MODÈLE A ERREURS SUR LES VARIABLES (G11, J)
(08 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

Lorsqu'une **représentation statistique** inclut des variables observées comportant des **erreurs** aléatoires, l'**inférence statistique** doit être adaptée.

(i) En effet, une **erreur aléatoire** est généralement de l'un des types suivants :

(a) erreur d'**observation** ou erreur de **mesure** proprement dite : celle-ci peut affecter tout, ou partie, des variables ou des « **donnée** » observées ;

(b) erreur d'**échantillonnage** : celle-ci affecte, par construction, l'ensemble des données. Le **statisticien** organise certaines **procédures statistiques** en sorte que ce type d'erreur soit sous contrôle ;

(c) une **erreur de concept** : l'un des concepts utilisés dans le modèle se traduit par le choix d'une **variable** qui n'est qu'un concept « proche » (« **proxy** ») ou « voisin » du « vrai » concept (cf **modèle**, **variable proche**, **spécification**). Cette proximité peut parfois être « large » (concept erroné, variable inadéquate).

(ii) Ainsi, un **modèle image** de la forme $(\mathcal{Z}, \mathcal{D}, \mathcal{P}^{\zeta})$ dans lequel ζ est observé avec erreur peut s'interpréter comme un modèle inapproprié, le véritable modèle étant $(\mathcal{Z}, \mathcal{D}, \mathcal{P}^{\zeta^*})$, où ζ^* désigne la variable sans erreur. Dans cette **situation statistique**, la **modélisation** de la relation entre ζ et ζ^* joue donc un rôle important.

Le cadre adapté à cette prise en compte est celui du **mélange légal** : chaque loi $P^{\zeta} \in \mathcal{P}^{\zeta}$ régissant l'**observation** peut donc se décomposer sous une forme tq :

(0) $\quad P^{\zeta} = (1 - \varepsilon) . P^{\zeta^*} + \varepsilon . P^{\upsilon}$,

où $\varepsilon \in]0, 1[$ (**taux d'erreur**) et où désigne d'une variable perturbatrice (ou contaminante) υ, à valeurs dans le même espace $(\mathcal{Z}, \mathcal{D})$ (cf aussi **perturbation aléatoire**).

(iii) Dans un **modèle statistique**, on distingue souvent entre **variables exogènes** ξ et **variables endogènes** η (cf **relation fonctionnelle**, **modèle de régression**, **modèle d'interdépendance**), ce qui conduit à « partitionner » ζ eg selon (ξ, η), ξ étant à valeurs dans un ensemble $\mathcal{X}$ et η dans un ensemble $\mathcal{Y}$, avec $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Par suite, (0) s'écrit sous la forme :

(1) $\quad P^{(\xi, \eta)} = (1 - \varepsilon) . P^{(\xi^*, \eta^*)} + \varepsilon . P^{\upsilon}$,

dans laquelle υ est à valeurs dans l'espace $(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{C})$, où $\mathcal{B}$ (resp $\mathcal{C}$) est une **tribu de parties** de $\mathcal{X}$ (resp de $\mathcal{Y}$).

Si l'erreur affecte seulement certaines des variables du modèle, le formalisme se simplifie, et υ peut n'affecter que ces variables (**erreur spécifique**).

(iv) Au lieu de définir l'erreur au niveau de la **famille** des **lois de probabilité** sous-jacente au **phénomène**, on la modélise souvent directement au niveau des variables (**espace des variables**) ou au niveau des observations (**espace des observations**), sachant qu'il en existe une loi de probabilité sous-jacente.

Ainsi, soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** de base, $(\mathcal{X}, \mathcal{B})$ un **espace d'observation**, $X : \Omega \mapsto \mathcal{X}$ une **va** (ou un **échantillon**) **observable**, $(\mathcal{X}^*, \mathcal{B}^*)$ un **espace d'inobservation** (ie un espace de valeurs **inobservables**) et $X^* : \Omega \mapsto \mathcal{X}^*$ une va (ou un échantillon) inobservable.

