le cnam

Mathématiques Appliquées pour le Génie des Procédés et l'Energétique

Paris, automne 2018

Droite de régression

Notes du cours 07

Amélie Danlos, Marie Debacq, François Dubois

MA-GPE (7)

Droite de régression.

F. Dubois
28 novembre 2018

- On se donne une famille de N points (x_j, y_j) qu'on suppose "à peu près alignés" (Figure 1)

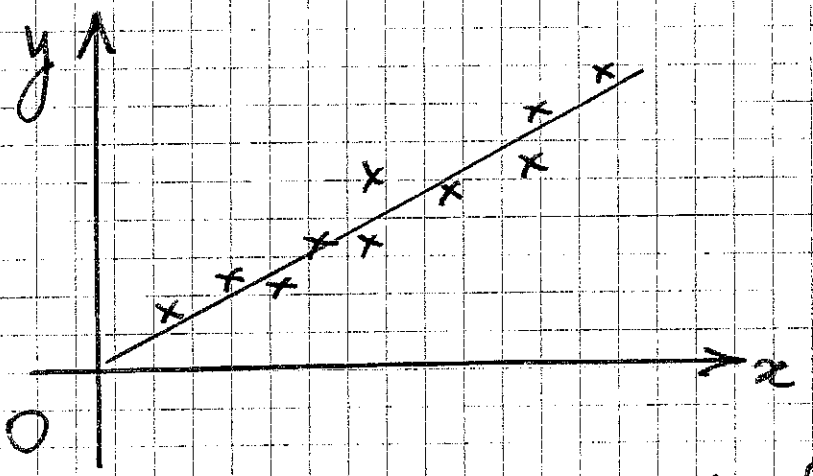

Figure 1 Famille de points quasi alignés.

On cherche à définir (et calculer !) la "meilleure" droite qui passe aussi près que possible de ces points.

- Pour fixer les idées, on cherche une droite de la forme

$$y = \alpha x + \beta$$

Mais (voir la figure 2), cette droite ne passe pas (sauf exception !) par tous les points (x_j, y_j) (voir la Figure 2).

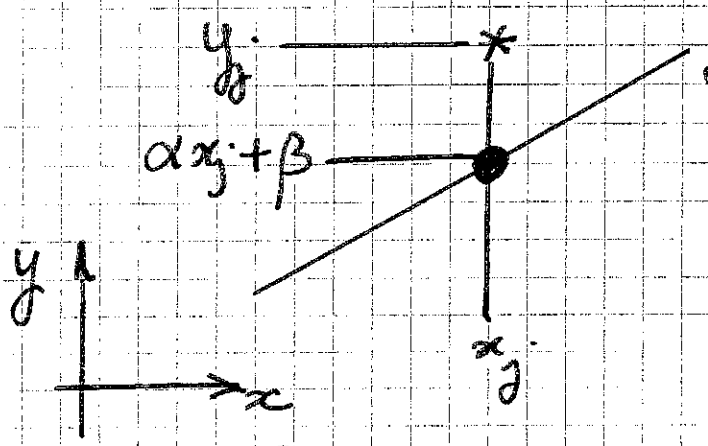

Figure 2. Pour l'abscisse x_j, le point de la droite d'ordonnée $\alpha x_j + \beta$ diffère du point d'ordonnée y_j. L'erreur vaut $|y_j - (\alpha x_j + \beta)|$.

Pour chaque indice j, on a une erreur

$$\varepsilon_j = |y_j - (\alpha x_j + \beta)|, \quad 1 \leq j \leq N.$$

- La méthode des moindres carrés se propose de minimiser la somme des carrés de ces erreurs. On pose, pour une droite de pente α donnée et d'ordonnée à l'origine β également donnée

$$J(\alpha, \beta) = \frac{1}{2N} \sum_{j=1}^{N} |\varepsilon_j|^2.$$

On cherche un couple (α^*, β^*) qui minimise cette erreur, c'est à dire tel que

$$J(\alpha, \beta) \geq J(\alpha^*, \beta^*), \quad \forall (\alpha, \beta) \in \mathbb{R}^2.$$

Le problème ci-dessus a une solution unique. La droite d'équation

$$y = \alpha^* x + \beta^*$$

s'appelle la droite de régression associée au nuage de points (x_j, y_j) $(1 \leq j \leq N)$.

