7. Análise de correlação e regressão

USP-ICMC-SME

2013

Introdução

Estudo da relação linear entre duas variáveis quantitativas.
Exemplos:

- Tempo de prática de esportes e ritmo cardíaco.
- Número de usuários e tempo de resposta de um sistema.
- Número de vendedores e resultados de vendas.
- Tempo de estudo e nota em uma prova.

Pontos de vista:

- Quantificando a força e o sinal dessa relação: **correlação**.
- Explicitando a forma dessa relação: **regressão**.

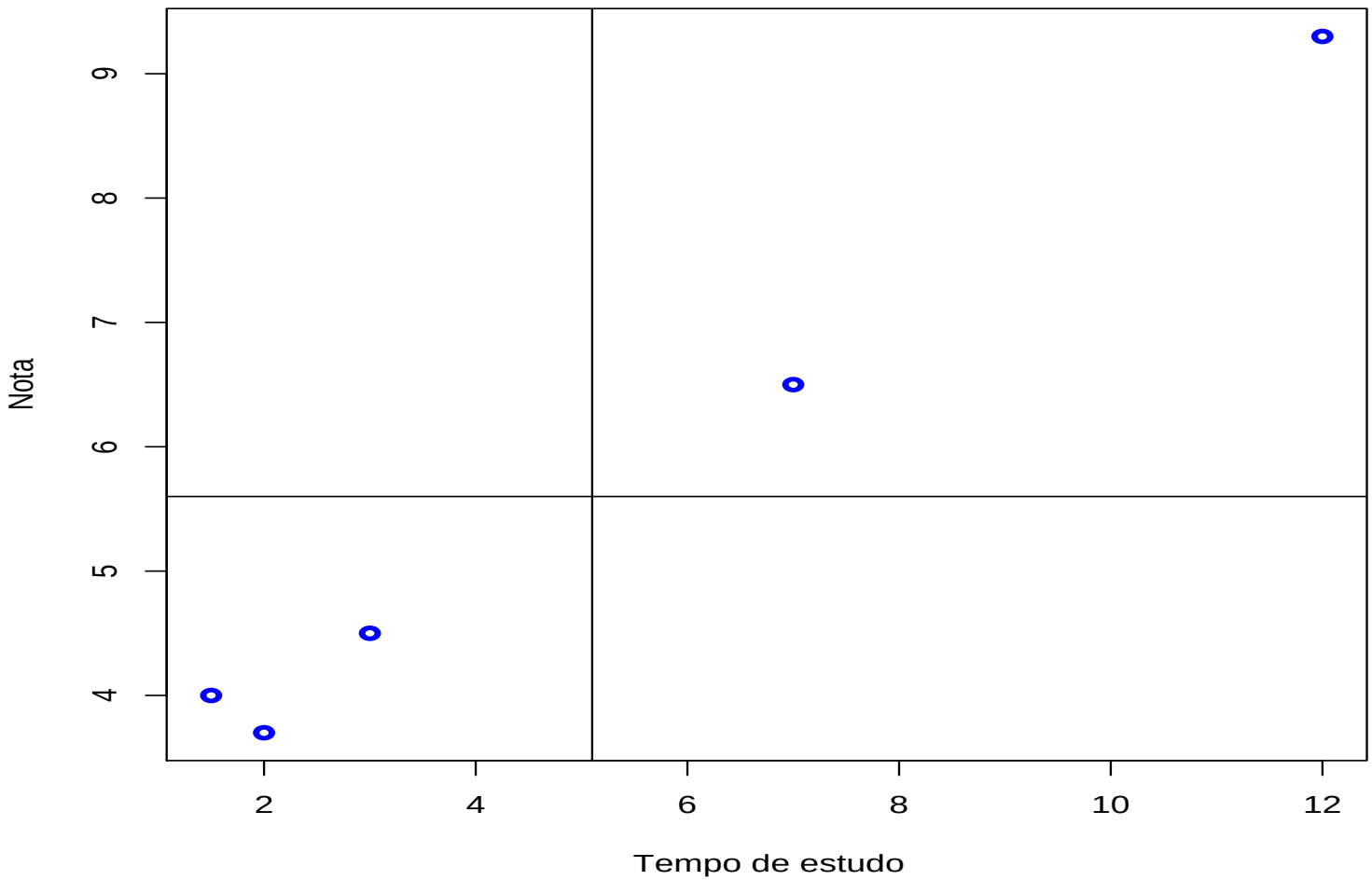
Exemplo

- X: *tempo de estudo (em horas) e*
- Y *nota em uma prova.*

Pares de observações (X_i, Y_i)*:*

Tempo	Nota
3,0	4,5
7,0	6,5
2,0	3,7
1,5	4,0
12,0	9,3

Diagrama de dispersão



Coeficiente de correlação

O coeficiente de correlação linear (de Pearson) é definido como

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

em que $\bar{X}$ e $\bar{Y}$ são as médias amostrais de X e Y.

Notações

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

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

e

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y})y_i = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

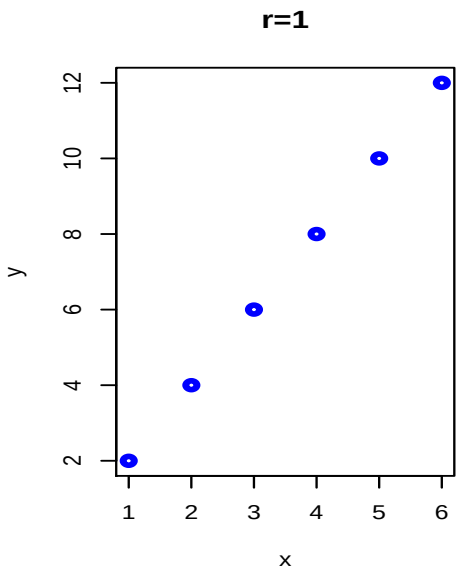
Propriedades

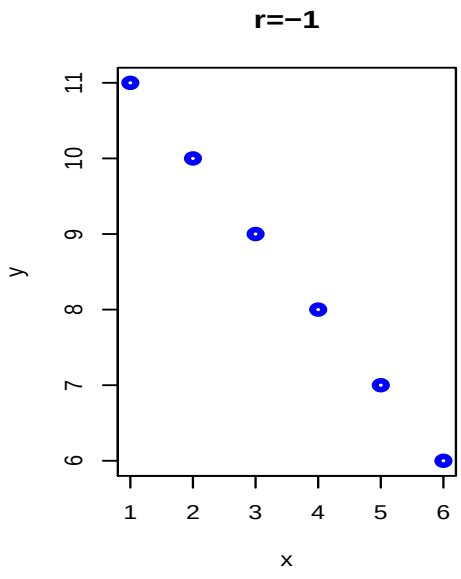
$$-1 \leq r \leq 1$$

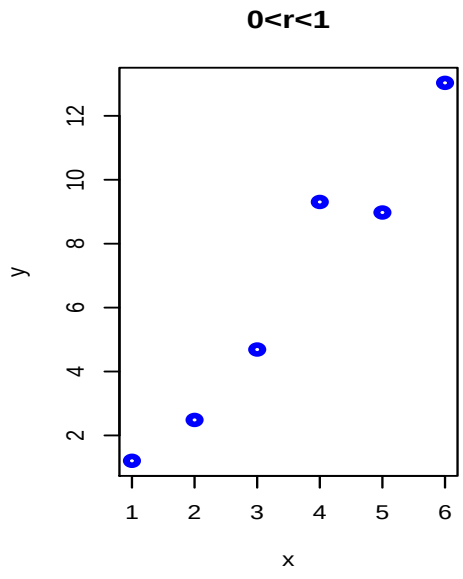
Classificação:

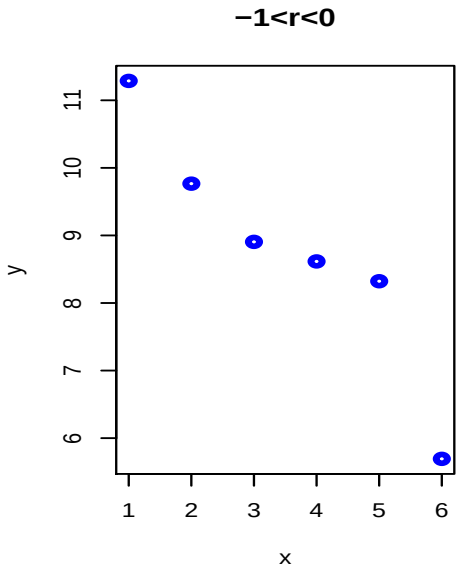
- $r = 1$: correlação linear positiva e perfeita.
- $r = -1$: correlação linear negativa e perfeita.
- $r = 0$: inexistência de correlação linear.

