Response set imbalance and non-response bias: a theoretical study with full use of auxiliary information

Kaur Lumiste

BaNoCoSS 2019

Introduction

In sample surveys we are interested in estimates of unknown parameters of a population, based on a **selected sample**

Introduction

In sample surveys we are interested in estimates of unknown parameters of a population, based on a **selected sample**, but often **non-response** occurs, the full sample cannot be collected.

Introduction

In sample surveys we are interested in estimates of unknown parameters of a population, based on a **selected sample**, but often **non-response** occurs, the full sample cannot be collected.

In practice troubles from non-response are treated in the estimation stage, usually with the aid of **auxiliary information**.

Introduction

In sample surveys we are interested in estimates of unknown parameters of a population, based on a **selected sample**, but often **non-response** occurs, the full sample cannot be collected.

In practice troubles from non-response are treated in the estimation stage, usually with the aid of **auxiliary information**.

Responsive (or adaptive) designs: Action should be taken during the data collection and with the aid of auxiliary information, the goal is to obtain in the end a **well balanced set of respondents**.

Introduction

In sample surveys we are interested in estimates of unknown parameters of a population, based on a **selected sample**, but often **non-response** occurs, the full sample cannot be collected.

In practice troubles from non-response are treated in the estimation stage, usually with the aid of **auxiliary information**.

Responsive (or adaptive) designs: Action should be taken during the data collection and with the aid of auxiliary information, the goal is to obtain in the end a **well balanced set of respondents**.

The crucial question: Will better balanced response guarantee better accuracy (lower variance and/or bias) in the estimates?

Notation

Let $U = (1, 2, ..., N)$ denote a finite **population**.

Notation

Let $U = (1, 2, ..., N)$ denote a finite **population**.

We take a random **sample** s of size n

Notation

Let $U = (1, 2, ..., N)$ denote a finite **population**.

We take a random **sample** s of size n to estimate the **population total** $Y = \sum_U y_k$ of the **study variable** y.

Notation

Let $U = (1, 2, ..., N)$ denote a finite **population**.

We take a random **sample** s of size n to estimate the **population total** $Y = \sum_U y_k$ of the **study variable** y.

The sampling design, which is used to select sample s, generates for each element $k \in U$ a known **inclusion probability** $\pi_k = Pr(k \in s)$

Notation

Let $U = (1, 2, ..., N)$ denote a finite **population**.

We take a random **sample** s of size n to estimate the **population total** $Y = \sum_U y_k$ of the **study variable** y.

The sampling design, which is used to select sample s, generates for each element $k \in U$ a known **inclusion probability** $\pi_k = Pr(k \in s)$

and a **design weight** $d_k = 1/\pi_k$.

Notation

Non-response occurs

Notation

Non-response occurs and values y_k are only recorded for a subset of units - **response set**, $r \subset s$.

Notation

Non-response occurs and values y_k are only recorded for a subset of units - **response set**, $r \subset s$.

It is assumed that we have access to **auxiliary variables** $\mathbf{x}_k = (x_{k1}, x_{k2}, ..., x_{kJ})'$ that are known $\forall k \in s$ and we know the population totals $\mathbf{X} = \sum_U \mathbf{x}_k$.

Notation

Non-response occurs and values y_k are only recorded for a subset of units - **response set**, $r \subset s$.

It is assumed that we have access to **auxiliary variables** $\mathbf{x}_k = (x_{k1}, x_{k2}, ..., x_{kJ})'$ that are known $\forall k \in s$ and we know the population totals $\mathbf{X} = \sum_U \mathbf{x}_k$.

