NOTAS DE AULA

ASSOCIAÇÃO ENTRE VARIÁVEIS

Prof.: Idemauro Antonio Rodrigues de Lara

Associação entre variáveis

Em um estudo observacional ou experimental, frequentemente tem-se interesse na **análise conjunta de duas** (ou mais) **variáveis.**

Neste curso, vamos considerar as seguintes situações:

1. Variável qualitativa $\times$ Variável qualitativa: tabelas de contingência.
2. Variável qualitativa $\times$ Variável quantitativa: duas ou mais amostras estratificadas por um atributo (fator).
3. Variável quantitativa $\times$ Variável quantitativa: correlação e regressão linear simples.

Associação entre variáveis qualitativas

Suponha que tenha uma amostra de n elementos, os quais foram classificados por dois atributos A e B (variáveis qualitativas). Se o atributo A tem r categorias e o atributo B s categorias, as frequências conjuntas observadas podem ser dispostas em uma tabela de contingência $r \times s$.

Tabela 1: Estrutura geral de uma tabela de contingência $r \times s$.

	B				
A	1	2	...	s	Total
1	n_{11}	n_{12}	...	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2s}	$n_{2.}$
...	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.s}$	n

Associação entre variáveis qualitativas

As hipóteses mais comuns acerca de uma tabela de contingência são as de **homogeneidade** e **independência**.

Observe que a estrutura da Tabela 1 nos permite estimar três tipos de frequências percentuais, a saber:

Em relação ao total geral: $\frac{n_{ij}}{n}$.

Em relação ao total linha: $\frac{n_{ij}}{n_{i.}}$.

Em relação ao total coluna: $\frac{n_{ij}}{n_{.j}}$.

A análise destas frequências percentuais podem dar um "indicativo" sobre a questão do estudo.

Exemplo 1. - (Hoffmann, 1999) Em uma pesquisa envolvendo 360 agricultores, deseja-se investigar se o fato de ser cooperado ou não independe do tipo de posse de terra. Os resultados obtidos estão dispostos a seguir.

Tabela 2: Distribuição conjunta dos agricultores segundo a posse da terra e cooperação.

	Cooperação		
Posse da terra	Sim	Não	Total
Arrendatário	43	107	150
Parceiro	25	65	90
Proprietário	52	68	120
Total	120	240	360

Com base nos resultados, há indicativos de dependência entre as variáveis?

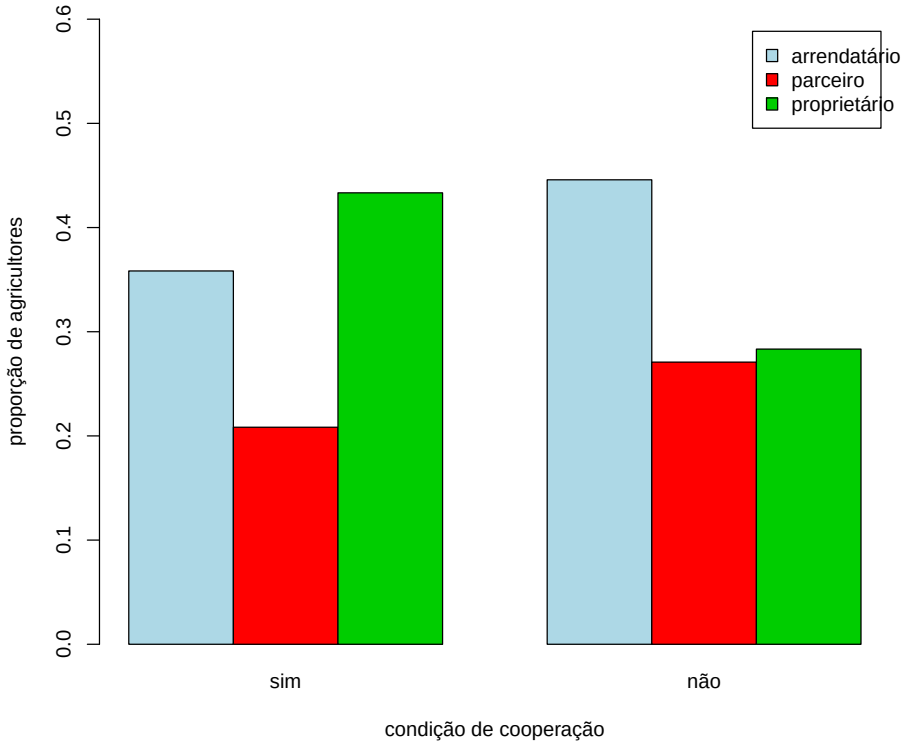

Figura 1: Distribuição do tipo de posse de terra dado à condição ser ou não cooperado.

