DN9 : La régression linéaire

Leçons : 230, 231, 437 ; Référence : NESC

Notation : Si X est une variable aléatoire qui admet une espérance on note $\bar{X} = X - \mathbb{E}(X)$ la variable centrée.

Théorème 1 : On suppose que $\Omega = \{\omega_i | i \in [\![1,n]\!]\}$ est fini de cardinal $n \geq 2$ et muni de la probabilité uniforme $\mathbb{p}$ sur $\mathcal{P}(\Omega)$, soit $M = (X,Y)$ un couple de variables aléatoires non constantes définies sur Ω, on note $\forall i \in [\![1,n]\!], x_i = X(\omega_i), y_i = Y(\omega_i)$ et $M_i = M(\omega_i)$, on se place dans un repère orthonormé $(O,\vec{\imath},\vec{\jmath})$ et on note pour toute droite $\mathcal{D}$ du repère $\phi_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^{n}(M_iH_i)^2$ ou H_i est le projeté de M_i sur $\mathcal{D}$ parallèlement au vecteur $\vec{\jmath}$, il existe une unique droite Δ dans le repère tel que $\phi_\Delta = \mathrm{Min}\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\}$ et alors $\phi_\Delta = \mathbb{v}(Y)(1 - \mathrm{Cor}(X,Y)^2)$ avec $\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y)$. L'équation de la droite Δ est $t \to \mathrm{Cov}(X,Y)/\mathbb{v}(X).(t - \mathbb{E}(X)) + \mathbb{E}(Y)$.

• Si on note $\langle . | . \rangle$ l'application $(u,v) \to \frac{1}{n}\sum_{i=1}^{n} u_iv_i$ de $\mathbb{R}^n$ vers $\mathbb{R}$ alors on peut vérifier facilement qu'il s'agit d'un produit scalaire sur $\mathbb{R}^n$ et on note $\| . \|$ la norme associée. Si on note $x = (x_i)_{1\leq i\leq n}$ et $y = (y_i)_{1\leq i\leq n}$ ainsi que $\mathcal{P} = \mathrm{Vect}\{\hat{1}, x\}$ ou $\hat{1}$ désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1 alors on a $\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\} = \left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2 \middle| (a,b) \in \mathbb{R}^2\right\} = \{\|y - ax - b.\hat{1}\|^2 | (a,b) \in \mathbb{R}^2\} = \{\|y - z\|^2 | z \in \mathcal{P}\}$ qui est non vide et minoré par 0.

• On a $\mathrm{Inf}\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\} = \mathrm{Inf}\{\|y - z\|^2 | z \in \mathcal{P}\} = \mathrm{Inf}\{\|y - z\| | z \in \mathcal{P}\}^2 = d(y,\mathcal{P})^2$ et d'après le théorème de projection sur un convexe complet on sait que $d(y,\mathcal{P}) = \|y - y'\|$ ou y' est le projeté orthogonal de y sur $\mathcal{P}$. Le vecteur y' est déterminé par les équations suivantes :

$$\begin{cases} y' \in \mathcal{P} \\ y - y' \perp \hat{1} \\ y - y' \perp x \end{cases} \Leftrightarrow \begin{cases} \exists (a,b) \in \mathbb{R}^2, y' = ax + b.\hat{1} \\ \langle y - y' | \hat{1} \rangle = 0 \\ \langle y - y' | x \rangle = 0 \end{cases} \Leftrightarrow \begin{cases} \exists (a,b) \in \mathbb{R}^2, y' = ax + b.\hat{1} \\ \langle y | \hat{1} \rangle = \langle y' | \hat{1} \rangle \\ \langle y | x \rangle = \langle y' | x \rangle \end{cases}$$

• Avec la deuxième équation du système on obtient $\mathbb{E}(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i = \langle y | \hat{1} \rangle = \langle ax + b.\hat{1} | \hat{1} \rangle = a.\langle x | \hat{1} \rangle + \langle b.\hat{1} | \hat{1} \rangle = a.\mathbb{E}(X) + b$ donc $b = \mathbb{E}(Y) - a.\mathbb{E}(X)$. La dernière équation du système donne $\mathbb{E}(YX) = \langle y | x \rangle = \langle ax + b.\hat{1} | x \rangle = a.\langle x | x \rangle + b.\langle x | \hat{1} \rangle = a.\mathbb{E}(X^2) + b.\mathbb{E}(X) = a.\mathbb{E}(X^2) + \mathbb{E}(Y).\mathbb{E}(X) - a.\mathbb{E}(X)^2 = a.\mathbb{v}(X) + \mathbb{E}(Y).\mathbb{E}(X)$ donc $a = \mathrm{Cov}(X,Y)/\mathbb{v}(X)$. La droite Δ est déterminée et son équation est $t \to a.t + b$.

• Puisque $y - y' \perp y'$ on a $\phi_\Delta = \mathrm{Min}\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\} = \|y - y'\|^2 = \langle y - y' | y - y' \rangle = \langle y - y' | y \rangle = \mathbb{E}(Y^2) - \langle y' | y \rangle = \mathbb{E}(Y^2) - \langle ax + b.\hat{1} | y \rangle = \mathbb{E}(Y^2) - a.\mathbb{E}(XY) - b.\mathbb{E}(Y) = \mathbb{E}(Y^2) - a.\mathbb{E}(XY) - \big(\mathbb{E}(Y) - a.\mathbb{E}(X)\big).\mathbb{E}(Y) = \mathbb{v}(Y) - a.\mathrm{Cov}(X,Y) = \mathbb{v}(Y)\left(1 - \frac{\mathrm{Cov}(X,Y)^2}{\mathbb{v}(X)\mathbb{v}(Y)}\right)$ de plus on remarque en notant $\bar{x} = (x_i - \mathbb{E}(X))_{1\leq i\leq n}$ et $\bar{y} = (y_i - \mathbb{E}(Y))_{1\leq i\leq n}$ que :

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathbb{E}(\bar{X}\bar{Y})}{\sqrt{\mathbb{E}(\bar{X}^2)\mathbb{E}(\bar{Y}^2)}} = \frac{\langle \bar{x} | \bar{y} \rangle}{\|\bar{x}\|\|\bar{y}\|} = \mathrm{Cos}(\bar{x},\bar{y})$$

A partir d'un angle de $30°$ la corrélation linéaire est jugée bonne.

DEVELOPPEMENT SOUS LA FORME D'UN EXERCICE :

1) Déterminer un produit scalaire $\langle . | . \rangle$ sur $\mathbb{R}^n$ tel que $\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\} = \{\|y - z\|^2 | z \in \mathcal{P}\}$ ou y est un vecteur de $\mathbb{R}^n$ associé à Y et $\mathcal{P}$ un plan vectoriel de $\mathbb{R}^n$.
2) Montrer que $\mathrm{Min}\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\} = \|y - y'\|^2$ ou y' est le projeté orthogonal de y sur $\mathcal{P}$.
3) Déterminer l'équation de la droite Δ et $\phi_\Delta = \mathrm{Min}\{\phi_{\mathcal{D}} | \mathcal{D}$ droite du plan$\}$.
4) Montrer que $\mathrm{Cor}(X,Y)$ est le cosinus d'un certain angle que l'on déterminera.