

## ON THE TREATMENT OF INFLUENTIAL OBSERVATIONS IN HOUSEHOLD SURVEYS

Asma Alavi and Jean-François Beaumont<sup>1</sup>

### ABSTRACT

Household expenditure or income surveys often deal with highly skewed distributions, which potentially lead to samples with some extreme observations. The problem is aggravated by the fact that there usually is a low amount of useful auxiliary information available at the design stage and that the sampling design is complex most of the time, leading to widely dispersed design weights. Therefore, it could happen that a large value is associated with a large design weight and that, this combination has a great influence on the estimates produced by the survey. Design consistent estimators, such as the Generalized REGression (GREG) estimator, are usually highly variable in the presence of influential observations but they have a low bias whereas model-based estimators are more stable but they are generally not consistent and more biased. In this paper, a compromise between these two types of estimators is proposed and a simulation study shows that it performs well with respect to the bias and mean squared error (MSE) criteria in comparison with some other robust estimators.

KEY WORDS: GREG estimator; Model-based estimator; M-estimator; Outliers; Robust estimator; Synthetic estimator.

### RÉSUMÉ

Les enquêtes sur les dépenses ou sur le revenu des ménages font souvent face à des distributions très asymétriques, ce qui conduit potentiellement à des échantillons avec des observations extrêmes. Le problème est amplifié par le fait qu'il n'y a généralement qu'une faible quantité d'information auxiliaire utile disponible lors de la conception du plan de sondage et que le plan de sondage est la plupart du temps complexe, ce qui conduit à des poids de sondage très dispersés. Il pourrait donc arriver qu'une grande valeur soit associée à un grand poids de sondage et que cette combinaison ait une grande influence sur les estimations produites par l'enquête. Les estimateurs convergents par rapport au plan de sondage, tel que l'estimateur par la RÉgression Généralisée (REGG), sont généralement très variables en présence d'observations influentes mais ont un faible biais tandis que les estimateurs basés sur un modèle sont plus stables mais ne sont généralement pas convergents et plus biaisés. Dans cet article, un compromis entre ces deux types d'estimateurs est proposé et une étude de simulation montre que ce compromis donne de bons résultats en regard du biais et de l'erreur quadratique moyenne (EQM) en comparaison avec d'autres estimateurs robustes.

MOTS CLÉS: Données aberrantes ; estimateur basé sur un modèle; estimateur-M; estimateur REGG; estimateur robuste; estimateur synthétique.

### 1. INTRODUCTION

Household expenditure or income surveys often deal with highly skewed distributions, which potentially lead to samples with some extreme observations. The problem can be aggravated when such extreme observations are associated with large design weights. In order to avoid this, a large selection probability (therefore, a small design weight) should be assigned to units with large values of the variables of interest and vice-versa. This can be achieved with proper stratification or with probability proportional to size sampling when the auxiliary variables available at the design stage are

well correlated with the variables of interest. In household surveys, however, such useful auxiliary variables are often not available. Moreover, several variables are usually collected in the same survey and appropriate auxiliary variables for one variable of interest may not necessarily be appropriate for another.

Design consistent estimators, such as the Generalized REGression (GREG) estimator, may be highly variable in the presence of influential observations. Since useful auxiliary information is often not available at the design and at the estimation stage of a survey, more robust (to

<sup>1</sup>Asma Alavi and Jean-François Beaumont, Labour Force Survey Methods Section, Household Survey Methods Division, Methodology Branch, Statistics Canada, Ottawa, Ontario, K1A 0T6, [asma.alavi@statcan.ca](mailto:asma.alavi@statcan.ca).

influential observations) estimators may be needed. Modifying the value (for example, the *Winsorization* technique) or modifying the weight (see, for example, Hidiroglou and Srinath (1981) of an influential observation are the two traditional approaches that have been used in sample surveys to obtain robust estimators. More recently, the M-estimation technique has been considered to form robust estimators (see, among others, Chambers, 1986; Gwet and Rivest., 1992; Lee, 1991, 1995; and Hulliger, 1995).

