Probabilistic Approach to Editing

Maiki Ilves*

Abstract

Editing all inconsistent data records is time consuming and costly. To save resources, alternative editing methods are sought by survey practitioners. In this paper, an editing procedure where the responses are selected for editing through Poisson sampling according to their impact to final estimates is proposed. Probabilistic approach gives simple tools known from sampling theory to describe the effect of editing on the survey estimates. A two-phase design approach is applied for bias estimation, and a bias corrected generalized regression (GREG) estimator and an estimator of its variance are presented. The effectiveness of the proposed editing procedure is illustrated using empirical data from Statistics Sweden.

Key Words: measurement error, two-phase sampling design, bias estimation

1. Introduction

Measurement errors can occur during data collection as well as during data processing. During the data collection phase, measurement errors can be caused by respondents, interviewers or measurement instruments. There is lot of literature available how to prevent mentioned errors happening by using well designed questionnaires and well trained interviewers (e.g. Biemer et al. (2004) and Lyberg et al. (1997)). In addition, coding errors, programming errors, scanning errors, and other errors during data handling can also cause measurement errors. Measurement errors are part of nonsampling errors, as are nonresponse errors and frame errors.

Regardless how good prevention methods are in place, data editing needs to be part of the survey process to check for measurement errors and so assure the quality of data. In addition to enabling to correct erroneous entries, editing also helps to find reasons way errors occurred and this way improve the measurement process.

It is known that measurement errors when not dealt with increase mean square error (MSE)(Biemer and Lyberg (2003)). Measurement errors can be systematic, thus increasing the bias part, and random, thus adding to the variance part of the MSE.

Different editing procedures can be applied on micro and macro levels. Usually editing means checking for inconsistencies between variables, doing logical checks, outlier detection, comparisons with historical data etc. Because errors can occur for so many different reasons tracking down the errors takes lot of human resources, is time consuming and costly.

Often editing is considered only as part of data processing and not part of estimation. However, the last two decades the effect of editing to survey estimates is getting more attention. Lawrence and McDavitt (1994) and Lawrence and McKenzie (2000) describe non-probabilistic editing approach called significance editing, also called selective editing, where main focus is how to select only the most influential errors in terms of influence to the final estimates.

*Swedish Business School at Örebro University, 701 82 Örebro, Sweden

This paper introduces a probabilistic editing procedure which in some sense is very similar to selective editing but different in nature. Combination of selective editing and probabilistic editing was introduced in Ilves and Laitila (2009). In the paper, purely probabilistic editing approach is discussed. Probabilistic approach to editing gives known tools from design-based sampling theory for evaluating the influence of measurement errors on survey estimates.

The paper consists of three parts. Probabilistic approach to editing, together with bias corrected estimator and its variance, is introduced in section 2. Section 3 describes an empirical study where the performance of proposed editing procedure is evaluated. The paper ends with a discussion and planned future work.

2. Probabilistic editing

Editing procedure covered in this paper can be incorporated to the everyday survey process or it can be used as evaluation tool for estimating the measurement bias. For a probabilistic editing procedure two measured values are assumed to be available or possible to obtain for small subsample of units and it is assumed that the second measurement contains the true value.

The editing process is interpreted as a two-phase sampling procedure in which the original sample is obtained in the first phase and the observations for editing are probability selected in the second phase.

Let us consider a population, $U = \{1, 2, \ldots, N\}$, from which sample s_a of size n_a is drawn according to sampling design $p_a(\cdot)$. Let us denote true values by z_k and observed values by y_k. This section aims to derive an unbiased estimator of the population total of variable z, i.e., $t_z = \sum_{k=1}^{N} z_k$, in the case of measurement error in the observed sample units. In this paper, full response is assumed everywhere i.e. measurement errors are the only nonsampling errors occurring here.

