WP. 36
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (vi): New and emerging methods

PROBABILITY EDITING

Prepared by Thomas Laitila[1] and Maiki Ilves[2]
[1)]Statistics Sweden and Örebro university, [2)]Statistics Estonia

Preliminary version 2012-08-24

I. Introduction

1. A major activity in a survey is the editing process. Developments of new theories and methods useful for reducing the resources spent on editing are of interest as it may provide with substantial cost savings and increase in timeliness. One alley for finding potential improvements is development of more efficient tools for identification of erroneous observations. Another is to reduce the number of observations edited. Indeed the traditional approach to edit all observations is not generally necessary for appropriate statistical inference.

2. Conditioning on the response set of observations, traditional sample survey methodology can be applied for inference on errors in the response set. Relaxing the conditioning, results can be generalized to population and domain estimates. This probability based approach to editing (probability editing) is suggested by Ilves and Laitila (2009) and the theory is further developed by Ilves (2011, 2012).

3. Empirical applications of the probability editing approach indicate good performance. Bias due to measurement errors is eliminated and the variances of estimates are on acceptable levels while only a fraction of the data set is edited.

4. This paper includes i) a presentation of the probability editing approach, ii) a summary of results presented in Ilves and Laitila (2009) and Ilves (2012a,2012b), iii) a comparison of probability and selective editing, and, iv) a discussion of implementing probability editing in a statistics production environment.

II. The probability editing approach

5. A simple illustration of the idea of probability editing is to consider a sample survey estimator $\hat{t}_y$ of a population total, t_z, calculated upon a response set *r*. Due to measurement errors in the observations the estimator is potentially biased, i.e. $E(\hat{t}_y) = t_z + b_y$. For a correction of measurement error bias, a subsample of units, $r_2 \subset r$, is selected from the response set using a probability sampling design. For the units in the subsample, true values of the study variable

are obtained, and an estimator $\hat{b}_y$ of the bias of $\hat{t}_y$ can be defined. A bias corrected estimator of the population total t_z is then obtained by the estimator $\hat{t}_z = \hat{t}_y - \hat{b}_y$.

6. A major advantage of selecting units for editing using a probability sampling design is the ability to address the statistical properties of the resulting estimator. The estimator of the bias $\hat{b}_y$ can be made unbiased, or approximately so, by appropriate choice of sampling design and estimator. An unbiased estimator of the measurement error bias gives, with respect to measurement errors, an unbiased estimator of the population total.

7. For statistical inference on the population total an estimate of the estimator variance is needed along with the point estimate. The variance of the bias corrected estimator $\hat{t}_z$ equals

$$V(\hat{t}_z) = V(\hat{t}_y) + V(\hat{b}_y) - 2Cov(\hat{t}_y, \hat{b}_y) \qquad (1)$$

An estimator of the first term on the l.h.s, $V(\hat{t}_y)$, is given by the original estimator used, e.g the Horvitz-Thompson (HT) or the generalized regression (GREG) estimators. For the second term, $V(\hat{t}_y)$, an estimator can be derived by considering the set r_2 as resulting from a two-phase sampling design. Finally, an estimator of the covariance term has to be derived considering the two estimators involved and the choice of sampling design for selection of units into r_2.

8. The probability sampling design chosen for selecting units for editing, i.e. units selected into r_2, is a major concern. In most standard sampling designs, fixed size sampling designs requires access to the whole response set *r* before sampling of units into r_2, reducing timeliness of statistics. An alternative is to consider sampling designs with independent selection of units, e.g. Bernoulli or Poisson sampling. Sampling of units for editing can then be integrated with the data collection process, adding minimum to survey time.

9. Independent sampling designs also have the benefit of reducing the complexity of the variance and covariance terms of the bias corrected estimator, yielding less complex variance estimators.