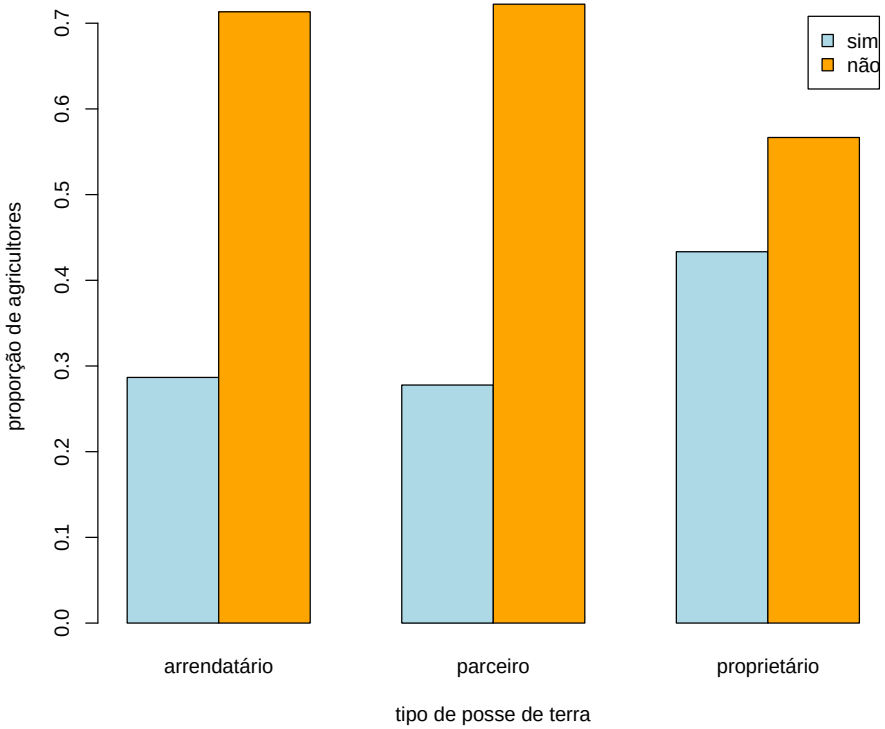

Figura 2: Distribuição de cooperados e não cooperados dado à condição de posse da terra.

Medidas de Associação

Estatística Qui-quadrado

É uma medida de discrepância entre os valores observados ($fo_{ij} = n_{ij}$) e os valores esperados sob a hipótese de independência (fe_{ij}).

$$Q^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

sendo:

$$fe_{ij} = \frac{n_{i.} n_{.j}}{n} \quad \forall \quad ij$$

Coeficiente de Contingência (Karl Pearson)

$$C = \sqrt{\frac{Q^2}{Q^2 + n}}$$

ou ainda,

$$C^* = \frac{C}{\sqrt{\frac{t-1}{t}}}$$

em que $t = \min(r, s)$.

$$0 \leq C^* \leq 1$$

Para o exemplo disposto na Tabela 2 temos:

$$Q^2 = 8,12$$

$$t = \min(3, 2) = 2$$

$$C = \sqrt{\frac{8,12}{8,12 + 360}} = 0,1485$$

$$C^* = \frac{0,1485}{\sqrt{\frac{2-1}{2}}} = 0,21$$

Correlação e Regressão linear simples

Objetivo:

Identificar e descrever associação e relação funcional entre duas variáveis quantitativas.

Correlação linear simples

Considere uma situação típica de análise de dados, em que a n unidades amostrais, estão associadas duas variáveis quantitativas.

Unidade amostral	X	Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Na análise de correlação tem-se interesse em verificar se há algum tipo de associação entre as variáveis, em particular:

i. Se para maiores valores de X, também se observam, em média, maiores valores de Y, diz-se que há uma associação do tipo direta ou positiva;

ii. Se para maiores valores de X, também se observam, em média, menores valores de Y, diz-se que há uma associação do tipo inversa ou negativa.

