A regressão linear

Alfred Stadler
Departamento de Física da Universidade de Évora

6 de Outubro de 2009

Resumo

Apresentam-se as fórmulas que permitem calcular a regressão linear para um conjunto de pontos, incluíndo os erros associados aos parâmetros do modelo linear. Os cálculos necessários são ilustrados num exemplo concreto em todos os pormenores.

1 Como calcular a regressão linear

1.1 Os parâmetros da melhor recta e os seus erros

Em muitos casos sabe-se (ou suspeita-se) que a relação entre duas grandezas medidas é linear, ou pelo menos aproximadamente linear. Nestas circunstâncias é frequentemente importante estabelecer esta relação linear quantitativamente, no sentido que deve ser deduzida a equação do modelo linear (a recta) que aproxima de forma mais estreita os pontos experimentais. É isso o objectivo da regressão linear. O método mais utilizado para definir o significado de "a melhor recta" é o chamado "método dos mínimos quadrados". Neste método, a melhor recta é aquela que minimiza a soma dos quadrados das diferenças entre os pontos dados e os respectivos pontos calculados pelo modelo linear (pela equação da recta).

Veremos agora em mais pormenor como isso é feito. Suponhamos que um conjunto de N pontos (x_i, y_i) é dado. Procuramos os parâmetros a e b da recta

$$y(x) = a + bx\,, \tag{1}$$

de forma que a soma dos quadrados dos desvios, $\sum_{i=1}^{N} [y_i - y(x_i)]^2$, seja mínima. Supondo que os erros dos x_i sejam desprezáveis e os erros dos y_i todos iguais, a melhor estimativa destes parâmetros é

$$a = \frac{S_{x^2} S_y - S_x S_{xy}}{\Delta}\,, \quad b = \frac{N S_{xy} - S_x S_y}{\Delta}\,, \tag{2}$$

onde

$$S_x = \sum_{i=1}^{N} x_i\,, \qquad S_y = \sum_{i=1}^{N} y_i\,, \qquad S_{x^2} = \sum_{i=1}^{N} x_i^2\,, \tag{3}$$

$$S_{xy} = \sum_{i=1}^{N} x_i y_i\,, \qquad \Delta = N S_{x^2} - (S_x)^2\,. \tag{4}$$

Como já mencionado, supõe-se que os erros das grandezas medidas y_i sejam todos iguais, mas eles não têm de ser conhecidos. De facto, a própria regressão linear fornece uma estimativa desta incerteza, nomeadamente

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - a - bx_i)^2}\,. \tag{5}$$

Este erro σ_y é determinado a partir da dispersão dos pontos experimentais em torno da melhor recta. Se existir uma estimativa independente do erro dos y_i, então ela devia ser comparável com a estimativa (5).

Com isso, os erros dos parâmetros a e b são

$$\sigma_a = \sigma_y\sqrt{\frac{S_{x^2}}{\Delta}}\,, \qquad \sigma_b = \sigma_y\sqrt{\frac{N}{\Delta}}\,, \tag{6}$$

e o resultado da regressão linear pode ser escrito na forma $a \pm \sigma_a$ para a intersecção com o eixo dos y, e $b \pm \sigma_b$ para o declive da recta.

1.2 Um exemplo

Veremos agora num exemplo concreto como a regressão linear é calculada na prática.

Consideremos uma série de 10 medições da posição dum corpo que se movimenta ao longo do eixo dos x em função do tempo. Os resultados são apresentadas na tabela 1 e na figura 1.

Suspeitamos que o movimento seja uniforme. Por exemplo, podiamos ter a certeza que nenhuma força actua sobre o corpo na direcção x durante o período da observação. Neste caso, a posição do corpo devia ser representada pela equação

$$x(t) = x_0 + vt\,, \tag{7}$$

onde x_0 é a posição inicial (em $t = 0$ s), e v é a velocidade ao longo da direcção x. A nossa tarefa é determinar a posição inicial x_0 e a velocidade v do corpo.

Tempo [s]	Posição [m]
3.3	7.3
4.1	9.3
4.8	11.1
5.5	11.5
6.2	11.1
7.0	14.5
7.9	17.7
8.6	19.4
9.1	19.1
9.9	21.0

Tabela 1: Posição (em metros) dum corpo em movimento a uma dimensão medido em função do tempo (em segundos).

