Une droite de régression méconnue: la droite de régression orthogonale

Alain Brobecker, mars 2008.

Mes remerciements à Philippe Domergue pour son aide.

Dans toutes mes références connues, y compris pour le supérieur, on donne les formules de $\mathcal{D}_1$, la droite de régression de Y en X qui minimise les carrés des distances verticales entre la droite et les points du nuage, et de $\mathcal{D}_2$, la droite de régression de X en Y qui minimise les carrés des distances horizontales entre la droite et les points du nuage.

Par contre la droite de régression orthogonale $\mathcal{D}_0$, qui minimise les carrés des distances, se fait étrangement très discrète: seule une brève mention en est faite dans le Dictionnaire des Mathématiques (A. Bouvier, M. George, F. Le Lionnais, PUF). Je vous propose de la découvrir ici.

1. Méthode des moindres carrés

Pourquoi minimiser les carrés des distances plutôt que les distances?
Une réponse, immédiate et assez convaincante, m'a été apportée par un des formateurs au concours de professorat: si on minimise les distances, on n'a pas obligatoirement unicité du minimum. Par exemple, prenons en dimension 1 le nuage constitué des 4 points $\{2;4;6;8\}$. Tous les points x compris entre 4 et 6 minimisent les distances $\sum|x-x_i|$, mais seul $x=5$ minimise les carrés des distances $\sum(x-x_i)^2$.
Notons toutefois que, même en minimisant les carrés, il peut arriver que le minimum ne soit pas unique en dimension 2, mais c'est plutôt rare.

2. Rappel des formules usuelles en statistiques

Soient un entier naturel $n \in \mathbb{N}-\{0;1\}$, une population $\Omega=\{\omega_1;...;\omega_n\}$ et deux variables aléatoires réelles $X:\Omega\to\mathbb{R}$ et $Y:\Omega\to\mathbb{R}$. On notera $\forall i\in[1;n]$, $x_i=X(\omega_i)$ et $y_i=Y(\omega_i)$. L'ensemble $\{P_i(x_i;y_i)\in\mathbb{R}^2 \ / \ i\in[1;n]\}$ sera appelé nuage de points associé à la série statistique double $(X;Y)$.

La moyenne de X est: $\bar{x}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i$

La variance de X est: $var(X)=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2=\dfrac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2$

La covariance de X et Y est: $cov(X;Y)=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\dfrac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}\bar{y}$

Notons que dans certains logiciels (notamment sur la TI-89), la variance est parfois divisée par $n-1$ au lieu de n. J'en ignore la raison.

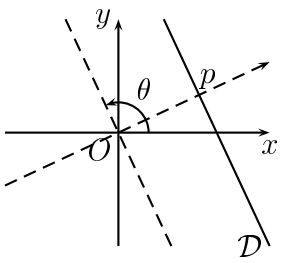
Le point de coordonnées $(\bar{x};\bar{y})$ est appelé point moyen du nuage. Avec toutes ces définitions, des équations des droites d'ajustement $\mathcal{D}_1$ et $\mathcal{D}_2$ sont:

$$\mathcal{D}_1 : y - \bar{y} = \frac{cov(X;Y)}{var(X)}(x - \bar{x}) \qquad \mathcal{D}_2 : x - \bar{x} = \frac{cov(X;Y)}{var(Y)}(y - \bar{y})$$

Ces deux formules sont simples car leur utilisation ne fait appel qu'aux définitions vues ci-dessus, et peuvent être abordées assez tôt dans la scolarité.

Notons enfin la méthode de Mayer qui consiste à couper le nuage de points en deux (par exemple en triant selon une des deux séries), puis en faisant passer la droite de régression $\mathcal{D}_M$ par les points moyens de chacun des sous nuages.

