

INFLUENTIAL OBSERVATIONS: IDENTIFICATION AND TREATMENT BY M-ESTIMATORS

J. Philippe Gwet¹

ABSTRACT

M-estimators were officially introduced by Huber (1964), and are often used to produce more precise finite population estimates when the underlying model is outlier-prone. However, it is a well-known fact that these estimators offer their best performance when the outliers are symmetrically distributed, because of the potential bias that may be due to asymmetry. This is a very strong assumption that is not likely to be satisfied often in practice. In order to implement M-estimators routinely, it is essential to modify them in order to accommodate asymmetric distribution to some extent. This paper investigates such an alternative that is referred to here as CM-estimators. The main property of this alternative is consistency within the design-based inference framework. The CM-estimator is still biased, but even if applied to a skewed population, its bias decreases gradually towards 0 as the sample size increases. Some asymptotic properties of this estimator will be studied and ways of estimating the bias and the variance will be proposed. Results from a small simulation study will also be presented.

KEY WORDS: Asymptotic; CM-estimator; Consistency; M-estimator; Outlier; Robust.

RÉSUMÉ

Les M-estimateurs ont officiellement été introduit par Huber (1964), et sont souvent utilisés pour produire des estimations plus précises pour une population finie lorsque le modèle sous-jacent est sujet à des données aberrantes. Toutefois, c'est un fait bien connu que ces estimateurs offrent une meilleure performance lorsque les données aberrantes sont symétriquement distribuées, parce qu'un biais potentiel dû à l'asymétrie peut en découler. Il s'agit d'une très forte hypothèse qui n'est probablement pas très souvent satisfaite en pratique. Pour pouvoir utiliser de façon régulière les M-estimateurs, il est essentiel de les modifier dans le but de les adapter, dans une certaine mesure, à une distribution asymétrique. Cet article étudie une telle alternative qu'on appellera ici CM-estimateur. La priorité principale de cette alternative est la cohérence à l'intérieur du cadre de travail où l'inférence est basée sur le plan de sondage. Le CM-estimateur est toutefois biaisé, mais même s'il est appliqué à une population asymétrique, son biais décroît graduellement vers 0 à mesure que la taille échantillonnale augmente. Quelques propriétés asymptotiques de cet estimateur seront étudiées et différentes façons d'estimer le biais et la variance seront proposées. On présente également les résultats d'une étude provenant d'une petite simulation.

MOTS-CLÉS: Asymptotique; CM-estimateur; consistance; M-estimateur; données aberrantes; robustesse.

1. INTRODUCTION

This paper focuses on the estimation of the finite population mean of a single variable of interest. For the sake of simplicity we will confine ourselves to the single-stage sampling design with no auxiliary variable available. In the model-assisted survey sampling framework, this would be referred to as the estimation under the common mean superpopulation model. Within this simple framework, the natural estimator often used in practice, is the Horvitz-Thompson estimator. It is however known that this

simple estimator can become highly unstable if applied to a sample from a skewed population. This reality has led many authors to develop alternative estimators.

Within the pure model-based inference framework, P. Huber (1964) suggested the maximum likelihood type estimator, commonly referred to as M-estimator, by replacing the square function of the least square approach with an alternative function that is similar in the middle of the distribution but bounded for extreme values. As pointed out by Fuller (1993) these robust

¹ J. Philippe Gwet, Westat, 1650 Research Blvd., Rockville, MD 20850, USA.
E-mail: gwetj1@westat.com

procedures generally produce heavily biased estimators of the mean in skewed populations.

Although M-estimators are often used to estimate theoretical parameters such as the location parameter of a theoretical distribution, Chambers (1986) attempted to apply this technique in finite population sampling where he formalized for the first time, the concepts of representative and non-representative outliers. Representative outliers in Chambers' terminology represent legitimate population units and should be treated accordingly. On the other hand, non-representative outliers are essentially caused by errors in data processing. Chambers and Kokic(1993) suggests either to correct them or to assign them a weight of 1.

Another attempt to apply M-estimators in sample surveys is from Gwet and Rivest (1992). They developed their robust estimator within the design-based inference framework, and under a ratio superpopulation model assuming essentially that there is no non-sample outlier in the population. Rivest and Rouillard (1991) propose other interesting forms of M-estimators.

