A Geometria da Regressão Linear

Carlos Gomes
Escola Secundária de Amarante

CITAÇÃO
Gomes, C.(2020)
A Geometria da Regressão Linear,
Rev. Ciência Elem., V8(04):054.
doi.org/10.24927/rce2020.054

EDITOR
José Ferreira Gomes,
Universidade do Porto

EDITOR CONVIDADO
João Lopes dos Santos
Universidade do Porto

RECEBIDO EM
25 de abril de 2020

ACEITE EM
28 de abril de 2020

PUBLICADO EM

rce.casadasciencias.org

A regressão linear é um tema normalmente explorado (nas escolas) com recurso a uma calculadora científica gráfica ou software da moda (*GeoGebra* ou *Desmos*, por exemplo), ficando os estudantes com a tarefa aborrecida de introduzir números em listas e obter como recompensa uma equação que utilizam para fazer previsões num dado contexto. O que aqui se trata é de mostrar o grande valor didático deste problema, mobilizando conhecimentos que os alunos detêm para aclarar, do ponto de vista geométrico, o que está em causa em todo este processo que decorre nos "bastidores" da tecnologia.

A geometria do problema

O problema que consiste na determinação da reta que melhor se ajusta a uma dada nuvem de n pontos (x_i, y_i) é tradicionalmente tratado como o problema de encontrar os parâmetros a e b da equação $y = ax + b$ que minimizam a soma $S = \sum_{i=1}^{n} d_i^2$, em que os d_i são as diferenças entre os valores observados e os valores do modelo, isto é, $d_i = y_i - ax - b$.

Sejam (x_1, y_1), (x_2, y_2),..., (x_n, y_n) os dados observados (nuvem de pontos na FIGURA 1). Para a determinação do parâmetro a (declive da reta), seria "simpático" que a nuvem tivesse o seu centro de massa na origem do referencial, isto é, no ponto de coordenadas (0; 0). Isto porque libertar-nos-íamos do parâmetro b da equação da reta, o que parece reduzir a dificuldade do problema, pois, nesta condições, o modelo associado à reta de regressão seria $y = ax$. Para fazer com que o centro de massa da nuvem se desloque para a origem, é suficiente efetuarmos uma translação de toda a nuvem de pontos segundo o vetor $(-\bar{x}, -\bar{y})$, ou seja, basta subtrairmos o centro de massa $(\bar{x}, \bar{y})$ a todos os pontos da nuvem. Obtém-se assim uma nova nuvem de pontos da forma $(x_i - \bar{x}, y_i - \bar{y})$ cujo centro de massa é (0; 0).

Fazendo $x_i - \bar{x} = \tilde{x}_i$ e $y_i - \bar{y} = \tilde{y}_i$, a nuvem sobre a qual o trabalho prossegue será $(\tilde{x}_i, \tilde{y}_i)$, com $i = 1, 2, ..., n$, cuja reta de regressão tem o mesmo declive que a reta de regressão da nuvem original, em consequência da translação efetuada.

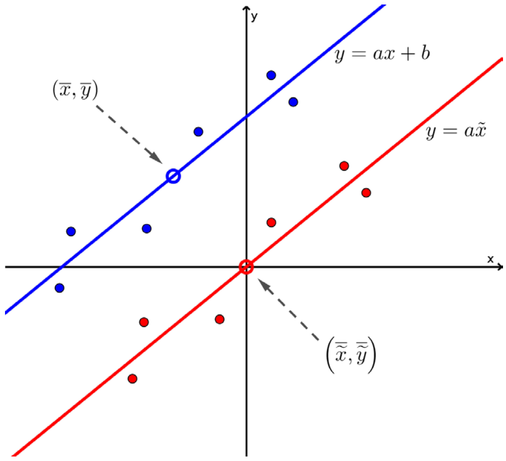

FIGURA 1. Translação da nuvem de pontos.

