La droite des moindres carrés par l'algèbre élémentaire

Maxime Zuber, Haute École Pédagogique BEJUNE, maxime.zuber@hep-bejune.ch

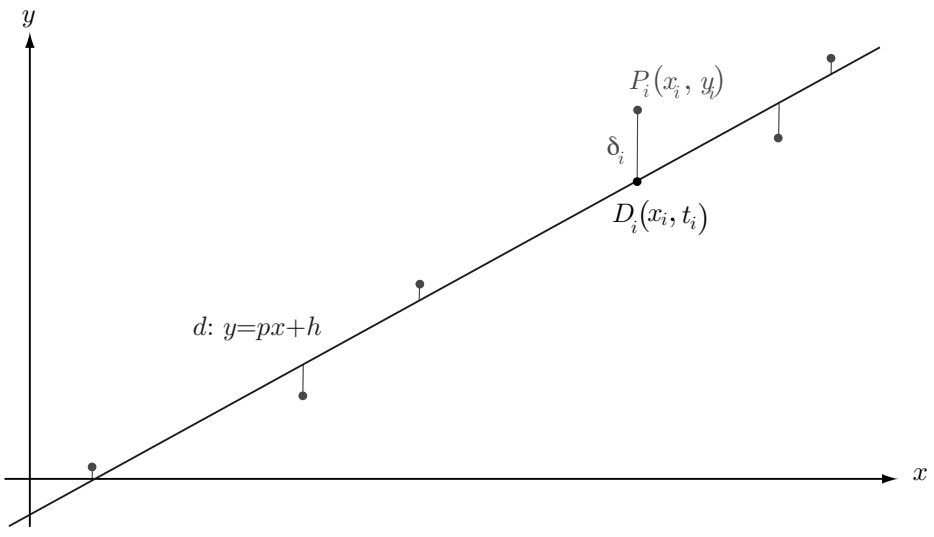
1 Introduction

Les enseignants qui, dans le cadre d'un cours de mathématiques appliquées aux sciences expérimentales ou à l'économie, souhaitent introduire la régression linéaire, s'en trouvent empêchés. Ils se considèrent à tort contraints d'enseigner des recettes dont les bases théoriques relèvent d'un niveau supérieur. Or, il n'en est rien. Nul n'est besoin en effet de faire appel au calcul différentiel, aux fonctions à deux variables ou à l'algèbre linéaire [1] pour établir l'équation de la droite des moindres carrés et évoquer la notion de corrélation. Les propriétés élémentaires de la parabole suffisent. C'est ce que nous proposons d'exposer ici. Les éléments théoriques présentés étant éminemment connus, le présent article ne prétend à aucune originalité quant au fond. Son intérêt réside dans la simplicité didactique de la démarche proposée.

Un nuage de n points $P_i(x_i; y_i)$ du plan devrait, en théorie, se situer exactement sur une droite. Les coordonnées y_i étant entachées d'erreur (elles représentent par exemple les mesures d'une expérience), les n points ne sont pas alignés. On cherche la droite d d'équation cartésienne $y = p \cdot x + h$ qui minimise la somme des carrés des distances verticales entre les points donnés $P_i(x_i; y_i)$ et les points théoriques $D_i(x_i; px_i + h)$ (cf. dessin ci-après), c'est-à-dire, la somme

$$E = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (t_i - y_i)^2 = \sum_{i=1}^{n} (px_i + h - y_i)^2.$$

Cette droite est appelée *droite des moindres carrés*.



Soit le barycentre $G(\bar{x}; \bar{y})$ de ce nuage, c'est-à-dire le point ayant les coordonnées moyennes $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. Nous allons démontrer que la droite d contient le point G et que la pente p et l'ordonnée

à l'origine h sont données par

$$\boxed{p = \frac{\mathrm{Cov}(x\,,y)}{\mathrm{Var}(x)} = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum(x_i-\bar{x})^2}}$$

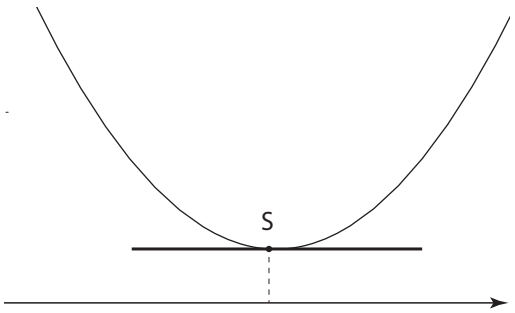
$$\boxed{h = \bar{y} - p\cdot\bar{x} = \bar{y} - \frac{\mathrm{Cov}(x\,,y)}{\mathrm{Var}(x)}\cdot\bar{x}}$$

(Dans ces formules et celles qui suivent, le symbole $\sum$ dénote une sommation pour l'indice i allant de 1 à n).