3. Équation d'une droite



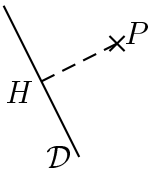
Toute droite $\mathcal{D}$ du plan est associée à un unique couple formé d'un angle $\theta \in [0;\pi[$ donnant la direction de la droite et d'une position $p \in \mathbb{R}$ perpendiculairement à cette direction, comme représenté sur la figure ci-contre. La droite $\mathcal{D}$ passe alors par le point de coordonnées $\left(p\cos(\theta - \frac{\pi}{2}); p\sin(\theta - \frac{\pi}{2})\right)$ et admet $\overrightarrow{u}(\cos\theta;\sin\theta)$ comme vecteur directeur.
Cherchons alors une équation cartésienne de $\mathcal{D}$:

$$M(x;y) \in \mathcal{D} \iff \exists\lambda \in \mathbb{R} \text{ tel que } \begin{cases} x = p\cos(\theta - \frac{\pi}{2}) + \lambda\cos\theta = p\sin\theta + \lambda\cos\theta \\ y = p\sin(\theta - \frac{\pi}{2}) + \lambda\sin\theta = -p\cos\theta + \lambda\sin\theta \end{cases}$$

En éliminant λ par combinaison linéaire, on obtient:

$$M(x;y) \in \mathcal{D} \iff x\sin\theta - y\cos\theta = p(\sin^2\theta + \cos^2\theta) \iff \boxed{x\sin\theta - y\cos\theta - p = 0}$$

4. Distance d'un point à une droite



Étant donnés une droite $\mathcal{D}$ du plan définie par le couple $(\theta;p) \in [0;\pi[\times\mathbb{R}$ et un point $P(x;y)$, on définit le point $H(x';y')$ comme le projeté orthogonal de P sur $\mathcal{D}$, et on recherche l'expression de $dist(P;\mathcal{D}) = dist(P;H)$.
Recherchons d'abord les coordonnées du point H:

$$H(x';y') \in \mathcal{D} \text{ tel que } \overrightarrow{PH}.\overrightarrow{u} = 0 \iff \begin{cases} x'\sin\theta - y'\cos\theta - p = 0 \\ (x'-x)\cos\theta + (y'-y)\sin\theta = 0 \end{cases}$$

$$\iff \begin{cases} x'\sin^2\theta - y'\cos\theta\sin\theta - p\sin\theta + (x'-x)\cos^2\theta + (y'-y)\sin\theta\cos\theta = 0 \\ x'\sin\theta\cos\theta - y'\cos^2\theta - p\cos\theta - (x'-x)\cos\theta\sin\theta - (y'-y)\sin^2\theta = 0 \end{cases}$$

$$\iff \begin{cases} x' = x\cos^2\theta + y\cos\theta\sin\theta + p\sin\theta \\ y' = x\cos\theta\sin\theta + y\sin^2\theta - p\cos\theta \end{cases}$$

On peut maintenant calculer la distance:

$$\begin{aligned}
dist(P;\mathcal{D}) &= dist(P;H) \\
&= \sqrt{(x-x')^2+(y-y')^2} \\
&= \sqrt{(x\sin^2\theta - y\cos\theta\sin\theta - p\sin\theta)^2 + (y\cos^2\theta - x\cos\theta\sin\theta + p\cos\theta)^2} \\
&= \sqrt{(x\sin\theta - y\cos\theta - p)^2\sin^2\theta + (y\cos\theta - x\sin\theta + p)^2\cos^2\theta} \\
&= \sqrt{(x\sin\theta - y\cos\theta - p)^2 \times (\cos^2\theta + \sin^2\theta)} \\
&= \sqrt{(x\sin\theta - y\cos\theta - p)^2} \\
&= \boxed{|x\sin\theta - y\cos\theta - p|}
\end{aligned}$$