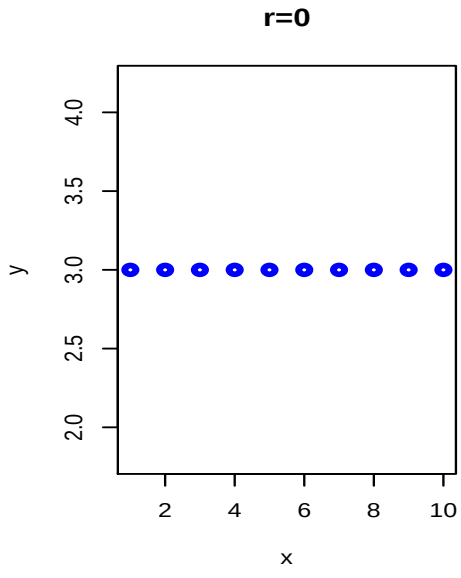
Exemplos

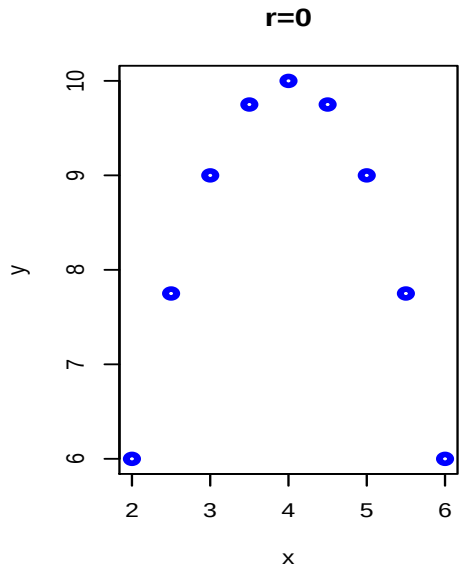

r=1
y
x
2
4
6
8
10
12
1
2
3
4
5
6


r=−1
y
x
6
7
8
9
10
11
1
2
3
4
5
6


0<r<1
y
x
2
4
6
8
10
12
1
2
3
4
5
6


−1<r<0
y
x
6
7
8
9
10
11
1
2
3
4
5
6


r=0
y
x
2.0
2.5
3.0
3.5
4.0
2
4
6
8
10


r=0
y
x
6
7
8
9
10
2
3
4
5
6

Exemplo 2

Variáveis:

- Y - consumo de bebidas em um dia (em 100 litros) e
- X - temperatura máxima (em ^{0}C).

Estas variáveis foram observadas em nove localidades com as mesmas características demográficas e sócio-econômicas.
Os dados amostrais foram os seguintes:

X	16	31	38	39	37	36	36	22	10
Y	290	374	393	425	406	370	365	320	269

Determinar a correlação de X e Y.

Exemplo 2

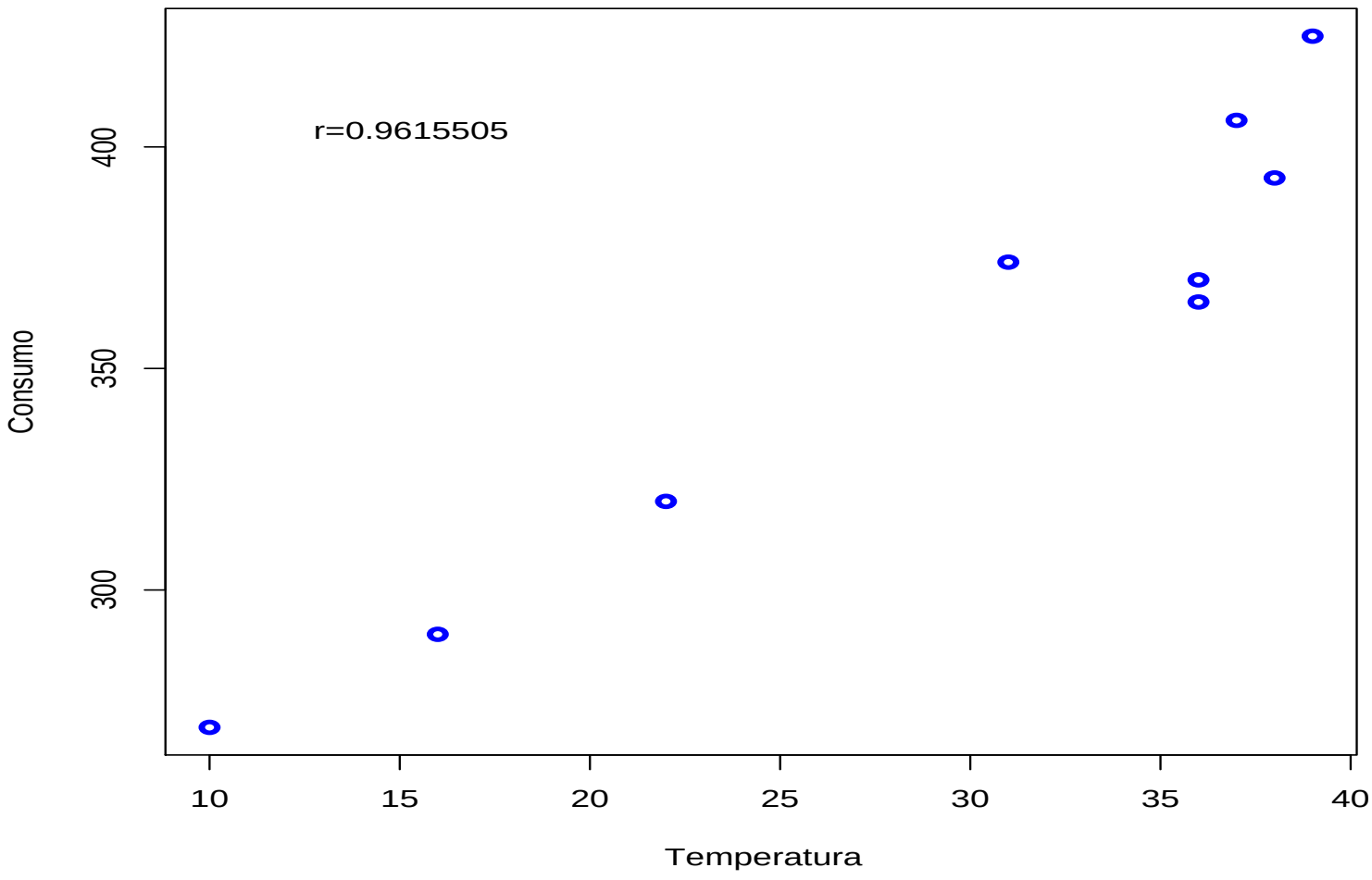
	y_i	x_i	$x_i y_i$	y_i^2	x_i^2
	290	16	4640	84100	256
	374	31	11594	139876	961
	393	38	14934	154449	1444
	425	39	16575	180625	1521
	406	37	15022	164836	1369
	370	36	13320	136900	1296
	365	36	13140	133225	1296
	320	22	7040	102400	484
	269	10	2690	72361	100
Total	3212	265	98955	1168772	8727

