MODÈLE DE RÉGRESSION (MULTIPLE) (D2, J1)
(17 / 04 / 2018)

Le terme, devenu « technique », de ***régression*** est assez inadapté. Il provient d'une étude « *R*egression *towards mediocrity in hereditary stature* » (in Journal of the Anthropological Institute, volume 15, 1886) dans laquelle F. GALTON (cousin de C. DARWIN) et ses collaborateurs décrivent un phénomène de « retour », vers un « état moyen » (médiocrité), de la descendance d'une unité vivante (transmission de caractères génétiques héréditaires).

Le mot « régression » suggère donc l'existence d'une relation « négative » (cf régression économique ou sociale, régression psychologique, etc).

Or, entendue dans son sens technique actuel, une « régression simple » (comportant eg une endogène et une exogène) peut aussi bien avoir une « pente » positive qu'une pente négative.

La signification profonde de cette notion se rapporte à celle de **loi scientifique** (cf aussi **loi**), dont les concepts généraux sont ceux de **loi multivariée** ou de **relation fonctionnelle**.

La simplicité (relative) des calculs qui se relient au concept de régression, ainsi que sa souplesse d'adaptation à des **contextes statistiques** variés en ont fait un outils d'usage courant en **Statistique**.

Ainsi, lorsque le language ordinaire fait état des « lois de la physique » (physique), des « lois de la **Nature** » (physique, biologie, écologie), ou encore des « lois économiques » (sociologie), c'est la notion de **régression** (ou ses extensions : eg l **modèle d'interdépendance**) qui en constitue(nt) le(s) concept(s) (unificateur(s)) sous-jacent(s).

(i) La notion (probabiliste) « théorique » de **régression** se relie étroitement à celle (statistique) de **modèle de régression** : ce dernier est un **modèle statistique** qui exprime l'existence d'une relation, ou simplement d'une **dépendance**, moyenne (ou centrale, ou « typique ») entre deux **ensembles** (ou « listes ») de **variables aléatoires** supposées jouer des rôles différents : les « exogènes » et les « endogènes ».

On peut ainsi distinguer entre :

(a) le concept de ***« régression »***, qui est une caractéristique conditionnelle d'une loi de probabilité multivariée. Ce concept représente la loi d'un **phénomène** à observer, ie l' « idée » ou la « représentation » que l'**homme de l'art** se fait de la **structure** et du fonctionnement de ce phénomène : liste et nature des variables pertinentes (« descripteurs » du phénomène), schéma des intéractions (sens, intensités, délais, etc) entre celles-ci.

Un modèle statistique exprimé à l'aide d'une régression suppose en effet dissymétriques les rôles joués par deux groupes de variables, dont les « listes » sont données a priori (eg par l'homme de l'art). Une régression exprime, de façon naturelle, que l'une des **va**, dite **variable endogène**, possède une caractéristique qui

admet, conditionnellement à l'autre va, dite **variable exogène**, une représentation analytique, paramétrique ou non paramétrique (cf **caractéristique conditionnelle**, **espérance conditionnelle**).

Une régression (ou une **fonction de régression**) est donc un objet mathématique qui se représente géométriquement dans l'**espace des variables** concernées (ou même, parfois, dans un espace de variables concernées ou potentielles) ;

(b) le concept de ***« modèle de régression »*** qui résulte de l'« accouplement », ou de l' « imbrication », des concepts de régression et d'**observation**. En effet, une représentation a priori tq la précédente n'est pas nécessairement explicitable ou explicitée. Elle doit donc être précisée (estimée, testée, validée, utilisée, etc) à l'aide de l'**information** provenant de l'observation du phénomène. Cette observation produit des « valeurs » prises par les **variables d'intérêt** et qui sont observées (« **observables** »), à l'aide de moyens d'investigation divers, soit sur des **« unités » d'observation** appelées **unités statistiques**, soit en des zones de l'espace $\mathbf{R}^3$, soit au cours du temps (T), soit même sur des unités statistiques observées dans l'espace-temps $\mathbf{R}^3$ x T.

Un modèle de régression est donc un objet statistique qui se représente géométriquement dans l'**espace des observations**.