Observação: Pode acontecer de não existir associação entre as variáveis. Neste curso, tem-se interesse em **identificar** e **quantificar as relações dos tipos lineares**.

As técnicas usadas para descrever e quantitificar o grau de associação entre duas variáveis quantitiativas são:

i. Tabelas

ii. Gráfico: diagrama de dispersão;

iii. Medidas descritivas: covariância e coeficiente de correlação

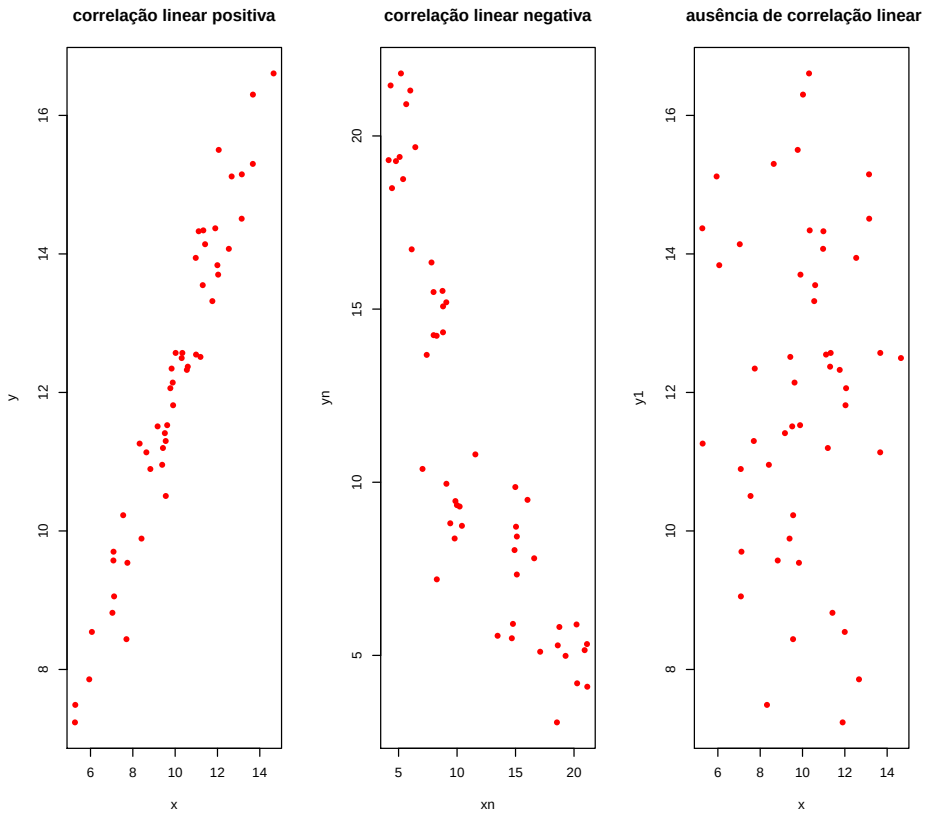

Figura 3: Diagramas de dispesão e tipos de associação.

Um exemplo

Exemplo: Os dados a seguir são referentes à produção de matéria seca de uma cultura (Y) e a quantidade de radiação fotossintética ativa (X).

Tabela 3: Dados de produção de matéria seca e a radiação fotossintética ativa

Produção	10	60	110	160	220	280	340	400	460	520
Radiação	18	55	190	300	410	460	570	770	815	965