2. THE M-ESTIMATOR AND ITS LIMITATIONS

Let U be a finite population of size N . The y character denotes the variable of interest and, y_i the y -value taken by the i^{th} unit. The parameter of interest that we like to estimate is the population total T_y given by:

$$T_y = \sum_{i=1}^N y_i.$$

Note that $T_y = N\bar{Y}$, where \bar{Y} is the population mean. Therefore an estimate of the mean can easily be obtained by dividing the total estimate by the population size or its estimate.

We also assume that the finite population U is a simple random sample from an infinite superpopulation that can be described by the following theoretical model:

$$(\xi_0) \begin{cases} y_i = \beta + \varepsilon_i, \\ E_{\xi_0}(\varepsilon_i \varepsilon_j) = 0, \forall i \neq j, \\ E_{\xi_0}(\varepsilon_i) = 0 \text{ and } V_{\xi_0}(\varepsilon_i) = \sigma^2. \end{cases}$$

Under the simple random sampling design, a natural estimator of T_y in this framework is the expansion estimator \hat{T}_y given by,

$$\hat{T}_y = \frac{N}{n} \sum_{i=1}^n y_i,$$

where n is the sample size. This estimator is also identical to the Generalized Regression (GREG) estimator under the common mean superpopulation model.

It is however recognized that the sample mean, although design unbiased can have an unduly large variance when the population is very skewed. One can see a skewed population as being generated by a skewed distribution of ε_i . The alternative proposed by Hidiroglou and Srinath (1981) was to assign a predetermined weight of λ to all sample units identified as being outlying and to adjust the remaining weights so that all the weights can sum to the population size N . Let s_1 and s_2 denote the non-outlier and the outlier samples respectively and n_0 the number of sample outlier. Then the Hidiroglou-Srinath estimator is given by:

$$\hat{T}_y = \sum_{i=1}^n w_i y_i \text{ where } w_i = \begin{cases} \lambda & \text{if } i \in s_2, \\ \frac{N - \lambda}{n - n_0} & \text{if } i \in s_1. \end{cases}$$

The determination of a good value for λ is difficult and will usually be chosen subjectively. Such a value will probably lead to interesting results if it is close to the ratio of the number of population outliers by the number of sample outliers.

Chambers (1986) suggested using an M-estimator in the finite population total estimation. His model-based estimator is given by:

$$\hat{T}_{yCH} = \sum_{i=1}^n y_i + (N - n) \left[\hat{\beta}_M + \frac{\hat{\sigma}}{n} \sum \psi_{2H} \left(\frac{y_i - \hat{\beta}_M}{\hat{\sigma}} \right) \right]$$

where $\hat{\beta}_M$ is an M-estimator of β , and $(\hat{\beta}, \hat{\sigma})$ is implicitly defined as solution to the following estimating equation:

$$\sum_{i=1}^n \psi_H \left(\frac{y_i - \beta}{\sigma} \right) = 0 \text{ where}$$

$$\sigma = 1/\Phi^{-1}(3/4) \text{ median}_{1 \leq k \leq n} |y_k - \beta|$$

One popular choice for the ψ_H function was suggested by Huber (1964) and is defined by

$$\psi_H(t) = \begin{cases} K & \text{if } t > K, \\ t & \text{if } -K \leq t \leq K, \\ -K & \text{if } t < -K. \end{cases}$$

The tuning constant K is usually taken between 1 and 2 except for the ψ_{2H} function for which Chambers (1986) suggested to use a tuning constant as large as 10. A comprehensive development of the theory of robust estimation can be found in Huber (1981) and Hampel et al. (1986).

Gwet and Rivest (1992) suggested a simpler M-estimator with the design-based framework. The Gwet-Rivest estimator is obtained in a two-step process. (1) Considering the superpopulation model (ξ_0), a finite population parameter B_M is introduced first. This parameter is supposed to be a good estimator of β under the assumption that the finite population is a simple random sample from the superpopulation model. The parameter B_M is implicitly defined by the following equation:

$$\sum_{i=1}^N \psi_H \left(\frac{y_i - B_M}{\sigma_M} \right) = 0 \text{ where}$$

$$\sigma_M = 1/\Phi^{-1}(3/4) \text{ median}_{1 \leq k \leq N} |y_k - B_M|.$$

If the sample is selected according to an arbitrary design where the selection probability of the i^{th} unit is given by π_i , then a consistent estimator \hat{B}_M of B_M can be obtained as solution to the following equation:

$$\sum_{i=1}^n \frac{1}{\pi_i} \psi_H \left(\frac{y_i - \hat{B}_M}{\hat{\sigma}_M} \right) = 0 \text{ where}$$

$$\hat{\sigma}_M = 1/\Phi^{-1}(3/4) \text{ median}_{1 \leq k \leq n} |y_k - \hat{B}_M|.$$

