Modèle linéaire – Questions complémentaires et corrections
Régression linéaire orthogonale et régression linéaire multiple

5.5. Régression linéaire orthogonale

Soit X et Y deux variables réelles définies sur n individus et $\left(X_i, Y_i\right)_{i=1,\dots,n}$ la série double associée.
On représente dans un diagramme cartésien les points M_i de coordonnées $\left(X_i, Y_i\right)$, $i = 1,\dots,n$.
Pour toute droite D, d'équation $y = a\,x + b$, on note H_i la projection orthogonale de M_i sur D, $i = 1,\dots,n$. Déterminer les coefficients a et b qui minimisent $\sum_{i=1}^{n} M_iH_i^2$.
La droite solution de ce problème est appelée droite de régression linéaire orthogonale.

5.6. Régression linéaire multiple

On considère $p + 1$ variables réelles $X_1,\dots,X_p, Y$ observées sur n individus. Pour $j = 1,\dots, p$ et $i = 1,\dots, n$, on note X_{ji} la valeur de la variable X_j sur l'individu i et Y_i la valeur de la variable Y sur l'individu i.
On se propose d'ajuster sur ces données une relation linéaire $Y = a_1X_1 + \dots + a_pX_p$.
Pour cela, on utilise le critère des moindres carrés des erreurs d'ajustement, c'est-à-dire, on cherche $\left(a_1,\dots,a_p\right)$ minimisant l'expression : $C = \sum_{i=1}^{n}\left(Y_i - a_1X_{1i} - \dots - a_pX_{pi}\right)^2$.
Pour $j = 1,\dots, p$, on identifie la variable X_j au vecteur $\left(X_{j1},\dots,X_{jn}\right)$ de $\mathbb{R}^n$ et à la matrice colonne de ses coordonnées par rapport à la base canonique de $\mathbb{R}^n$; de même pour Y.
Soit $X = \left(X_1 \dots X_p\right)$ la matrice (n, p) dont les colonnes sont les matrices colonnes X_j, X' sa transposée et A le vecteur $\left(a_1,\dots,a_p\right)$ de $\mathbb{R}^p$ (identifié également à la matrice colonne $(p,1)$ de ses coordonnées par rapport à la base canonique de $\mathbb{R}^p$).
On suppose que les vecteurs $X_1,\dots,X_p$ sont linéairement indépendants.
1) Montrer que la matrice $X'X$ est inversible.
2) Montrer que le critère des moindres carrés revient à déterminer le vecteur A de $\mathbb{R}^p$ qui minimise $\|Y - XA\|$, la norme étant la norme euclidienne classique de $\mathbb{R}^n$.
3) On en déduit que XA est la projection orthogonale de Y sur le sous-espace vectoriel engendré par $X_1,\dots,X_p$. Vérifier alors l'égalité : $A = \left(X'X\right)^{-1}X'Y$.
4) Montrer que l'on a, après ajustement : $\|Y\|^2 = \|XA\|^2 + \|Y - XA\|^2$.
En déduire que le rapport $q = \|XA\| / \|Y\|$ est un indice de qualité de l'ajustement compris entre 0 et 1. Quelle est l'interprétation de $q = 1$?
5) On suppose $p = 2$ et $X_{2i} = 1$ pour tout i de $\{1,\dots,n\}$.
Retrouver, en calculant a_1 et a_2, les résultats concernant l'ajustement par les moindres carrés d'une relation affine $Y = a_1X_1 + a_2$.

5.5. Régression linéaire orthogonale. Correction.

Soit X et Y deux variables réelles (non constantes) observées sur n individus statistiques.
On note $\overline{x}, \overline{y}, s_X^2, s_Y^2, s_{XY}$ les moyennes, variances et la covariance de X et Y et on suppose $s_{XY} \neq 0$.
On représente les n points $M_i\left(x_i, y_i\right)$, i = 1, …, n, dans un repère orthogonormé, on note (D) une droite d'équation $y = ax + b$ et H_i la projection orthogonale de M_i sur $\left(D\right)$, i = 1, …, n.
On vérifie alors que, pour i = 1, … , n, le carré de la distance du point M_i à la droite (D) est :
$M_iH_i^2 = \left(y_i - ax_i - b\right)^2 \frac{1}{a^2+1}$.
La droite de régression orthogonale est la droite dont les coefficients a et b minimisent la moyenne (ou la somme) des carrés des distances des points à la droite, c'est-à-dire, le critère :
$$f\left(a,b\right) = \frac{1}{a^2+1}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - ax_i - b\right)^2$$
Comme on a :
$$\begin{aligned}\left(y_i - ax_i - b\right)^2 &= \left(y_i - \overline{y} - a\left(x_i - \overline{x}\right) + \overline{y} - a\overline{x} - b\right)^2\\ &= \left(y_i - \overline{y} - ax_i + a\overline{x}\right)^2 + 2\left(y_i - \overline{y} - ax_i + a\overline{x}\right)\left(\overline{y} - a\overline{x} - b\right) + \left(\overline{y} - a\overline{x} - b\right)^2\end{aligned}$$
on en déduit :
$f\left(a,b\right) = \frac{1}{a^2+1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y} - ax_i + a\overline{x}\right)^2 + \left(\overline{y} - a\overline{x} - b\right)^2\right]$ car la somme des seconds termes est nulle.

