Correlação e Regressão

Luan D. Fiorentin

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação

08/11/2019

Sumário

Introdução

- Considere a existência de uma **variável quantitativa** X a qual acreditamos apresentar algum tipo de relação com uma outra **variável quantitativa** Y. Exemplo:
  - Consumo de eletricidade (X) e valor da conta de energia elétrica (Y).
  - Idade (X) e tempo de reação a um estímulo (Y).
  - Temperatura (X) e tempo de uma reação química (Y).
  - entre outras...
- É bastante comum o interesse em estudar a relação entre duas (ou mais) variáveis X e Y.
- Na prática, procura-se encontrar uma variável X que explique a variável Y:

$$Y \cong f(X)$$

Introdução

- Uma das preocupações estatísticas ao analisar dados é a de **criar modelos** do fenômeno em observação.
- As observações frequentemente estão misturadas com variações acidentais ou **aleatórias**.
- Assim, é conveniente supor que cada observação é formada por duas partes: uma **previsível** (ou controlada) e outra **aleatória** (ou não previsível ou não controlada), ou seja

$$\text{observação} = \text{previsível} + \text{aleatório}.$$

Introdução

$$\text{observação} = \text{previsível} + \text{aleatório}$$

- A parte previsível, incorpora o **conhecimento sobre o fenômeno**, sendo usualmente expressa por uma função matemática com parâmetros desconhecidos.
- A parte aleatória deve obedecer algum **modelo de probabilidade**.
- Com isso, o trabalho é produzir **estimativas** para os parâmetro desconhecidos, com base em amostras observadas.

Introdução

- Matematicamente, podemos descrever a relação como

$$y_i = \theta + e_i,$$

onde:

- y_i = observação i.
- θ = efeito fixo, comum a todos os indivíduos.
- e_i = "erro" da observação i, ou um efeito resídual ou aleatório.
- O e_i pode ser resultante de outras variáveis que **não foram controladas** (ou não são controláveis). Logo, essas variáveis não estão explícitas no modelo.

Introdução

Exemplo:

Considerando que o peso médio das plantas é de $\mu = 23$ kg, então o peso de cada planta y_i pode ser descrita pelo seguinte modelo:

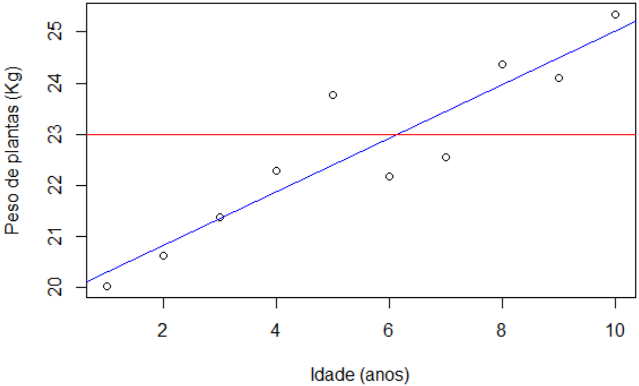
$$y_i = 23 + e_i,$$

onde: $\theta = \mu$ e cada e_i determinará o peso de cada pessoa em função de diversos fatores como: altura, sítio, idade, . . . , ou seja:

$$e_i = f(\text{altura}, \text{sítio}, \text{idade}, \text{adubação}, \ldots)$$

Assim, à medida que relacionamos o peso com outras variáveis, ganhamos informação e diminuimos o erro.

Introdução



Introdução

- Um **modelo linear** entre duas variáveis, X e Y, é definido matematicamente como uma equação com dois parâmetros desconhecidos:

$$Y = \beta_0 + \beta_1 X.$$

- Sendo assim, o modelo anterior onde conheciamos só a média, μ,

$$y_i = \mu + e_i$$

pode ser reescrito como

$$\text{Peso}_i = \beta_0 + \beta_1 \text{Idade}_i + e_i.$$

- **Importante**: o erro deverá diminuir, pois incorporamos uma informação para explicar o peso, que antes estava inserida no erro.

