

Regressão Linear Simples

Víctor Hugo Lachos Dávila

`hlachos@ime.unicamp.br`

Departamento Estatística-IMECC
Universidade Estadual de Campinas
Campinas, São Paulo, Brasil

Objetivos

Estudar a relação linear entre duas variáveis quantitativas.
Veja alguns exemplos:

- ⑥ Altura dos pais e altura dos filhos(Fig 1);

Objetivos

Estudar a relação linear entre duas variáveis quantitativas.
Veja alguns exemplos:

- ⑥ Altura dos pais e altura dos filhos(Fig 1);
- ⑥ Renda semanal e despesas de consumo;

Objetivos

Estudar a relação linear entre duas variáveis quantitativas.
Veja alguns exemplos:

- ⑥ Altura dos pais e altura dos filhos(Fig 1);
- ⑥ Renda semanal e despesas de consumo;
- ⑥ Variação dos salários e taxa de desemprego (Fig 2);

Objetivos

Estudar a relação linear entre duas variáveis quantitativas.
Veja alguns exemplos:

- ⑥ Altura dos pais e altura dos filhos(Fig 1);
- ⑥ Renda semanal e despesas de consumo;
- ⑥ Variação dos salários e taxa de desemprego (Fig 2);
- ⑥ Demanda dos productos de uma firma e publicidade;

Objetivos

Estudar a relação linear entre duas variáveis quantitativas.
Veja alguns exemplos:

- ⑥ Altura dos pais e altura dos filhos(Fig 1);
- ⑥ Renda semanal e despesas de consumo;
- ⑥ Variação dos salários e taxa de desemprego (Fig 2);
- ⑥ Demanda dos productos de uma firma e publicidade;

Sob dois pontos de vista:

- ⑥ Explicitando a forma dessa relação: **regressão**.
- ⑥ Quantificando a força dessa relação: **correlação**.

1) Regressão vs Causação

- ⑥ Uma relação estatística por si própria não implica uma causação
- ⑥ Para atribuir causação, devemos invocar a alguma teoria (p.e. económica)

2) Regressão (AR) vs Correlação (AC)

- ⑥ na AC há tratamento simétrico das variáveis
- ⑥ na AR a variável explanatoria é fixa
- ⑥ na AC presume-se que as duas variáveis são aleatórias

Dados Hipotéticos

Os dados se referem à renda semanal (X) e as despesas de consumo (Y) (em *US\$*), de uma população total de 60 famílias. As 60 famílias foram divididas em 10 grupos de renda (Fig 3 e 4).

Y	80	100	120	140	160	180	200	220	240	260
X	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	-	88	-	113	125	140	-	160	189	185
	-	-	-	115	-	-	-	162	-	191
Total	325	462	445	707	678	750	685	1043	966	1211
E(Y X)	65	77	89	101	113	125	137	149	161	173

Função de Regressão Populacional

É razoável supor que a média da variável aleatória Y , está relacionada com X pela seguinte relação

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

onde β_0 e β_1 , são respectivamente, o intercepto e a inclinação da reta e recebem o nome de coeficientes de regressão.

Função de Regressão Populacional

É razoável supor que a média da variável aleatória Y , está relacionada com X pela seguinte relação

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

onde β_0 e β_1 , são respectivamente, o intercepto e a inclinação da reta e recebem o nome de coeficientes de regressão.

Cada valor individual Y_i será determinado pelo valor médio da função linear ($\mu_{Y|x}$) mais um termo que representa um erro aleatório,

Função de Regressão Populacional

É razoável supor que a média da variável aleatória Y , está relacionada com X pela seguinte relação

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

onde β_0 e β_1 , são respectivamente, o intercepto e a inclinação da reta e recebem o nome de coeficientes de regressão.

Cada valor individual Y_i será determinado pelo valor médio da função linear ($\mu_{Y|x}$) mais um termo que representa um erro aleatório,

$$Y_i = \mu_{Y|x} + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

onde ε_i é o erro estocástico que satisfaz $E(\varepsilon_i|x_i) = 0$

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

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon,$$

A equação é denominada modelo de regressão linear múltipla.

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

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon,$$

A equação é denominada modelo de regressão linear múltipla.

O adjetivo "linear" é usado para indicar que o modelo é linear nos parâmetros β_1, \dots, β_k e não porque Y é função linear dos X 's. Por exemplo, 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, mas o mesmo não acontece com a equação $Y = \beta_0 + \beta_1 X_1^{\beta_2} + \beta_3 X_2^2 + \varepsilon$.

Significado do erro estocástico

- ⑥ Caráter vago da teoria
- ⑥ Falta de dados disponíveis
- ⑥ Variáveis essenciais vs variáveis periféricas
- ⑥ Caráter aleatório da natureza
- ⑥ Princípio da parcimônia
- ⑥ Forma funcional equivocada

Função de Regressão Amostral(FRA)

A tarefa agora é estimar a FRP com base em informações amostrais

$$Y_i = \hat{Y}_i + \hat{\varepsilon}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i, i = 1, \dots, n,$$

onde $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores de β_0 e β_1 , respectivamente e $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ a componente residual (Fig 5). Precisamos formular uma regra ou método que torne tal aproximação o mais próximo possível!


Exercício: Resolva o problema 2.16 do livro texto.

Estimação: Método de MQO

Suponha que tem-se n pares de observações amostrais $(x_1, y_1), \dots, (x_n, y_n)$. A soma de quadrados dos desvios das observações em relação à FRA é:

$$Q = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

O método de mínimos quadrados ordinários (MQO) escolhe $\hat{\beta}_1$ e $\hat{\beta}_2$ (únicos) de forma que, para qualquer amostra, Q é o menor possível. Após uma simple algebra tem-se


$$(1) \quad \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$
$$\hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.$$

As equações (1) recebem o nome de equações normais de mínimos quadrados.

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

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

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

onde $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$.

Notações especiais

$$\begin{aligned} S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} = \sum_{i=1}^n x_i^2 - n\bar{x}^2, \\ S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n} \\ &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}, \\ S_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})y_i = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = \sum_{i=1}^n y_i^2 - n\bar{y}^2. \end{aligned}$$

Os EMQ de β_0 e β_1 em termos da notação acima são:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}).$$

Observações sobre os EMQ

- Os EMQ dependem só de quantidades observáveis
- São estimadores pontuais
- A linha de regressão amostral é facilmente obtida

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

- O valor médio do resíduo $\hat{\epsilon}_i$ é zero
- Os resíduos $\hat{\epsilon}_i$ são não correlacionados com X_i e \hat{Y}_i .

