Lack of Balance Indicator for Data Collection

Maiken Mätik

University of Tartu, e-mail: maiken.matik@gmail.com

Abstract

The purpose of this paper is to study novel tools that measure balance of the response set against the full sample with respect to auxiliary variables. A measure called "lack of balance" is introduced. Its statistical properties are explored and an instrument called "balance indicator" is defined. Illustrative examples about the special cases of the balance indicator and related matters are given. A practical experiment was carried out on real data to illustrate theory about balance indicators. The experiment confirmed that balance indicator really shows balance under random or independent nonresponse and imbalance under dependent (on auxiliary variables) nonresponse.

Keywords: Auxiliary information, balance indicator, balanced response set, lack of balance

1 Introduction

The purpose of survey sampling is to give information about unknown parameters in the population $U = \{1, \dots, N\}$. Depending on the purpose and scope of the survey, special sampling design is used in U. With the design, inclusion probabilities, weights and other design characteristics are defined. For every object $k \in U$ we have positive inclusion probabilities $\pi_k = P(k \in s) > 0$ and for every object $k \in s$ we have a design weight $d_k = 1/\pi_k$.

Nowadays, nonresponse is a very common issue in survey sampling. There are always objects, from whom information is not received. We refer to the response set with symbol r. For example, many people with higher salary will not give their income data which leads to imbalanced response set with respect to the full sample. Survey estimates from respondents will then have nonresponse bias. Special efforts should be made already at the data collection stage to measure nonresponse effect, and possibly to reduce this effect. In this paper we introduce and study the tools given in Särndal (2011), and explored in Mätik (2012).

2 Response Rate and Response Probabilities

Lets assume we have the probability sample s with size n which means we have the objects with some auxiliary information that we gathered from the registers. But only a subset r with size m from s responds. Response rate is defined as

$$P = \frac{\sum_r d_k}{\sum_s d_k}. \tag{1}$$

We see that for equal d_k, $P = m/n$. The response indicator I is the binary random variable, observed for $k \in s$, with value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in (s - r)$.

Definition 2.1 The response probability for object $k \in s$ is defined through response indicator in a following way,

$$E(I_k|s) = P(I_k = 1|s) = \theta_k. \tag{2}$$

Response probabilities for all $k \in s$ are unknown parameters.

3 Measuring Lack of Balance

Assume we know a $J-$dimensional auxiliary variable vector $\mathbf{x}_k$ for each $k \in s$.

Definition 3.1 We call the response set r balanced when the means for appropriate auxiliary variables in r equal to corresponding means in the sample s.

We consider auxiliary vectors $\mathbf{x}_k$ which for some constant vector $\boldsymbol{\mu} \neq 0$, satisfy

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \qquad \text{for all} \quad k \in U. \tag{3}$$

We define two J-dimensional mean vectors and two computable $J \times J$ nonsingular weighting matrices:

$$\bar{\mathbf{x}}_{r;d} = \sum\nolimits_r d_k\mathbf{x}_k \Big/ \sum\nolimits_r d_k, \tag{4}$$

$$\boldsymbol{\Sigma}_r = \sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k^{'} \Big/ \sum\nolimits_r d_k, \tag{5}$$

$$\bar{\mathbf{x}}_{s;d} = \sum\nolimits_s d_k \mathbf{x}_k \Big/ \sum\nolimits_s d_k, \tag{6}$$

$$\boldsymbol{\Sigma}_s = \sum\nolimits_s d_k \mathbf{x}_k \mathbf{x}_k^{'} \Big/ \sum\nolimits_s d_k. \tag{7}$$

Auxiliary vectors that satisfy (3) also satisfy on all outcomes $(s{,}r)$:

$$\bar{\mathbf{x}}_{r;d}^{'} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{r;d}^{'} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}'_{r;d} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}_{s;d}^{'} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{s;d} = 1. \tag{8}$$

Definition 3.2 A measure

$$\boldsymbol{D}^{'} \boldsymbol{\Sigma}_s^{-1} \boldsymbol{D} = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})^{'} \boldsymbol{\Sigma}_s^{-1} (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}), \tag{9}$$

is defined as lack of balance indicator. It is a quadratic form in the differences in auxiliary variable means between the response set and the whole sample.

The lack of balance indicator refers to balance when the auxiliary variable means between the response set and the whole sample are equal, then $\mathbf{D} = 0$ and $\mathbf{D}^{'} \boldsymbol{\Sigma}_s^{-1} \mathbf{D} = 0$.
For one dimensional auxiliary vector $\mathbf{x}_k = \mathrm{x}_k$, the lack of balance indicator is

$$\mathbf{D}^{'} \boldsymbol{\Sigma}_s^{-1} \mathbf{D} = (\bar{\mathrm{x}}_{r;d} - \bar{\mathrm{x}}_{s;d})^2 \cdot \frac{\sum_s d_k}{\sum_s d_k \mathrm{x}_k^2}.$$

4 Estimated Response Probabilities

Looking for an estimator of θ_k, linearly depending on $\mathbf{x}_k$,

$$\hat{\theta}_k = \boldsymbol{\lambda}^{'} \mathbf{x}_k, \tag{10}$$

one gets,

$$\hat{\theta}_k = t_k = \left(\sum\nolimits_r d_k \mathbf{x}_k\right)^{'} \left(\sum\nolimits_s d_k \mathbf{x}_k \mathbf{x}_k^{'}\right)^{-1} \mathbf{x}_k. \tag{11}$$