Introdução

- O problema da análise de regressão consiste em definir a **forma funcional** de relação existente entre as variáveis.
- Tipos de relações:
    - Relação Linear: $Y = \beta_0 + \beta_1 X$
    - Relação Polinomial: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$
    - Relação de Potência: $Y = \beta_0 X^{\beta_1}$
- Em todos os casos, a variável Y será dita **dependente**, pois ela que será predita a partir da sua relação com a variável X, chamada de variável **independente**.

Sumário

Regressão Linear Simples

- A **análise de regressão** é a técnica estatística que analisa as relações existentes entre uma única variável dependente e uma ou mais variáveis independentes.
- Em uma **análise de regressão linear** considera-se apenas as variáveis que possuem uma relação linear entre si.
- Uma análise de **regressão linear simples** associa uma única variável independente (X) com uma variável dependente (Y):

$$Y = \beta_0 + \beta_1 X_1 + e$$

- Uma análise de **regressão linear múltipla** associa k variáveis independente (X) com uma variável dependente (Y):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$$

Regressão Linear Simples

- Se for admitido que Y é função linear de X, pode-se estabelecer uma regressão linear simples:
$$Y_i = \beta_0 + \beta_1 X_i + e_i,$$
onde:
- $i = 1, 2, ..., n$;
- Y_i é a variável resposta (ou **dependente**);
- X_i é a variável explicativa (ou **independente**);
- β_0 é o **intercepto** da reta (valor de Y quando $X = 0$);
- β_1 é o **coeficiente angular** da reta (efeito de X sobre Y);
- $e_i \sim N(0, \sigma^2)$ é o **erro**, ou desvio, ou resíduo.

Regressão Linear Simples

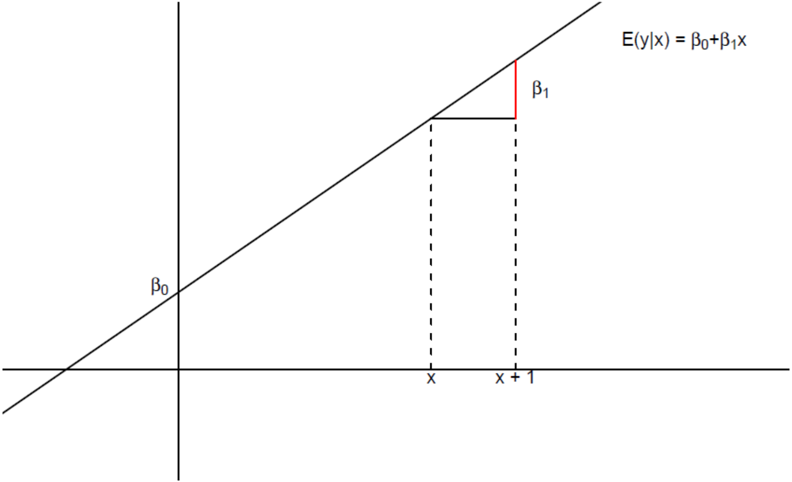
- Interpretação dos Parâmetros:

β_0 representa o ponto em que a reta corta o eixo Y (na maioria das vezes não possui interpretação prática).

β_1 representa a variabilidade em Y causada pelo aumento de uma unidade em X.

- $\beta_1 > 0$ mostra que com o aumento de X, há um aumento em Y.
- $\beta_1 = 0$ mostra que não há efeito de X sobre Y.
- $\beta_1 < 0$ mostra que com a aumento de X, há uma diminuição em Y.

Regressão Linear Simples



Regressão Linear Simples

- A estimação dos parâmetros:
  - **Método dos Mínimos Quadrados (MMQ)**.
  - Método da Máxima Verossimilhança.
- Através de uma amostra, obtem-se uma **estimativa da verdadeira equação de regressão**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

  ou seja, $\hat{Y}_i$ é o valor estimado (predito) de Y_i, por meio das estimativas de β_0 e β_1, que chama-se $\hat{\beta}_0$ e $\hat{\beta}_1$.
- Para cada valor de Y_i, temos um valor $\hat{Y}_i$ estimado pela equação de regressão

