PERTURBATION ALÉATOIRE (C1, D1, E, J, N)
(27 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'expression générale **perturbation aléatoire** désigne une **variable aléatoire** inobservable (cf **inobservable**, **variable observable**) qui agit :

(a) soit sur une équation (**« erreur » sur l'équation**) ;

(b) soit sur une autre **va** (**« erreur » sur la variable**).

On l'appelle aussi **erreur aléatoire**, **variable latente** ou **résidu stochastique**.

Cette terminologie intervient dans la construction de divers modèles de **relation fonctionnelle** : **modèle de régression**, **modèle d'interdépendance**, **modèle à erreurs sur les variables**.

Une perturbation aléatoire est alors censée « résumer » tous les **effets non pris en compte** dans le modèle ou dans les variables utilisées : simplification du modèle pr à la description, différences de concepts, etc. Sa spécificité est d'être inobservable et sa justification peut se concevoir selon deux lignes directrices, en pratique équivalentes (cf **phénomène**, **loi**, **loi scientifique**, **système**).

On présente ici ces deux lignes dans le cadre d'un **modèle de régression multiple** linéaire à perturbation additive.

(i) L'**homme de l'art** conçoit initialement un ***modèle déterministe*** (écriture dans l'**espace des variables**) :

(1) $\eta = \sum_{k=1}^{K} b_k \xi_k = \xi' b,$

visant à expliquer une **vars** endogène η à l'aide d'une combinaison linéaire de vars exogènes ξ_k , $\forall$ k $\in$ N_K^* (cf **variable endogène**, **variable exogène**, **loi**, **loi scientifique**).

On dispose de N **observations** y de η et X de ξ, y étant à valeurs dans $\mathbf{R}^N$ et X à valeurs dans M_{NK} (**R**), avec K < N (cf **degré de liberté**). Si l'équation (1) était vérifiée par toutes les observations (indicées par n $\in$ N_N^*), on aurait, dans l'**espace des observations** :

(2) $y_n = X_n b, \quad \forall\ n \in N_N^*$ (ie sous forme vectorielle : y = X b),

et b pourrait être déterminé par le **système linéaire** constitué avec K quelconques des observations (X_n , y_n), eg les K premières :

(3) $b^*(K) = X(K)^{-1} y(K),$

avec $X(K) = [X_1 / ... / X_K]$ (matrice en colonne) et $y(K) = (y_1 ... y_K)'$ (vecteur colonne).

L'**estimation** de b à l'aide de (X, y) dépend a priori du sous-ensemble des K observations choisies parmi les N : a priori, l'**estimateur** résultant de la **procédure** d'estimation n'est pas unique. Pour tenir compte de toutes les observations disponibles (indicées par $n \in N_N^*$), une méthode peut consister à calculer eg la **moyenne** (ou toute autre **caractéristique** de **centralité**) empirique des C_N^K valeurs ainsi obtenues, ou encore à extraire des ***observations « douteuses »*** (cf **aberration**, **erreur**, **modèle à erreurs sur les variables**) et à recommencer la procédure.

Les propriétés statistiques d'un tel estimateur peuvent exiger des calculs combinatoires complexes. Par ailleurs, cette démarche ne s'étend pas simplement à d'autres **situations statistiques** : eg **modèle de régression** non linéaire.

En pratique, pour diverses raisons (**erreur** de **mesure**, d'**observation**, de **spécification**, omission de variables exogènes, etc), l'équation (1) n'est généralement pas vérifiée par toutes les observations.

(ii) Aussi le **statisticien** réécrit-il (1) sous une **forme à perturbation additive** (dans l'espace des variables) (cf **relation fonctionnelle**) :

(4) $\eta = \Sigma_{k=1}^{K} b_k \xi_k = \xi' b + \varepsilon = \xi' b + \varepsilon,$

qui peut s'interpréter comme un changement (affine) de **variable aléatoire** $\varepsilon \mapsto \eta$ (si ξ est considéré comme fixé, ou sinon conditionnellement à ξ) et reçoit ainsi un contenu probabiliste simple et direct : on passe d'une **variable inobservable** ε à une **variable observable** η à travers une **application affine** dont la « **constante** » est ici une **forme linéaire** qui dépend d'une variable (observable) ξ par l'intermédiaire d'un **paramètre** (inobservable car inconnu a priori) b.

Corrélativement, (2) se réécrit (dans l'**espace des observations**) :

(5) $y_n = X_n b + u_n , \quad \forall n \in N_N^*$ (ie sous forme vectorielle : $y = X b + u$),

les variables u_n (resp y_n) étant supposées indépendantes ou non. Les variables ε, u_1 ,..., u_N sont appelées **perturbations** (ou **erreurs**) **aléatoires sur l'équation**. Si l'on suppose eg que leur moyenne est nulle (ie que $E(\varepsilon / \xi) = 0$), c'est essentiellement pour signifier que l'***effet moyen*** des variables ξ_1 ,..., ξ_K est « correctement » représenté par la grandeur $\xi' b = E(\eta / \xi)$ (cf **relation fonctionnelle**, **fonction de régression**, **régression**).

