

LCE0216 - Introdução à Bioestatística Florestal

11. Correlação e Regressão Linear

Profa. Dra. Clarice Garcia Borges Demétrio
Monitor: Silvio Gomes

Escola Superior de Agricultura "Luiz de Queiroz"
Universidade de São Paulo

Piracicaba, 18 de Junho 2020

Correlação

Análise do comportamento conjunto de duas ou mais variáveis **quantitativas**.

Exemplos:

- Relação entre altura da árvore e diâmetro à altura do peito;
- Relação entre doses de nitrogênio e produção de determinada cultura;
- Relação entre a porcentagem de nucleotídeos totais e a temperatura em graus centígrados;
- ...

Diagrama de dispersão

Representação gráfica dos pares de valores num sistema cartesiano.

Exemplo: Os dados a seguir são referentes à altura da árvore (Y) e seu diâmetro à altura do peito (X).

Tabela: Dados de altura da árvore (Y) e seu diâmetro à altura do peito (DAP)

Altura	8,1	9,2	8,7	12,7	13,2	12,4	15,7	17,0	18,9	20,1
DAP	5,9	6,3	7,0	9,4	12,0	12,5	15,4	17,0	20,0	23,0

Fonte: Dados simulados

Diagrama de dispersão

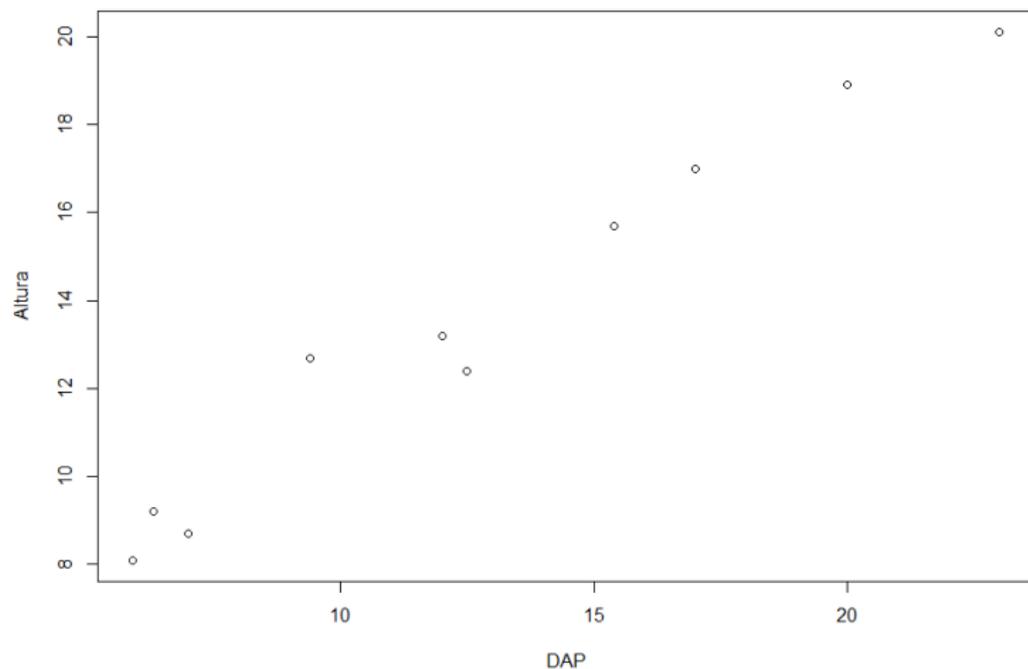


Figura: Diagrama de dispersão das variáveis altura e DAP

Exemplo: Os dados a seguir são referentes ao espaçamento das linhas na cultura de soja (X) e à fração da radiação solar extinta pela planta (Y).