A nova nuvem é constituída por pontos da forma $(\tilde{x}_i, \tilde{y}_i)$ e os pontos da forma $(\tilde{x}_i, a\tilde{x}_i)\,, i = 1, 2, ..., n$, são os pontos sobre a reta $\tilde{y} = a\tilde{x}$, que coincidiriam com os primeiros caso a correlação fosse perfeita. Os n vetores $\vec{u}_i = (\tilde{x}_i, a\tilde{x}_i)$ determinados por estes pontos são colineares. Mas aqui, uma mudança de dimensão vai tornar o trabalho mais simples: em vez de considerarmos estes n vetores de dimensão 2, utilizamos os dados organizados em vetores de dimensão n:

$$\vec{i} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)\,,$$
$$\vec{j} = (a\tilde{x}_1, a\tilde{x}_2, ..., a\tilde{x}_n)\,,$$

e

$$\vec{u} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n)\,.$$

Os vetores $\vec{i}$ e $\vec{j}$ são colineares:

$$\begin{aligned} \vec{j} &= (a\tilde{x}_1, a\tilde{x}_2, ..., a\tilde{x}_n) \\ &= a\,(\tilde{x}_1\tilde{x}_2, ..., \tilde{x}_n) \\ &= a\vec{i}. \end{aligned} \tag{1}$$

Para além do mais, o escalar a em (1) é precisamente o declive da reta procurada! Assim, determinar a será equivalente a determinar (algo sobre) $\vec{j}$, agora num espaço de dimensão n, (veja-se o apêndice da versão eletrónica para clarificação deste ponto).

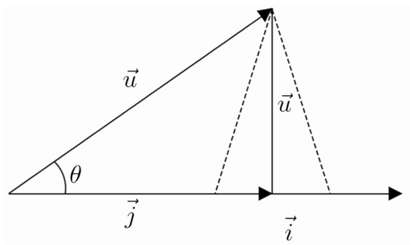

FIGURA 2. Vetores num espaço de dimensão n.

Repare-se que $\vec{u}-\vec{j}=(\tilde{y}_1-a\tilde{x}_1,...,\tilde{y}_n-a\tilde{x}_n)$ não é mais do que o vetor dos resíduos, isto é, o vetor cujas componentes são as diferenças entre os dados observados e os dados teóricos da nova nuvem. Ora, o que se pretende é que a norma (ou distância) $\left\|\vec{u}-\vec{j}\right\|$ seja mínima. Isto só acontecerá se $\vec{u}-\vec{j}$ for normal a $\vec{i}$ (como sugere a FIGURA 2). Para que tal aconteça, $\vec{j}$ tem de ser a projeção de $\vec{u}$ sobre $\vec{i}$. Logo, o produto escalar de $\vec{u}-\vec{j}$ com $\vec{i}$ tem de ser nulo, retirando-se desta condição o valor do multiplicador a, declive da reta de regressão:

$$\begin{aligned}&\left(\vec{u}-\vec{j}\right)\cdot\vec{i}=0\\ \Leftrightarrow\ &\left(\vec{u}-a\vec{i}\right)\cdot\vec{i}=0 \quad \left(\vec{j}=a\vec{i}, de\,(1)\right)\\ \Leftrightarrow\ &\vec{u}\cdot\vec{i}-a\vec{i}\cdot\vec{i}=0\\ \Leftrightarrow\ &a=\frac{\vec{u}\cdot\vec{i}}{\left\|\vec{i}\right\|} \quad \left(\vec{i}\cdot\vec{i}=\left\|\vec{i}\right\|^2\right).\end{aligned} \tag{2}$$

Depois de se calcular a através de (2), a determinação do parâmetro b é um simples exercício: dado que $(\bar{x},\bar{y})$ pertence à reta procurada, ele terá de satisfazer a condição $y=ax+b$. Daqui se retira que $b=\bar{y}-a\bar{x}$.

Exemplos de aplicação

Exemplo 1

Vejamos a aplicação destes resultados a um exercício típico de um manual escolar.

Existirá alguma relação entre a temperatura e a quantidade de chuva que cai em Amarante? Para responder a esta pergunta vamos comparar num gráfico de correlação as temperaturas médias (ºC) dos vários meses do ano com a pluviosidade média (mm).

TABELA 1. Valores de temperatura e pluviosidade; à esquerda, dados originais, à direira dados transladados.

Temperatura	Pluviosidade
11.3	122
12.0	108
13.5	101
15.2	54
17.6	44
20.0	22
22.2	4
22.5	6
21.3	29
18.3	80
14.2	102
11.6	107

Temperatura $(\vec{i})$	Pluviosidade $(\vec{u})$
-5.3417	57.0833
-4.6417	43.0833
-3.1417	36.0833
-1.4417	-10.917
0.9583	-20.9167
3.3583	-42.9167
5.5583	-60.9167
5.8583	-58.9167
4.6583	-35.9167
1.6583	15.08333
-2.4417	37.08333
-5.0417	42.08333

Neste exemplo, a tabela da esquerda é dada e a da direita foi calculada por nós. O centróide da nuvem de pontos é $(\bar{x}, \bar{y}) = (16.6417, 64.9167)$. Os vetores $\vec{u}$ e $\vec{i}$ são as colunas da tabela da direita, depois de efetuada a translação da nuvem original: são vetores num espaço de dimensão 12.

