OUTLIERS IN SAMPLE SURVEYS – INVITED SESSION DISCUSSION

Hyunshik Lee, Westat
Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Robust estimation, Working model, Variance-bias trade-off

First of all, I would like to congratulate the authors for their excellent papers on the subject. To my best knowledge, this is the first invited session devoted to the outlier problem in sample surveys. I hope this is the beginning rather than the end. I will discuss Papers 1 and 3 first because they are somewhat related.

Paper 1: "Simple and Robust Estimator for Sampling" by Beat Hulliger

A vast amount of robust estimation literature for handling of outliers under non-survey setting is available and there have been some attempts to transfer this knowledge to survey sampling. Notably some forms of M-estimation technique have been proposed to use (Chambers, 1986; Lee, 1991; Gwet and Rivest, 1992; Hulliger, 1995). The weighting procedure of these estimators is complicated or appears so and there seems some resistance in the survey community in adopting these methods besides other cultural barriers. The current author of Paper 1 is trying to break this barrier by proposing simpler and more understandable ways of deriving the estimation weights yet based on the traditional robust estimation principle.

To estimate the population mean $\bar{y}_U = \sum_U y_i / N$, often is used the Hajék-type estimator given as

$$T_M = \frac{\sum_S w_i y_i}{\sum_S w_i} \tag{1}$$

where $w_i = \pi_i^{-1}$. It can be considered that the estimator is obtained from the following estimating equation

$$\sum_S a_i \psi\left(\frac{y_i - T}{\sigma}\right) = 0 \tag{2}$$

with $a_i = w_i$ and $\psi = I$ (identity function), which allows unbounding influence of outlying y-values. To robustify T_M, we need to use a bounding ψ -function such as Huber's proposal,

$$\psi_H(t) = \begin{cases} c & \text{if } t > c \\ t & \text{if } -c \le t \le c \\ -c & \text{if } t < -c \end{cases} \tag{3}$$

To solve the estimating equation (2) with (3), an iterative procedure is needed. A popular iterative algorithm is the weighted least square algorithm. It is often the case that the estimator obtained from the first iteration is as good as the one obtained from full iteration (Lee, 1991). The current author is exploring this possibility further. The one-step robustified estimator obtained this way is given by

$$T_M = \frac{\sum_S w_i u_i y_i}{\sum_S w_i u_i} \tag{4}$$

where, with some robust estimator $\hat{\sigma}$ for the scale,

$$u_i = \begin{cases} 1 & \text{if } |y_i - T_0| \le c\hat{\sigma} \\ c\hat{\sigma}/|y_i - T_0| & \text{otherwise} \end{cases}$$

and T_0 is an appropriate initial estimator used in the iteration.

The estimator in (4) appears to be a (sampling) weight-modified version of the estimator in (1). However, the basic underlying principle here is not to modify the sampling weight but to curb the influence of large $|r|$ where $r = (y_i - T)/\sigma$ with no involvement of the sampling weight. Therefore, if an outlying weight causes a problem, it is not treated in this formulation.

When an auxiliary variable (x) is available, the usual model is

$$y_i = \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim \text{iid}\left(0, \eta(x_i)\sigma^2\right) \tag{5}$$

Under this model, the projection estimator is given by

$$T_{\text{PROJ}} = \sum_U \hat{\beta} x_i / N = \hat{\beta} \bar{X}_U \tag{6}$$

for some estimator $\hat{\beta}$. The current author uses this estimator. Other estimators, however, can be constructed using the $\hat{\beta}$. For example, the prediction estimator (model-based) is given by

$$T_{\text{PRED}} = \frac{1}{N}\left(\sum_S y_i + \sum_{\bar{S}} \hat{\beta} x_i\right), \quad \bar{S} = U - S \qquad (7)$$

and the generalized regression (GREG) estimator (model-assisted) has the form of

$$T_{\text{GREG}} = \frac{1}{N}\left[\sum_S \pi_i^{-1} y_i + \hat{\beta}\left(X_U - \sum_S \pi_i^{-1} x_i\right)\right] \qquad (8)$$

