ESTIMATEUR DES MOINDRES CARRÉS ORDINAIRES (H1, J1)
(31 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **méthode des moindres carrés ordinaires** (mco) est très souvent associée à un **modèle de régression** standard (écrit dans l'**espace des observations**) :

(1) $y = z + u$, avec $E\,u = 0$ et $V\,u = \sigma^2 . I_N$.

On appelle **estimateur des moindres carrés ordinaires**, ou **prévision des moindres carrés ordinaires**, de la **variable endogène** y la solution en z du problème de **programmation mathématique** suivant :

(2) $\inf_{z \in V} ||y - z||^2$, où $V \subset \mathbf{R}^N$.

Comme y est aléatoire et non pas « certain », la solution ne définit pas un « estimateur » au sens usuel. C'est pourquoi le terme de « prévision » lui est parfois préféré.

(i) Si (1) est un **modèle de régression linéaire**, on pose $V = V_K = \{z \in \mathbf{R}^N : z = X\,b, \forall\, b \in \mathbf{R}^K\}$. Sous l'hypothèse rg X = K, la solution de (1) est alors :

(3) $\hat{y} = X\,\hat{b}$, avec $\hat{b} = (X'\,X)^{-1}\,X'\,y$.

On appelle **estimateur des moindres carrés ordinaires**, ou **estimateur de C.F. GAUSS - A.A. MARKOV**, du paramètre b le vecteur $\hat{b}$ défini en (3).

On appelle **« estimateur » des mco de la perturbation** u, ou **prévision des mco de la perturbation** u, le vecteur $\hat{u}$ des **résidus** défini par :

(4) $\hat{u} = y - \hat{y} = y - X\,\hat{b} = M\,y = M\,u$, avec $M = I_N - X\,(X'\,X)^{-1}\,X'$.

L'estimateur des mco possède de nombreuses propriétés optimales (tant du point de vue mathématique - commodité des calculs - que statistique) qui ont fait de la méthode des moindres carrés une méthode d'estimation très utilisée :

(a) $E\,\hat{b} = b$ (**estimateur sans biais**), $V\,\hat{u} = \sigma^2\,(X'\,X)^{-1}$;

(b) si le modèle (1) est avec terme constant (ie si la (N,K)-matrice $X = [x_1, ..., x_K]$ est tq eg $x_1 = e_N$) (cf **constante**), on établit l'**équation d'analyse de la variance**, ou **théorème de PYTHAGORE** dans $\mathbf{R}^N$:

(5) $||y - \bar{y}\,e_N||^2 = ||\hat{y} - \bar{y}\,e_N||^2 + ||\hat{u}||^2$,

ainsi que les propriétés suivantes :

(6) $e_N'\,\hat{u} = 0$, $e_N'\,\hat{y} = e_N'\,y$.

On peut alors définir le **coefficient de détermination** de la régression :

(7) $R^2 = ||\hat{y} - \bar{y}\, e_N||^2 / ||y - \bar{y}\, e_N||^2 = 1 - (y' P y)^{-1} ||\hat{u}||^2 \in [0, 1]$,

qui représente le rapport entre le premier terme du membre de droite de (5) et le premier membre de (5) (expression dans laquelle P est la **matrice de centrage par rapport à la moyenne**) ;

(c) $E\,\hat{u} = 0$, $V\,\hat{u} = \sigma^2 M$, avec $M = I_N - X (X' X)^{-1} X'$;

(d) $E (\hat{b}\, u') = 0 \in M_{KN}$ (**R**) ;

(e) $E\,\hat{y} = E\,y = X b$, $V\,\hat{y} = V (X \hat{b}) = \sigma^2 (X' X)^{-1} X'$;

(f) $E (\sigma^2)\hat{} = \sigma^2$, avec $(\sigma^2)\hat{} = ||\hat{u}||^2 / (N-K)$ (estimateur des mco, sans biais, de σ^2) ;

(g) si $u \sim \mathcal{N}_N (0, \sigma^2 I_N)$ (**loi normale multidimensionnelle** centrée), alors $\hat{b} \sim \mathcal{N}_K (b, \sigma^2 (X' X)^{-1})$ et $\hat{b}$ est aussi l'**estimateur du maximum de vraisemblance** (gaussienne) de b. On montre que :

(8)
$(N - K)\, \sigma^{-2}\, (\sigma^2)\hat{} \sim \mathcal{X}^2_{N-K}$ (**loi du chi-deux** à N-K **dl**),

$(b_k\hat{} - b_k) / \sigma_k\hat{} \sim \mathcal{T}_{N-K}$, (**loi de STUDENT** à N-K **dl**),

où $\sigma_k\hat{}$ est le k-ième terme de la diagonale principale de la **matrice** :

(9) $v (\hat{b}) = (\sigma^2)\hat{} (X' X)^{-1}$,

laquelle est un **estimateur sans biais** de $V\,\hat{b}$;