- **Préliminaires algébriques**

L'erreur ε_j contient trois termes, qu'il s'agit de mettre au carré. On a la relation générale

$$(a+b+c)^2 = a^2+b^2+c^2+2(ab+bc+ca).$$

La preuve est un excellent exercice, laissé au lecteur

- Quand on a une fonction polynomiale de degré deux, de la forme

$$p(\xi) = a\xi^2 + b\xi + c \quad , a \neq 0, \xi \in \mathbb{R},$$

on peut toujours l'écrire sous la forme

$$p(\xi) = a(\text{polynome de degré } 1)^2 + \text{constante}.$$

Il suffit de factoriser a : $p(\xi) = a\left(\xi^2 + \frac{b\xi}{a} + \frac{c}{a}\right)$ et de remarquer que $\xi^2 + \frac{b}{a}\xi$ est le début du carré parfait $\left(\xi + \frac{b}{2a}\right)^2$:

$$\xi^2 + \frac{b}{a}\xi = \left(\xi + \frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2$$

Donc

$$p(\xi) = a\left(\xi + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$$

on va se rendre compte que la fonction $J(\alpha, \beta)$ à minimiser est en fait un polynôme du second degré par rapport au couple de variables (α, β). Afin d'alléger le plus possible les notations, nous posons un

• Préliminaire statistique

Si $(\varphi_j)_{1 \leq j \leq N}$ est une famille de N nombres absolument quelconque, on pose

$$\overline{\varphi} = \frac{1}{N} \sum_{j=1}^{N} \varphi_j \ ,$$

moyenne des φ_j. On a en particulier

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \ ; \quad \overline{y} = \frac{1}{N} \sum_{j=1}^{N} y_j \ .$$

Mais on peut également former des moyennes de degré deux :

$$\overline{x^2} = \frac{1}{N} \sum_{j=1}^{N} (x_j)^2 \ ; \quad \overline{y^2} = \frac{1}{N} \sum_{j=1}^{N} (y_j)^2 \ ,$$

moyennes de "x au carré" et de "y au carré".

• Si on forme la moyenne des carrés des écarts à la moyenne de la famille x_j, c'est à dire

$$\overline{(x - \overline{x})^2} = \frac{1}{N} \sum_{j=1}^{N} (x_j - \overline{x})^2$$

on définit un nombre positif (en général non nul, sauf en cas extrême que nous limitons le lecteur à étudier), la variance de la famille $(x_j)_{1 \leq j \leq N}$:

$$\sigma_x^2 = \overline{(x - \bar{x})^2} \geq 0.$$

on a $\sigma_x^2 = \overline{x^2} - (\bar{x})^2$.

En effet, $\sigma_x^2 = \frac{1}{N} \sum_{j=1}^{N} \left((x_j)^2 - 2\bar{x} x_j + (\bar{x})^2 \right)$

$$= \overline{x^2} - 2\bar{x} \cdot \frac{1}{N} \sum_{j=1}^{N} x_j + (\bar{x})^2 \frac{1}{N} \sum_{j=1}^{N} 1$$

$$= \overline{x^2} - 2\bar{x} \cdot \bar{x} + (\bar{x})^2$$

$$= \overline{x^2} - (\bar{x})^2$$

on a bien sûr une définition analogue pour la variable y :

$$\sigma_y^2 = \overline{(y - \bar{y})^2} = \overline{y^2} - (\bar{y})^2 \geq 0.$$

