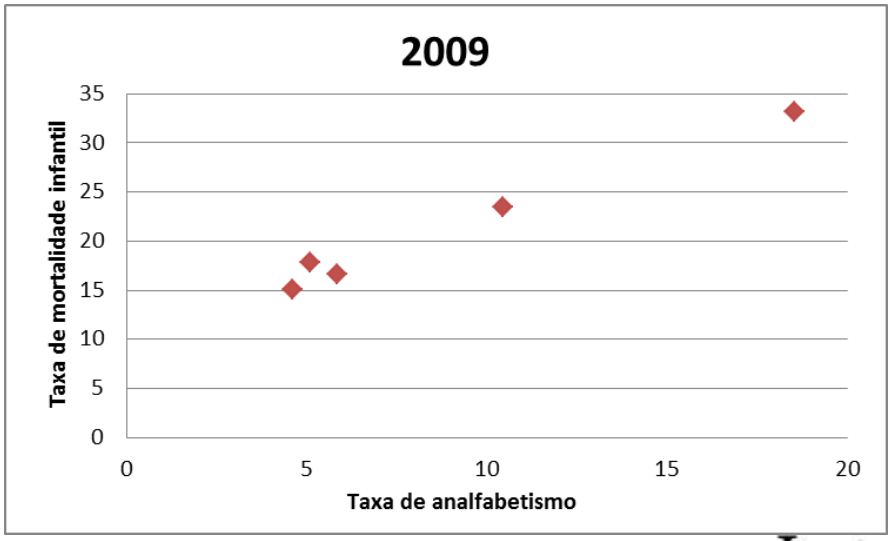
r = -1

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



Covariância	101,72
Correlação	0,976

Covariância	34,45
Correlação	0,993



Fonte: IBGE.

Insper

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

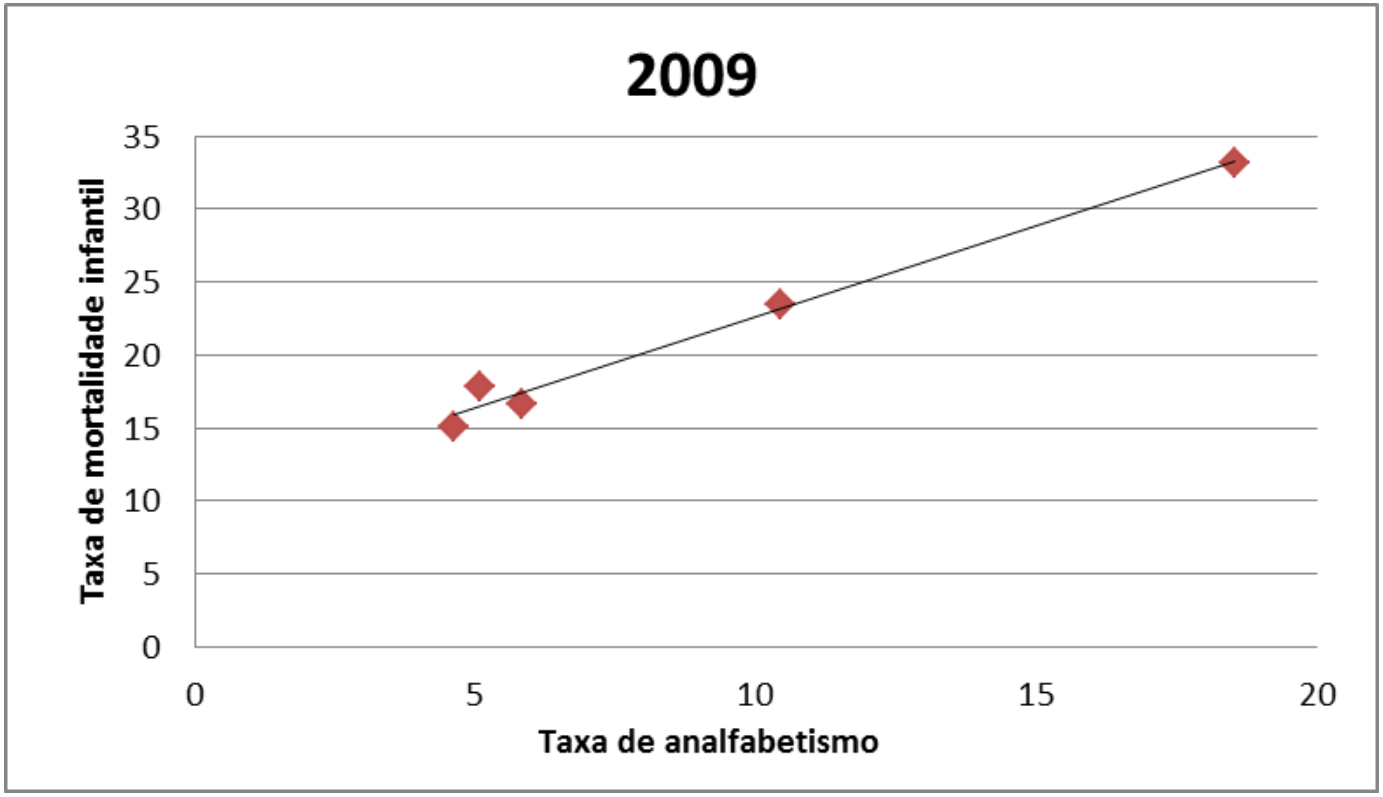
Objetivo: Explicar como uma variável se comporta em função de outra.

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)

