Imputation équilibrée pour la non-réponse en fromage suisse

Audrey-Anne Vallée
et
Yves Tillé

Université de Neuchâtel

Journées de méthodologie statistique de l'Insee
14 juin 2018
Paris

Non-réponse partielle

- **Une seule variable** est sujette à la non-réponse.

Sexe	Taille	Poids
H	175	68
F	160	55
H	180	?
F	165	?

Non-réponse partielle

- **Une seule variable** est sujette à la non-réponse.

Sexe	Taille	Poids
H	175	68
F	160	55
H	180	?
F	165	?

- **Toutes les variables** de l'enquête sont sujettes à la non-réponse.

Sexe	Taille	$P_{t=1}$	$P_{t=2}$	$P_{t=3}$
H	175	68	67	68
F	160	55	58	?
H	180	70	?	?
F	165	?	?	?

Monotone

Sexe	Taille	Poids
H	175	68
F	160	?
H	?	70
?	165	?

Non-Monotone

Non-réponse en fromage suisse

Non-réponse en fromage suisse (non-monotone)

Toutes les variables d'une enquête contiennent des valeurs manquantes sans schéma particulier.

Traitements

- Méthodes d'imputation par donneur (Andridge et Little, 2010; Judkins, 1997).
- Méthodes d'imputation itératives: une séquence de modèles de régression entre les variables (Raghunathan et coll., 2001).

Non-réponse en fromage suisse

Propriétés souhaitées d'une méthode d'imputation

- Préserver les distributions des variables;
- Préserver les relations entre les variables;
- Imputer par des valeurs réalistes.

Non-réponse en fromage suisse

Propriétés souhaitées d'une méthode d'imputation

- Préserver les distributions des variables;
- Préserver les relations entre les variables;
- Imputer par des valeurs réalistes.

Imputation équilibrée par les K plus proches voisins (Hasler et Tillé, 2016)

- Imputation pour une variable;
- Méthode par donneur (aléatoire);
  - Variables continues et catégorielles;
  - Qu'un donneur par non-répondant;
- Imputation par donneurs proches (voisins);
- Échantillonnage équilibré;
- Si les valeurs observées étaient imputées, les estimations des totaux imputés et des totaux des valeurs connues devraient être les mêmes.

Non-réponse en fromage suisse

Propriétés souhaitées d'une méthode d'imputation

- Préserver les distributions des variables;
- Préserver les relations entre les variables;
- Imputer par des valeurs réalistes.

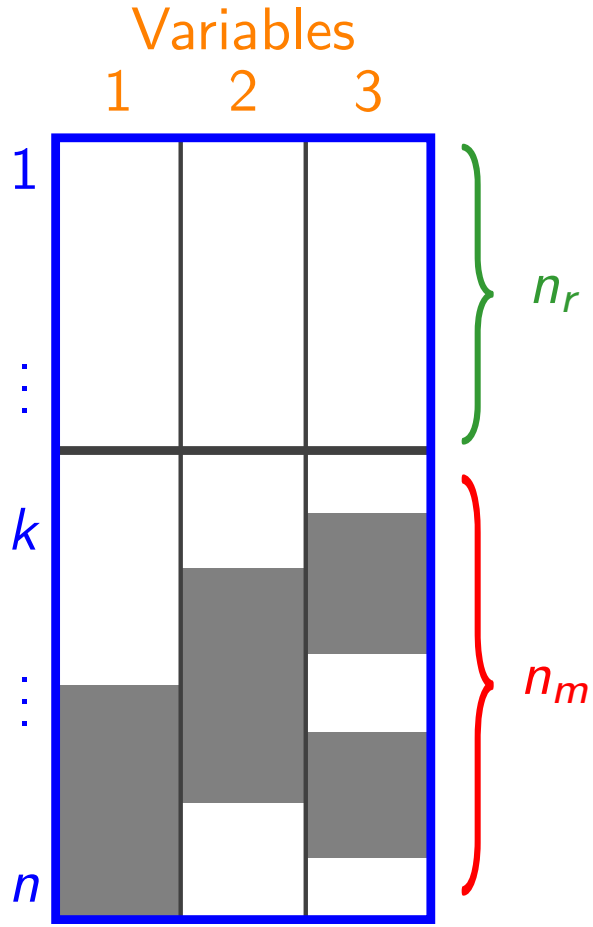
Imputation équilibrée par les K plus proches voisins (Hasler et Tillé, 2016)