Tabela: Valores de radiação e espaçamento na cultura de soja

Radiação	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1
Espaçamento	0,53	0,51	0,48	0,45	0,44	0,41	0,40	0,39	0,36	0,30

Fonte: Andrade e Ogliari, 2007

Diagrama de dispersão

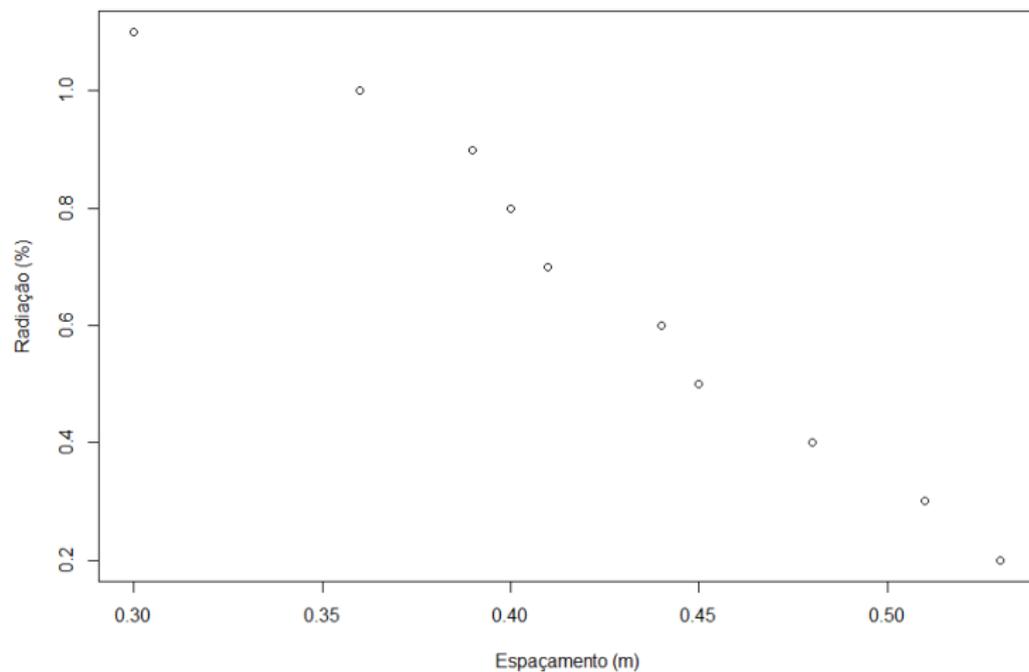


Figura: Diagrama de dispersão das variáveis radiação e espaçamento

Exemplo: Os dados a seguir são referentes à salinidade (g/l) e à temperatura na região III da Lagoa da Conceição, Florianópolis, SC.

Tabela: Valores de salinidade e temperatura na região III da Lagoa da Conceição, Florianópolis, SC

Estação	23	23A	24	25	26	27	27A	28
Temperatura	24,0	23,0	23,0	26,0	25,5	25,0	24,3	23,0
Salinidade	3,85	9,61	2,26	2,06	2,89	9,61	10,58	11,40