5. Droite de régression orthogonale

Soient un entier naturel $n \in \mathbb{N}-\{0;1\}$ et un nuage de points $\{P_i(x_i;y_i) \in \mathbb{R}^2 \ / \ i \in [1;n]\}$. On recherche la droite $\mathcal{D}_0$ qui minimise la somme des carrés des distances de $\mathcal{D}_0$ aux points P_i, c'est à dire le couple $(\theta_0;p_0) \in [0;\pi[\times\mathbb{R}$ qui minimise la fonction:

$$\varphi(\theta;p) = \sum_{i=1}^{n}(x_i\sin\theta - y_i\cos\theta - p)^2$$

• *Existence d'un minimum pour* $\varphi(\theta;p)$

La fonction φ est une fonction périodique en θ, de période π. On peut donc la prolonger par continuité sur $[0;\pi]\times\mathbb{R}$, et si un extremum est atteint sur cet intervalle il y en aura aussi un atteint sur $[0;\pi[\times\mathbb{R}$ puisque $\forall p \in \mathbb{R}\ ,\ \varphi(\pi;p) = \varphi(0;p)$.

On a $\forall\theta \in [0;\pi]\ ,\ \lim_{p\to+\infty}\varphi(\theta;p) = +\infty$ et $\lim_{p\to-\infty}\varphi(\theta;p) = +\infty$, donc $\exists P \in \mathbb{R}_+$ tel que $\forall p \in \mathbb{R}\ ,\ |p| > P \Rightarrow (\forall\theta \in [0;\pi]\ ,\ \varphi(\theta;p) > \varphi(0;0))$.

Comme $[0;\pi]\times[-P;P]$ est un compact, son image par la fonction continue φ est un compact, la fonction φ admet donc un minimum $\varphi(\theta_0;p_0)$ sur $[0;\pi]\times[-P;P]$ et on vient de montrer que les valeurs atteintes en dehors de cet ensemble sont supérieures à $\varphi(0;0)$ qui est elle même supérieure ou égale à $\varphi(\theta_0;p_0)$. Donc φ admet un minimum sur $[0;\pi]\times\mathbb{R}$, donc sur $[0;\pi[\times\mathbb{R}$.

• *Passage de* $\mathcal{D}_0$ *par le point moyen*

φ est différentiable sur $[0;\pi[\times\mathbb{R}$, donc lorsque $\varphi(\theta;p)$ est minimal ses dérivées partielles sont nulles, donc en dérivant par rapport à la variable p on obtient:

$$\frac{\partial\varphi}{\partial p}(\theta_0;p_0) = 0 \quad\Rightarrow\quad \sum_{i=1}^{n}(2p_0 - 2x_i\sin\theta_0 + 2y_i\cos\theta_0) = 0$$

$$\Rightarrow\quad 2np_0 - 2\sin\theta_0\sum_{i=1}^{n}x_i + 2\cos\theta_0\sum_{i=1}^{n}y_i = 0 \quad\Rightarrow\quad \boxed{p_0 = \bar{x}\sin\theta_0 - \bar{y}\cos\theta_0}$$

On en conclut que $\mathcal{D}_0$ passe par le point moyen du nuage, de coordonnées $(\bar{x};\bar{y})$, et que son équation peut s'écrire sous la forme $(x-\bar{x})\sin\theta_0 - (y-\bar{y})\cos\theta_0 = 0$.

• *Calcul d'une équation cartésienne de $\mathcal{D}_0$*
Commençons d'abord par réécrire $\varphi(\theta;p_0)$ à l'aide du résultat précédent, et en changeant sa notation pour introduire la fonction partielle $f:\theta\in[0;\pi[\ \mapsto\varphi(\theta;p_0)$:

$$\begin{aligned} f(\theta) &= \sum_{i=1}^{n}\left((x_i-\bar{x})\sin\theta-(y_i-\bar{y})\cos\theta\right)^2 \\ &= \sum_{i=1}^{n}\left((x_i-\bar{x})^2\sin^2\theta-2(x_i-\bar{x})(y_i-\bar{y})\sin\theta\cos\theta+(y_i-\bar{y})^2\cos^2\theta\right) \\ &= \sin^2\theta\sum_{i=1}^{n}(x_i-\bar{x})^2-2\sin\theta\cos\theta\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+\cos^2\theta\sum_{i=1}^{n}(y_i-\bar{y})^2 \end{aligned}$$