(h) $\hat{b}$ vérifie le **théorème de GAUSS-MARKOV** ;

(i) si les perturbations u_n ($n = 1, ..., N$) sont indépendantes en probabilité et si (**moment algébrique** d'ordre 4) $\mu_{4n} = E\, u_n^4 < +\infty$, on a :

(10) $V (\sigma^2)\hat{} = 2 (N - K)^{-1} \sigma^4 + (N - K)^{-2} \Sigma_{n=1}^{N} (\mu_{4n} - 3 \sigma^4)\, m_{nn}^2$,

où m_{nn} est le n-ième terme de la diagonale principale de la matrice M définie en (c) ;

(j) si $u \sim \mathcal{N}_N (0, \sigma^2 I_N)$, alors :

(11) $\sigma^{-2} (\hat{b} - b)' X' X (\hat{b} - b) \sim \mathcal{X}^2_K$ (**loi du chi-deux** à K **degrés de liberté**) ;

(12) $\hat{u} \sim \mathcal{N}_N (0, \sigma^2 M)$ (**loi dégénérée** car rg M < N) ;

(13) $N^{-1} (\hat{b} - b)' X' X (\hat{b} - b) / (\hat{\sigma^2}) \sim \mathcal{F}_{K,N-K}$ (**loi de FISHER** à K et N-K dl) ;

Si, de plus, X contient un terme constant, on a :

(14) $\{(N - K) / (K - 1)\} . R^2 / (1 - R^2) \sim \mathcal{F}(K-1,N-K)$,

où R^2 est le coefficient de détermination défini en (b) ;

(k) **propriétés asymptotiques** de $\hat{b}$ (noté ici $\hat{b_N}$) :

$(k)_1$ si X_n est la n-ième ligne de X, on a la **convergence en loi** (en supposant que $n \leq N$) :

(15) $X_n (X' X)^{-1} X_n' \rightarrow_{(n,N) \rightarrow (+\infty, +\infty)} 0 \Rightarrow (X' X)^{1/2} (\hat{b_N} - b) \rightarrow^{\mathcal{L}} \mathcal{N}_K (0, \sigma^2 I_N)$;

$(k)_2$ s'il existe $\psi : \mathbf{N}^* \mapsto \mathbf{R}_+$ tq $\lim_N \psi(N) = +\infty$ et tq il existe une **matrice définie positive** $A \in D_K^{++}(\mathbf{R})$ vérifiant $X' X / \psi(N) \rightarrow_{N \rightarrow +\infty} A$, on a la **convergence en moyenne quadratique** (donc aussi la **convergence en probabilité**) :

(16) $\hat{b_N} \rightarrow^{m.q.}_{N \rightarrow +\infty} b$;

$(k)_3$ les trois convergences suivantes sont équivalentes :

$(X' X)^{-1} \rightarrow_{N \rightarrow +\infty} 0$ ((K,K)-matrice nulle),

(17) $\min_{k=1}^{K} l_{N,k} \rightarrow_{N \rightarrow +\infty} +\infty$,

$\hat{b_N} \rightarrow^{m.q.}_{N \rightarrow +\infty} b$,

où les $l_{N,k}$ désignent les **valeurs propres** de X' X ;

(l) la matrice $H = I_N - M = X (X' X)^{-1} X'$ possède les propriétés suivantes :

$(l)_1$ $\hat{y} = H y$;

$(l)_2$ $0 \leq h_{nn} \leq 1$ et $\Sigma_{n=1}^{N} h_{nn} = K$;

$(l)_3$ $h_{nn} + ||u||^{-2} (\hat{u_n})^2 \leq 1, \forall n \in N_N^*$;

$(l)_4$ si le vecteur $\xi = (\xi_1, ..., \xi_K)'$ des **variables exogènes** est gaussien, ie si $\xi \sim \mathcal{N}_K (\mu_\xi , \Sigma_\xi)$, alors :

(18) $(K - 1)^{-1} (1 - h_{nn})^{-1} (N - K) (h_{nn} - N^{-1}) \sim \mathcal{F}(K-1, N-K)$ (**loi de FISHER**) ;

$(I)_5$ $V\,\hat{u} = \sigma^2 (I_N - H)$ et $V\,\hat{u_n} = \sigma^2 (1 - h_{nn})$.

(ii) Si le modèle (de régression) (1) est un **modèle non linéaire**, on pose $V = V_Q = \{z \in \mathbf{R}^N : z = F(b), \forall\, b \in \mathbf{R}^Q\}$ et la solution des mco $\hat{b_F}$ s'obtient comme solution du programme mathématique :

(19) $\inf_{b \in \mathbf{R}Q} (y - F(b))' (y - F(b))$.

Cet **estimateur des moindres carrés ordinaires non linéaire** $\hat{b_F}$ de b ne s'exprime pas, en général, de façon explicite en fonction de F (qui dépend de X) et de y.