Fonte: Andrade e Ogliari, 2007

Diagrama de dispersão

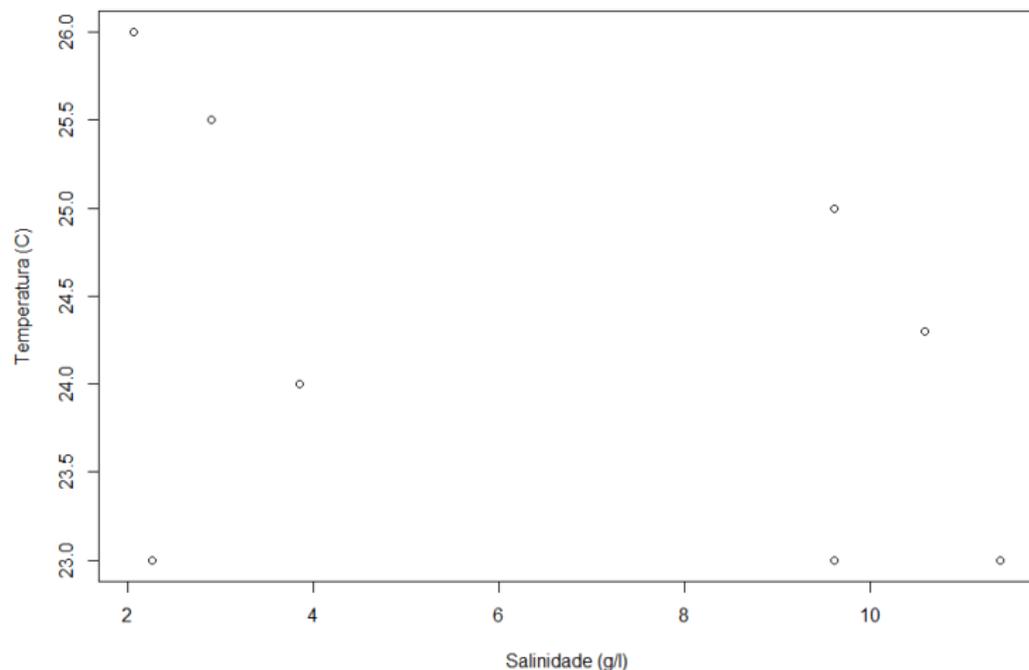


Figura: Diagrama de dispersão das variáveis salinidade e temperatura

Coeficiente de correlação linear de Pearson

Quantifica a correlação entre duas variáveis quantitativas.

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

$r = -1$



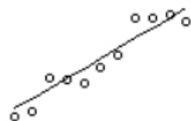
$-1 < r < 0$



$r = 0$



$0 < r < 1$



$r = 1$



Coeficiente de correlação linear de Pearson

$$r = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$$r = \text{Corr}(X, Y) = \frac{n(\sum xy) - (\sum x \sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro à altura do peito (X), calcular o valor do coeficiente de correlação de Pearson:

Tabela: Etapas intermediária para o cálculo do coeficiente de correlação de Pearson

Observação	x	y	x^2	y^2	xy
1	5,9	8,1			
2	6,3	9,2			
3	7,0	8,7			
4	9,4	12,7			
5	12,0	13,2			
6	12,5	12,4			
7	15,4	15,7			
8	17,0	17,0			
9	20,0	18,9			
10	23,0	20,1			
Total					

Tabela: Etapas intermediárias para o cálculo do coeficiente de correlação de Pearson

Observação	x	y	x^2	y^2	xy
1	5,9	8,1	34,8	65,9	47,9
2	6,3	9,2	39,7	84,6	57,9
3	7,0	8,7	49,0	74,9	60,6
4	9,4	12,7	88,4	161,7	119,5
5	12,0	13,2	144,0	174,8	158,6
6	12,5	12,4	156,2	154,0	155,1
7	15,4	15,7	237,2	246,2	241,6
8	17,0	17,0	289,0	290,0	289,5
9	20,0	18,9	400,0	357,4	378,1
10	23,0	20,1	529,0	402,6	461,5
Total	128,5	136,0	1967,23	2011,94	1970,51

Tabela: Etapas intermediárias para o cálculo do coeficiente de correlação de Pearson

Observação	x	y	x ²	y ²	xy
1	5,9	8,1	34,8	65,9	47,9
2	6,3	9,2	39,7	84,6	57,9
3	7,0	8,7	49,0	74,9	60,6
4	9,4	12,7	88,4	161,7	119,5
5	12,0	13,2	144,0	174,8	158,6
6	12,5	12,4	156,2	154,0	155,1
7	15,4	15,7	237,2	246,2	241,6
8	17,0	17,0	289,0	290,0	289,5
9	20,0	18,9	400,0	357,4	378,1
10	23,0	20,1	529,0	402,6	461,5
Total	128,5	136,0	1967,23	2011,94	1970,51

$$\begin{aligned}r = \text{Corr}(X, Y) &= \frac{10(1970,51) - (128,5)(136,0)}{\sqrt{10(1967,23) - 128,5^2} \sqrt{10(2011,94) - 136,0^2}} \\ &= \frac{2229,1}{2264,956} = 0,9842\end{aligned}$$

Regressão Linear

- Equação matemática linear;
- Representação de um conjunto de dados;
- Relação de causa e efeito;
- Interpolação e Extrapolação.

