TRANSFORMATION DE BOX - COX (C2, K2)
(21 / 09 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **transformation de BOX - COX** est, en réalité, une **famille** de transformations à un « **paramètre** », souvent utilisée pour la **transformation des données**.

(i) Cette **famille**, indexée par un paramètre $\lambda \in \mathbf{R}$, est définie selon :

(1) $\beta_\lambda : u \mapsto u_\lambda = \lambda^{-1} (u^\lambda - 1), \quad \forall u \in \mathbf{R}_+^*.$

Par continuité, on a :

(2) $\beta_0 : u \mapsto u_0 = \text{Log } u, \quad \forall u \in \mathbf{R}_+^*.$

(ii) Cette **transformation de G.E.P. BOX - D.R. COX** permet eg :

(a) d'obtenir un **modèle de régression** dont les **perturbations** sont indépendantes et gaussiennes. Ainsi, si $\eta = \xi ' b + \varepsilon$ est un **modèle de régression linéaire** multiple défini dans un **espace de variables** (ξ, η) et observé dans un **espace d'observation** (X, y) selon y = X b + u, avec $u = (u_1 ,..., u_N)$, on peut rechercher le paramètre de la transformation pour lequel les perturbations u sont des **copies** indépendantes d'une **variable parente** dont la **loi** est $\mathcal{N}(0, 1)$ (**loi normale réduite**) ;

(b) de simplifier un modèle de **régression non linéaire**. Ainsi, le modèle $\eta = f(\xi , b) + \varepsilon$ peut être transformé selon :

(3) $\lambda^{-1} (\eta^\lambda - 1) = f \{\mu_1^{-1} (\xi^{\mu(1)} - 1) ,..., \mu_K^{-1} (\xi^{\mu(K)} - 1), b\} + \varepsilon.$

Par suite, un choix approprié des paramètres $\lambda, \mu_1 ,..., \mu_K$ peut conduire à une **forme** plus simple, eg :

(4) $\eta_\lambda = \varphi (\xi_{1,\mu(1)} ,..., \xi_{K,\mu(K)})' \psi (b) + \varepsilon$ (modèle à « **forme séparée** »),

dans laquelle, par commodité, $\mu(K)$ désigne μ_K . Ceci peut conduire à des fonctions φ ou ψ plus simples : eg lorsque ψ est quasiment linéaire au voisinage de la **vraie valeur** b* de b.

(iii) Une version plus générale de la transformation (1) contient deux paramètres :

(1) $\beta_{\alpha\lambda} : u \mapsto u_{\alpha\lambda} = \lambda^{-1} \{(u - \alpha)^\lambda - 1\}, \quad \forall u > \alpha.$

Lorsque $\lambda \to 0+$, on obtient, par continuité :

(2)' $\beta_{\alpha 0} : u \mapsto u_{\alpha 0} = \text{Log } (u - \alpha), \quad \forall u > \alpha.$

(iv) Les transformations de BOX - COX peuvent replacer un **modèle statistique** dans un **contexte** « standard ». Ainsi, lorsque le **modèle de régression multiple** linéaire y = X b + u ne vérifie pas les hypothèses usuelles :

(3) $E\ u = 0, \qquad V\ u = \sigma_u^2 \ .\ I_N$,

des transformations analogues à celles de l'exemple $(ii)_{(b)}$ ci-dessus peuvent conduire à un modèle transformé tq :

(4) $\eta\ (\lambda) \ = \ \Sigma_{k=1}^{K}\ b_k\ .\ x_k\ (\mu_k) + v$,

en sorte que (hypothèses à tester) (cf **test préliminaire**) :

(3)' $E\ v = 0, \qquad V\ v = \sigma_v^2\ .\ I_N$.

Par suite, le modèle (4) nécessite l'estimation des 3 + K paramètres (b, σ_v^2, λ, μ_1 ,..., μ_K) : la **méthode du maximum de vraisemblance** peut alors permettre de résoudre le problème initial.

Il en va de même, dans certains cas, lorsque le vecteur u des perturbations n'est pas gaussien, ie $P^u \neq \mathcal{N}_N\ (0,\ \sigma_u^2\ .\ I_N)$. Une transformation tq l'une des précédentes peut permettre d'obtenir $P^v = \mathcal{N}_N\ (0,\ \sigma_v^2\ .\ I_N)$. On l'appelle parfois **transformation « normalisante »**.