Exemplo 2

$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 98955 - (265)(3212)/9 = 4379,444$
$S_{yy} = \sum y_i^2 - (\sum y_i)^2/n = 1168772 - (3212)/9 = 22444,89,$
$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 8727 - (265)/9 = 924,222$ e

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{4379,444}{\sqrt{22444,89 \times 924,222}} = 0,9615.$$

Diagrama de dispersão



Cálculo de r com o pacote R

$x = c(16, 31, 38, 39, 37, 36, 36, 22, 10)$
$y = c(290, 374, 393, 425, 406, 370, 365, 320, 269)$
Diagrama de dispersão
$> plot(x, y, xlab = "Temperatura", ylab = "Consumo", col = "blue", lwd = 3)$
Calcula a correlação:
$> cor(x, y)$

Análise de regressão

Exemplo

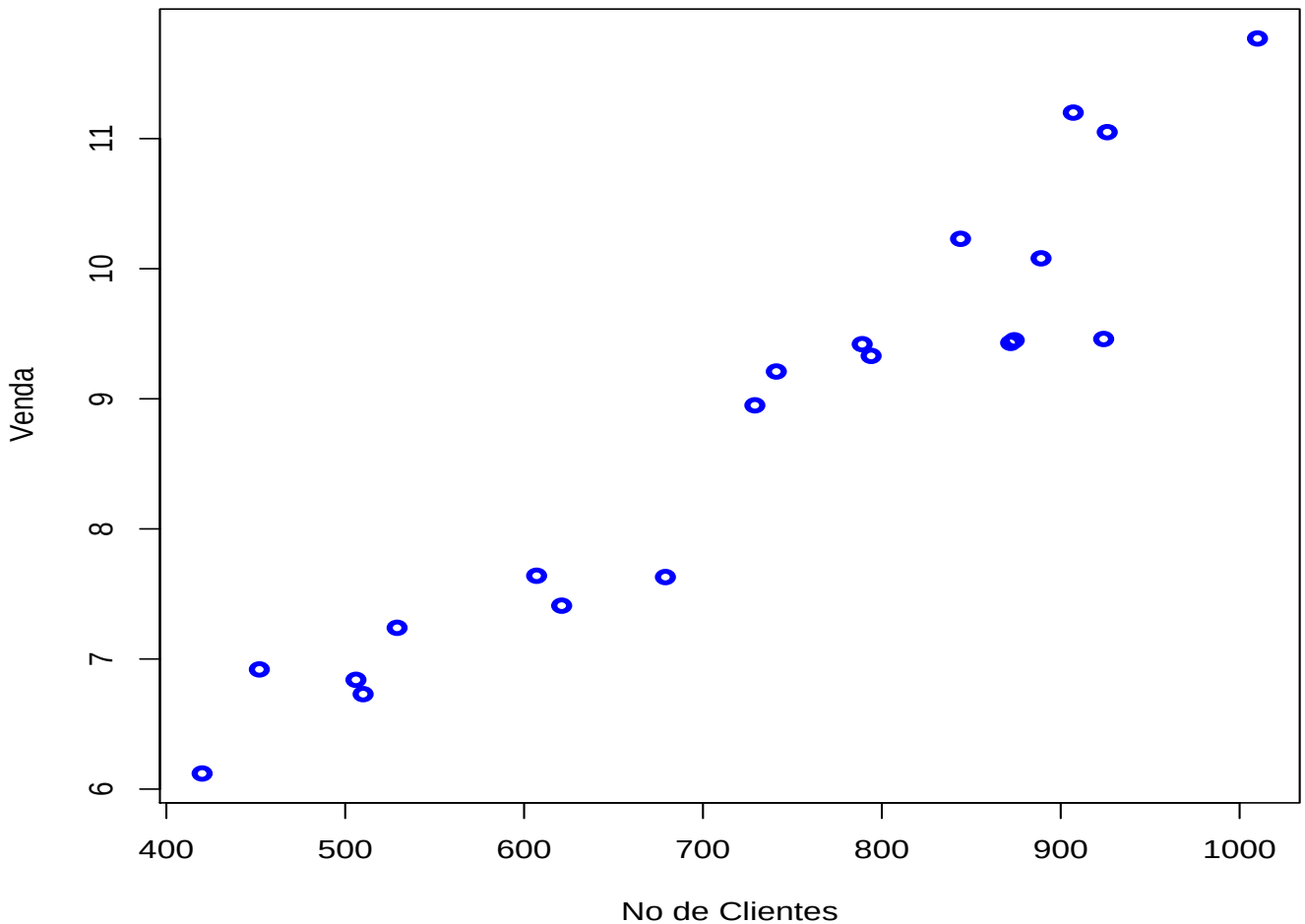
O gerente de uma grande cadeia de lojas deseja desenvolver um modelo com a finalidade de estimar as vendas médias semanais (em milhares de reais).

- *Y : vendas semanais e*
- *X : número de clientes.*

Estas variáveis foram observadas em 20 lojas escolhidas aleatoriamente.

X	*907*	*926*	*506*	*741*	*789*	*889*	*874*	*510*	*529*	*420*
Y	*11,20*	*11,05*	*6,84*	*9,21*	*9,42*	*10,08*	*9,45*	*6,73*	*7,24*	*6,12*
X	*679*	*872*	*924*	*607*	*452*	*729*	*794*	*844*	*1010*	*621*
Y	*7,63*	*9,43*	*9,46*	*7,64*	*6,92*	*8,95*	*9,33*	*10,23*	*11,77*	*7,41*

Diagrama de dispersão



Análise de regressão

Em muitos problemas a média da variável aleatória Y está relacionada com X pela relação

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x,$$

em que β_0 e β_1 são, respectivamente, o intercepto e a inclinação da reta e recebem o nome de coeficientes de regressão.
O valor de Y será determinado pelo valor médio da função linear ($\mu_{Y|x}$) mais um termo que representa um erro aleatório.

$$Y = \mu_{Y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon,$$

em que ε é o erro aleatório não observável.

Análise de regressão

Em geral, a variável resposta pode estar relacionada com k variáveis explicativas $X_1, \ldots X_k$ obedecendo à equação

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

A equação é denominada modelo de regressão linear múltipla.

Análise de regressão

O adjetivo "linear" é usado para indicar que o modelo é linear nos parâmetros $\beta_1, \ldots, \beta_k$ e não porque Y é função linear dos X's.
Uma expressão da forma

$$Y = \beta_0 + \beta_1 \log X_1 + \beta_2 X_2^3 + \varepsilon$$

é um modelo de regressão linear múltipla.
A equação

$$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \beta_3 X_2^2 + \varepsilon$$

representa um modelo de regressão não linear.

Análise de regressão linear simples

Um modelo de regressão linear simples (MRLS) descreve uma relação entre uma variável independente (explicativa ou regressora) X e uma variável dependente (resposta) Y:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1}$$

em que β_0 e β_1 são constantes (parâmetros) desconhecidas e ε é o erro aleatório.

Suposições do MRLS

(i) $E(\varepsilon) = 0$ $Var(\varepsilon) = \sigma^2$ (desconhecida).

(ii) Os erros são não correlacionados.