Exemplo 1

O gerente de uma cadeia de supermercados deseja desenvolver um modelo com a finalidade de estimar as vendas médias semanais (em milhares de dólares)

- ⑥ Y - Vendas semanais; e
- ⑥ X - Número de clientes.

Estas variáveis foram observadas em 20 supermercados escolhidos 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

Considerando os dados do exemplo 1

$$n = 20$$

$$\sum_{i=1}^n x_i = 907 + 926 + \dots + 621 = 14.623; \quad \bar{x} = 731,15$$

$$\sum_{i=1}^n y_i = 11,20 + 11,05 + \dots + 7,41 = 176,11; \quad \bar{y} = 8,8055$$

$$\sum_{i=1}^n x_i^2 = (907)^2 + (926)^2 + \dots + (621)^2 = 11.306.209$$

$$\sum_{i=1}^n y_i^2 = (11,20)^2 + (11,05)^2 + \dots + (7,41)^2 = 1.602,0971$$

$$\sum_{i=1}^n x_i y_i = (907)(11,20) + (11,05)(926) \dots + (7,41)(621) = 134.127,90$$

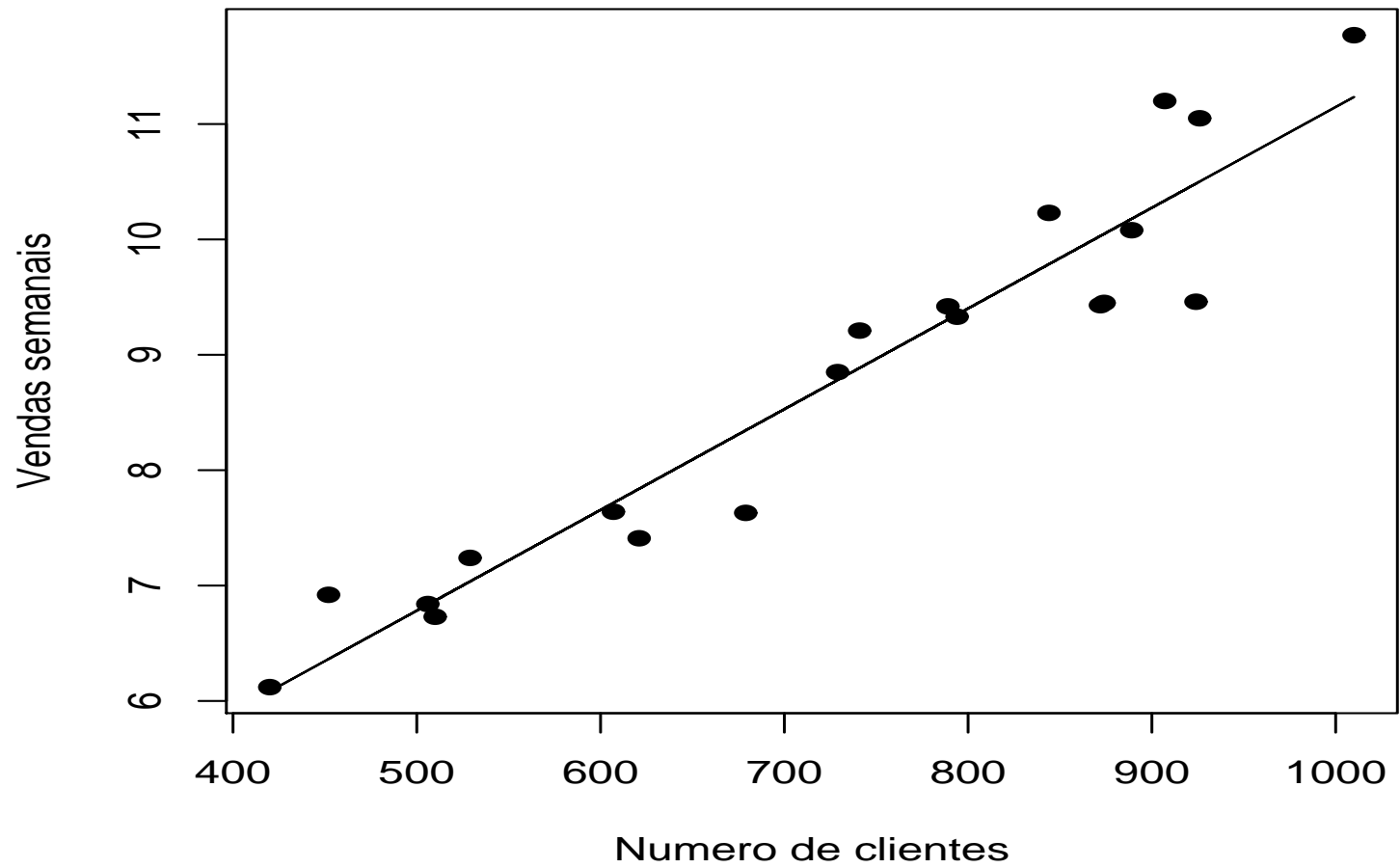
$$\begin{aligned}
 S_{xx} &= \sum_{i=1}^n x_i^2 - n(\bar{x})^2 = 11.306.209 - 20(731,15)^2 = 614.603 \\
 S_{xy} &= \sum_{i=1}^n x_i y_i - n(\bar{x})(\bar{y}) = 134.127,90 - 20(8,8055)(731,15) = 5.365,08 \\
 S_{yy} &= \sum_{i=1}^n y_i^2 - n(\bar{y})^2 = 1.609,0971 - 20(8,8055)^2 = 51,3605.
 \end{aligned}$$

As estimativas dos parâmetros do MRLS são:

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

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

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



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

$$\hat{y} = 2,423 + (0,00873)(600) = 7,661.$$

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

$$\hat{y} = 2,423 + (0,00873)(600) = 7,661.$$

A venda semanal de 7,661 mil dólares pode ser interpretada com uma estimacão da venda média semanal verdadeira dos supermercados com $X = 600$ clientes,

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

$$\hat{y} = 2,423 + (0,00873)(600) = 7,661.$$

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

Suposições do método de MQO

- (i) $E(\varepsilon|X) = 0$, $Var(\varepsilon|X) = \sigma^2$ (desconhecido).
- (ii) Os erros são não correlacionados
- (iii) A variável explicativa X é controlada pelo experimentador.
- (iv) o modelo de regressão está especificado da forma correta
- (v) $n >$ número de variáveis explanatorias
- (iv) não há multicolinearidade perfeita

Propriedades dos EMQ

Se as suposições do método de MQO são válidas, então

$$\textcircled{6} \quad E(\hat{\beta}_1) = \beta_1, \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \sigma_{\hat{\beta}_1}^2.$$

$$\textcircled{6} \quad E(\hat{\beta}_0) = \beta_0, \quad Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] = \sigma_{\hat{\beta}_0}^2.$$

$$\textcircled{6} \quad Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

Exercicio 2.

