

CORRELAÇÃO E REGRESSÃO

Aulas 22 e 23

Objetivos



- ❑ Correlação
- ❑ Regressão linear
- ❑ Medidas de regressão e intervalos de predição
- ❑ Regressão múltipla

Objetivos da Aula

- Introdução à correlação linear, variáveis dependentes e independentes e tipos de correlação.
- Encontrar o coeficiente de correlação.
- Testar o coeficiente de correlação ρ de uma população usando uma tabela.
- Realizar um teste de hipótese para o coeficiente de correlação ρ de uma população.
- Distinguir entre correlação e causalidade.

Correlação

4

- Inspetor de segurança quer determinar se existe uma relação entre o número de horas de treinamento para um funcionário e o número de acidentes com o mesmo funcionário
- Psicóloga quer saber se existe uma relação entre o número de horas que uma pessoa dorme e o tempo de reação da pessoa

Correlação

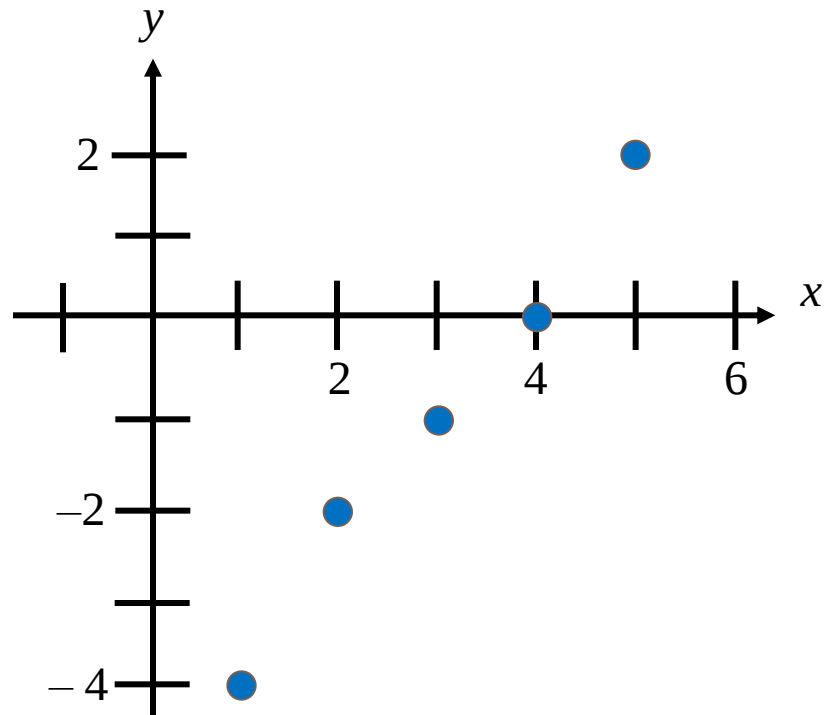
- Relação entre duas variáveis.
- Os dados podem ser representados por pares ordenados (x, y) :
 - ▣ x é a **variável independente** (ou **explanatória**).
 - ▣ y é a **variável dependente** (ou **resposta**).

Correlação

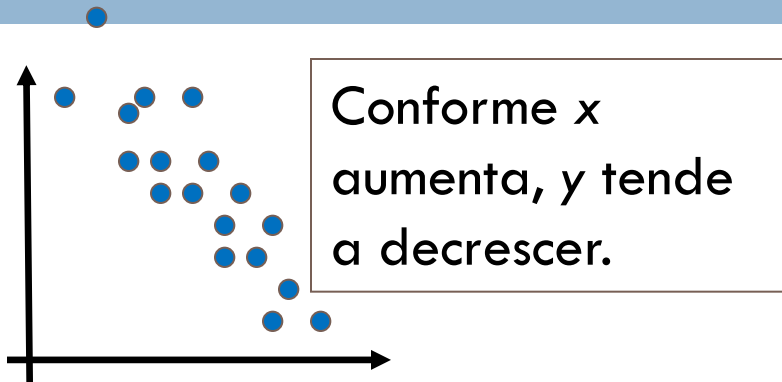
Um **diagrama de dispersão** pode ser usado para determinar se uma correlação linear (linha reta) existe entre duas variáveis.

Exemplo:

