HÉTÉROSCÉDASTICITÉ (C5, J)
(09 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion d'**hétéroscédasticité** (K. PEARSON) traduit le fait que des **variables** ou des **observations** peuvent être :

(a) indépendantes (ou non corrélées) (cf **indépendance stochastique**, **dépendance**, **corrélation**) ;

(b) sans cependant être homogènes entre elles, en termes de **variabilité propre**, ou **variabilité « intrinsèque »** (cf **homogénéité**, **scédasticité**, **test d'hétéroscédasticité**).

(i) Soit $X = \{(\Omega, \mathcal{T}, P), (\mathbf{R}, \mathcal{B}_{\mathbf{R}}), (X_t)_{t \in T}\}$ un **processus stochastique** réel scalaire de carré intégrable (ie $X_t \in L_{\mathbf{R}}^2 (\Omega, \mathcal{T}, P)$, $\forall\ t \in T$).

On dit que X est un **processus hétéroscédastique** ssi il existe (au moins) une **suite** finie $(t_1, ..., t_N)$ d'indices $t_n \in T$ distincts tq la **matrice des covariances** $V_{t(1)... t(N)}$ du **vecteur aléatoire** $(X_{t(1)}, ..., X_{t(N)})'$ soit diagonale sans être scalaire (où t(n) désigne t_n). Le plus souvent, le processus X est indépendant (mais non équidistribué) (cf **processus purement aléatoire**).

Le principal problème peut consister :

(a) à estimer des **matrices** tq $V_{t(1)... t(N)}$, dont le nombre d'éléments diagonaux varie avec le nombre N d'observations ;

(b) ou à tester une **hypothèse d'hétéroscédasticité**.

Lorsque les termes diagonaux $V\ X_{t(n)}$ dépendent (de façon connue) de n (ou, plus généralement, d'un processus auxiliaire $Z = (Z_{t(n)})_{n \in \mathbf{N}^*}$), l'**inférence statistique** est possible : ainsi en est-il si $V\ X_{t(n)} = f(n)$ (resp si $V\ X_{t(n)} = F(Z_{t(n)})$, $\forall\ n \in \mathbf{N}^*$), où f (resp F) est connue (ou dépend d'un nombre fixé de « **paramètres** »).

(ii) L'**hétéroscédasticité** peut parfois s'analyser comme une propriété selon laquelle un processus $X = (X_t)_{t \in T}$ résulte d'un tirage aléatoire selon un **mélange légal** $P^{\xi} = \Sigma_{i \in I}\ \alpha_i\ P_i^{\xi}$, les va X_t étant indépendantes (entre elles) et les lois composantes P_i^{ξ} ne différant entre elles que par leurs variabilités (ou **variances**) propres, $V\ \xi_i$, $\forall\ i \in I$. Une « proportion » α_i des va X_t possède ainsi (asymptotiquement) une variance $V\ \xi_i$. le problème consiste à « identifier » le mélange en question (du moins au second ordre) (cf **dissection d'un mélange de lois**).

(iii) Un contexte dans lequel la notion intervient souvent est celui du modèle de régression. Soit :

(1) $y = F(b) + u$, avec $E\ u = 0$,

un **modèle de régression** non linéaire multiple, écrit dans l'**espace d'observation** $\mathbf{R}^N$.

On dit qu'il y a **hétéroscédasticité des perturbations** u_n , ou que la **perturbation** (vectorielle) u est hétéroscédastique, ou parfois aussi que les perturbations sont hétérogènes au second ordre (cf **hétérogénéité**), ssi V u est une **matrice diagonale** mais non scalaire, ie ssi, à la fois :

$$(2) \quad \begin{cases} V\,u \in D_N(\mathbf{R}), \\ V\,u \neq \sigma^2 . I_N \end{cases}$$

(ie il n'existe pas de nombre réel $\sigma > 0$ tq $V\,u = \sigma^2 . I_N$). Autrement dit, on peut écrire V u sous la forme explicite :

$$(3) \quad \begin{cases} C(u_\alpha , u_\beta) = 0 & \text{si } \beta \neq \alpha, \\ C(u_n , u_n) = V\,u_n = \sigma_n^2 & \text{sinon, } \forall\, n \in N_N^*, \end{cases}$$

où $n \mapsto \sigma_n$ n'est pas une **suite constante**.