Estimação de σ^2

Os resíduos,

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

são empregados na estimação de σ^2 . A soma de quadrados residuais ou soma de quadrados dos erros, denotado por SQR é:

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

Pode-se demonstrar que o valor esperado da soma de quadrados dos residuais SQR , é dado por: (Exercício 3)

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

Portanto, um estimador não viciado de σ^2 , é

$$\hat{\sigma}^2 = \frac{SQR}{n - 2} = QMR \quad (\text{Quadrado médio residual}),$$

Uma fórmula mais conveniente para o cálculo da SQR é dada por:

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

A estimativa de σ^2 para o exemplo 1.

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

Seja x_p o valor para o qual deseja-se prever (ou projetar) o valor médio $E(Y|x_p)$ e o valor individual de Y .

- Previsão média

⑥ \hat{Y}_i é um estimador não viciado de $E[Y|x_p]$, dado que

$$E(\hat{Y}_i) = E(\hat{\beta}_0 + \hat{\beta}_1 x_p) = \beta_0 + \beta_1 x_p = E(Y|x_p)$$

⑥ $Var(\hat{Y}_i) = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right]$

- Previsão individual (Exercício 4.)

⑥ $Var(\hat{Y}_{part}) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right]$

Na prática substituímos σ^2 (desconhecido), pelo estimador consistente $\hat{\sigma}^2$

Coeficiente de Determinação (r^2)

O r^2 é uma medida de qualidade do ajustamento. No caso de regressão linear simples o coeficiente de determinação é o quadrado do coeficiente de correlação.(Fig 6)

$$\begin{aligned}(Y_i - \bar{Y}) &= (Y_i - \hat{Y}_i - \bar{Y} + \hat{Y}_i) \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2\end{aligned}$$

$$SQT = SQM + SQR$$

$$1 = \frac{SQM}{SQT} + \frac{SQR}{SQT} \Rightarrow r^2 = \frac{SQM}{SQT} = \frac{s_{xy}^2}{s_{xx}s_{yy}}$$

Teorema de Gauss-Markov

Se as suposições MQO são satisfeitas, os EMQ da classe de estimadores lineares não viesados têm variância mínima, isto é, são os melhores estimadores lineares não viesados. (Prova)

Para que normalidade?

A estimação é a metade do caminho, a outra metade é teste de hipóteses, para isto, suposições adicionais são necessárias.

- ⑥ uma alternativa é considerar tamanhos de amostra o suficientemente grandes (estimação de máxima verossimilhança)
- ⑥ a outra é supor que $\epsilon_i \sim N(0, \sigma^2)$ (O modelo de regressão normal simple clássico)

Propriedades dos EMQ sob Normalidade

A justificação teórica da premissa de normalidade é o TLC

$$\beta_1 = \sum_{i=1}^n k_i Y_i = \sum_{i=1}^n k_i (\beta_1 + \beta_2 x_i + \epsilon_i) \sim N(.)$$

- ⑥ $\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2)$, $\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2)$,
- ⑥ $(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)$
- ⑥ A distribuição de $\hat{\beta}_0$ e $\hat{\beta}_1$ é independente de $\hat{\sigma}^2$ (Exercício 5.)
- ⑥ $\hat{\beta}_0$ e $\hat{\beta}_1$ têm variância mínima dentro de toda classe dos estimadores não viesados, sejam ou não lineares (Rao)
- ⑥ $Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

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:

$$H_0 : \beta_1 = \beta_{1,0}, \text{ vs } H_1 : \beta_1 \neq \beta_{1,0}$$

A estatística

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

tem distribuição t -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

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

A estatística

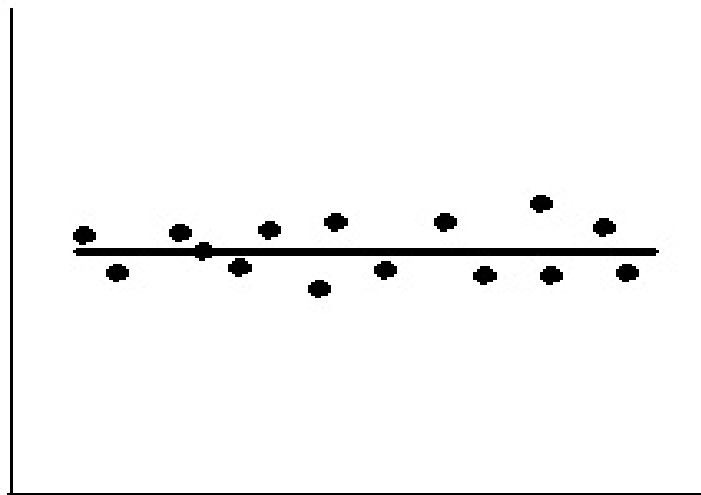
$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}}$$

