Lack of Balance Indicator for Data Collection

Workshop of Baltic-Nordic-Ukrainian Network on Survey Statistics

August 2012

Maiken Mätik

Nonresponse ➡ Biased estimates

Special efforts should be made at the data collection stage to measure nonresponse effect, and possibly to reduce this effect.

What helps and what doesn't?

Response rate P on its own doesn't carry enough information.

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

$$E(I_k|s) = P(I_k = 1|s) = \theta_k$$

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

We need to use registers to gather auxiliary information about our sample

Balanced Response Set

Definition 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 mark the difference with $\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}$.

We have balance, when $\bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{s;d}$,

where

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

Lack of Balance Indicator

Definition A measure

$$D'\Sigma_s^{-1}D = (\bar{\mathrm{x}}_{r;d} - \bar{\mathrm{x}}_{s;d})'\Sigma_s^{-1}(\bar{\mathrm{x}}_{r;d} - \bar{\mathrm{x}}_{s;d}),$$

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.

Lets mark here that a weighing matrix is calculated in a following way:

$$\Sigma_s = \sum\nolimits_s d_k \mathrm{x}_k \mathrm{x}_k' \Big/ \sum\nolimits_s d_k$$

When

$$\bar{\mathrm{X}}_{r;d} = \bar{\mathrm{X}}_{s;d} \quad \Rightarrow \quad \mathrm{D} = 0 \quad \Rightarrow \quad \mathrm{D}'\Sigma_s^{-1}\mathrm{D} = 0$$

Overall, $\mathrm{D}'\Sigma_s^{-1}\mathrm{D}$ is defined on unit interval scale.

Relation with Estimated Response Probabilities

We can estimate the response probabilities in the following way:

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

It can now be shown that the mean over *r* and the mean and variance over *s* of the estimated response probabilities t_k are related to response rate *P* and lack of balance indicator in the following way:

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

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

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

Balance Indicators

We have now reached the stage when the balance indicators can be defined. It can be seen that while they are functions of lack of balance indicators, they are also connected to the estimated response probabilities t_k through their variance.

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

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

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

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

Simulation Example

- Data about Estonian health care employees
  - N=21764
  - 29 variables measured for each individual.
    - In this example one cathegorical variable, *education*, and one continuous variable, *age*, were used.
    - Simple random sample *s* with size n=1000 was drawn
    - Response set *r* was with size m=700.
- Experiment was carried out in two parts:
  - Response set independent of the variables,
  - Response set dependent on the variables.

Part 1

- The response set was also drawn with simple random sample.
- Thus the theoretical response probabilities were equal for all $k \in r, \theta_k = 0.7$.
- Auxiliary vectors were built with stepwise forward selection, by adding *education* cathegories and finally variable *age.*

Table 1: Independent nonresponse

Auxiliary vector x_k	Estimates t_k in sample s		BI_1	BI_2
	mean	sd		
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

Part 2

- The response set was generated as dependent on variable *age.*
- Auxiliary vectors were again built with stepwise forward selection, by adding *education* cathegories and finally variable *age.*
- Results show certain imbalance in the response set with respect to chosen auxiliary vector.

Table 2: Dependent nonresponse

Auxiliary vector $\mathbf{x}_k$	Estimates t_k in sample s		BI_1	BI_2
	mean	sd		
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

- **Result** Calculated balance indicators approve with theory.

Thank you for your attention!