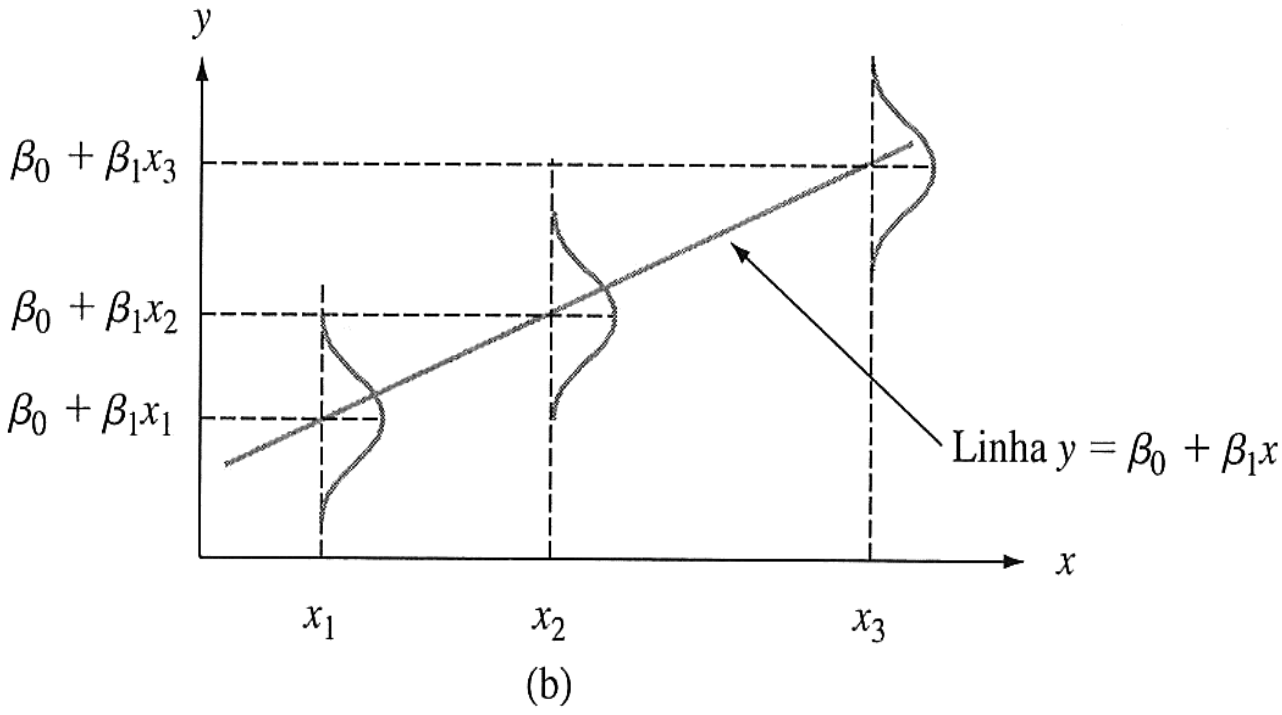
(iii) A variável explicativa X é controlada pelo experimentador.

(iv) $\varepsilon \sim N(0, \sigma^2)$.

Se (i)-(iv) se verificarem, então a variável dependente Y é uma v.a. com distribuição normal com variância σ^2 e média $\mu_{Y|x}$, sendo

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x.$$

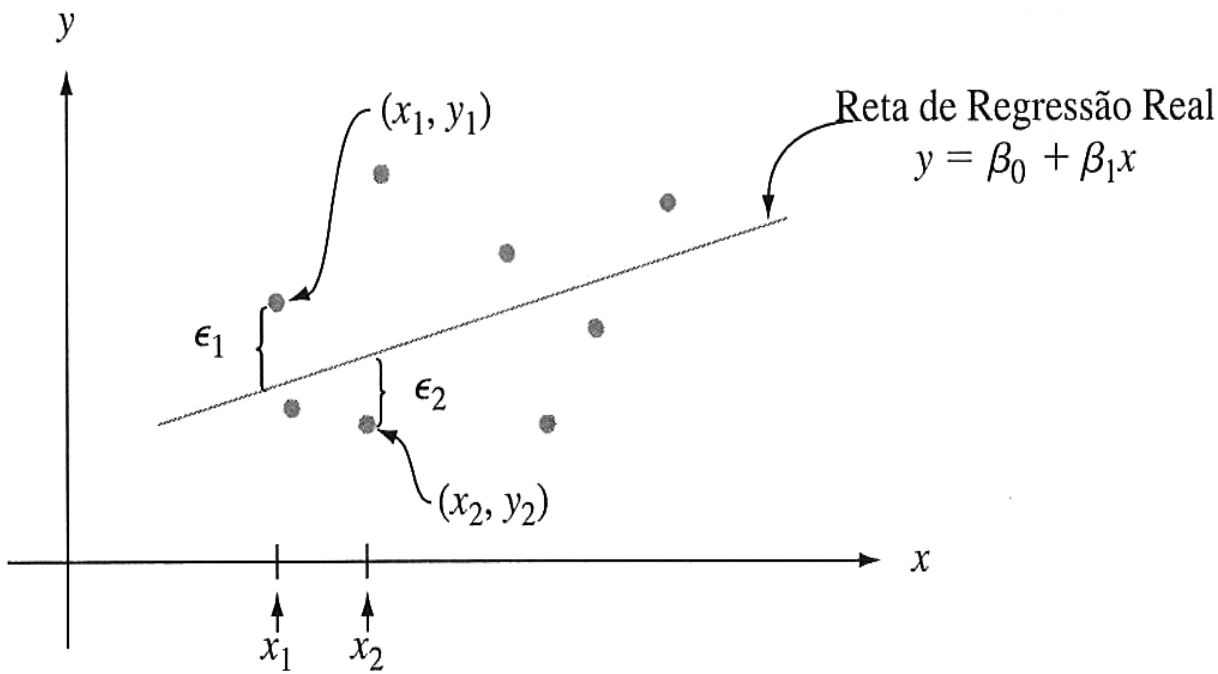
Suposições do MRLS



(b)

Estimação pelo método de mínimos quadrados (MQ)

Foram coletados n pares de observações $(x_1, y_1), \ldots, (x_n, y_n)$. A figura mostra uma representação gráfica dos dados observados e a linha de regressão.



Estimação pelo método de mínimos quadrados(MQ)

Ao utilizar o modelo (1), é possível expressar as n observações da amostra como

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

E a soma de quadrados dos desvios das observações em relação à linha de regressão é

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Estimação

Os estimadores de mínimos quadrados (EMQ) de β_0 e β_1, denotados por $\widehat{\beta}_0$ e $\widehat{\beta}_1$, devem satisfazer às equações

$$\frac{\partial Q}{\partial \beta_0} \mid_{\widehat{\beta}_0, \widehat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

e

$$\frac{\partial Q}{\partial \beta_1} \mid_{\widehat{\beta}_0, \widehat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) x_i = 0.$$

Estimação

Após simplificar as expressões anteriores obtemos

$$\begin{aligned} \widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \qquad (3) \\ \text{e} \quad \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned}$$

As equações (3) recebem o nome de equações normais de mínimos quadrados.

Estimação

A solução dessas equações fornece os EMQ, $\widehat{\beta}_0$ e $\widehat{\beta}_1$, dados por

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

e

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}.$$

Estimação

Portanto, a linha de regressão estimada ou ajustada é

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

e estima a média da variável dependente para um valor da variável explicativa $X = x$ $(\mu_{Y|x})$.
Cada par de observações é tal que

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i, \;\; i = 1, \ldots, n,$$

em que $e_i = y_i - \widehat{y}_i$ recebe o nome de **resíduo**.