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

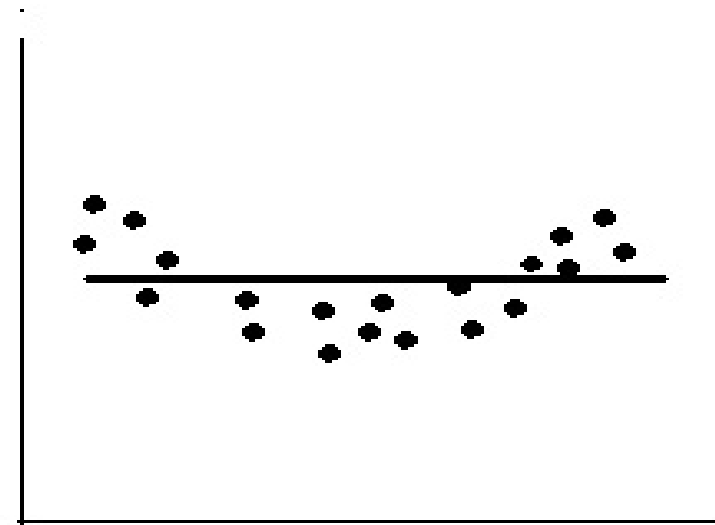
Teste de significância do MRLS

$$H_0 : \beta_1 = 0, \text{ vs } H_1 : \beta_1 \neq 0,$$

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

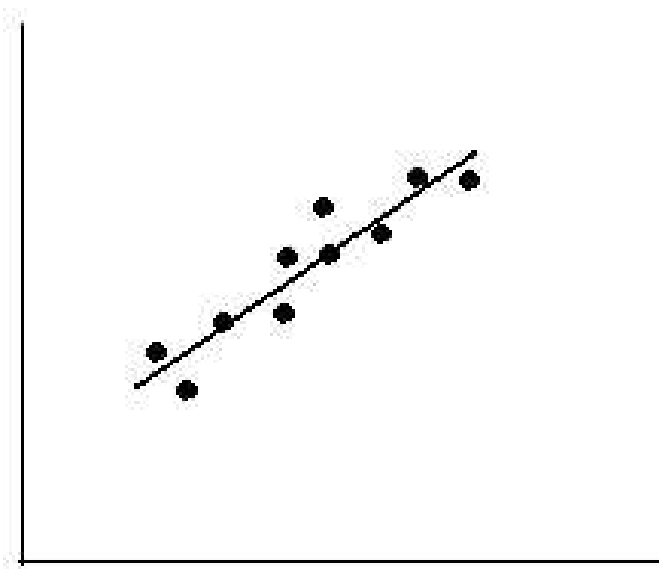


(a)

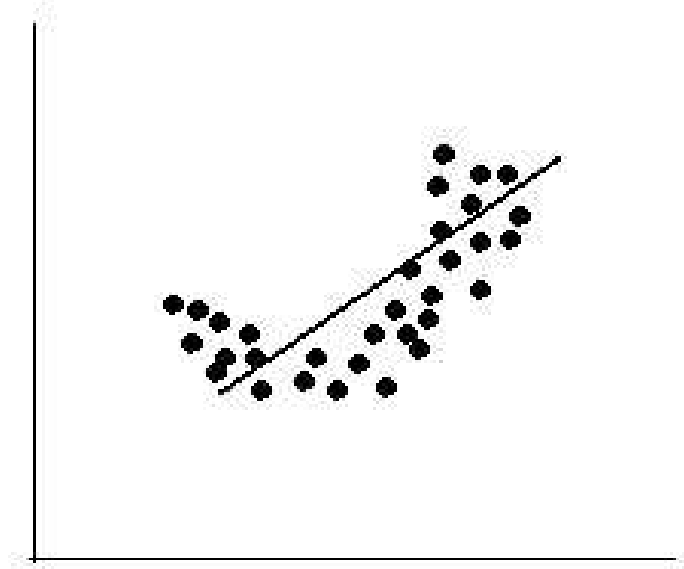


(b)

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



(a)



(b)

Exemplo

Teste de significância para o MRLS para os dados do exemplo 1, com $\alpha = 0,05$.

As hipóteses são $H_0 : \beta_1 = 0$, vs $H_1 : \beta_1 \neq 0$

Do exemplo tem-se:

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

De modo que a estatística de teste, é:

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

Como $T_{obs} = 13,65 > t_{0,03,18} = 2,101$, rejeita-se a hipótese $H_0 : \beta_1 = 0$.

Análise de variância

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

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

Portanto, rejeita-se H_0 se $F_{obs} > F_{\alpha, 1, n-2}$.

As quantidades

$QMreg = \frac{SQM}{1}$, (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	Quadrado Médio	F
Regressão	SQM	1	QM_{reg}	$\frac{QM_{reg}}{QMR}$
Residual	SQR	$n - 2$	QMR	
Total	SQT	$n - 1$		

Tabela de ANOVA

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	SQM	1	QM_{reg}	$\frac{QM_{reg}}{QMR}$
Residual	SQR	$n - 2$	QMR	
Total	SQT	$n - 1$		

Exemplo: o procedimento de análise de variância para testar se de fato existe relação linear entre o número de clientes (X) e as vendas semanais (Y), no modelo proposto para os dados do exemplo 1. Relembre que $S_{yy} = 51,3605$, $\hat{\beta}_1 = 0,00873$, $S_{xy} = 5.365,08$ e $n = 20$.

A soma de quadrados da regressão é

$$SQM = \hat{\beta}_1 S_{xy} = (0,00873)(5.365,08) = 46,8371$$

enquanto a soma de quadrados dos residuais é:

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

A soma de quadrados da regressão é

$$SQM = \hat{\beta}_1 S_{xy} = (0,00873)(5.365,08) = 46,8371$$

enquanto a soma de quadrados dos residuais é:

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

A ANOVA para testar $H_0 : \beta_1 = 0$. Nesse caso, a estatística de teste é

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

A soma de quadrados da regressão é

$$SQM = \hat{\beta}_1 S_{xy} = (0,00873)(5.365,08) = 46,8371$$

enquanto a soma de quadrados dos residuais é:

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

A ANOVA para testar $H_0 : \beta_1 = 0$. Nesse caso, a estatística de teste é

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

Como $F_{obs} = 186,4536 > F_{0,05,1,18} = 4,41$ rejeita-se H_0 , ao nível de significância de 5%.

Tabela de ANOVA para Ex. 1

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	46,8371	1	46,8371	186,45
Residual	4,5234	18	0,2513	
Total	51,3605	19		

Intervalo de confiança para β_0 e β_1

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

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

são variáveis aleatórias com distribuição t -Student com $n - 2$ graus de liberdade.

Um intervalo de $100(1 - \alpha)\%$ de confiança para β_1 :

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

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

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

A seguir é obtido um intervalo de 95% de confiança para a inclinação do MRLS com os dados do exemplo 1,

Relembre que $n = 20$, $\hat{\beta}_1 = 0,00873$, $S_{xx} = 614,603$ e $QMR = 0,2513$. Para $1 - \alpha = 0,95$, tem-se $t_{0,025,18} = 2,101$.

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

$$E = t_{0,025,18} \sqrt{\frac{QMR}{S_{xx}}} = 2,101 \sqrt{\frac{0,2513}{614,603}} = 0,00134$$

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

Intervalo de confiança para resposta média

O interesse consiste em estimar 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

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

Se $\varepsilon_i \sim NID(0, \sigma^2)$ é válida, pode-se demonstrar

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

Int. conf. $100(1 - \alpha)\%$ para $\mu_{Y|x_0}$

$$IC(\hat{\mu}_{Y|x}; 1 - \alpha) = (\hat{\mu}_{Y|x_0} - E; \hat{\mu}_{Y|x_0} + E)$$

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

Exemplo: Suponha que tem-se interesse em construir um intervalo de 95% de confiança da venda, média, semanal para todos supermercados com 600 clientes.