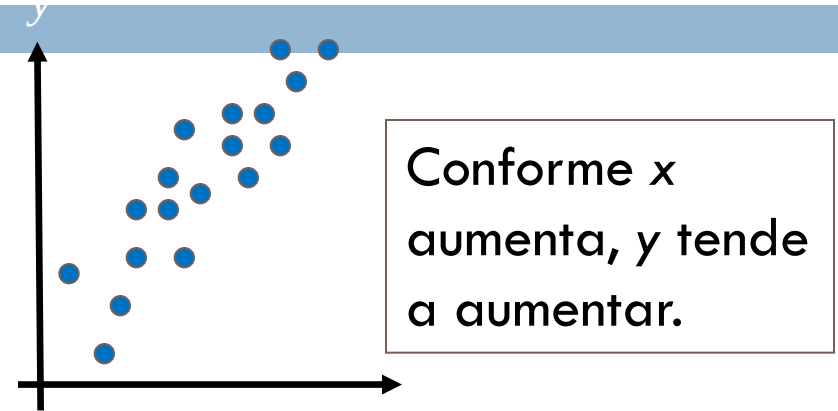
x	1	2	3	4	5
y	-4	-2	-1	0	2



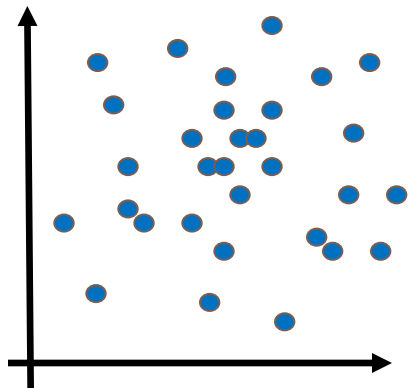
Tipos de correlação



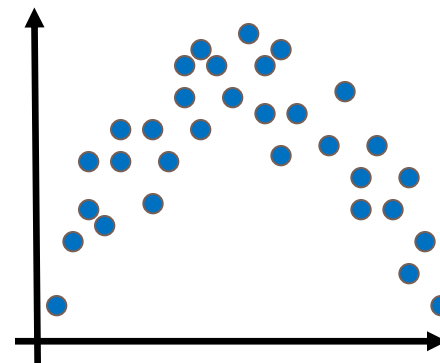
Correlação linear negativa



Correlação linear positiva



Sem correlação



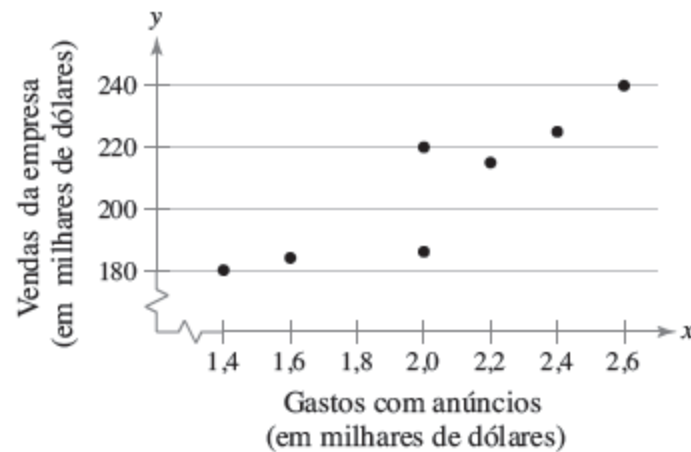
Correlação não linear

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
2,0	186
2,2	215

Exercício: construindo um diagrama de dispersão



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

Exemplo: construindo um diagrama de dispersão usando tecnologia

O Old Faithful, localizado no Parque Nacional Yellowstone, é o gêiser mais famoso do mundo. A duração (em minutos) de diversas erupções do Old Faithful e os tempos (em minutos) até que as próximas erupções aconteçam são mostrados na tabela à direita. Represente os dados em um diagrama de dispersão. Determine o tipo de correlação. (fazer no Excel – para casa)

Duração x	Tempo, y	Duração x	Tempo, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Coeficiente de correlação

- Uma medida da força e direção de uma relação linear entre duas variáveis.
- O símbolo r representa o coeficiente de correlação amostral.
- Uma fórmula para r é:

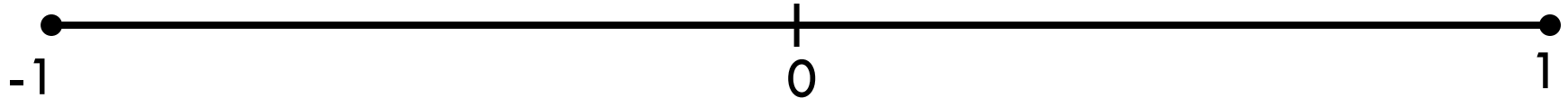
$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

n é o número de dados emparelhados

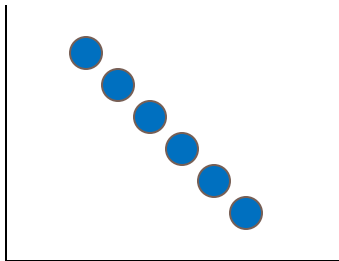
- O coeficiente de correlação populacional é representado por ρ (rô).

Coeficiente de correlação

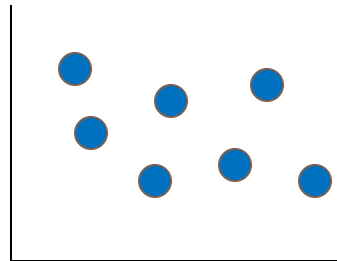
- A amplitude do coeficiente de correlação é -1 para 1.



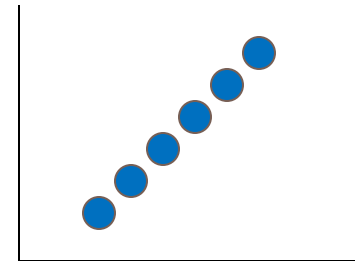
Se $r = -1$ existe uma correlação negativa perfeita.



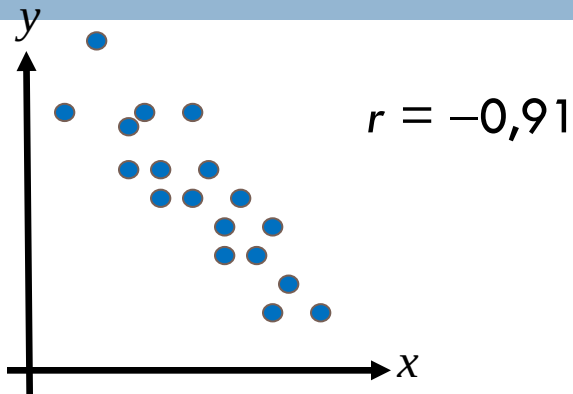
Se r está próximo de 0 não existe correlação linear.