Based on the model given in (5), a $\hat{\beta}$ can be obtained by solving the following estimating equation:

$$\sum_S a_i \psi\left(\frac{y_i - \beta x_i}{\sigma\sqrt{\eta(x_i)}}\right)\frac{x_i}{\sqrt{\eta(x_i)}} = 0 \qquad (9)$$

If $\eta(x_i) = x_i$ (the usual ratio model), with $\psi = I$ and $a_i = w_i = \pi_i^{-1} = X_U/(nx_i)$, the solution for β is given by

$$\hat{\beta} = \frac{\sum_S w_i x_i y_i/\eta(x_i)}{\sum_S w_i x_i^2/\eta(x_i)} = \frac{1}{n}\sum_S \frac{y_i}{x_i} \qquad (10)$$

Then $T_{\text{PROJ}} = \hat{\beta}\bar{X}_U = T_{\text{HT}}$. For this reason, the author asserts that the motivating model for the HT estimator is (5) with $\eta(x_i) = x_i$. However, it has been shown and can be easily seen that the HT estimator is a BLUE under the model (5) with $\eta(x_i) = x_i^2$ - the HT estimator is obtained using $\hat{\beta} = \frac{1}{n}\sum_S \frac{y_i}{x_i}$, which is the solution of (9) with $\psi = I$ and $a_i = 1$. The robustification of the HT estimator under the model with $\eta(x_i) = x_i^2$ is given by the same form as T_{HTS} but the u_i is defined differently as

$$u_i = \begin{cases} 1 & \text{if } |r_i| \le c\hat{\sigma} \\ c\hat{\sigma}/|r_i| & \text{if } |r_i| > c\hat{\sigma} \end{cases} \qquad (11)$$

where $r_i = y_i/x_i - \beta_0$ with some initial estimator β_0 such as $\text{med}(y_i/x_i)$. (If the sampling weight (w_i) is incorporated in the estimating equation, a different estimator than the HT estimator will be obtained, which can be robustified similarly.)

Robustification of an estimator by the M-estimation technique is inherently model-based or at least model-assisted. Therefore, it is important to choose a plausible working model. If the chosen model fits well with the set of the nonoutliers in the population, a robustified estimator based on that working model should be better than other estimators based on a different working model.

Some good practical advices in application of the one-step procedures are provided. For example, when outlying weights are present, the author advises to treat them separately. However, separate treatment of outlying y-values and outlying (sampling) weights could still leave some outlying weighted y-values. Also variable-by-variable treatment can destroy the relational structure of a multivariate data set.

The proposed procedures may work well for symmetric populations but they could have a serious bias for skewed populations, which are often encountered, in establishment surveys. This was clearly demonstrated in the simulation study.

The availability of a variance estimator for the procedure is an important advantage. It is yet to see how well the procedures work in interval estimation.

Paper 3: "Down-weighting Influential Clusters in Surveys, with Application to the 1990 Post-Enumeration Survey" by Thomas R. Belin, Gregg J. Diffendal, Nathaniel Schenker, and Alan Zaslavsky

This paper is somewhat related to the first in the sense that it discusses how to down-weight data with outliers.

The PES uses T_M-type of estimator, which is not robust. This paper presents the first attempt to use a systematic robust estimation technique in social surveys. Weight truncation (or Winsorization) is sometimes used in an ad-hoc fashion. To robustify T_M, the down-weighting factors u_i are derived assuming that the underlying distribution is a t_ν-distribution as

$$u_i = \left[1 + \frac{\{(y_i - T_0)/\sigma\}^2}{\nu - 2}\right]^{-1} \qquad (12)$$

It is not uncommon that a t-distribution fits actual data better than a normal distribution. Note, however, the distribution here is that of weighted cluster values. In this way, the sampling weights are incorporated in the formulation. If we look at the estimation problem in terms of an estimation equation given in (2), a_i are set to 1, and y_i are weighted cluster values. If these weighted values follow closely the normal distribution, T_M is the best estimator. However, the authors found that the weighted values

follow more closely a Student t-distribution. The M-estimator under a t-distribution is then given by (4) with u_i defined by (12). If weighted values can be modeled like this, we can obtain a more efficient estimator than T_M by using the down-weighting factor derived from the model.