Int. conf. $100(1 - \alpha)\%$ para $\mu_{Y|x_0}$

$$IC(\hat{\mu}_{Y|x}; 1 - \alpha) = (\hat{\mu}_{Y|x_0} - E; \hat{\mu}_{Y|x_0} + E)$$

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

Exemplo: Suponha que tem-se interesse em construir um intervalo de 95% de confiança da venda, média, semanal para todos supermercados com 600 clientes.

No modelo ajustado $\hat{\mu}_{Y|x_0} = 2,423 + 0,00873x_0$. Para $x_0 = 600$, obtém-se $\hat{\mu}_{Y|x_0} = 7,661$.

Int. conf. $100(1 - \alpha)\%$ para $\mu_{Y|x_0}$

$$IC(\hat{\mu}_{Y|x}; 1 - \alpha) = (\hat{\mu}_{Y|x_0} - E; \hat{\mu}_{Y|x_0} + E)$$

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

Exemplo: Suponha que tem-se interesse em construir um intervalo de 95% de confiança da venda, média, semanal para todos supermercados com 600 clientes.

No modelo ajustado $\hat{\mu}_{Y|x_0} = 2,423 + 0,00873x_0$. Para $x_0 = 600$, obtém-se $\hat{\mu}_{Y|x_0} = 7,661$. Também, $\bar{x} = 731,15$, $QMR = 0,2513$, $S_{xx} = 614.603$, $n = 20$ e $1 - \alpha = 0,95 \Rightarrow t_{0,05,18} = 2,101$.

Int. conf. $100(1 - \alpha)\%$ para $\mu_{Y|x_0}$

$$IC(\hat{\mu}_{Y|x}; 1 - \alpha) = (\hat{\mu}_{Y|x_0} - E; \hat{\mu}_{Y|x_0} + E)$$

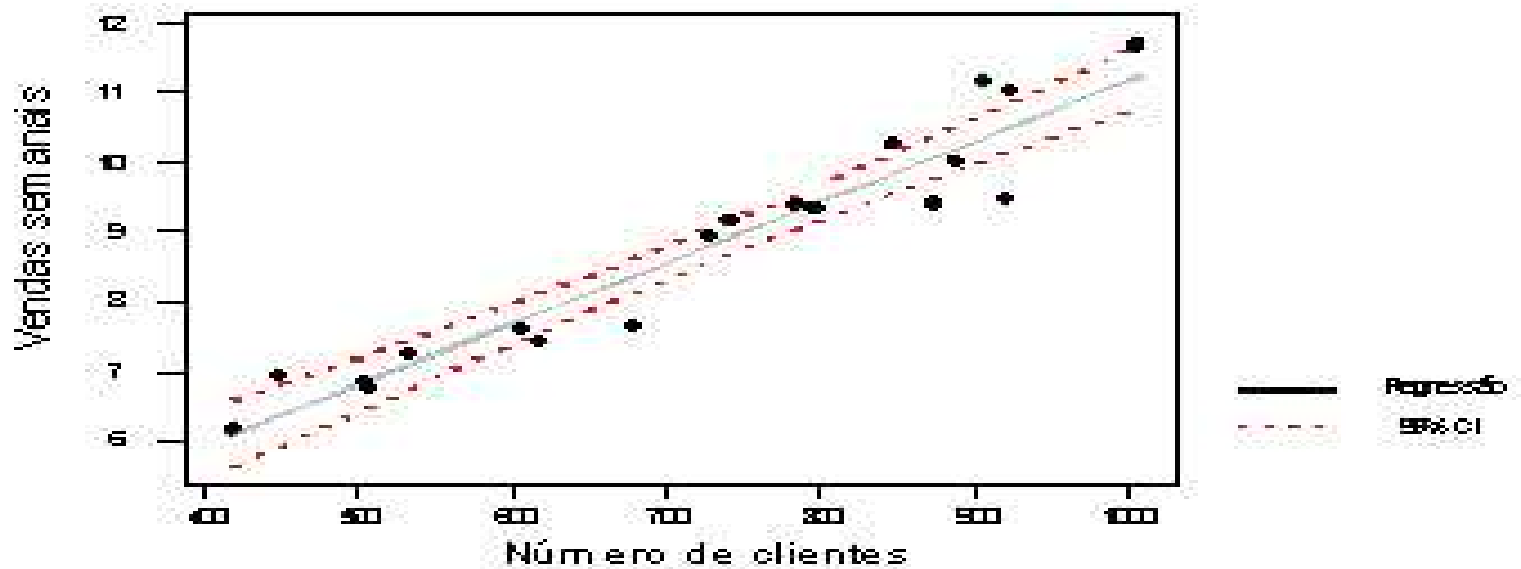
onde $E = t_{\frac{\alpha}{2}, n-2} \sqrt{QMR \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}$

Exemplo: Suponha que tem-se interesse em construir um intervalo de 95% de confiança da venda, média, semanal para todos supermercados com 600 clientes.

No modelo ajustado $\hat{\mu}_{Y|x_0} = 2,423 + 0,00873x_0$. Para $x_0 = 600$, obtém-se $\hat{\mu}_{Y|x_0} = 7,661$. Também, $\bar{x} = 731,15$, $QMR = 0,2513$, $S_{xx} = 614.603$, $n = 20$ e $1 - \alpha = 0,95 \Rightarrow t_{0,05,18} = 2,101$.

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

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



Previsão de novas observações

Uma aplicação muito importante de um modelo de regressão é a previsão de novas ou futuras observações de Y , (Y_0) correspondente a um dado valor da variável explicativa X , x_0 , então

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

é o melhor estimador pontual de Y_0 .