Exemplo de aplicação

Considerando os dados do exemplo,

$$
\begin{aligned}
n &= 20, \\
\sum_{i=1}^{n} x_i &= 907 + 926 + \cdots + 621 = 14.623; \quad \bar{x} = 731,15, \\
\sum_{i=1}^{n} y_i &= 11,20 + 11,05 + \cdots + 7,41 = 176,11; \quad \bar{y} = 8,8055, \\
\sum_{i=1}^{n} x_i^2 &= (907)^2 + (926)^2 + \cdots + (621)^2 = 11.306.209, \\
\sum_{i=1}^{n} y_i^2 &= (11,20)^2 + (11,05)^2 + \cdots + (7,41)^2 = 1.602,0971 \\
&\text{e} \\
\sum_{i=1}^{n} x_i y_i &= 907 \times 11,20 + 11,05 \times 926 + \cdots + 7,41 \times 621 = 134.127,90.
\end{aligned}
$$

Exemplo de aplicação

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 11.306.209 - 20(731,15)^2 = 614.603, \\
S_{xy} &= \sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y}) = 134.127,90 - 20(8,8055)(731,15) = 5.365,08 \\
&\text{e} \\
S_{yy} &= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 1.609,0971 - 20(8,8055) = 51,3605.
\end{aligned}
$$

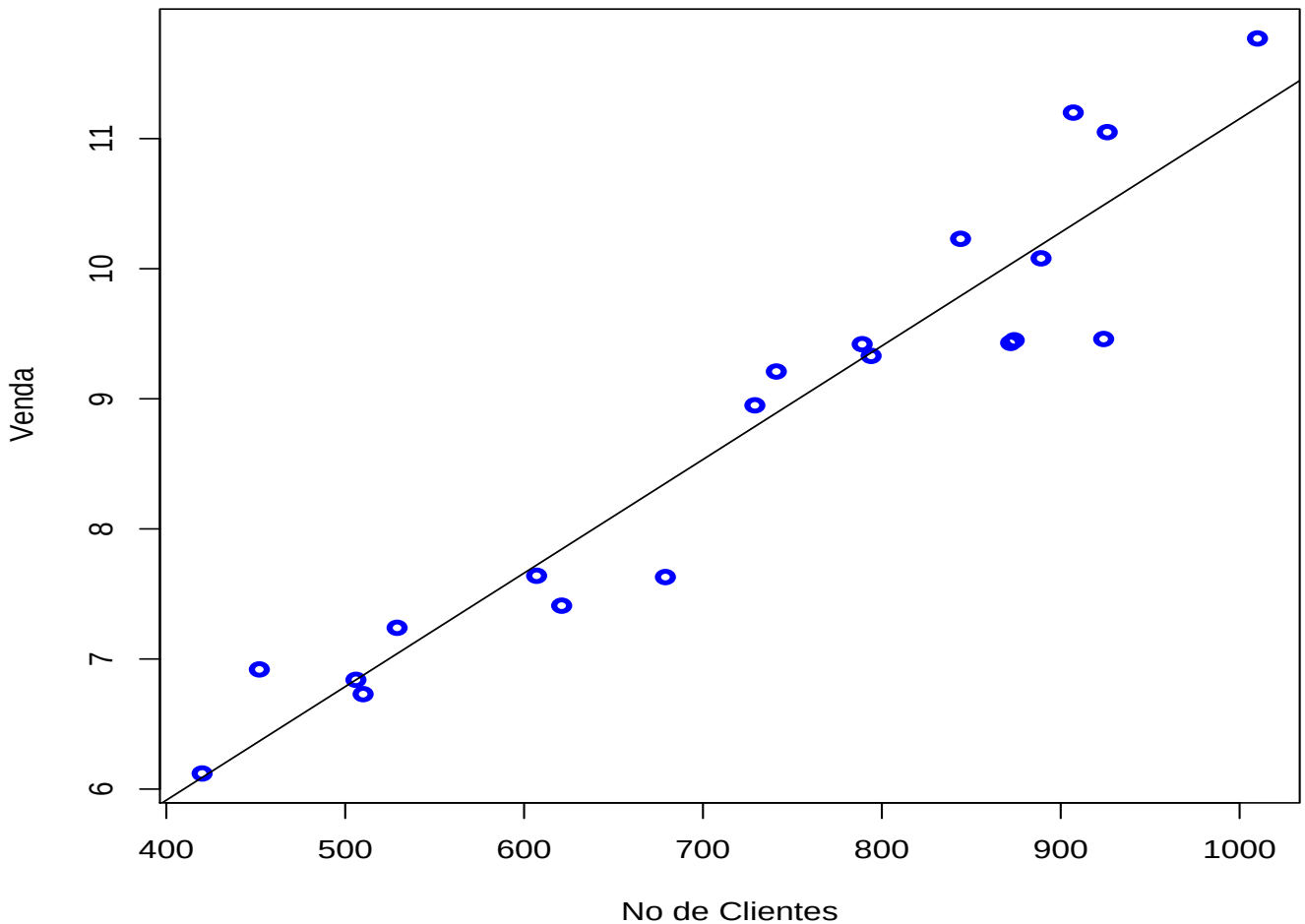
As estimativas dos parâmetros do MRLS são:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5.365,08}{614.603} = 0,00873 \quad \text{e} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 8,8055 - (0,00873)(731,15) = 2,423.$$

Portanto, a linha de regressão ajustada ou estimada para esses dados é

$$\widehat{y} = 2,423 + 0,00873x.$$

Diagrama de dispersão



Suponha que temos interesse em prever as vendas semanais para um supermercado com 600 clientes.
No modelo de regressão ajustado basta substituir $X = 600$, isto é,

$$\widehat{y} = 2,423 + 0,00873 \times 600 = 7,661.$$

A venda semanal de 7,661 mil reais pode ser interpretada com uma estimativa da venda média semanal verdadeira dos supermercados com $X = 600$ clientes, ou como uma estimativa de uma futura venda de um supermercado quando o número de clientes for $X = 600$.

Propriedades dos EMQ

Se as suposições do MRLS sejam válidas é possível demonstrar que

- $E(\widehat{\beta}_1) = \beta_1, \quad Var(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.
- $E(\widehat{\beta}_0) = \beta_0, \quad Var(\widehat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$.
- $Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$.
- $\widehat{\beta}_j \sim N(\beta_j, Var(\widehat{\beta}_j)), \quad j = 0, 1$.

Estimação de σ^2

Os resíduos

$$e_i = y_i - \widehat{y}_i$$

são empregados na estimação de σ^2. A soma de quadrados residuais, denotada por SQR, é

$$SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Pode-se demonstrar que o valor esperado da soma de quadrados dos residuais é dado por

$$E(SQR) = (n-2)\sigma^2.$$

Portanto,

$$\widehat{\sigma^2} = \frac{SQR}{n-2} = QMR \quad (\textbf{Quadrado médio residual})$$

é um estimador não viesado de σ^2.
Uma expressão mais conveniente para o cálculo da SQR é dada por

$$SQR = S_{yy} - \widehat{\beta}_1 S_{xy}.$$

Exemplo

Com os dados do exemplo obtemos a estimativa da variância σ^2. Nesse caso, $S_{yy} = 51,3605$, $S_{xy} = 5.365,08$ e $\widehat{\beta}_1 = 0,00873$. Portanto, a estimativa de σ^2 para o exemplo é

$$\begin{aligned} \widehat{\sigma^2} &= \frac{SQR}{n-2} = \frac{S_{yy} - \widehat{\beta}_1 S_{xy}}{n-2} \\ &= \frac{51,3605 - 0,00873 \times 5.365,08}{20-2} = 0,2513. \end{aligned}$$

Teste de hipóteses sobre β_1

Suponha que se deseje testar a hipótese de que a inclinação é igual a uma constante representada por $\beta_{1,0}$. As hipóteses apropriadas são

$$H_0 : \beta_1 = \beta_{1,0} \quad \text{contra} \quad H_1 : \beta_1 \neq \beta_{1,0}.$$

A estatística

$$T = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{\widehat{\sigma}^2 / S_{xx}}}$$

tem distribuição t de Student com $n - 2$ graus de liberdade sob $H_0 : \beta_1 = \beta_{1,0}$. Rejeita-se H_0 se

$$|T_{obs}| > t_{\alpha/2,\, n-2}.$$

Teste de hipóteses sobre β_0

$$H_0 : \beta_0 = \beta_{0,0} \quad \text{contra} \quad H_1 : \beta_0 \neq \beta_{0,0}$$