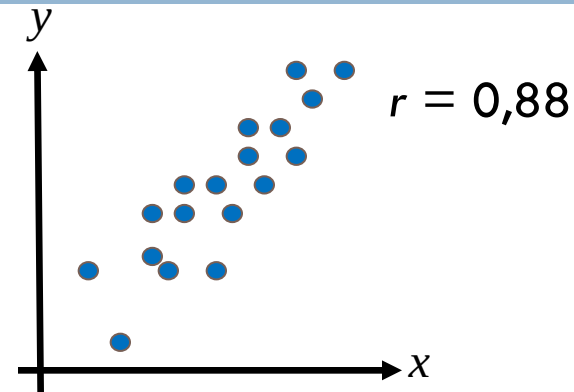
Se $r = 1$ Existe uma correlação positiva perfeita.



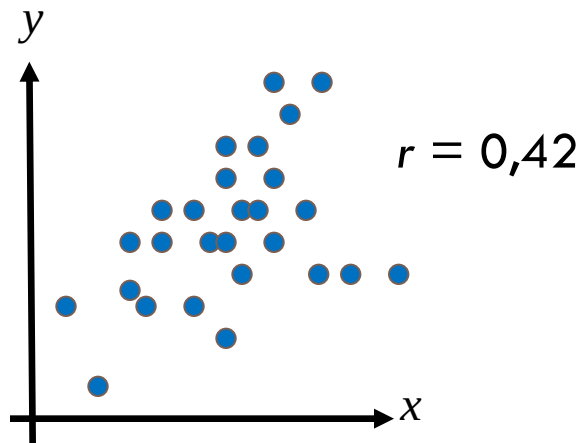
Correlação linear



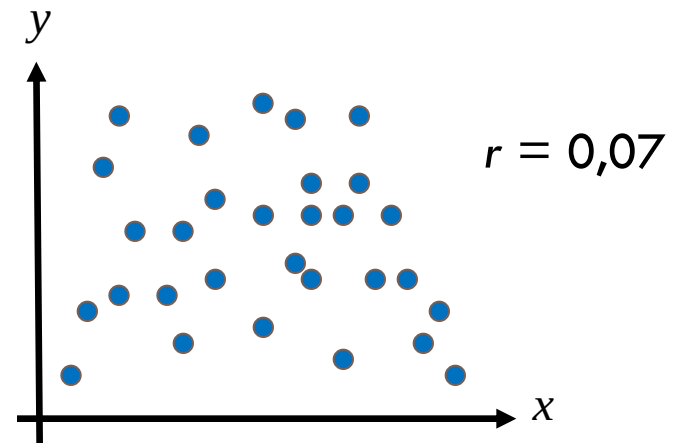
Correlação negativa forte



Correlação positiva forte



Correlação positiva fraca



Correlação não linear

Calculando um coeficiente de correlação

Em palavras

1. Encontre a soma dos valores x .
2. Encontre a soma dos valores y .
3. Multiplique cada valor x pelo y correspondente e encontre a soma.

Em símbolos

$$\sum x$$

$$\sum y$$

$$\sum xy$$

Calculando um coeficiente de correlação

Em palavras

4. Faça o quadrado de cada valor x e encontre a soma.
5. Faça o quadrado de cada valor y e encontre a soma.
6. Use as cinco somas para calcular o coeficiente de correlação.

Em símbolos

$$\sum x^2$$

$$\sum y^2$$

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

Exemplo: encontrando o coeficiente de correlação

Calcule o coeficiente de correlação para os 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
2,0	186
2,2	215

Solução: encontrando o coeficiente de correlação

x	y	xy	x^2	y^2
2.4	225	540	5.76	50,625
1.6	184	294.4	2.56	33,856
2.0	220	440	4	48,400
2.6	240	624	6.76	57,600
1.4	180	252	1.96	32,400
1.6	184	294.4	2.56	33,856
2.0	186	372	4	34,596
2.2	215	473	4.84	46,225
$\Sigma x = 15.8$	$\Sigma y = 1634$	$\Sigma xy = 3289.8$	$\Sigma x^2 = 32.44$	$\Sigma y^2 = 337,558$

Solução: encontrando o coeficiente de correlação

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

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

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

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

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

Exemplo: usando tecnologia para encontrar o coeficiente de correlação

Use a ferramenta tecnológica para calcular o coeficiente de correlação para os dados do Old Faithful. O que podemos concluir? (para casa)

EXCEL			
	A	B	C
26	CORREL(A1:A25,B1:B25)		
27			0.978659

Duração x	Tempo, y	Duração x	Tempo, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Usando uma tabela para testar o coeficiente de correlação populacional ρ

- Uma vez calculado o coeficiente de correlação amostral r , precisamos determinar se há evidência suficiente para decidir se o coeficiente de correlação populacional ρ é significativo em um nível de significância especificado.
- Use a tabela (Correlação de Pearson)
- Se $|r|$ é maior que o valor crítico, há evidência o bastante para decidir se o coeficiente de correlação ρ é significativo.

- Determine se ρ é significativa para cinco dados emparelhados ($n = 5$) em um nível de significância $\alpha = 0.01$.

Número n de pares de dados na amostra

Valores críticos para $\alpha = 0,05$ e $\alpha = 0,01$

n	$\alpha = 0,05$	$\alpha = 0,01$
4	0,950	0,990
5	0,878	0,959
6	0,811	0,917

- Se $|r| > 0,959$, a correlação é significativa. Do contrário, não há evidência suficiente para concluir que a correlação é significativa.

Em palavras

1. Determine o número de dados emparelhados em uma amostra.
2. Especifique o nível de significância.
3. Encontre o valor crítico.

Em símbolos

Determine n .

Identifique α .

Use a tabela
coeficiente de
correlação de Pearson

Em palavras

4. Decida se a correlação é significativa.
5. Interprete a decisão no contexto da afirmação original.

Em símbolos

Se $|r| > \text{valor crítico}$, a correlação é significativa. Se não, não há evidência suficiente para afirmar que a correlação é significativa.

