Doubly robust inference for the distribution function in the presence of missing survey data

Guillaume Chauvet (ENSAI)

Graybill Conference on Modern Survey Statictics
Fort Collins, 11/06/2013

This talk is based on joint work with :

1. Jean-Claude Deville (ENSAI), David Haziza (Univ. Montréal)
   On balanced random imputation in surveys, Biometrika, 98.
2. Hélène Boistard (TSE), David Haziza (Univ. Montréal)
   Doubly robust inference for complex parameters in the presence of missing survey data, work in progress.

Notation

The Imputation Model

The Non-Response Model

Notation

We are interested in a finite population $U = \{1, \ldots, N\}$. Denote by y_i the value taken by a characteristic y on some unit i in U.

A sample S is selected in U by means of a sampling design $p(\cdot)$. We note $\pi_i = \mathbb{P}(i \in S)$ and $d_i = 1/\pi_i$ the design weight.

In case of full response for the y-variable, we may estimate unbiasedly

$$t_y = \sum_{i \in U} y_i \quad \text{by} \quad \hat{t}_{y\pi} = \sum_{i \in S} d_i y_i,$$

$$F_{N,y}(t) = \frac{1}{N} \sum_{i \in U} 1(y_i \leq t) \quad \text{by} \quad \hat{F}_{N,y}(t) = \frac{1}{N} \sum_{i \in S} d_i 1(y_i \leq t).$$

In case of item non-response, the variable y is observed on a sub-sample S_r only. Denote by r_i the response indicator for unit i, and by $q(\cdot)$ the response mechanism.

The Imputation Model

The IM approach

A missing y_i may be replaced by an (artificial) imputed value y_i^*. This imputation can be motivated by an *imputation model* (IM ; see Särndal, 1992)

$$m: \ y_i = f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i,$$

where

- $f(\cdot\ ;\cdot)$ is a given function,
- z_i is a K-vector of auxiliary variables known for any $i \in s$,
- v_i is a known constant,
- β and σ are unknown parameters,
- the ϵ_i are assumed i.i.d., standardized with common df $F_\epsilon(\cdot)$.

Assumption on the response mechanism $q(\cdot)$: MAR data.

The IM approach

An estimator of β is obtained by solving the estimating equations

$$\sum_{i \in s} \omega_i r_i v_i^{-1} \{y_i - f(z_i; \beta)\} h_i = 0,$$

where

- $h_i = \partial f(z_i; \beta)/\partial \beta$,
- ω_i is an imputation weight attached to unit i (Haziza, 2009).

This leads to the *deterministic imputation mechanism* (DRI)

$$I: \; y_i^* = f(z_i; \hat{B}_r),$$

and under the model m the imputed estimator of the total

$$\hat{t}_{y,DRI} = \sum_{i \in S} d_i r_i y_i + \sum_{i \in S} d_i (1 - r_i) f(z_i; \hat{B}_r)$$

is approximately mpq-unbiased.

Estimation of the distribution function

The imputed estimator of the population df

$$\hat{F}_{y,DRI}(t) = \frac{1}{N}\sum_{i\in S} d_i r_i 1(y_i \leq t) + \frac{1}{N}\sum_{i\in S} d_i (1-r_i) 1\{f(z_i;\hat{B}_r) \leq t\}$$

is usually biased.

Possible solutions are to use :

- a bias-corrected estimator (Chambers and Dunstan, 1986),
- or a *random imputation mechanism* (RRI)

$$I: \; y_i^* = f(z_i;\hat{B}_r) + \hat{\sigma}\sqrt{v_i}\epsilon_i^*$$

with random residuals ϵ_i^*.

