Mercredi 16 mars 2005 : 15h00-15h25 *(en parallèle avec la session 15)*

# Imputation Multiple : conditions de mise en oeuvre avec des « variables-flags » .

*Modou DIA*
Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques (CEPS),
International Networks for Studies in Technology, Environment, Alternatives, Development (INSTEAD)
Rue Emile Mark 44, B.P.48, L-4501 Differdange - Grand-Duché du Luxembourg.
Tel. : (+352) 585855-544 | Fax: (+352) 585560 E-mail: *modou.dia@ceps.lu*

La méthode d'imputation proposée découle de l'algorithme de EM. Sa formalisation au niveau théorique revient à Rubin. Elle a été développée sous forme de macros SAS par l'Université de Michigan (http://www.isr.umich.edu/src/smp/ive/ ).

Sous sa forme simple c'est-à-dire en une seule passe, elle peut être considérée comme un modèle séquentiel de régression. Son algorithme peut être décrit de la façon suivante :
soit U un ensemble de variables explicatives sans données manquantes,
soit $Y_1$, $Y_2$,…, $Y_k$, un ensemble de variables dépendantes ordonnées selon le taux croissant de données manquantes,
la séquence des imputations est déterminée par la factorisation que voici :
$[Y_1/U]$, $[Y_2/ U, Y_1]$, $[Y_k/ U, Y_1, .., Y_{k-1}]$ où $[Y_i/U]$ est la distribution conditionnelle jointe de Y sachant U. Autrement dit après chaque itération i, la variable $Y_i$ qui vient d'être imputée s'ajoute à l'ensemble des variables explicatives.

Selon la nature de la variable $Y_i$ à imputer, le modèle de régression peut revêtir cinq formes :

a) une régression linéaire généralisée si la variable $Y_i$ est continue ;
b) une régression logistique si la variable $Y_i$ est binaire ;
c) une régression polytomique si $Y_i$ est une variable catégorielle ;
d) une régression log-linéaire (loi de Poisson) si $Y_i$ est une variable discrète finie ;
e) une régression pour variables mixte si $Y_i$ est une variable mixte telle qu'une quantité de stocks (=0, ou >0 s'il y a lieu) .

La diversité des types de variables qui peuvent être traités ainsi que la prise en compte de leur structure corrélative constituent un point fort de cette approche. Cependant, la pratique courante qui consiste à imputer une valeur unique de la donnée manquante présente de graves lacunes dans la mesure où elle ne restitue pas toute l'incertitude de la distribution des données imputées. La conséquence est la sous-estimation de la variance de la variable imputée par l'assimilation des données imputées à des données observées.

L'imputation multiple, en générant plusieurs valeurs issues d'une distribution adéquate, permet d'éviter cet écueil. Mais cela ne suffit pas dans la mesure où la présence de variables-flags dans sa mise en œuvre peut influer sur les résultats. Ce cas de figure se présente tout particulièrement lorsque la variable à imputer comporte des observations dites «non-concernées» comme par exemple dans le cas de retraités dans une variable relative au salaire. Le fait de leur attribuer une valeur nulle ou un code de donnée manquante diminue ou augmente les moments d'ordre 1 et d'ordre 2 qui servent à initialiser l'algorithme EM.

…/.

La dernière partie de cette contribution présente deux procédures d'imputation appliquées aux données individuelles du *Panel Socio-Economique Liewen zu Lëtzebuerg* (PSELL3) :

- La première impute les salaires en attribuant des valeurs nulles aux données manquantes des individus « non-concernés »
- La seconde réalise l'imputation de ces dernières, c'est à dire impute des valeurs pour les non-salariés, quitte à remplacer leurs résultats par des valeurs nulles après la procédure d'imputation grâce aux « variables-flags ».

Les résultats de ces modèles seront ensuite confrontés à une distribution des salaires issus d'une source administrative.