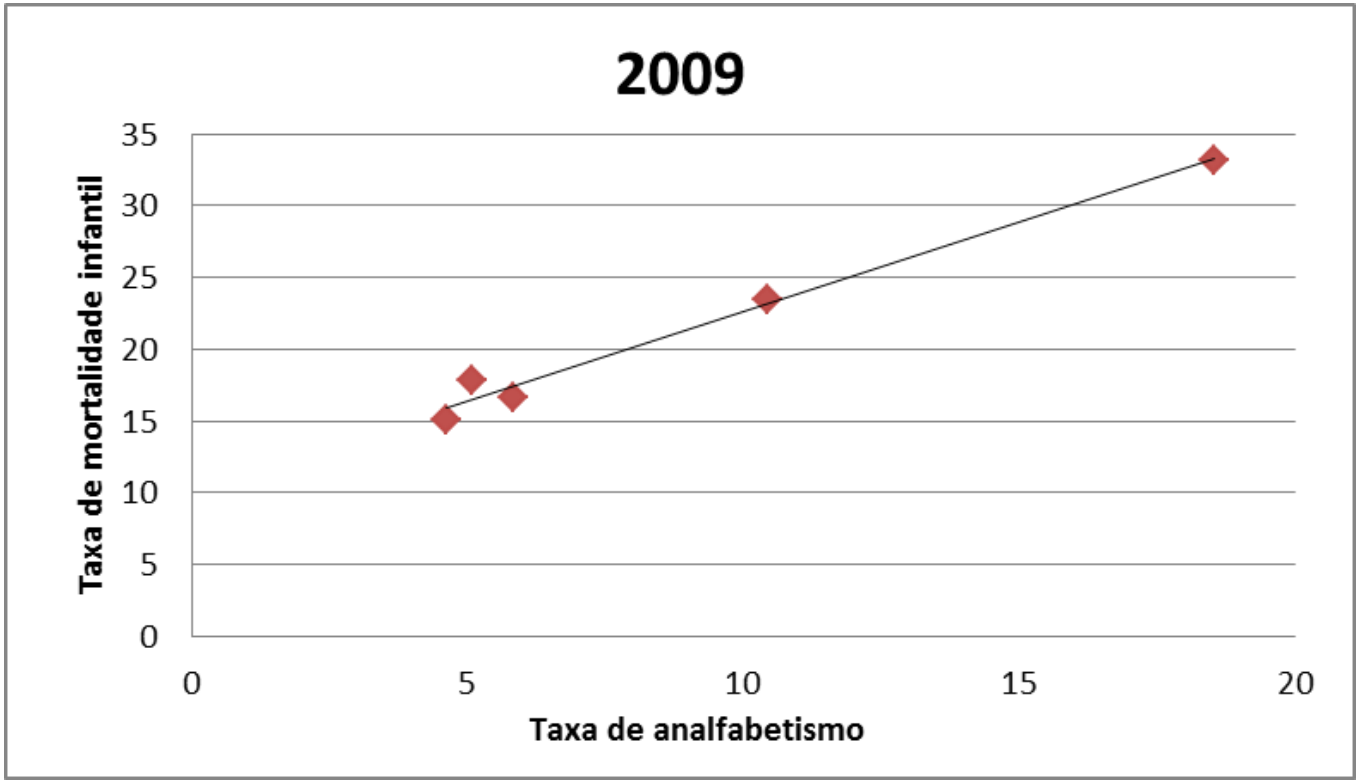
x: variável explicativa

$\hat{y}$: reta ajustada

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

Como obter a e b?

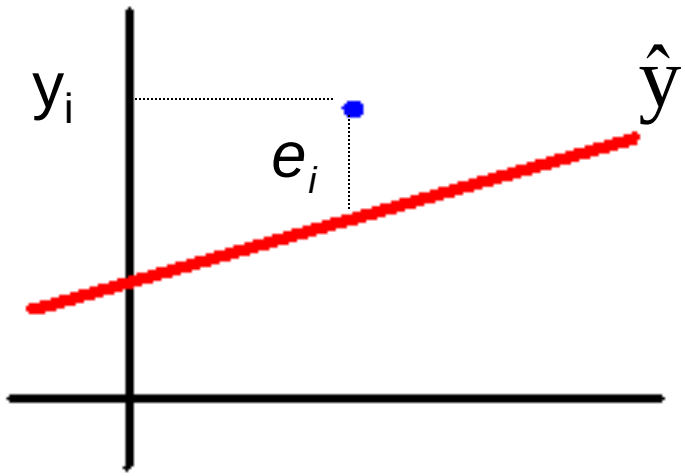
Diagrama de Dispersão



Como estimar essa reta?

Regressão Linear Simples

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



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

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:

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

Como encontrar a e b?

Método dos Mínimos Quadrados

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

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

$$b = \overline{y} - a\,\overline{x}$$

Método dos Mínimos Quadrados

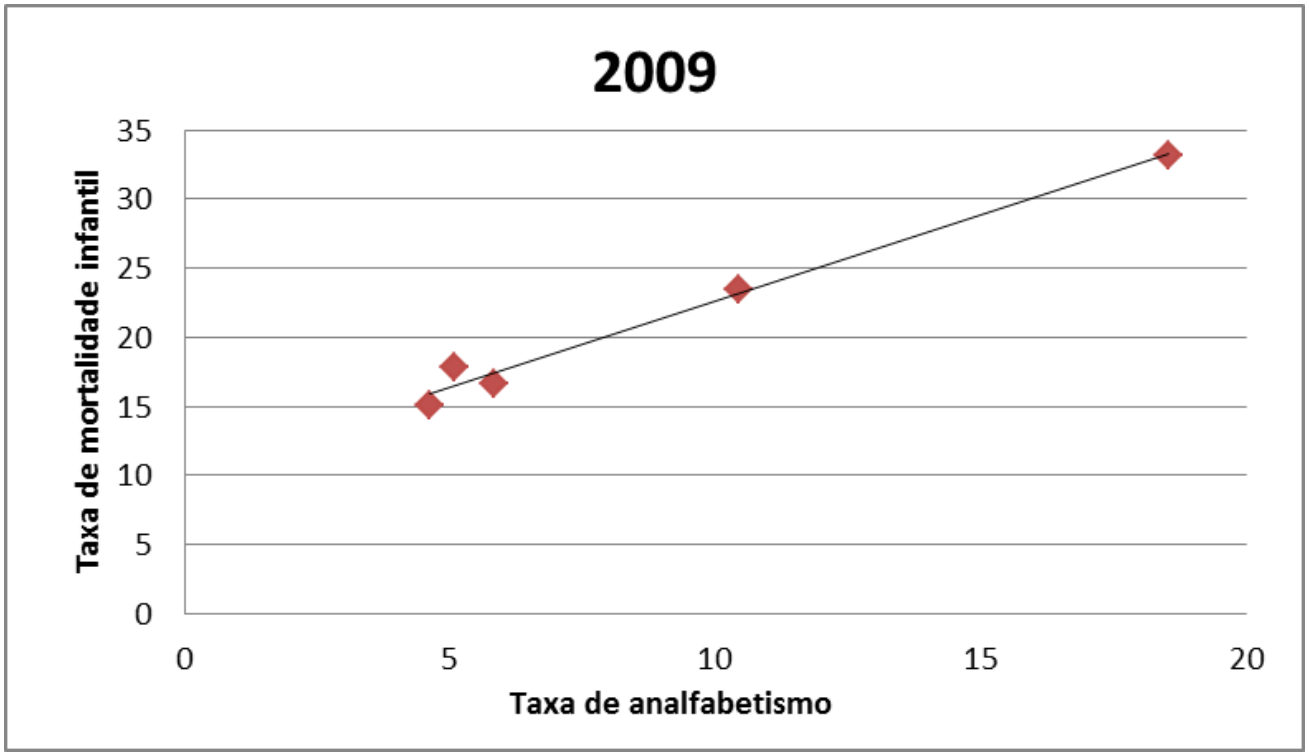
A melhor reta que descreve a relação linear entre **y** e **x** ocorre quando **a** e **b** assumem as seguintes expressões:

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

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

sendo DP(Y) e DP(X) os desvios padrões de Y e X, respectivamente; Cov(X,Y) e 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:

