Contribution à l'estimation du coefficient de corrélation en présence d'imputation par hot-deck aléatoire dans le cas d'enquêtes stratifiées à plusieurs degrés

Daniel Yapi, Catherine Vermandele, Jean-Jacques Droesbeke

▶ **To cite this version:**

Daniel Yapi, Catherine Vermandele, Jean-Jacques Droesbeke. Contribution à l'estimation du coefficient de corrélation en présence d'imputation par hot-deck aléatoire dans le cas d'enquêtes stratifiées à plusieurs degrés. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386646

HAL Id: inria-00386646
https://inria.hal.science/inria-00386646

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Contribution à l'estimation du cœfficient de corrélation en présence d'imputation par hot-deck aléatoire dans le cas d'enquêtes stratifiées à plusieurs degrés

Daniel Yapi, Catherine Vermandele & Jean-Jacques DROESBEKE

Av. F.D. Roosevelt 50 B-1050 Bruxelles - Belgium

Résumé

Les paramètres qui ont souvent fait l'objet d'estimation en présence d'imputation sont les moyennes ou les totaux qui sont généralement estimés sans biais par un nombre important de méthodes d'imputation. Cependant, en pratique, on est souvent amené à estimer des paramètres plus complexes tels que des cœfficients de corrélation, de régression, etc., dont l'estimation nécessite d'estimer des composantes multivariées. L'imputation a tendance à déformer les relations entre variables. Une méthode d'imputation hot-deck aléatoire modifiée qui permet d'obtenir un estimateur non biaisé du cœfficient de corrélation est proposée. Une étude de simulation a été effectuée afin de montrer l'efficacité de la nouvelle méthode. Un estimateur convergent de la variance de l'estimateur proposé a été obtenu par la méthode du Jackknife.

Mots clés: Jackknife, imputation hot-deck aléatoire modifiée, estimation de variance.

Abstract

The parameters commonly estimated in the presence of imputation are means and totals. They are generally unbiased by most imputation methods. However, in practice, it is often required to estimate more complex parameters such as regression coefficients, correlation coefficients and these latter are not unbiased by the commonly used methods. A modified hot-deck imputation method is proposed that preserves unbiasedness for correlation cœfficient. A small simulation study show that modified hot-deck imputation methods produces an approximatively unbiased estimator. An asymptotically unbiased and consistent variance estimator for the proposed estimator of correlation cœfficient is derived using a jackknife method. .

KEY WORDS: Jackknife, modified hot-deck imputation, variance estimation.

Présentation

L'imputation est le processus utilisé pour déterminer et attribuer des valeurs "plausibles" de remplacement aux données manquantes. Elle a pour avantage de produire un fichier de données complet et cohérent. Une grande partie de la littérature sur l'imputation traite des effets de celle-ci sur l'estimation de paramètres simples tels que moyennes, totaux et d'autres paramètres univariés, où elle mène généralement à des estimateurs sans biais (Bailar et Bailar, 1978; Ford, 1976; Kalton, 1981).
En pratique, lors des enquêtes, on est régulièrement amené à estimer des paramètres plus complexes tels une moyenne d'un domaine, une différence de moyennes ou de proportions, un cœfficient de régression ou de corrélation et bien d'autres paramètres impliquant deux ou plusieurs variables.
Santos (1981), Kalton et Kasprzyk (1982) ont montré que les estimateurs imputés de moyennes de domaines ou de cœfficients de corrélation peuvent être biaisés par certaines méthodes d'imputation.

Dans une population finie U de taille N, un cœfficient de corrélation entre deux variables x et y est donné par

$$\rho_{xy} = \frac{\left(\sum_{i\in U} x_i y_i - \frac{XY}{N}\right)}{\left[\left(\sum_{i\in U} x_i^2 - \frac{X^2}{N}\right)\left(\sum_{i\in U} y_i^2 - \frac{Y^2}{N}\right)\right]^{1/2}} \tag{1}$$

où $X = \sum_{i\in U} x_i$ et $Y = \sum_{i\in U} y_i$.
En l'absence de non réponse, un estimateur approximativement sans biais de (1) est donné par

$$\hat{\rho}_{xy} = \frac{\left(\sum_{i\in s} w_i x_i y_i - \frac{\hat{X}\hat{Y}}{\hat{N}}\right)}{\left[\left(\sum_{i\in s} w_i x_i^2 - \frac{\hat{X}^2}{\hat{N}}\right)\left(\sum_{i\in s} w_i y_i^2 - \frac{\hat{Y}^2}{\hat{N}}\right)\right]^{1/2}} \tag{2}$$

où $\hat{N} = \sum_{i\in s} w_i$, $\hat{X} = \sum_{i\in s} w_i x_i$ et $\hat{Y} = \sum_{i\in s} w_i y_i$, w_i désignant le poids de sondage de l'unité i dans l'échantillon s.
Quand il y a des valeurs manquantes parmi les valeurs observées des variables x et y, l'estimateur imputé de (1) est donné par