On dit alors que $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$, **modèle image** du modèle de base par X, est un **modèle à erreur sur la variable** X, ou **modèle à variable erronée**, ssi le « vrai » modèle est $(\mathcal{X}^*, \mathcal{B}^*, \mathcal{P}^{X^*})$ et s'il existe une application mesurable, dite **fonction d'erreur**, $\varphi : \mathcal{X}^* \mapsto \mathcal{X}$ tq :

(2) $\quad X = \varphi(X^*)$ et $\varphi(X) = X$.

La première relation $X = \varphi(X^*)$ exprime un lien entre la variable observable X et la variable inobservable X^*. La seconde relation $\varphi(X) = X$ exprime que le lien précédent se réduit à une **identité** lorsque X^* est observée sans erreur (ie lorsque $X = X^*$).

(v) Plus généralement, le lien φ peut aussi lui-même être considéré comme aléatoire, ie de la forme eg $\psi : \mathcal{X}^* \times \mathcal{E} \mapsto \mathcal{X}$, où $(\mathcal{E}, \mathcal{B}_{\mathcal{E}})$ est un **espace mesurable**, que l'on peut appeler **espace d' « erreur »**. On suppose alors qu'il existe une va $E : \Omega \mapsto \mathcal{E}$, elle-même inobservable et appelée **erreur sur la variable**, tq :

(3) $\quad X = \varphi(X^*) = \psi(X^*, E)$,

avec $\varphi(X^*) = \psi(X^*, 0) = X^*$ lorsque la notion d' « **erreur nulle** » a un sens (eg si $\mathcal{E}$ est un **groupe algébrique** ou un **espace vectoriel**).

(vi) Le plus souvent, $(\mathcal{E}, \mathcal{B}_{\mathcal{E}}) = (\mathcal{X}^*, \mathcal{B}^*)$ est un **espace vectoriel**, et ψ est de forme additive, ie $X = \psi(X^*, E) = X^* + E$: on parle alors d'**erreur additive sur la variable**. La variable X (resp X^*) peut être un **vecteur aléatoire**, ou une « liste » constitué(e) de plusieurs va, eg des **variables endogènes** et des **variables exogènes** : on parle alors de **modèle à erreurs sur les variables**. Elle peut aussi (cas général) être un ensemble d'observations portant sur ces variables. Par suite, au lieu d'étudier le modèle « inobservable » $(\mathcal{X}^*, \mathcal{B}^*, \mathcal{P}^{X^*})$, on se contente d'étudier le modèle observable $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ (cf aussi **coefficient d'atténuation**, **robustesse**, **courbe de sensibilité**, **courbe d'influence**).

(vii) La fonction ψ est généralement connue ou donnée (eg linéaire ou, le plus souvent, additive). Le problème formel est alors simple si l'on suppose que ψ est partiellement inversible (ie inversible pr à son premier argument X*) (cf **application inverse**, **fonction inverse**) : dans ce cas, on peut écrire $X^* = \psi^*(X, E)$ (grandeur alors observable et calculable) et l'inférence statistique porte sur la **loi** de X*, ou seulement sur une **caractéristique légale** (eg **relation fonctionnelle**, **fonction de régression** ou d'**interdépendance**). Cette loi est une loi $\mathscr{L}(X^*)$, image par ψ^* de l'une des lois $P^X \in \mathscr{P}^X$ et de la loi propre $\mathscr{L}(E)$ de E. Lorsque $\mathscr{P}^X = (P_\theta{}^X)_{\theta \in \Theta}$ (modèle paramétré), $\mathscr{L}(E)$ est généralement supposée indépendante de θ.

(viii) En pratique, des modèles couramment associés à des erreurs sur les variables sont le **modèle de régression**, le **modèle d'analyse de la variance** ou le **modèle d'analyse de la covariance**, et le **modèle d'interdépendance**.