Exemplo: usando uma tabela para testar um coeficiente de correlação populacional ρ

Com os dados do Old Faithful, você usou 25 dados emparelhados para encontrar $r \approx 0.979$. O coeficiente de correlação é significativo? Use $\alpha = 0.05$.

Duração x	Tempo, y	Duração x	Tempo, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Exemplo: usando uma tabela para testar um coeficiente de correlação populacional ρ

- $n = 25, \alpha = 0,05$
- $|r| \approx 0,979 > 0,396$
- No nível de significância 5%, há evidência suficiente para concluir que há correlação linear significativa entre a duração das erupções do Old Faithful e o tempo entre as erupções.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.669	0.798
10	0.635	0.766
11	0.604	0.737
12	0.576	0.710
13	0.550	0.685
14	0.526	0.661
15	0.503	0.639
16	0.482	0.618
17	0.462	0.598
18	0.443	0.579
19	0.425	0.561
20	0.408	0.544
21	0.392	0.528
22	0.377	0.513
23	0.362	0.498
24	0.348	0.484
25	0.335	0.471
26	0.322	0.458
27	0.310	0.446
28	0.298	0.434

Teste de hipótese para um coeficiente de correlação populacional ρ

- Um teste de hipótese também pode ser usado para determinar se o coeficiente de correlação da amostra r fornece evidência suficiente para concluir que o coeficiente de correlação populacional ρ é significativo em um nível de significância especificado.
- Um teste de hipótese pode ser unicaudal ou bicaudal.

Teste de hipótese para um coeficiente de correlação populacional ρ

□ Teste unicaudal à esquerda

$H_0: \rho \geq 0$ (não há correlação negativa significativa)

$H_a: \rho < 0$ (há correlação negativa significativa)

□ Teste unicaudal à direita

$H_0: \rho \leq 0$ (não há correlação positiva significativa)

$H_a: \rho > 0$ (há relação positiva significativa)

□ Teste bicaudal

$H_0: \rho = 0$ (não há correlação significativa)

$H_a: \rho \neq 0$ (há correlação significativa)

Teste t para um coeficiente de correlação

- Usado para testar se a correlação entre duas variáveis é significativa.
- O estatística de teste é r .
- O estatística de teste padronizada segue uma distribuição t com g.l. = $n - 2$.

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

- No curso somente testes de hipótese bicaudais para ρ são considerados.

Usando o teste t para ρ

Em palavras

1. Expresse as hipóteses nula e alternativa.
2. Especifique o nível de significância.
3. Identifique os graus de liberdade.
4. Determine o(s) valor(es) crítico(s) e a(s) região(ões) de rejeição.

Em símbolos

Expresse H_0 e H_a .

Identifique α .

g.l. = $n - 2$.

Use a tabela da distribuição t

Usando o teste t para ρ

Em palavras

5. Encontre o teste estatístico padronizado.
6. Tome a decisão de rejeitar ou não rejeitar a hipótese nula.
7. Interprete a decisão no contexto da afirmação original.

Em símbolos

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Se t está na região de rejeição, rejeite H_0 . Se não, não rejeite H_0 .

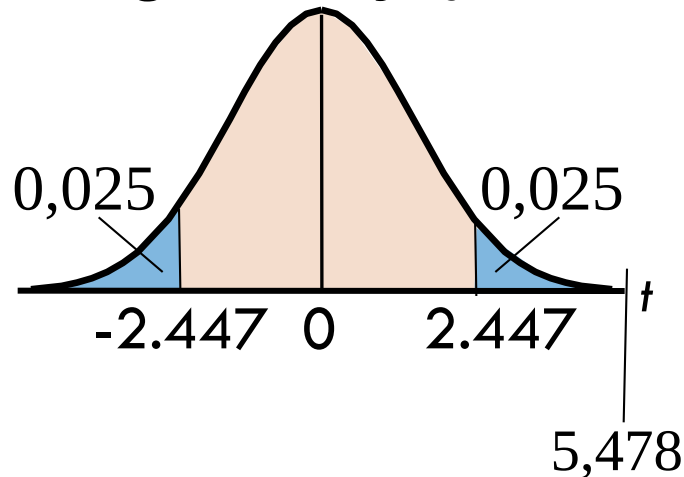
Exemplo: teste t para um coeficiente de correlação

Anteriormente você calculou
 $r \approx 0.9129$. Teste a significância
desse coeficiente de correlação. Use
 $\alpha = 0.05$.

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: teste t para um coeficiente de correlação

- $H_0: \rho = 0$
- $H_a: \rho \neq 0$
- $\alpha = 0,05$
- g.l. = $8 - 2 = 6$
- Região de rejeição:



- Teste estatístico:

$$t = \frac{0.9129}{\sqrt{\frac{1 - (0.9129)^2}{8 - 2}}} \approx 5.478$$

- Decisão: Rejeitar H_0

No nível de significância 5%, existe evidência suficiente para concluir que há uma correlação linear significativa entre os gastos com propaganda e as vendas da empresa.

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 significativa entre duas variáveis, você deve considerar as seguintes possibilidades:
 1. Existe uma relação direta de causa e efeito entre as variáveis?
 - x causa y ?

Correlação e Causalidade

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?

Objetivos

- Introduzir a correlação linear, variáveis dependentes e independentes e tipos de correlação.
- Encontrar o coeficiente de correlação.
- Testar o coeficiente de correlação $\tilde{\rho}$ de uma população usando uma tabela.
- Realizar um teste de hipótese para o coeficiente de correlação $\tilde{\rho}$ de uma população.
- Distinguir entre correlação e causalidade.