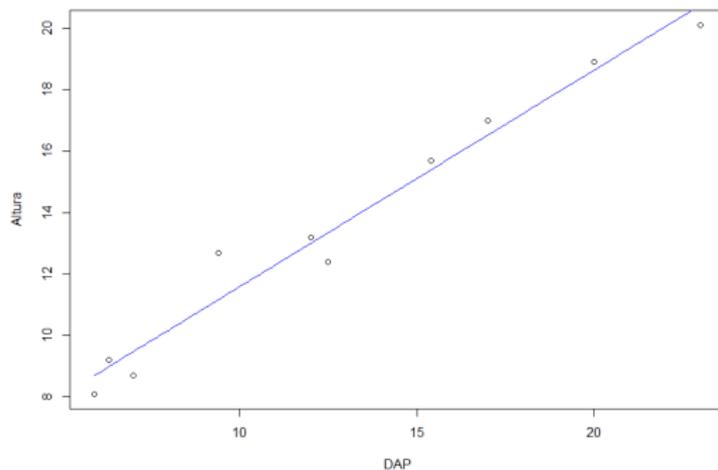


Figura: Diagrama de dispersão das variáveis altura e DAP

Variáveis:

X \Rightarrow Variável **Independente**

Y \Rightarrow Variável **Dependente**

Equação matemática:

$$y = \alpha + \beta x,$$

em que α representa o intercepto e β o coeficiente angular.

Interpretação prática do parâmetro β : o quanto varia a resposta y para um acréscimo de uma unidade na variável x .

Modelo Estatístico

$$y = \alpha + \beta x + \epsilon$$

Reta ajustada:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

ou

$$\hat{y} = a + bx$$

em que $\hat{\alpha}$ (ou a) e $\hat{\beta}$ (ou b) são as estimativas dos parâmetros α e β .

Estimativas pelo método dos mínimos quadrados:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Isolando ε_i tem-se:

$$\varepsilon_i = y_i - (\alpha + \beta x_i)$$

Para obter a reta com o menor erro (resíduo) possível em relação ao conjunto de dados, devemos minimizar a soma de quadrados dos erros (resíduos)

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

Estimativas pelo método dos mínimos quadrados:

$$\frac{\partial Q}{\partial \alpha} = 0 \rightarrow 2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)(-1) = 0 \quad (I)$$

$$\frac{\partial Q}{\partial \beta} = 0 \rightarrow 2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)(-x_i) = 0 \quad (II)$$

Isolando $\hat{\alpha}$ em (I) tem-se:

$$\sum_{i=1}^n y_i - n\hat{\alpha} - \hat{\beta} \sum_{i=1}^n x_i = 0$$

$$\sum_{i=1}^n y_i - \hat{\beta} \sum_{i=1}^n x_i = n\hat{\alpha}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^n y_i}{n} - \hat{\beta} \frac{\sum_{i=1}^n x_i}{n}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Estimativas pelo método dos mínimos quadrados:

Substituindo (I) em (II), tem-se:

$$\sum_{i=1}^n x_i (y_i - (\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}x_i) = 0$$

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} + \hat{\beta} \sum_{i=1}^n x_i \bar{x} - \hat{\beta} \sum_{i=1}^n x_i^2 = 0$$

$$\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \frac{\sum_{i=1}^n y_i}{n} = -\hat{\beta} \sum_{i=1}^n x_i \frac{\sum_{i=1}^n x_i}{n} + \hat{\beta} \sum_{i=1}^n x_i^2$$

$$\hat{\beta} \left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}$$

$$\hat{\beta} \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right) = n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$

$$\hat{\beta} = \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

Estimativas pelo método dos mínimos quadrados:

Portanto, as estimativas de $\hat{\alpha}$ e $\hat{\beta}$ pelo método de mínimos quadrados são dadas por:

$$\hat{\beta} = \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

e

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

em que n corresponde ao tamanho da amostra.