- On peut enfin regarder le moment d'ordre 2 qui fait intervenir les variables x et y :

$$\overline{xy} = \frac{1}{N} \sum_{j=1}^{N} x_j y_j.$$

Le moment obtenu en retranchant d'abord à x sa moyenne et à y sa moyenne s'appelle

la <u>covariance</u> de x et y :

$$\mathrm{Cov}(x,y) = \overline{(x-\bar{x})(y-\bar{y})}.$$

Elle n'a pas de signe particulier, mais son expression peut être transformée :

$$\mathrm{Cov}(x,y) = \overline{xy} - \bar{x}\,\bar{y}.$$

En effet,

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{1}{N}\sum_{j=1}^{N}(x_j-\bar{x})(y_j-\bar{y}) \\
&= \frac{1}{N}\sum_{j=1}^{N}\left(x_j y_j - \bar{x}\,y_j - \bar{y}\,x_j + \bar{x}\,\bar{y}\right) \\
&= \overline{xy} - \bar{x}\,\frac{1}{N}\sum_{j=1}^{N} y_j - \bar{y}\,\frac{1}{N}\sum_{j=1}^{N} x_j + \bar{x}\,\bar{y} \\
&= \overline{xy} - \bar{x}\,\bar{y} - \bar{y}\,\bar{x} + \bar{x}\,\bar{y} \\
&= \overline{xy} - \bar{x}\,\bar{y} \quad \text{comme annoncé.}
\end{aligned}$$

- Expression de la fonction d'erreur avec une somme de carrés.

Nous allons montrer que

$$J(\alpha,\beta) = \frac{1}{2}\left(\beta + \alpha\bar{x} - \bar{y}\right)^2 + \frac{1}{2}\sigma_x^2\left(\alpha - \frac{\mathrm{cov}(x,y)}{\sigma_x^2}\right)^2 + \frac{1}{2\sigma_x^2}\left(\sigma_x^2\sigma_y^2 - \mathrm{cov}(x,y)^2\right)$$

On a

$$J(\alpha,\beta) = \frac{1}{2N}\sum_{j=1}^{N}\left(y_j - \alpha x_j - \beta\right)^2$$

$$= \frac{1}{2N}\sum_{j=1}^{N}\left\{y_j^2 + \alpha^2 x_j^2 + \beta^2 - 2\alpha x_j y_j - 2\beta y_j + 2\alpha\beta x_j\right\}$$

$$= \frac{1}{2}\left[\overline{y^2} + \overline{x^2}\alpha^2 + \beta^2 - 2\alpha\overline{xy} - 2\beta\overline{y} + 2\alpha\beta\overline{x}\right]$$

On considère que cette expression est un polynome du second degré par rapport à la variable β. On fait apparaître un carré (voir le préliminaire algébrique):

$$J(\alpha,\beta) = \frac{1}{2}\left[\beta^2 + 2\beta(\alpha\overline{x} - \overline{y}) + \overline{x^2}\alpha^2 - 2\overline{xy}\alpha + \overline{y^2}\right]$$

$$= \frac{1}{2}\left[\left(\beta + \overline{x}\alpha - \overline{y}\right)^2 - \left(\overline{x}\alpha - \overline{y}\right)^2 + \overline{x^2}\alpha^2 - 2\overline{xy}\alpha + \overline{y^2}\right]$$

$$= \frac{1}{2}\left(\beta + \alpha\overline{x} - \overline{y}\right)^2 + \frac{1}{2}\left[-(\overline{x})^2\alpha^2 - (\overline{y})^2 + 2\alpha\overline{x}\,\overline{y} + \overline{x^2}\alpha - 2\overline{xy}\alpha + \overline{y^2}\right]$$

$$= \frac{1}{2}\left(\alpha\overline{x} + \beta - \overline{y}\right)^2 + \frac{1}{2}\left[\sigma_x^2\alpha^2 - 2\alpha\,\mathrm{cov}(x,y)\alpha + \sigma_y^2\right]$$

en utilisant les définitions introduites lors du préliminaire statistique.