Um intervalo de $100(1 - \alpha)\%$ de confiança para uma futura observação é dado por:

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

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

Exemplo

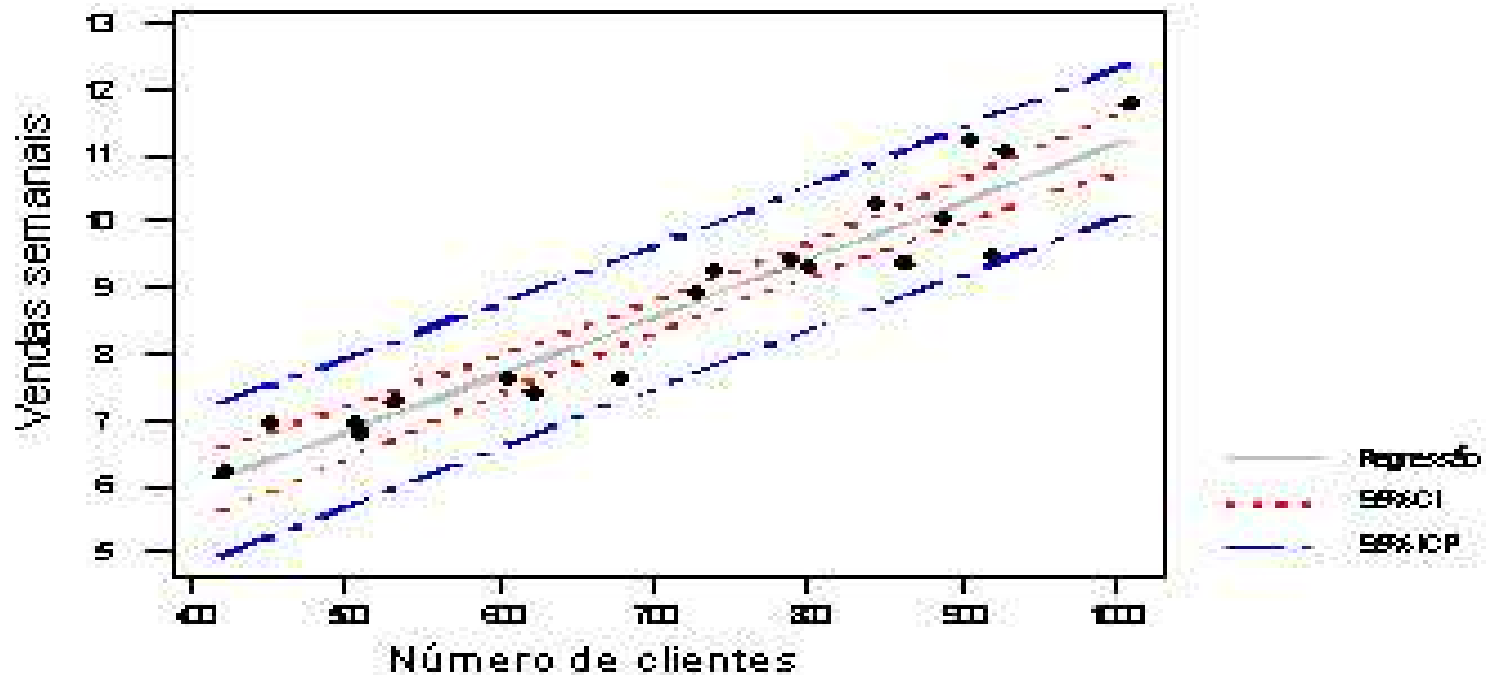
Suponha agora, tem-se interesse em encontrar um intervalo de previsão de 95% das vendas semanais de um supermercado com 600 clientes.

Considerando os dados do exemplo 1, $\hat{Y} = 7,661$ e o intervalo de predição é:

$$E = 2,101 \sqrt{0,2513 \left[1 + \frac{1}{20} + \frac{(600 - 731,15)^2}{614.603} \right]} = 1,084$$

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

Bandas de confiança do 95% para $\mu_{Y|x_0}$ (CI) e Y_0 (ICP)



Adequação do modelo de regressão



- ⑥ Análise residual,

Adequação do modelo de regressão



- ⑥ Análise residual,
- ⑥ Coeficiente de determinação

Adequação do modelo de regressão

- ⑥ Análise residual,
- ⑥ Coeficiente de determinação

Os resíduos de um modelo de regressão são definidos como

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

onde y_i é uma observação real de Y e \hat{y}_i é o valor correspondente estimado através do modelo de regressão.

Adequação do modelo de regressão

- ⑥ Análise residual,
- ⑥ Coeficiente de determinação

Os resíduos de um modelo de regressão são definidos como

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

onde y_i é uma observação real de Y e \hat{y}_i é o valor correspondente estimado através do modelo de regressão.
Resíduos padronizados

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

Adequação do modelo de regressão

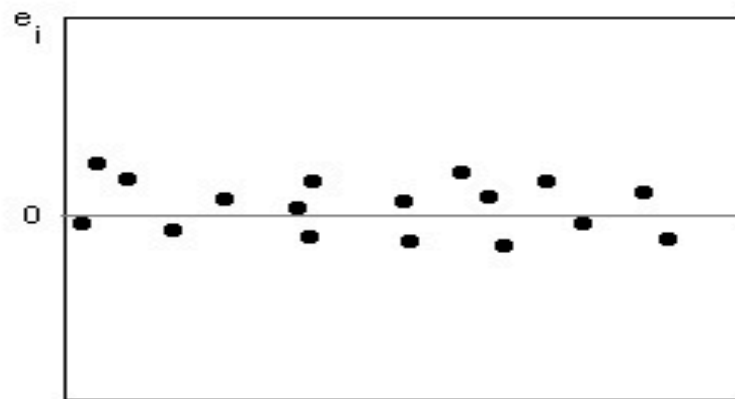
- ⑥ Análise residual,
- ⑥ Coeficiente de determinação

Os resíduos de um modelo de regressão são definidos como

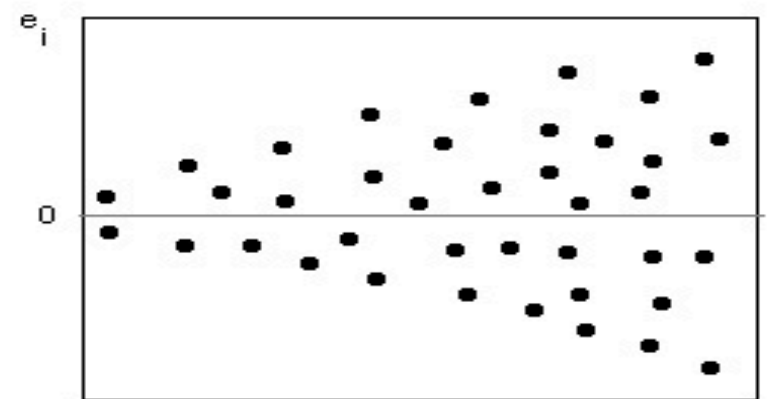
$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

onde y_i é uma observação real de Y e \hat{y}_i é o valor correspondente estimado através do modelo de regressão.
Resíduos padronizados