Obs: Para comprovar que as estimativas encontradas correspondem ao ponto de mínimo da função, deve-se fazer o estudo do sinal do determinante da matriz hessiana e das derivadas parciais de segunda ordem.

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro à altura do peito (X):

Tabela: Etapas intermediárias

Observação	x	y	x^2	y^2	xy
1	5,9	8,1	34,8	65,9	47,9
2	6,3	9,2	39,7	84,6	57,9
3	7,0	8,7	49,0	74,9	60,6
4	9,4	12,7	88,4	161,7	119,5
5	12,0	13,2	144,0	174,8	158,6
6	12,5	12,4	156,2	154,0	155,1
7	15,4	15,7	237,2	246,2	241,6
8	17,0	17,0	289,0	290,0	289,5
9	20,0	18,9	400,0	357,4	378,1
10	23,0	20,1	529,0	402,6	461,5
Total	128,5	136,0	1967,23	2011,94	1970,51

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro à altura do peito (X):

Tabela: Etapas intermediárias

Observação	x	y	x^2	y^2	xy
1	5,9	8,1	34,8	65,9	47,9
2	6,3	9,2	39,7	84,6	57,9
3	7,0	8,7	49,0	74,9	60,6
4	9,4	12,7	88,4	161,7	119,5
5	12,0	13,2	144,0	174,8	158,6
6	12,5	12,4	156,2	154,0	155,1
7	15,4	15,7	237,2	246,2	241,6
8	17,0	17,0	289,0	290,0	289,5
9	20,0	18,9	400,0	357,4	378,1
10	23,0	20,1	529,0	402,6	461,5
Total	128,5	136,0	1967,23	2011,94	1970,51

$$\hat{\beta} = \frac{10(1970,51) - (128,5)(136,0)}{10(1967,23) - (128,5)^2} = 0,7053$$

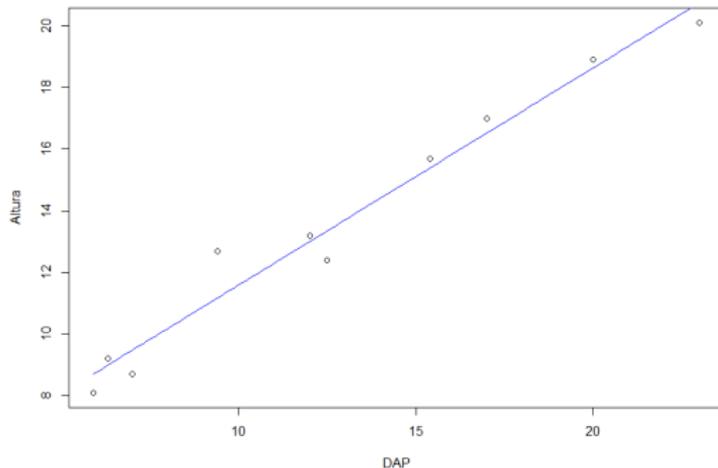
$$\hat{\alpha} = \frac{136,0 - 0,7053(128,5)}{10} = 4,5368$$

Ajuste de uma reta

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro à altura do peito (X):

Reta ajustada

$$\hat{y}_i = 4,5368 + 0,7053x_i.$$



- A análise de variância representa o desdobramento da soma de quadrados total (SQtotal) em diversos componentes que podem explicar o fenômeno em questão.
- No caso da regressão linear simples, pode-se desdobrar a SQTotal em soma de quadrados da regressão (SQReg) e soma de quadrados dos resíduos (SQres).