Estimation of the distribution function

This leads to the imputed estimator

$$\hat{F}_{y,RRI}(t) = \frac{1}{N}\sum_{i\in S} d_i r_i 1(y_i \leq t) + \frac{1}{N}\sum_{i\in S} d_i(1-r_i)1(y_i^* \leq t).$$

In the random imputation mechanism

$$I: \; y_i^* = f(z_i;\hat{B}_r) + \hat{\sigma}\sqrt{v_i}\epsilon_i^*,$$

the residuals ϵ_i^* may be generated from a parametric distribution, or selected from the set of observed estimated residuals

$$E_r = \left\{ e_j = \frac{y_j - f(z_j;\hat{B}_r)}{\hat{\sigma}\sqrt{v_j}};\; j \in S_r \right\} \quad \text{with} \quad \mathbb{P}(\epsilon_i^* = e_j) = \frac{\omega_j}{\sum_{k\in S}\omega_k r_k}.$$

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$\hat{F}_{y,I}(t) - \hat{F}_{y,N}(t) \quad = \quad N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(y_i^* \leq t) - 1(y_i \leq t)\}$$

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$\begin{aligned}
\hat{F}_{y,I}(t) - \hat{F}_{y,N}(t) &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(y_i^* \leq t) - 1(y_i \leq t)\} \\
y_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i. \\
y_i^* &= f(z_i; \hat{B}_r) + \hat{\sigma} \sqrt{v_i} \epsilon_i^*
\end{aligned}$$

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$
\begin{aligned}
T_1 &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(y_i^* \leq t) - 1(\tilde{y}_i \leq t)\} \\
y_i^* &= f(z_i; \hat{B}_r) + \hat{\sigma} \sqrt{v_i} \epsilon_i^* \\
\tilde{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i^*
\end{aligned}
$$

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$
\begin{aligned}
T_1 &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(y_i^* \leq t) - 1(\tilde{y}_i \leq t)\} \\
y_i^* &= f(z_i; \hat{B}_r) + \hat{\sigma} \sqrt{v_i} \epsilon_i^* \\
\tilde{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i^*
\end{aligned}
$$

Consistency of $(\hat{B}_r, \hat{\sigma})$ + $f(\cdot; \beta)$ continuous $\Rightarrow E|T_1| \to 0$.

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$
\begin{aligned}
T_2 &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(\tilde{y}_i \leq t) - 1(\hat{y}_i \leq t)\} \\
\tilde{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i^* \\
\hat{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \hat{\epsilon}_i, \quad \mathbb{P}(\epsilon_i^* = e_j, \hat{\epsilon}_i = \epsilon_j) = \frac{\omega_j}{\sum_{k \in S} \omega_k r_k}.
\end{aligned}
$$

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$
\begin{aligned}
T_2 &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(\tilde{y}_i \leq t) - 1(\hat{y}_i \leq t)\} \\
\tilde{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i^* \\
\hat{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \hat{\epsilon}_i, \quad \mathbb{P}(\epsilon_i^* = e_j, \hat{\epsilon}_i = \epsilon_j) = \frac{\omega_j}{\sum_{k \in S} \omega_k r_k}.
\end{aligned}
$$

$F_\epsilon(\cdot)$ absolutely continuous $\Rightarrow E|T_2| \to 0$.

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$
\begin{aligned}
T_3 &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(\hat{y}_i \leq t) - 1(y_i \leq t)\} \\
\hat{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \hat{\epsilon}_i \\
y_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i .
\end{aligned}
$$

Theorem (CDH, 2011)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions : $E|\hat{F}_{y,I}(t) - F_{y,N}(t)| \to 0$.

Main steps of the proof :
$E|\hat{F}_{y,N}(t) - F_{y,N}(t)| \to 0$ under standard conditions (Isaki and Fuller, 1982)

$$\begin{aligned} T_3 &= N^{-1} \sum_{i \in S} d_i (1 - r_i) \{1(\hat{y}_i \leq t) - 1(y_i \leq t)\} \\ \hat{y}_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \hat{\epsilon}_i \\ y_i &= f(z_i; \beta) + \sigma \sqrt{v_i} \epsilon_i . \end{aligned}$$

$$\max \left(\frac{d_i}{\sum_{j \in S} d_j} \right) = O(n^{-1}), \ \max \left(\frac{\omega_i}{\sum_{j \in S} \omega_j} \right) = O(n^{-1}) \Rightarrow V(T_3) \to 0.$$

Reducing the imputation variance

The imputed estimators suffer from an additional imputation variance, e.g.

$$\hat{t}_{y,RRI} = \hat{t}_{y,DRI} + \sum_{i \in S} d_i (1 - r_i) \hat{\sigma} \sqrt{v_i} \epsilon_i^* .$$

