Capítulo 9 | Correlação e regressão

Descrição do capítulo

- *9.1 Correlação*
- *9.2 Regressão linear*
- *9.3 Medidas de regressão e intervalos de predição*
- *9.4 Regressão múltipla*

Seção 9.1

Correlação

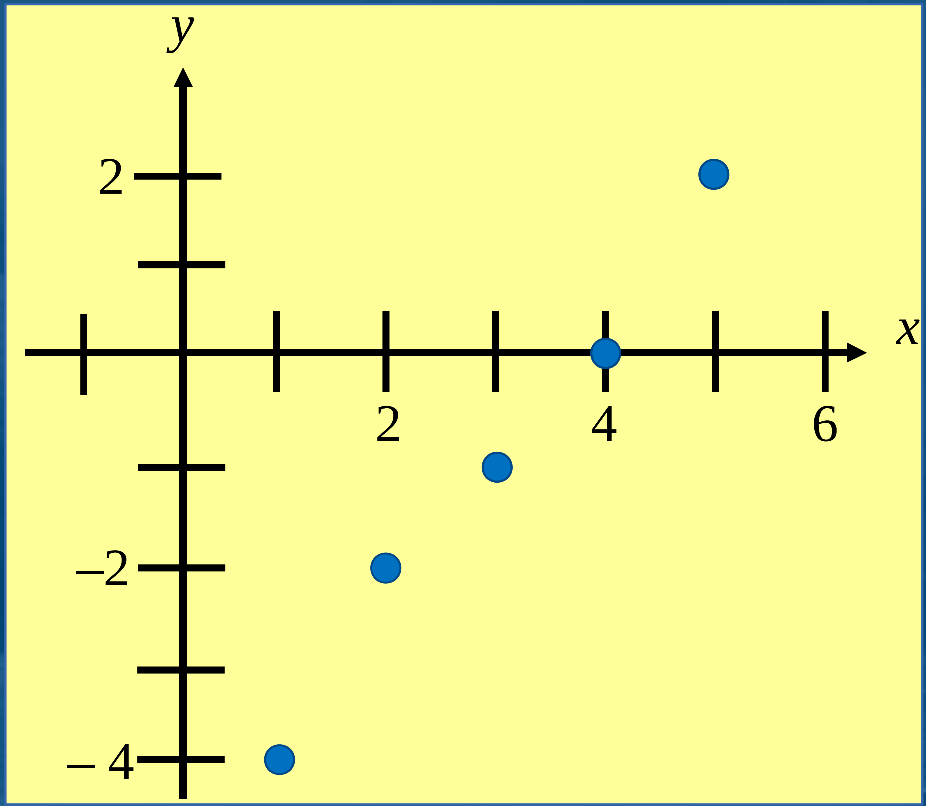
Correlação

- ***Uma relação entre duas variáveis.***
- ***Os dados podem ser representados por pares ordenados (x, y):***
  - *x* é a **variável independente** (ou **explanatória**).
  - *y* é a **variável dependente** (ou **resposta**).

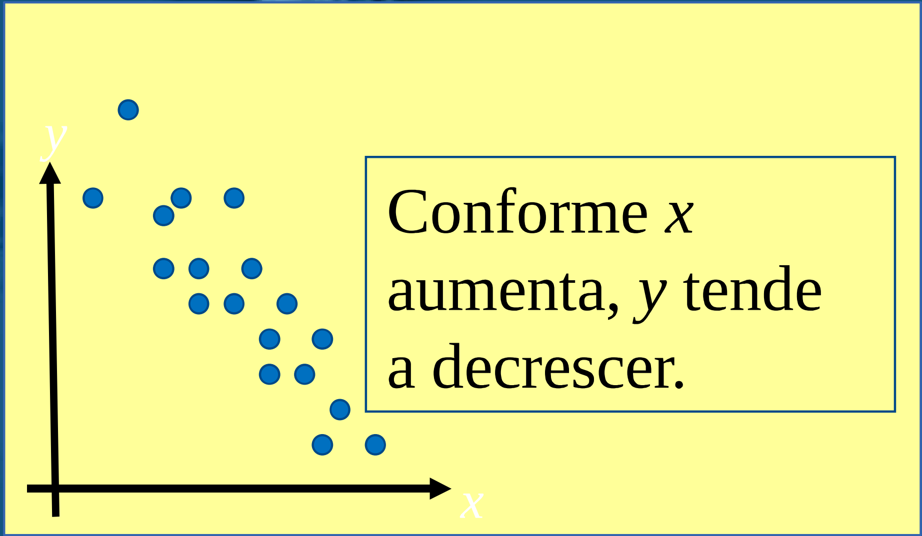
Um **diagrama de dispersão** pode ser usado para determinar se uma correlação linear (linha reta) existe entre duas variáveis.