$$SQ_{total} = SQ_{Reg} + SQ_{res}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Pode-se, então, verificar estatisticamente se as variáveis X e Y apresentam a suposta relação linear por meio da análise de variância, o que equivale a testar a hipótese:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

Para isso, devem-se obter as seguintes quantidades

Causa de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F
Regressão linear	glreg	SQReg	QMReg	F_{cal}
Resíduo	glres	SQRes	QMRes	-
Total	gltotal	SQTotal	-	-

Análise de variância

$$gl_{reg} = 1$$

$$gl_{res} = n - 2$$

$$gl_{total} = n - 1$$

$$SQ_{Reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{(\sum_{i=1}^n x_i y_i - 1/n \sum_{i=1}^n x_i \sum_{i=1}^n y_i)^2}{\sum_{i=1}^n x_i^2 - 1/n (\sum_{i=1}^n x_i)^2}$$

$$SQ_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SQ_{Total} - SQ_{Reg}$$

$$SQ_{Total} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 / n$$

$$QM_{Reg} = SQ_{Reg} / gl_{reg}$$

$$QM_{Res} = SQ_{Res} / gl_{res}$$

$$F_{cal} = QM_{Reg} / QM_{Res}$$

$$F_{tab}(1; n - 2; \alpha)$$

Se $F_{cal} > F_{tab}$, rejeita-se a hipótese nula.

Análise de variância

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro à altura do peito (X):

Tabela: Etapas intermediárias

Observação	x	y	\hat{y}	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	5,9	8,1	8,7	24,03	0,36
2	6,3	9,2	9,0	21,34	0,05
3	7,0	8,7	9,5	17,02	0,60
4	9,4	12,7	11,2	5,92	2,35
5	12,0	13,2	13,0	0,36	0,04
6	12,5	12,4	13,4	0,06	0,91
7	15,4	15,7	15,4	3,23	0,09
8	17,0	17,0	16,5	8,57	0,22
9	20,0	18,9	18,6	25,43	0,07
10	23,0	20,1	20,8	51,25	0,43
Soma		136		157,21	5,12

$$F_{tab}(1; 8; \alpha = 0,05) = 5,32$$

Causa de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F
Regressão linear	1	157,21	157,21	245,7
Resíduo	8	5,12	0,64	-
Total	9	162,33	-	-

Como $F_{cal} > F_{tab}$, rejeita-se a hipótese nula. Há indícios de que o parâmetro β é importante para explicar o fenômeno estudado.

Resíduo:

Diferença entre o valor observado (y_i) e o valor predito (\hat{y}_i), para um determinado valor x_i :

$$e_i = y_i - \hat{y}_i.$$

O primeiro resíduo (simples) é dado por:

Resíduo:

Diferença entre o valor observado (y_i) e o valor predito (\hat{y}_i), para um determinado valor x_i :

$$e_i = y_i - \hat{y}_i.$$

O primeiro resíduo (simples) é dado por:

$$e_1 = 8,1 - (4,5368 + 0,7053 \times 5,9) = 8,1 - 8,7 = -0,6$$

Modelo bem ajustado:

é aquele que apresenta resíduos pequenos.

Resíduo simples \Rightarrow depende das unidades de medida



Resíduos Padronizados $\Rightarrow z_i = \frac{e_i}{\sqrt{\sum_{i=1}^n e_i^2 / (n-2)}}$

Na prática: erro pequeno \Rightarrow resíduo padronizado entre -2 e 2.