The generalized regression estimator (GREG) is considered, thus assuming that relevant auxiliary information is available for the initial sample s_a. The GREG estimator for t_y is a weighted sum of variable y:

$$\hat{t}_{yG} = \sum_{k \in s_a} w_k y_k, \tag{1}$$

where

$$\begin{aligned} w_k &= g_k \frac{1}{\pi_{ak}}, \\ g_k &= 1 + \left(\sum_{k \in U} x_k - \sum_{k \in s_a} \frac{x_k}{\pi_{ak}} \right)^T \left(\frac{\sum_{k \in s_a} x_k x_k^T}{\pi_{ak}} \right)^{-1} x_k, \end{aligned} \tag{2}$$

and π_{ak} being the first order inclusion probability for the initial sample.

GREG is a model-assisted estimator where model describes the relationship between the variable y and auxiliary information x. More information about the model assumed in GREG and its influence to the properties of GREG can be found in Särndal et al. (1992).

GREG is a nearly unbiased estimator of the total but due to the measurement errors in the data, (1) is a biased estimator of t_z. Using the notation introduced earlier, the bias of the total estimate can be expressed as:

$$B(\hat{t}_{yG}) = E(\hat{t}_{yG}) - t_z = \sum_{k \in U} q_k, \tag{3}$$

where $q_k = g_k(y_k - z_k)$. The bias in (3) is unknown because the difference between observed value and the true value is not known for all population units.

In order to estimate the size of the bias a subsample s_2 of size n_2 is drawn from s_a and all the units in the subsample s_2 are edited. The second measurement is denoted by $\tilde{y}_k$, $k \in s_2$ and is assumed to be the true value.

The Sampling design used for selecting the units in the second phase is a Poisson design. The use of Poisson design in the second phase is advantageous in many ways. Poisson sampling allows units to be sampled simultaneously with data collection, and different inclusion probabilities can be assigned to the units to reflect the likelihood and influence of errors. In addition, the independent sampling of units simplifies the derivation of variance formulae.

Remark 1: In order to carry out real time sampling, one needs to somehow estimate the total influence of errors i.e. total of global scores (see 2.2). In case of repeated survey, historical data can be used for estimating total influence.

Remark 2: When no information is available for helping to distinguish between erroneous records and correct records, Bernoulli design can be used as the sampling design in the second phase. Bernoulli design is a special case of Poisson design where all units having equal inclusion probabilities.

The bias (3) can be estimated by:

$$\hat{t}_q = \sum_{k \in s_2} \frac{q_k}{\pi_{ak}\pi_{k|s_a}}, \tag{4}$$

where $q_k = g_k(y_k - \tilde{y}_k)$ for $k \in s_2$, and $\pi_{k|s_a}$ denotes the first-order inclusion probability in the second phase. Estimator (4) is also called the $\pi^\star$-estimator (Särndal et al. (1992)).

An unbiased estimator of t_z is now obtained by subtracting the estimated bias from the biased total estimate:

$$\hat{t}_z = \hat{t}_{yG} - \hat{t}_q = \sum_{k \in s_a} \frac{g_k y_k}{\pi_{ak}} - \sum_{k \in s_2} \frac{g_k(y_k - \tilde{y}_k)}{\pi_{ak}\pi_{k|s_a}} \tag{5}$$

where g_k is given by (2).

The variance of (5) is approximately

$$var(\hat{t}_z) = var(\hat{t}_{yG}) + var(\hat{t}_q) - 2cov(\hat{t}_{yG}, \hat{t}_q) \tag{6}$$

where

$$\begin{aligned}
var(\hat{t}_{yG}) &= \sum_{k,l \in U}\sum \Delta_{akl} \frac{y_k - x_k^T B}{\pi_{ak}} \frac{y_l - x_l^T B}{\pi_{al}}, \\
var(\hat{t}_q) &= \sum_{k,l \in U}\sum \Delta_{akl} \frac{q_k}{\pi_{ak}} \frac{q_l}{\pi_{al}} + E_a \left[\sum_{k \in U} \pi_{k|s_a}(1 - \pi_{k|s_a}) \left(\frac{I_k q_k}{\pi_{ak}\pi_{k|s_a}} \right)^2 \right], \\
cov(\hat{t}_{yG}, \hat{t}_q) &= \sum_{k,l \in U}\sum \Delta_{akl} \frac{g_k y_k}{\pi_{ak}} \frac{q_l}{\pi_{al}},
\end{aligned}$$

$B = (\sum_{k \in U} x_k x_k^T)^{-1}(\sum_{k \in U} x_k y_k)$ is a vector of regression coefficients, $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$ is a covariance between sampling indicators, and π_{akl} is a second order inclusion probability for the first phase sample units.