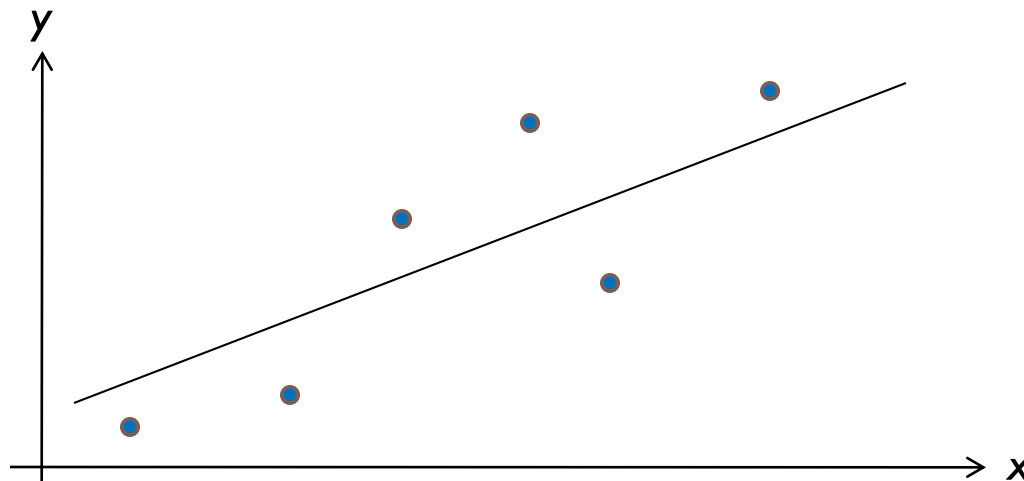
REGRESSÃO LINEAR

Objetivos

- Encontrar a equação da linha de regressão.
- Prever valores y usando uma equação de regressão.

Linhas de regressão

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

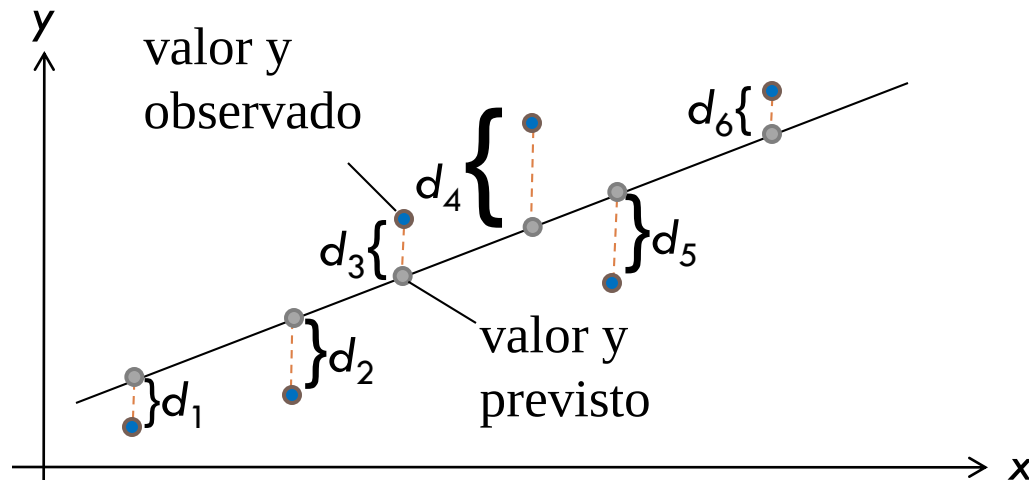


Resíduos

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

Para um dado valor x ,

$$d_i = (\text{valor y observado}) - (\text{valor y previsto})$$



Resíduos

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: $\hat{y} = mx + b$

Diagram illustrating the components of the regression equation:

- $\hat{y}$: valor y previsto para um dado valor x
- m : inclinação
- b : interseção y

Equação da linha de regressão

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

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

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

- $\bar{y}$ 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 $(\bar{x}, \bar{y})$

Encontrando a equação da linha 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

Encontrando a equação da linha de regressão

Exercício anterior:

x	y	xy	x^2	y^2
2,4	225	540	5,76	50.625
1,6	184	294,4	2,56	33.856
2,0	220	440	4	48.400
2,6	240	624	6,76	57.600
1,4	180	252	1,96	32.400
1,6	184	294,4	2,56	33.856
2,0	186	372	4	34.596
2,2	215	473	4,84	46.225
$\Sigma x = 15,8$	$\Sigma y = 1634$	$\Sigma xy = 3289,8$	$\Sigma x^2 = 32,44$	$\Sigma y^2 = 337.558$