A estatística

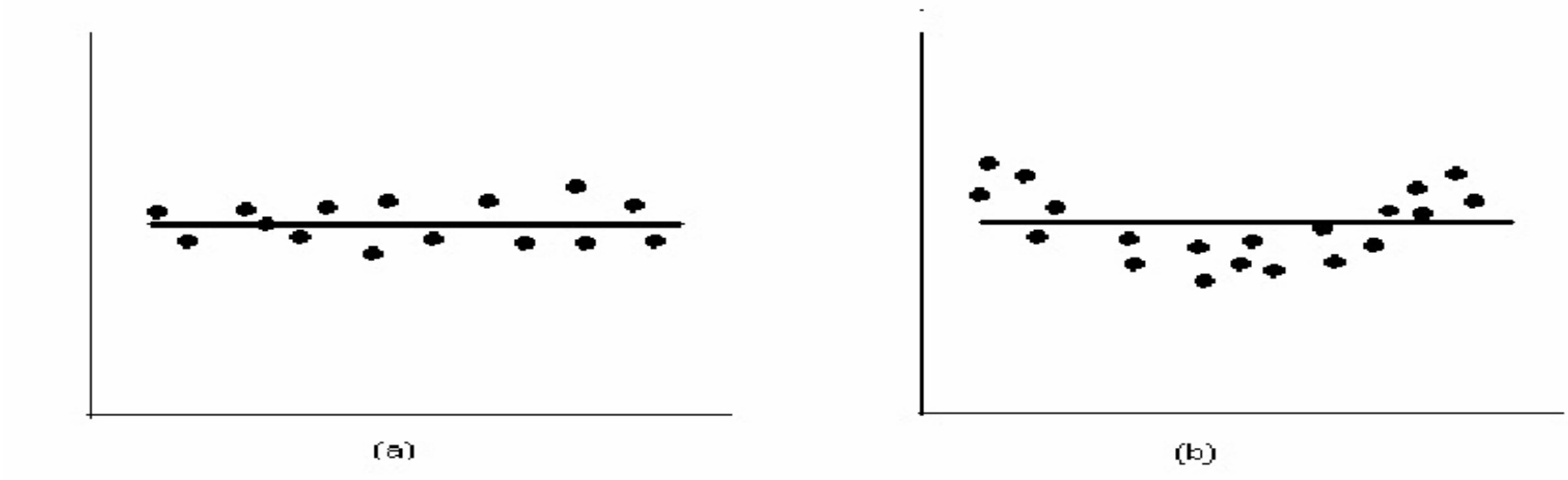
$$T = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}}$$

tem distribuição t de Student com $n - 2$ graus de liberdade.
Rejeitamos a hipóteses nula se $|T_{obs}| > t_{\alpha/2,\, n-2}$.

Teste de significância do MRLS

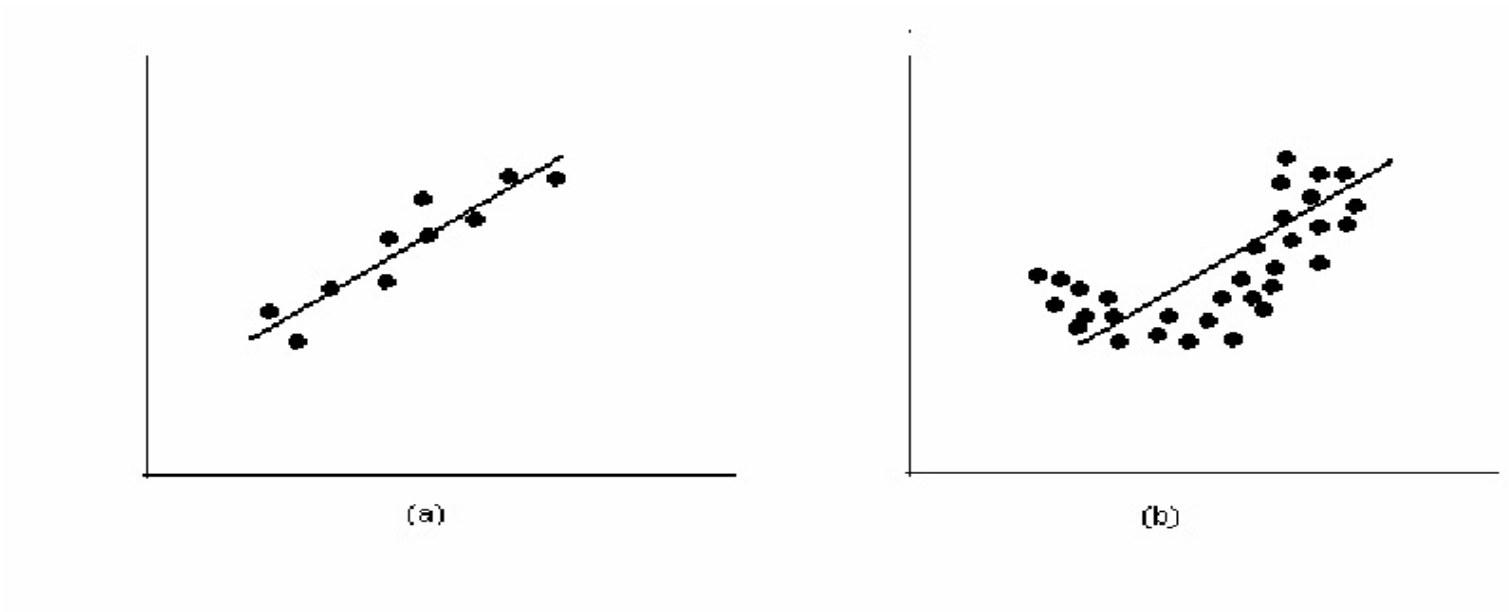
$$H_0 : \beta_1 = 0 \quad \text{contra} \quad H_1 : \beta_1 \neq 0.$$

Deixar de rejeitar $H_0 : \beta_1 = 0$ é equivalente a concluir que não há relação linear entre X e Y.



Teste de significância do MRLS

Se $H_0 : \beta_1 = 0$ é rejeitada, implica que X tem importância ao explicar a variabilidade de Y



Exemplo

Teste de significância para o MRLS para os dados do exemplo 1, com $\alpha = 0,05$.
As hipóteses são $H_0 : \beta_1 = 0 \quad$ contra $\quad H_1 : \beta_1 \neq 0$.
Do exemplo tem-se

$$n = 20, \quad S_{xx} = 614,603, \quad \widehat{\beta}_1 = 0,00873 \quad \text{e} \quad \widehat{\sigma}^2 = 0,2512,$$

de modo que a estatística de teste é

$$T_{obs} = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2/S_{xx}}} = \frac{0,00873}{\sqrt{0,2513/614.603}} = 13,65.$$

Como $T_{obs} = 13,65 > t_{0,025;18} = 2,101$, rejeita-se a hipótese $H_0 : \beta_1 = 0$.

Análise de variância

O método consiste em decompor a variabilidade da variável resposta em relação à sua média ($\bar{Y}$). Considere a identidade

$$Y_i - \bar{Y} = Y_i - \widehat{Y}_i + \widehat{Y}_i - \bar{Y}.$$

Elevando ao quadrado a igualdade e somando as n observações tem-se

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2.$$

Análise de variância

$SQR = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$:*soma de quadrados dos resíduos* e
$SQreg = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2$:*soma de quadrados da regressão*. Logo,

$$S_{YY} = SQreg + SQR,$$

em que $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ é a soma de quadrados total de Y, representada por SQT.

Análise de variância

Se a hipótese nula $H_0 : \beta_1 = 0$ é verdadeira, a estatística

$$F = \frac{SQreg/1}{SQR/(n-2)} = \frac{QMreg}{QMR} \sim F(1, n-2).$$

Portanto, rejeita-se H_0 se $F_{obs} > F_{\alpha,\, 1,\, n-2}$. Equivale a rejeitar H_0 se $\sqrt{F_{obs}} > t_{\alpha/2, n-2}$.