Le critère $f\left(a,b\right)$ est la somme de deux termes positifs, le deuxième est minimum (et même nul) pour $b = \overline{y} - a\overline{x}$ et le premier ne dépend que de a. La droite $\left(D\right)$ passe donc par le point moyen $G\left(\overline{x}, \overline{y}\right)$.

Il reste à minimiser sur a la fonction :
$$g\left(a\right) = \frac{1}{a^2+1}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y} - ax_i + a\overline{x}\right)^2$$
Comme on a :
$$\left(y_i - \overline{y} - a\left(x_i - \overline{x}\right)\right)^2 = a^2\left(x_i - \overline{x}\right)^2 - 2a\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) + \left(y_i - \overline{y}\right)^2,$$
on en déduit :
$$g\left(a\right) = \frac{1}{a^2+1}\left(a^2 s_X^2 - 2a\,s_{XY} + s_Y^2\right).$$
L'application g est définie pour tout a de $\mathbb{R}$, continue, dérivable et a pour limite s_X^2 lorsque a tend vers $+\infty$ ou $-\infty$.
On obtient pour la dérivée :

$$g'(a) = \frac{1}{\left(a^2+1\right)^2}\left[\left(2as_X^2 - 2s_{XY}\right)\left(a^2+1\right) - \left(a^2 s_X^2 - 2as_{XY} + s_Y^2\right)2a\right]$$

$$= \frac{1}{\left(a^2+1\right)^2}\left[2a^3 s_X^2 - 2a^2 s_{XY} + 2as_X^2 - 2s_{XY} - 2a^3 s_X^2 + 4a^2 s_{XY} - 2as_Y^2\right]$$

$$= \frac{2}{\left(a^2+1\right)^2}\left[a^2 s_{XY} + a\left(s_X^2 - s_Y^2\right) - s_{XY}\right]$$

Le discriminant du trinôme du second degré en a est alors : $\Delta = \left(s_X^2 - s_Y^2\right)^2 + 4s_{XY}^2$; étant strictement positif, le trinôme a deux racines : $a = \dfrac{-s_X^2 + s_Y^2 \pm \sqrt{\Delta}}{2s_{XY}}$.

Selon le signe de s_{XY}, on a le tableau de variation suivant :

si $s_{XY} > 0$

a		$\frac{-s_X^2 + s_Y^2 - \sqrt{\Delta}}{2s_{XY}}$		$\frac{-s_X^2 + s_Y^2 + \sqrt{\Delta}}{2s_{XY}}$		
$g'(a)$	+	0	–	0	+	
g	s_X^2 ↗		↘	min	↗	s_X^2

si $s_{XY} < 0$

a		$\frac{-s_X^2 + s_Y^2 + \sqrt{\Delta}}{2s_{XY}}$		$\frac{-s_X^2 + s_Y^2 - \sqrt{\Delta}}{2s_{XY}}$		
$g'(a)$	–	0	+	0	–	
g	s_X^2 ↘	min	↗		↘	s_X^2

Dans les deux cas le minimum de *g(a)* est obtenu pour $a = \dfrac{-s_X^2 + s_Y^2 + \sqrt{\Delta}}{2s_{XY}}$.

La droite de régression orthogonale est la droite (*D*) d'équation $y = ax + b$ avec :

$$a = \frac{-s_X^2 + s_Y^2 + \sqrt{\left(\left(s_X^2 - s_Y^2\right)^2 + 4s_{XY}^2\right)}}{2s_{XY}} \quad \text{et} \quad b = \overline{y} - a\overline{x}\,.$$

On vérifie ci-après que cette droite est engendrée par les vecteurs propres associés à la plus grande valeur propre de la matrice de covariance de X et Y. Cette matrice étant à coefficients réels, symétrique et positive, elle est diagonalisable et admet des valeurs propres réelles positives.

Soit $V=\begin{pmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{pmatrix}$ la matrice de covariance.

On peut écrire, pour tout λ réel :

$|V-\lambda I|=\left(s_X^2-\lambda\right)\left(s_Y^2-\lambda\right)-s_{XY}^2=\lambda^2-\lambda\left(s_X^2+s_Y^2\right)+s_X^2\,s_Y^2-s_{XY}^2$.

Le discriminant du trinôme en λ vérifie :

$\Delta=\left(s_X^2+s_Y^2\right)^2-4\left(s_X^2\,s_Y^2-s_{XY}^2\right)=\left(s_X^2-s_Y^2\right)^2+4s_{XY}^2>0$.

Le trinôme a donc deux racines :

$\lambda_1=\dfrac{s_X^2+s_Y^2+\sqrt{\left(s_X^2-s_Y^2\right)^2+4s_{XY}^2}}{2}$ $\quad\lambda_2=\dfrac{s_X^2+s_Y^2-\sqrt{\left(s_X^2-s_Y^2\right)^2+4s_{XY}^2}}{2}$ et on a $\lambda_1>\lambda_2$

Soit (α,β) un vecteur propre associé à λ_1.

