Balanced imputation for swiss cheese nonresponse

Audrey-Anne Vallée
and
Yves Tillé

University of Neuchâtel

UNECE Workshop on Statistical Data Editing
September 20th 2018
Neuchâtel

Item nonresponse

- **Only one variable** is subject to nonresponse.

Sex	Height	Weight
M	175	68
F	160	55
M	180	?
F	165	?

Item nonresponse

- **Only one variable** is subject to nonresponse.

Sex	Height	Weight
M	175	68
F	160	55
M	180	?
F	165	?

- **All variables** of the survey are subject to nonresponse.

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

Monotone

Sex	Height	Weight
H	175	68
F	160	?
H	?	70
?	165	?

Non Monotone

Swiss cheese nonresponse

Swiss cheese nonresponse (non monotone)

All variables of the survey contain missing values, without a particular pattern.

Treatments

- Donor imputation methods (Andridge & Little, 2010; Judkins, 1997).
- Iterative imputation methods: a sequence of regression models between the variables (Raghunathan *et al.* , 2001).

Swiss cheese nonresponse

Properties of interest for an imputation method

- Preserve the distributions of the variables;
- Preserve the relationships between variables;
- Impute by realistic values.

Swiss cheese nonresponse

Properties of interest for an imputation method

- Preserve the distributions of the variables;
- Preserve the relationships between variables;
- Impute by realistic values.

Balanced K-nearest neighbor imputation (Hasler & Tillé, 2016)

- Imputation for one variable;
- Donor imputation method (random);
  - Continuous and discrete variables;
  - Only one donor per nonrespondent;
- Imputation by near donors (neighbors);
- Balanced sampling;
- If the observed values were imputed, the estimators with imputed values and the estimators with the observed values would be the same.

Swiss cheese nonresponse

Properties of interest for an imputation method

- Preserve the distributions of the variables;
- Preserve the relationships between variables;
- Impute by realistic values.

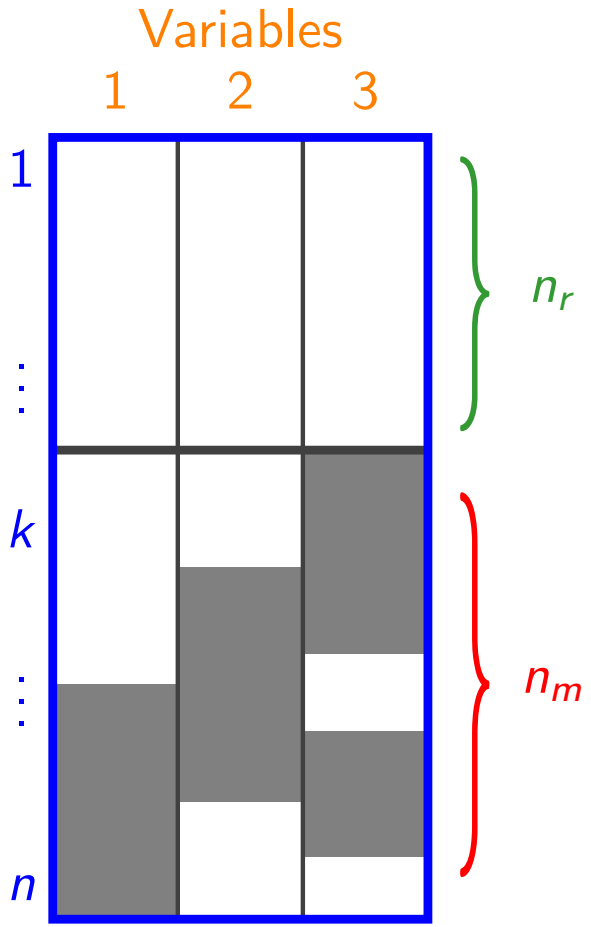
Balanced K-nearest neighbor imputation (Hasler & Tillé, 2016)

- Imputation for one variable;
- Donor imputation method (random);
  - Continuous and discrete variables;
  - Only one donor per nonrespondent;
- Imputation by near donors (neighbors);
- Balanced sampling;
- If the observed values were imputed, the estimators with imputed values and the estimators with the observed values would be the same.

→ Extend this method for swiss cheese nonresponse !

Swiss cheese nonresponse

- Population U of size N.
- J variables of interest, $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$.
- Sample s of size n.
- π_k, inclusion probabilities of the unit k.
- $s_r \subset s$, n_r completely observed units.
- $s_m = s - s_r$, $n_m = n - n_r$ units with missing values.
- Non monotone nonresponse.