Exemplo:

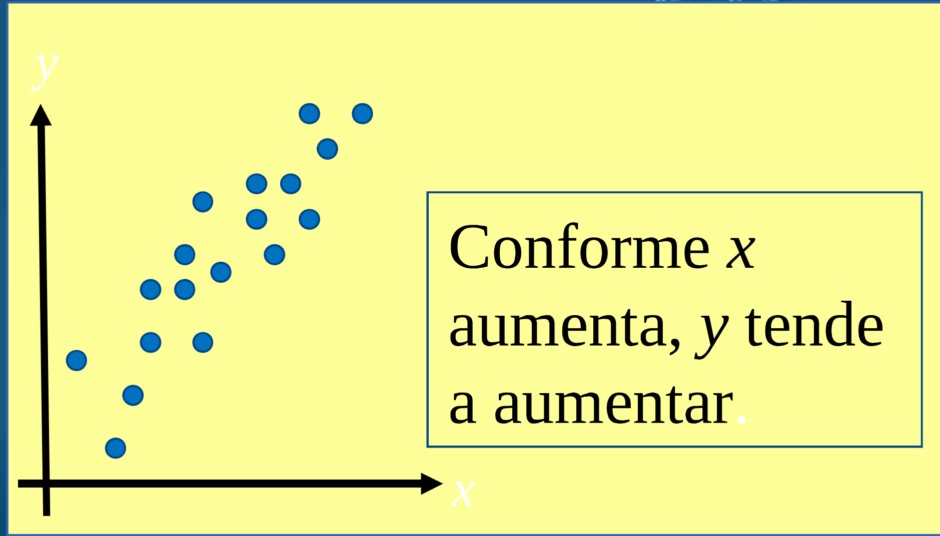
x	1	2	3	4	5
y	– 4	– 2	– 1	0	2



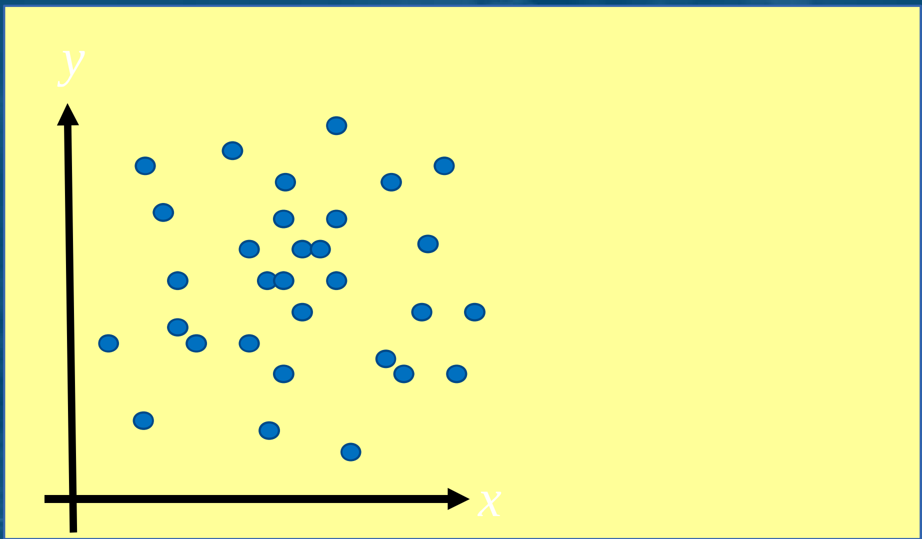
Tipos de correlação



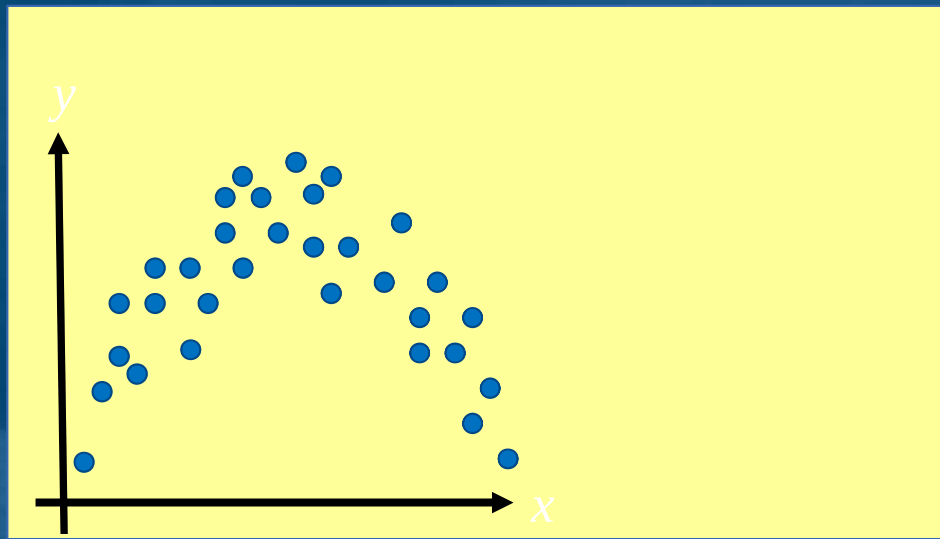
Correlação linear negativa



Correlação linear positiva



Sem correlação



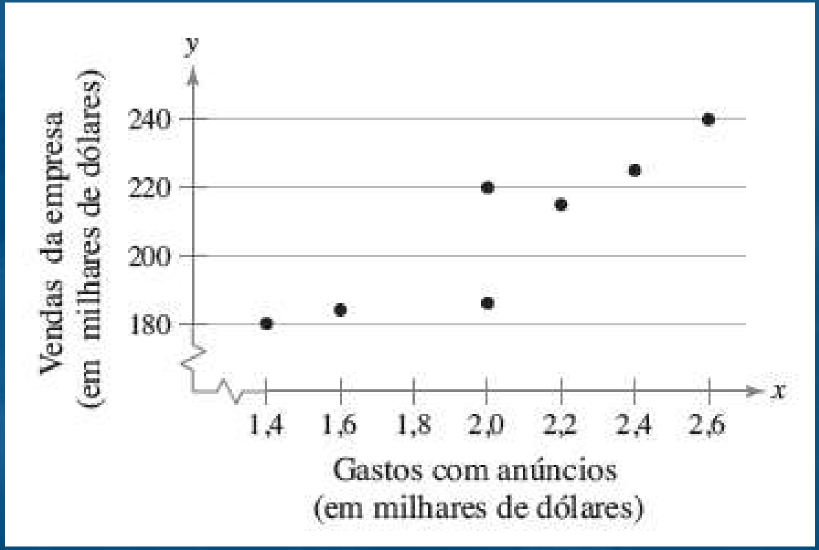
Correlação não linear

Exemplo: construindo um diagrama de dispersão

Um gerente de marketing conduziu um estudo para determinar se há uma relação entre o dinheiro gasto com propaganda e as vendas da empresa. Os dados são mostrados na tabela ao lado. Coloque os dados em um diagrama de dispersão e determine se parece haver uma correlação linear positiva e negativa ou se parece não haver correlação linear.