We assume that the auxiliary vector can be constructed as such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1, \forall k \in s, \text{for some vector } \boldsymbol{\mu} \text{ independent on } k.$$

Balance and imbalance

The response set is **balanced** if

$$\bar{\mathbf{x}}_r = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k} = \bar{\mathbf{x}}_s.$$

Balance and imbalance

The response set is **balanced** if

$$\bar{\mathbf{x}}_r = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k} = \bar{\mathbf{x}}_s.$$

We measure **imbalance** with

$$IMB = P^2(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \Sigma_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s),$$

Balance and imbalance

The response set is **balanced** if

$$\bar{\mathbf{x}}_r = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k} = \bar{\mathbf{x}}_s.$$

We measure **imbalance** with

$$IMB = P^2 (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \Sigma_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s),$$

where

$$P = \sum_r d_k / \sum_s d_k, \quad \Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k.$$

Balance and imbalance

The response set is **balanced** if

$$\bar{\mathbf{x}}_r = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k} = \bar{\mathbf{x}}_s.$$

We measure **imbalance** with

$$IMB = P^2(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\Sigma_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s),$$

where

$$P = \sum_r d_k / \sum_s d_k, \quad \Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k.$$

IMB takes values between $0 \leq IMB \leq P(1-P)$.

Balance and imbalance

The response set is **balanced** if

$$\bar{\mathbf{x}}_r = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k} = \bar{\mathbf{x}}_s.$$

We measure **imbalance** with

$$IMB = P^2(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\Sigma_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s),$$

where

$$P = \sum_r d_k / \sum_s d_k, \quad \Sigma_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k.$$

IMB takes values between $0 \leq IMB \leq P(1-P)$.

Guiding data collection with *IMB* - **monitoring response**.

Estimation based on s

Horvitz-Thompson estimator (HT):

$$\hat{Y}_{FUL} = \sum_{s} d_k y_k = \hat{N}\bar{y}_s.$$

Estimation based on s

Horvitz-Thompson estimator (HT):

$$\hat{Y}_{FUL} = \sum_s d_k y_k = \hat{N}\bar{y}_s.$$

Calibration estimator

$$\hat{Y}^*_{CAL} = \sum_s d_k w_k y_k,$$

where $w_k = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$.

Estimation based on s

Horvitz-Thompson estimator (HT):

$$\hat{Y}_{FUL} = \sum_{s} d_k y_k = \hat{N}\bar{y}_s.$$

Calibration estimator

$$\hat{Y}^*_{CAL} = \sum_{s} d_k w_k y_k,$$

where $w_k = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$.

Weights w_k satisfy calibration requirements:

$$\sum_{s} d_k w_k \mathbf{x}_k' = \sum_{U} \mathbf{x}_k'.$$

Estimation under non-response

The expansion estimator:

$$\hat{Y}_{EXP} = \hat{N} \sum_r d_k y_k / \sum_r d_k = \hat{N} \bar{y}_r,$$

where $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$.

Estimation under non-response

The expansion estimator:

$$\hat{Y}_{EXP} = \hat{N} \sum_r d_k y_k / \sum_r d_k = \hat{N} \bar{y}_r,$$

where $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$.

The calibration estimator under non-response:

$$\hat{Y}_{CAL} = \sum_r d_k g_k y_k,$$

where

$$g_k = \left(\sum_s d_k \mathbf{x}_k \right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k.$$

Imbalance of study variable

The study variable imbalance is characterised by

$$\bar{y}_r - \bar{y}_s,$$

where $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ and $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$.

Imbalance of study variable

The study variable imbalance is characterised by

$$\bar{y}_r - \bar{y}_s,$$

where $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ and $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$.

If we multiple with $\hat{N}$ we get:

$$\hat{N}\left(\bar{y}_r - \bar{y}_s\right) = \hat{Y}_{EXP} - \hat{Y}_{FUL}.$$

Imbalance of study variable

The study variable imbalance is characterised by

$$\bar{y}_r - \bar{y}_s,$$

where $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ and $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$.

If we multiple with $\hat{N}$ we get:

$$\hat{N}\,(\bar{y}_r - \bar{y}_s) = \hat{Y}_{EXP} - \hat{Y}_{FUL}.$$

Let us expand the right side by $\pm\hat{Y}_{CAL}$:

$$\hat{N}\,(\bar{y}_r - \bar{y}_s) \;=\; \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right)$$

Non-response bias

$$\hat{N}\left(\bar{y}_r - \bar{y}_s\right) = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right)$$

Non-response bias

$$\begin{aligned}\hat{N}(\bar{y}_r - \bar{y}_s) &= \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right) \\ &= \hat{N}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \mathbf{b}_r + \hat{N}(\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s.\end{aligned}$$

where $\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k$ and
$\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$.