D'après ce qui précède, on peut dire qu'**une régression est « observée » à l'aide un modèle de régression**.

Dans ce qui suit, on suppose, le cas échéant, que les diverses notions ou opérations (eg caractéristiques conditionnelles, opérations additives, etc) ont un sens.

(ii) On appelle (modèle de) **régression (multiple) (non linéaire) (théorique)** une **structure statistique** $(\mathcal{X} \times \mathcal{Y}, \mathcal{B} \otimes \mathcal{C}, \mathcal{P}^{(\xi,\eta)})$ tq :

(a) l'**espace mesurable** $(\mathcal{X}, \mathcal{B})$, appelé **espace des variables exogènes**, est un **espace vectoriel** (eg un **espace de BANACH**) réel. C'est l'espace des valeurs de la va ξ, dite **variable exogène**, ou **régresseur** : ξ peut désigner une « liste » de telles variables ;

(b) l'espace mesurable $(\mathcal{Y}, \mathcal{C})$, appelé **espace des variables endogènes**, est un espace vectoriel (eg un espace de BANACH) réel. C'est l'espace des valeurs de la va η, dite **variable endogène**, ou **régressande** : η peut désigner une « liste » de telles variables (cf **régression multidimensionnelle**) ;

(c) la **famille** des **lois de probabilité** possibles (les « lois candidates ») $P^{(\xi,\eta)}$ du modèle, notée $\mathcal{P}^{(\xi,\eta)}$, vérifie la condition suivante :

(1) $\quad P^{(\xi,\eta)} \in \mathcal{P}^{(\xi,\eta)} \Rightarrow C(\eta / \xi) = \rho(\xi)$,

où $C(\eta / \xi)$ désigne un représentant quelconque de la classe définie par une **caractéristique** de **centralité** conditionnelle C de η sachant ξ et où $\rho : \mathcal{X} \mapsto \mathcal{Y}$ est

une **application mesurable** donnée, appelée **fonction de régression** (ou **fonction d'interdépendance** dans le cas de plusieurs endogènes).

Elle est souvent aussi notée r, f ou φ.

Si C = E (resp C = S, resp C = Q_p), cette caractéristique n'est autre que l'**espérance conditionnelle** (resp le **mode** conditionnel, resp un **quantile conditionnel** d'ordre p $\in$ [0,1]) (cf aussi **valeur typique de FRÉCHET**).

La fonction de régression de η en ξ (donc la **variété de régression** qui lui est associée géométriquement) est ainsi définie par l'application $\rho : \mathcal{X} \mapsto \mathcal{Y}$ tq :

(2) $\quad x \mapsto y = \rho(x) = C(\eta / \xi = x)$,

où $C(\eta / \xi = x)$ est un représentant quelconque de la classe $C(\eta / \xi)$. En général, ρ est une fonction non linéaire pr à x.

On dit qu'on exprime la variable endogène par « régression » sur la variable exogène, ou encore que l'on « régresse l'endogène sur l'exogène ».

L'équation (2) précédente, dite **équation de régression**, (parfois écrite au pluriel) représente donc la variation en **tendance centrale** (ie selon la caractéristique de centralité conditionnelle retenue) de η en fonction des valeurs possibles $\xi(\omega) = x$ de ξ.

Quand $P^{(\xi,\eta)}$ parcourt $\mathcal{P}^{(\xi,\eta)}$, la relation ρ, en général inconnue, varie dans un espace de fonctions. L'**inférence statistique** (eg **estimation** ou **test** d'hypothèses) porte principalement sur cette relation ρ.

(iii) Le modèle (1) ou (2) ainsi décrit est parfois appelé **modèle « théorique »**, car il correspond au concept de régression et ne fait donc pas (encore) explicitement appel à des observations du **couple aléatoire** (ξ, η). On parle aussi de représentation dans l'**espace des variables**.

La représentation géométrique de ce type de modèles s'effectue généralement à l'aide d'objets géométriques tq une **courbe de régression**, une hyper-**surface de régression**, ou une carte d'une **« variété de régression »** simple. La fonction de régression exprimée en (2) définit ainsi une **variété de régression** (théorique).