10. The examples of Poisson and Bernoulli sampling indicate a general applicability of probability editing and an option of designing editing in relation to the survey conducted. In cases where information is available on potential size or importance of a measurement error, Poisson sampling can be utilized for selecting some units with larger probabilities than others. If there is no such information available, Bernoulli sampling can be used. Thus, probability editing can be undertaken in unique surveys with no prior information. It can also be designed to utilize available prior information in repeated surveys.

11. Also note that Bernoulli sampling is a special case of Poisson sampling where all units have the same inclusion probability. This implies that theory on probability editing based or not based on prior information can be integrated into a single theory. The theory will also encompass use of other independent, unequal probability sampling designs. An example is the PoMix design, where the Bernoulli design and the Poisson design is combined.

Table 1: Relative bias and RMSE ratio of the HT estimator after selective editing (SE) and editing using a two-step procedure (TP), respectively. Results generated from a simulation study with true values generated from a Po(5) distribution and erroneous observations generated from either a Po(2) or a Po(10) distribution. The rate of observations with measurement errors = 0.4. Case 1 = erroneous observations are taken from Po(2) or Po(10) with equal probability (0.2). Case 2 = all erroneous observations are generated from the Po(2) distribution. Table reproduced from Ilves and Laitila (2009).

		Correlation between score values and observations			
		1.0	0.7	0.5	0.3
Case 1					
SE	Edited	12%	16%	14%	17%
	Rel. Bias	2%	2%	4%	6%
TP	Edited	4% + 8%	4%+12%	6%+8%	7%+10%
	Rel. Bias	0%	0%	0%	0%
	RMSE ratio (TP/SE)	2.03	1.88	1.36	1.02
Case 2					
SE	Edited	14%	14%	15%	17%
	Rel. Bias	-18%	-21%	-20%	-20
TP	Edited	4% + 10%	4%+10%	4%+11%	5%+12%
	Rel. Bias	0%	0%	0%	0%
	RMSE ratio (TP/SE)	0.26	0.23	0.23	0.22

III. Results on Probability editing

12. The first suggestion of using probability editing was made by Ilves and Laitila (2009) who considered probability editing performed on units not edited after selective editing, an editing process they named a “two-step procedure”. This two-step procedure was developed for the HT estimator under general sampling designs (design for selecting the sample and design for selecting edited units) and with full sample response. As a special case they also considered simple random sampling (SRS) of the sample combined with Poisson sampling for selection of units for editing.

13. In a simulation study, Ilves and Laitila (2009) studied the properties of the two-step procedure in relation to pure selective editing and a pure probability editing. The simulation setup considered SRS combined with Bernoulli sampling of units for editing. The population consisted of observations generated from a Poisson distribution and measurement errors were generated by replacing true values with random draws from alternative Poisson distributions. Score values used in selective editing was generated as deviations from the mean of the “true”

Poisson distribution with a random number added. This allowed for a control of the correlation between the observed value of the study variable and the score value.

14. The general results from the simulation study show that probability editing works either on its own or in combination with selective editing in a two step procedure. The relative performance of probability editing is especially good when the measurement errors in general yield larger (smaller) observations. This is illustrated in Table 1, where parts of the results presented in Ilves and Laitila (2009) are reproduced.

15. For Case 1 in Table 1 it is observed that the bias of the estimator under selective editing is increasing as the correlation between the score value, upon which selection of units for editing is made, and the observed value of the study variable decreases. This is an expected result as when the correlation decreases the score value is less useful for detecting large deviating observations. The result has implications for the use of global scores, where the score used for selection of edited units are based on a summary statistic of a set of study variables.

16. In Case 2 in Table 2, there is no pattern of an increasing relative bias with a decreasing correlation between the score value and the observation. This is unexpected. A possible explanation is that the when the correlation is low, the selection of units for editing in selective editing mimics a random selection of units, i.e. mimics probability editing.