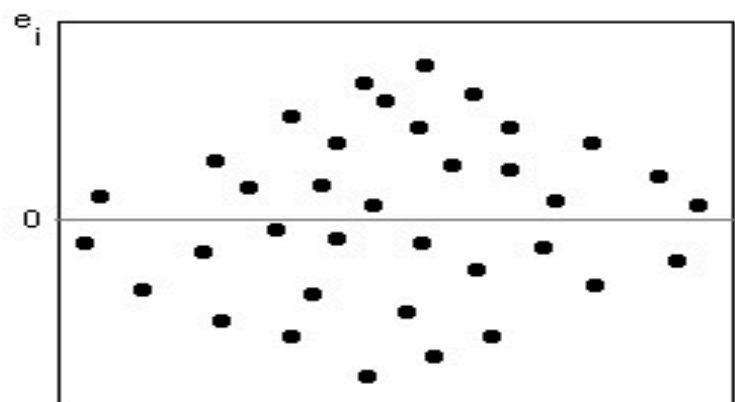
$$d_i = \frac{e_i}{\sqrt{QMR}}, \quad i = 1, \dots, n$$



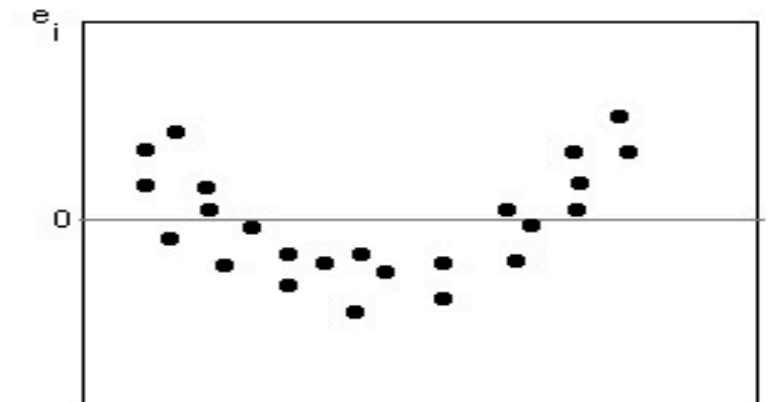
(a)



(b)

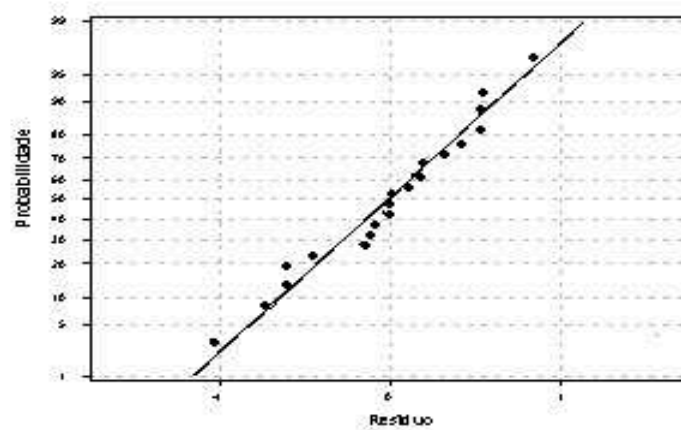


(c)

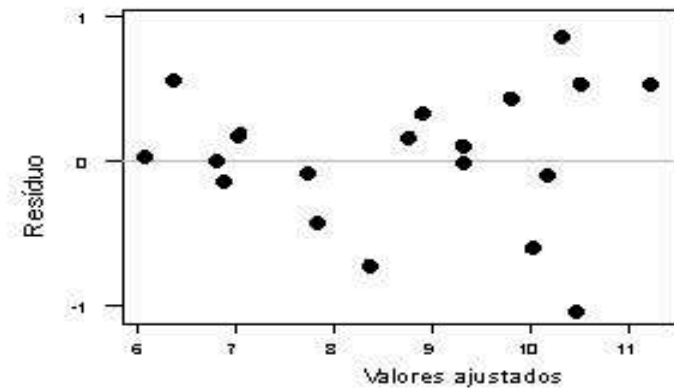


(d)

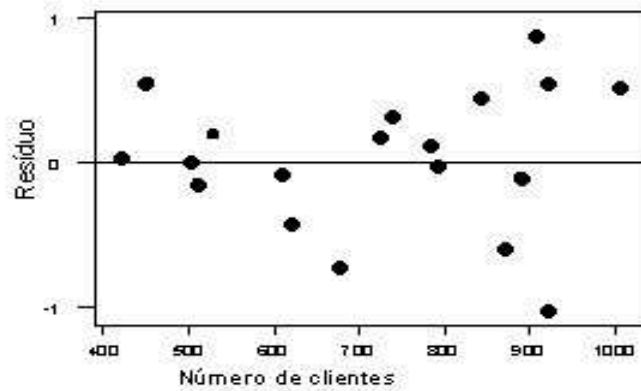
Gráfico de resíduos do exemplo 1



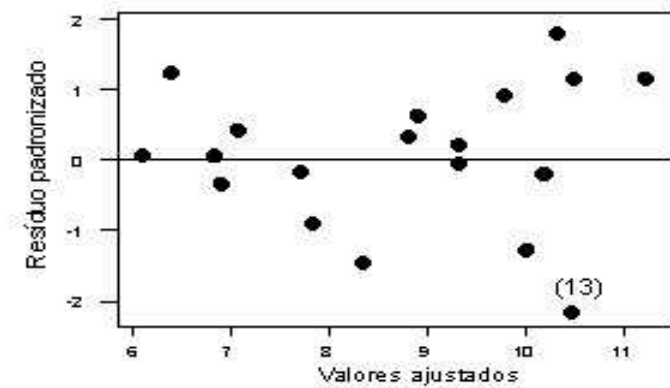
(a)



(b)



(c)



(d)

Exemplo: Coeficiente de Determinação

Para os dados dos supermercados do exemplo1,
determinar R^2 .

Exemplo: Coeficiente de Determinação

Para os dados dos supermercados do exemplo1,
determinar R^2 . Da definição tem-se:

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

Exemplo: Coeficiente de Determinação

Para os dados dos supermercados do exemplo1, determinar R^2 . Da definição tem-se:

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

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).