$$Y_i = \hat{Y}_i + e_i.$$

Regressão Linear Simples

- Portanto, o erro (ou desvio) de cada observação em relação ao modelo adotado será

$$e_i = Y_i - \hat{Y}_i$$

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

- Qual a **melhor** combinação de β_s? A melhor é aquela que **minimiza a soma de quadrados dos resíduos** (erro) (SQR)

$$SQR = \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2.$$

Regressão Linear Simples

- MMQ baseia-se na determinação de $\hat{\beta}_0$ e $\hat{\beta}_1$ de tal forma que a soma de quadrados dos erros seja mínima.
- Para se encontrar o ponto mínimo de uma função, temos que obter as derivadas parciais em relação a cada parâmetro como

$$\frac{\partial SQR}{\partial \beta_0} = 2\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_i)](-1)$$

$$\frac{\partial SQR}{\partial \beta_1} = 2\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_i)](-X_i).$$

Regressão Linear Simples

Posteriormente, devemos igualar ambas a zero

$$\frac{\partial SQR}{\partial \beta_0} = 2\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)](-1) = 0$$

$$\frac{\partial SQR}{\partial \beta_1} = 2\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)](-X_i) = 0.$$

Regressão Linear Simples

- A solução do sistema apresentado resulta nos seguintes **estimadores de mínimos quadrados**:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

e

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2},$$

onde

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Exercício 1

A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível (litros), em uma amostra de 10 carros novos. Estime os parâmetros de um modelo de regressão linear simples.

Distância	Consumo	Distância	Consumo
20	1,33	80	6,15
60	5,45	70	4,11
15	1,66	73	5,00
45	3,46	28	2,95
35	2,92	85	6,54

Exercício 1

$n = 10;$

$\bar{x} = 51,1;$

$\bar{y} = 3,957;$

$\sum_{i=1}^{n} x_i^2 = 32113;$

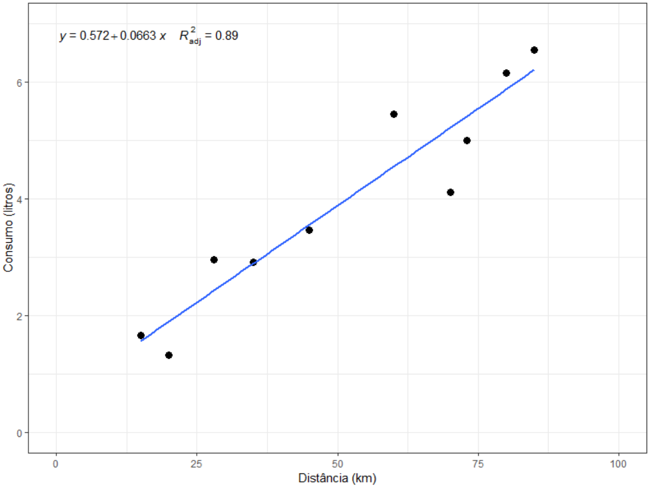
$\sum_{i=1}^{n} x_i \cdot y_i = 2419,6;$

$\hat{\beta}_1 = \frac{2419{,}6 - 10 \cdot 51{,}1 \cdot 3{,}957}{32113 - 10 \cdot 51{,}1^2} = 0,0663;$

$\hat{\beta}_0 = 3,957 + 0,0663 \cdot 51,1 = 0,572;$

$\hat{y}_i = 0,572 + 0,0663 \cdot x_i.$

Exercício 1



y = 0.572 + 0.0663 x   R²adj = 0.89
Consumo (litros)
0
2
4
6
0
25
50
75
100
Distância (km)

Sumário

Correlação Linear

- O objetivo em **regressão linear** é estudar qual a influência de uma V.A. X sob uma V.A. Y, por meio de uma **relação linear**.
- Em uma análise de regressão é indispensável identificar qual variável é **dependente**.
    - Exemplo: o valor da conta de energia elétrica *depende* do consumo de eletricidade (*independente*).
- Na **análise de correlação** isto não é necessário, pois queremos estudar o **grau de relacionamento** entre duas variáveis X e Y, ou seja, uma medida de associação entre elas.
- A **correlação** é considerada como uma medida de **influência mútua** entre variáveis, por isso não é necessário especificar quem influencia e quem é influenciado.