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

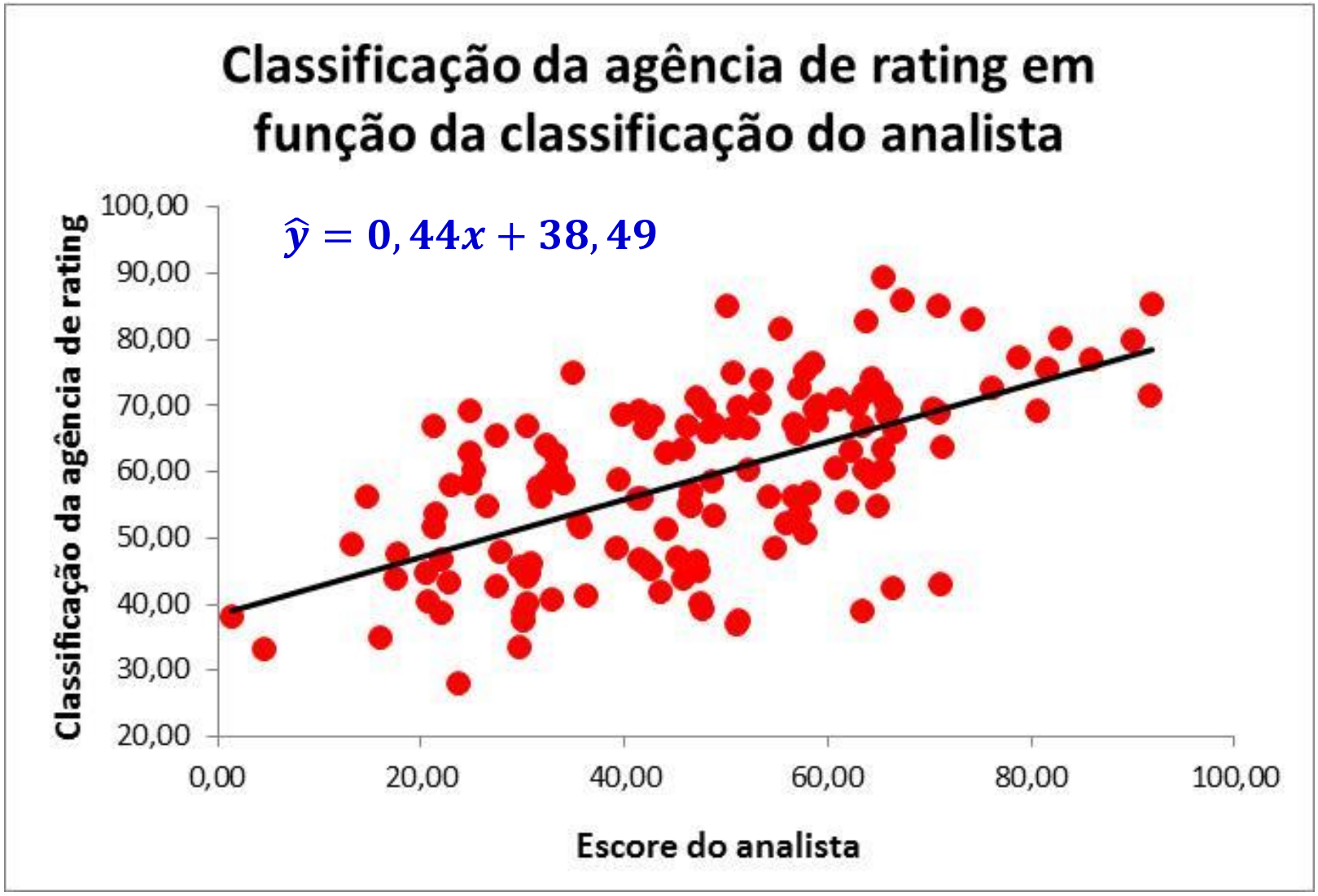
Interpretação dos coeficientes

- 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.

Exemplo: Rating



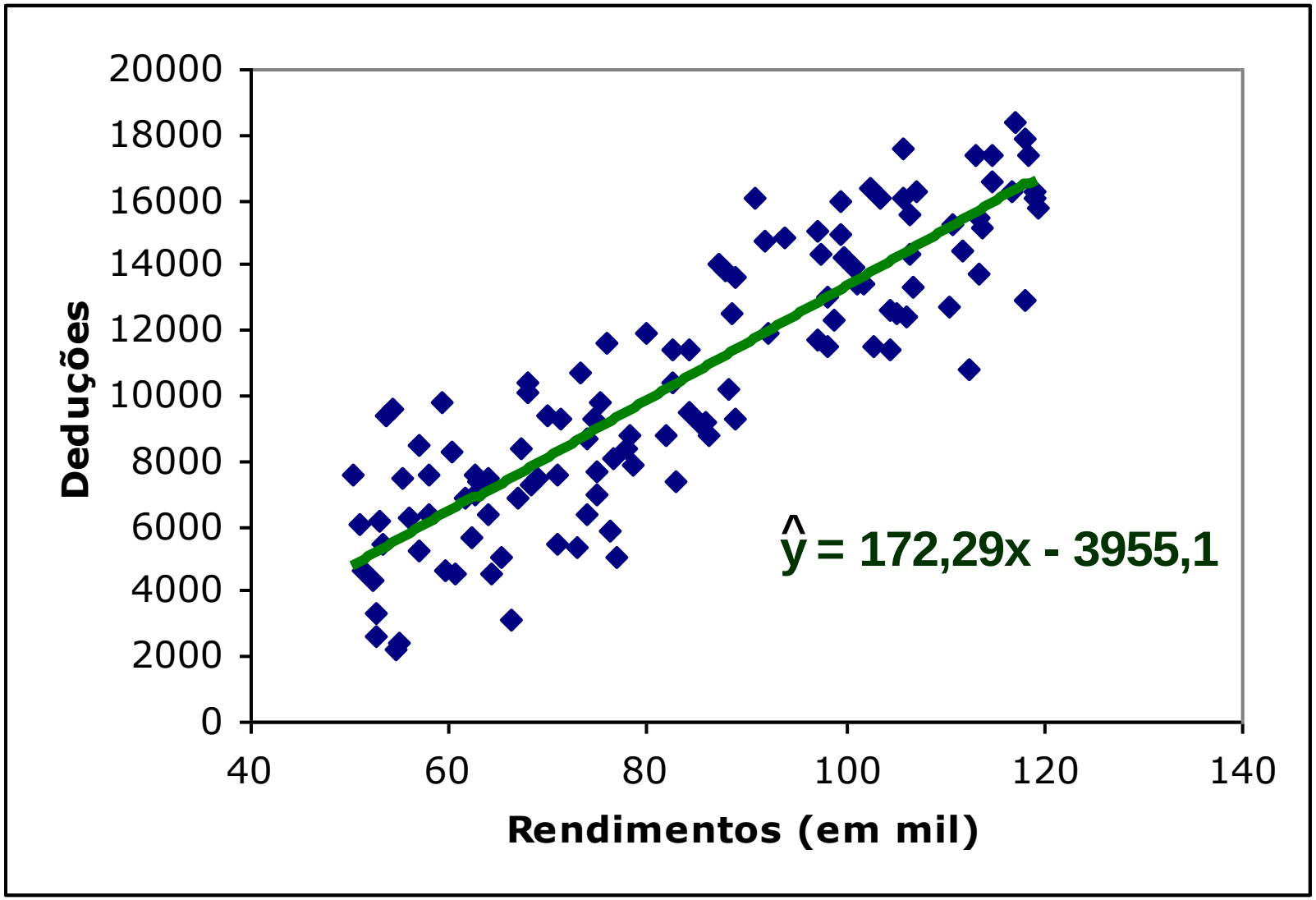
Interpretação dos coeficientes

- 0,44: variação média (esperada) no rating fornecido pela agência quando o escore do analista aumenta 1 unidade.

- 38,49: rating médio (esperado) calculado pela agência de rating de um investimento que recebe escore zero do analista.

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

Exemplo: IR


20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Deduções
40
60
80
100
120
140
Rendimentos (em mil)
ŷ = 172,29x - 3955,1

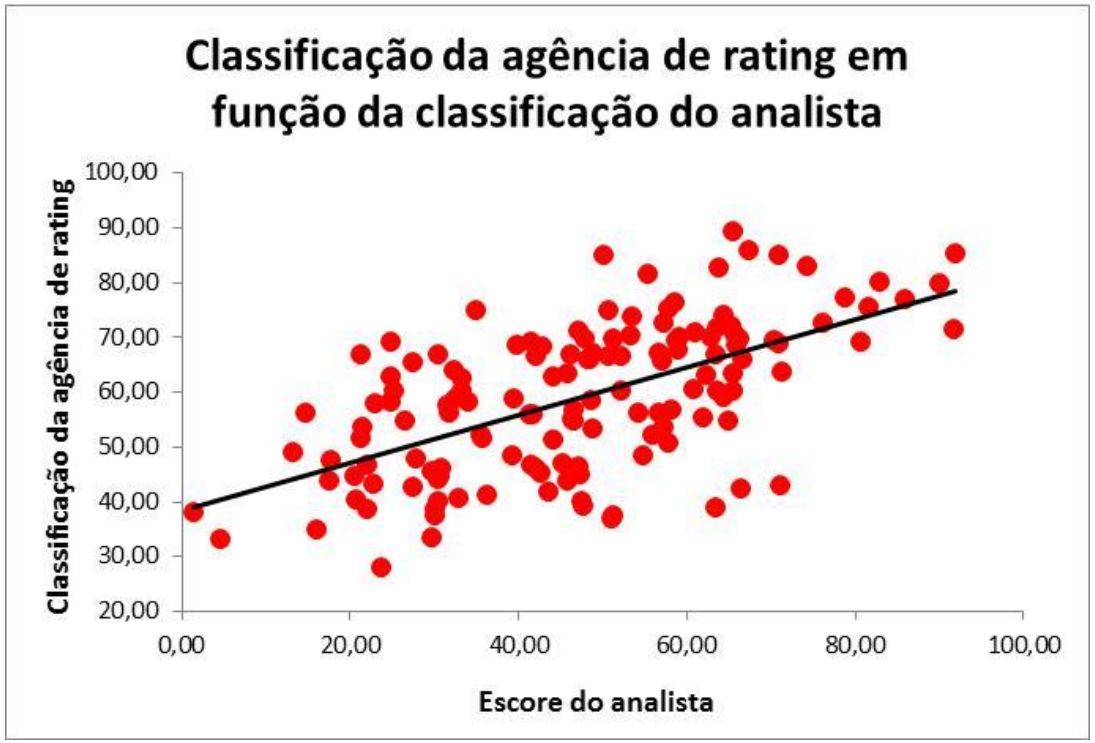
Interpretação dos coeficientes

- 172,29: acréscimo médio (esperado) na dedução do IR a cada aumento de mil reais (uma unidade) na renda do contribuinte.

- -3955,1: **não tem interpretação** na linguagem do problema.

Exemplo: Rating

$$\hat{y} = 0,44x + 38,49$$



O analista atribuiu escore 80 a um investimento. Qual seria o valor esperado (médio) do rating oferecido pela agência de classificação de risco?

Previsão de valores

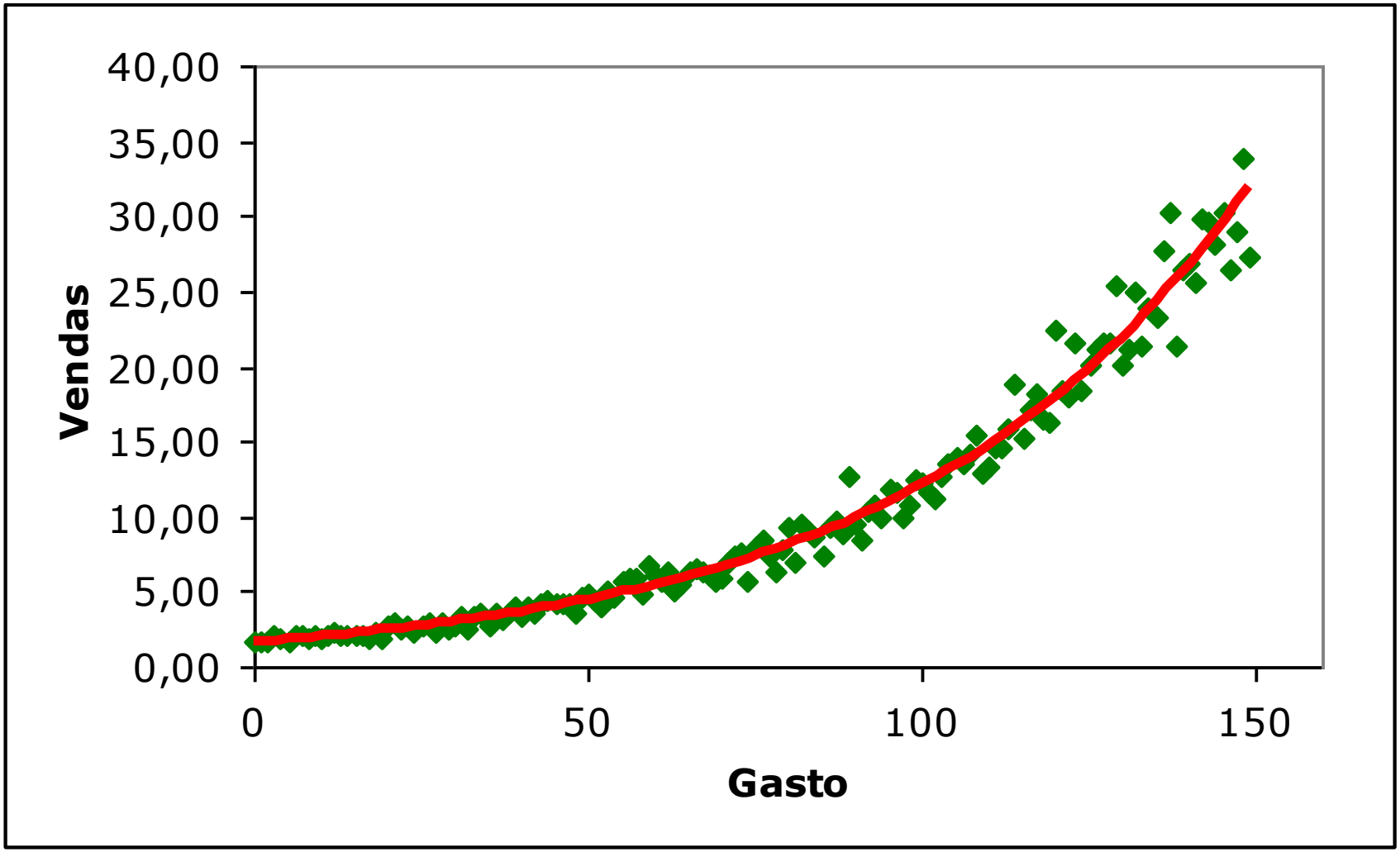
Previsor natural: $\hat{y}_p = a\, x_p + b$

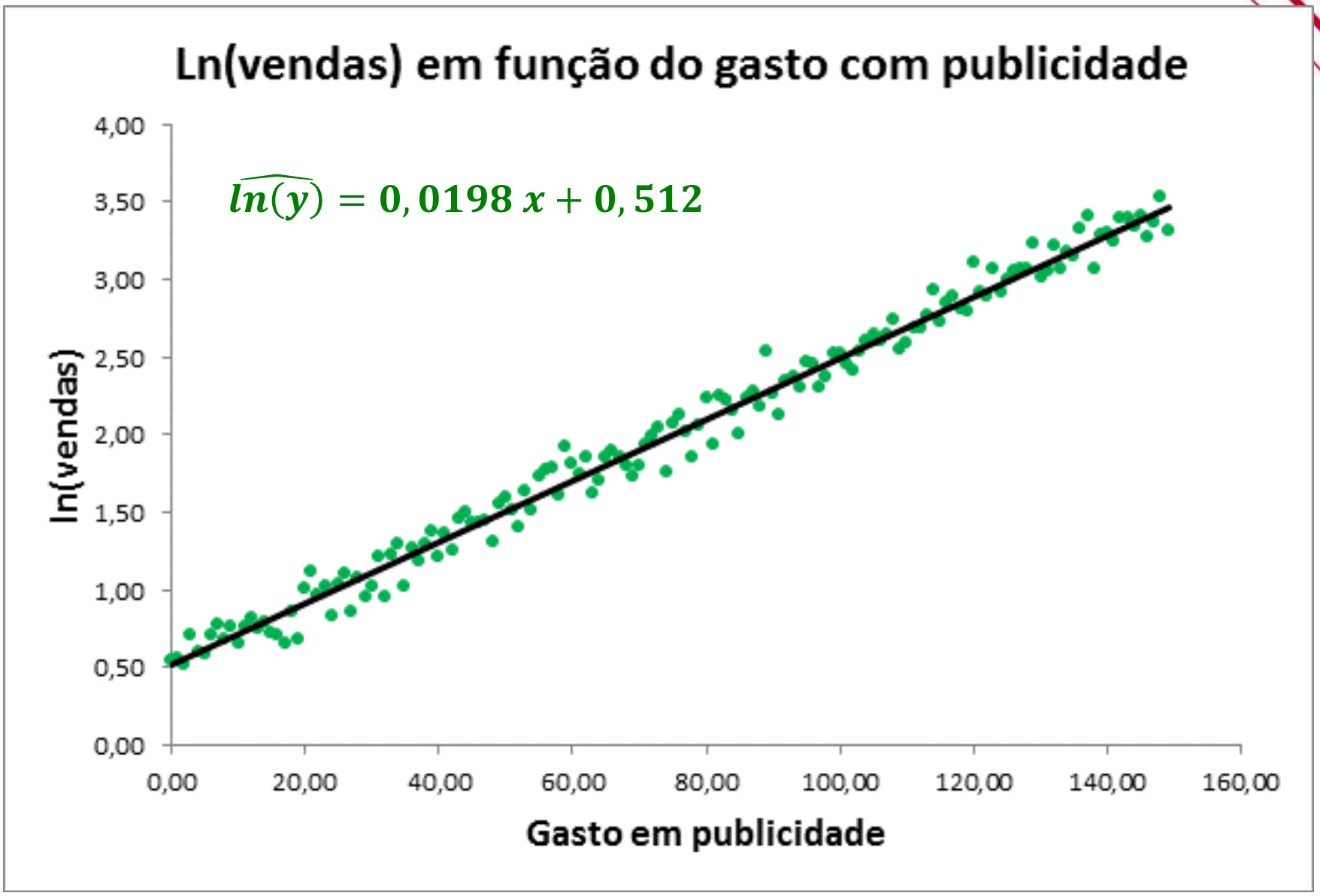
No exemplo:

$$\hat{y}_p = 0{,}44 * 80 + 38{,}49 = 73{,}69$$

Desempenho previsto

EXTRA: Como estimar a curva abaixo?




Ln(vendas) em função do gasto com publicidade
$\widehat{ln(y)} = 0,0198\,x + 0,512$
ln(vendas)
4,00
3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
0,00
20,00
40,00
60,00
80,00
100,00
120,00
140,00
160,00
Gasto em publicidade