An unbiased estimator of (6) is

$$\hat{var}(\hat{t}_z) = \hat{var}(\hat{t}_{yG}) + \hat{var}(\hat{t}_q) - 2\hat{cov}(\hat{t}_{yG}, \hat{t}_q). \tag{7}$$

where

$$\begin{aligned} \hat{var}(\hat{t}_{yG}) &= \sum_{k,l\in s_a}\sum \frac{\Delta_{akl}}{\pi_{akl}}\frac{y_k - x_k^T\hat{B}}{\pi_{ak}}\frac{y_l - x_l^T\hat{B}}{\pi_{al}}, \\ \hat{var}(\hat{t}_q) &= \sum_{k,l\in s_2}\sum \frac{\Delta_{akl}}{\pi_{akl}\pi_{kl|s_a}}\frac{q_k}{\pi_{ak}}\frac{q_l}{\pi_{al}} + \sum_{k\in s_2}(1-\pi_{k|s_a})\left(\frac{q_k}{\pi_{ak}\pi_{k|s_a}}\right)^2, \\ \hat{cov}(\hat{t}_{yG},\hat{t}_q) &= \sum_{k\in s_a}\sum_{l\in s_2}\frac{\Delta_{akl}}{\pi_{akl}}\frac{g_k y_k}{\pi_{ak}}\frac{q_l}{\pi_{al}\pi_{l|s_a}}, \end{aligned}$$

and

$$\hat{B} = \left(\sum_{k\in s}\frac{x_k x_k^T}{\pi_k}\right)^{-1}\left(\sum_{k\in s}\frac{x_k y_k}{\pi_k}\right) \tag{8}$$

is a vector of estimated regression coefficients.

Each term in (7) is an unbiased estimate of the corresponding term in (6).

2.1 Stratified Simple Random Sampling

As an example, let's consider stratified simple random sampling as a first phase sampling design. Then, the population and the sample are partitioned into H non-overlapping subgroups, $U = U_1 \cup \ldots \cup U_H$ and $s_a = s_{a1} \cup \ldots \cup s_{aH}$, respectively, and the first-order inclusion probabilities and covariance of sampling indicators are:

$$\begin{aligned} \pi_{ak} &= \frac{n_{ah}}{N_{ah}} = f_{ah}, \\ \Delta_{akl} &= -f_{ah}\frac{1-f_{ah}}{N_{ah}-1},\ k \neq l \\ \Delta_{akk} &= f_{ah}(1-f_{ah}),\ k = l. \end{aligned}$$

The unbiased estimator of the total is:

$$\hat{t}_z = \sum_{h=1}^{H}\frac{N_{ah}}{n_{ah}}\left[\sum_{k\in s_{ah}} g_k y_k - \sum_{k\in s_{2h}}\frac{q_k}{\pi_{k|s_a}}\right]. \tag{9}$$