Encontrando a equação da linha de regressão

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

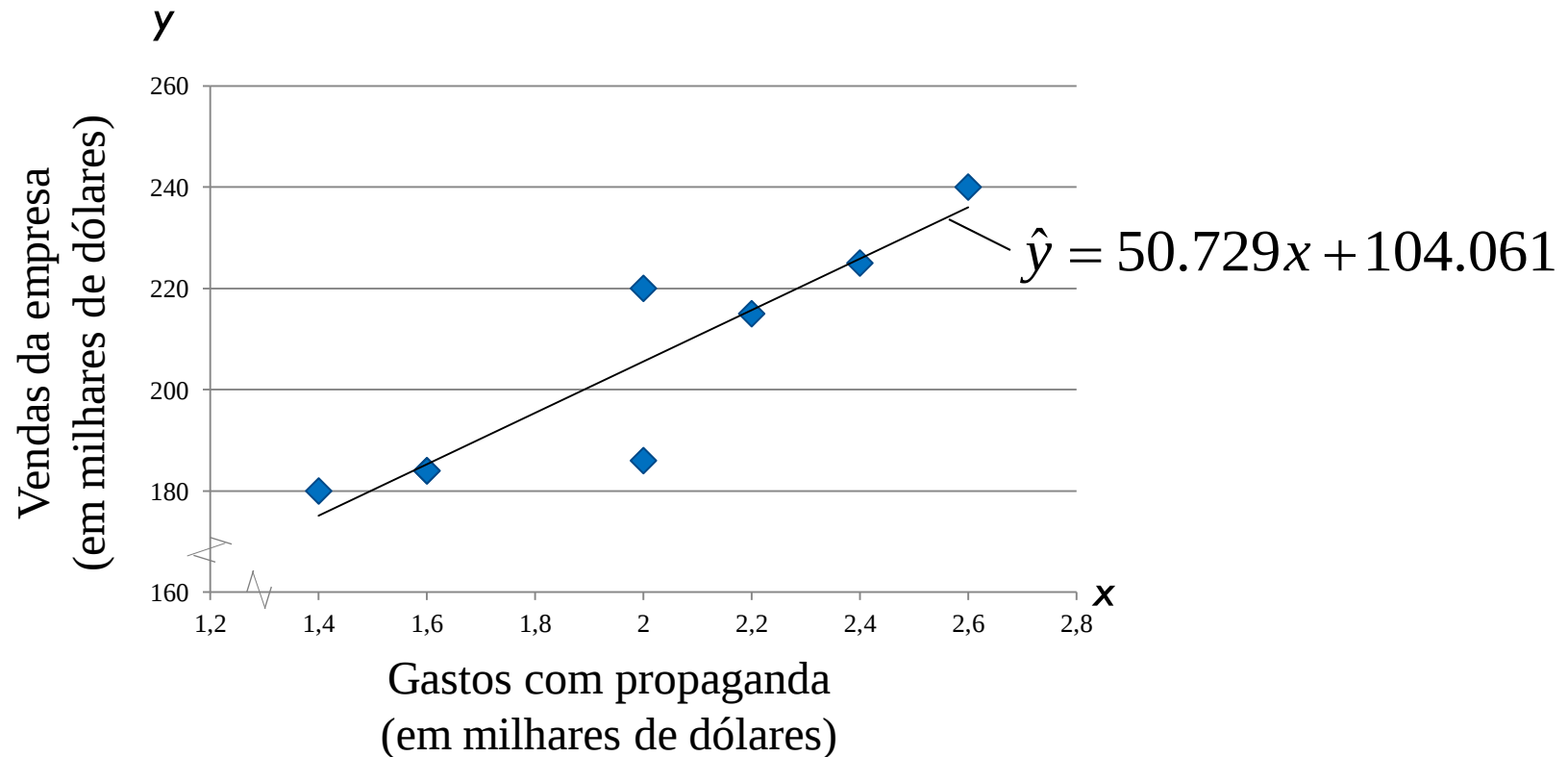
$$m = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2} = \frac{8(3289.8) - (15.8)(1634)}{8(32.44) - 15.8^2}$$
$$= \frac{501.2}{9.88} \approx 50.72874$$

$$b = \bar{y} - m\bar{x} = \frac{1634}{8} - (50.72874)\frac{15.8}{8}$$
$$= 204.25 - (50.72874)(1.975) \approx 104.0607$$

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

Encontrando a equação da linha de regressão

- 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.



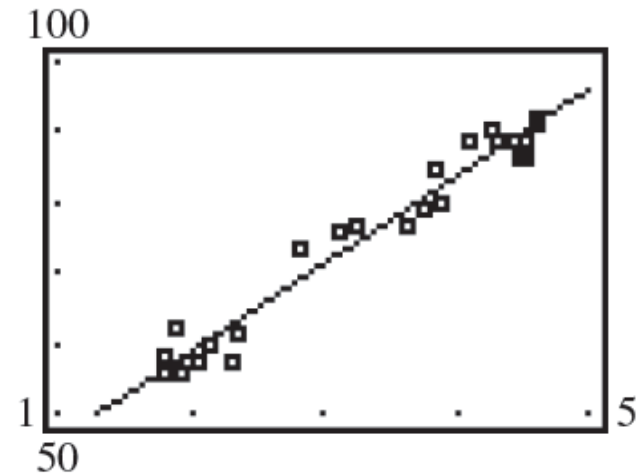
Use a tecnologia para encontrar uma equação de regressão

Use Excel para encontrar a equação da linha de regressão para os dados do Old Faithful. (para casa)

Duração x	Tempo, y	Duração x	Tempo, y
1.8	56	3.78	79
1.82	58	3.83	85
1.9	62	3.88	80
1.93	56	4.1	89
1.98	57	4.27	90
2.05	57	4.3	89
2.13	60	4.43	89
2.3	57	4.47	86
2.37	61	4.53	89
2.82	73	4.55	86
3.13	76	4.6	92
3.27	77	4.63	91
3.65	77		

Use a tecnologia para encontrar uma equação de regressão

EXCEL				
	A	B	C	D
1	Slope:			
2	INDEX(LINEST(known_y's,known_x's),1)			
3				12.48094
4				
5	Y-intercept:			
6	INDEX(LINEST(known_y's,known_x's),2)			
7				33.6829



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.

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

Prever 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.

Prever valores y usando equações de regressão

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 $\hat{y} = 50,729x + 104,061$ para prever as vendas da empresa por gastos com propaganda, tais como 0,5 (\$ 500) ou 5,0 (\$ 5.000).

Objetivos

- Encontrar a equação da linha de regressão.
- Prever valores y usando uma equação de regressão.