Fonte: Andrade e Ogliari, 2007

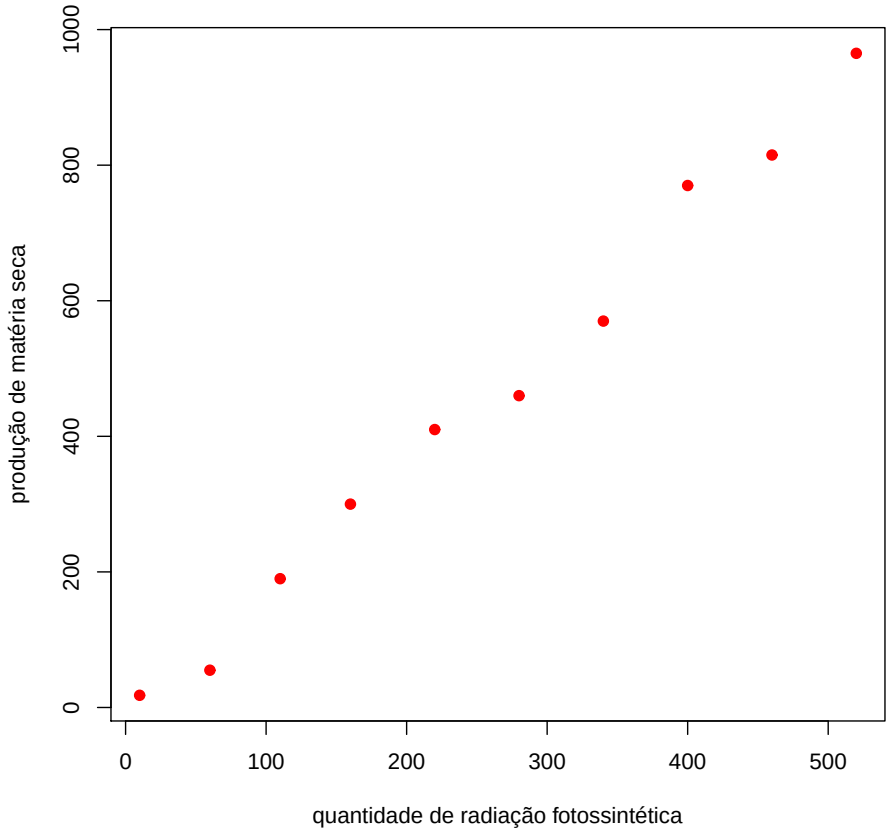

Figura 4: Diagrama de dispersão entre a radiação fotossintética e a produção de matéria seca de uma cultura

A inspeção dos dados e do diagrama de dispersão nos indica que há uma associação do tipo direta, ou seja, uma correlação linear positiva entre condutividade e salinidade. Mas qual o grau desta associação?

Coeficiente de Correlação linear simples

É uma medida de associação que quantifica a **correlação linear** entre duas variáveis quantitativas.

$$\text{Corr}(X,Y) = r = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{\displaystyle\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left[\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\displaystyle\sum_{i=1}^{n} y_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

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

Coeficiente de Correlação linear simples

Alternativamente:

$$\text{Corr}(X,Y) = r = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left[\displaystyle\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]\left[\displaystyle\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]}}$$