$QMreg = \frac{SQreg}{1}$: quadrado médio devido à regressão e

$QMR = \frac{SQR}{n-2}$: quadrado médio residual.

Tabela de ANOVA

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrados médios	F
Regressão	$SQreg$	1	$QMreg$	$\frac{QMreg}{QMR}$
Residual	SQR	$n-2$	QMR	
Total	SQT	$n-1$		

Exemplo

Exemplo Relação linear entre o número de clientes (X) e as vendas semanais (Y). Relembre que $S_{yy} = 51,3605$, $\widehat{\beta}_1 = 0,00873$, $S_{xy} = 5.365,08$ *e* $n = 20$.

A soma de quadrados da regressão é

$$SQreg = \widehat{\beta}_1 S_{xy} = 0,00873 \times 5.365,08 = 46,8371,$$

enquanto a soma de quadrados dos resíduos é

$$SQR = SQT - \widehat{\beta}_1 S_{xy} = 51,3605 - 46,8371 = 4,5234.$$

A estatística de teste é

$$F_{obs} = QMreg/QMR = 46,837148/0,2512 = 186,4536.$$

Como $\sqrt{F_{obs}} = T_{obs} = 13,65 > t_{0,025;18} = 2,101$, rejeita-se H_0 ao nível de significância de 5%.

Tabela de ANOVA

Fonte de variação	Soma de quadrados	Graus de qiberdade	Quadrados médios	F
Regressão	$46,8371$	1	$46,8371$	186,45
Residual	$4,5234$	18	$0,2513$	
Total	$51,3605$	19		

Análise de regressão com o R

$X = c(907, 926, 506, 741, 789, 889, 874, 510, 529, 420, 679, 872, 924, 607, 452, 729, 794, 844, 1010, 621)$
$Y =$
$c(11.20, 11.05, 6.84, 9.21, 9.42, 10.08, 9.45, 6.73, 7.24, 6.12, 7.63, 9.43, 9.46, 7.64, 6.92, 8.95, 9.33, 10.23, 11.77, 7.41)$

$> fit = lm(Y \sim X)$
>summary(fit)
Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	2.4230444	0.4809646	5.038	8.55e-05 ***
X	0.0087293	0.0006397	13.646	6.21e-11 ***

Residual standard error: 0.5015 on 18 degrees of freedom Multiple
R-Squared: 0.9119, Adjusted R-squared: 0.907 F-statistic:

186.45 on 1 and 18 DF, p-value: $6.206e - 11$

Análise de regressão com o R

```
> anova(fit)
Analysis of Variance Table
Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X           1  46.834  46.834  186.45 6.206e-11 ***
Residuals  18   4.527   0.251
```

Intervalo de confiança para β_0 e β_1

Se é válida a suposição de que $\varepsilon_i \sim NID(0, \sigma^2)$, então

$$(\widehat{\beta}_1 - \beta_1)/\sqrt{QMR/S_{xx}} \quad \text{e} \quad (\widehat{\beta}_0 - \beta_0)/\sqrt{QMR\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

são variáveis aleatórias com distribuição t de Student com $n-2$ graus de liberdade.
Intervalo de confiança de $100(1-\alpha)\%$ para β_1 :

$$IC(\beta_1; 1-\alpha) = \left(\widehat{\beta}_1 - t_{\frac{\alpha}{2},\, n-2}\sqrt{\frac{QMR}{S_{xx}}}\ ;\ \widehat{\beta}_1 + t_{\frac{\alpha}{2},\, n-2}\sqrt{\frac{QMR}{S_{xx}}}\right).$$

Intervalo de confiança para β_0

De modo similar, um intervalo de $100(1-\alpha)\%$ de confiança para β_0 é dado por

$$IC(\beta_0; 1-\alpha) = \left(\widehat{\beta}_0 - t_{\frac{\alpha}{2},\, n-2} \sqrt{QMR \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \; ; \right.$$
$$\left. \widehat{\beta}_0 + t_{\frac{\alpha}{2},\, n-2} \sqrt{QMR \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \right).$$

Intervalo de 95% de confiança para a inclinação com os dados do exemplo 1.
Relembre que $n = 20$, $\widehat{\beta}_1 = 0,00873$, $S_{xx} = 614,603$ e $QMR = 0,2513$. Para $1 - \alpha = 0,95$, tem-se $t_{0,025;18} = 2,101$.

$$IC(\beta_1; 0,95) = (\widehat{\beta}_1 - E \ ; \ \widehat{\beta}_1 + E),$$

em que $E = t_{0,025,18}\sqrt{\frac{QMR}{S_{xx}}} = 2,101\sqrt{\frac{0,2513}{614.603}} = 0,00134$

$$\begin{aligned} IC(\beta_1; 0,95) &= (0,00873 - 0,00134; 0,00873 + 0,00134) \\ &= (0,00739; 0,01007). \end{aligned}$$

Intervalo de confiança para a resposta média

O interesse consiste em obter um intervalo de confiança para

$$E(Y|X = x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0.$$

Um estimador pontual de $\mu_{Y|x_0}$ é

$$\widehat{\mu}_{Y|x_o} = \widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Se $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, pode-se demonstrar que

$$T = \frac{\widehat{\mu}_{Y|x_o} - \mu_{Y|x_o}}{\sqrt{QMR\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t(n-2).$$

Intervalo de confiança para a resposta média

$$IC(\widehat{\mu}_{Y|x}; 1-\alpha) = \left(\widehat{\mu}_{Y|x_o} - E; \widehat{\mu}_{Y|x_o} + E\right),$$

em que $E = t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}]}$

Exemplo. Suponha que há interesse em construir um intervalo de 95% de confiança para a venda média semanal de supermercados com 600 clientes.

No modelo ajustado, $\widehat{\mu}_{Y|x_0} = 2,423 + 0,00873x_0$. Para $x_0 = 600$, obtém-se

$$\widehat{\mu}_{Y|x_0} = 7,661.$$

Intervalo de confiança para a resposta média

Também, $\bar{x} = 731,15$, $QMR = 0,2513$, $S_{xx} = 614.603$, $n = 20$, $1 - \alpha = 0,95$ e $t_{0,025;18} = 2,101$.

$E = 2,101\sqrt{0,2513[\frac{1}{20} + \frac{(600-731,15)^2}{614.603}]} = 0,292$, de forma que

$$\begin{aligned} IC(\mu_{Y|x_0}; 0,95) &= (7,661 - 0,292; 7,661 + 0,292) \\ &= (7,369; 7,935). \end{aligned}$$

Previsão de novas observações

Uma aplicação frequente de um modelo de regressão é a previsão de uma nova ou futura observação de Y (Y_0) correspondente a um dado valor da variável explicativa X (x_0). Temos que

$$\widehat{Y_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

é um preditor de Y_0.
Um intervalo de $100(1-\alpha)\%$ de previsão para uma futura observação é dado por

$$IC(Y_0; 1-\alpha) = (\widehat{Y} - E; \widehat{Y} + E),$$

em que $E = t_{\frac{\alpha}{2}, n-2} \sqrt{QMR \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$.

Exemplo

Agora temos interesse em encontrar um intervalo de previsão de 95% das vendas semanais de um supermercado com 600 clientes. Considerando os dados do exemplo 1, $\widehat{Y_0} = 7,661$, $E = 2,101\sqrt{0,2513\left[1+\frac{1}{20}+\frac{(600-731,15)^2}{614.603}\right]} = 1,084$ e o intervalo de predição é

$$\begin{aligned} IC(Y_0; 0,95) &= (7,661 - 1,084; 7,661 + 1,084) \\ &= (6,577; 8,745). \end{aligned}$$

Adequação do modelo de regressão

- Análise dos resíduos.
- Coeficiente de determinação.