Verificação da qualidade do ajuste

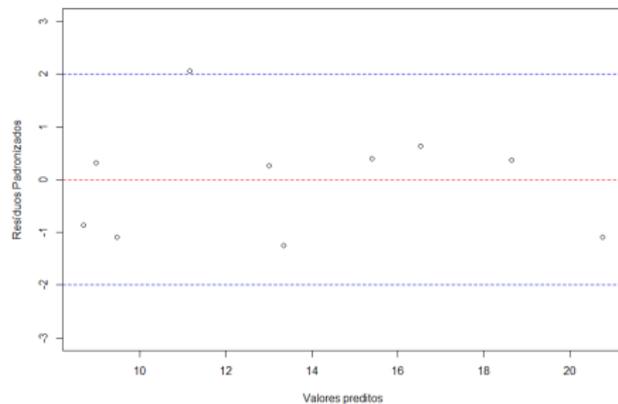


Figura: Gráfico dos valores preditos *versus* resíduos padronizados

Ideal: Gráfico sem padrão!

Verificação da qualidade do ajuste

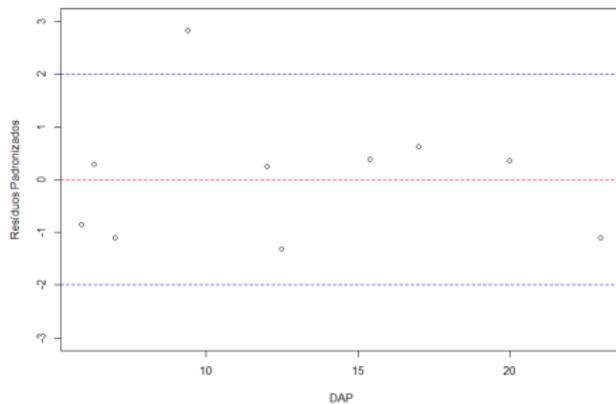


Figura: Gráfico dos valores de DAP *versus* resíduos padronizados

Ideal: Gráfico sem padrão!

Coefficiente de determinação R^2

- Este coeficiente indica quanto da variabilidade na variável dependente Y está sendo “explicada” pela variável independente X .

$$R^2 = \frac{SQ_{\text{Reg}}}{SQ_{\text{total}}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- O valor de R^2 varia no intervalo de 0 a 1. Valores próximos de 1 indicam que o modelo proposto é adequado para descrever o fenômeno.

Obs: Esse coeficiente apresenta uma relação diretamente proporcional ao número de parâmetros do modelo de regressão.

Verificação da qualidade do ajuste

$$\hat{y}_i = 4,5368 + 0,7053x_i.$$

Exemplo: Considerando-se o exemplo de altura da árvore (Y) e o diâmetro à altura do peito (X):

Tabela: Etapas intermediárias

Observação	x	y	\hat{y}	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	5,9	8,1	8,7	24,03	0,36
2	6,3	9,2	9,0	21,34	0,05
3	7,0	8,7	9,5	17,02	0,60
4	9,4	12,7	11,2	5,92	2,35
5	12,0	13,2	13,0	0,36	0,04
6	12,5	12,4	13,4	0,06	0,91
7	15,4	15,7	15,4	3,23	0,09
8	17,0	17,0	16,5	8,57	0,22
9	20,0	18,9	18,6	25,43	0,07
10	23,0	20,1	20,8	51,25	0,43
Soma		136		157,21	5,12

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{(8,7 - 13,6)^2 + \dots + (20,8 - 13,6)^2}{(8,1 - 13,6)^2 + \dots + (20,1 - 13,6)^2} \\ &= \frac{157,21}{162,33} = 0,97 \end{aligned}$$

Exercício:

Um inventário florestal foi realizado em uma floresta de *Eucalyptus grandis*, no município de Itatinga (SP). Os dados que se seguem representam os valores de diâmetro, altura total e volume das árvores em parcelas de $500m^2$.