The unbiased estimator of variance is given by (7), where

$$\hat{var}(\hat{t}_y) = \frac{(1-f_{ah})N_{ah}^2}{n_{ah}} S^2_{es_{ah}},$$

with $S^2_{es_{ah}} = \sum_{k\in s_{ah}}(e_k - \sum_{k\in s_{ah}} e_k/n_{ah})^2/(n_{ah}-1)$, $e_k = y_k - x_k^T\hat{B}$, and $\hat{B}$ is given by (8),

$$\hat{var}(\hat{t}_q) = \sum_{h=1}^{H}\frac{(1-f_{ah})N_{ah}^2}{n_{ah}(n_{ah}-1)}\left[\sum_{k\in s_{2h}}\breve{q}_k^2 - \frac{(\sum_{k\in s_{2h}}\breve{q}_k)^2}{n_{ah}}\right] + \sum_{h=1}^{H}\frac{1}{f_{ah}}\sum_{k\in s_{2h}}(1-\pi_{k|s_a})\breve{q}_k^2,$$

and

$$\hat{cov}(\hat{t}_{yG},\hat{t}_q) = \sum_{h=1}^{H}\frac{(1-f_{ah})N_{ah}^2}{n_{ah}^2}\left[\sum_{k\in s_{2h}} g_k y_k \breve{q}_k - \frac{1}{n_{ah}-1}\sum_{k\in s_{ah}} g_k y_k \sum_{l\in s_{2h}}\breve{q}_l\right],$$

where $\breve{q}_k = \frac{q_k}{\pi_{k|s_a}}$.

To increase the effectiveness of the editing process one should include available information (e.g. from a previous survey) to the selection process. This can be done by constructing score function which summarizes the existing information for each observation and carrying out selection with inclusion probabilities proportional to score function. The score function in the probabilistic editing serves the same purpose as the score function in the selective editing: it should distinguish between possible errors and correct values.

In practice, several variables are measured in a survey, and score values, referred to as local scores, are calculated for each variable obtained from a sampled unit. Instead of editing single variables separately, global score functions are constructed from the local scores. Then, selection is based on the global scores and, if observation is selected for editing, all or a subset of variables are edited simultaneously. Thus, using global scores the effect is not based on the influence of a specific variable on a single total estimate.

In Latouche and Berthelot (1992) different ways of constructing global scores for selective editing procedure are suggested. The global score considered in this paper has following general form:

$$gscore_k = \sum_{i=1}^{I} \frac{d_k |y_{ik} - \hat{y}_{ik}| z_k v_i}{\hat{t}_{y_i}} \tag{10}$$

where I denotes total number of variables, d_k is design weight, y_i is the value of unedited variable i, $\hat{y}_i$ is an estimate of y_i based on available information, z_k is indicator variable indicating whether record k was flagged for the editing or not and v_i denotes the importance of variable i.

Score function (10) is the best suited for quantitative variables, but can be used also for dichotomous variables. Unordered categorical variables can be handled by creating set of new dichotomous variables (one for each category).

In case no additional information is available to predict the erroneous observations, equal inclusion probabilities (i.e Bernoulli design) can be used in the second phase.

3. Empirical Study

An empirical study is performed to examine the effectiveness of the described probabilistic editing approach on a specific data. Real data from the short-term employment survey carried out by Statistics Sweden are used in the simulation. Short-term employment survey is a quarterly survey with rotating sampling units. The sampling design employed is stratified simple random sampling. The main variable collected during the survey is the total number of employees in the local unit of the enterprise. The data used here is from the second quarter of 2008 and 2009. Data from 2008 is used only for constructing global scores. In addition, business register information about the size of activity group and total number of employees per activity group is used as auxiliary information in the regression estimator.

Dataset contains 22 448 local units (out of 298 728 units) and unedited and edited values for the variable total number of short-term employees were recorded. In total, 2.2% of units changed the value during editing process which amounted to 4.2% of change in the estimate of total number of short-term employees. It is assumed that all measurement errors were found and corrected.

For empirical study following set-up is used

1. Data from the second quarter of 2009 is considered as first phase sample s_a. Unedited values are used to compute (1).

2. From s_a subsample of expected size n_2 according to Poisson design is drawn. Three different values of n_2 were considered: $0.05 * n_a$, $0.1 * n_a$, and $0.15 * n_a$, where $n_a = 22448$ is the size of the first phase sample. Two different sets of inclusion probabilities were used: equal inclusion probabilities which corresponds to the Bernoulli design, and inclusion probabilities proportional to the global score (10) which corresponds to the Poisson design. In computation of global scores $I = 2$, y_1 and y_2 denote the total number of long-term and short term employees, respectively, and $\hat{y}_1$ and $\hat{y}_2$ are the average number of long-term and short-term employees, respectively, in the stratum based on second quarter 2008 data, and $z_k = v_i = 1$.