Non-response bias

$$\hat{N}\left(\bar{y}_r - \bar{y}_s\right) = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right)$$
$$= \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r + \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s.$$

where $\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k$ and $\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$.

This decompositions highlights two undesirable differences:

- Difference due to imbalance in the response

- Difference due to biased regression

Previous results

Let's denote

$$\Delta_r = (\mathbf{b}_r - \mathbf{b}_s)' \, \bar{\mathbf{x}}_s = \left(\hat{Y}_{CAL} - \hat{Y}_{FUL} \right) / \hat{N}$$

and investigate the effect of imbalance on Δ_r.

Previous results

Let's denote

$$\Delta_r = (\mathbf{b}_r - \mathbf{b}_s)' \, \bar{\mathbf{x}}_s = \left(\hat{Y}_{CAL} - \hat{Y}_{FUL} \right) / \hat{N}$$

and investigate the effect of imbalance on Δ_r.

Särndal et. al (2016) showed that under certain simplifying conditions, the conditional mean $E(\Delta_r | \bar{\mathbf{x}}_r, m, s) = 0$

Previous results

Let's denote

$$\Delta_r = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s = \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right) / \hat{N}$$

and investigate the effect of imbalance on Δ_r.

Särndal et. al (2016) showed that under certain simplifying conditions, the conditional mean $E(\Delta_r|\bar{\mathbf{x}}_r, m, s) = 0$ and the conditional variance

$$V\left(\Delta_r|\bar{\mathbf{x}}_r, m, s\right) \approx \frac{S_y^2}{m}\left(1 - p + \frac{IMB}{p^2}\right)$$

where m is the number of respondents, $p = m/n$ is the response rate, $S_y^2 = \sum_{j=1}^{J} n_j/nS_{yj}^2$ and $S_{yj}^2 = \sum_{s_j}(y_k - \bar{y}_{s_j})^2/(n_j - 1), j = 1, ..., J$.

Further exploration

For simplification let us redefine the calibration estimator under non-response:

$$\hat{Y}_{CAL2} = \sum_r d_k g_{Uk} y_k,$$

where $g_{Uk} = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$.

Further exploration

For simplification let us redefine the calibration estimator under non-response:

$$\hat{Y}_{CAL2} = \sum_r d_k g_{Uk} y_k,$$

where $g_{Uk} = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$.

Let us expand the $(\bar{y}_r - \bar{y}_s)$ decomposition with $\pm \hat{Y}^*_{CAL}$:

$$\hat{N}\left(\bar{y}_r - \bar{y}_s\right) = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL2}\right) + \left(\hat{Y}_{CAL2} - \hat{Y}^*_{CAL}\right) + \left(\hat{Y}^*_{CAL} - \hat{Y}_{FUL}\right)$$

Further exploration

For simplification let us redefine the calibration estimator under non-response:

$$\hat{Y}_{CAL2} = \sum_r d_k g_{Uk} y_k,$$

where $g_{Uk} = \left(\sum_U \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k$.

Let us expand the $(\bar{y}_r - \bar{y}_s)$ decomposition with $\pm \hat{Y}^*_{CAL}$:

$$\begin{aligned}
\hat{N}\left(\bar{y}_r - \bar{y}_s\right) &= \left(\hat{Y}_{EXP} - \hat{Y}_{CAL2}\right) + \left(\hat{Y}_{CAL2} - \hat{Y}^*_{CAL}\right) + \left(\hat{Y}^*_{CAL} - \hat{Y}_{FUL}\right) \\
&= \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_U\right)' \mathbf{b}_r + \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_U + \hat{N}\left(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_s\right)' \mathbf{b}_s,
\end{aligned}$$

where $\bar{\mathbf{x}}_U = \sum_U \mathbf{x}_k / \sum_s d_k$.