Il vérifie : $\dfrac{s_X^2-s_Y^2-\sqrt{\left(s_X^2-s_Y^2\right)^2+4s_{XY}^2}}{2}\alpha+s_{XY}\beta=0$

Le coefficient directeur de la droite engendrée par ce vecteur est donc :

$\dfrac{\beta}{\alpha}=\dfrac{-s_X^2+s_Y^2+\sqrt{\left(s_X^2-s_Y^2\right)^2+4s_{XY}^2}}{2s_{XY}}$.

On retrouve le coefficient directeur de la droite de régression orthogonale.

5.6. Régression linéaire multiple. Correction.

1) Soit U un vecteur de $\mathbb{R}^p$ tel que : $X'XU=0$.

On a aussi $U'X'XU=0$, c'est-à-dire $\|XU\|^2=0$, donc $XU=0$.

Les vecteurs $X_1,\ldots,X_p$ étant linéairement indépendants, on en déduit $U=0$.

Le système $X'XU=0$ n'ayant que la solution nulle, la matrice $X'X$ est donc inversible.

2) On a $C=\sum_{i=1}^n\left(Y_i-a_1X_{1i}-\ldots-a_pX_{pi}\right)^2=\left\|Y-a_1X_1-\ldots-a_pX_p\right\|^2=\|Y-XA\|^2$.

Minimiser C revient donc à minimiser $\|Y-XA\|$.

3) Le vecteur XA est la projection orthogonale de Y sur le sous-espace vectoriel engendré par les X_j, on a donc, $\forall j\in\{1,\ldots,p\}$, $\langle X_j,Y-XA\rangle=X_j'\left(Y-XA\right)=0$ et $X'Y=X'XA$.

La matrice $X'X$ étant inversible, on en déduit : $A=\left(X'X\right)^{-1}X'Y$.

4) Pour le vecteur A réalisant l'ajustement, les vecteurs XA et $Y-XA$ sont orthogonaux et de la décomposition : $Y=XA+\left(Y-XA\right)$ on peut déduire :

$$\|Y\|^2=\|XA\|^2+\|Y-XA\|^2$$

Le rapport $q=\|XA\|/\|Y\|$ est donc la part de norme de Y conservée par projection sur le sous-espace vectoriel engendré par les X_j. Puisque $\|Y\|^2\geq\|XA\|^2$, ce rapport est compris entre 0 et1.

Il est égal à 1 lorsque $\|Y - XA\|^2 = 0$, c'est-à-dire, lorsque $Y = XA$. En d'autres termes, ce rapport est égal à 1 lorsque Y appartient au sous-espace vectoriel engendré par les X_j.

5) Cas $p = 2$ avec $X_{2i} = 1$ pour tout i de $\{1, \dots, n\}$.

On pose $m(X_1) = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$ et $m_2(X_1) = \frac{1}{n}\sum_{i=1}^{n}(X_{1i})^2$

On a alors :

$$X'X = n\begin{bmatrix} m_2(X_1) & m(X_1) \\ m(X_1) & 1 \end{bmatrix}$$

On a $\det(X'X) = n^2\left(m_2(X_1) - \left[m(X_1)\right]^2\right) = n^2\, V(X_1)$.

Les vecteurs X_1 et X_2 étant supposés linéairement indépendants, on en déduit que la variable réelle X_1 n'est pas une constante et donc que sa variance est non nulle. On vérifie bien que cette hypothèse entraîne l'inversibilité de la matrice $X'X$.

On a donc :

$$(X'X)^{-1} = \frac{1}{n\,V(X_1)}\begin{bmatrix} 1 & -m(X_1) \\ -m(X_1) & m_2(X_1) \end{bmatrix}$$

Par ailleurs, si on pose $m(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ et $m_{11}(X_1, Y) = \frac{1}{n}\sum_{i=1}^{n} X_{1i} Y_i$, on a :

$$X'Y = n\begin{bmatrix} m_{11}(X_1, Y) \\ m(Y) \end{bmatrix}.$$

On en déduit :

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = (X'X)^{-1} X'Y,$$

c'est-à-dire :

$$a_1 = \frac{1}{V(X_1)}\left[m_{11}(X_1, Y) - m(X_1)m(Y)\right] \text{ et}$$

$$a_2 = \frac{1}{V(X_1)}\left[-m(X_1)m_{11}(X_1, Y) + m_2(X_1)m(Y)\right]$$

En simplifiant on obtient :

$a_1 = \frac{COV(X_1, Y)}{V(X_1)}$ et $a_2 = \frac{1}{V(X_1)}\left[-m(X_1)COV(X_1, Y) + V(X_1)m(Y)\right] = m(Y) - a_1 m(X_1)$.

On retrouve bien l'équation de la droite de régression linéaire de Y en X_1, obtenue selon le critère des moindres carrés C :

$$Y - m(Y) = \frac{COV(X_1, Y)}{V(X_1)}\left(X_1 - m(X_1)\right)$$