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
Introduction - Swiss cheese nonresponse

Item nonresponse

Swiss cheese nonresponse

Context

Swiss cheese nonresponse

Requirements

Matrix of imputation probabilities

Imputation

Imputation matrix

Imputation
Item nonresponse

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

<table>
<thead>
<tr>
<th>Sex</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>175</td>
<td>68</td>
</tr>
<tr>
<td>F</td>
<td>160</td>
<td>55</td>
</tr>
<tr>
<td>M</td>
<td>180</td>
<td>?</td>
</tr>
<tr>
<td>F</td>
<td>165</td>
<td>?</td>
</tr>
</tbody>
</table>
Item nonresponse

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

<table>
<thead>
<tr>
<th>Sex</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>175</td>
<td>68</td>
</tr>
<tr>
<td>F</td>
<td>160</td>
<td>55</td>
</tr>
<tr>
<td>M</td>
<td>180</td>
<td>?</td>
</tr>
<tr>
<td>F</td>
<td>165</td>
<td>?</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sex</th>
<th>Height</th>
<th>$P_{t=1}$</th>
<th>$P_{t=2}$</th>
<th>$P_{t=3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>175</td>
<td>68</td>
<td>67</td>
<td>68</td>
</tr>
<tr>
<td>F</td>
<td>160</td>
<td>55</td>
<td>58</td>
<td>?</td>
</tr>
<tr>
<td>H</td>
<td>180</td>
<td>70</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>F</td>
<td>165</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

**Monotone**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Height</th>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>175</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>F</td>
<td>160</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>H</td>
<td>?</td>
<td>?</td>
<td>70</td>
</tr>
<tr>
<td>F</td>
<td>?</td>
<td>165</td>
<td>?</td>
</tr>
</tbody>
</table>

**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!
Introduction - Swiss cheese nonresponse

Item nonresponse
Swiss cheese nonresponse

Context
Swiss cheese nonresponse
Requirements

Matrix of imputation probabilities

Imputation
Imputation matrix
Imputation
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 $n$.
- $\pi_k$, 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.
Introduction - Swiss cheese nonresponse

Item nonresponse
Swiss cheese nonresponse

Context
Swiss cheese nonresponse
Requirements

Matrix of imputation probabilities

Imputation
Imputation matrix
Imputation
Matrix of imputation probabilities

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

\[ \psi_{ik} : \text{Probability that respondent } i \text{ gives its values to nonrespondent } k; \]
\[ \psi_{ik} \geq 0. \]

\[
\psi = \begin{pmatrix}
\psi_{11} & \psi_{12} & \psi_{13} \\
\psi_{21} & \psi_{22} & \psi_{23} \\
\psi_{31} & \psi_{32} & \psi_{33}
\end{pmatrix} = \begin{pmatrix}
0 & 0.5 & 0.5 \\
0.5 & 0.5 & 0 \\
0.5 & 0 & 0.5
\end{pmatrix}
\]
Matrix of imputation probabilities

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

▶ $\psi_{ik}$: 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 $\psi = (\psi_{ik})$, where $(i, k) \in s_r \times s_m$.
- $\psi_{ik}$: 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:

\[
\sum_{k \in s} d_k r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj},
\]

where \(d_\ell = \pi_\ell^{-1}\) and \(r_{\ell j}\) is 1 if unit \(\ell\) 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:

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

Where

\[
d_{\ell} = \pi - 1 \quad \text{and} \quad r_{j\ell} = 1 \text{ if unit } \ell \text{ responded to variable } j, \quad 0 \text{ otherwise, for } j = 1, \ldots, J.
\]

<table>
<thead>
<tr>
<th></th>
<th>Incomplete</th>
<th>Imputed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_1$</td>
<td>$x_2$</td>
</tr>
<tr>
<td>$s_r$</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>$s_m$</td>
<td>?</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>3</td>
</tr>
<tr>
<td>Total in $s_m$</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
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: \( \psi_{ik} \) are calculated by calibration:

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

Iterations: calibrate, normalize.
<table>
<thead>
<tr>
<th>Introduction - Swiss cheese nonresponse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item nonresponse</td>
</tr>
<tr>
<td>Swiss cheese nonresponse</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss cheese nonresponse</td>
</tr>
<tr>
<td>Requirements</td>
</tr>
</tbody>
</table>

| Matrix of imputation probabilities     |

<table>
<thead>
<tr>
<th>Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imputation matrix</td>
</tr>
<tr>
<td>Imputation</td>
</tr>
</tbody>
</table>
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

\[ \phi = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \]

- \( \phi_{ik} \): 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 \( \psi_{ik} \);
- 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: \( \hat{X}_j = \sum_{k \in s_r} d_k x_{kj} + \sum_{k \in s_m} r_{kj} d_k x_{kj} + \sum_{k \in s_m} (1 - r_{kj}) d_k x_{kj}^* \)
Discussion

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