(iii) La seconde justification vient de l'hypothèse selon laquelle une **famille** donnée de **variables exogènes** est censée « expliquer » (eg sous forme linéaire) une **variable endogène** η, mais la théorie retenue (ou diverses limitations tenant au nombre K de variables exogènes ou N d'observations disponibles) se restreint à n'introduire (après ***sélection***) que K de ces variables exogènes.

Soit alors $L \neq \varnothing$ un ensemble d'**indices**, $b = (b_l)_{l \in L}$ une famille de scalaires réels, V un **espace normé** réel et $(\xi_l)_{l \in L}$ une famille de vars à valeurs dans V.

Par définition, $(b_l\ ,\ \xi_l)_{l\ \in\ L}$ est une **famille sommable** et de somme η ssi, en notant $\mathcal{B}(L)$ l'ensemble des **parties** finies de L et $z_M = \Sigma_{l\ \in\ M}\ b_l\ \xi_l$, on a :

$$\forall\ \varepsilon > 0,\ \exists\ L_\varepsilon \in \mathcal{B}(L)\ \ tq\ :\ L_\varepsilon \subset M \text{ et } M \in \mathcal{B}(L)\ \Rightarrow\ ||\eta - \zeta_M|| \leq \varepsilon.$$

On note alors :

(6) $\quad \eta\ =\ \Sigma_{l\ \in\ L}\ b_l\ \xi_l\ .$

Si seulement $K \geq 1$ variables $\xi_1\ ,...,\ \xi_K$ sont retenues et si l'application $\eta : \Omega \mapsto V$ définie par (6) définit une véritable va, alors (6) se réécrit (**espace des variables**) :

(7) $\quad \eta\ =\ \Sigma_{k=1}^{K}\ b_k\ \xi_k + \Sigma_{j=1}^{J}\ b_j\ \xi_j\ =\ \Sigma_{k=1}^{K}\ b_k\ \xi_k + \varepsilon\ =\ \xi'\ b + \varepsilon,$

où l'on note $N_K^* = \{1,..., K\}$ certains indices de L, $J = L \setminus N_K^*$ les indices restants et :

(8) $\quad \varepsilon\ =\ \Sigma_{j=1}^{J}\ b_j\ \xi_j\ .$

Lorsque le modèle (7) est bien spécifié (ie notamment lorsque les variables ξ_k représentent bien l'effet « moyen » exercé sur la va η), on peut écrire :

(9) $\quad \xi'\ b\ =\ E\ (\eta\ /\ \xi),\qquad$ avec $\xi = (\xi_1\ ,...,\ \xi_K)'$,

ce qui implique :

(10) $\quad E\ (\varepsilon\ /\ \xi)\ =\ E\ (\Sigma_{j=1}^{J}\ b_j\ \xi_j\ /\ \xi)\ =\ 0.$

La perturbation aléatoire ε ainsi définie est donc nulle en moyenne, quelles que soient les valeurs des b_j .

(iv) On remarque que :

(a) dans les deux présentations précédentes, on considère d'emblée les variables $\xi_1\ ,...,\ \xi_K$ comme aléatoires. Cependant, seules certaines **caractéristiques conditionnelles** de η (souvent l'**espérance**) sont nécessaires. Le cas où ξ est un vecteur (presque) certain est donc un cas particulier ;

(b) dans le cas d'un modèle correctement spécifié (cf **spécification**), on dit parfois, que $\xi'\ b$ représente l'effet moyen « majeur » exercé sur η, ie que (cf aussi **test de HAUSMANN**, **test de spécification**) :

(11) $\quad E\ (\xi'\ b\ /\ \xi)\ =\ \xi'\ b\ \neq\ 0$, tandis que $E\ (\varepsilon\ /\ \xi)\ =\ 0.$

Une autre interprétation, nécessitant l'introduction de la notion de **variance** conditionnelle, conduit à ajouter à (11) l'hypothèse :

(12) $V(\xi' b / \xi) = V(\xi' b) >> V(\varepsilon / \xi)$

(hypothèse non vérifiée si ξ est certaine) ;

(c) si J est dénombrable (eg si J = **N** \ N_K*) et si V = $\mathbf{R}^p$, on admet souvent que la va ε, définie en (8) selon une **formalisation** linéaire, est générée par une suite $(\xi_i)_{i \in J}$ de va satisfaisant aux hypothèses du **théorème de la limite centrale**. Ceci peut justifier l'hypothèse de **normalité** (ou de **normalité asymptotique**) souvent faite sur ε (donc sur η), notamment pour effectuer un **test d'hypothèses**.