Correlação Linear

- O **coeficiente de correlação linear de Pearson** (r) expressa a **associação linear** entre duas variáveis quantitativa Y e X:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2][\sum_{i=1}^{n} y_i^2 - n\bar{y}^2]}} = \frac{COV(XV)}{DP(X) \cdot DP(Y)},$$

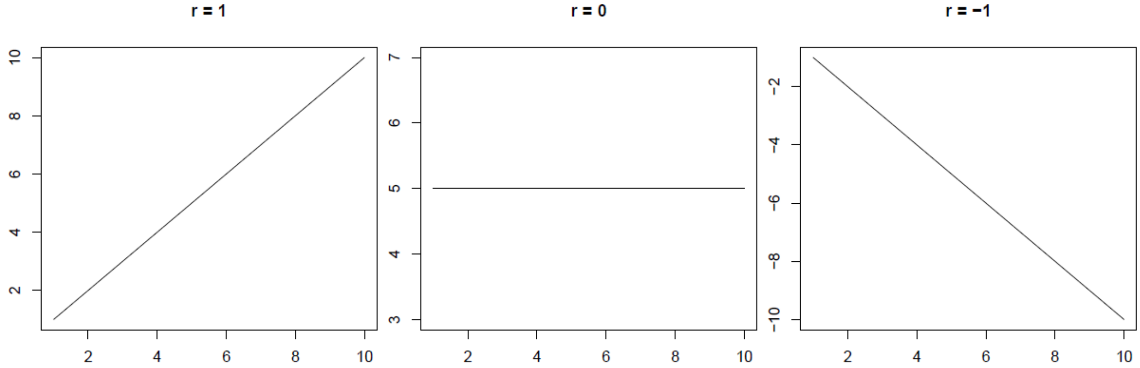
onde:

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

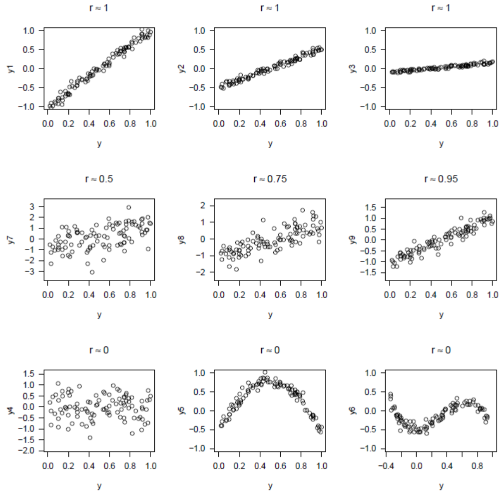
- Logo, se
  - $r = 1$ há correlação positiva perfeita entre as variáveis.
  - $r = 0$ não há correlação.
  - $r = -1$ há correlação negativa perfeita entre as variáveis.

Correlação Linear

Existem muitos tipos de associações possíveis, e o coeficiente de correlação avalia o quanto uma nuvem de pontos no gráfico de dispersão se aproxima de uma reta.



Correlação Linear



Correlação Linear

- Teste de hipótese da existência de correlação linear:
  - Usualmente definimos o coeficiente de correlação para uma **amostra**, pois desconhecemos esse valor para a população.
  - Uma população que tenha duas variáveis **não correlacionadas** pode produzir uma amostra com coeficiente de correlação diferente de zero.
  - Para testar se uma amostra foi colhida de uma população para o qual o coeficiente de correlação entre duas variáveis é nulo, precisamos obter a **distribuição amostral da estatística** r.

Correlação Linear

- Seja ρ o **verdadeiro** coeficiente de correlação populacional **desconhecido**.
- Para testar se o **coeficiente de correlação populacional é igual a zero**, realiza-se um teste de hipótese:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

- A **estatística de teste** é

$$t_{cal} = r\sqrt{\frac{n-2}{1-r^2}},$$

e tem distribuição t de Student com $n-2$ graus de liberdade.