- Imputation pour une variable;
- Méthode par donneur (aléatoire);
  - Variables continues et catégorielles;
  - Qu'un donneur par non-répondant;
- Imputation par donneurs proches (voisins);
- Échantillonnage équilibré;
- Si les valeurs observées étaient imputées, les estimations des totaux imputés et des totaux des valeurs connues devraient être les mêmes.

$\rightarrow$ Développons cette méthode pour la non-réponse en fromage suisse !

Non-réponse en fromage suisse

- Population U de taille N.
- J variables d'intérêt, $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$.
- Échantillon s de taille n.
- π_k, probabilité d'inclusion de l'unité k.
- $s_r \subset s$, n_r unités complètement observées.
- $s_m = s - s_r$, $n_m = n - n_r$ unités avec valeurs manquantes.
- Non-réponse non monotone.



Exigences de la méthode d'imputation

(i) Méthode par donneur: choisir les donneurs parmi s_r.

(ii) Un seul donneur par unité.

(iii) Donneur sélectionné parmi les K plus proches voisins de l'unité avec des valeurs manquantes.

(iv) Si les valeurs observées des non-répondants étaient imputées, l'estimateur du total de toutes les valeurs observées devraient rester inchangés.

Matrice de probabilités d'imputation

(i) Méthode par donneur: choisir les donneurs parmi s_r:

Matrice de probabilités d'imputation $\boldsymbol{\psi} = (\psi_{ik})$, où $(i, k) \in s_r \times s_m$.

- ψ_{ik}: probabilité que le répondant i donne ses valeurs au non-répondant k;
- $\psi_{ik} \geq 0$.

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{pmatrix} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

Matrice de probabilités d'imputation

(i) Méthode par donneur: choisir les donneurs parmi s_r:
Matrice de probabilités d'imputation $\boldsymbol{\psi} = (\psi_{ik})$, où $(i, k) \in s_r \times s_m$.

- ψ_{ik}: probabilité que le répondant i donne ses valeurs au non-répondant k;
- $\psi_{ik} \geq 0$.

(ii) Un seul donneur par unité non-répondante:

$$\sum_{i \in s_r} \psi_{ik} = 1.$$

Matrice de probabilités d'imputation

(i) Méthode par donneur: choisir les donneurs parmi s_r:
Matrice de probabilités d'imputation $\boldsymbol{\psi} = (\psi_{ik})$, où $(i,k) \in s_r \times s_m$.

- ψ_{ik}: probabilité que le répondant i donne ses valeurs au non-répondant k;
- $\psi_{ik} \geq 0$.

(ii) Un seul donneur par unité non-répondante:

$$\sum_{i \in s_r} \psi_{ik} = 1.$$

(iii) Donneur sélectionné parmi les K plus proches voisins de l'unité avec des valeurs manquantes:

$$\psi_{ik} = 0 \text{ si } i \notin \text{kpp}(k)$$

où $\text{kpp}(\ell) = \{j \in s_r | \, \text{rang}(d(j,\ell)) \leq K\}$ et $d(.,.)$ est une fonction de distance.