The Gwet-Rivest estimator of the population total T_y under the simple random sampling design is given by:

$$\hat{T}_{yGR} = N \hat{B}_M$$

$$= \sum_{i=1}^n \left(\frac{N \alpha_i / \pi_i}{\sum_{k=1}^n \alpha_k / \pi_k} \right) y_i,$$

where the α_i 's are defined as follows:

$$\alpha_i = \begin{cases} \frac{\psi_H(e_{Mi}^*)}{e_{Mi}^*} & \text{if } |e_{Mi}^*| > 0, \\ 1 & \text{if } e_{Mi}^* = 0. \end{cases}$$

$$\text{for } e_{Mi}^* = \frac{e_{Mi}}{\hat{\sigma}_M} \text{ and } e_{Mi} = y_i - \hat{B}_M.$$

The Gwet-Rivest estimator is biased. Its bias can even be quite substantial if the population under study is generated from a very skewed superpopulation distribution. In fact this estimator works well when the underlying superpopulation distribution is symmetric. More formally let the function g_0 be defined for all real number b by:

$$g_0(b) = \int \psi_H \left(\frac{y-b}{\sigma} \right) dF_{\xi_0}(y),$$

where F_{ξ_0} is the common theoretical distribution of the superpopulation units y_i , and σ their common standard deviation. If the superpopulation model location parameter β is the unique solution of the equation $g_0(b) = 0$, and g_0 has a bounded first-order derivative, then the Gwet-Rivest is consistent with respect to the ξ_0 model and the design. Gwet (1997) proves this result. The use of theoretical models in conjunction with the sampling design seems to be the only way to develop a useful asymptotic theory on robust estimators. Although one can make a model-free statement on the variance of robust estimators, it is difficult to make such a statement on their overall precision. This is due to the fact that most robust estimators such as M-estimators are design biased, the magnitude of which is model dependent.

In the next section we will introduce CM-estimators, where C stands for consistent. This new class of estimators is obtained from the class of M-estimators by letting the tuning constant go to infinity at a certain rate. Gwet (1997) proves that CM-estimators are design consistent. However, the order of magnitude of the bias depends on the limiting distribution of the finite population, which is that of the superpopulation model.

3. THE CM-ESTIMATOR AND ITS LARGE SAMPLE PROPERTIES

There are at least two possible approaches towards developing consistent robust estimators. One approach would be model-based. The robustness constant would then be optimally chosen under a completely specified parametric model. Rivest and Hurtubise (1995) used that approach to derived a cut-off for winsorized estimators that would be near optimal for a wide variety of distributions. Some of their findings will guide our choice of robustness constant for the CM-estimator.

The first step in defining the CM-estimator will be to introduce the following sample size dependent theoretical quantity $B_{M(n)}$ implicitly defined as solution to the following equation:

$$\sum_{i=1}^N \psi_H \left(\frac{y_i - B_{M(n)}}{\sigma_{M(n)}}, K(n) \right) = 0 \text{ where}$$

$$\sigma_{M(n)} = 1 / \Phi^{-1}(3/4) \text{median}_{1 \leq i \leq N} |y_i - B_{M(n)}|,$$

where $K(n)$ is a robustness constant to be specified. The only constraint on this constant at the moment is its limiting value that must be infinity. Other choices for $K(n)$ that would increase the estimator efficiency will be further discussed later in this section. It should be noted that as n goes to infinity, the ψ_H function will get closer and closer to the identity function and the difference between $B_{M(n)}$ and the population mean will be smaller.

The CM-estimator \hat{T}_{yCM} of the population total is given by:

$$\hat{T}_{yCM} = N\hat{B}_{M(n)},$$

where $\hat{B}_{M(n)}$ is a consistent estimator of $B_{M(n)}$ in the sense that $\left| \hat{B}_{M(n)} - B_{M(n)} \right| \xrightarrow[n \rightarrow \infty]{P} 0$. It is defined by the following equation:

$$\sum_{i=1}^n \frac{1}{\pi_i} \psi_H \left(\frac{y_i - \hat{B}_{M(n)}}{\hat{\sigma}_{M(n)}}, K(n) \right) = 0 \text{ where}$$

$$\hat{\sigma}_{M(n)} = 1 / \Phi^{-1}(3/4) \text{median}_{1 \leq i \leq N} |y_i - \hat{B}_{M(n)}|.$$

This estimator can be computed in the same way as standard M-estimators. Huber (1964) and Hampel et al. (1981) discuss several ways of computing M-estimators. Rivest (1989) also discusses some interesting limitations to the standard computation methods.