O que significa *regressão linear?*

(1) $y_i = a\, x_i + b + \varepsilon_i$

(2) $y_i = a \log(x_i) + b + \varepsilon_i$

(3) $y_i = \exp(a^{Xi} + b) + \varepsilon_i$

(4) $y_i = \exp(a^{Xi} + b)\, \varepsilon_i$

Modelo (1) é de regressão linear

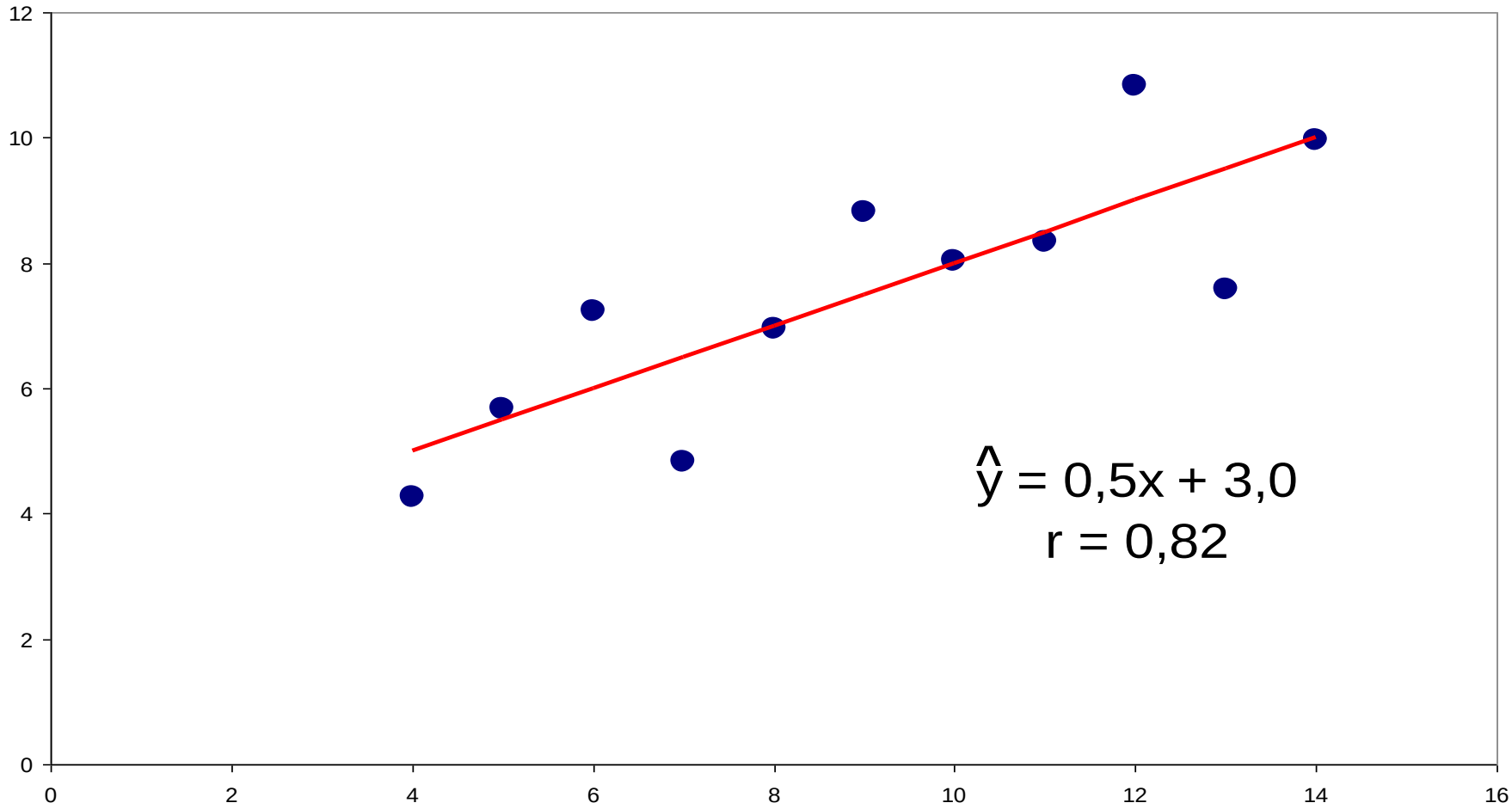
Modelo (2): defina $z_i = \ln(x_i)$

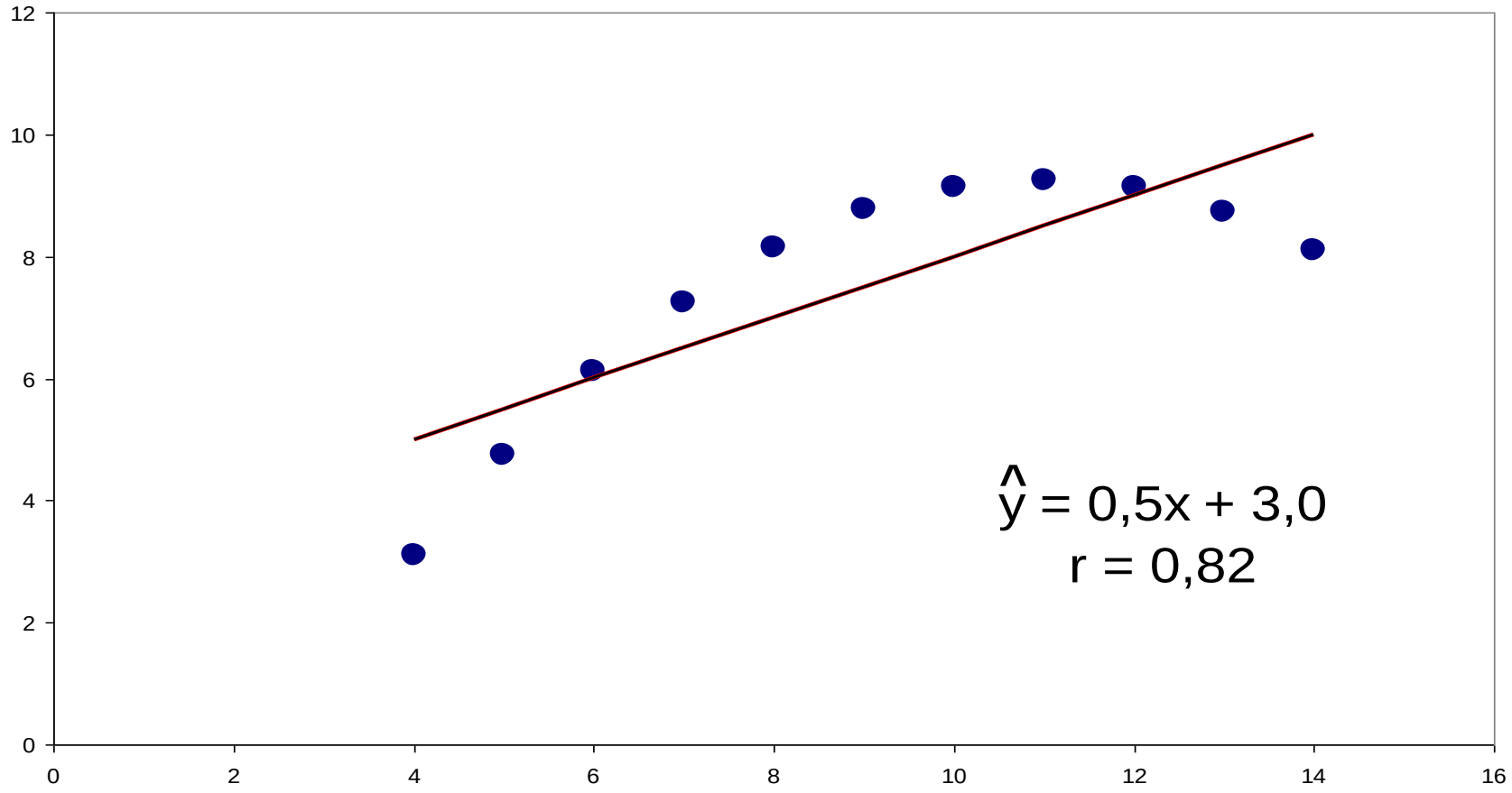
$y_i = a\, z_i + b + \varepsilon_i$, portanto é de regressão linear