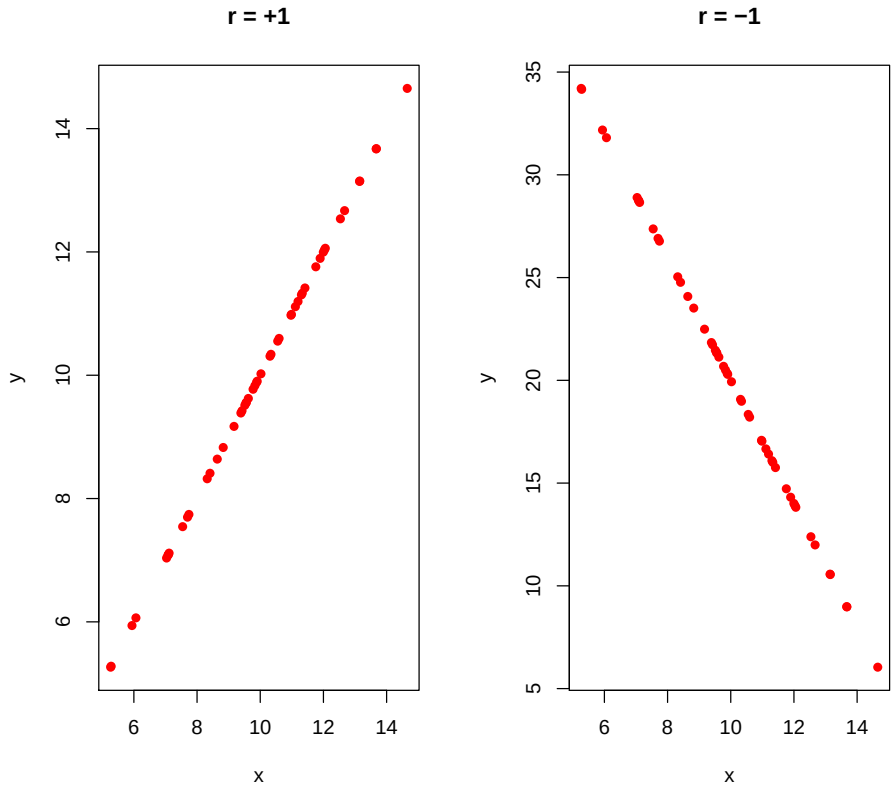

Figura 5: Diagramas de dispersão com correlações lineares perfeitas.

Exemplo: Considere os dados a seguir são referentes à produção de matéria seca de uma cultura (Y) e a quantidade de radiação fotossintética ativa (X).

Tabela 4: Etapas intermediária para o cálculo do coeficiente de correlação de Pearson

Observação	x	y	x^2	y^2	xy
1	18	10			
2	55	60			
3	190	110			
4	300	160			
5	410	220			
6	460	280			
7	570	340			
8	770	400			
9	815	460			
10	965	520			
Total					

Tabela 5: Etapas intermediária para o cálculo do coeficiente de correlação de Pearson

Observação	x	y	x^2	y^2	xy
1	18	10	324	100	180
2	55	60	3025	3600	3300
3	190	110	36100	12100	20900
4	300	160	90000	25600	48000
5	410	220	168100	48400	90200
6	460	280	211600	78400	128800
7	570	340	324900	115600	193800
8	770	400	592900	160000	308000
9	815	460	664225	211600	374900
10	965	520	931225	270400	501800
Total	4553	2560	3022399	925800	1669880

$$
\begin{aligned}
r = \mathrm{Corr}(X, Y) &= \frac{1669880 - 10(45,53)(25,60)}{\sqrt{3022399 - 10(45,53)^2}\sqrt{925800 - 10(25,60)^2}} \\
&= \frac{5043120}{5067155,33} = 0,9953
\end{aligned}
$$

Regressão linear simples

Um Modelo de Regressão linear simples é uma relação funcional entre as duas variáveis quantitativas.

$$X \quad \Rightarrow \text{Variável } \textbf{Independente} \text{ (covariável)}$$
$$Y \quad \Rightarrow \text{Variável } \textbf{Dependente} \text{ (resposta)}$$

Relação de causa e efeito;

Ajuste do modelo a um conjunto de dados: método dos mínimos quadrados (Visto no Cálculo II);

Previsão por interpolação e extrapolação.

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

$$y_i = \alpha + \beta x i_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

Ajuste de uma reta

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \qquad i = 1, 2, \ldots, n$$

ou

$$\hat{y}_i = a + b x_i$$

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

Estimadores (funções) dos parâmetros pelo método dos mínimos quadrados

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

e

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

em que n corresponde ao tamanho da amostra.

RLS: Decomposição da Variação Total

A variação total dos dados pode ser decomposta em duas parcelas, a saber:

Varição Residual ou "Variação não Explicada" $\Rightarrow$ fica quantificada pela Soma de Quadrados Residual:

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

Varição devida à Regressão ou "Variação Explicada" $\Rightarrow$ fica quantificada pela Soma de Quadrados de Regressão:

$$SQReg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Desta forma, tem-se:

$$\text{Variação Total} = \text{Variação Explicada} + \text{Variação Residual}$$

$$SQtotal = SQReg + SQRes$$

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

Coeficiente de Determinação do modelo

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

Exemplo

(Magalhães e Lima, 2002, pág. 332) Em uma dada região de Bocaina-SP, acredita-se que o gado alimentado com determinado pasto tem um ganho de peso maior do que o usual. Estudos de laboratório detectaram uma substância no pasto e deseja-se verificar se ela pode ser utilizada para melhorar o ganho de peso dos bovinos. Foram escolhidos 15 animais de mesma raça e idade e cada animal recebeu uma determinada concentração X (em mg/l). O ganho de peso, denotado por Y foi anotado e os resultados são os seguintes (em kg).