On peut parfois interpréter l'expression de ***« modèle de régression »*** de deux manières :

(a) soit comme signifiant un **modèle statistique** (**modèle produit**) dans lequel on s'intéresse à un paramètre « fonctionnel » de la famille $\mathcal{P}^{(\xi,\eta)}$: ce paramètre n'est autre que la fonction de régression ρ elle-même, et celle-ci prend ses valeurs dans la famille des fonctions de régression associée à $\mathcal{P}^{(\xi,\eta)}$;

(b) soit comme signifiant directement un modèle de régression ρ « observé » dans un espace d'observation produit $(\mathcal{X}, \mathcal{B}) \times (\mathcal{Y}, \mathcal{C})$ (cf **produit d'espaces**

probabilisés) sous la forme tq eg : $y_n = \rho\ (x_n) + u_n$, où n appartient à un ensemble d'**indices** (unités statistiques, zones de l'espace, dates ou périodes) (cf (6) infra).

(iv) On complète généralement le cadre précédent en analysant eg la « **variabilité** » (conditionnelle, ou « relativement » à ξ) de η autour de ces valeurs centrales lorsque $\xi\ (\omega) = x$ varie dans $\mathcal{X}$.

Cette variabilité peut être définie selon :

(3) $\quad D\ (\eta / \xi) = C\ \{(\eta - C\ (\eta / \xi)) / \xi\}$,

ie à l'aide de la même caractéristique conditionnelle C (. / ξ) retenue, ici appliquée à l'écart η - C (η / ξ).

En particulier, si C = E (espérance), la **dispersion** conditionnelle est généralement la **matrice de covariance** conditionnelle :

(4) $\quad V\ (\eta / \xi) = \sigma^2 > 0$,

ou encore le **rapport de corrélation** de η pr à ξ.

(v) La fonction de régression « théorique » ρ précédente est définie (et représentable géométriquement) dans l'espace des variables concernées. Ainsi, un modèle de régression (multiple) est très souvent défini à partir des espaces « réels » $(\mathcal{X}, \mathcal{B}) = (\mathbf{R}^K, \mathcal{B}(\mathbf{R}^K))$ (K variables exogènes ξ_1 ,..., ξ_K) et $(\mathcal{Y}, \mathcal{C}) = (\mathbf{R}, \mathcal{B}_{\mathbf{R}})$ (une variable endogène η).

Par ailleurs, le modèle théorique ci-dessus définit alors, de façon générale, une **régression paramétrée**, ou **régression non paramétrique**. En effet, la fonction ρ est généralement inconnue au premier abord : elle ne fait donc l'objet d'aucune hypothèse a priori et varie dans un espace de fonctions quelconques, ou une **famille** de fonctions ad hoc. Par suite, ρ joue le rôle d'un **indice** (d'où l'expression de régression « paramétrée » par ρ), et cet indice constitue le **paramètre d'intérêt** ;

Cependant, on peut souvent discerner, dans ρ, un paramètre vectoriel (réel et fixe) b qui possède une interprétation concrète : pression, élasticité ou viscosité moyennes (physique), tension superficielle minimale (biologie), taux de reproduction (écologie), propension marginale à consommer (sociologie), etc.

La fonction de régression s'écrit alors sous une forme (généralement non linéaire pr à b) :

(5) $\quad P^{(\xi,\eta)} \in \mathcal{P}^{(\xi,\eta)} \Rightarrow C\ (\eta / \xi) = f\ (\xi, b), \quad \forall\ b \in B$,

dans laquelle (eg) $B \in \mathcal{O}(\mathbf{R}^Q)$ (donc b appartient à un **espace vectoriel** réel de dimension finie) et $f : \mathcal{X} \times B \mapsto \mathcal{Y}$ est une fonction dont la forme analytique est :

(a) soit connue : on parle alors de **régression paramétrique** car seul b est à estimer ;

(b) soit inconnue : on parle alors de **régression semi-paramétrique**, car b et f sont tous les deux à estimer, généralement à l'aide de méthodes spécifiques.