On recherche maintenant les extremums de la fonction f. Notons que le minimum et le maximum seront atteints par le prolongement par continuité de f sur le compact $[0;\pi]$, puis par périodicité de f ces deux extremums seront aussi atteints sur $[0;\pi[$. Comme f est dérivable sur $[0;\pi[$, lorsqu'elle atteindra un extremum on aura $f'(\theta_0)=0$, donc:

$$2\cos\theta_0\sin\theta_0\sum_{i=1}^{n}(x_i-\bar{x})^2-2(\cos^2\theta_0-\sin^2\theta_0)\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-2\cos\theta_0\sin\theta_0\sum_{i=1}^{n}(y_i-\bar{y})^2=0$$

$$\Rightarrow\quad\boxed{(\cos^2\theta_0-\sin^2\theta_0)\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\cos\theta_0\sin\theta_0\left(\sum_{i=1}^{n}(x_i-\bar{x})^2-\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}$$

Posons $A=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ qui vaut n fois la covariance et $B=\sum_{i=1}^{n}(x_i-\bar{x})^2-\sum_{i=1}^{n}(y_i-\bar{y})^2$ qui vaut n fois la différence des variances. On est maintenant amené à résoudre l'équation $(\cos^2\theta_0-\sin^2\theta_0)\times A=\cos\theta_0\sin\theta_0\times B$ selon les cas suivants:

(1) Si on a $A=0$ et $B=0$, c'est à dire que la covariance est nulle et que les variances sont égales, alors toute valeur de θ_0 convient puisqu'on a $f(\theta_0)=\sum(x_i-\bar{x})^2=\sum(y_i-\bar{y})^2$ qui ne dépend pas de θ_0. Toutes les droites passant par le point moyen du nuage minimisent les carrés des distances.

(2) Si on a $A=0$ et $B\neq0$, càd que la covariance est nulle, alors on cherche θ_0 vérifiant $\cos\theta_0\sin\theta_0\times B=0\iff(\cos\theta_0=0\text{ ou }\sin\theta_0=0)\iff\theta_0\in\left\{0;\frac{\pi}{2}\right\}$. L'équation de $\mathcal{D}_0$ est alors $x=\bar{x}$ ou $y=\bar{y}$.

(3) Si on a $A\neq0$ et $B=0$, càd que les variance sont égales, alors on cherche θ_0 vérifiant $(\cos^2\theta_0-\sin^2\theta_0)\times A=0\iff\cos\theta_0=\pm\sin\theta_0\iff\theta_0\in\left\{\frac{\pi}{4};\frac{3\pi}{4}\right\}$. L'équation de $\mathcal{D}_0$ est alors $y-\bar{y}=x-\bar{x}$ ou $y-\bar{y}=-x+\bar{x}$.

(4) Enfin si $A\neq0$ et $B\neq0$ il y a plusieurs manières de résoudre l'équation. On peut par exemple la réécrire sous la forme $2\cos(2\theta_0)\times A=\sin(2\theta_0)\times B$, et comme aucun terme n'est nul on trouve $\tan(2\theta_0)=\dfrac{2A}{B}\iff\theta_0=\dfrac{1}{2}\arctan(\dfrac{2A}{B}+k\pi),k\in\mathbb{Z}$, ce qui donne deux droites perpendiculaires.
Si on est allergique à l'utilisation de arctan, comme mon compilateur C, on posera plutôt

$t = \tan\theta_0$ et l'équation se résoudra avec la méthode du discriminant:

$$2 \times \frac{1-t^2}{1+t^2} \times A = \frac{2t}{1+t^2} \times B \iff t^2 \times A + t \times B - A = 0 \iff t = \frac{-B \pm \sqrt{B^2+4A^2}}{2A}$$

On a alors des équations de deux droites perpendiculaires qui sont:

$$\boxed{\boldsymbol{y = \frac{-B \pm \sqrt{B^2+4A^2}}{2A} \times (x - \bar{x}) + \bar{y}}}$$

avec $\boldsymbol{A = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$ et $\boldsymbol{B = \sum_{i=1}^{n}(x_i - \bar{x})^2 - \sum_{i=1}^{n}(y_i - \bar{y})^2}$

Dans les cas (2), (3) et (4) on trouve deux solutions qui correspondent aux deux extremums de la fonction $f(\theta)$ sur $[0;\pi[$, càd qui correspondent à deux droites perpendiculaires passant par le point moyen du nuage, l'une minimisant les carrés des distances, l'autre les maximisant (pour une droite passant par le point moyen). Il faut donc encore calculer de manière effective les carrés des distances pour savoir laquelle des deux droites est $\mathcal{D}_0$.