Correlação Linear

Etapas de um teste de hipótese:

1. Formular as hipóteses nula e alternativa.
2. Fixar um valor para o nível de significância α.
3. Construir a Região Crítica (RC) com base no t_{crit} (com $n-2$ graus de liberdade) e estabelecer a Regra de Decisão (RD).
4. Calcular a estatística de teste, sob hipótese nula.
5. Concluir o teste: se a estimativa do parâmetro pertencer à Região Crítica, rejeitamos a Hipótese Nula. Caso contrário, não.

Correlação Linear

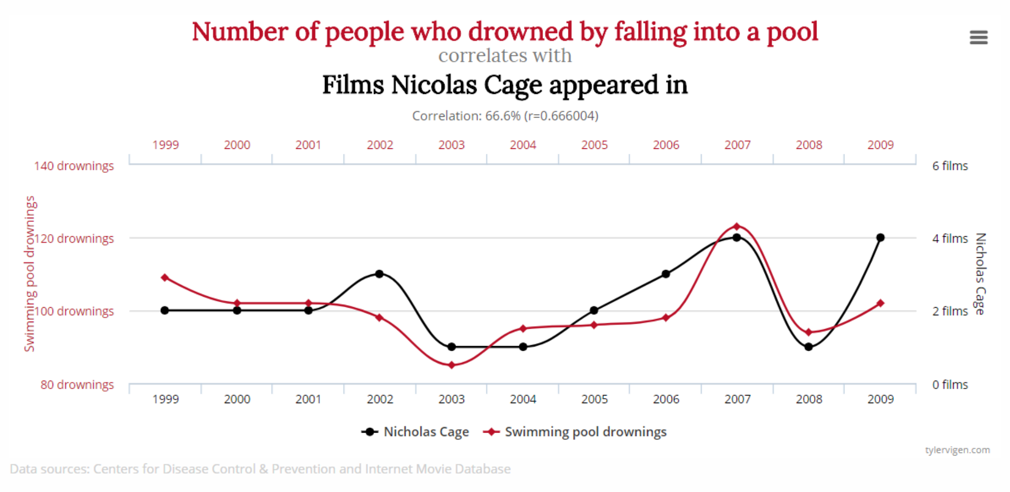
ATENÇÃO! - Correlação não implica causação

Existir uma correlação (positiva ou negativa) entre duas variáveis aleatórias X e Y, mesmo que significativa, não implica que X causa Y.

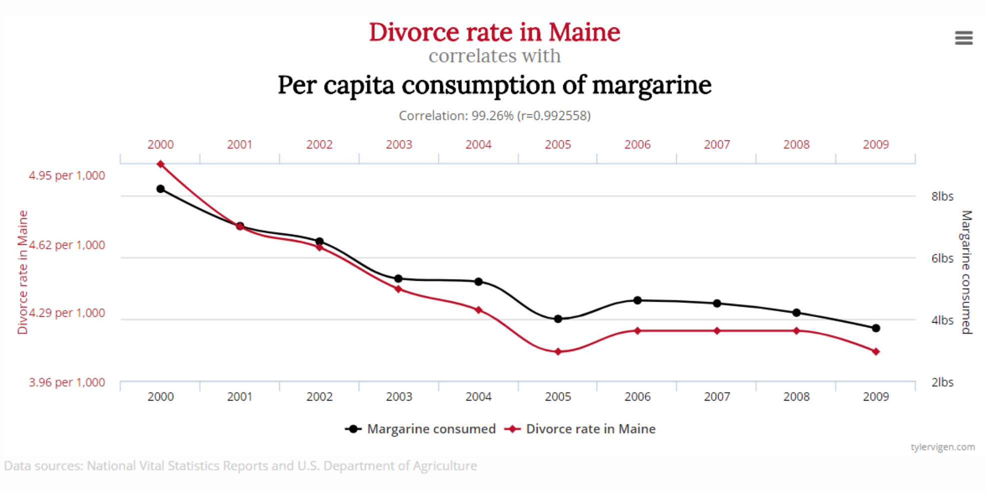
ATENÇÃO! - Correlação não implica causação