MEDIDAS DE REGRESSÃO E INTERVALOS DE PREVISÃO

Objetivos

- Interpretar os três tipos de variação sobre uma linha de regressão.
- Encontrar e interpretar o **coeficiente de determinação**.
- Encontrar e interpretar o **erro padrão de estimativa** para uma linha de regressão.
- Construir e interpretar um intervalo de previsão para y .

Variação sobre uma linha de regressão

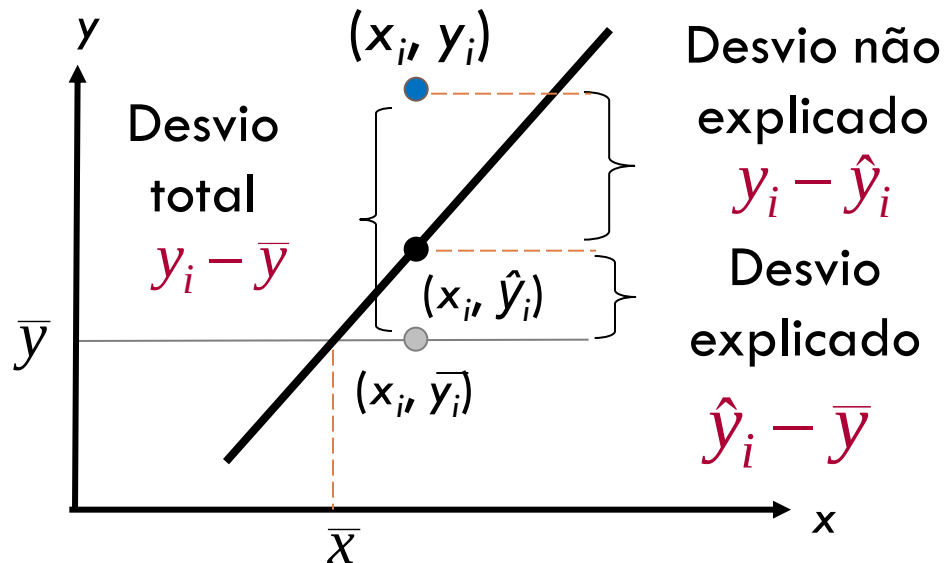
- Três tipos de variação sobre uma linha de regressão:
 - ▣ Variação total.
 - ▣ Variação explicada.
 - ▣ Variação não explicada.
- Para encontrar as variações precisa calcular:
 - ▣ **desvio total.**
 - ▣ **desvio explicado.**
 - ▣ **desvio não explicado.**

Variação sobre uma linha de regressão

Desvio total = $y_i - \bar{y}$

Desvio explicado = $\hat{y}_i - \bar{y}$

Desvio não explicado = $y_i - \hat{y}_i$



Variação sobre uma linha de regressão

Variação total

- A soma dos quadrados das diferenças entre o valor y de cada par ordenado e a média de y .

$$\text{Variação total} = \sum (y_i - \bar{y})^2$$

Variação explicada

- A soma dos quadrados das diferenças entre cada valor y previsto e a média de y .

$$\text{Variação explicada} = \sum (\hat{y}_i - \bar{y})^2$$

Variação sobre uma linha de regressão

Variação não explicada

- A soma dos quadrados das diferenças entre o valor y de cada par ordenado e cada valor y correspondente previsto.

$$\text{Variação não explicada} = \sum (y_i - \hat{y}_i)^2$$

A soma das variações não explicada e explicada é igual à variação total.

$$\text{Variação total} = \text{Variação explicada} + \text{Variação não explicada}$$

Coeficiente de determinação

- A proporção da variação explicada para a variação total.
- Denotada por r^2 .

$$r^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

Exemplo: coeficiente de determinação

O coeficiente de correlação para os gastos de propaganda e dados sobre as vendas da empresa conforme calculado é $r = 0,913$. Encontre o coeficiente de determinação. O que isso lhe diz sobre a variação explicada dos dados sobre a linha de regressão? E sobre a variação não explicada?

Exemplo: coeficiente de determinação

O coeficiente de correlação para os gastos de propaganda e dados sobre as vendas da empresa conforme calculado no Exemplo 4 da Seção 9.1 é $r = 0,913$. Encontre o coeficiente de determinação. O que isso lhe diz sobre a variação explicada dos dados sobre a linha de regressão? E sobre a variação não explicada?

$$\begin{aligned} r^2 &\approx (0.913)^2 \\ &\approx 0.834 \end{aligned}$$

Cerca de **83,4%** da variação nas vendas da empresa podem ser explicadas pela variação dos gastos com propaganda. Cerca de **16,9%** da variação é não explicada.

Erro padrão da estimativa

- O desvio padrão dos valores y_i observados sobre o valor $\hat{y}$ previsto para um dado valor x_i .
- Denotado por s_e .

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

n é o número de pares pedidos no conjunto de dados

- Quanto mais próximos os valores y observados estiverem dos valores y previstos, menor será o erro padrão da estimativa.

Em palavras

1. Faça uma tabela que inclua a coluna exibida a seguir.
2. Use a equação de regressão para calcular os valores y previstos.
3. Calcule a soma dos quadrados das diferenças entre cada valor y observado e o correspondente valor y previsto.
4. Encontre o erro padrão da estimativa.

Em símbolos

$$x_i, y_i, \hat{y}_i, (y_i - \hat{y}_i), (y_i - \hat{y}_i)^2$$

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

$$\Sigma(y_i - \hat{y}_i)^2$$

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n - 2}}$$

Erro padrão da estimativa

A equação de regressão para os gastos de propaganda e dados sobre as vendas da empresa conforme calculado é:

$$\hat{y} = 50,729x + 140,061.$$

Encontre o erro padrão da estimativa.