Reducing the imputation variance

The imputed estimators suffer from an additional imputation variance, e.g.

$$\hat{t}_{y,RRI} = \hat{t}_{y,DRI} + \sum_{i \in S} d_i (1 - r_i) \hat{\sigma} \sqrt{v_i} \epsilon_i^* .$$

Possible solutions are :

- to use several imputed values for each missing value : fractional imputation (Kim and Fuller, 2004 ;Fuller and Kim, 2005),
- to adjust the imputed values to eliminate the imputation variance (Chen, Rao and Sitter, 2000),
- to select the random residuals ϵ_i^* at random so that *balancing equations* are respected, e.g.

$$\sum_{i \in S} d_i (1 - r_i) \hat{\sigma} \sqrt{v_i} \epsilon_i^* = 0.$$

We may use the cube method (Deville and Tillé, 2004 ; Deville, 2006) or the rejective method of Fuller (2009).

The Non-Response Model

The NM approach

A modeling of the non-response mechanism may be useful, e.g. to prevent an incorrect specification of the imputation model. This leads to a *non-response model* (NM) for the response probability, e.g.

$$p_i \equiv p(z_i;\ \alpha)$$

for some unknown parameter α. The estimated response probability is $\hat{p}_i = p(z_i;\ \hat{\alpha})$ where $\hat{\alpha}$ is a consistent estimator of α.

In case of DRI with the imputation weight $\omega_i = d_i\ \frac{1-p_i}{p_i}$ (Haziza and Rao, 2006), the imputed estimator of the total $\hat{t}_{y,DRI}$ is approximately

- mpq-unbiased (with respect to the Imputation Model),
- pq-unbiased (with respect to the Non-response Model).

$\Rightarrow$ *double robustness*.

Estimation of the distribution function

We consider doubly robust inference for the distribution function $F_{y,N}(\cdot)$.

We consider the mean imputation model within classes, where the population U is divided into g mutually disjoint imputation cells $U_1, \ldots, U_G$ and

$$m: \; y_i \sim (\mu_g, \sigma_g^2), \qquad i \in U_g.$$

The corresponding random imputation mechanism is the hot-deck within classes, where for $i \in S_{mg}$

$$I: \; y_i^* = y_j \text{ for } j \in S_{rg} \quad \text{with} \quad \mathbb{P}(y_i^* = y_j) = \frac{\omega_j}{\sum_{k \in S_{rg}} \omega_k}.$$

Estimation of the distribution function

Theorem (BCH, 2013)

Assume that $\omega_i = d_i \frac{1-\hat{p}_i}{\hat{p}_i}$, *where* $\hat{p}_i = p(z_i;\ \hat{\alpha})$ *and* $\hat{\alpha}$ *is a consistent estimator of* α. *Then under mild assumptions :* $E|\hat{F}_I(t) - F_N(t)| \to 0$ *under the IM approach and under the NM approach.*

Estimation of the distribution function

Theorem (BCH, 2013)

Assume that $\omega_i = d_i \frac{1-\hat{p}_i}{\hat{p}_i}$, *where* $\hat{p}_i = p(z_i;\ \hat{\alpha})$ *and* $\hat{\alpha}$ *is a consistent estimator of* α. *Then under mild assumptions :* $E|\hat{F}_I(t) - F_N(t)| \to 0$ *under the IM approach and under the NM approach.*

Idea of the proof :
Under the NM approach, if p_i is assumed to be known ($\omega_i = d_i \frac{1-p_i}{p_i}$), the proof is roughly similar to the first one.