De acordo com as conclusões da secção anterior, os parâmetros da equação da reta de regressão $y = ax + b$ podem ser calculados do seguinte modo:

$$\begin{aligned} a &= \frac{\vec{u} \cdot \vec{i}}{\left\|\vec{i}\right\|^2} \\ &\approx \frac{-1895.4583}{195.2692} \\ &\approx -9.7069, \\ b &= \bar{y} - a\bar{x} \\ &\approx 64.9167 + 9.7069 \times 16.6417 \\ &\approx 226.4557. \end{aligned}$$

Assim, $y \approx -9.7069x + 226.4557$ será a equação da reta de regressão e, com ela, podemos fazer estimativas no contexto do problema.

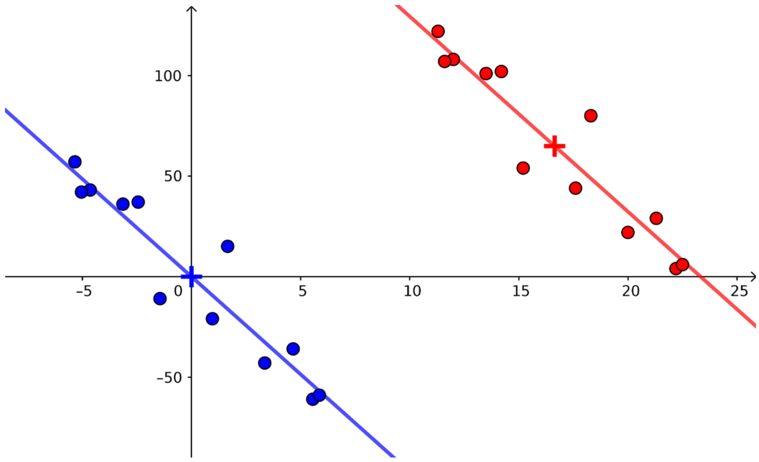

FIGURA 3. Retas de ajuste a dados de temperatura e pluviosidade.

Note-se que o produto escalar de dois vectores de dimensão n não é mais do que a soma dos produtos das correspondentes componentes desses vectores (uma generalização do que se faz para $n = 2$ ou $n = 3$, na disciplina de Matemática A no Ensino Secundário), ou seja, se $\vec{a} = (a_1, a_2, ..., a_n)$ e $\vec{b} = (b_1, b_2, ..., b_n)$,

$$\vec{a} \cdot \vec{b} = a_1 \times b_1 + a_2 \times b_2 + \cdots + a_n \times b_n = \sum_{i=1}^{n} a_i \times b_i$$

Também a norma de um vector de dimensão n é uma generalização da norma de vetores em 2 e 3 dimensões, isto é,

$$\|\vec{a}\| = \sqrt{{a_1}^2 + {a_2}^2 + \cdots + {a_n}^2} = \sqrt{\sum_{i=1}^{n} {a_i}^2}$$

assim, no presente exemplo, $\vec{u} \cdot \vec{i}$ corresponde a efectuar a soma dos produtos dos elementos correspondentes de cada linha da tabela da direita.

Exemplo 2

Neste exemplo, aplicaremos os conceitos anteriores à construção de um modelo linear do número de infetados pelo novo coronavírus em função do tempo decorrido no período de 8 a 31 de maio. Aqui, o centro de massa é dado pelas coordenadas do ponto $(\bar{x}, \bar{y}) = (11.5, 29648.583)$ e os vetores $\vec{i}$ e $\vec{u}$ habitam um espaço de dimensão 24 (colunas da tabela da direita).

TABELA 2. Total de infectados em função dos dias; à esquerda dados originais; à direita dados transladados.

Nº de dias	Nº de infetados
67	27268
68	27406
69	27581
70	27679
71	27913
...	...
87	31596
88	31946
89	32203
90	32500

Nº de dias $\vec{i}$	Nº de infetados $\vec{u}$
-11.5	-2380.583
-10.5	-2242.583
-9.5	-2067.583
-8.5	-1969.583
-7.5	-1735.583
...	...
8.5	1947.417
9.5	2297.417
10.5	2554.417
11.5	2851.417

O produto escalar é $\vec{u} \cdot \vec{i} \simeq 261980$ (soma dos produtos dos elementos de cada linha da tabela de baixo). O quadrado da norma do vetor $\vec{i}$ (quadrância de $\vec{i}$) é $\|i\|^2 = 1150$.

