A GEOMETRIA DA REGRESSÃO LINEAR

Carlos Gomes
Escola Secundária de Amarante
carlosgomes@esamarante.edu.pt

A regressão linear é um tema normalmente explorado (nas escolas) com recurso a uma calculadora científica gráfica ou a um *software* da moda (GeoGebra, 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.

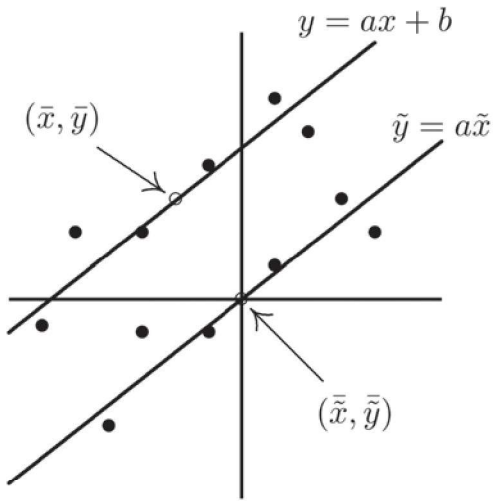

Figura 1. Translação da nuvem de pontos.

1. 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 teóricos, isto é, $d_i = y_i - ax_i - b$ (veja-se [3]).

Sejam (x_1, y_1), (x_2, y_2), $\ldots$, (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 $(-\overline{x}, -\overline{y})$, ou seja, basta subtrairmos o centro de gravidade $(\overline{x}, \overline{y})$ a todos os pontos da nuvem. Obtém-se assim uma nova nuvem de pontos da forma $(x_i - \overline{x}, y_i - \overline{y})$ cujo centro de gravidade é $(0, 0)$.

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

A nova nuvem é constituída por pontos da forma $(\tilde{x}_i, \tilde{y}_i)$ e os pontos da forma $(\tilde{x}_i, a\tilde{x}_i)$, $i = 1, 2, \ldots, n$, são os pontos sobre a reta $\tilde{y} = a\tilde{x}$, que coincidiriam com os primeiros caso a correlação fosse perfeita. Os n vectores $\vec{u}_i = (\tilde{x}_i, a\tilde{x}_i)$ 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**:

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

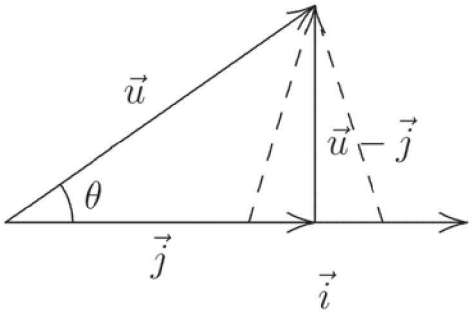

Figura 2. Vetores num espaço de dimensão *n*.

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

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

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***.

Repare-se que $\vec{u}-\vec{j} = (\tilde{y}_1 - a\tilde{x}_1, \ldots, \tilde{y}_n - a\tilde{x}_n)$ 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) $\|\vec{u}-\vec{j}\|$ 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:

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

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 $b = \bar{y} - a\bar{x}$.

Apesar de querermos focar-nos nos aspetos marcadamente geométricos do problema, vale a pena notar aqui que o resultado (2) pode ainda ser obtido pela combinação da geometria analítica com a aplicação das derivadas a problemas de otimização (assuntos tratados no 11.º ano, antes da regressão linear): depois da translação dos dados, o objetivo é o de minimizar a soma

$$S = \sum_{i=1}^{n} e_i^2,\ (e_i = \tilde{y}_i - a\tilde{x}_i),$$

como aparece no problema original. Assim,

$$\begin{aligned}\frac{dS}{da} &= \sum_{i=1}^{n} \frac{dS}{de_i}\frac{de_i}{da} = \sum_{i=1}^{n} 2e_i\frac{de_i}{da}\\ &= \sum_{i=1}^{n} 2e_i(-\tilde{x}_i) = \sum_{i=1}^{n} 2(\tilde{y}_i - a\tilde{x}_i)(-\tilde{x}_i)\\ &= -2\left(\vec{u}-\vec{j}\right)\cdot\vec{i}. \end{aligned}$$

Segue-se (como sugerido geometricamente), $\left(\vec{u}-\vec{j}\right)\cdot\vec{i} = 0$, o que leva a (2).