Estimation of the distribution function

Theorem (BCH, 2013)

Assume that $\omega_i = d_i \frac{1-\hat{p}_i}{\hat{p}_i}$, *where* $\hat{p}_i = p(z_i;\ \hat{\alpha})$ *and* $\hat{\alpha}$ *is a consistent estimator of* α. *Then under mild assumptions :* $E|\hat{F}_I(t) - F_N(t)| \to 0$ *under the IM approach and under the NM approach.*

Idea of the proof :
Under the NM approach, if p_i is assumed to be known ($\omega_i = d_i \frac{1-p_i}{p_i}$), the proof is roughly similar to the first one.
When p_i is unknown, we use a coupling procedure with :

$$y_i^{**} = y_j \text{ for } j \in S_{rg} \quad \text{with} \quad \mathbb{P}(y_i^{**} = y_j) = \frac{d_i \frac{1-p_i}{p_i}}{\sum_{k \in S_{rg}} d_i \frac{1-p_k}{p_k}},$$

$$y_i^{*} = y_j \text{ for } j \in S_{rg} \quad \text{with} \quad \mathbb{P}(y_i^{*} = y_j) = \frac{d_i \frac{1-\hat{p}_i}{\hat{p}_i}}{\sum_{k \in S_{rg}} d_i \frac{1-\hat{p}_k}{\hat{p}_k}},$$

$$|1(y_i^{*} \le t) - 1(y_i^{**} \le t)| \le \eta_g \quad \text{with} \quad E(\eta_g) \to 0.$$

Simulation study

Set-up

We consider a population of size $N = 10\ 000$, generated according to the model

$$y_i \quad = \quad 10 + x_{1i} + x_{2i} + \eta_i,$$

where the x_{1i}, x_{2i} are generated according to a Gamma distribution, and the η_i according to a standard normal distribution. We consider $R^2 = 0.70$.

The sample S of size $n = 500$ is selected by means of simple random sampling. The non-response mechanism is generated by means of Poisson sampling, with

$$Pr(r_i = 1|x_{1i}, x_{2i}) \quad = \quad \frac{\exp\left(-1 + 1.6\ x_{1i} + 1.6\ x_{2i}\right)}{1 + \exp\left(-1 + 1.6\ x_{1i} + 1.6\ x_{2i}\right)}.$$

The average response probability equals 0.60.

Hot-deck imputation

We repeat $B = 1,000$ times the sampling design + response mechanism + imputation mechanism. We are interested in estimating $F_N(t_\alpha)$, with $\alpha = 0.05,\ 0.25,\ 0.50,\ 0.75,\ 0.95$. The non-response mechanism is modeled :

- correctly : $\mathbf{x} = (1, x_1, x_2) \Rightarrow \hat{p}_1$,
- not correctly : $\mathbf{x} = (1, x_1) \Rightarrow \hat{p}_2$.

We consider three imputed estimators :

- unweighted hot-deck : $\omega_k = 1$,
- correct weighted hot-deck : $\omega_k = d_k \ \frac{1-\hat{p}_1}{\hat{p}_1}$,
- wrong weighted hot-deck : $\omega_k = d_k \ \frac{1-\hat{p}_2}{\hat{p}_2}$.

Results

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Table: Relative Bias and Monte Carlo Mean Square Error (in percent) for three hot-deck imputed estimators

Results

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Table: Relative Bias and Monte Carlo Mean Square Error (in percent) for three hot-deck imputed estimators

Results

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Table: Relative Bias and Monte Carlo Mean Square Error (in percent) for three hot-deck imputed estimators

References

- Chambers, R.L., Dunstan, R. (1986). Estimating distribution functions from survey data, *Biometrika*, **73**, 597–604.
- Chen, J., Rao, J. N. K., Sitter, R. R. (2000). Efficient random imputation for missing survey data in complex surveys. *Stat. Sinica*, **10**, 1153–1169.
- Deville, J-C. (2006). Random imputation using balanced sampling. Presentation to the Joint Statistical Meeting of the ASA, Seattle, USA.
- Deville, J-C. & Tillé, Y. (2004). Efficient balanced sampling : the Cube method. *Biometrika*, **91**, 893–912.
- Fuller, W.A. & Kim, J.K. (2005). Hot-deck imputation for the response model. *Survey Methodology*, **31**, 139–149.
- Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, **96**, 933–944.
- Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Volume 29, Sample Surveys : Theory Methods and Inference*, Editors : C.R. Rao and D. Pfeffermann, 215-246.
- Haziza, D., Rao, J.N.K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, **32**, 53–64.
- Isaki, C.T., Fuller, W.A. (1982). Survey design under a regression superpopulation model. *JASA*, **77**, 89–96.
- Kim, J.K. & Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika*, **91**, 559–578.
- Särndal, C. E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241–252.