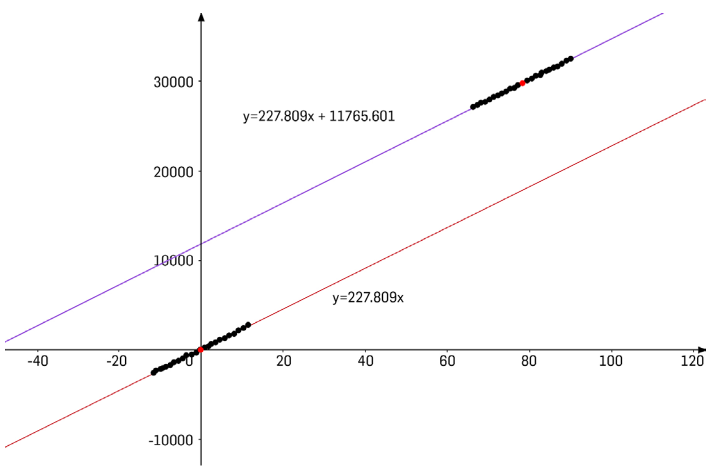

FIGURA 4. Análise de dados de infetados com modelo linear.

Assim, com $a = \dfrac{261980}{1150} \simeq 227.809$ e $b = \bar{y} - a\bar{x} \simeq 11765.601$, obtemos a equação da reta mostrada na figura acima.

O leitor pode criar uma lição no *Geogebra Classroom* com este exemplo, seguindo para https://www.geogebra.org/m/ncpffvne.

Coeficiente de correlação linear

O *coeficiente de correlação* é uma medida que pretende determinar o grau de alinhamento dos dados. Sobre ele costumam ser colocadas duas questões:

- Por que razão varia no intervalo $[-1, 1]$*?*

- Por que razão a correlação entre as variáveis é tanto mais forte quanto mais próximo de -1 *ou de* 1 *se encontra o coeficiente? Não seria razoável pensarmos que quanto mais próximo de zero mais forte será a correlação, uma vez que ele mede o grau de proximidade dos dados em relação à reta?!*

Repare-se que o coeficiente de correlação, sendo uma medida do alinhamento dos dados, deve estar relacionado com o "grau de colinearidade" entre os vetores $\vec{u}$ e $\vec{i}$, referentes aos dados transladados (note que a correlação não depende da nuvem que se considera, uma vez que a operação de translação efetuada à nuvem inicial garante a manutenção das relações entre os dados observados e os teóricos). E uma forma natural de medir este "grau de colinearidade" é estudando o ângulo θ que $\vec{u}$ e $\vec{i}$ formam entre si (ver FIGURA 2). (Note que em tudo o que se segue se pode substituir a unidade grau por rad.). Assim, θ poderia ser usado com legitimidade como medida do grau de alinhamento dos dados, ou seja, como coeficiente de correlação. O diagrama da FIGURA 5 resume a variação deste coeficiente de correlação.

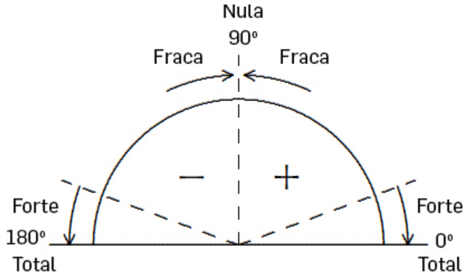

FIGURA 5. Coeficiente de correlação θ.

Visto que $cos\theta = \dfrac{\vec{u} \cdot \vec{i}}{\|\vec{u}\| \left\|\vec{i}\right\|}$, θ pode ser obtido através de

$$\theta = \arccos\left(\frac{\vec{u} \cdot \vec{i}}{\|\vec{u}\| \left\|\vec{i}\right\|}\right). \tag{3}$$

No exemplo 1 da secção anterior, o coeficiente de correlação θ é

$\theta =$ arcos $\left(\dfrac{\vec{u} \cdot \vec{i}}{\|\vec{u}\| \left\|\vec{i}\right\|}\right) =$ arcos $\left(\dfrac{-1895.4583}{143.7391 \times 13.9739}\right) = 160.68^o$ (forte Negativa?).

e no segundo exemplo, $\theta = \arccos\left(\frac{2,61980}{2,62579.265}\right) = \arccos(0.998) \simeq 3.62^o$ (Muito forte, positiva?).