(ix) Dans le cas d'un **modèle de régression linéaire** multiple (dans l'**espace des variables**) :

$$(4) \qquad \eta^* = (\xi^*)' b + \varepsilon^*, \qquad \text{avec } E\,\varepsilon^* = 0,\ V\,\varepsilon^* = \sigma^2,$$

η^* représente une **vars** endogène inobservable, ξ^* un vecteur de vars exogènes inobservables et ε^* une **perturbation aléatoire** (inobservable), appelée **erreur sur l'équation** (4). On observe alors les variables ξ et η, correspondant resp aux précédentes et tq :

$$(5) \qquad \begin{aligned} \xi &= \xi^* + \varphi, && \text{avec } E\,\varphi = 0, \\ \eta &= \eta^* + \psi, && \text{avec } E\,\psi = 0. \end{aligned}$$

De plus, on admet souvent les hypothèses de non **corrélation** suivantes :

$$(3) \qquad \begin{aligned} &C(\xi^*, \varphi) = 0, \quad C(\eta^*, \psi) = 0, \quad C(\xi^*, \varepsilon^*) = 0, \\ &C(\varphi, \varepsilon^*) = 0, \quad C(\psi, \varepsilon^*) = 0, \quad C(\varphi, \psi) = 0. \end{aligned}$$

La relation « observable » se déduit alors de {(4),(5)} :

$$(6) \qquad \eta = \xi' b + \varepsilon, \qquad \text{avec } \varepsilon = \varepsilon^* + \psi - \varphi' b.$$

Si (X, y) représente les N observations du **couple aléatoire** (ξ, η), le modèle précédent admet la **forme « initiale »** suivante (dans l'**espace d'observation**) :

$$(7)_a \qquad y^* = X^* b + u^*, \qquad \text{avec } E\,u^* = 0, \qquad C(X^*, u^*) = 0,$$

dans laquelle :

$$X = X^* + V, \qquad \text{avec } E\,V = 0,$$

$$y = y^* + w, \qquad \text{avec } E\,w = 0,$$

$(7)_b$

$$C(X^*, V) = 0, \qquad C(y^*, w) = 0,$$

$$C(V, u^*) = 0, \qquad C(w, u^*) = 0, \qquad C(V, w) = 0.$$

Sous la forme $\{(7)_a , (7)_b\}$, ce modèle admet pour **paramètre d'intérêt** le triplet (b, X^*, y^*) : ce paramètre augmente avec le nombre N des observations (cf **paramètre incident**).

La **forme « finale »** du modèle s'écrit :

$$(8) \qquad y = X b + u, \qquad \text{avec } u = u^* - V b + w.$$

Bien que $E u = 0$, la **méthode des moindres carrés** (ordinaires ou généralisés) n'est pas applicable car :

$$(9) \qquad C(\xi, \varepsilon) = -(V \varphi) b \neq 0 \quad \text{(sur la forme initiale)},$$

$$C(X, u) = -(V V) b \neq 0 \quad \text{(sur la forme finale)}.$$

L'estimateur des moindres carrés est alors biaisé et divergent en probabilité (cf **biais**, **convergence en probabilité**).

Deux méthodes d'estimation de b sont principalement utilisées : la **méthode des variables instrumentales** et une **méthode à distance minimale** (cf **régression pondérée**).

(x) Dans le cas d'un **modèle d'interdépendance linéaire** sous la forme implicite :

$$(10) \qquad B \eta^* + C \xi^* = \varepsilon^*, \qquad \text{avec } E \varepsilon^* = 0,$$

on suppose que :

$$(11) \qquad \xi = \xi^* + \varphi, \qquad \text{avec } E \varphi = 0, \qquad C(\xi^*, \varphi) = 0,$$

$$\eta = \eta^* + \psi, \qquad \text{avec } E \psi = 0, \qquad C(\eta^*, \psi) = 0.$$

Par suite, en insérant (9) dans (8), on obtient :

$$(12) \qquad B \eta + C \xi = \varepsilon, \qquad \text{avec } \varepsilon = B \psi + C \varphi + \varepsilon^*,$$

modèle d'interdépendance dont la perturbation ε dépend du couple de paramètres d'intérêt (B, C).

On peut développer des calculs semblables aux précédents, à partir du modèle « observé » correspondant à un N-**échantillon** (X, Y) de (ξ, η).