Les méthodes de calcul numérique utilisées sont donc itératives. On procède souvent par **linéarisation** (supposée licite) de F au voisinage d'une valeur (d'amorçage) donnée de b, soit $b^{(0)}$, ce qui conduit au premier **modèle linéarisé** :

$(20)_{(0)}$ $z^{(0)} = F(b^{(0)}) + D\,F(b^{(0)}) \,.\, (b - b^{(0)})$,

qui permet d'estimer b par $b^{(1)}$ en appliquant la méthode des mco (linéaires) pour estimer b ; puis, on « rentre » cette nouvelle valeur dans $(20)_{(0)}$ et l'on estime le modèle :

$(20)_{(1)}$ $z^{(1)} = F(b^{(1)}) + D\,F(b^{(1)}) \,.\, (b - b^{(1)})$,

par la méthode des mco. La p-ième itération conduit ainsi à :

$(20)_{(p)}$ $z^{(p)} = F(b^{(p)}) + D\,F(b^{(p)}) \,.\, (b - b^{(p)})$.

Cette méthode d'estimation itérative conduit aussi à « prévoir » y et u. Ses deux principaux inconvénients sont :

(a) la difficulté de localiser une **valeur initiale** (ou **valeur d'amorçage**) pour b (l'espace $\mathbf{R}^Q$ auquel b appartient est « grand ») qui permette d'assurer la convergence (numérique) de $b^{(p)}$ vers une valeur « plausible » $b^{(\infty)}$ lorsque $p \to +\infty$. L'**homme de l'art** peut, parfois, suggérer une telle valeur au **statisticien** ;

(b) la possibilité de solutions multiples de l'un des programmes :

(21) $\inf_{b \in \mathbf{R}Q} ||y - z^{(p)}||^2$

à résoudre (où **R**Q désigne $\mathbf{R}^Q$). Cette éventualité dépend de la forme de F, ie de la variété V_Q .

Les propriétés à distance finies ($N << +\infty$) de l'estimateur $\hat{b_F} = b^{(\infty)}$ de b ainsi calculé (en fait, seulement « approché ») sont peu connues, car elles dépendent en général de la « forme » de chaque équation de régression. Elles peuvent être étudiées par **simulation**. En général, $\hat{b_F}$ est un estimateur biaisé.

Comme dans le cas linéaire, $b_F^{\wedge}$ est aussi l'estimateur du maximum de vraisemblance (gaussienne) de b lorsque u (ou y) est gaussienne. Ses propriétés asymptotiques sont donc généralement optimales.

(iii) Si, dans le modèle (1) supposé linéaire, rg X = K - L (avec $L > 0$), X' X n'est pas inversible et la méthode tombe en défaut (cf **singularité**). Pour estimer b par la méthode des mco (linéaire), on doit introduire une restriction, qui est généralement de la forme :

(22) $H\, b = 0$, avec $H \in M_{LK}(\mathbf{R})$ et rg H = L.

L'estimateur des mco $b^{\wedge}$ de b est alors solution du système (en β) :

(23)
$$X' X \beta = X' y \quad \text{(cf \textbf{équations normales})},$$
$$H \beta = 0,$$

d'où l'estimateur :

(24) $b_H^{\wedge} = (X' X + H' H)^{-1} X' y.$

On montre que :

(25)
$$b_H^{\wedge} = \sigma^2 (X' X + H' H)^{-1} X' X (X' X + H' H)^{-1},$$
$$E\, q(b_H^{\wedge}) = (N + K - L)\, \sigma^2,$$

où $q(b_H^{\wedge}) = ||y||^2 - y' X b_H^{\wedge}$ est la somme des carrés des résidus (cf **forme quadratique résiduelle**).

Cette situation de singularité se rencontre souvent avec un **modèle d'analyse de la variance** ou un **modèle d'analyse de la covariance**. La solution mathématique fait donc usage des notions de **matrice inverse généralisée** ou de **matrice pseudo-inverse**.

(iv) A titre d'exemple, le modèle de régression linéaire le plus simple s'écrit (modèle « constant ») :

(26) $y = b \,.\, e_N + u.$

La **moyenne empirique** vaut :

(27) $\bar{y}_N = N^{-1} e_N' y = b\, N^{-1} e_N' e_N + N^{-1} e_N' u,$

ie :

(27)' $\bar{y}_N = b + \bar{u}_N$, avec $\bar{u}_N = N^{-1} e_N' u.$

Par suite :

(a) si E u / e_N = E u = 0, alors E $\bar{y}_N$ = b ;

(b) si V u / e_N = V u = σ^2 . I_N , alors V $\bar{y}_N$ = V $\bar{u}_N$ = N^{-1} . σ^2.

La moyenne empirique $\bar{y}_N$ constitue ici un estimateur (naturel) sans biais de b (cf **statistique naturelle**), dont la variance tend asymptotiquement vers zéro (cf **échantillonnage**, **tirage bernoullien**).