Árvore	DAP (cm)	Altura (m)	Volume (m^3)	Árvore	DAP (cm)	Altura (m)	Volume (m^3)
1	12,4	21,1	0,13	11	16,2	27,7	0,26
2	14,7	25,7	0,22	12	17,6	28,4	0,28
3	12,4	22,4	0,13	13	17,0	27,5	0,29
4	12,4	21,7	0,12	14	13,8	23,0	0,15
5	15,7	25,4	0,24	15	16,4	27,1	0,23
6	16,7	26,5	0,27	16	18,5	27,3	0,31
7	16,1	26,7	0,27	17	20,5	27,3	0,42
8	15,9	27,0	0,28	18	17,7	25,4	0,29
9	18,1	27,5	0,34	19	17,8	26,7	0,30
10	17,1	27,0	0,29	20	17,9	27,3	0,34

: Sabe-se que o volume (y) da árvore possui relação com a combinação do diâmetro e a altura ($x = DAP^2 Alt$). Obtenha as estimativas dos parâmetros do seguinte modelo de regressão linear simples:

$$y_i = \alpha + \beta x_i + \varepsilon_i, i = 1, 2, \dots, 20$$

Usando a equação estimada, obtenha os volumes estimados das árvores faltantes da parcela:

Árvore	DAP (cm)	Altura (m)	Volume (m^2)	Árvore	DAP (cm)	Altura (m)	Volume (m^2)
21	13,8	25,1		25	16,9	26,0	
22	15,6	26,5		26	20,2	26,3	
23	19,2	27,5		27	14,7	24,4	
24	11,0	20,1		28	12,4	25,2	

Qual o volume total da parcela?

Exercício: Uma das maneiras de calcular o volume de árvores é por meio do método do xilômetro. Esse método consiste em medir o volume de um sólido a partir do deslocamento da água quando o sólido é mergulhado num recipiente com água. As dimensões do recipiente com água são: 3m metros de altura e raio de 0,8m. Sabendo-se que o recipiente é cilíndrico e está 50% preenchido com água, calcule os volumes das árvores, sendo Δy o deslocamento observado de água:

Árvore	DAP	Altura	Δy
1	13,4	19,8	0,17
2	16,3	23,6	0,22
3	11,8	19,4	0,15
4	21,7	25,1	0,26
5	9,5	16,2	0,11
6	15,8	23,9	0,20
7	12,9	18,7	0,16
8	23,1	25,7	0,29
9	12,3	22,6	0,18
10	19,4	21,6	0,22

Com os volumes calculados, obtenha as estimativas dos parâmetros dos seguintes modelos:

$$y_i = \alpha + \beta DAP_i + e_i$$

$$y_i = \alpha + \beta(DAP_i^2 Alt) + e_i$$

Pede-se:

- Calcule o coeficiente de determinação e gráfico de resíduos padronizado de ambos os modelos. Qual modelo explica melhor a variação do volume? Obs.: utilize no gráfico os valores preditos no eixo das abcissas.
- Calcule a correlação a partir da escolha da variável independente do melhor modelo.
- Represente graficamente a reta ajustada dos dois modelos.
- Construa o quadro da Análise de variância do modelo escolhido. As variáveis X e Y possuem relação linear?

Exercício: Os dados que se seguem referem-se a porcentagens de sementes de *Pinus taeda* que flutuam em água e porcentagens de germinação das mesmas. O pesquisador está interessado em saber se se deve jogar fora as sementes que flutuam (acredita-se que as sementes que flutuam não servem para produção de mudas).

Árvore	Flutuação	Germinação	Árvore	Flutuação	Germinação
1	85,67	90,07	13	75,00	87,61
2	79,67	88,71	14	72,67	88,99
3	72,33	96,58	15	71,00	91,96
4	68,33	93,20	16	80,67	91,37
5	42,33	88,89	17	43,67	100,00
6	42,67	96,43	18	40,33	90,48
7	81,00	86,71	19	83,00	84,93
8	80,33	84,06	20	77,00	78,40
9	77,00	78,40	21	66,67	93,88
10	77,67	90,23	22	77,33	90,77
11	65,33	98,95	23	67,67	93,14
12	56,67	97,10	24	65,67	90,63