Modelos (3) e (4) não são de regressão linear

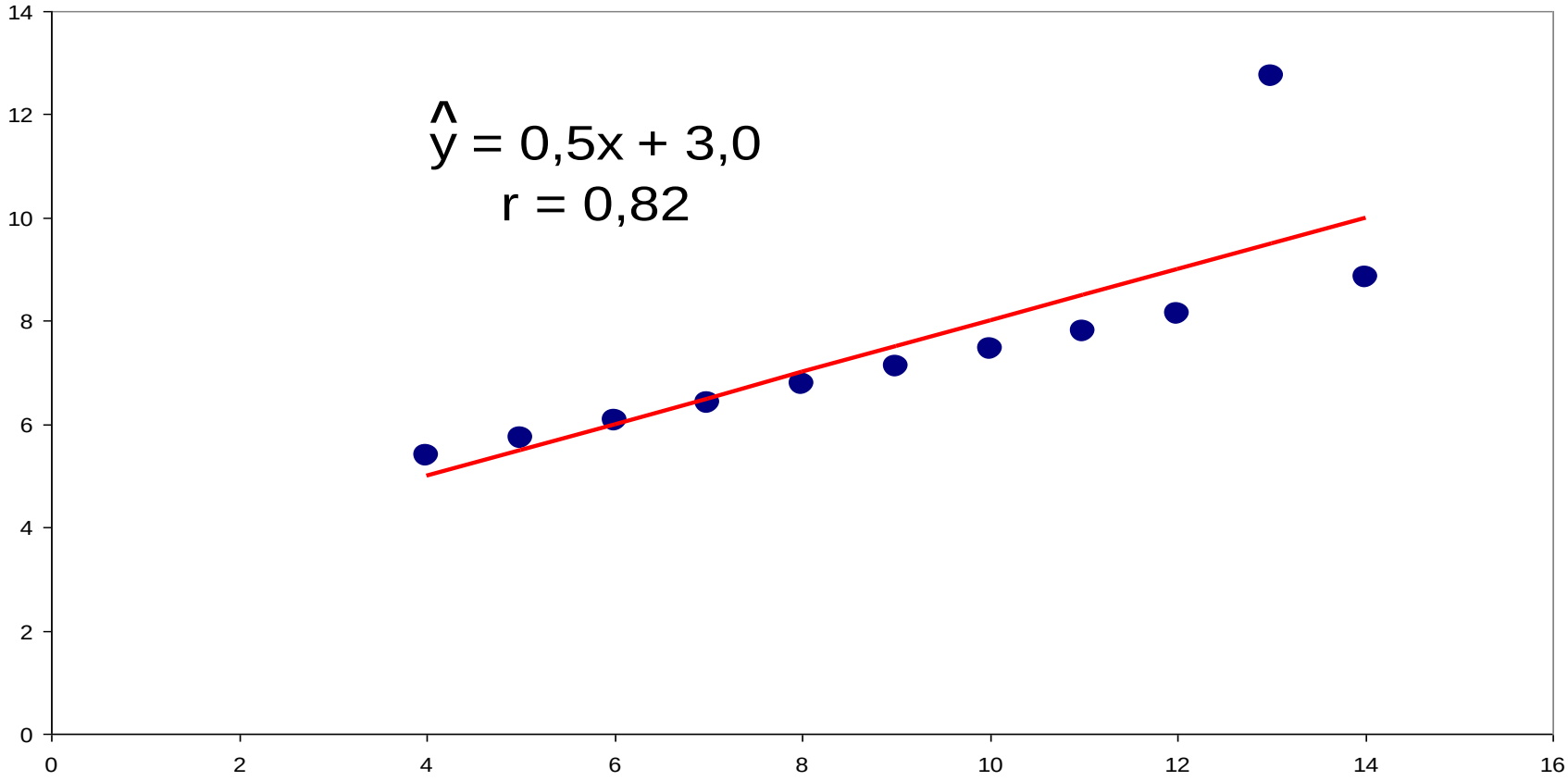
Limitações no uso do coeficiente de correlação (r)

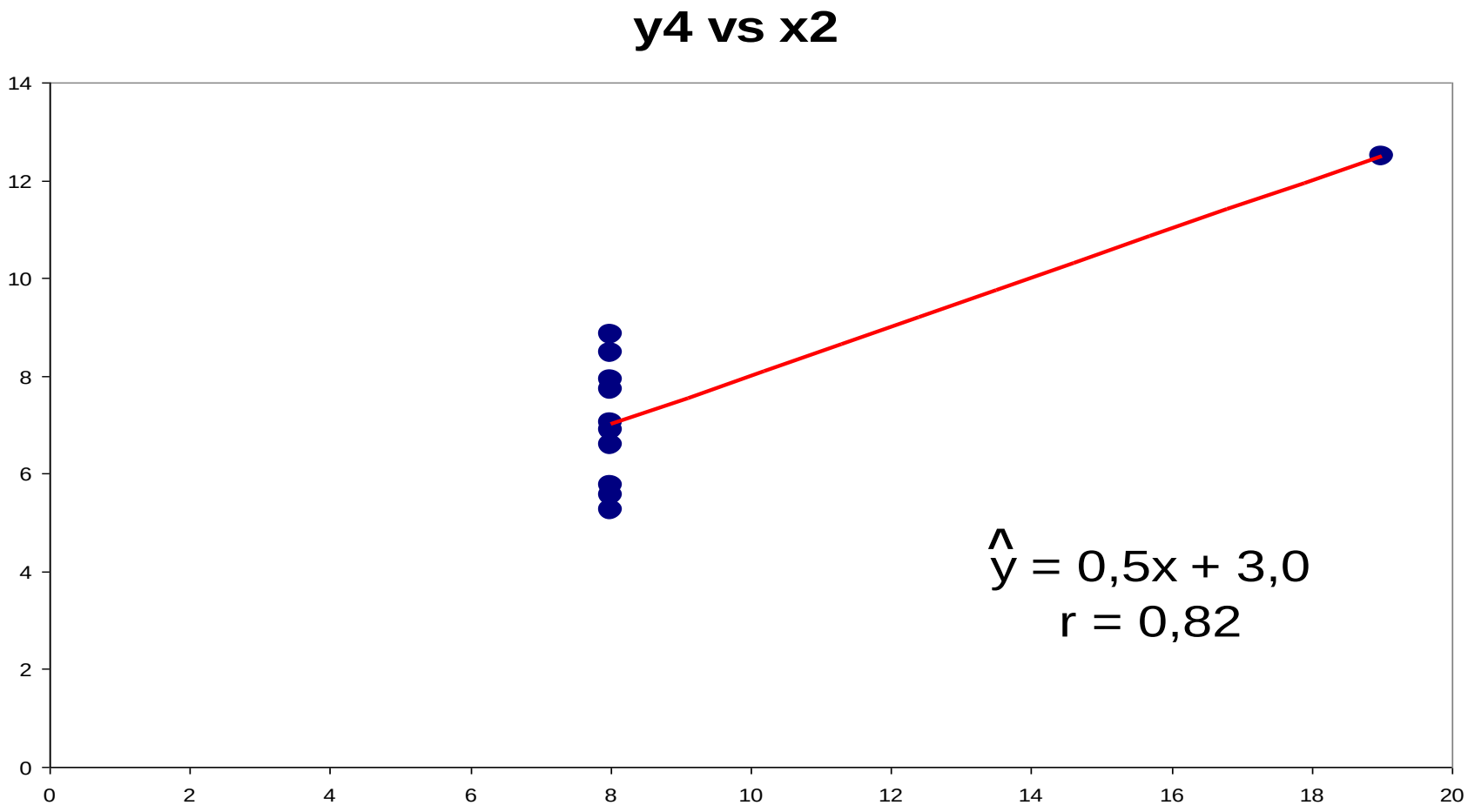
y1 vs x1




y2 vs x1
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
ŷ = 0,5x + 3,0
r = 0,82