Step 2 was repeated 10 000 times and for each repetition the bias corrected estimate (9) was computed. Two estimates were of interest: the total estimate, $\hat{t}_z$, and one domain estimate, $\hat{t}_{zd}$, where domain being a middle sized county in Sweden (Örebro län). Table 1 gives the average number of records that changed the value and the empirical relative bias (RB) after probabilistic editing for different settings.

Table 1: Empirical relative bias (RB) and average number of records corrected under different second phase inclusion probabilities and second phase sample size for the estimate of population total and domain total.

Estimate		Bernoulli design			Poisson design		
	n_2/n_a	5%	10%	15%	5%	10%	15%
$\hat{t}_z$	Records corrected	24	48	73	52	95	131
	RB (%)	0.04	0.03	0.03	0.00	0.00	0.00
$\hat{t}_{zd}$	Records corrected	0.8	1.7	2.5	1.2	2.4	3.4
	RB (%)	0.00	0.00	0.00	0.00	0.00	0.00

The relative difference between the total estimate and the true value before editing was 4.2% and, as seen from Table 1, unbiased estimate is obtained regardless the design used for sampling records for editing. The relative difference between the domain estimate and the true domain value before editing was 0.4% mainly due to one big error and after editing the unbiased estimate is again obtained.

The coefficient of variation, in this example, is quite large. For the total estimate the coefficient of variation is 28% under Poisson design and 30% under Bernoulli design for the smallest sample size observed in the simulation study.

4. Discussion

Probabilistic editing enables to correct some errors and estimate the influence of errors not corrected during the editing and still get unbiased estimates. In addition, when no information is available for distinguishing between erroneous and correct

records, using Bernoulli sampling for selecting records for editing is good alternative to Poisson sampling, as one could see from simulation study results.

The high variance in the study is caused by outliers in the study variable and by skewly distributed measurement errors with lot of zeros and heavy tails due to few big errors. Outliers in the study variable increase the variability of model fit residuals in the GREG variance estimator making it rather sensitive to the outliers. In addition, the distribution of measurement errors together with choice of score function are important in determining the variance of bias estimator. One needs to investigate possibilities which enable to reduce the both kind of variation. One possibility is to apply probabilistic editing after obvious errors, i.e. outliers, are taken care of.

As seen from Table 1, when using score function in inclusion probabilities, more erroneous records are selected compared to when score function is not used. This assures us that used score function works well for given study variable. However, more work needs to be done about the choice of score function and its influence to the properties of the estimator.

All in all, the probabilistic editing can be used as an alternative to selective editing. It retains all desirable properties of selective editing by saving time and resources. Probabilistic editing does complicate the estimation, but it is compensated by enabling to evaluate the size of bias for any variable of interest in specific dataset.

Acknowledgements

I would like to thank Anders Eklund from Statistics Sweden for his valuable help regarding short-term employment survey and putting together dataset suitable for my needs.

References

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., & Sudman, S. (Eds.). (2004). *Measurement errors in surveys*. New York: Wiley.

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality*. New York: Wiley.

Ilves, M., & Laitila, T. (2009). Probability-sampling approach to editing. *Austrian Journal of Statistics*, *38*, 171-183.

Latouche, M., & Berthelot, J.-M. (1992). Use of score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, *8*, 389-400.

Lawrence, D., & McDavitt, C. (1994). Significance editing in the australian survey of average weekly earnings. *Journal of Official Statistics*, *10*, 437-447.

Lawrence, D., & McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, *16*, 243-253.

Lyberg, L., et al. (Eds.). (1997). *Survey measurement and process quality*. New York: Wiley.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.