Measuring and Monitoring Balance of Response Set

Imbi Traat and Nora Roosileht
University of Tartu, Tartu, Estonia
Corresponding author: Imbi Traat, e-mail: imbi.traat@ut.ee

Abstract

Today, non-response is a common issue in each sample survey. The resulting response set is usually out of balance with respect to the full sample, and causes biased estimators. Even special adjustment methods cannot remove the entire bias. In Särndal (2011) new indicators are proposed to measure balance of the response set. In this paper we use these balance indicators to monitor sample selection process. We argue that response rate alone is not enough to express quality of the sample. Moreover, increasing response rate by the efforts to get data from arbitrary non-respondents may be even harmful to the balance of the response set. We describe methodology to identify those non-respondents who would bring highest increase towards balance. This knowledge could be used in the data collection process; directing more efforts for incorporating the most influential non-respondents. An example with three strategies is given.

Keywords: auxiliary information, balance indicator, monitoring data collection, non-response.

1 Introduction

Today, non-response is a common issue in each sample survey. The resulting response set is usually out of balance with respect to the full sample. Special adjustment methods need to be made in the estimation stage to reduce nonresponse bias. More and more auxiliary information becomes available for this purpose. Calibration estimators (Särndal, 2007) based on auxiliary information can be formed. However, some bias remains in the estimators. This calls for other actions. Steps can be made in data collection process.

Data collection extends over quite a large length of time. Usually it proceeds to an original plan, with compulsory repeated trials to contact units. However, the deadline approaches and the final response set remains considerably smaller than the desired sample. Furthermore, it usually differs from the full sample by its various characteristics. In this situation prior actions during data collection process can be made to achieve better balanced response set. Here, again, auxiliary information offers the possibilities.

Särndal (2011) uses an auxiliary vector with fixed composition of auxiliary variables, known also for nonrespondents, and constructs indicators that measure balance of the response set with respect to the full sample. These indicators can be used to monitor data collection process, and to take steps towards better balanced response set.

In this paper we study methodology to identify those non-respondents who would bring highest increase in balance indicator. In the later stages of data collection process the interviewers could put their efforts for incorporating those most influential non-respondents. We give an example where a balance indicator is monitored.

The effect to the indicator from incorporating new respondents is illustrated. As for comparison two other strategies are illustrated, the black and the neutral scenarios.

2 Balance indicators

Let s be the full sample extracted from the population by some sampling design with inclusion probabilities π_k for unit k. Let r denote the response set, $r \subset s$. Let $\mathbf{x}_k$ be the auxiliary vector with dimensionality J for unit k. The components of $\mathbf{x}_k$ may be both the numeric and the categorical variables, the latter broken down to binary variables for each category. One category of each categorical variable needs to be dropped for singularity reasons.

Balance of the response set is measured by the difference of the auxiliary variable means in the response set and in the full sample. The design-weighted means are used with weights $d_k = 1/\pi_k$:

$$\bar{\mathbf{x}}_{r;d} = \frac{\sum_r d_k \mathbf{x}_k}{\sum_r d_k}, \quad \bar{\mathbf{x}}_{s;d} = \frac{\sum_s d_k \mathbf{x}_k}{\sum_s d_k}.$$

Denote the difference $\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}$. Särndal (2011) introduces three indicators

$$\begin{aligned} BI_1 &= 1 - (Q-1)^{-1}\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}, \\ BI_2 &= 1 - 4P^2\mathbf{D}'\boldsymbol{\Sigma}_\mathbf{s}^{-1}\mathbf{D}, \\ BI_3 &= 1 - 2P(\mathbf{D}'\boldsymbol{\Sigma}_\mathbf{s}^{-1}\mathbf{D})^{\frac{1}{2}}, \end{aligned}$$

where $P = \sum_r d_k / \sum_s d_k$ is the response rate, $Q = 1/P$ and $\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k$. All indicators involve $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ which is called the lack of balance indicator. This term is the one that involves $\mathbf{x}_k$, therefore all the indicators change in the same direction when including new units into the response set (the factors $(Q-1)^{-1}$, P, P^2 are increasing together with response). All the indicators identify the same most influential element for the balance. By absolute value the indicators differ from each other for each realization of s, r and auxiliary vector $\mathbf{x}_k$:

$$0 \leq BI_1 \leq BI_2 \leq 1, \quad 0 \leq BI_3 \leq BI_2 \leq 1.$$

3 Monitoring data collection

We have a sample of size $n = 770$ drawn by simple random sampling from a population composed on real data. Auxiliary variables available for us are sex (1 male, 0 female), language (0 Estonian, 1 other), education (4 categories from basic to University degree), living place (4 categories from big city to countryside), age (numeric).

We have informative response with logit of non-normed response probabilities

$$logit(p_k) = -0.5\ sex + 0.05\ age^*,$$

where age^* is centered age. With this model for a middle-aged female the odds to respond are $e^0 = 1$, for a male they are $1/e^{-0.5} = 1.65$ times lower. The odds to respond increase with every ten years of age 1.65 times.

We have 50% response rate and size of the response set $m = 335$. Our normed response probabilities are $\theta_k = mp_k / \sum_s p_k$. With these probabilities we get our

response set through order sampling scheme by Rosén (1997). The response set is unbalanced with respect to age as can be seen from Table 1, and with respect to sex; proportion of males is 0.45 in sample and 0.37 in response set.

