Fiche de synthèse : Régression linéaire

Cette feuille regroupe l'essentiel de ce que nous avons vu sur les sujets MMI 2017 et 2019 dans la feuille 24 d'informatique.

Cette notion apparait dans plusieurs chapitres du programme de mathématiques :

- Cours de statistiques : Coefficients de la droite de régression linéaire d'une série double de données.
- Projection orthogonale : Calcul de la distance d'un vecteur u à un plan Vect(v_1, v_2).
- Fonction de $\mathbb{R}^2$ dans $\mathbb{R}$: Calcul d'un point critique de $(a,b) \longmapsto F(a,b)$.

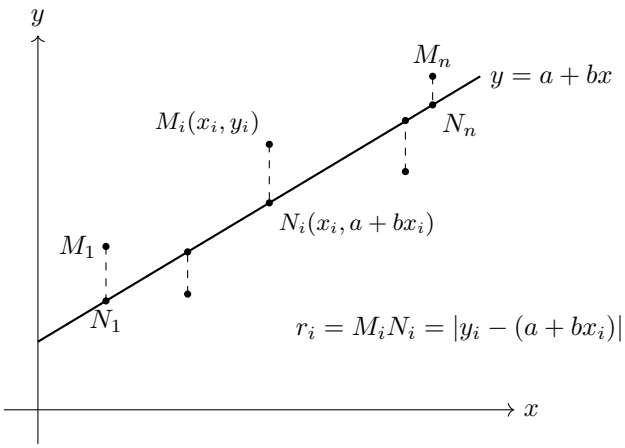
Objectif :

Déterminer le meilleur ajustement affine au sens des moindres carrés du nuage de points :

$((x_1,y_1),(x_2,y_2),...,(x_n,y_n))$ *(on suppose que les x_i ne sont pas tous égaux)*

Autrement dit :

on veut trouver la droite d'équation $y = a + bx$ qui minimise la somme $\boldsymbol{\sum_{i=1}^{n} r_i^2}$ où les r_i sont définis sur la figure ci-dessous :



Notations usuelles :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$\sigma_x^2 = \overline{x^2} - (\bar{x})^2$ *(variance empirique de (x_i))* et $\text{cov}(x,y) = \overline{xy} - \bar{x}\,\bar{y}$ *(covariance empirique de (x_i, y_i))*

Remarque : les x_i ne sont pas tous égaux donc $\sigma_x^2 \neq 0$.

Résultat fondamental

Les coefficients de cette droite appelée droite de régression linéaire sont : $b = \dfrac{\text{cov}(x,y)}{\sigma_x^2}$ et $a = \bar{y} - b\,\bar{x}$

• Première approche.

On note $u = (y_1, ..., y_n)$, $v_1 = (1, ..., 1)$ et $v_2 = (x_1, ..., x_n)$

(la famille (v_1, v_2) est une famille libre car les x_i ne sont pas tous égaux)

on note H le plan vect(v_1, v_2).

Remarque : J'appelle ici ce plan H et non F pour éviter les confusions avec la fonction F.

On sait que la distance de u à H est le minimum de $\left\{||u - av_1 - bv_2|| \mid (a,b) \in \mathbb{R}^2\right\}$ et ce minimum est atteint lorsque $av_1 + bv_2$ est le projeté orthogonal de u sur H.

Or, pour tout $(a,b) \in \mathbb{R}^2$, $\|u - av_1 - bv_2\|^2 = \sum_{i=1}^{n} \big(y_i - (a + bx_i)\big)^2 = \sum_{i=1}^{n} r_i^2$