The mean over r, and the mean and variance over s of the estimated response probabilities t_k are now related to the response rate P and lack of balance indicator in the following way:

$$\bar{t}_{r;d} = P \times \bar{\mathbf{x}}_{r;d}^{'} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{r;d}, \tag{12}$$

$$\bar{t}_{s;d} = P, \tag{13}$$

$$S_{t|s;d}^2 = \bar{t}_{s;d}(\bar{t}_{r;d} - \bar{t}_{s;d}) = P^2 \times \mathbf{D}^{'} \boldsymbol{\Sigma}_s^{-1} \mathbf{D}. \tag{14}$$

For constant response probability estimates, $\hat{\theta}_k = t_k = c$, the variance of the estimates is zero. Consequently, for $P \neq 0$ the lack of balance indicator is zero for $t_k = c$. Thus, for constant response probabilities the response set r is always balanced and represents the whole sample s.

We see that (13) and (14) now define the lack of balance indicator as the coefficient of variation of estimated response probabilities,

$$cv_{t|s;d} = \frac{S_{t|s;d}}{\bar{t}_{s;d}} = \frac{\sqrt{P^2 \times \mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D}}}{P} = (\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2}.$$

The upper bound of the lack of balance indicator is

$$\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq Q - 1,$$

where Q is inverse value of response rate P. We call $Q - 1$ nonresponse odds.

5 Balance Indicators

We consider three types of the balance indicators, all of them measured on the unit interval scale:

$$BI_1 = 1 - \frac{\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D}}{Q-1} = 1 - \frac{S^2_{t|s;d}}{P(1-P)}, \tag{15}$$

$$BI_2 = 1 - 4P^2\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D} = 1 - 4S^2_{t|s;d}, \tag{16}$$

$$BI_3 = 1 - 2P(\mathbf{D}^{'}\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2} = 1 - 2S_{t|s;d}. \tag{17}$$

For every outcome (s, r) and a fixed auxiliary vector $\mathbf{x}_k$ we have

$$0 \leq BI_1 \leq BI_2 \leq 1 \quad \text{ja} \quad 0 \leq BI_3 \leq BI_2 \leq 1.$$

These indexes show complete imbalance with the value 0, and complete balance with the value 1. It is important to remember, that balance/imbalance is measured with respect to chosen auxiliary vector.

6 Simulation Example

In this simulation example we used data about Estonian health care employees. There were 21761 objects in the register and 29 variables were measured for each individual. In our experiment we used one categorical variable, *education* (5 categories), and one continuous variable, *age*.

In the first part of the experiment we considered a response set that was independent form any of the variables. Both, the sample s (with size $n = 1000$) and the response set r (with size $m = 700$) were drawn with simple random sampling. Thus, the response rate was $P = 0.7$. The theoretical response probabilities were equal for all $k \in s$, so $\theta_k = m/n = 0.7$. We calculated the estimated response probabilities using three auxiliary vectors $\mathbf{x}_k$, extended stepwise. The results are shown in Table 1. The estimates t_k had very small variation around their mean 0.7 which equals theoretical θ_k. The calculated balance indicators approve theory that for independent from the variables response, the response set is balanced and represents the whole sample.

Table 1: Independent nonresponse

Auxiliary vector $\mathbf{x}_k$	Estimates t_k in sample s			
	mean	sd	BI_1	BI_2
One education category	0.7	0.0020	1.0000	1.0000
Four education categories	0.7	0.0103	0.9995	0.9996
Four education categories and age	0.7	0.0292	0.9959	0.9966

In the second part of the simulation exercise we drew a simple random sample s (with size $n = 1000$) but the response set r (with size $m = 700$) was generated as dependent on the variable *age*. Older people had bigger response probability. Thus our response set is imbalanced and the balance indicators should approve it. Again, we calculated the estimated response probabilities using three auxiliary vectors $\mathbf{x}_k$ built step by step. The results are shown in Table 2. The mean of t_k is still 0.7 but their variability is now bigger. For the first two $\mathbf{x}_k$ vectors, the indicators show balance because the response was not dependent on the variable *education*. For the third auxiliary vector, that includes *age*, the indicators approve that the response set is imbalanced.

Table 2: Dependent nonresponse

Auxiliary vector $\mathbf{x}_k$	Estimates t_k in sample s			
	mean	sd	BI_1	BI_2
One education category	0.7	0.0260	0.9968	0.9973
Four education categories	0.7	0.0272	0.9965	0.9970
Four education categories and age	0.7	0.1871	0.8333	0.8600

The experiment confirmed that balance indicators show balance under random or independent nonresponse. They show imbalance if the variables related to the response mechanism are included in $\mathbf{x}_k$.

References

Särndal, C.-E., 2011. The 2010 Morris Hansen Lecture. Dealing with Survey Nonresponse in Data Collection, in Estimation. *Journal of Official Statistics.* 27(1): 1-21.

Mätik, M., 2012. Dealing with Survey Nonresponse in Data Collection and in Estimation. Bachelor thesis (in Estonian). University of Tartu.