No entanto, na literatura sobre o assunto, θ é convenientemente substituído pelo seu cosseno (porquê?), e assim se compreende a sua variação tal como encontramos nos manuais:

$$0^o \leq \theta \leq 180^o \Rightarrow -1 \leq \cos\theta \leq 1 \Leftrightarrow -1 \leq \frac{\vec{u} \cdot \vec{i}}{\|\vec{u}\| \left\|\vec{i}\right\|} \leq 1.$$

Uma fórmula que normalmente acompanha os manuais para determinar o valor do coeficiente de correlação, r, é

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \tag{4}$$

Sendo (4) equivalente a

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

fica estabelecida a igualdade

$$r = \frac{\vec{u} \cdot \vec{i}}{\|\vec{u}\| \left\|\vec{i}\right\|} = \cos\theta$$

Apêndice

A interpretação geométrica que se explora neste texto tem como elemento essencial a translação da nuvem de pontos original para uma nuvem de pontos com centro de massa na origem do referencial. Esta operação faz com que os dados transladados cumpram

$$\sum_{i=1}^{n} \tilde{x}_i = 0 \quad \text{e} \quad \sum_{i=1}^{n} \tilde{y}_i = 0 \ .$$

Reescrevendo estas condições, ficamos com

$$\begin{aligned}
\sum_{i=1}^{n} \tilde{x}_i = 0 &\Leftrightarrow 1 \times \tilde{x}_1 + 1 \times \tilde{x}_2 + \cdots + 1 \times \tilde{x}_n = 0 \\
&\Leftrightarrow \vec{w} \cdot \vec{i} = 0 \\
\sum_{i=1}^{n} \tilde{y}_i = 0 &\Leftrightarrow 1 \times \tilde{y}_1 + 1 \times \tilde{y}_2 + \cdots + 1 \times \tilde{y}_n = 0 \\
&\Leftrightarrow \vec{w} \cdot \vec{u} = 0
\end{aligned}$$

que, do ponto de vista geométrico, permitem afirmar que os vectores $\vec{i}$ e $\vec{u}$ (e, consequentemente, $\vec{j}$) são perpendiculares ao vector unitário $w = (1, 1, \cdots, 1)$. Assim, $\vec{i}$, $\vec{j}$ e $\vec{u}$ habitam o hiperplano de dimensão $n-1$, normal ao vector unitário $\vec{w}$. Este facto não altera a argumentação seguida pois no hiperplano de dimensão $n-1$ continuamos a querer reduzir ao mínimo a norma de $\vec{u}-\vec{j}$ e a condição continua a ser a ortogonalidade deste vector a $\vec{j}$.

No caso em que a amostra observada é constituída apenas por dois pontos, $\vec{i}$, $\vec{j}$ e $\vec{u}$ são colineares e a correlação é perfeita, como seria de esperar. Para a situação em que $n = 3$, pode manipular e descarregar a animação GeoGebra em https://www.geogebra.org/m/muxygsbz.

Conclusão

Ao longo dos anos, o tema da regressão linear tem sido tratado nas nossas escolas, quase exclusivamente, como uma manipulação de fórmulas, à qual a tecnologia veio retirar algum desse desprazer salvando, por um lado, os alunos dos cálculos fastidiosos, mas atirando-os, por outro, para uma cegueira determinada pela calculadora gráfica. O que aqui se quis mostrar foi que essas abordagens tradicionais ao tema podem, com enormes vantagens, serem substituídas por uma abordagem geométrica sólida, coerente e palpável, em que a única novidade (mas não surpresa) reside na generalização de conceitos de geometria analítica a espaços de dimensão superior a três. Para além disso, abre também espaço à compreensão dos “bastidores” da calculadora gráfica, permitindo que os alunos olhem para ela como uma biblioteca de algoritmos que podem compreender e até criar.

REFERÊNCIAS

[1] SIMON, S., *http://www.pmean.com/10/LeastSquares.html*. 2019.
[2] SALAS, J.M., *Elementos de Matematicas*, 6.a edición, págs 177-190.
[3] RIBEIRO, H., *et al.*, *A Regressão Linear Simples no Ensino Secundário*. *Gazeta de Matemática da SPM*, no 168, pág. no 42. 2012.
[4] MARTINS, M., *Regressão Linear Simples*. 2019.