(vi) Pour mettre en oeuvre une **procédure** d'**inférence statistique** (eg estimation, prévision, etc), le recours à des données observées (ou observables) est nécessaire.

Par suite, le modèle à considérer se définit (et est représentable géométriquement) dans l'**espace d'observation** des variables concernées. On peut alors définir la notion de **modèle de régression multiple**, parfois appelé **modèle « observé »** (par abus de language, puisque f ou b ne sont pas connus).

En effet, on dispose souvent d'**observations** du couple (ξ, η), eg supposé à valeurs dans $\mathbf{R}^K$ x $\mathbf{R}$. La notion de modèle de régression multiple est alors définie comme suit :

(a) $(\mathcal{X}, \mathcal{B}) = (M_{NK}(\mathbf{R}), \mathcal{B}(M_{NK}(\mathbf{R}))$ (N observations des K variables exogènes ξ) (avec $K < N$), d'où une (N,K)-**matrice d'observation** notée $X = (X_1, ..., X_N) = (X_1 /// X_N)$ (écriture en lignes) ou $X = [x_1, ..., x_K]$ (écriture en colonnes) ;

(b) $(\mathcal{Y}, \mathcal{C}) = (\mathbf{R}^N, \mathcal{B}(\mathbf{R}^N))$ (N observations de la variable endogène η), d'où un **vecteur aléatoire** y à valeurs dans $\mathbf{R}^N$;

(c) la famille des lois possibles $P^{(X,y)}$ du modèle est alors notée $\mathcal{P}^{(X,y)}$ et l'expression (1) est remplacée par la fonction de régression (forme paramétrée) :

(6) $P^{(X,y)} \in \mathcal{P}^{(X,y)} \Rightarrow C(y / X) = F,$

(où F dépend, en général, de X). Ceci implique que la caractéristique conditionnelle C (. / X) ait un sens, et soit calculable.

De même, l'expression (5) est remplacée par :

(7) $P^{(X,y)} \in \mathcal{P}^{(X,y)} \Rightarrow E(y / X) = F(b), \quad \forall b \in B,$

où $F : B \mapsto \mathbf{R}^N$ dépend de X et $B \in \mathcal{O}(\mathbf{R}^Q)$ (**ouvert** de $\mathbf{R}^Q$) (avec $Q \in \mathbf{N}^*$).

(vii) F doit vérifier une **« condition de marginalisation »** ou **« condition de projection »** importante. En effet, F doit être définie à partir de f et sa n-ième coordonnée doit coïncider avec la fonction de régression, évaluée au « point » (X_n, y_n), ie :

(8) $F_n(b) = f(X_n, b), \quad \forall n \in N_N^*, \forall b \in B,$

où $F_n : B \mapsto \mathbf{R}$ est la n-ième application coordonnée de F et X_n la n-ième observation de ξ (n-ième ligne de X).

Autrement dit, la régression « projetée » sur le n-ième espace facteur de l'espace d'observation produit doit posséder la même « forme analytique » que la régression initiale.

En effet, chaque loi $P^{(X,y)}$ contient, en général, des informations utiles et « exploitables » :

(a) relations entre X et y (eg corrélations croisées) ;

(b) relations internes à X (eg autocorrélation, etc) ;

(c) relations internes à y (eg autocorrélation, etc).

Cependant, dans tous les cas, chaque **loi marginale** en (X_n , y_n) de $P^{(X,y)}$ doit être identique à $P^{(\xi, \eta)}$ lorsque (X_n , y_n) est remplacé par (ξ, η), et ceci $\forall\ P^{(X,y)} \in \mathscr{P}^{(X,y)}$, puisque la fonction de régression théorique est la seule interprétable par l'**homme de l'art**.

On définit ainsi un **modèle d'échantillonnage** particulier. Ce modèle admet aussi une représentation géométrique en termes de variétés, qui est alors une représentation dans l'**espace des observations**.

(viii) Si les coordonnées y_n de y forme une **suite iid** et du second ordre, on admet souvent une hypothèse d'**homoscédasticité** :

(9) $V(y / X) = \sigma^2 . I_N$ (**matrice scalaire**),

ie (cf **covariance conditionnelle**) :

(10) $C((u_\alpha , u_\beta) / X) = \delta_{\alpha\beta} . \sigma^2, \qquad \forall\ (\alpha, \beta) \in (N_N^*)^2.$

Dans le cas général, la définition (3) est remplacée par :

(11) $D(y / X) = C(y - C(y / X)) = \Sigma,$

où Σ est une **matrice de dispersion** ad hoc, qui s'interprète eg comme une **matrice des covariances**.

Dans d'autres cas, le N-**échantillon** $((X_1 , y_1) ,..., (X_N , y_N))$ du couple (ξ, η) n'est pas nécessairement supposé iid : eg **processus** quelconque $(X_n , y_n)_{n=1,...,N}$ observé en temps discret $T = N_N^*$.

(ix) Les exogènes notées ξ (resp X) peuvent être des variables non aléatoires (ie des **variables dégénérées**), auquel cas les moments précédents doivent s'interpréter comme des moments inconditionnels (les notations étant alors allégées dans ce sens) (cf **moment**).

Ainsi, les expressions (1) ou (5) s'interprètent comme suit. La « liste » des va exogènes ξ est censée « expliquer » une (**valeur centrale** de la) **variable d'intérêt** (endogène) η. C'est cette forme que l'homme de l'art « demande » au statisticien de traiter (estimation, notamment). Les expressions (6) ou (7) sont celles que ce dernier spécifie afin de mettre en oeuvre les méthodes d'**inférence statistique** (allers et

retours entre la théorie et les observations) (cf **inférence conditionnelle**, **spécification**).

Lorsqu'il est explicité, le **paramètre d'intérêt** du modèle est donc $b \in B$, ou encore $(b, \sigma^2) \in B \times \mathbf{R}_+^*$.

Le paramètre σ est parfois considéré comme un **paramètre « nuisible »**, car la variabilité des données affecte la « **précision intrinsèque** » de la régression, « imposée » par les lois de probabilité $P^{(\xi,\eta)}$ (cf **paramètre importun**).

(x) Pour tout $g \in G$ (**groupe** additif), tout élément $h \in G$ admet la décomposition triviale : $h = g + (h - g)$. La forme vectorielle (6) peut donc aussi s'exprimer selon :

(12) $C(y / X) = F \Rightarrow y = C(y / X) + \{y - C(y / X)\}$,

ie :

(13) $y = C(y / X) + u$,

où $u = y - C(y / X)$ est un vecteur aléatoire à valeurs dans $\mathbf{R}^N$, appelé **perturbation**, ou **erreur sur l'équation**, ou encore **choc sur l'équation** (ou parfois **résidu de l'équation**).

De même, la forme vectorielle (7) s'exprime selon :

(14) $E(y / X) = F(b) \Rightarrow y = E(y / X) + \{y - E(y / X)\}$,

ie :

(15) $y = E(y / X) = F(b) + u$,

où u s'interprète de façon analogue. Par suite, les hypothèses {(11),(12),(13)} s'écrivent :

(16) $C(u / X) = 0, \qquad D\, u \in S_N(\mathbf{R})$,

et les hypothèses {(9),(15)} s'écrivent :

(17) $E(u / X) = 0, \qquad V\, u = \sigma^2 . I_N$.

On note souvent σ_u^2 (ou σ_ε^2) (ou encore parfois σ_y^2 ou σ_η^2) au lieu de σ^2.

Une interprétation analogue aux précédentes s'applique :

(a) au modèle théorique (1), avec la décomposition évidente :

(18) $C(\eta / \xi) = \rho(\xi) \Rightarrow \eta = C(\eta / \xi) + \{\eta - C(\eta / \xi)\}$,

ie :

(19) $\eta = C(\eta / \xi) + \varepsilon, \qquad$ avec $C(\varepsilon / \xi) = 0$;

(b) au modèle théorique (5) au sens de l'espérance (ie où C = E) :

(20) $E(\eta / \xi) = f(\xi, b) \Rightarrow \eta = E(\eta / \xi) + \{\eta - E(\eta / \xi)\}$,

ie :

(21) $\eta = E(\eta / \xi) + \varepsilon = f(\xi, b) + \varepsilon$, avec $E(\varepsilon / \xi) = 0$,

et :

(22) $V(\varepsilon / \xi) = \sigma^2$ (ou σ_ε^2).

(xi) L'équation (1) (resp (5)) du modèle théorique peut encore s'interpréter comme un **changement de variable aléatoire** $(\xi, \varepsilon) \mapsto \eta$, la loi de la va η s'analysant comme l'image de la loi d'un couple aléatoire (ξ, ε) par la transformation (1) (resp (5)). Il en va de même pour le modèle de régression multiple (« modèle observé ») : le changement de va est alors de la forme $C(X, u) \mapsto y$ et la loi de y est l'image de celle de (X, u) par la transformation (13) (resp (15)).

(xii) Une situation très courante est celle du **modèle de régression (multiple) linéaire** (cf **modèle de régression linéaire**), dans lequel Q = K (autant de paramètres b_q que de variables exogènes ξ_k) et où f est bilinéaire en (ξ, b), ie (cf **forme multilinéaire**) :

(23) $f(\xi, b) = \xi' b$, $\forall (\xi, b)$.

Par suite, l'équation (15) correspond à l' « observation » :

(24) $F(b) = X b$, $\forall (X, b)$,

où X désigne à la fois la (N,K)-matrice d'observation des va exogènes et l'**opérateur** $b \mapsto X b = E(y / X)$. On l'appelle parfois **matrice du modèle** ou même **matrice du plan** (cf **théorie des plans d'expérience**).

(xiii) L'**estimation** d'un modèle de régression multiple s'effectue à l'aide de méthodes générales : méthodes « paramétriques » lorsque seul le paramètre b est à estimer, méthodes « semi-paramétriques » lorsque, à la fois, ρ, f (ou F) et b sont à estimer, méthodes non paramétriques si aucun paramètre b n'est explicité et que l'on doit estimer ρ ou f (ou F) (qui joue un rôle d'indice ou de « paramétrage »).

Par suite :

(a) dans l'**approche paramétrique**, on spécifie a priori la forme analytique de ρ, qui dépend d'un paramètre b (ie $\rho(\xi) = f_0(\xi, b)$), ainsi que la **loi de probabilité** de η (ou de y). En effet, dans cette approche, tout est connu dès que f_0 et b le sont. Si f_0 n'est pas spécifiée, on définit une autre approche, parfois dite **approche « semi-paramétrique »**. Le problème se ramène alors à l'estimation du **paramètre principal** b, et les méthodes usuelles peuvent être : une **méthode de moindre norme** (eg **méthode des moindres carrés ordinaires**), une **méthode du maximum de**

vraisemblance (**vraisemblance** gaussienne, le plus souvent), ou encore une **méthode des moments**.

Une extension du modèle précédent consiste à remplacer l' « hypothèse de scalarité » $V\,y = V\,u = \sigma_u^2\, I_N$ par une hypothèse plus générale :

(25) $V\,y = V\,u = \Sigma$ (ou $\sigma_u^2\,\Omega$),

dans laquelle Σ (resp Ω) est une **matrice symétrique** (semi-)définie positive (cf **matrice définie positive**).

Cette matrice est, en général, indépendante de (X, y). Elle comporte moins de termes (ou paramètres) inconnus que le nombre maximum de termes indépendants (au sens de l'analyse), ie (pour cette matrice symétrique) $\{(N^2 - N) / 2\} + N = N\,(N + 1) / 2 = C_N^2$ termes. On écrit alors explicitement $\Sigma\,(\lambda)$ (ou $\Omega\,(\lambda)$) pour exprimer que Σ (resp Ω) dépend du paramètre vectoriel $\lambda \in \Lambda$ (avec eg $\Lambda \in \mathcal{P}(\mathbf{R}^L)$ et $L < C_N^2$). L'estimation du modèle s'effectue alors par la **méthode des moindres carrés généralisés** ou par la **méthode du mv** ;

(b) dans l'**approche non paramétrique**, on peut, dans certains cas, définir un **estimateur** non paramétrique de la **densité** de $P^{(\xi,\eta)}$ pr à une **mesure positive** donnée (cf **estimation non paramétrique**). Une **méthode non paramétrique** habituelle est la **méthode du noyau**, la **méthode des fonctions orthogonales** ou la **méthode des fonctions splines**. Le modèle de régression « estimé » se déduit alors de l'estimateur de la densité précédente, par **conditionnement** relativement aux exogènes.

(xiv) Les divers **test d'hypothèses** (portant sur b ou sur σ) sont souvent fondés, en pratique, sur l'**hypothèse de normalité** de y (ou, de façon équivalente, de u) suivante (cf **loi normale multidimensionnelle**) :

(26) $y \sim \mathcal{N}_N\,(F\,(b), \sigma_u^2\, I_N)$ (ou $u \sim \mathcal{N}_N\,(0, \sigma_u^2\, I_N)$),

dans le cas non linéaire général, ou :

(27) $y \sim \mathcal{N}_N\,(X\,b, \sigma_u^2\, I_N)$ (ou $u \sim \mathcal{N}_N\,(0, \sigma_u^2\, I_N)$),

dans le cas linéaire.

Les développements des moyens matériels (ordinateurs) aussi bien que des méthodes de calcul numérique permettent aussi de traiter des **situations de non-normalité**.

(xv) Le **domaine de validité** d'un modèle de régression est un problème important.

Dans le cas linéaire où $y = X\,b + u$, on peut admettre que, une fois validé à l'aide des tests spécifiques, le modèle est valide pour toute valeur $x_0 \in$ Co X (**enveloppe convexe** de l'ensemble de points $X = \{X_1, ..., X_N\}$ de $\mathbf{R}^K$). Lorsque $x_0 \in$ Co X, le calcul

de la valeur de l'endogène y_0 correspondante s'appelle **interpolation**. A l'extérieur de Co X, on parle d'**extrapolation**.

Si le modèle de régression est non paramétrique, l'extrapolation peut être moins simple à définir (eg **approximation linéaire** ou **approximation quadratique** locales, à la **frontière**).

(xvi) En **théorie bayésienne**, b n'est plus un paramètre certain mais une **va** dont la loi est donnée a priori. L'**inférence statistique** (estimation, tests, prévision, etc) s'effectue à partir de la **loi a posteriori** (qui résulte des hypothèses stochastiques du modèle ainsi que de la **loi a priori** de b). Une extension possible porte sur le paramètre σ.

(xvii) De nombreux types de modèles statistiques peuvent se ramener au modèle de régression, ou à ses extensions (cf **relation fonctionnelle**). Les variables ξ et η (donc X et y) peuvent alors être de type quantitatif aussi bien que qualitatif (cf **loi multivariée**). Ainsi en est-il du **modèle d'interdépendance**, du modèle des **régressions multiples**, du **modèle d'analyse de la variance**, du **modèle d'analyse de la covariance**, du **modèle multi-indicé**, etc.

(xviii) Les « observations » de (ξ, η) ont été supposées indicées par $n \in N_N^*$. On peut encore étendre le modèle au cas d'observations indicées par $t \in T$, où T est un ensemble d'indices a priori quelconque, souvent muni d'une **structure** d'**ordre** (eg le « **temps** » physique) (cf **relation d'ordre**) ou d'une structure mesurée tq $(T, \mathcal{B}_T, \nu)$ (cf **espace mesuré**).

Certaines méthodes d'estimation s'étendent dans ce sens (eg la méthode des moindres carrés). La méthode des moindres carrés ordinaires conduit à déterminer ici l'estimateur $\hat{b}$ de b tq :

(28) $\quad ||y - X\,\hat{b}||_2^2 = \inf_{b \in \mathbf{R}^K} ||y - X\,b||_2^2,$

où $y : T \mapsto \mathbf{R}^N$, $X : T \mapsto M_{NK}(\mathbf{R})$, $||u||_2^2 = \int_T u_t^2 \, d\nu(t)$ et $\mathbf{R}^K$ désigne ici, par commodité, $\mathbf{R}^K$. On obtient alors une **équation normale** de la forme $M_{xx}\, b = m_{xy}$.

En particulier, les observations peuvent être en temps continu. Ainsi, la notation $n \in N_N^*$ est remplacée par $t \in T$, avec $T \subset \mathbf{R}$, d'où :

(29) $\quad y_t = f(X_t, b) + u_t ,$

avec $E\,u_t = 0$ et $C(u_s, u_t) = \delta_{st} \cdot \sigma_\varepsilon^2$, $\forall\, t \in T$, où X_t est à valeurs dans $\mathbf{R}^K$ et $b \in \mathbf{R}^Q$.

(xix) Les types les plus classiques de modèles de régression sont :

(a) le modèle de régression « ordinaire » (ie au sens de l'**espérance**) ;

(b) celui de **régression modale** (ie au sens du **mode**) ;

(c) ou encore le modèle de **régression quantilaire** (ie au sens d'un **quantile**).

Le modèle de régression multiple intervient dans de nombreux contextes :

(a) **forme réduite** d'un **modèle d'interdépendance** ;

(b) **modèle d'analyse de la variance** ;

(c) **modèle d'analyse de la covariance** (associés à un **plan d'expérience**) ;

(d) **analyse des données**.

Ce modèle de base connaît aussi de nombreux développements :

(a) extensions : modèle multidimensionnel, traitement des **singularités**, **modèle d'analyse de la variance** ou **modèle d'analyse de la covariance** ;

(b) généralisations : **modèle d'interdépendance**, **modèle à erreurs sur les variables**, **relation fonctionnelle**.

(xx) On appelle souvent **modèle de régression** (multiple) un **modèle de régression** comportant une seule variable endogène η et plusieurs (eg K) variables exogènes ξ. La **régression** qui définit un tel modèle comporte ainsi une équation scalaire dans l'espace des variables. La forme générale en est (cas non paramétré) :

$$(30) \quad \eta = \rho(\xi) + \varepsilon = \rho(\xi_1, ..., \xi_K) + \varepsilon,$$

ou (cas paramétré ou paramétrique) :

$$(31) \quad \eta = f(\xi, b) + \varepsilon = f(\xi_1, ..., \xi_K, b) + \varepsilon,$$

avec $\xi : \Omega \mapsto \mathbf{R}^K$ et $\eta : \Omega \mapsto \mathbf{R}$ (cas de variables numériques), où $b \in \mathbf{R}^Q$ et f est une fonction de régression scalaire dépendant de b ;

Lorsque le modèle de régression comporte plusieurs (eg G > 1) endogènes η, on parle plutôt de **régressions multiples** (au pluriel) ou de **modèle à équations apparemment sans relations** (ie corrélations) (appellation anglo-saxonne). La régression qui génère un tel modèle comporte donc G équations scalaires dans l'espace des variables, ie une seule équation multidimensionnelle (dans ce même espace). La forme générale en est (cas non paramétré) :

$$(32) \quad \begin{aligned} \eta_1 &= \rho_G(\xi) + \varepsilon_G = \rho_G(\xi_1, ..., \xi_K) + \varepsilon_1 \\ &... \\ \eta_G &= \rho_G(\xi) + \varepsilon_G = \rho_G(\xi_1, ..., \xi_K) + \varepsilon_G \end{aligned}$$

ou (cas paramétré ou paramétrique) :

$$(33) \quad \begin{aligned} \eta_1 &= f_G(\xi) + \varepsilon_G = f_G(\xi_1, ..., \xi_K, b) + \varepsilon_1 \\ &... \\ \eta_G &= f_G(\xi) + \varepsilon_G = f_G(\xi_1, ..., \xi_K, b) + \varepsilon_G \end{aligned}$$

avec $\xi : \Omega \mapsto \mathbf{R}^K$ et $\eta : \Omega \mapsto \mathbf{R}^G$ (cas de variables numériques), où $b \in \mathbf{R}^Q$ et les f_g sont des fonctions de régression scalaires dépendant de b.

Des exemples typiques de modèles de régression multiple sont (avec cette terminologie) :

(a) le **modèle linéaire** et le **modèle non linéaire** ;

(b) le **modèle de régression multidimensionnel**.