2 Démonstration

2.1 Lemme algébrique banal

La parabole d'équation $y = ax^2 + bx + c$ a un sommet S d'abscisse $x = -\frac{b}{2a}$. Si $a > 0$, le point S correspond à un minimum de la fonction quadratique.



Ce résultat ne requiert pas de technique d'analyse. Il suffit d'observer que $y = \underbrace{a\cdot\left(x+\frac{b}{2a}\right)^2}_{\geq 0} - \frac{b^2}{4a} + c$ est minimal quand le terme positif $a\cdot\left(x+\frac{b}{2a}\right)^2$ est nul, c'est-à-dire quand $x = -\frac{b}{2a}$.

2.2 Les paramètres de la droite à partir du lemme

La droite optimale d ayant pour équation $y = p\cdot x + h$, toute autre droite conduit à une somme $\sum_{i=1}^{n}\delta_i^2$ supérieure à la valeur correspondante à d. Prouvons maintenant que la droite d contient le point G. Il se trouve que la droite de pente p a une ordonnée à l'origine h qui rend minimale la somme

$$\begin{aligned}
S(h) &= \sum \delta_i^2 \\
&= \sum (px_i + h - y_i)^2 \\
&= \sum (h + px_i - y_i)^2 \\
&= \sum \left[h^2 + 2(px_i - y_i)h + (px_i - y_i)^2\right] \\
&= \underbrace{n}_{a}\cdot h^2 + h\cdot \underbrace{2\sum(px_i - y_i)}_{b} + \cdots
\end{aligned}$$

Cette expression quadratique est minimale (cf. lemme) pour

$$h = -\frac{b}{2a} = -\frac{2\sum(px_i - y_i)}{2n} = \frac{1}{n}\sum y_i - p \cdot \frac{1}{n}\sum x_i = \bar{y} - p \cdot \bar{x}$$

Ainsi l'équation de d s'écrit

$$y = p \cdot x + \bar{y} - p \cdot \bar{x}$$

ce qui prouve qu'elle passe bien par le barycentre $G(\bar{x};\bar{y})$.

La droite optimale ayant pour équation $y = p \cdot x + \bar{y} - p \cdot \bar{x}$, sa pente p rend forcément minimale l'expression quadratique

$$\begin{aligned} Q(p) &= \sum \delta_i^2 \\ &= \sum (px_i + \bar{y} - p\bar{x} - y_i)^2 \\ &= \sum \left[(x_i - \bar{x}) \cdot p - (y_i - \bar{y})\right]^2 \\ &= \sum \left[(x_i - \bar{x})^2 \cdot p^2 - 2(x_i - \bar{x})(y_i - \bar{y})p + \cdots\right] \\ &= \underbrace{\left[\sum (x_i - \bar{x})^2\right]}_{a} \cdot p^2 + p \cdot \underbrace{(-2)\sum (x_i - \bar{x})(y_i - \bar{y})}_{b} + \cdots \end{aligned}$$

Ce qui signifie (cf. lemme) que

$$p = -\frac{b}{2a} = \frac{2\sum(x_i - \bar{x})(y_i - \bar{y})}{2\sum(x_i - \bar{x})^2} \cdot \frac{1/n}{1/n} = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x\,,y)}{\mathrm{Var}(x)}$$

Il se trouve en effet que $\mathrm{Var}(x) = \frac{1}{n}\sum(x_i - \bar{x})^2$ est appelée *variance* de la variable x et $\mathrm{Cov}(x\,,y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$, *covariance* des variables x, y. Enfin

$$h = \bar{y} - p \cdot \bar{x} = \bar{y} - \frac{\mathrm{Cov}(x\,,y)}{\mathrm{Var}(x)} \cdot \bar{x}$$

3 Corrélation

3.1 La droite de régression de x et y

La droite d'ajustement $d : y = p \cdot x + h$ exprimant y en fonction de x minimise la somme des carrés des écarts verticaux entre les points donnés et les points de mêmes abscisses situés sur la droite. Cette droite est appelée *droite de régression de y en x*. Si on permute les rôles respectifs de x et y, on obtient une relation $x = p'y + h'$ caractérisant la droite d' qui minimise la somme des carrés des écarts horizontaux séparant les points donnés de ceux de mêmes ordonnées sur la droite. Une telle droite est appelée *droite de régression de x en y*. On détermine son équation

$$d' \quad : \quad x = p'y + h'$$

de la même manière que celle appliquée pour la première droite mais en échangeant les rôles de x et y.