17. As is evident from the outline of the idea of probability editing, the proposal for the editing procedure takes the problem of estimation in sample surveys as a starting point. By using probability editing it is possible to integrate editing within the theory on estimation in sample surveys. The major questions to be addressed are how to correct the estimator for measurement errors and how to estimate the variance of the corrected estimator.

18. The results on the HT estimator in Ilves and Laitila (2009) are extended to the GREG estimator in Ilves (2012a) and to the calibration estimator in Ilves (2012b). In the Ilves (2012a) paper, an empirical illustration is given based on data from the Statistics Sweden short term statistics study on employment, the second quarter of 2008. The illustration is a simulation study where the use of Bernoulli, Poisson and the PoMix designs in probability editing are compared. For the Poisson and the PoMix design, score values are calculated based on data available from the previous year, the second quarter of 2007. An initial bias of 4.2% is reduced to zero irrespective of design for sampling of units for editing. The precision of estimates, in terms of the coefficient of variation (CV), is 2 – 2,5 times larger for the Bernoulli sampling design compared with the Poisson sampling design. The Poisson sampling design yields a higher rate of edited observations. The PoMix design yields slightly higher CV estimates. In the simulations, three levels of editing was studies, 5, 10 and 15%. For the Poisson design, the CV decreases from 2.5% to 2.2% when the rate of editing increases from 5 to 15%.

IV. Comparison of selective and probability editing

19. There are a number of advantages of probability editing over selective editing. One major advantage is the ability to evaluate the statistical properties of the estimator based on probability edited data. Probability selection of units for editing allows for assessment of unbiasdnessm and variance, and development of variance estimators. Another advantage is its general applicability. If valid prior information is available, it can be utilized to construct scores for indication of suspicious observations and used in the sampling design to oversample strongly suspicious observations. If no such information is available, probability editing can still be applied.

20. A third advantage is its applicability irrespective of the kind of data considered. There is no need for developing different procedures for different kind of data, i.e. for continuous, count or categorical variables.

21. Another important advantage is related to the randomness introduced in editing and utilized for the forming of unbiased estimators. Selective editing does not provide a control over the bias of the resulting estimator. Calculation of pseudo-bias in prior data sets may give an indication of the size of the bias. However, the pseudo-bias is valid for the prior data set and not the data set edited. Interpretation of the statistics has then to be made conditionally on an assumption on the size of the bias being on par with the pseudo-bias. Probability editing can be utilized for unbiased estimation and any assumptions regarding bias due to measurement errors can be avoided for the interpretation of the statistics.

22. Probability editing also have an advantage in secondary use of the data. The design of selective editing is based on assessment of the pseudo-bias in prior data sets. If the collected data is used for other purposes then those considered in the design of the scores for selection of units for editing, the pseudo-bias is unknown in general, unless it is possible to evaluate it with the prior data set. In probability editing, this is not a problem since using the sampling design for selection of units for editing and the observed measurement errors, unbiased estimators can be defined for new population parameters.

23. One disadvantage of probability editing, compared with selective editing, is that strongly suspected observations can be left unedited if selection probabilities of units for editing are restricted to be within the 0-1 interval. Leaving such observations unedited does not change the interpretation of the estimate obtained, the estimator is still derived from use of e.g. an unbiased estimator and the statistical inference can be made using confidence intervals. However, leaving an observation which deviates large from the rest of the observations leaves the estimator with a larger variance then if it was edited to a more “normal” value. Thus, editing such observations is of interest regarding the precision of the estimate derived.

24. Another aspect on leaving suspicious observations unedited is the secondary use of the data for analysis based on the microdata. Here, units with observations deviating largely from the rest may have serious impact on the result. Furthermore, the statistical agency may be discredited when external users discover large and obviously wrong observations in the data.