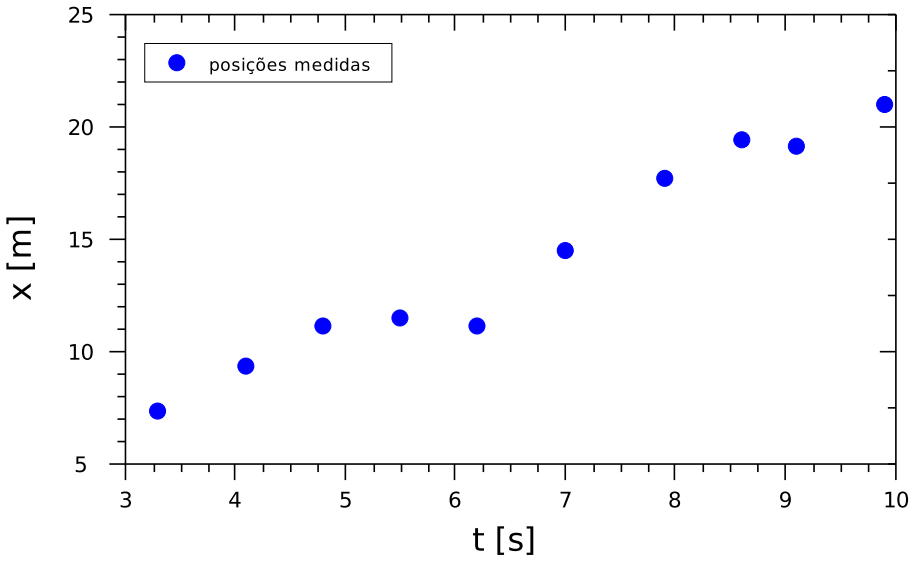

Figura 1: Posições medidas do corpo.

Para calcular x_0 e v aplicamos as regras da regressão linear da secção 1.1. Na aplicação das fórmulas temos de ter algum cuidado com a notação, porque no nosso caso t tem o papel da variável independente [x nas equações (1)-(6)] e a posição x assume o papél da variável dependente [y nas equações (1)-(6)].

Antes de começarmos os cálculos ainda um conselho geral: para minimizar a acumulação de erros de arredondamento, é importante que se façam os cálculos numéricos com mais algarismos do que são significativos, e que se efectue o devido arredondamento para os algarismos significativos apenas no resultado final.

Temos então de calcular as grandezas seguintes, com $N = 10$:

$$S_t = \sum_{i=1}^{10} t_i = (3.3 + 4.1 + \cdots + 9.9)\ \mathrm{s} = 66.4\ \mathrm{s} \tag{8}$$

$$S_{t^2} = \sum_{i=1}^{10} t_i^2 = (3.3^2 + 4.1^2 + \cdots + 9.9^2)\ \mathrm{s}^2 = 485.62\ \mathrm{s}^2 \tag{9}$$

$$S_x = \sum_{i=1}^{10} x_i = (7.3 + 9.3 + \cdots + 21.0)\ \mathrm{m} = 142.0\ \mathrm{m} \tag{10}$$

$$S_{tx} = \sum_{i=1}^{10} t_i x_i = (3.3 \times 7.3 + 4.1 \times 9.3 + \cdots + 9.9 \times 21.0)\ \mathrm{m \cdot s} = 1037.45\ \mathrm{m \cdot s} \tag{11}$$

$$\Delta = 10 S_{t^2} - (S_t)^2 = 10 \times 485.62\ \mathrm{s}^2 - (66.4\ \mathrm{s})^2 = 447.24\ \mathrm{s}^2\,. \tag{12}$$

Com estas grandezas auxiliares (repare que também eles têm unidades!) obte-

t_i [s]	x_i [m]	$x(t_i) = x_0 + vt_i$ [m]	$x_i - x_0 - vt_i$ [m]	$(x_i - x_0 - vt_i)^2$ [m^2]
3.3	7.3	7.13749	0.162512	0.0264102
4.1	9.3	8.82911	0.470893	0.22174
4.8	11.1	10.3093	0.790725	0.625247
5.5	11.5	11.7894	-0.289442	0.0837766
6.2	11.1	13.2696	-2.16961	4.7072
7.0	14.5	14.9612	-0.461229	0.212732
7.9	17.7	16.8643	0.835699	0.698393
8.6	19.4	18.3445	1.05553	1.11415
9.1	19.1	19.4017	-0.301731	0.0910414
9.9	21.0	21.0934	-0.0933503	0.00871428

Tabela 2: Elementos para o cálculo dos erros dos parâmetros da regressão linear do exemplo.