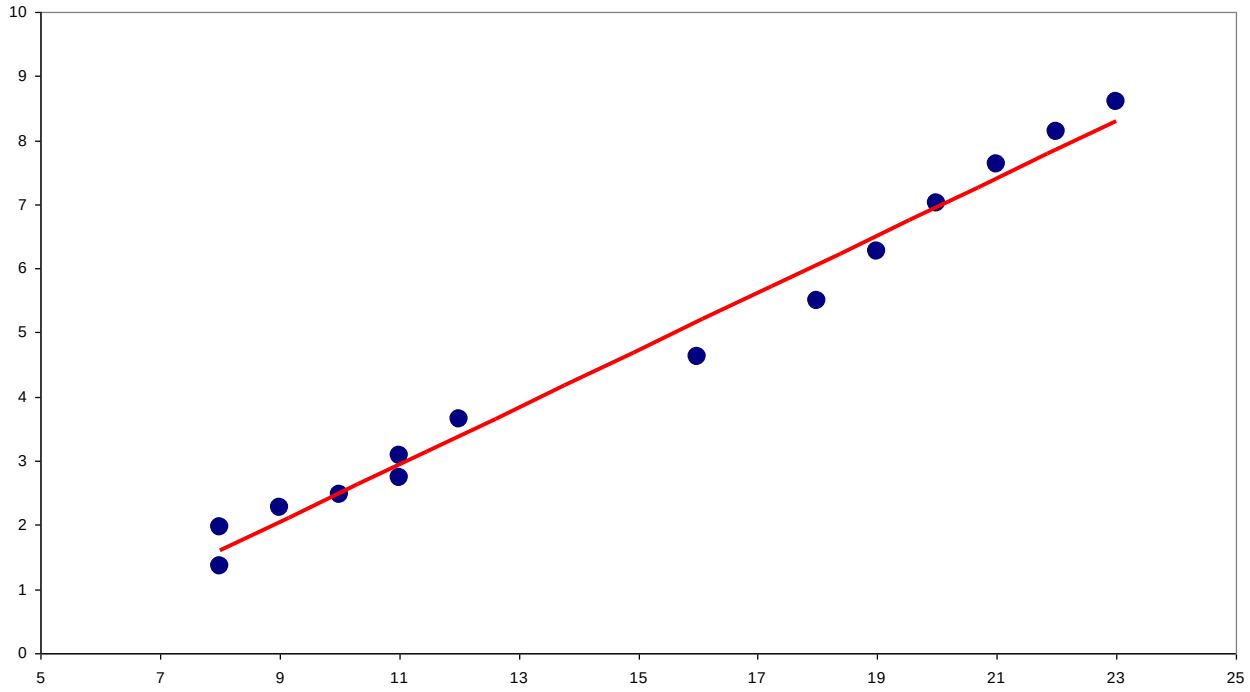
y3 vs x1




y4 vs x2
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18
20
ŷ = 0,5x + 3,0
r = 0,82

Mau uso da regressão

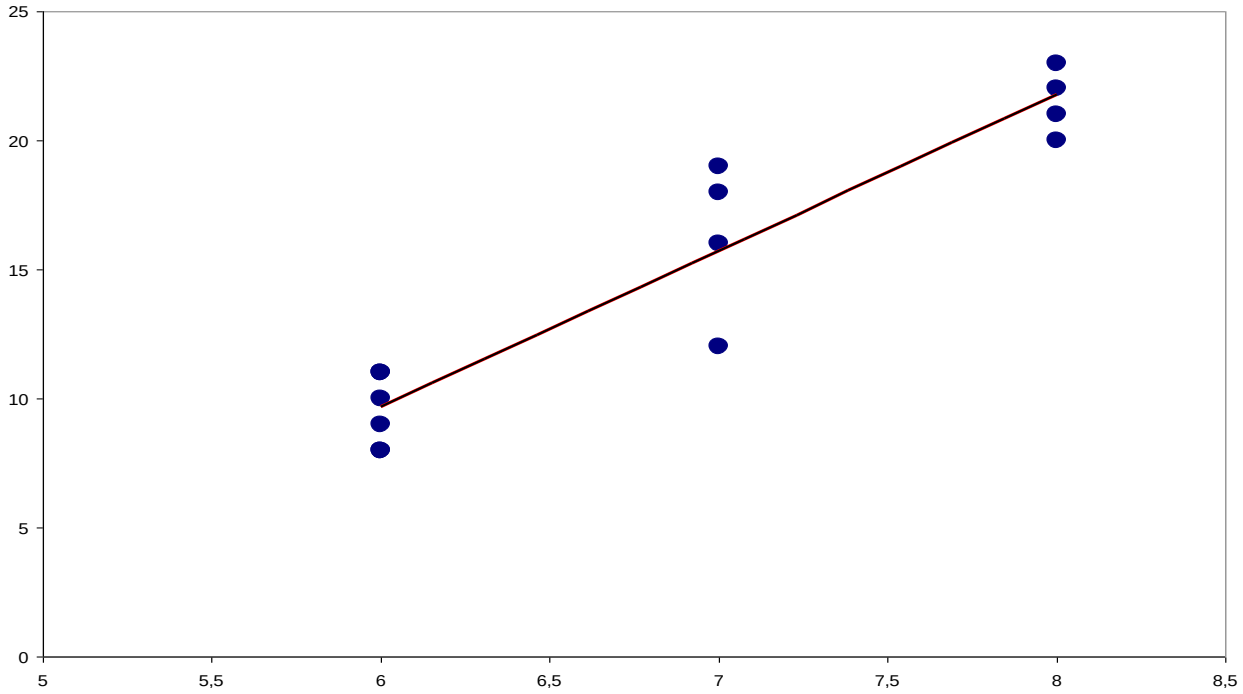
Regressão de y vs x_1



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

$$r = 0{,}99$$

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 Discriminação Salarial...

?? 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:

`Aula05 Atividade Variáveis Quantitativas Discriminacao.ipynb`

Preparo para próxima aula

Os alunos devem se preparar com:

1. Leitura de tudo!!
2. **Apresentar ideia do projeto (Grupos que serão analisados)**

Fiquem atentos aos Avisos no Blackboard!