25. The problem of leaving suspicious or strange observations unedited in probability editing can be circumvented by adapting the two-step procedure suggested in Ilves and Laitila (2009). In this, selective editing is first applied in order to edit the most suspicious observations. Thereafter, probability editing is applied to the rest of the data set. This two step procedure avoids leaving strongly erroneous observations in the data and provides with means for unbiased estimation.

26. From a theoretical point of view, the implementation of the two-step procedure is facilitated by the theory on probability editing derived. In the two-step procedure, some units are selected with probability one and others are selected with probabilities less than one. Thus, the two-step procedure is just a special case of probability editing using e.g. a Poisson sampling design where some probabilities are set to one. The results derived by Ilves (2012a) and Ilves (2012b) are therefore also valid for the two step procedure.

V. Implementation of probability editing

27. Application of probability editing requires development of new estimators in the estimation stage in statistical surveys. If this requirement is neglected, resulting estimator will be biased and calculated variance estimates will be misleading. Calculated estimates can be expected to have larger bias then what would be obtained under selective editing, since the latter is directed for editing the largest errors. Thus, resources have to be placed on developing appropriate software for estimators making use of probability edited data.

28. When it comes to the editing phase in a statistical survey, probability adds little to the editing process in comparison to selective editing. Using scores for Poisson sampling similar procedures as undertaken in the design of selective editing can be done for developing score functions. In the editing phase, a randomization mechanism has to be added for the decision on which units to edit.

29. The theory developed so far on probability editing shows it to be a powerful tool for reducing the amount of editing and giving interpretable estimators with good precision. The potentials of saving resources are expected to be particularly great in large surveys.

30. There are some further problems to address for more efficient implementation of probability editing. One problem in a Poisson sampling design based on scores, calculated from observations obtained, is the unknown level of editing obtained. For proper application and known expected level of editing, the total of the score values need to be known in advance of the study. One ad hoc solution would be to await the collection of a first set of observations and based on them make an estimate of the score total. Before application the statistical properties of such a method need to be evaluated, however.

31. When applying probability editing and adapting appropriate estimation methods, macro editing is unnecessary with respect to the bias of the estimator. Suppose the estimator above can be expressed as $\hat{t}_z = \hat{t}_y - \hat{b}_y = \sum_r w_k y_k - \sum_{r_2} w_k (z_k - y_k)$. This estimator is unbiased, or asymptotically so, with respect to measurement errors, if an appropriate estimator of the bias b_y is used. That is, $E(\hat{t}_z) = t_z$. Now, if a set of observations, $r_{me} \subset r - r_2$, are edited after the probability editing phase, it would yield the estimator $\hat{t}_{zme} = \hat{t}_z + \sum_{r_{me}} w_k (z_k - y_k)$, thus it would introduce the bias $\sum_{r_{me}} w_k (z_k - y_k)$. Thus, the effect of macro editing on the estimate need to be subtracted from the bias estimate, otherwise the macro edited estimator is biased.

32. An aspect neglected in the prior paragraph is that the set chosen for macro editing is decided by the set chosen in the probability editing stage. The set of macro edited observations are therefore stochastically chosen. This does not alter the conclusion above, however. Lets express the dependence of r_{me} on r_2 by the notation $r_{me}(r_2)$, where $r_{me}(r_2)$ is deterministically chosen from $r - r_2$ conditionally on r_2. Since the expectation of a sum is equal to the sum of expectations, then

$$E(\hat{t}_{zme}) = t_z + E_{p(r_2)}\left(\sum_{r_{me}(r_2)} w_k (z_k - y_k)\right)$$

where $p(r_2)$ denotes the sampling design for selection of r_2.

References

Ilves, M. and T. Laitila (2009). Probability-Sampling Approach to Editing, *Austrian Journal of Statistics*, 38:3, 171-182.

Ilves, M. (2012). GREG estimation and probabilistic editing, *Metron*, (To appear)

Ilves, M. (2012). Estimation in the presence of nonresponse and measurement errors, Mimeo, Department of Statistics, Örebro university, Sweden.