Solução:

Use a tabela para calcular a soma das diferenças quadradas de cada valor y observado e o correspondente valor y previsto.

Erro padrão da estimativa

x	y	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
2.4	225	225.81	$(225 - 225.81)^2 = 0.6561$
1.6	184	185.23	$(184 - 185.23)^2 = 1.5129$
2.0	220	205.52	$(220 - 205.52)^2 = 209.6704$
2.6	240	235.96	$(240 - 235.96)^2 = 16.3216$
1.4	180	175.08	$(180 - 175.08)^2 = 24.2064$
1.6	184	185.23	$(184 - 185.23)^2 = 1.5129$
2.0	186	205.52	$(186 - 205.52)^2 = 381.0304$
2.2	215	215.66	$(215 - 215.66)^2 = 0.4356$
			$\Sigma = 635.3463$

↑
Variação não
explicada

Erro padrão da estimativa

- $n = 8, \Sigma(y_i - \hat{y}_i)^2 = 635.3463$

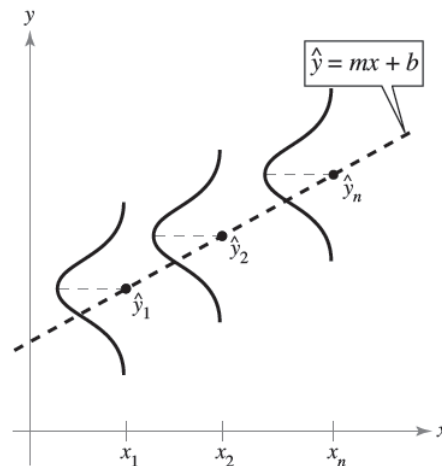
$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{635.3463}{8 - 2}} \approx 10.290$$

O erro padrão da estimativa das vendas da empresa para um gasto com propaganda específico é cerca de \$10.29.

Intervalos de previsão

- Duas variáveis têm uma **distribuição normal bivariada** se para qualquer valor fixo de x , os valores correspondentes de y são normalmente distribuídos, e para qualquer valor fixo de y , os correspondentes valores x são normalmente distribuídos.

Distribuição normal bivariada



Intervalos de previsão

- Um intervalo de previsão pode ser construído para qualquer valor real de y .
- Dada a equação de regressão linear $\hat{y} = mx + b$ e x_0 , um valor específico de x , um **intervalo de previsão c** para y é:

$$\hat{y} - E < y < \hat{y} + E \quad \text{onde:}$$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum x^2 - (\sum x)^2}}$$

- O ponto estimado é $\hat{y}$ e a margem de erro é E . A probabilidade que o intervalo de previsão contenha y é c .

Construindo um intervalo de predição para y para um valor x específico

Em palavras

1. Identifique o número de pares ordenados no conjunto de dados n e os graus de rejeição.
2. Use a equação de regressão e o dado valor x para encontrar o ponto estimado $\hat{y}$.
3. Encontre o valor crítico t_c que corresponda ao nível de confiança c dado.

Em símbolos

$$\text{g.l.} = n - 2$$

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

Use a tabela t

Em palavras

4. Encontre o erro padrão de estimativa s_e .
5. Encontre a margem de erro E .
6. Encontre os extremos esquerdo e direito e forme o intervalo de predição.

Em símbolos

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum x^2 - (\sum x)^2}}$$

Extremo esquerdo: $\hat{y} - E$

Extremo direito: $\hat{y} + E$

Intervalo: $\hat{y} - E < y < \hat{y} + E$

Construir um intervalo de previsão

Construa um intervalo de previsão 95% para vendas da empresa quando os gastos com propaganda são \$ 2,100. O que você pode concluir?

Lembre-se, $n = 8$, $\hat{y} = 50.729x + 104.061$, $s_e = 10.290$

$$\Sigma x = 15.8, \quad \Sigma x^2 = 32.44, \quad \bar{x} = 1.975$$

Construir um intervalo de predição

Construa um intervalo de previsão 95% para vendas da empresa quando os gastos com propaganda são \$ 2,100. O que você pode concluir?

Lembre-se, $n = 8$, $\hat{y} = 50.729x + 104.061$, $s_e = 10.290$

$$\Sigma x = 15.8, \quad \Sigma x^2 = 32.44, \quad \bar{x} = 1.975$$

Solução:

Ponto estimado:

$$\hat{y} = 50.729(2.1) + 104.061 \approx 210.592$$

Valor crítico:

$$\text{g.l.} = n - 2 = 8 - 2 = 6 \quad t_c = 2.447$$

Construindo um intervalo de predição

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum x^2 - (\sum x)^2}}$$
$$= (2.447)(10.290) \sqrt{1 + \frac{1}{8} + \frac{8(2.1 - 1.975)^2}{8(32.44) - (15.8)^2}} \approx 26.857$$

Extremo esquerdo: $\hat{y} - E$

$$210.592 - 26.857 \\ \approx 183.735$$

Extremo direito: $\hat{y} + E$

$$210.592 + 26.857 \\ \approx 237.449$$

$$\mathbf{183.735 < y < 237.449}$$

Você pode ter 95% de confiança que quando os gastos com propaganda forem \$2.100, as vendas da empresa estarão entre \$183.735 e \$237.449.

Objetivos

- Interpretar os três tipos de variação sobre uma linha de regressão.
- Encontrar e interpretar o coeficiente de determinação.
- Encontrar e interpretar o erro padrão de estimativa para uma linha de regressão.
- Construir e interpretar um intervalo de previsão para y .

REGRESSÃO MÚLTIPLA