Tabela 6: Dados sobre concentração de uma substância (X) e ganho de peso (Y)

X	0,2	0,5	0,6	0,7	1,0	1,5	2,0	2,5
Y	9,4	11,4	12,3	10,2	11,9	13,6	14,2	16,2
X	3,0	3,5	4,0	4,5	5,0	5,5	6,0	
Y	16,2	17,7	18,8	19,9	22,5	24,7	23,1	

Fonte: Magalhães e Lima, 2002

Tabela 7: Cálculos auxiliares para correlação e regressão

Unidade	x	y	x^2	y^2	xy
1	0,2	9,4	0,04	88,36	1,88
2	0,5	11,4	0,25	129,96	5,70
3	0,6	12,3	0,36	151,29	7,38
4	0,7	10,2	0,49	104,04	7,14
5	1,0	11,9	1,00	141,61	11,90
6	1,5	13,6	2,25	184,96	20,40
7	2,0	14,2	4,00	201,64	28,40
8	2,5	16,2	6,25	262,44	40,50
9	3,0	16,2	9,00	262,44	48,60
10	3,5	17,7	12,25	313,29	61,95
11	4,0	18,8	16,00	353,44	75,20
12	4,5	19,9	20,25	396,01	89,55
13	5,0	22,5	25,00	506,25	112,50
14	5,5	24,7	30,25	610,09	135,85
15	6,0	23,1	36,00	533,61	138,60
Total	**40,50**	**242,10**	**163,39**	**4239,43**	**785,55**

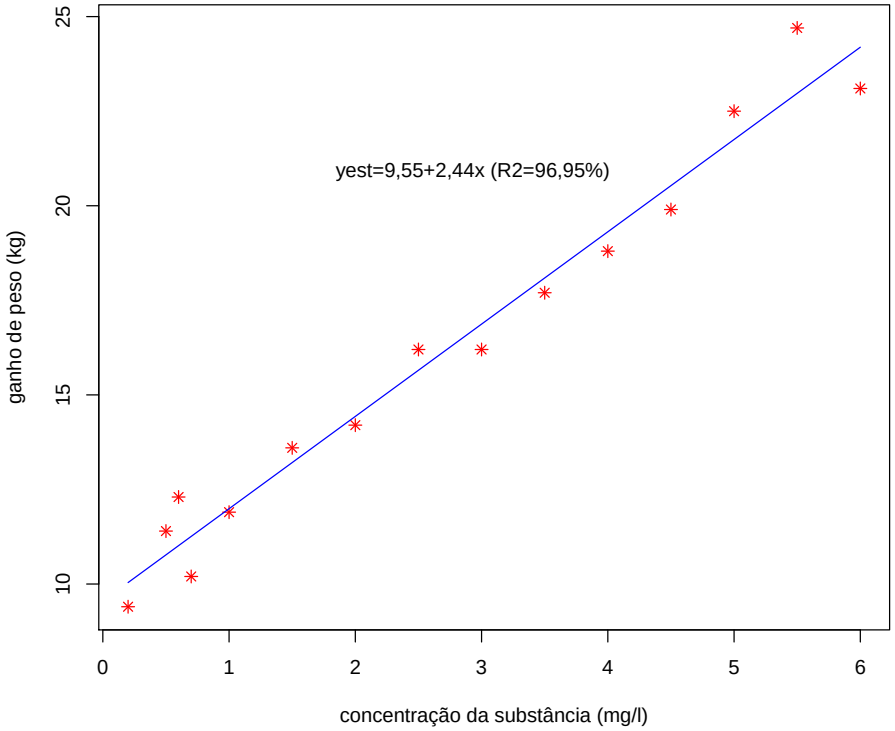

Figura 6: Modelo de regressão linear simples ajustado aos dados sobre ganho de peso e concentração de uma substância na alimentação.

REFERÊNCIAS BIBLIOGRÁFICAS

Andrade, D. F.; Ogliari, P.J. **Estatística para as ciências agrárias e biológicas**. Ed. UFSC, 2010.

Bussab, W. O; Morettin, P.A. **Estatística Básica**. São Paulo, Saraiva, 5 ed. 2002.

Magalhães, M. N; Lima, A. C. P. **Noções de Probabilidade e Estatística**, Edusp, 2002.