Gastos com propaganda, ($1000), *x*	Vendas da empresa ($1000), *y*
2,4	225
1,6	184
2,0	220
2,6	240
1,4	180
1,6	184

Solução: construindo um diagrama de dispersão



Parece haver uma **correlação linear positiva**. Conforme os gastos com propaganda aumentam, as vendas tendem a aumentar.

Coeficiente de correlação

- *Uma medida da força e direção de uma relação linear entre duas variáveis.*
- *O coeficiente de correlação é a razão da covariância entre as duas variáveis pelo produto de seus desvios padrão.*
- *Covariância: E[(x-E(x))(y-E(y))]*

Coeficiente de correlação

-
- *O símbolo r representa o coeficiente de correlação amostral.*
- *Uma fórmula para r é:*

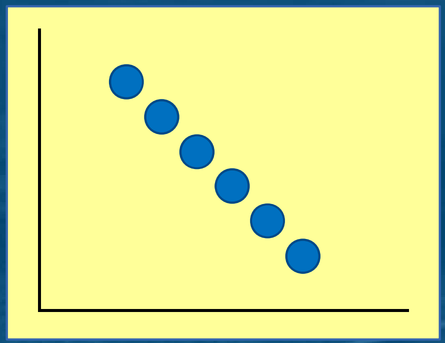
$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

