IMPUTATION PAR PRÉDICTION OU IMPUTATION AVEC ALÉA ?

Jean-Claude DEVILLE (*)

(*) CREST / ENSAI

La correction pour non-réponse peut se faire par repondération ou par imputation.

La première technique est la plus efficace car elle permet une estimation correcte de statistiques tant linéaires que non-linéaires. La seconde n'est qu'un pis-aller utilisée quand des contraintes non statistiques empêchent l'usage de la première.

Elle repose sur l'estimation de paramètres déterminant la loi de probabilité que la valeur manquante est supposée suivre. On a alors le choix entre deux types de stratégies :

-Imputer le meilleur prédicteur de la valeur manquante.

-Imputer une valeur prise au hasard dans sa loi estimée.

La première méthode permet d'obtenir des estimations correctes de statistiques linéaires ou semi-linéaires (ex: total sur un domaine) et d'en obtenir la variance (sorte d'estimation par régression partielle). En revanche, elle ne donne pas d'estimateur correct pour des statistiques non-linéaires. Elle est relativement robuste au modèle d'imputation.

La seconde permet des estimations correctes dans le cas non-linéaire (mais sans grande robustesse) au prix d'une augmentation artificielle de la variance due aux aléas mis dans les imputations. Celle-ci peut néanmoins être réduite dans de nombreux cas grâce à un échantillonnage équilibré des aléas.

En revanche, dans tous les cas, les liaisons (covariances par exemple) avec d'autres variables d'intérêt sont modifiées, ce qui rend fondamentalement vaine toute étude économétrique utilisant des données imputées.

Aucune des deux méthodes ne fait l'économie d'une étude soigneuse et d'une modélisation adéquate du mécanisme de réponse.