2. EXEMPLO DE APLICAÇÃO

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).*

Neste exemplo, a tabela da esquerda é dada e a da direita foi calculada por nós. O centroide da nuvem de pontos é

Tabela 1.

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

Tabela 2.

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

$(\bar{x}, \bar{y}) = (16.6417, 64.9167)$. 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:

$$\begin{aligned} a &= \frac{\vec{u} \cdot \vec{i}}{\|\vec{i}\|^2} \\ &\approx \frac{-1895.4583}{195.2692} \\ &\approx -9.7069\,, \end{aligned}$$

$$\begin{aligned} b &= \bar{y} - a\bar{x} \\ &\approx 64.9167 + 9.7069 \times 16.6417 \\ &\approx 226.4557\,. \end{aligned}$$

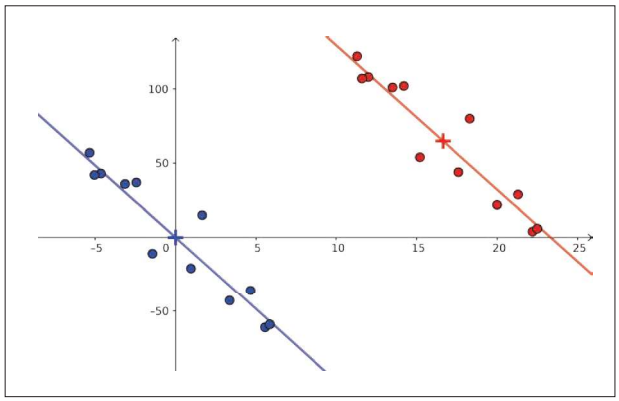

Figura 3. Translação da nuvem de pontos e centros de massa.

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

3. 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[1]. 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).[2] 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 4 resume a variação deste coeficiente de correlação.

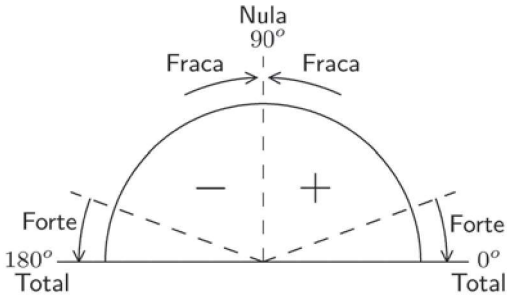

Figura 4. Coeficiente de correlação θ.

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

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

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

$$\begin{aligned} \theta &= \arccos\left(\frac{\vec{u} \cdot \vec{i}}{\|\vec{u}\|\,\|\vec{i}\|}\right) \\ &= \arccos\left(\frac{-1895.4583}{143.7391 \times 13.9739}\right) \\ &= 160.68^o\ (forte?). \end{aligned}$$

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:

[1] 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.

[2] Em tudo o que se segue pode-se substituir a unidade *grau* por *rad*.

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

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

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

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

$$r = \frac{\vec{u}\cdot\vec{i}}{\|\vec{u}\|\,\|\vec{i}\|} = \cos\theta\,.$$

4. 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, ser 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. 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] Steve Simon *http://www.pmean.com/10/LeastSquares.html*, visualizado em 15.08.2019.

[2] José Martínez Salas. *Elementos de Matematicas*, 6.ª edição, págs 177-190.

[3] Helena Ribeiro, Maria Alice Martins, Rui Santos. "A regressão linear simples no ensino secundário". *Gazeta de Matemática da SPM*, n.º 168, pág. n.º 42, novembro 2012.

SOBRE O AUTOR

Carlos Alberto da Silva Gomes. Licenciado em Ensino de Matemática pela Universidade do Minho. Professor de matemática na Escola Secundária de Amarante desde 2000.