Analise de Correlação

Suponha que se deseja desenvolver um modelo de regressão que relacione a resistência ao corte dos pontos de soldadura com o diâmetro dos mesmos. Neste caso, não é possível controlar o diâmetro de soldadura. O que pode ser feito é selecionar ao acaso n pontos de soldadura e observar o diâmetro (X_i) e a resistência ao corte (Y_i) de cada um deles. Portanto, (X_i, Y_i) são variáveis aleatórias distribuídas de maneira conjunta.

Suponha que a distribuição conjunta de X_i e Y_i tenha uma distribuição normal bivariada cuja função de densidade é dada por

Suponha que a distribuição conjunta de X_i e Y_i tenha uma distribuição normal bivariada cuja função de densidade é dada por

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ \frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) \right] \right\}$$

Suponha que a distribuição conjunta de X_i e Y_i tenha uma distribuição normal bivariada cuja função de densidade é dada por

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ \frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) \right] \right\}$$

onde μ_1 e σ_1^2 são a média e variância de X e μ_2 e σ_2^2 são a média e variância de Y e, ρ é *coeficiente de correlação* entre X e Y .

A densidade condicional de Y para um valor dado $X = x$ é dado por (exercício 5.)

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1 x}{\sigma_{Y|x}^2} \right)^2 \right\}$$

A densidade condicional de Y para um valor dado $X = x$ é dado por (exercício 5.)

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1 x}{\sigma_{Y|x}^2} \right)^2 \right\}$$

onde $\beta_0 = \mu_2 - \mu_1 \rho \frac{\sigma_2}{\sigma_1}$, $\beta_1 = \frac{\sigma_2}{\sigma_1} \rho$ e $\sigma_{Y|x}^2 = \sigma_2^2 (1 - \rho^2)$

A densidade condicional de Y para um valor dado $X = x$ é dado por (exercício 5.)

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp \left\{ -\frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1 x}{\sigma_{Y|x}^2} \right)^2 \right\}$$

onde $\beta_0 = \mu_2 - \mu_1 \rho \frac{\sigma_2}{\sigma_1}$, $\beta_1 = \frac{\sigma_2}{\sigma_1} \rho$ e $\sigma_{Y|x}^2 = \sigma_2^2 (1 - \rho^2)$

A distribuição condicional de Y dado $X = x$ é normal com média

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

e variância $\sigma_{Y|x}^2$.

Estimadores de β_0 , β_1 e ρ

⑥ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Estimadores de β_0 , β_1 e ρ

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

Estimadores de β_0 , β_1 e ρ

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\rho} = r = \frac{\sum_{i=1}^n Y_i (X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

Estimadores de β_0 , β_1 e ρ

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\rho} = r = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

$$\hat{\beta}_1 = \left(\frac{S_{YY}}{S_{XX}} \right)^{1/2} r$$

Teste de hipóteses

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

A estatística de teste apropriada é

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{\sim}{\text{sob } H_0} t(n-2)$$

Teste de hipóteses

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

A estatística de teste apropriada é

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{\sim}{\text{sob } H_0} t(n-2)$$

A hipótese nula deverá ser rejeitada se $|T_{obs}| \geq t_{\alpha/2, n-2}$.
Esse teste é equivalente ao teste de hipóteses $H_0 : \beta_1 = 0$.


$$H_0 : \rho = \rho_0 \text{ vs } H_1 : \rho \neq \rho_0$$

onde $\rho_0 \neq 0$.

$$H_0 : \rho = \rho_0 \text{ vs } H_1 : \rho \neq \rho_0$$

onde $\rho_0 \neq 0$.

Para amostras de tamanho moderado grande ($n \geq 30$), a estatística


$$H_0 : \rho = \rho_0 \text{ vs } H_1 : \rho \neq \rho_0$$

onde $\rho_0 \neq 0$.

Para amostras de tamanho moderado grande ($n \geq 30$), a estatística

$$Z_r = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r}$$

tem distribuição aproximadamente normal com média

$$\mu_{Z_r} = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

e variância $\sigma_{Z_r}^2 = (n-3)^{-1}$.

A estatística de teste apropriada é:

$$Z = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0) (n - 3)^{1/2}.$$

A estatística de teste apropriada é:

$$Z = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0) (n - 3)^{1/2}.$$

Se $H_0 : \rho = \rho_0$ é verdadeira, a estatística Z tem, aproximadamente, distribuição normal padrão. Portanto, H_0 deverá ser rejeitada se $|Z_{obs}| \geq z_{\alpha/2}$.

Intervalo de confiança para ρ

Um intervalo aproximado de $100(1 - \alpha)\%$ de confiança para o coeficiente de correlação ρ , que é dado por:


$$IC(\rho; 1 - \alpha) = \left(\tanh \left[\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n - 3}} \right]; \right. \\ \left. \tanh \left[\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n - 3}} \right] \right),$$

onde $\tanh w = \frac{e^w - e^{-w}}{e^w + e^{-w}}$.

Exemplo 2

Suponha que se tenha interesse em medir a força da relação linear de dois produtos diferentes com relação ao preço em várias cidades do mundo.

- ⑥ Y - Preço de uma libra de frango; e
- ⑥ X - Preço de uma caixa de suco.



Cidade	Caixa com seis sucos (X)	Uma libra de frango (Y)
Frankfurt	3,27	3,06
Hong Kong	2,22	2,34
Londres	2,28	2,27
Manila	3,04	1,51
México	2,33	1,87
Nova York	2,69	1,65
París	4,07	3,09
Sidney	2,78	2,36
Tokyo	5,97	4,85

Dos dados da tabela são obtidos os valores seguintes:

$$n = 9; \sum_{i=1}^n X_i = 28,65; \bar{X} = 3,183; \sum_{i=1}^n X_i^2 = 28,65 = 102$$

$$S_{XX} = 11,4594; \sum_{i=1}^n Y_i = 23,00; \bar{Y} = 2,5566; \sum_{i=1}^n Y_i^2 = 67$$

$$S_{YY} = 8,3522; \sum_{i=1}^n X_i Y_i = 81,854; S_{XY} = 8,6437$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{8,6437}{\sqrt{(11,4594)(8,3522)}} = 0,883.$$

$H_0 : \rho = 0$ (não relação linear entre X e Y)

$H_1 : \rho \neq 0$ (há relação linear entre X e Y)

O valor calculado para a estatística do teste foi

$$T_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,883\sqrt{9-2}}{\sqrt{1-(0,883)^2}} = 4,98.$$

Para $\alpha = 0,05$, tem-se que $t_{0,025,7} = 2,365 < T_{obs} = 4,98$, logo, rejeita-se $H_0 : \rho = 0$ ao nível de significância de $\alpha = 5\%$.