Os resíduos de um modelo de regressão são definidos como

$$e_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n,$$

em que y_i é uma observação de Y e $\widehat{y}_i$ é o valor correspondente predito através do modelo de regressão.
Resíduos padronizados:

$$d_i = \frac{e_i}{\sqrt{QMR}}, \quad i = 1, \ldots, n.$$

Gráfico de resíduos

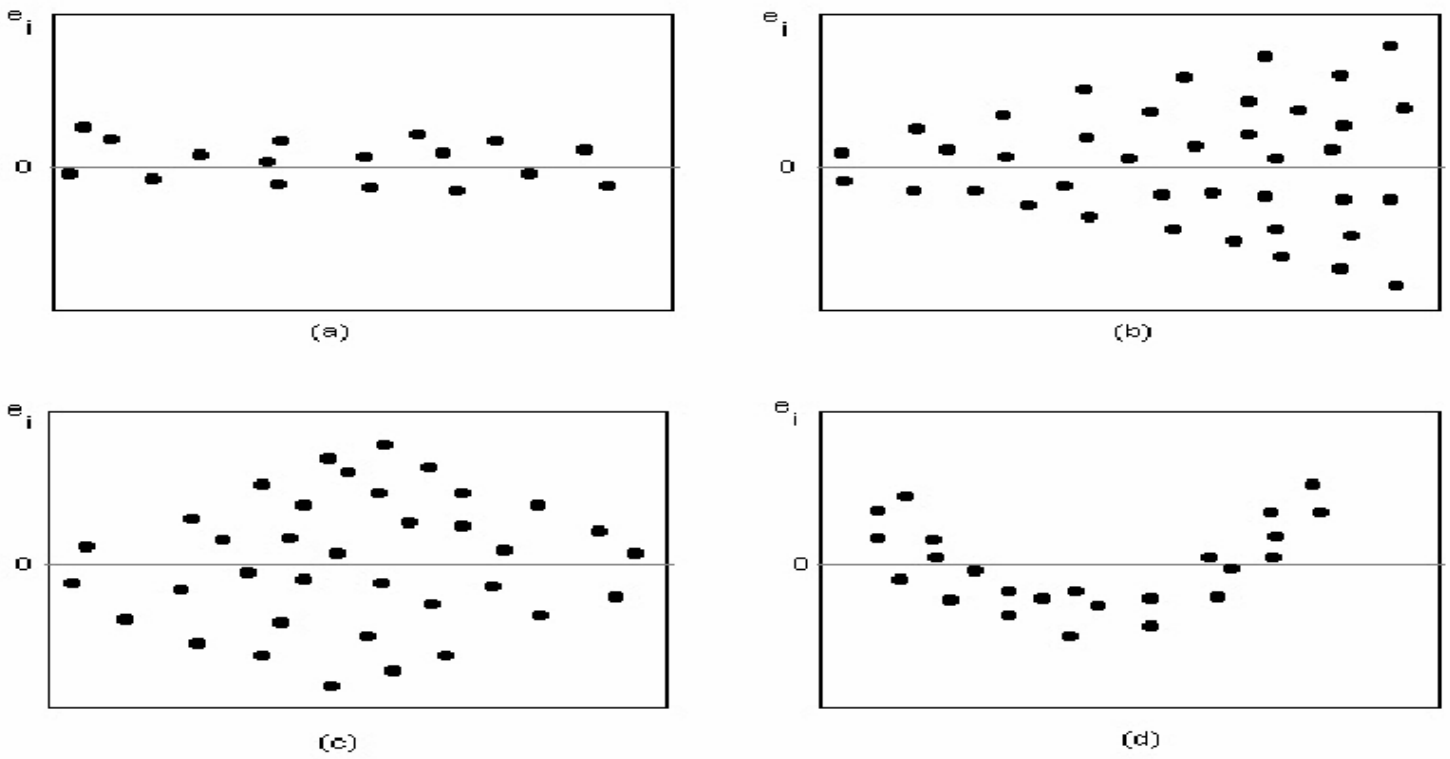
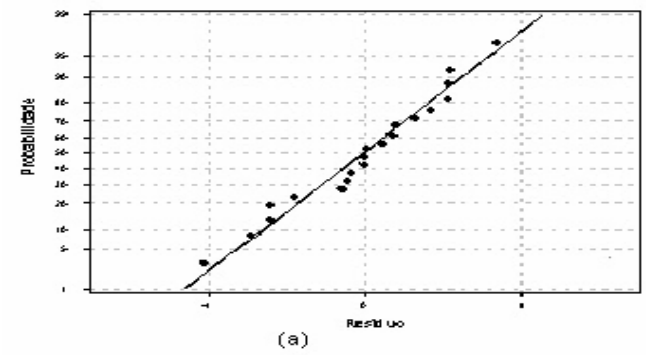
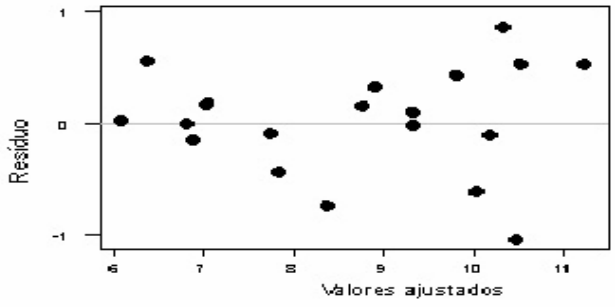


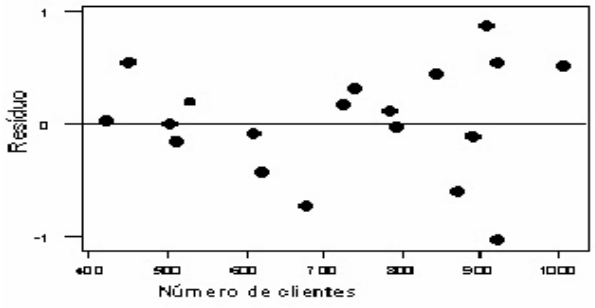
Gráfico de resíduos do exemplo



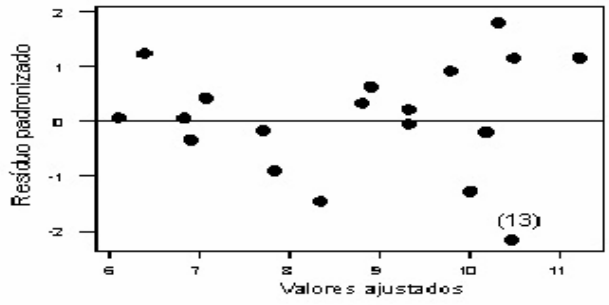
(a)



(b)



(c)



(d)

Coeficiente de determinação

A quantidade

$$R^2 = \frac{SQreg}{SQT} = 1 - \frac{SQR}{SQT}$$

recebe o nome de **coeficiente de determinação** e é usada para avaliar a adequação do modelo de regressão.
Pode ser interpretado como a proporção da variabilidade presente nas observações da variável resposta Y que é explicada pela variável independente X no modelo de regressão.

Exemplo

Para os dados do exemplo dos supermercados do exemplo, determinar R^2. Da definição tem-se

$$R^2 = \frac{SQreg}{SQT} = \frac{46,8371}{51,3605} = 0,912.$$

Esse resultado significa que o modelo ajustado explicou 91,2% da variação na variável resposta Y (vendas semanais). Isto é, 91,2% da variabilidade de Y é explicada pela variável regressora X (número de clientes).

Bibliografia

- Devore, J. L. Probabilidade e Estatística para Engenharia e Ciências. São Paulo: Pioneira Thomson Learning, 2006.
- Montgomery, D. C. & Runger, G. C. Estatística Aplicada e Probabilidade para Engenheiros. Rio de Janeiro: LTC, 2006.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.