Vários exemplos em Spurious correlations:
`http://www.tylervigen.com/spurious-correlations`

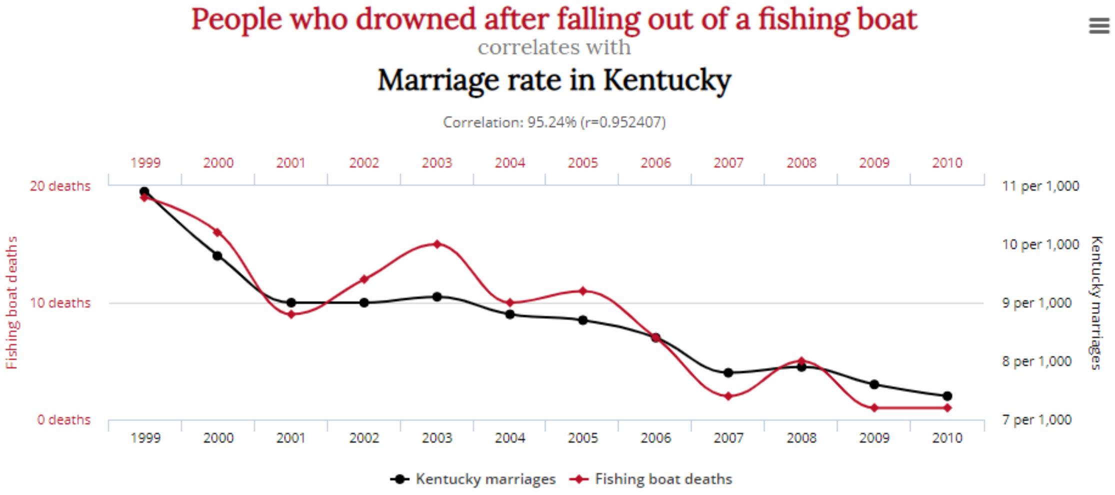
Correlação Linear



Correlação Linear



Correlação Linear



Sumário

Coeficiente de Determinação

- O **coeficiente de determinação** (R^2) é o quadrado do coeficiente de correlação, por consequência

$$0 \leq R^2 \leq 1$$

- O R^2 nos dá a porcentagem de variação em Y que pode ser explicada pela variável independente X.
- Quanto mais próximo de 1, maior é a explicação da variável Y pela variável X.

Exercício 2

A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível (litros), em uma amostra de 10 carros novos. i) Calcule o coeficiente de correlação linear de Pearson e faça o teste de hipótese considerando um nível de significância de 5%. ii) Calcule o coeficiente de determinação. iii) Faça uma predição para o consumo de combustível para uma distância de 50 km.

Distância	Consumo	Distância	Consumo
20	1,33	80	6,15
60	5,45	70	4,11
15	1,66	73	5,00
45	3,46	28	2,95
35	2,92	85	6,54

Exercício 2

$n = 10$; $\bar{x} = 51,1$; $\bar{y} = 3,957$; $\sum_{i=1}^{n} x_i^2 = 32113$; $\sum_{i=1}^{n} y_i^2 = 185,9137$; $\sum_{i=1}^{n} x_i \cdot y_i = 2419,6$;

$$r = \frac{2419{,}6-10{\cdot}51{,}1{\cdot}3{,}957}{\sqrt{(32113-10{\cdot}51{,}1^2)}\sqrt{185{,}9137-10{\cdot}3{,}957^2}} = 0,9476;$$

$$t_{cal} = 0,9476 \cdot \sqrt{\frac{10-2}{1-0{,}9476^2}} = 8,39;$$

- Como t_{cal} com 2 G.L. é $2,31$, então rejeitamos a hipótese nula de que a correlação é igual a zero.
- $R^2 = 0,9476^2 = 0,8975$.
- $\hat{y} = 0,572 + 0,0663 \cdot 50 = 3,887$.

Sumário

Exercícios recomendados

- Seção 9.5: Ex. 1, 2 e 3.
- Seção 9.6: Ex. 22, 25, 26 e 28.