3.2 Le coefficient de la corrélation

Si les points sont parfaitement alignés, alors les droites d et d' sont confondues. Dans ce cas, leurs pentes p et $1/p'$ étant égales, on a alors $p \cdot p' = 1$. En général, les points mesurés ne sont toutefois pas

alignés ; la corrélation sera alors d'autant plus forte que les deux droites de régression sont proches. Pour mesurer la corrélation, on introduit le *coefficient de corrélation linéaire* r défini comme étant la moyenne géométrique des «pentes» p et p'

$$\boxed{r = \pm\sqrt{p \cdot p'}}$$

Ce nombre, qui a le même signe que p et p', est toujours compris entre -1 et 1. Il mesure le degré d'alignement des points et est donc tel que $r = \pm 1$ si les points sont parfaitement alignés. Si $r = 0$, il n'y a aucune dépendance *linéaire* entre x et y. Comme $p = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$ et $p' = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(y)}$, on a donc

$$\boxed{r = \frac{\mathrm{Cov}(x,y)}{S(x) \cdot S(y)}}$$

relation dans laquelle $S(x) = \sqrt{\mathrm{Var}(x)}$ et $S(y) = \sqrt{\mathrm{Var}(y)}$ désignent les écarts-types de x et y.

4 Les équations normales

4.1 Démarche analytique

On obtient d'ordinaire la pente p et l'ordonnée h comme étant le couple $(p;h)$ correspondant au minimum de la fonction de deux variables

$$E(p;h) = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (px_i + h - y_i)^2$$

c'est-à-dire le point pour lequel les dérivées partielles $\frac{\partial E}{\partial h}$ et $\frac{\partial E}{\partial p}$ s'annulent. Les conditions $\frac{\partial E}{\partial h} = 0$ et $\frac{\partial E}{\partial p} = 0$ conduisent au système

$$\begin{cases} \dfrac{\partial E}{\partial h} = \sum 2(px_i + h - y_i) \cdot 1 &= 0 \\ \dfrac{\partial E}{\partial p} = \sum 2(px_i + h - y_i) \cdot x_i &= 0 \end{cases}$$

qui s'écrit aussi

$$\begin{cases} p \cdot \sum x_i + h \cdot \sum 1 &= \sum y_i \\ p \cdot \sum x_i^2 + h \cdot \sum x_i &= \sum x_i y_i \end{cases}$$

ou encore

$$\boxed{(S) : \begin{cases} p \cdot \sum x_i + n \cdot h &= \sum y_i \quad | \ (1) \\ p \cdot \sum x_i^2 + h \cdot \sum x_i &= \sum x_i y_i \ | \ (2) \end{cases}}$$

Les relations (1) et (2) sont appelées *équations normales*.

4.2 Démarche algébrique

On peut retrouver ces relations par une démarche purement algébrique.

Nous avons établi que

$$h = \bar{y} - p \cdot \bar{x}$$

Multipliée par n, cette relation s'écrit aussi

$$\begin{aligned} n \cdot h &= n \cdot \bar{y} - p \cdot n \cdot \bar{x} \\ n \cdot h &= \sum y_i - p \cdot \sum x_i \end{aligned}$$

ou encore

$$\boxed{p \cdot \sum x_i + n \cdot h = \sum y_i}$$

qui n'est autre que la première équation normale.

Quant à la relation $p \cdot \mathrm{Var}(x) = \mathrm{Cov}(x, y)$, multipliée par n, elle s'écrit aussi

$$\begin{aligned} p \cdot \sum (x_i - \bar{x})^2 &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ p \sum x_i^2 - 2p\bar{x} \sum x_i + np\bar{x}^2 &= \sum x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i + n\bar{x}\bar{y} \\ p \sum x_i^2 - 2pn\bar{x}^2 + np\bar{x}^2 &= \sum x_i y_i - \bar{x} n \bar{y} - \bar{y} n \bar{x} + n\bar{x}\bar{y} \\ p \sum x_i^2 - np\bar{x}^2 &= \sum x_i y_i - n\bar{x}\bar{y} \\ p \sum x_i^2 + \underbrace{(\bar{y} - p\bar{x})}_{h} \underbrace{n\bar{x}}_{\sum x_i} &= \sum x_i y_i \end{aligned}$$

ou encore

$$\boxed{p \sum x_i^2 + h \sum x_i = \sum x_i y_i}$$

qui n'est autre que la seconde équation normale.

Références

[1] *Méthode des moindres carrés sans calcul différentiel*, Bulletin de la SSPMP, no 77, 1998