Assessing CANCEIS imputation uncertainty

Li-Chun Zhang

S3RI & Social Statistics and Demography, UK

(L.Zhang@soton.ac.uk); Statistics Norway, Norway

Damião N. da Silva

Departamento de Estatística, UFRN, Brazil

Introduction

- UK Census 2011: sequential modular implementation of

CANCEIS *(Aldrich et al., 2012; Wardman et al., 2012)*

- Elementary outputs *(CANCEIS, 2009)*

  - near minimum change imputation actions (NMCIAs)

  - imputation (expectation, variance): conditional & unit-level

- Regarding the available values for a given module:

  - matching values (for donor identification): may be *imputed*

  - target values (for imputation): always *observed* and/or *accepted*

- **Unconditional** imputation expectation & variance?

NMCIAs elementary outputs: given action matrix $\mathbf{\Delta}_{jkm}$

Imputation of y_k module m unit j & donor $1, ..., M_{jkm}$:

$$\begin{pmatrix} 1 & y_{1,jkm} & p_{1,jkm} \\ \vdots & \vdots & \vdots \\ \ell & y_{\ell,jkm} & p_{\ell,jkm} \\ \vdots & \vdots & \vdots \\ M_{jkm} & y_{M_{jkm},jkm} & p_{M_{jkm},jkm} \end{pmatrix} \rightarrow \begin{array}{l} \hat{y}_{jkm} = y_{\ell,jkm} \\ \ell \sim \text{Multinomial}(1, \mathbf{p}_{jkm}) \end{array}$$

$$E_{jkm} \stackrel{def}{=} \text{E}_{imp}(\hat{y}_{jkm}|\mathbf{\Delta}_{jkm}) = \sum_{\ell=1}^{M_{jkm}} p_{\ell,jkm} y_{\ell,jkm}$$

$$V_{jkm} \stackrel{def}{=} \text{V}_{imp}(\hat{y}_{jkm}|\mathbf{\Delta}_{jkm}) = \sum_{\ell=1}^{M_{jkm}} p_{\ell,jkm} y_{\ell,jkm}^2 - E_{jkm}^2$$

Extended elementary outputs: variable k in module m

- Random typically large action matrix $\boldsymbol{\Delta}_{km} = (\delta_{ij} p_{ij})$

  $\delta_{ij} = 1$ if unit i is donor for imputation of y_{jkm}, and 0 otherwise

  (NB. For accepted y_{jkm}: $\delta_{jj} = 1$ and $\delta_{ij} = 0$ if $i \neq j$)

  $p_{ij} = p_{i,jkm}$ if $\delta_{ij} = 1$, and 0 otherwise

  (NB. For accepted y_{jkm}: $p_{ij} = \delta_{ij}$)

- Extended elementary outputs (N_g = size of relevant imputation group)

$$\begin{cases} e_{jkm} = \sum_{i=1}^{N_g} p_{ij} y_{ij} = \begin{cases} y_{jkm} & \text{if accepted} \\ E_{jkm} & \text{if imputed} \end{cases} \\ v_{jkm} = \sum_{i=1}^{N_g} p_{ij} y_{ij}^2 - e_{jkm}^2 = \begin{cases} 0 & \text{if accepted} \\ V_{jkm} & \text{if imputed} \end{cases} \end{cases}$$

Imputation uncertainty of CANCEIS: Monte Carlo

- Repeat imputation of all modules and variables B times

$$(\hat{y}_{jkm}^{(b)}, e_{jkm}^{(b)}, v_{jkm}^{(b)}) \qquad \text{where} \quad \hat{y}_{jkm}^{(b)} = y_{jkm} \text{ or not}$$

- For domain total estimator: $\hat{t}_{dkm} = \sum_{j \in U_d} \hat{y}_{jkm}$

$$\widehat{\mathrm{E}}_{imp}(\hat{t}_{dkm}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{j \in U_d} e_{jkm}^{(b)} \neq \frac{1}{B} \sum_{b=1}^{B} \hat{t}_{jkm}^{(b)}$$

$$\widehat{\mathrm{V}}_{imp}(\hat{t}_{dkm}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{j \in U_d} v_{jkm}^{(b)}$$

$$+ \frac{1}{B-1} \sum_{b=1}^{B} \sum_{g=1}^{G} \left(\sum_{j \in U_{gd}} e_{jkm}^{(b)} - \widehat{\mathrm{E}}_{imp}(\hat{t}_{gdkm}) \right)^2$$

NB. independent imputation groups $g = 1, ..., G$; applies as well to $U_d = \{j\}$

Poor-man's option: sampling?

Full repetitions for selected imputation groups only

$$\frac{1}{B}\sum_{b=1}^{B}\sum_{j\in U_d} e_{jkm}^{(b)} = \sum_{g=1}^{G}\sum_{j\in U_{gd}} \bar{e}_{jkm} \qquad \text{for} \quad \bar{e}_{jkm} = \frac{1}{B}\sum_{b=1}^{B} e_{jkm}^{(b)}$$

$$\frac{1}{B}\sum_{b=1}^{B}\sum_{j\in U_d} v_{jkm}^{(b)} = \sum_{g=1}^{G}\sum_{j\in U_{gd}} \bar{v}_{jkm} \qquad \text{for} \quad \bar{v}_{jkm} = \frac{1}{B}\sum_{b=1}^{B} v_{jkm}^{(b)}$$

$$\frac{1}{B-1}\sum_{b=1}^{B}\sum_{g=1}^{G}\left(\sum_{j\in U_{gd}}[e_{jkm}^{(b)} - \bar{e}_{jkm}]\right)^2 = \sum_{g=1}^{G}\sum_{i,j\in U_{gd}}\left(\frac{1}{B-1}\sum_{b=1}^{B} c_{ikm}^{(b)} c_{jkm}^{(b)}\right)$$

where $c_{jkm}^{(b)} = e_{jkm}^{(b)} - \bar{e}_{jkm}$. Weighing of sample imputation groups

- partition of accepted and imputed values & adjustment accordingly
- make use of census elementary outputs $(e_{jkm}, v_{jkm}) = (\hat{\bar{e}}_{jkm}, \hat{\bar{v}}_{jkm})$

References

[1] Aldrich, S., Wardman, L. and Rogers, S. (2012). The Practical Implementation of the 2011 UK Census Imputation Methodology. *UN/ECE Work Session on Statistical Data Editing, Oslo 24-26 September 2012.*

[2] Wardman, L. Aldrich, S., and Rogers, S. (2012). Item imputation of Census data in an automated production environment; advantages, disadvantages and diagnostics. *UN/ECE Work Session on Statistical Data Editing, Oslo 24-26 September 2012.*

[3] CANCEIS, Development Team (2009). *CANCEIS Version 4.5 User Guide.* Social Survey Methods Division, Statistics Canada.