

IMPUTATION, MIVQUE ET PRÉSERVATION DES RELATIONS ENTRE VARIABLES

Brigitte Gelein ¹ & David Causeur ² & David Haziza ³

¹ *Ensai, Rennes, France, bgelein@ensai.fr*

² *Agrocampus Ouest, Rennes, France, david.causeur@agrocampus-ouest.fr*

³ *Université de Montréal, Montréal, Canada, haziza@DMS.UMontreal.CA*

La non réponse touche les enquêtes par sondage, les recensements et les données administratives. Qu'elle soit totale ou partielle, la non réponse peut affecter de façon significative la qualité des études statistiques. Alors que la non réponse totale se traite souvent par des procédures d'ajustement de poids, la non-réponse partielle est généralement traitée par des méthodes d'imputation.

L'imputation par la régression est souvent utilisée lorsque des variables auxiliaires sont disponibles sur l'ensemble de la population et présentent un lien fort avec les variables d'intérêt. Les estimateurs sur données d'enquête sont généralement calculés en traitant les valeurs imputées comme des valeurs observées. C'est pourquoi une méthode d'imputation doit fournir des estimateurs qui préservent l'absence de biais des formules standard utilisées sur les cas complets.

On s'intéresse ici aux estimateurs des paramètres mesurant les relations entre variables. Shao et Wang (2002) ont proposé une méthode d'imputation jointe par régression qui préserve ces relations. Un inconvénient de cette méthode tient à la variance additionnelle, dite d'imputation. Chauvet et Haziza (2011) ont élaboré une version équilibrée de la méthode de Shao et Wang permettant de préserver les relations entre deux variables tout en réduisant la variance d'imputation.

Ces deux dernières méthodes reposent sur l'utilisation de coefficients qui peuvent être estimés grâce aux estimateurs MIVQUE : Minimum In Variance QUadratic Estimators (MIVQUE). La méthode MIVQUE permet l'estimation de paramètres d'un modèle linéaire multivarié dans un contexte Missing At Random (MAR) au sens défini par Rubin. Elle est basée sur une étude géométrique de la structure de covariance des variables.

Nous proposons une méthode d'imputation où le calcul de MIVQUE est utilisé pour définir des contraintes sur les valeurs imputées finales. Des simulations montrent que la prise en compte de ces contraintes permet d'améliorer les valeurs imputées initiales obtenues, par exemple, avec la méthode de Shao et Wang. Nous observons en effet une réduction supplémentaire de la variance d'imputation.

Bibliographie

Chauvet, G., et Haziza, D. (2011). Fully efficient estimation of coefficients of correlation in the presence of imputed survey data, *Canadian Journal of Statistics*, 40, 1, 124-149

Shao, J., et Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation, *Journal of the American Statistical Association*, 97, 544-552.