

10. ANÁLISE DE CORRELAÇÃO E REGRESSÃO



2019

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.
- Teor de carbono e resistência de uma peça produzida.

Pontos de vista:

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

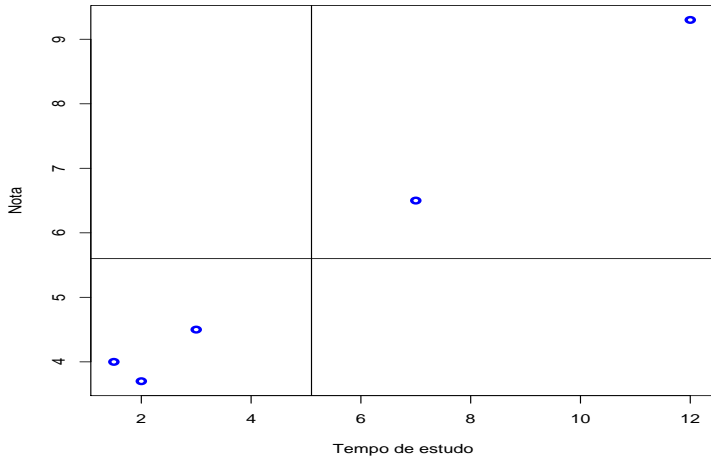
Exemplo

- 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

Gráfico de dispersão



Coeficiente de correlação

O coeficiente de correlação linear 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

$$S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} = \sum_{i=1}^n X_i^2 - n\bar{X}^2,$$

$$\begin{aligned} S_{XY} &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \bar{X})Y_i = \sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n} \\ &= \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} \end{aligned}$$

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

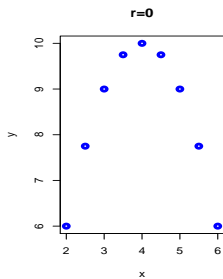
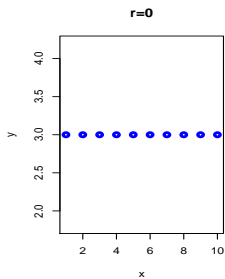
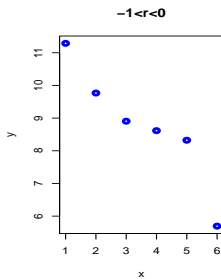
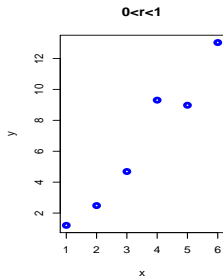
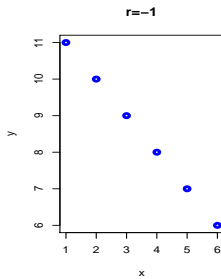
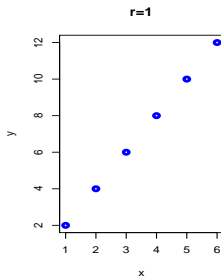
Intervalo de valores de r :

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

Classificação:

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

Exemplos



Exemplos

Variáveis:

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

Estas variáveis foram observadas em nove localidades com as mesmas características demográficas e sócio-econômicas.

Os dados coletados foram os seguintes:

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

Apresente o coeficiente de correlação entre X e Y .

Exemplos

	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
Soma	3212	265	98955	1168772	8727

Exemplos

$$S_{XY} = \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i / n = 98955 - (265)(3212)/9 = 4379,444,$$

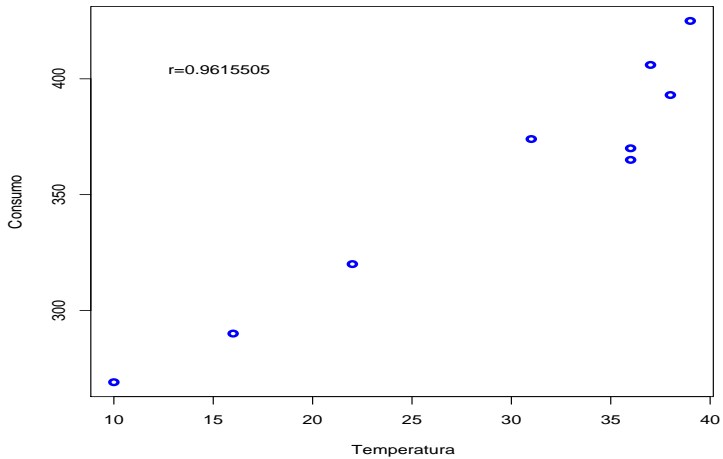
$$S_{YY} = \sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n Y_i)^2 / n = 1168772 - (3212)^2 / 9 = 22444,89,$$

$$S_{XX} = \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2 / n = 8727 - (265)^2 / 9 = 924,222$$

e

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{4379,444}{\sqrt{22444,89 \times 924,222}} = 0,9615.$$