Table 1: Distribution of age

	Min	Median	Mean	Max
Sample	19.00	43.00	43.06	74.00
Response	19.00	49.00	47.23	74.00

It is clear that if response mechanism depends on some variables then Balance indicators can discover unbalance only if they use the same variables, or correlated with them variables. Auxiliary vector not depending on the response mechanism can not discover unbalance. This is illustrated in Table 2 where auxiliary vector **b** of dimensionality 9 includes one binary variable for sex, one for language, three for both education and living place, and one for age; whereas auxiliary vector **a** lacks sex and age.

Table 2: Balance indicators

Auxiliary vector	BI_1	BI_2	BI_3
a	0.991	0.991	0.904
b	0.912	0.912	0.703

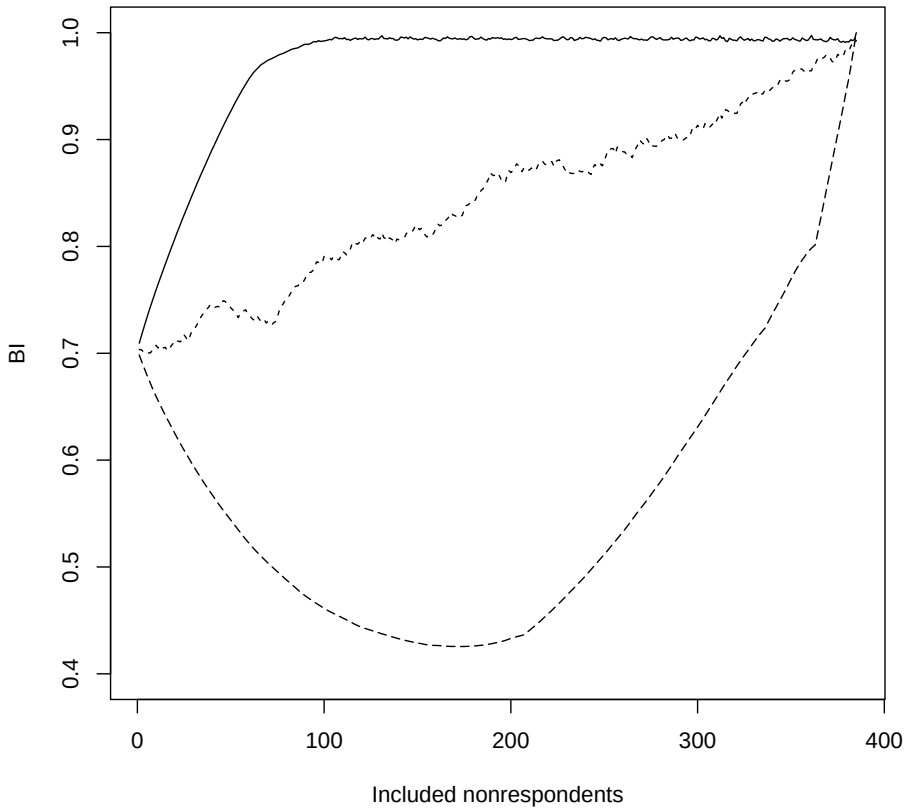
The indicators show balance for vector **a**, though we know that the response set is unbalanced with respect to sex and age.

Next we concentrate on the auxiliary vector **b** and on the balance indicator BI_3. We selected one by one a nonrespondent and found out its effect on BI_3 if this unit were included into response set. It turned out that a single unit can not be very influential. Comparing with the value 0.703 in Table 2, BI_3 varied between minimum 0.698 and maximum 0.707 with mean value 0.704. The indicator attained its maximum value for two nonrespondents, they were 21 year old men with all other values in the auxiliary vector **b** equal.

Our further calculations show the behavior of BI_3 in a repeated process. The three strategies were considered; the maximizing, the minimizing and the random strategy. Under each strategy a specific nonrespondent is identified, it is incorporated into response set and respective BI_3 is computed. Then a new search for the next specific unit is made. Under maximizing strategy the unit which maximizers balance indicator, and under minimizing strategy the unit which minimizers this indicator, is searched. Under random strategy the random nonrespondent is identified. The Figure 1 illustrates all these strategies.

We see that with maximizing strategy inclusion of only 100 nonrespondents among the 335 respondents gives fully balanced sample, $BI_3 \approx 1$. The response rate is 56% then. With random inclusion of 100 nonrespondents the indicator is still below 0.8. But with informative response that happens to minimize the balance indicator one may result with extremely unbalanced response set. One may get more and more respondents but the resulting set is more and more different from the target full sample.

Figure 1: Behavior of BI_3 when including nonrespondents into response set by three strategies



4 Conclusions

We studied a strategy in data collection process that increases balance of the final response set. At certain stage only those units were chosen for an interview that maximize the balance indicator. Even if the response rate remains low, these efforts guarantee the response set being close to the full sample. We also illustrated that increasing response rate by getting data from arbitrary non-respondents may be harmful to the balance of the response set.

References

Rosén, B, 1997. Asymptotic Theory for Order Sampling. *Journal of Statistical Planning and Inference.* 62: 135-158.

Särndal, C.-E. (2007), "The Calibration Approach in Survey Theory and Practice". *Survey Methodology*, 33, 99-119.

Särndal, C.-E. (2011) "The 2012 Morris Hansen Lecture. Dealing with survey non-response in data collection, in estimation," *Journal of Official Statistics*, 27(1), 1-21.