Note that with a robustness constant $K(n)$ going to infinity, the difference between the CM-estimator and the expansion estimator will be of order $O_p(n^{-1/2})$ for most limiting distributions. Therefore these two estimators will have the same asymptotic distributions. This is due to the fact that for many common distributions, there is no outliers at infinity. There is one class of distributions however for which this does not happen. It is described by the following equation:

$$1 - H(r) \approx \frac{C}{r^\alpha},$$

where $H(\cdot)$ is the distribution function. The Pareto family of distributions for example belongs to this family.

In order to develop an asymptotic theory that will be useful for finite samples, we suggest to derive $K(n)$ under the assumption that the limiting distribution belongs the Pareto family where $F_{\xi_0}(\cdot)$ is given by:

$$F_{\xi_0}(x) = \begin{cases} 1 - (x/\theta)^{-\alpha} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Using the customary framework for asymptotics in finite population sampling, one can assume a sequence of embedded finite populations $\{U_t\}_{t \geq 1}$, having each a distribution function $F_{N_t}(\cdot)$. It is further assumed that this sequence of distribution functions converges almost surely to $F_{\xi_0}(\cdot)$.

Rivest and Hurtubise (1995) have proved that the optimal cut-off for a winsorized mean under the simple random sampling design must satisfy the following equation:

$$1 - H(R) \approx \frac{(\alpha - 1)}{n}$$

For large samples, one can approximate the CM-estimator as follows:

$$\hat{T}_{y_{CM}} \approx \sum_{i=1}^n \frac{z_i^*}{N\pi_i} \text{ where}$$

$$z_i^* = \begin{cases} z_i & \text{if } z_i \leq \frac{\sigma_{CM}K(n)}{F(K(n))} + B_{CM}, \\ \frac{\sigma_{CM}K(n)}{F(K(n))} + B_{CM} & \text{otherwise,} \end{cases}$$

and

$$z_i = \frac{y_i - B_{CM}(1 - F(K(n)))}{F(K(n))}.$$

It follows from the Rivest-Hurtubise optimality result that a good choice for $K(n)$ is given by:

$$K(n) \approx (\alpha - 1)^{-1/\alpha} n^{1/\alpha}.$$

Simulation results not reported here indicate that for $\alpha \approx 2$, the CM-estimator will usually be better than most standard estimators in terms of mean square error.

4. CONCLUDING REMARKS

The robust procedures outlined in this paper can be used to compute more precise estimates in a wide variety of situations. It should be stressed that M-estimators can be heavily biased unless one has faith in the symmetric nature of the model that underlies the population under study. We have indicated how M-estimators can be used under a complex sampling design.

When the population under study is known to be skewed, the M-estimator performance is usually poor, because of its bias. In this case, we recommend the use

of CM-estimators. Although still biased, the CM-estimator is nevertheless consistent. Therefore its bias decreases as the sample size increases. However, the rate at which the tuning constant should go to infinity is a difficult problem the solution of which depends upon the theoretical limiting distribution of the population under study. We have assumed in this paper that this limiting distribution belongs to the Pareto family. This assumption is somewhat subjective but it guarantees that at infinity there will still be an outlier problem. This condition is needed to develop a useful asymptotic theory.

REFERENCES

- Chambers, R. L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- Chambers, R. L., and Kokic, P. N. (1993). Outlier Robust Sample Survey Inference. *Proceedings of the 49th Session, International Statistical Institute*.
- Fuller, W. A. (1993). Estimators for Long-Tailed Distribution. *Proceedings of the 49th Session, International Statistical Institute*.
- Gwet, J. P. (1997). *Robust Statistical Inference in Survey Sampling*, Ph.D. Thesis, Carleton University.
- Gwet, J. P., and Rivest, L. P. (1992). Outlier Resistant Alternatives to the Ratio Estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. E. (1986). *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.
- Hidiroglou, M. A., and Srinath, K. P. (1981). Some Estimators of a Population Total From Simple Random Samples Containing Large Units, *Journal of the American Statistical Association*, 76, 690-695.
- Huber, P.J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35, 73-101
- Rivest, L. P., and Rouillard, E. (1991). M-estimators and Outlier Resistant Alternatives to the Ratio Estimator, *Proceedings of the 1990 Symposium of Statistics Canada*, 271-285.