Matrice de probabilités d'imputation

(iv) Si les valeurs observées des non-répondants étaient imputées, l'estimateur du total de toutes les valeurs observées devraient rester inchangés:
Pour $j = 1, \ldots, J$,

$$\sum_{k \in s_m} d_k r_{kj} \underbrace{\sum_{i \in s_r} \psi_{ik} x_{ij}}_{x^*_{kj}} = \sum_{k \in s_m} d_k r_{kj} x_{kj},$$

où $d_\ell = 1/\pi_\ell$ et $r_{\ell j}$ vaut 1 si l'unité ℓ a répondu à la variable j, 0 sinon.

Matrice de probabilités d'imputation

(iv) Pour $j = 1, \ldots, J$,

$$\sum_{k \in s_m} d_k r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}.$$

Sexe	Taille
0	175
1	160
0	?
?	165
0	165

Sexe	Taille
0	175
1	160
0	175
1	**160**
0	**160**

Matrice de probabilités d'imputation

(iv) Pour $j = 1, \ldots, J$,

$$\sum_{k \in s_m} d_k r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}$$

$$\sum_{i \in s_r} \left(\sum_{k \in s_m} d_k r_{kj} \psi_{ik} \right) r_{ij} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}.$$

Algorithme: ψ_{ik} calculés par calage:

$$\text{Poids initiaux} \quad \psi_{ik}^0 = \begin{cases} \frac{1}{K} & \text{si } i \in \text{kpp}(k), \\ 0 & \text{sinon.} \end{cases}$$

Itérerations : Caler, normaliser.

Matrice d'imputation

Matrice de probabilités d'imputation

$$\psi = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

Matrice d'imputation

$$\phi = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- ϕ_{ik}: 1 si l'unité i est choisie comme donneur pour l'unité k, 0 sinon.
- Un seul donneur est par non-répondant, $\sum_{i \in s_r} \phi_{ik} = 1$.
- Exigence (iv): donneurs doivent être choisis de façon à respecter

$$\sum_{k \in s_m} \sum_{i \in s_r} \phi_{ik}\, d_k\, r_{kj}\, x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} \psi_{ik}\, d_k\, r_{kj}\, x_{ij} \left(= \sum_{k \in s_m} d_k\, r_{kj}\, x_{kj} \right).$$

Matrice d'imputation

Exigence (iv): donneurs doivent être choisis de façon à respecter

$$\sum_{k \in s_m} \sum_{i \in s_r} \phi_{ik} \, d_k \, r_{kj} \, x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} \psi_{ik} \, d_k \, r_{kj} \, x_{ij}.$$

- Échantillonnage stratifié équilibré (Chauvet, 2009; Hasler et Tillé, 2014);
- n_m strates (non-répondants) formées;
- Un donneur est choisi par strate.
- Probabilité d'inclusion utilisée dans l'échantillonnage stratifié équilibré est ψ_{ik};
- Variable d'équilibrage associée est $\psi_{ik} \, d_k \, r_{kj} \, x_{ij}$.

Imputation

Valeur imputée: $x_{kj}^* = \sum_{i \in s_r} \phi_{ik} x_{ij}$

Total imputé: $\widehat{X}_j = \sum_{k \in s_r} d_k x_{kj} + \sum_{k \in s_m} r_{kj} d_k x_{kj} + \sum_{k \in s_m} (1 - r_{kj}) d_k x_{kj}^*$

Variante déterministe: $x_{kj}^* = \sum_{i \in s_r} \psi_{ik} x_{ij}$.

Discussion

- Détermination de K (pas trop grand).
- Méthode pour variables qualitative/quantitatives.
- Possibilité de forcer $\psi_{ik} = 0$ pour une raison quelconque.
- Modèles et principes.
- Programme en R.
- Estimation de la variance.

Andridge, R. R. et Little, R. J. A. (2010). A review of dot deck imputation for survey non-response. *International Statistical Review*, **78**, 40–64.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, **35**, 115–119.

Hasler, C. et Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, **74**, 81–94.

Hasler, C. et Tillé, Y. (2016). Balanced *k*-nearest neighbor imputation. *Statistics*, **105**, 11–23.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium*, 97. Statistics Canada.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. et Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, **27**, 85–95.