Gráfico de dispersão



Análise de regressão

Exemplo

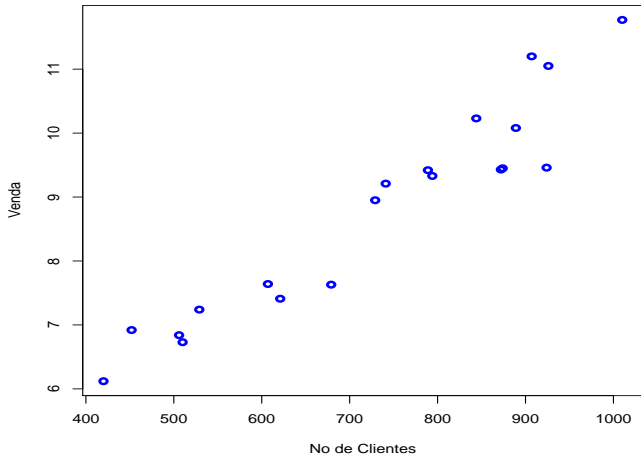
O gerente de uma grande cadeia de lojas deseja desenvolver um modelo com a finalidade de estimar as vendas médias semanais (em milhares de reais).

- *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

Gráfico 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, \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.

Análise de regressão

O adjetivo “linear” é usado para indicar que o modelo é linear nos coeficientes β_1, \dots, β_k e não porque Y é função linear dos X 's. Uma expressão da forma

$$Y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2^3 + \varepsilon$$

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 (ou explicativa ou regressora) X e uma variável dependente (ou resposta) Y :

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

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

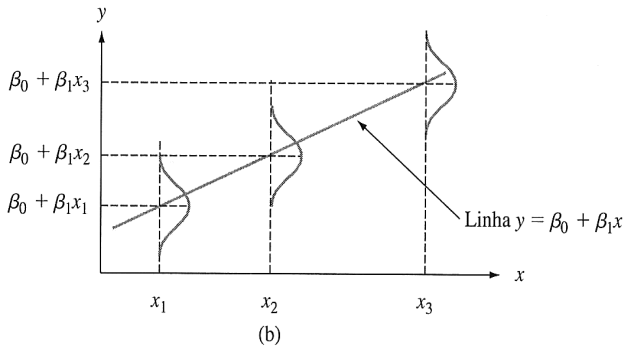
Suposições do MRLS

- (i) $E(\varepsilon) = 0$ e $Var(\varepsilon) = \sigma^2$ (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

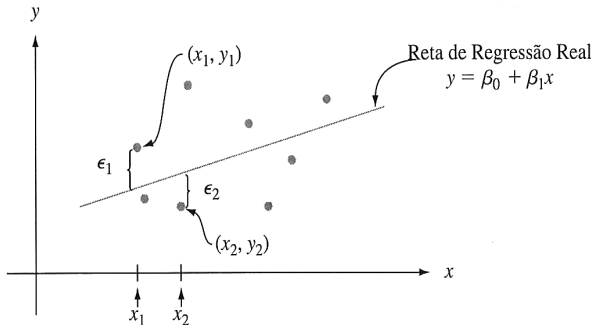
$$E(Y|X = x) = \mu_{Y|X} = \beta_0 + \beta_1 x.$$

Suposições do MRLS



Estimação pelo método de mínimos quadrados (MQ)

Foram coletados n pares de observações $(x_1, y_1), \dots, (x_n, y_n)$. 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

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n. \quad (2)$$

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 $\hat{\beta}_0$ e $\hat{\beta}_1$, devem resolver as equações

$$\frac{\partial Q}{\partial \beta_0} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

e

$$\frac{\partial Q}{\partial \beta_1} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0.$$

Estimação

Após simplificar as expressões anteriores obtemos

$$\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n Y_i \quad (3)$$

$$\text{e } \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 x_i^2 = \sum_{i=1}^n x_i Y_i.$$

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

Estimação

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

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

e

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

Estimação

Portanto, a linha de regressão estimada ou ajustada é

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

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

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \dots, n,$$

em que $e_i = Y_i - \hat{Y}_i$ recebe o nome de **resíduo**.

