

Regressão linear simples

CITAÇÃO

Graça Martins, M. E. (2019)
Regressão linear simples,
Rev. Ciência Elem., V7 (03):045.
doi.org/10.24927/rce2019.045

EDITOR

José Ferreira Gomes,
Universidade do Porto

RECEBIDO EM

16 de fevereiro de 2019

ACEITE EM

23 de março de 2019

PUBLICADO EM

16 de outubro de 2019

COPYRIGHT

© Casa das Ciências 2019.
Este artigo é de acesso livre,
distribuído sob licença Creative
Commons com a designação
[CC-BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), que permite
a utilização e a partilha para fins
não comerciais, desde que citado
o autor e a fonte original do artigo.

rce.casadasciencias.org



Maria Eugénia Graça Martins

Universidade de Lisboa

Um modelo de Regressão é um modelo matemático que descreve a relação entre duas ou mais variáveis de tipo quantitativo. Se o estudo incidir unicamente sobre duas variáveis e o modelo matemático for a equação de uma reta, então designa-se por regressão linear simples.

Quando o diagrama de dispersão sugere a existência de uma associação linear entre duas variáveis x e y , é possível resumir através de uma reta a forma como a variável dependente ou *variável resposta* (ou *variável a prever*) y é influenciada pela variável independente ou *variável explanatória* (ou *variável preditora*) x . A esta reta dá-se o nome de **reta de regressão**.

Dado um conjunto de dados bivariados (x_i, y_i) , $i = 1, \dots, n$, do par de variáveis (x, y) , pode ter interesse ajustar uma reta da forma $y = a + bx$, que dê informação sobre como se refletem em y as mudanças processadas em x . Um dos métodos mais conhecidos de ajustar uma reta a um conjunto de dados é o *método dos mínimos quadrados* (FIGURA 1), que consiste em determinar a reta que minimiza a soma dos quadrados dos desvios (ou erros) entre os verdadeiros valores das ordenadas e os obtidos a partir da reta que se pretende ajustar

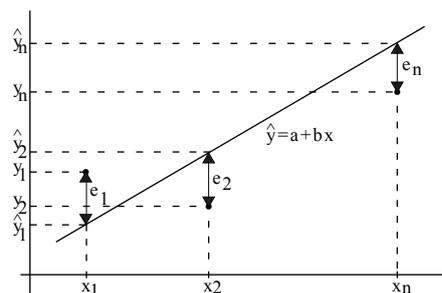


FIGURA 1. No método dos mínimos quadrados, minimiza-se a soma dos quadrados dos erros assinalados aqui pela distância segundo y entre o valor da ordenada do ponto dado e o obtido a partir da reta de regressão.

Esta técnica, embora muito simples, é pouco resistente, já que é muito sensível a dados “estranhos” - valores que se afastam da estrutura da maioria, normalmente designados por *outliers*. Efetivamente, quando se pretende minimizar

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

pode-se mostrar que os estimadores do declive e da ordenada da origem da reta de regressão são, respetivamente:

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{e} \quad a = \bar{y} - b\bar{x}$$

onde se representa por \bar{x} e \bar{y} as médias dos x_i 's e dos y_i 's. O facto de dependerem da média, que é uma medida não resistente, faz com que a recta de regressão seja também não resistente. Assim, é necessário proceder a uma análise prévia do diagrama de dispersão para ver se não existem alguns *outliers*. À reta de regressão obtida por este processo também se dá o nome de **reta dos mínimos quadrados**.

Pode-se mostrar que $r^2 = 1 - \frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{\sum_{i=0}^n (y_i - \bar{y})^2}$ onde r é o coeficiente de correlação amostral entre x e y .

Esta quantidade r^2 é o **coeficiente de determinação** e é referida como a quantidade de variabilidade dos dados explicada pelo modelo de regressão. Esta medida é normalmente utilizada como uma indicação da adequação do modelo de regressão ao conjunto de pontos inicialmente dado ², mas deve ser usada com precaução, pois nem sempre um valor de r^2 grande (próximo de 1) é sinal de que um modelo esteja a ajustar bem os dados. Do mesmo modo, um valor baixo de r^2 , pode ser provocado por um *outlier*, enquanto a maior parte dos dados se ajustam razoavelmente bem a uma reta ¹. Uma visualização prévia dos dados num diagrama de dispersão é fundamental.

Uma forma de verificar se o modelo ajustado é bom é através dos resíduos, isto é, das diferenças entre os valores observados y e os valores ajustados \hat{y} :

$$\text{resíduos} = \text{dados observados} - \text{valores ajustados}$$

pois se estes não se apresentarem muito grandes, nem com nenhum padrão bem determinado, é sintoma de que o modelo que estamos a ajustar é bom.

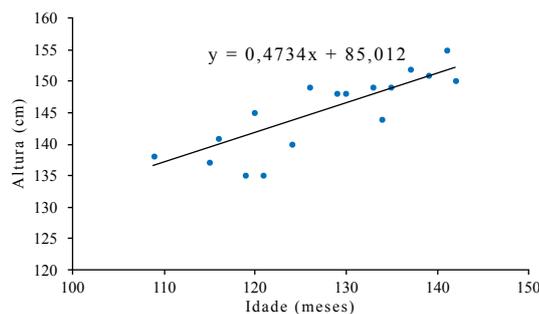
Nota

A reta de regressão é utilizada em predições, isto é, para predizer o valor de y , para um dado valor de x . No entanto estas predições não devem contemplar valores de x fora do intervalo dos x_i 's, uma vez que o facto de a reta se ajustar bem aos pontos dados não significa que sirva para fazer *extrapolações*.

Suponha que se recolheu o seguinte conjunto de dados referentes à idade (em meses) e à altura (em centímetros) de 18 crianças de uma escola:

Criança	Idade (meses)	Altura (cm)
1	109	138
2	113	145
3	115	137
4	116	141
5	119	135
6	120	145
7	121	135
8	124	140
9	126	149
10	129	148
11	130	148
12	133	149
13	134	144
14	135	149
15	137	152
16	139	151
17	141	155
18	142	150

O diagrama de dispersão dos dados sugere a existência de uma relação linear entre a idade e a altura, pelo que se vai ajustar aos dados uma reta dos mínimos quadrados, cuja equação está no gráfico seguinte (obtida no Excel):



O coeficiente de correlação é igual a 0,793, donde o coeficiente de determinação vem aproximadamente igual a 63% ($\approx 100 \times 0,793^2$)%, o que significa que a variabilidade que não é explicada pela reta de regressão anda à volta de 37% ($= 100 - 63$)%.

Se se tentar extrapolar a altura de um jovem com cerca de 17 anos (200 meses) obter-se-á uma altura de 180 cm e para um jovem adulto de cerca de 21 anos mais de 2 metros de altura, o que ilustra o problema referido na nota anterior.

REFERÊNCIAS

¹ DE VEAUX, R. D. *et al.*, *Intro stats*. Pearson Education Inc. ISBN 0-201-70910. 2004.

² GRAÇA MARTINS, M. E., *Introdução à Probabilidade e à Estatística. Com complementos de Excel*. Edição da SPE, ISBN-972-8890-03-6. Depósito Legal 228501/05. 2005.

³ MONTGOMERY, D. C. & RUNGER, G. C., *Applied statistics and probability for engineers*, 6th ed. John Wiley & Sons, Inc. ISBN 0-471-20454-4. 2003.