The authors showed that the gain in variance efficiency is very large with the t-distribution of a lower degrees of freedom – the variance reduction is over 90 percent for the national estimate. However, bias seems to be quite serious and bias-variance trade-off is needed. The robust estimator based on t_{20} seems good – the bias is limited and the variance reduction is still large (70 percent for the national estimate).

The Q-Q plot also suggests that other robust estimators (Huber type or bisquare M-estimators) may work well as well. Particularly, I suggest studying the Huber-type because it provide a systematic weight truncation procedure, which will be more readily acceptable to survey practitioners because such a procedure is already being used in an ad-hoc fashion.

Paper 2: "Stratum Jumpers: Can We Avoid Them?" by Louis-Paul Rivest

This paper discusses an important problem for business surveys, where stratum jumpers are very troublesome for repeated business surveys. The usual treatment methods of stratum jumpers are down-weighting or Winsorization.

Stratum jumpers occur due to faulty size measure. The size measure used for stratification might be good at the time of the sample design but might have been deteriorated over time.

Optimal stratification procedure such as Lavallée and Hidiroglou (1988) based on the assumption that size measure (x-variable) is a perfect predictor of the study variable (y-variable) is vulnerable to the problem. The author is trying to address this problem by generalizing Lavallée/Hidiroglou algorithm based on two stratum jumper models: multiplicative model and random replacement model.

Hidiroglou and Srinath (1993) considered a similar algorithm based on the model given in (5). The algorithms proposed by Hidiroglou and Srinath and the current author provide more realistic stratification boundaries than the one by Lavallée and Hidiroglou (1988). The principle of the algorithms is the same between these alternatives but the underlying models are very different. Therefore, the resulting boundaries can be quite different as well. It would be interesting to compare these procedures.

The proposed procedure focuses more on smaller size strata and tends to allocate more to the smaller size strata and less to the larger size strata than the procedures that do not consider stratum jumpers. This makes sense since it is more likely that a unit with a small size measure jumps over to a large size stratum than the other way around. In addition, such jumpers are more trouble some because of its large weight. On the other hand, a unit in a large size stratum that jumps down to a smaller size stratum is less problematic because its weight is small.

If the model (5) is correct, model-based stratification also works well (Särndal et al., 1992). However, the model decays quickly over time in repeated business surveys, which creates stratum jumpers (outliers).

It is important to prevent the occurrence of stratum jumpers (outliers) at the design stage as much as possible. However, can we avoid them completely? I believe the answer is negative. If we cannot prevent them completely, we need a strategy to handle them when they occur. Therefore, a stratification strategy such as this together with an effective robust estimation method should be used. If the core part of the data follows closely the assumed model, stratification using the procedure by Hidiroglou and Srinath or the model-based method with an effective method of stratum jumpers treatment at the estimation stage would be another workable alternative.

References

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of American Statistical Association*, 81, 1063-1069.

Gwet, J.-P. and Rivest, L.-P. (1992). Outlier resistent alternative to the ratio estimator. *Journal of American Statistical Association*, 87, 1174-1182.

Hidiroglou, M., and Srinath, K.P. (1993). Problems associated with designing sub-annual business surveys. *Journal for Business and Economic Statistics*, 11, 397-405.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimator. *Survey Methodology*, 21, 79-87.

Lavallée and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

Lee, H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the 1991 Annual Research Conference*, Washington, D.C., 178-202.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer.