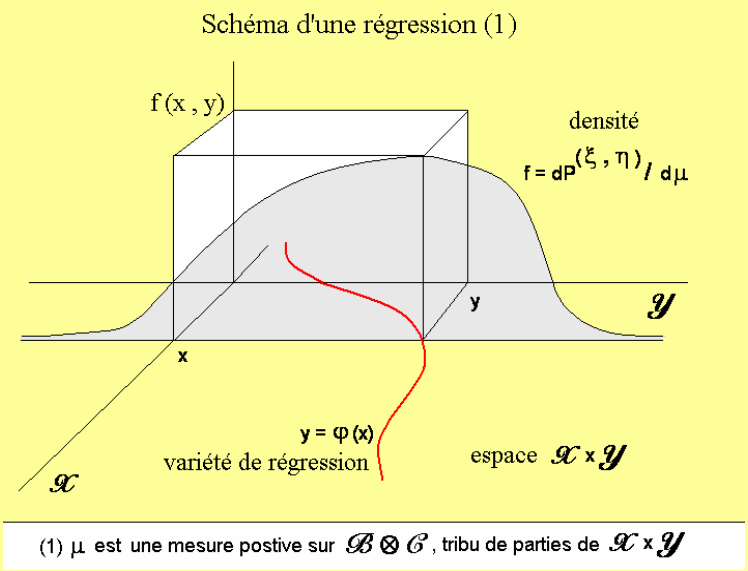
SURFACE DE RÉGRESSION (A13, D2, J7)
(09 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

Les importantes notions de **régression** et de **modèle de régression** peuvent s'interpréter comme des représentations du concept de **loi scientifique** (cf aussi **relation fonctionnelle**).

Elles s'interprètent aussi géométriquement à l'aide du concept de **variété de régression** : **courbe de régression** dans $\mathbf{R}^2$, « hyper »-**surface de régression** dans $\mathbf{R}^3$ ou $\mathbf{R}^{K+1}$ (cf graphique ci-après, dans lequel la variété se réduit à une **courbe** dans l'espace des variables, représenté par le « plan » $\mathcal{X}$ x $\mathcal{Y}$).



(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $\zeta = (\xi , \eta) : \Omega \mapsto \mathbf{R}^K$ x **R** un **couple aléatoire** constitué d'un **vecteur aléatoire** ξ à valeurs dans E et d'une **vars** η (ie à valeurs dans **R**). $P^{(\xi,\eta)}$ désignant la **loi conjointe** du couple (ξ , η), on définit l'**espérance conditionnelle** E (η / ξ) de η sachant ξ (supposée exister) :

(1) $\quad x \in \mathbf{R}^K \mapsto y = \varphi (x) = E (\eta / \xi = x)$ ou $E^{\xi = x} \eta$,

où φ est la **fonction de régression** et E $(\eta / \xi = x)$ un représentant quelconque de la classe E (η / ξ).

On appelle alors **surface de régression**, ou parfois **hypersurface de régression**, le graphe dans ($\mathbf{R}^{K+1} = \mathbf{R}^K$ x **R**) de la fonction de régression φ, graphe noté Γ.

(ii) Sous certaines **conditions de régularité**, la surface précédente constitue une (sous-)**variété différentielle** de E de dimension K (son **espace tangent** en tout point $(x , y) \in \mathbf{R}^K$ x **R** est de dimension K).

En particulier, ceci définit une **courbe de régression** (K = 1) ou une **surface de régression** (K = 2).

Si $\zeta \in \mathcal{L}_{\mathbf{R}K+1}^{1}(\Omega, \mathcal{T}, P)$, la surface de régression existe toujours ($\mathbf{R}K+1$ désignant, par commodité, l'espace $\mathbf{R}^{K+1}$).

Si $\zeta \in \mathcal{L}_{\mathbf{R}K+1}^{2}(\Omega, \mathcal{T}, P)$, la surface de régression vérifie une importante propriété d'**optimalité**. En effet, dans ce cas, φ est (l'équation de) la surface de $\mathbf{R}^{K+1}$ qui minimise l'**écart quadratique moyen** :

(2) $\quad Q^2(\phi) = E(\eta - \phi(\xi))^2$

par rapport à la fonction $\phi : \mathbf{R}^K \mapsto \mathbf{R}$. Autrement dit, φ est optimale au sens des moindres carrés car elle est solution du problème de **programmation mathématique** « fonctionnel » :

(3) $\quad \min_{\phi} Q^2(\phi)$.

Ceci résulte du **théorème de KOENIG** de décomposition de l'écart quadratique moyen, ie : $Q^2(\phi) = E(\eta - \varphi(\xi))^2 + E(\varphi(\xi) - \phi(\xi))^2$.

La solution φ est donc unique P-presque sûrement : c'est pourquoi on l'appelle souvent aussi **surface des moindres carrés**.

(iii) La notion de régression conduit à définir des **rapports de corrélation** multiples. En effet, si $\{k_1, ..., k_L\} \subset \{1, ..., K\}$ (avec $L \leq K$), le **rapport de corrélation multiple** de η en $(\xi_{k1}, ..., \xi_{kL})$ est défini selon (on note ξkl pour désigner ξ_{kl}) :

(4) $\quad \eta^2_{\eta / (\xi k1, ..., \xi kL)} = V_a / V_b$,

où $V_a = V_{(\xi k1, ..., \xi kL)} E\{\eta / (\xi_{k1} = x_{k1}) \cap \ ... \ \cap (\xi_{kL} = x_{kL})\}$ est la **variance** de l'espérance de η conditionnelle au L-uple $(\xi_{k1}, ..., \xi_{kL})$, et $V_b = V_\eta \eta$ est la variance propre de η.

Ces rapports sont au nombre de $K(2^{K-1} - 1)$.

(iv) Une fonction de régression est souvent estimée à l'aide d'une **méthode non paramétrique**. Etant donné un modèle de base $(\Omega, \mathcal{T}, \mathcal{P})$ et un **espace d'observation** $(\mathbf{R}^{K+1}, \mathcal{B}(\mathbf{R}^{K+1}))$, le couple aléatoire ζ associe à toute **probabilité** $P \in \mathcal{P}$ la lp $P^{(\xi,\eta)} \in \mathcal{P}^{(\xi,\eta)}$ et la fonction de régression φ associée à $P^{(\xi,\eta)}$. Par suite, si $Z = ((X_1, Y_1) \ ... \ (X_N, Y_N))$ est un **échantillon aléatoire** constitué de **copies** (X_n, Y_n) de (ξ, η) indépendantes entre elles et de loi commune $P^{(\xi,\eta)}$ (**échantillon iid**), une classe générale d'**estimateurs** de φ est définie par :

(5) $\quad \tilde{\varphi_N}(x) = \Sigma_{n=1}^{N} a_{nN}(x, X) . Y_n$,

où $a_{nN} : \mathbf{R}^K \times \mathbf{R}^N \mapsto \mathbf{R}_+$ $(n = 1, ..., N)$ constitue une suite finie de fonctions mesurables positives tq $\Sigma_{n=1}^{N} a_{nN}(x, X) = 1$, $\forall x \in \mathbf{R}^K$ (cf **score**) et où $X = (X_1, ..., X_N)$ est l'échantillon « marginal » associé à ξ.

En particulier, si p : $\mathbf{R}^K \mapsto \mathbf{R}_+$ est une fonction donnée (**noyau**) et $h_N > 0$, l'**estimateur par le noyau** (cf **méthode du noyau**) de φ est donné par (5), avec :

(6) $a_{nN}(x , X) = (\Sigma_{n=1}^{N} p_n)^{-1} . p_n ,$

où $p_n = p \{h_N^{-1} . (X_n - x)\}$.

La notion de surface (ou de variété) de régression est donc associée à celle de **modèle de régression**, qui définit un cadre d'**estimation non paramétrique** (ou semi-paramétrique lorsque φ dépend d'un paramètre vectoriel $b \in \mathbf{R}^Q$).

(vi) On généralise la notion de surface de régression de plusieurs façons :

(a) à l'aide d'autres caractéristiques conditionnelles que l'espérance : cf **régression modale**, **régression quantilaire**. Autrement dit, on remplace l'espérance conditionnelle, à partir de laquelle est définie la notion de régression précédente, par une **caractéristique conditionnelle** (eg mode conditionnel, médiane conditionnelle, lorsque ces notions sont définies). La notion de régression est donc essentiellement fondée sur une caractéristique légale conditionnelle ;

(b) en définissant des fonctions du type :

(7) $x \in \mathbf{R}^K \mapsto y = \varphi_p(x) = E\{(\eta - E(\eta / \xi = x)^p) / \xi = x\}$ (**régression dans L^p**),

où $p \in \mathbf{N}^*$ est un entier positif donné (cf **régression dans L^2** lorsque p = 2) ;

(c) à partir de variétés plus générales que les « hypersurfaces » précédentes. Ainsi, si $\zeta = (\xi , \eta) : (\xi , \eta) : \Omega \mapsto \mathbf{R}^K \times \mathbf{R}^G$, on définit une **variété (paramétrée) de régression** (parfois improprement appelée **variété des moindres carrés**) selon :

(8) $x \in \mathbf{R}^K \mapsto y = \varphi(x) = E(\eta / \xi = x) \in \mathbf{R}^G$ (P - p.s.).

Dans ce cadre, la propriété (3) ci-dessus peut se généraliser. La « variable » (vectorielle) ξ (resp η) s'interprète alors comme une « **liste** » **de variables exogènes** (resp **liste de variables endogènes**).

On étend ainsi la notion de régression (sous la première forme explicite (1)), en une notion de **régression multidimensionnelle**, dans laquelle la **variable endogène** est un **vecteur aléatoire** à valeurs dans $\mathbf{R}^G$ ou, plus généralement, dans un espace puissance $\mathcal{Y}^G$.

(vii) De même, selon que les observations X de ξ et Y de η sont mono-indicées ou multi-indicées, on définit des modèles de régression de structure plus ou moins simple, prenant en compte d'éventuelles **corrélations** entre variables ou observations d'**indices** (ou de multi-indices) différents.