On peut concevoir d'autres lois limites (cf **loi asymptotique**), mais les hypothèses de linéarité précédentes, liées au caractère bilinéaire de la forme $(\xi, b) \mapsto \xi' b$ s'y conforment moins naturellement ;

(d) on appelle parfois **résidu** la variable ε (resp u_n) précédente. On distingue souvent entre **perturbation** et **résidu** : le résidu est une variable observable (ou estimée), ie calculable après **estimation** du modèle. Si, dans l'exemple ci-dessus, une méthode conduit à estimer b à l'aide d'un **estimateur ponctuel** $b^\#$, alors y est « estimée » (ou « prévue ») à l'aide de la va $y^\# = X b^\#$, et le résidu sera défini par la va « observable » (ie calculable) $u^\# = y - y^\#$;

(e) on qualifie ε (resp u) d'**erreur sur l'équation** pour la distinguer de la notion d'**erreur sur une variable**. Cette dernier type d'erreur peut cependant recevoir des justifications analogues aux précédentes ;

(f) ce qui précède est commun à d'autres modèles statistiques : cf eg **relation fonctionnelle**, **modèle non linéaire**, **modèle d'interdépendance**.

(v) D'un point de vue formel, dès qu'une « liste » $(\xi, \eta) = (\xi_1, ..., \xi_K, \eta)$ constituée de K+1 va est donnée (ou retenue), on peut toujours écrire l'**espérance conditionnelle** de η pr à ξ (si elle existe) selon la décomposition triviale (tautologie) :

(13) $\eta = E(\eta / \xi) + \{\eta - E(\eta / \xi)\} = E(\eta / \xi) + \varepsilon.$

Cette décomposition est licite car l'opérateur **espérance** E est linéaire, donc compatible avec une décomposition additive. Elle joue un rôle important en **Statistique** (cf eg **analyse générale des données**). Elle conduit à définir ε comme perturbation aléatoire centrée (cf **variable centrée**).

Si (X, y) est une observation du **couple aléatoire** (ξ, η), on peut écrire une décomposition analogue à (13) sous forme vectorielle (espace des observations) :

(14) $y = E(y / X) + \{y - E(y / X)\} = E(y / X) + u.$

Un aspect essentiel de la **modélisation** (cf « ***hasard et nécessité*** », **hasard**, **Statistique et hasard**), consiste alors à exprimer, le cas échéant, l'effet moyen conditionnel E (y / X) en fonction de X, eg :

(a) E (y / X) = X b pour un **modèle linéaire** ;

(b) E (y / X) = F (b) pour un **modèle non linéaire**,

ainsi que diverses **hypothèses statistiques** (outre celle du premier ordre E (u / X) = 0 portant sur la perturbation aléatoire u) : eg V (u / X) = Σ, ou encore u ~ $\mathcal{N}_N$ (0, Σ) (**loi normale multidimensionnelle** centrée), etc.

(vi) Dans le cadre de la **fonction de régression**, et plus généralement d'une **relation fonctionnelle**, le concept de perturbation peut recevoir deux interprétations importantes, dont le **contexte statistique** est différent, et qu'il convient de distinguer :

(a) soit comme un **écart** pr à une **valeur centrale**. C'est donc essentiellement une notion probabiliste (voire simplement descriptive) tenant à la **variabilité** propre à tout **phénomène** observable ;

(b) soit comme écart dû à une **erreur de spécification** du modèle, et notamment :

$(b)_1$ à une erreur sur la liste des variables exogènes. Dans ce cas, la loi de probabilité d'ensemble (**loi conjointe**) est supposée connue (en particulier, la liste des variables qu'elle solidarise) ;

$(b)_2$ à l'effet de variables non prises en compte dans le modèle : soit intentionnellement (simplification, **parcimonie**), soit par impossibilité (variables inobservables ou non déterminables par la théorie). Ceci conduit à introduire une hypothèse de type **théorème de la limite centrale**. Dans de cas, la loi gouvernant le phénomène est imparfaitement connue (liste de variables incomplète) ou est trop complexe (nombre important de variables décrivant le phénomène, **complexité** de leurs **intéractions**), ce qui conduit à « regrouper » (ou agréger) toutes celles non prises en compte, et à les synthétiser en une variable « résiduelle » (variable « fourre-tout », ou « variable omnibus ») appelée perturbation aléatoire ou erreur sur l'équation.

Les deux interprétations précédentes peuvent intervenir dans d'autres situations statistiques. En effet, une démarche analogue peut être développée pour d'autres **structures statistiques** : eg **modèle d'interdépendance**.