Il suffit maintenant de travailler avec le polynôme à une seule variable (α) qui figure entre crochet:

$$J(\alpha,\beta) = \frac{1}{2}(\alpha\bar{x}+\beta-\bar{y}) + \frac{1}{2}\sigma_x^2\left(\alpha^2 - 2\alpha\frac{cov(x,y)}{\sigma_x^2}\right) + \frac{1}{2}\sigma_y^2$$

$$= \frac{1}{2}(\alpha\bar{x}+\beta-\bar{y})^2 + \frac{\sigma_x^2}{2}\left[\left(\alpha - \frac{cov(x,y)}{\sigma_x^2}\right)^2 - \left(\frac{cov(x,y)}{\sigma_x^2}\right)^2\right] + \frac{1}{2}\sigma_y^2$$

$$= \frac{1}{2}(\alpha\bar{x}+\beta-\bar{y})^2 + \frac{1}{2}\sigma_x^2\left(\alpha - \frac{cov(x,y)}{\sigma_x^2}\right)^2 + \frac{1}{2}\left(\sigma_y^2 - \frac{1}{\sigma_x^2}(cov(x,y))^2\right)$$

$$= \frac{1}{2}(\alpha\bar{x}+\beta-\bar{y}) + \frac{1}{2}\sigma_x^2\left(\alpha - \frac{cov(x,y)}{\sigma_x^2}\right)^2 + \frac{1}{2\sigma_x^2}\left(\sigma_x^2\sigma_y^2 - cov(x,y)^2\right)$$

expression identique à celle proposée page 6.

- <u>Minimisation de la fonction d'erreur.</u>

L'expression ci-dessus contient deux carrés, donc deux termes positifs. On a donc

$$J(\alpha,\beta) \geq \frac{1}{2\sigma_x^2}\left(\sigma_x^2\sigma_y^2 - cov(x,y)^2\right)$$

De plus, on peut trouver α et β de façon à annuler ces termes quadratiques. Les satisfont aux équations

$$\begin{cases} \alpha^* \bar{x} + \beta^* - \bar{y} = 0 \\ \alpha^* - \frac{1}{\sigma_x^2} \operatorname{cov}(x,y) = 0 \end{cases}$$

Pour ces deux valeurs particulières

$$\alpha^* = \frac{1}{\sigma_x^2} \operatorname{cov}(x,y) \; ; \; \beta^* = \bar{y} - \bar{x}\alpha^*$$

on a bien entendu

$$J(\alpha^*, \beta^*) = \frac{1}{2\sigma_x^2}\left(\sigma_x^2 \sigma_y^2 - \operatorname{cov}(x,y)^2\right)$$

- on reporte cette valeur dans l'inégalité précédente et on a établi que

$$J(\alpha,\beta) \geq J(\alpha^*, \beta^*), \quad \forall (\alpha,\beta) \in \mathbb{R}^2.$$

on a trouvé le point de minimum (α^*, β^*). Il définit la droite de régression.

- Inégalité.

A partir de l'expression de $J(\alpha^*, \beta^*)$ et du fait que ce nombre est positif, on déduit

$$\frac{1}{2\sigma_x^2}\left(\sigma_x^2 \sigma_y^2 - \operatorname{cov}(x,y)^2\right) \geq 0$$

inégalité qu'on peut écrire (sans oublier qu'on suppose les écarts types σ_x et σ_y positifs) :

$$|\mathrm{cov}(x,y)| \leqslant \sigma_x \sigma_y .$$

- On appelle <u>coefficient de corrélation</u> des distributions x_j et y_j le nombre

$$r = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} .$$

alors l'inégalité précédente (qui est une forme particulière de l'inégalité de Cauchy-Schwarz) montre que

$$-1 \leqslant r \leqslant 1$$

□ <u>Approche par l'analyse.</u>

On sait que si une fonction $f : \mathbb{R} \to \mathbb{R}$ régulière admet un point x_0 de minimum, c'est à dire

$$f(x) \geqslant f(x_0), \quad \forall x \in \mathbb{R},$$

alors la dérivée de f au point x_0 est nulle :

$$f'(x_0) = 0 .$$

- Pour une fonction de deux variables $J(\alpha, \beta)$ qui est minimale en (α^*, β^*), c'est à dire

$$J(\alpha, \beta) \geqslant J(\alpha^*, \beta^*), \quad \forall (\alpha, \beta) \in \mathbb{R},$$

on peut raisonner comme plus haut par exemple à β fixé, en faisant varier α.