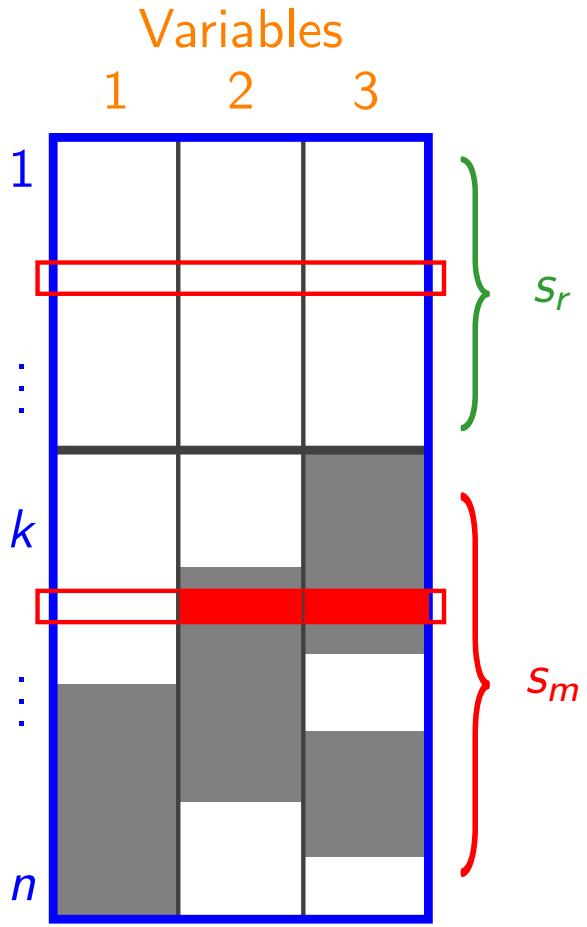
Requirements for the imputation methods

(i) Donor imputation method: select the donors in s_r.

(ii) Only one donor per unit.

(iii) Donor selected in the set of K nearest neighbors of the unit with missing values.

(iv) Described in the following slides.



Matrix of imputation probabilities

(i) Donor imputation: select the donors in s_r:
Matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.

- ψ_{ik}: Probability that respondent i gives its values to nonrespondent k;
- $\psi_{ik} \geq 0$.

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

Matrix of imputation probabilities

(i) Donor imputation: select the donors in s_r:
Matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.

- ψ_{ik}: Probability that respondent i gives its values to nonrespondent k;
- $\psi_{ik} \geq 0$.

(ii) Only one donor per nonrespondent:

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

Matrix of imputation probabilities

(i) Donor imputation: select the donors in s_r:
Matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.

- ψ_{ik}: Probability that respondent i gives its values to nonrespondent k;
- $\psi_{ik} \geq 0$.

(ii) Only one donor per nonrespondent:

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

(iii) Donor selected in the set of K nearest neighbors of the unit with missing values:

$$\psi_{ik} = 0 \text{ if } i \notin \text{knn}(k)$$

where $\text{knn}(\ell) = \{j \in s_r | \, \text{rang}(d(j, \ell)) \leq K\}$ and $d(.,.)$ is a distance function.

Matrix of imputation probabilities

(iv) If the observed values of the nonrespondents were imputed, the total estimator of each variable should remain unchanged:

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

where $d_\ell = \pi_\ell^{-1}$ and $r_{\ell j}$ is 1 if unit ℓ responded to variable j, 0 otherwise, for $j = 1, \ldots, J$.

Matrix of imputation probabilities

(iv) If the observed values of the nonrespondents were imputed, the total estimator of each variable should remain unchanged:

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

	Incomplete		Imputed	
	x_1	x_2	x_1	x_2
s_r	8	4	8	4
	6	2	6	2
s_m	?	3	*8*	*4*
	?	3	*6*	*2*
Total in s_m		6		6

Matrix of imputation probabilities

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

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

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

Algorithm: ψ_{ik} are calculated by calibration:

$$\text{Initial weights} \quad \psi_{ik}^0 = \begin{cases} \frac{1}{K} & \text{if } i \in \text{knn}(k), \\ 0 & \text{otherwise.} \end{cases}$$

Iterations : calibrate, normalize.

Imputation matrix

Matrix of imputation probabilities

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

Imputation matrix

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

- ϕ_{ik}: 1 if unit i is the selected donor for unit k, 0 otherwise.
- Only one donor per nonrespondent, $\sum_{i \in s_r} \phi_{ik} = 1$.
- Requirement (iv): donors are selected in order to respect

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

Imputation matrix

Requirement (iv): donors are selected in order to respect

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

- Stratified balanced sampling (Chauvet, 2009; Hasler & Tillé, 2014);
- n_m strata (nonrespondent) are created;
- One donor is selected per stratum.
- Inclusion probability used in the stratified balanced sampling is ψ_{ik};
- Balancing variable is $\psi_{ik}\, d_k r_{kj} x_{ij}$.

Imputation

Imputed value: $x^*_{kj} = \sum_{i \in s_r} \phi_{ik} x_{ij}$

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

Discussion

- Determine K (not to large).
- Method for qualitative/quantitative variables.
- Possibility to force $\psi_{ik} = 0$ for any reason.
- Models and principles.
- R program.
- Variance estimation.

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

Chauvet, Guillaume. 2009. Stratified Balanced Sampling. *Survey Methodology*, **35**, 115–119.

Hasler, Caren, & Tillé, Yves. 2014. Fast Balanced Sampling for Highly Stratified Population. *Computational Statistics and Data Analysis*, **74**, 81–94.

Hasler, Caren, & Tillé, Yves. 2016. Balanced k-Nearest Neighbor Imputation. *Statistics*, **105**, 11–23.

Judkins, David R. 1997. Imputing for Swiss cheese patterns of missing data. *Page 97 of: Proceedings of Statistics Canada Symposium*.

Raghunathan, Trivellore E., Lepkowski, James M., van Hoewyk, John, & Solenberger, Peter W. 2001. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, **27**(1), 85–95.