Ainsi, on rencontre des situations selon lesquelles :

$$(4) \quad \sigma_n^2 = g(Z_n , \theta),$$

où Z_n est la n-ième ligne d'une (N,L)-matrice Z constituée de N observations de L variables ζ_1 ,..., ζ_L , g une fonction strictement positive et θ un paramètre à déterminer. Les variables peuvent figurer dans la liste des **variables exogènes** du modèle (1) (on parle parfois d'**exoscédasticité**) ; si L = 1 et Z = y, on parle d'**endoscédasticité** (la matrice de dispersion V u dépend alors d'un vecteur aléatoire y).

L'**estimation** d'un modèle avec hétéroscédasticité relève généralement de la **méthode des moindres carrés généralisés** ou de la **méthode du maximum de vraisemblance**.

Il existe des **tests d'hétéroscédasticité** préalables qui permettent d'en déceler l'existence. A l'inverse, un **test d'homoscédasticité** est un test de l'**hypothèse d'homoscédasticité**, ie de l'hypothèse $H_0 : V\,u = \sigma^2 . I_N$ (ie $\sigma_1^2 = ... = \sigma_N^2 = \sigma_u^2$). Un tel test est souvent fondé sur l'examen du graphe $(y_n , \hat{u}_n)_{n=1,..,N}$, où $\hat{u} = (\hat{u}_n)_{n=1,..,N}$ est le **résidu** des moindres carrés ordinaires (mco) appliqués au modèle (1).

Un exemple de test simple est le suivant, où l'on suppose que le modèle (1) est linéaire, avec Q = K paramètres $b = (b_1 ,..., b_K)$ à estimer) (cf aussi **test de sphéricité**) :

(a) **partition** de l'**échantillon** (X, y) en deux sous-échantillons indépendants (X_A , y_A) et (X_B , y_B) de tailles resp M et N - M ;

(b) application de la **méthode des mco** à chacun d'eux, ce qui fournit des résidus indépendants, $u_A^{\wedge}$ et $u_B^{\wedge}$;

(c) utilisation de la propriété suivante, valable sous les hypothèses de **normalité** $u_A^{\wedge} \sim \mathcal{N}_M (0, \sigma_A^2 I_M)$ et $u_B^{\wedge} \sim \mathcal{N}_{N-M} (0, \sigma_B^2 I_{N-M})$ et d'homoscédasticité H_0 :

(5) $(N - M - K) . (N - K)^{-1} . ||u_B^{\wedge}||^{-2} ||u_A^{\wedge}||^2 \sim \mathcal{F} (M-K, N-M-K)$

(**loi de FISHER-SNEDECOR** à M-K et N-M-K **degrés de liberté**), ce qui fonde le test en question.

Si l'hétéroscédasticité est avérée, l'estimation de b nécessite alors la connaissance de V u :

(a) si V u (resp g et θ) est connue, on utilise la **méthode des mcg**, qui est équivalente à la méthode des mco appliquée au modèle homoscédastique suivant, parfois appelé **modèle sphéricisé** :

(6) $(V u)^{-1/2} y = (V u)^{-1/2} X b + (V u)^{-1/2} u,$

dans le cas d'un modèle linéaire ;

(b) sinon, on utilise les alternatives correspondantes de la méthode des mcg : **méthode du maximum de vraisemblance**, **méthode des moindres carrés quasi-généralisés**, **méthodes à distance minimale**.

Il arrive souvent qu'une variable ζ, observée selon $z = (z_1 ,..., z_N) \in \mathbf{R}^N$, soit corrélée avec u (en fait, avec $u^{\wedge}$). S'il existe une relation de la forme :

(7) $V u_n = \sigma^2 . h (z_n),$

où h est une fonction positive donnée, on obtient :

(8) $\Omega = \sigma^{-2} V u = \text{Diag} \{h (z_1) ... h (z_N)\}.$

(iv) Enfin, l'existence d'une hétéroscédasticité est à distinguer de l'existence d'**aberrations** dans un ensemble de données, lorsque ces aberrations tiennent au fait qu'une proportion donnée des observations possède une variance importante (cf **aberration**, **mélange de lois**).