Ainsi, minimiser $\sum_{i=1}^{n} r_i^2$ revient à projeter u sur H. Autrement dit trouver (a,b) tel que : $p_H(u) = av_1 + bv_2$

$$\begin{aligned}
av_1 + bv_2 = p_{_H}(u) &\iff u - (av_1 + bv_2) \in H^{\perp} \\
&\iff \left\{ \begin{array}{l} u - (av_1 + bv_2) \perp v_1 \\ u - (av_1 + bv_2) \perp v_2 \end{array} \right. \\
&\iff \left\{ \begin{array}{l} < u - (av_1 + bv_2) | v_1 >= 0 \\ < u - (av_1 + bv_2) | v_2 >= 0 \end{array} \right. \\
&\iff \left\{ \begin{array}{l} < u \,|\, v_1 >= a < v_1 \,|\, v_1 > + b < v_2 \,|\, v_1 > \\ < u \,|\, v_2 >= a < v_1 \,|\, v_2 > + b < v_2 \,|\, v_2 > \end{array} \right. \\
&\iff \begin{pmatrix} < v_1 \,|\, v_1 > & < v_1 \,|\, v_2 > \\ < v_2 \,|\, v_1 > & < v_2 \,|\, v_2 > \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} < u \,|\, v_1 > \\ < u \,|\, v_2 > \end{pmatrix}
\end{aligned}$$

on remarque que

$$< v_1 \,|\, v_1 >= n \qquad < v_1 \,|\, v_2 >= n\bar{x} \qquad < v_2 \,|\, v_2 >= n\overline{x^2} \qquad < u \,|\, v_1 >= n\bar{y} \qquad < u \,|\, v_2 >= n\overline{xy}$$

on obtient le système :

$$\begin{aligned}
\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ n\overline{xy} \end{pmatrix} &\iff \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix} \\
&\iff \left\{ \begin{array}{l} a \;+\; \bar{x}\, b = \bar{y} \\ \bar{x}\, a + \overline{x^2}\, b = \overline{xy} \end{array} \right. \\
&\iff \left\{ \begin{array}{ll} a \;+\; \bar{x}\, b = \bar{y} & \\ \big(\overline{x^2} - (\bar{x})^2\big) b = \overline{xy} - \bar{x}\,\bar{y} & L_2 \leftarrow L_2 - \bar{x} L_1 \end{array} \right.
\end{aligned}$$

On obtient alors (car $\sigma_x^2 \neq 0$) :

$$av_1 + bv_2 = p_{_H}(u) \iff \left\{ \begin{array}{l} a = \bar{y} \;-\; b\,\bar{x} \\ \\ b = \dfrac{\mathrm{cov}(x,y)}{\sigma_x^2} \end{array} \right.$$

• Deuxième approche.

On définit $F : (a,b) \longmapsto \sum_{i=1}^{n} (y_i - a - bx_i)^2$,

On remarque que $F(a,b) = \sum_{i=1}^{n} r_i^2$, on cherche donc a et b tels que $F(a,b)$ est minimal.

La première approche justifie que F admette un minimum donc on cherche le(s) point(s) critique(s) de F.

$F : (a,b) \longmapsto \sum_{i=1}^{n} (y_i - a - bx_i)^2$ est polynomiale en a et en b donc elle admet des dérivées partielles.

$$\begin{aligned}
\frac{\partial F}{\partial a}(a,b) &= \sum_{i=1}^{n} -2(y_i - a - bx_i) \\
&= -2\left(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} 1 - b \sum_{i=1}^{n} x_i \right)
\end{aligned} \qquad \begin{aligned}
\frac{\partial F}{\partial b}(a,b) &= \sum_{i=1}^{n} -2x_i(y_i - a - bx_i) \\
&= -2\left(\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} x_i^2 \right)
\end{aligned}$$

donc

$$\begin{aligned}
\left\{ \begin{array}{l} \dfrac{\partial F}{\partial a}(a,b) = 0 \\ \dfrac{\partial F}{\partial b}(a,b) = 0 \end{array} \right. &\iff \left\{ \begin{array}{rcl} n\bar{y} - na - bn\bar{x} & = & 0 \\ \displaystyle\sum_{i=1}^{n} x_i y_i \;-\; an\bar{x} \;-\; b \sum_{i=1}^{n} x_i^2 & = & 0 \end{array} \right. \\
&\iff \left\{ \begin{array}{l} a \;+\; \bar{x}\, b = \bar{y} \\ \bar{x}\, a + \overline{x^2}\, b = \overline{xy} \end{array} \right. \\
&\iff \left\{ \begin{array}{l} a = \bar{y} \;-\; b\,\bar{x} \\ \\ b = \dfrac{\mathrm{cov}(x,y)}{\sigma_x^2} \end{array} \right.
\end{aligned}$$