Exemplo de aplicação

Considerando os dados do exemplo (vendas semanais),

$$n = 20,$$

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

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

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

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

e

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

Exemplo de aplicação

$$S_{XX} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 11.306.209 - 20 \times 731,15^2 = 614.603,$$

$$S_{XY} = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = 134.127,90 - 20 \times 8,8055 \times 731,15 = 5.365,08$$

e

$$S_{YY} = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = 1.609,0971 - 20 \times 8,8055^2 = 51,3605.$$

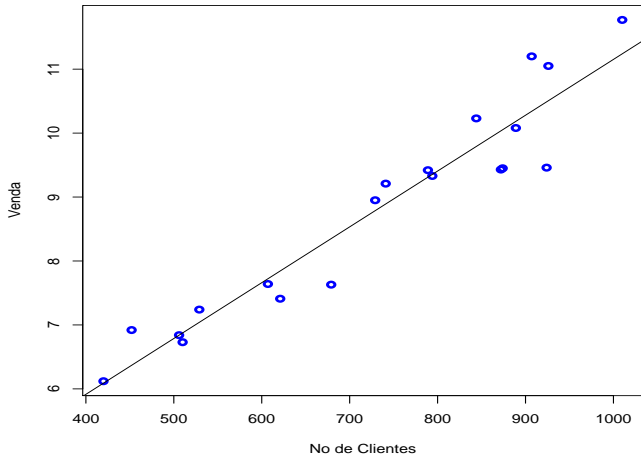
As estimativas dos coeficientes do MRLS são

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{5.365,08}{614.603} = 0,00873 \quad \text{e} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = 8,8055 - 0,00873 \times 731,15 = 2,423.$$

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

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

Gráfico 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,

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

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

Propriedades dos EMQ

Se as suposições do MRLS forem válidas é possível demonstrar que

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

$\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores **não viesados**.

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, denotada 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 é dado por

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

Portanto,

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

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 $\hat{\beta}_1 = 0,00873$. Portanto, a estimativa de σ^2 para o exemplo é

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

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

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

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

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

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

Rejeitamos a hipóteses nula se $|T_{obs}| > t_{\alpha/2, n-2}$.

Exemplo

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

As hipóteses são $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$.

Do exemplo tem-se

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

de modo que a estatística de teste é

$$T_{obs} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{XX}}} = \frac{0,00873}{\sqrt{0,2512/614,603}} = 13,65.$$

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

Intervalo de confiança para β_0 e β_1

Se forem válidas as suposições do MRLS, então

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

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

Intervalo de confiança (IC) de $100(1 - \alpha)\%$ para β_1 :

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

Intervalo de confiança para β_0

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

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

Intervalo de 95% de confiança para a inclinação com os dados do exemplo.

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

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

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

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

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

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

Intervalo de confiança para a resposta média

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

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

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

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

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

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 $t_{0,025;18} = 2,101$.

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

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

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

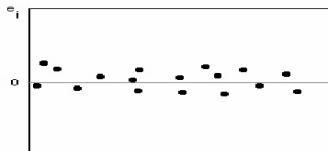
$$e_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n,$$

em que Y_i é uma observação de Y e \hat{Y}_i é o valor correspondente predito através do modelo de regressão.

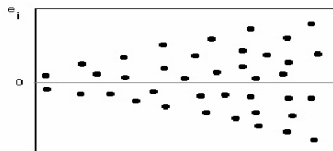
Resíduos padronizados:

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

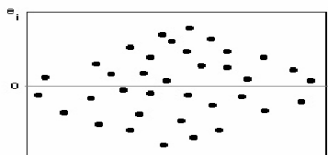
Gráfico de resíduos



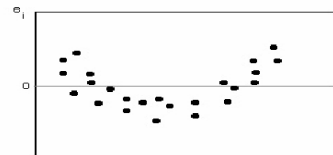
(a)



(b)

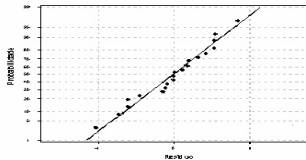


(c)

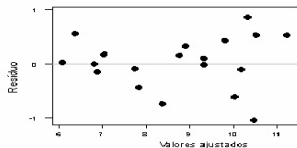


(d)

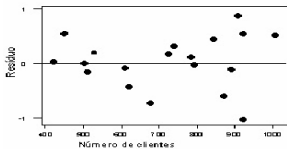
Gráfico de resíduos do exemplo



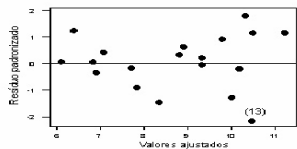
(a)



(b)



(c)



(d)

Coeficiente de determinação

A quantidade

$$R^2 = \frac{SQ_{reg}}{SQT} = 1 - \frac{SQR}{SQT}, \quad \text{em que } 0 \leq R^2 \leq 1,$$

recebe o nome de **coeficiente de determinação** e é usada para avaliar a adequação do modelo de regressão, sendo que

$SQ_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$: soma de quadrados da regressão e

$$SQT = S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

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

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

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.