La dérivée (partielle) de la fonction J par rapport à la variable α est nécessairement nulle au point (α^*, β^*) :

$$\frac{\partial J}{\partial \alpha}(\alpha^*, \beta^*) = 0$$

De même en échangeant le rôle des variables :

$$\frac{\partial J}{\partial \beta}(\alpha^*, \beta^*) = 0 .$$

On vient d'écrire deux équations nécessaires pour le point de minimum (α^*, β^*). Si on part maintenant de l'une des expressions de J obtenue page 7, c'est à dire

$$J(\alpha, \beta) = \frac{1}{2}\left[\overline{x^2}\alpha^2 + 2\overline{x}\alpha\beta + \beta^2 - 2\overline{xy}\alpha - 2\overline{y}\beta + \overline{y^2}\right],$$

le calcul des dérivées partielles est facile :

$$\frac{\partial J}{\partial \alpha} = \overline{x^2}\alpha + \overline{x}\beta - \overline{xy}$$

$$\frac{\partial J}{\partial \beta} = \overline{x}\alpha + \beta - \overline{y}$$

et le système d'équations satisfait par (α^*, β^*) s'écrit

$$\begin{cases} \overline{x^2}\,\alpha^* + \overline{x}\,\beta^* = \overline{xy} & \left\{ \begin{matrix} 1 \\ -\overline{x} \end{matrix} \right. \\ \overline{x}\,\alpha^* + \beta^* = \overline{y} \end{cases}$$

On multiplie la première équation par 1, la seconde par $-\overline{x}$, et on ajoute. Le terme en β disparaît et il vient

$$\left(\overline{x^2} - (\overline{x})^2\right) \alpha^* = \overline{xy} - \overline{x}\,\overline{y}$$

c'est à dire

$$\sigma_x^2 \, \alpha^* = cov(x,y)$$

ce qui redonne l'expression de la pente de la droite de regression trouvée à la page 9.

□ <u>Application</u> : Ordre de convergence d'un calcul d'intégrale

En travaux pratiques, nous prendrons des "abscisses" comme un logarithme (base 10) du nombre de mailles n lors d'un calcul approché d'une intégrale et les "ordonnées" comme le logarithme (base 10) de la valeur absolue de l'écart entre la solution exacte et la solution approchée.

Pour fixer les idées, on veut calculer de façon approchée l'intégrale

$$I = \int_a^b f(t)\,dt$$

avec une méthode des trapèzes sur l'intervalle [a,b] coupé en n morceaux :

$$I_n = \frac{b-a}{2n} \sum_{j=0}^{n-1} \left[f(t_j) + f(t_{j+1}) \right]$$

avec $t_{ij} = a + j\,\frac{b-a}{n}$.

- on choisit N valeurs de l'entier n, par exemple en progression géométrique :

$$n_i = 2^i\, m \quad , \quad 1 \leqslant i \leqslant N$$

et on calcule l'erreur $E_n \equiv |I - I_n|$ correspondante :

$$\varepsilon_i = |I - I_{n_i}| = E_{n_i}$$

- la famille de points (X_i, Y_i) est définie ainsi :

$$X_i = \log_{10}(n_i) \; ; \; Y_i = \log_{10}(\varepsilon_i)$$

on les génère à l'aide d'une double-boucle d'un programme Python. Puis on calcule les valeurs α^* et β^* de la droite de régression associée :

$$Y_i \simeq \alpha^* X_i + \beta^*$$

L'opposé du nombre α^* donne alors l'ordre de convergence de la méthode d'intégration numérique :

$$\varepsilon_i \simeq 10^{\beta^*} \left(\frac{1}{n_i}\right)^{-\alpha^*} \quad , \quad 1 \leqslant i \leqslant N$$

Pour la méthode des trapèzes, $-\alpha^* \simeq 2$.

Fribourg, 28 novembre 2018.