BIAIS D'UNE RÉGRESSION (J1)
(23 / 09 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) Dans le **modèle de régression multiple** (linéaire), écrit dans l'**espace des observations** (X, y) selon :

(1) $y = X b + u,$

on admet généralement l'**hypothèse d'absence de biais**, au sens où :

(2) $E y = X b$ (resp $E (y / X) = X b$), $\forall b \in \mathbf{R}^K$,

ie :

(3) $E u = 0$ (resp $E (u / X) = 0$).

Si l'on utilise le modèle {(1),(3)} alors que le **vrai modèle** est {(1),(4)}, avec :

(4) $E u = m \in \mathbf{R}^N$ (resp $E (u / X) = m \in \mathbf{R}^N$),

on commet une **erreur de spécification** (cf **spécification**). En effet, on a $E y = X b + m$ alors qu'on traite le problème comme si (2) était vérifié, omettant ainsi une **variable** μ, observée à travers m, alors que la « vraie » **matrice des observations** des **exogènes** est la (N,K+1)-matrice « augmentée » [X, m].

En particulier, si μ est une variable exogène **constante** (ie une **variable dégénérée** $\mu : \Omega \mapsto \{c\}$, avec $c \in \mathbf{R}$), alors $m = c . e_N$. En posant $v = u - m$, le modèle {(1),(4)} devient :

(5) $y = (X b + c . e_N) + v,$ avec $E v = 0$.

Par suite, si X contient déjà le vecteur e_N correspondant au terme constant b_1 , soit $X = [e_N , x_2 ,..., x_K]$, on peut écrire :

(6) $y = X b + m + v = (c + b_1) e_N + \Sigma_{k=2}^K b_k x_k + v,$

et l'hypothèse (4) revient à modifier le seul « terme constant » (éventuellement déjà nul) dans (1). Par suite, l'estimation de b eg par la **méthode des moindres carrés ordinaires** n'est que partiellement biaisée : $E b_1^\wedge \neq b_1$, mais $E b_k^\wedge = b_k$, $\forall k = 2,..., K$.

(ii) Le même problème peut se poser :

(a) pour un **modèle non linéaire** (à **perturbation** additive) (ie $y = F (b) + u$) ;

(b) pour un **modèle d'interdépendance** linéaire (ie $Y B' + X C' = U$, avec $E U = M$) ou non linéaire (ie $F (X, Y) = U$, avec $E U = M$).