D'un point de vue informatique je conseille de mélanger les cas (1) et (2) avec le test "si $A = 0$", à condition d'accepter que la particularité du cas (1) ne soit pas visible. De même on peut mélanger les cas (3) et (4) avec le "si $A \neq 0$", puisque la formule du (4) fonctionne aussi lorsque B=0.

6. Remarques et questions diverses

• Dans le cas ou le nuage de point est étudié dans un système de coordonnées homogène, la droite de régression orthogonale semble la plus pertinente. Elle est notamment invariante lors d'une rotation, ce qui n'est pas le cas de $\mathcal{D}_1$ et $\mathcal{D}_2$.

• Par contre si les deux variables sont hétérogènes (par exemple si elles expriment une distance en fonction du temps), la notion de distance n'a pas de sens, or la droite de régression orthogonale repose sur cette notion de distance! Mais même dans ce cas il n'est pas sûr qu'elle soit disqualifiée puisque l'utilisation d'une pseudo-distance ne semble pas plus gênante qu'un choix qui ne serait pas motivé entre $\mathcal{D}_1$ et $\mathcal{D}_2$.

• Un cas particulier du cas hétérogène apparait lorsqu' une des variables doit être modifiée (par exemple à l'aide d'un logarithme) avant de procéder à l'ajustement affine.

• Il est possible d'arriver aux formules précédentes en gardant les équations de droite sous la forme $y = ax + b$. On trouve alors que:

$$dist(P(x;y);\mathcal{D}) = \frac{|ax + b - y|}{\sqrt{a^2+1}}$$

Puis on pose:

$$\varphi(a;b) = \sum_{i=1}^{n} \frac{(ax_i + b - y_i)^2}{a^2+1}$$

Enfin en cherchant les valeurs de a et b pour lesquelles les dérivées partielles de $\varphi(a;b)$ par rapport à b et a s'annulent, on trouvera que $b = \bar{y} - a\bar{x}$ puis les formules déjà présentées.

Cela avait été fait par mes soins en janvier 2003, mais n'avait pas permis de prouver l'unicité du minimum ni d'étudier tous les cas. Le passage aux équations de droites de la forme $ax+by+c=0$ s'est avéré compliqué, d'où le passage aux coordonnées "polaires".

• Existe t'il une manière simple de déterminer laquelle des deux droites trouvées est celle qui minimise le carré des distances, sans avoir à calculer effectivement celui-ci?

• Lorsqu'on minimise les carrés on donne une grande importance aux valeurs éloignées de la moyenne. En pratique ne vaudrait-il pas mieux les éliminer puisqu'elles risquent davantage de résulter d'erreurs de mesures?

• Les deux droites de régression trouvées ont-elles un lien avec les axes de l'ellipse d'aire minimale englobant le nuage de points? Il faudrait vérifier en premier lieu que son centre est bien $(\bar{x};\bar{y})$ le point moyen du nuage. A ma connaissance il n'existe que des méthodes approchées pour la recherche de l'ellipse d'aire minimale englobant un nuage de points.

7. Exemple

x	1	9	4	7	13	0	6	3	7	13
y	2	8	4	9	7	0	6	2	5	10

$\mathcal{D}_0$: Droite de regression orthogonale, $y = 0.703708114x + 0.866638880$
$\mathcal{D}_1$: Droite de regression de Y en X, $y = 0.643053267x + 1.248764415$
$\mathcal{D}_2$: Droite de regression de X en Y, $y = 0.837745517x + 0.022203245$