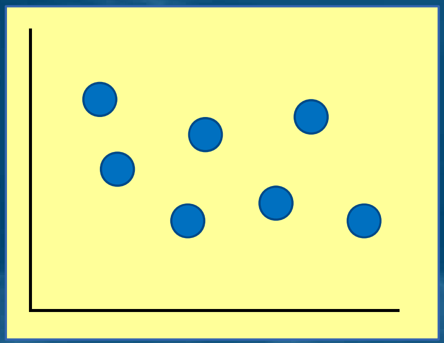
n é o número de dados emparelhados

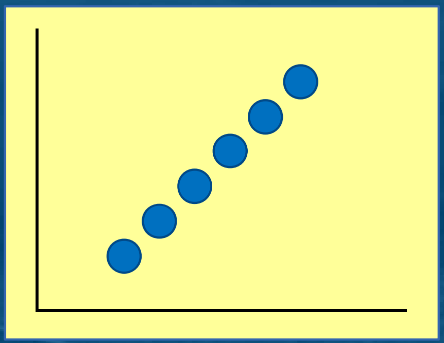
- *O coeficiente de correlação populacional é representado por* ***ρ*** *(rô).*

- ***A amplitude do coeficiente de correlação é -1 para 1.***

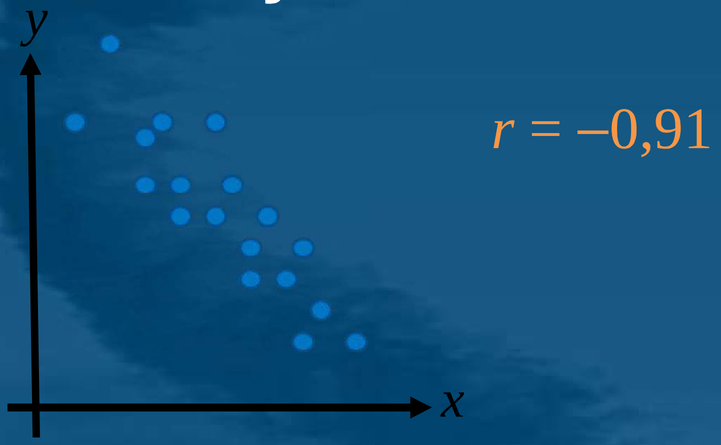
Se *r* = -1 existe uma correlação negativa perfeita.	Se *r* está próximo de 0 não existe correlação linear.	Se *r* = 1 Existe uma correlação positiva perfeita.



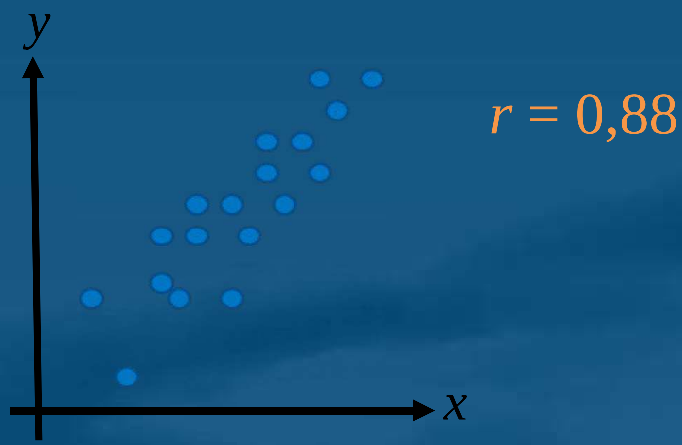




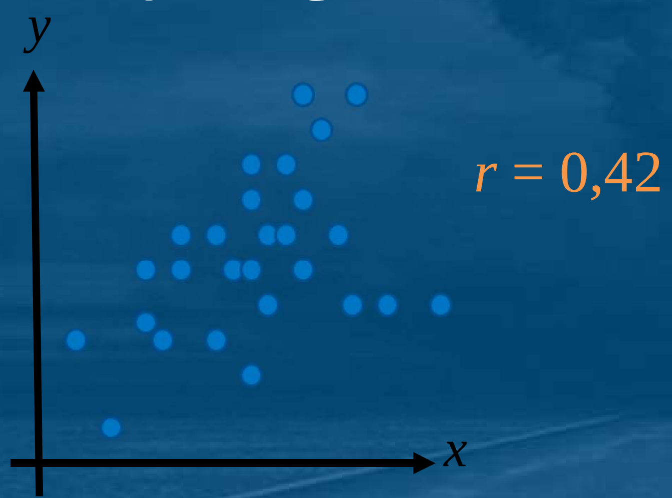
Correlação linear



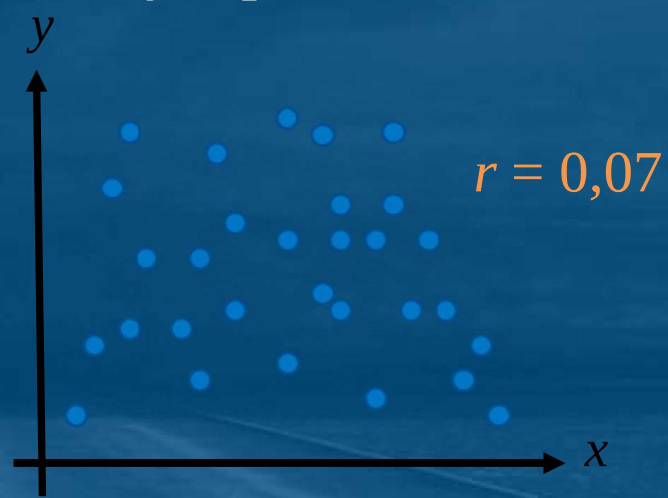
Correlação negativa forte



Correlação positiva forte



Correlação positiva fraca



Correlação não linear

Calculando um coeficiente de correlação

Em palavras	*Em símbolos*
1. Encontre a soma dos valores *x*.	$\sum x$
2. Encontre a soma dos valores *y*.	$\sum y$
3. Multiplique cada valor *x* por seu valor *y* correspondente e encontre a soma.	$\sum xy$

Em palavras	*Em símbolos*
4. Faça o quadrado de cada valor *x* e encontre a soma.	$\sum x^2$
5. Faça o quadrado de cada valor *y* e encontre a soma.	$\sum y^2$
6. Use as cinco somas para calcular o coeficiente de correlação.	$r = \dfrac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$

Exemplo: encontrando o coeficiente de correlação

Calcule o coeficiente de correlação para os dados dos gastos com propaganda e vendas da empresa informados no Exemplo 1. O que podemos concluir?

Gastos com propaganda, ($1000), *x*	Vendas da empresa ($1000), *y*
2,4	225
1,6	184
2,0	220
2,6	240
1,4	180
1,6	184

Solução: encontrando o coeficiente de correlação

$\Sigma x = 15.8 \quad \Sigma y = 1634$

$\Sigma xy = 3289.8$

$\Sigma x^2 = 32.44 \quad \Sigma y^2 = 337{,}558$

$\Sigma x = 15.8 \qquad \Sigma y = 1634 \qquad \Sigma xy = 3289.8 \qquad \Sigma x^2 = 32.44 \qquad \Sigma y^2 = 337{,}558$

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

$$= \frac{8(3289.8) - (15.8)(1634)}{\sqrt{8(32.44) - 15.8^2}\sqrt{8(337{,}558) - 1634^2}}$$

$$= \frac{501.2}{\sqrt{9.88}\sqrt{30{,}508}} \approx 0.9129$$

r ≈ 0.913 sugere uma correlação linear positiva forte. Conforme aumenta o gasto com propaganda, as vendas da empresa também aumentam.

Correlação e causalidade

- ***O fato de duas variáveis serem fortemente correlacionadas não implica uma relação de causa e efeito entre elas.***
- ***Se há uma correlação significante entre duas variáveis, você deve considerar as seguintes possibilidades:***
  1. Existe uma relação direta de causa e efeito entre as variáveis?
     - x causa y?

2. Existe uma relação de causa e efeito reversa entre as variáveis?
   - *y* causa *x*?

3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável ou por uma combinação de várias outras variáveis?

4. É possível que a relação entre as duas variáveis possa ser uma coincidência?

Seção 9.2

Regressão linear

Linhas de regressão

- Após verificar se a correlação linear entre duas variáveis é significante, o próximo passo é determinar a equação da linha que mais bem modela os dados (linha de regressão).
- Pode ser usada para prever o valor de y para um dado valor de x.

Resíduos

- A diferença entre o valor y observado e o valor y previsto para um dado valor x na linha.

Para um dado valor *x*,

d_i = (valor *y* observado) – (valor *y* previsto)

Linha de regressão (linha de melhor ajuste)

- ***A linha para a qual a soma dos quadrados dos resíduos é um mínimo.***
- ***A equação de uma linha de regressão para uma variável independente x e uma variável dependente y é:***

$$\hat{y}_i = mx_i + b$$

valor y previsto para um dado valor x

inclinação

interseção y

Linha de regressão (linha de melhor ajuste)

- ***Soma dos quadrados dos resíduos:***

$c = \Sigma d_i^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - mx_i - b)^2$

- *Para minimizar (segundo Cálculo II):*

$\partial c/\partial m = 2\Sigma(y_i - mx_i - b)\Sigma(-x_i) = 0$

$\partial c/\partial b = 2\Sigma(y_i - mx_i - b)\Sigma(-1) = 0$

- ***Temos um sistema de equações a resolver:***

$$\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix} \begin{vmatrix} m \\ b \end{vmatrix} = \begin{vmatrix} \sum y_i x_i \\ \sum y_i \end{vmatrix}$$

A equação da linha de regressão

- $\hat{y} = mx + b$ *onde*

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

$$b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

- $\bar{y}$ *é a média dos valores y no conjunto de dados*
- $\bar{x}$ *é a média dos valores x no conjunto de dados*
- *A linha de regressão sempre passa pelo ponto* $\left(\bar{x}, \bar{y}\right)$

Exemplo: encontrando a equação da reta de regressão

Encontre a equação da reta de regressão para os gastos com propaganda e dados sobre as vendas da empresa.

Gastos com propaganda, ($1000), *x*	Vendas da empresa ($1000), *y*
2,4	225
1,6	184
2,0	220
2,6	240
1,4	180
1,6	184
2,0	186
2,2	215

Solução: encontrando a equação da linha de regressão

Lembrando da seção 9.1:

$\Sigma x = 15{,}8$ $\Sigma y = 1634$

$\Sigma xy = 3289{,}8$

$\Sigma x^2 = 32{,}44$ $\Sigma y^2 = 337.558$

$\Sigma x = 15{,}8$ $\Sigma y = 1634$ $\Sigma xy = 3289{,}8$ $\Sigma x^2 = 32{,}44$ $\Sigma y^2 = 337.558$

$$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{8(3289.8) - (15.8)(1634)}{8(32.44) - 15.8^2}$$

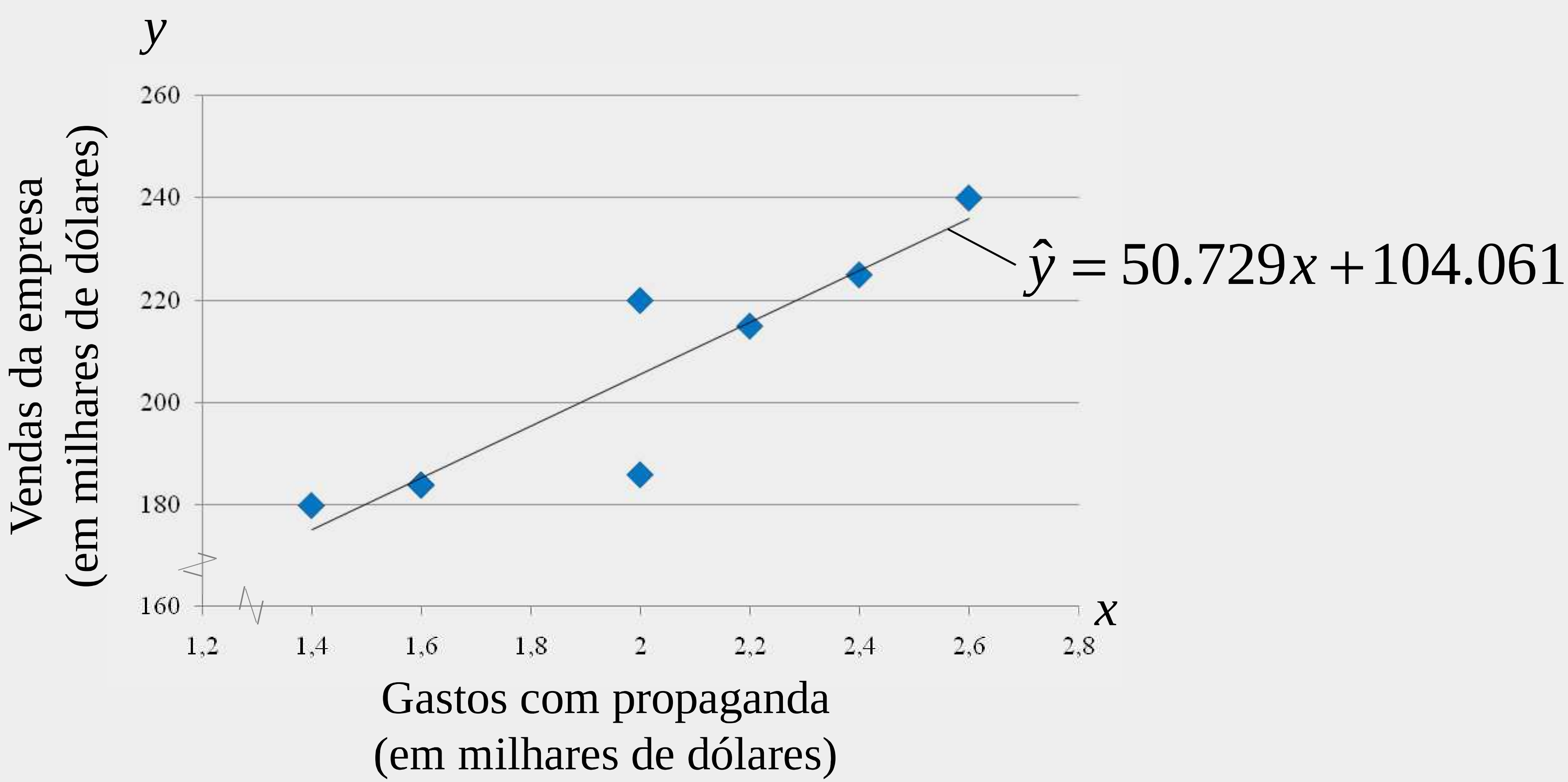
$$= \frac{501.2}{9.88} \approx 50.72874$$

$$b = \overline{y} - m\overline{x} = \frac{1634}{8} - (50.72874)\frac{15.8}{8}$$

$$= 204.25 - (50.72874)(1.975) \approx 104.0607$$

Equação da linha de regressão $\hat{y} = 50.729x + 104.061$

- ***Para desenhar a linha de regressão, use quaisquer dois valores x dentro da faixa de dados e calcule seus valores y correspondentes a partir da linha de regressão.***



Exemplo: prevendo valores y usando equações de regressão

- A equação de regressão para os dados sobre gastos com propaganda (em milhares de dólares) e vendas da empresa (em milhares de dólares) é: $\hat{y} = 50{,}729x + 104{,}061$. Use essa equação para prever as vendas esperadas da empresa para os seguintes gastos com propaganda. (Reveja o Exemplo 7 da Seção 9,1, no qual x e y têm uma correlação linear significante.)

1. ***1,5 mil dólares***
2. ***1,8 mil dólares***
3. ***2,5 mil dólares***

Solução: prevendo valores y usando equações de regressão

$\hat{y} = 50{,}729x + 104{,}061$

1. 1,5 mil dólares

   $\hat{y} = 50{,}729(1{,}5) + 104{,}061 \approx 180{,}155$

   Quando os gastos com propaganda são de $1500, as vendas da empresa são cerca de $180,155.

2. 1,8 mil dólares

   $\hat{y} = 50{,}729(1{,}8) + 104{,}061 \approx 195{,}373$

   Quando os gastos com propaganda são de $1800, as vendas da empresa são cerca de $195,373.

3. ***2,5 mil dólares***

$\hat{y} = 50{,}729(2{,}5) + 104{,}061 \approx 230{,}884$

Quando os gastos com propaganda são de $2500, as vendas da empresa são cerca de $230,884.

Valores de previsão são significantes somente para valores x na (ou próximos à) faixa dos dados. Os valores x do conjunto original de dados variam de 1,4 a 2,6. Portanto, não seria apropriado usar a linha de regressão y^ = 50,729x + 104,061 para prever as vendas da empresa por gastos com propaganda, tais como 0,5 ($ 500) ou 5,0 ($ 5.000).