Further exploration

Let the auxiliary vector be a grouping vector, so that $\mathbf{x}_k = (0, \dots, 1, \dots, 0)'$, where the only 1 indicates the unique group (out of J possible) to which k belongs. Then

$$\hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_U\right)' \mathbf{b}_r = \hat{N} \sum_{j=1}^{J} \bar{y}_{rj} \left(\frac{m_j}{m} - \frac{N_j}{\hat{N}} \right),$$

Further exploration

Let the auxiliary vector be a grouping vector, so that $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, where the only 1 indicates the unique group (out of J possible) to which k belongs. Then

$$\hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_U\right)' \mathbf{b}_r = \hat{N} \sum_{j=1}^{J} \bar{y}_{rj} \left(\frac{m_j}{m} - \frac{N_j}{\hat{N}} \right),$$

$$\hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_U = \sum_{j=1}^{J} N_j \left(\bar{y}_{rj} - \bar{y}_{sj}\right),$$

Further exploration

Let the auxiliary vector be a grouping vector, so that $\mathbf{x}_k = (0, \dots, 1, \dots, 0)'$, where the only 1 indicates the unique group (out of J possible) to which k belongs. Then

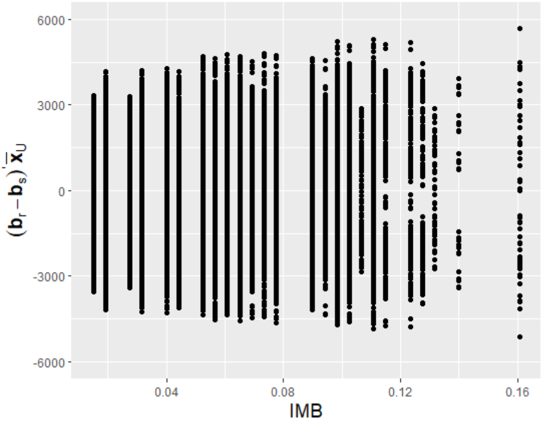
$$\hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_U\right)' \mathbf{b}_r = \hat{N}\sum_{j=1}^{J} \bar{y}_{rj}\left(\frac{m_j}{m} - \frac{N_j}{\hat{N}}\right),$$

$$\hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_U = \sum_{j=1}^{J} N_j\left(\bar{y}_{rj} - \bar{y}_{sj}\right),$$

$$\hat{N}\left(\bar{\mathbf{x}}_U - \bar{\mathbf{x}}_s\right)' \mathbf{b}_s = \hat{N}\sum_{j=1}^{J} \bar{y}_{sj}\left(\frac{N_j}{\hat{N}} - \frac{n_j}{n}\right).$$

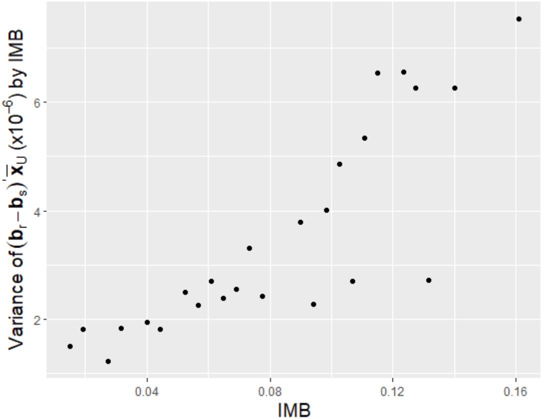
Simulations

A sample of $n = 20$ is fixed and all possible response sets are considered where $m = 12$. The auxiliary vector is a group vector, *IMB* and $(\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_U$ is calculated for 56 576 response sets.



Simulations

Variance of $(\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_U$ by *IMB* value.



References:

- Särndal, C.E., Lumiste, K., and Traat, I. (2016) Reducing the Response Imbalance: Is the Accuracy of the Survey Estimates Improved? *Survey Methodology*, 42 (2): 219–238.
- Lumiste, K. (2018) *Improving accuracy of survey estimators by using auxiliary information in data collection and estimation stages.* Dissertation, University of Tartu.