mos

$$x_0 = \frac{S_{t^2} S_x - S_t S_{tx}}{\Delta} = \frac{(485.62 \times 142.0 - 66.4 \times 1037.45)\ \mathrm{m \cdot s^2}}{447.24\ \mathrm{s^2}} = 0.159556\ \mathrm{m} \tag{13}$$

$$v = \frac{10 S_{tx} - S_t S_x}{\Delta} = \frac{(10 \times 1037.45 - 66.4 \times 142.0)\ \mathrm{m \cdot s}}{447.24\ \mathrm{s^2}} = 2.11452\ \mathrm{m/s}\,. \tag{14}$$

Agora os melhores parâmetros x_0 e v são conhecidos, e podemos avançar para calcular os erros.

O primeiro passo consiste no cálculo dos desvios entre as previsões do modelo linear $x(t_i) = x_0 + vt_i$ e os valores verdadeiramente medidos, x_i, para todos os i. Calculam-se também os quadrados dos desvios, que a seguir são somados. Isso é mostrado na tabela 2.

A soma dos elementos da última coluna da tabela leva à estimativa do desvio-padrão σ_x das medidas x_i,

$$\sigma_x = \sqrt{\frac{1}{8} \sum_{i=1}^{10} (x_i - x_0 - vt_i)^2} = 0.98675\ \mathrm{m} \tag{15}$$

Isto significa que o erro das medições da posição, x_i, é relativamente grande, da ordem de 1 m.

Substituir este resultado para σ_x em (6) dá os erros de x_0 e v:

$$\sigma_{x_0} = \sigma_x \sqrt{\frac{S_{t^2}}{\Delta}} = (0.98675\ \mathrm{m}) \sqrt{\frac{485.62\ \mathrm{s^2}}{447.24\ \mathrm{s^2}}} = 1.02822\ \mathrm{m} \tag{16}$$

$$\sigma_v = \sigma_x \sqrt{\frac{N}{\Delta}} = (0.98675\ \mathrm{m}) \sqrt{\frac{10}{447.24\ \mathrm{s^2}}} = 0.147549\ \mathrm{m/s}\,. \tag{17}$$

Com isso chegamos a

$$x_0 = (0.159556 \pm 1.02822)\ \mathrm{m}\,, \qquad v = (2.11452 \pm 0.147549)\ \mathrm{m/s}\,. \tag{18}$$

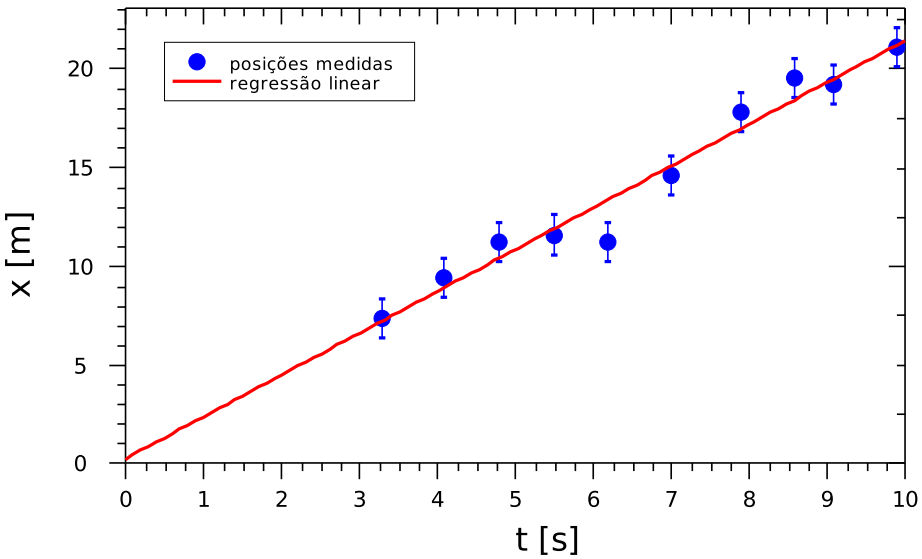

Figura 2: Resultado da regressão linear. As barras de erro nos dados medidos correspondem à estimativa obtida pela equação (15).

Falta ainda o arredondamento para os algarismos significativos. Porque o primeiro dígito é 1 em ambos os casos, ficamos com dois algarismos significativos nos erros. O resultado final é

$$\boxed{x_0 = (0.2 \pm 1.0)\ \mathrm{m}\,, \qquad v = (2.11 \pm 0.15)\ \mathrm{m/s}\,.} \tag{19}$$