$$\hat{\rho}^*_{xy} = \frac{\left(\sum_{i\in s} w_i x^*_i y^*_i - \frac{\hat{X}^*\hat{Y}^*}{\hat{N}}\right)}{\left[\left(\sum_{i\in s} w_i x^{*2}_i - \frac{\hat{X}^*}{\hat{N}}\right)\left(\sum_{i\in s} w_i y^{*2}_i - \frac{\hat{Y}^*}{\hat{N}}\right)\right]^{1/2}} \tag{3}$$

où

$$\hat{X}^* = \sum_{i\in s} w_i a_i x_i + \sum_{i\in s} w_i (1-a_i) x^*_i \tag{4}$$

et

$$\hat{Y}^* = \sum_{i\in s} w_i b_i y_i + \sum_{i\in s} w_i (1-b_i) y^*_i \tag{5}$$

sont les estimateurs imputés respectifs de X et Y, a_i et b_i sont des variables indicatrices qui sont égales à 1 si les valeurs x_i et y_i, respectivement, ont été observées (c'est-à-dire) si l'unité i est répondante pour la variable x et la variable y respectivement, et égales à 0 sinon, x^*_i et y^*_i sont les valeurs imputées de x_i et y_i, respectivement.
Sous les méthodes d'imputation usuelles, l'estimateur (3) qui tient compte des valeurs imputées comme si elles étaient observées est biaisé du fait de la difficulté à estimer sans biais la composante $\sum_{i\in U} x_i y_i$ de (1) qui est une mesure de la relation entre les variables x et y. Kalton et Kasprzyk (1986) ont en effet montré que les méthodes standards d'imputation ne préservent pas les relations entre variables: elles ont plutôt tendance à les atténuer.

Dans cet article, nous présentons la méthode hot-deck aléatoire modifiée dérivée de la méthode d'imputation hot-deck aléatoire standard et qui permettra d'estimer sans biais le cœfficient de corrélation. L'approche adoptée est identique à celle présentée par Haziza et Rao (2004) pour l'estimation sans biais du cœfficient de corrélation à la seule différence que les termes résiduels issus de l'échantillonnage aléatoire sont obtenus selon un modèle qui préserve le lien originel entre les variables x et y.

La méthode d'imputation hot-deck aléatoire standard peut être représentée par le modèle général suivant:

$$x_i = f(\mathbf{z_i}) + \varepsilon_i, \qquad E(\varepsilon_i) = 0, \quad E(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j, \quad E(\varepsilon_i^2) = \sigma_\varepsilon^2\ , \tag{6}$$

$$y_i = g(\mathbf{z_i}) + \eta_i, \qquad E(\eta_i) = 0, \quad E(\eta_i, \eta_j) = 0,\ i \neq j, \quad E(\eta_i^2) = \sigma_\eta^2\,, \tag{7}$$

où $\mathbf{z}$ est un vecteur de variables auxiliaires disponibles pour toutes les unités dans l'échantillon s. Dans le cas de l'imputation hot-deck, $\mathbf{z}$ est un vecteur de variables indicatrices qui servent à indexer les cellules "hot-deck", donc à la formation des classes d'imputation. La valeur imputée pour l'item i est donnée par

$$x_i^* = \hat{f}(\mathbf{z_i}) + \varepsilon_i^* \,, \tag{8}$$

$$y_i^* = \hat{g}(\mathbf{z_i}) + \eta_i^* \,, \tag{9}$$

où ε_i^* et η_i^* sont tirés au hasard dans l'ensemble des résidus standardisés correspondant aux items répondants.
Pour la méthode hot-deck modifiée, quand x_i et y_i sont manquants, ε_i^* et η_i^* sont obtenus selon un modèle qui garantit que les valeurs imputées x_i^* et y_i^* conservent la même corrélation que celle de la paire originale x_i et y_i.

Bibliographie

[1] Bailar, III, J.C., et Bailar, B.A. (1978), Comparison of two procedures for imputing missing survey values, *American Statistical Association*, Proceedings of the Section on Survey Research Methods.
[2] Ford, B. (1976), Missing data procedures: a comparative study, *American Statistical Association*, proc. sect. Soc. stat., 324–329.
[3] Haziza, D. et Rao, J.N.(2004), Inférence pour des statistiques bivariées en présence d'imputation dans le cas d'enquêtes stratifiées à degrés multiples, *Echantillonnage et Méthodes d'enquêtes*, Ardilly. P., 189–196.
[4] Kalton, G. (1981), *Compasating for missing data*, Survey Research Center, University of Michigan.
[5] Kalton, G. et Kasprzyk, D.(1982), Imputation for missing survey responses, *American Statistical Association*, Proceedings of the Section on Survey Research Methods.
[6] Kalton, G. et Kasprzyk, D.(1986), The treatment of missing survey data. *Survey Methodology*, **12**, 1–16.
[7] Santos, R. (1981), *The effects of imputation on complex statistics*, Survey Research Center, University of Michigan.