In this paper, an observation is defined as being influential if its exclusion or inclusion in the sample affects the estimates greatly. Therefore, an observation can be influential because of a large design weight, a large value or the combination of both. An extreme observation is an observation isolated from the bulk of the data. An extreme observation is often associated to a large value of the variable of interest (or a large regression residuals if a regression estimator is used). Note that an extreme observation is not necessarily influential in a large sample. The term *outlier* is also frequently seen in the literature and it usually refers to either an extreme or an influential observation. These definitions are very closely related to those of Lee (1995) and are suited to survey sampling.

## 2. BACKGROUND

In this section we discuss the GREG estimator along with some of the estimators that need detection of outliers before treating them.

Suppose we want to estimate the total of a variable of interest  $y$  for a population  $U$  and denote this unknown population parameter by  $t_y = \sum_{k \in U} y_k$ .

Let us also assume that we have a vector of auxiliary variables,  $\mathbf{x}_k$ , available for all units of the sample  $s$  and for which the population totals,  $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$ , are known. The GREG estimator can then be used to estimate the unknown total  $t_y$ :

$$\hat{t}_y^G = \hat{t}_y^{HT} + \hat{\mathbf{B}}'(\mathbf{t}_x - \hat{\mathbf{t}}_x^{HT}), \quad (2.1)$$

where,  $\hat{t}_y^{HT} = \sum_{k \in s} w_k y_k$  and  $\hat{\mathbf{t}}_x^{HT} = \sum_{k \in s} w_k \mathbf{x}_k$  are Horvitz-Thompson estimators,  $w_k$  is the design weight and

$$\hat{\mathbf{B}} = \left( \sum_{k \in s} \frac{a_k}{c_k} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} \frac{a_k}{c_k} \mathbf{x}_k y_k. \quad (2.2)$$

In most practical cases,  $a_k = w_k$  and  $c_k = \lambda' \mathbf{x}_k$ , where  $\lambda$  is a vector of known constants. In these cases, it can easily be shown that (2.1) reduces to the synthetic estimator  $\hat{t}_y^S = \hat{\mathbf{B}}' \mathbf{t}_x$ . Regarding the determination of  $a_k$ , the usual choice  $a_k = w_k$  makes  $\hat{\mathbf{B}}$  a design consistent estimator for the population parameter  $\mathbf{B}$  by

$$\mathbf{B} = \left( \sum_{k \in U} \frac{1}{c_k} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in U} \frac{1}{c_k} \mathbf{x}_k y_k.$$

On the other hand, the choice  $a_k = 1$  will make  $\hat{\mathbf{B}}$  the best linear unbiased estimator for the vector of parameters  $\beta$  under the model,  $y_k = \beta' \mathbf{x}_k + \varepsilon_k$ , with  $E_m(\varepsilon_k) = 0$ ,  $E_m(\varepsilon_k \varepsilon_l) = 0$ , for  $k \neq l$ ,  $E_m(\varepsilon_k^2) = \sigma^2 c_k$ , (where  $\sigma^2$  is an unknown parameter) provided that the sampling mechanism is ignorable. It is also interesting to note that  $\hat{t}_y^G$  is design consistent whether the model holds or not and no matter how  $a_k$  is specified.

In the presence of influential data, the GREG estimator is highly variable. In order to treat influential data several methods have been proposed in the literature such as *Winsorization*, and *weighted Winsorization* described in Tambay (1988). These methods are based on a modification of the values of the variable of interest. They could also be transformed such that we modify the estimation weight instead of the  $y$ -values. However, this would require a different weight for each variable of interest, which is not very appealing to data users. To avoid the production of more than one weight, these methods can be applied to only one key variable or to some linear combination of the key variables.

A very simple method for reducing the effect of influential observations consists of giving a weight of one to all influential observations. This method can be too drastic in practice, especially if most of the estimation weights are large. A more interesting method is used by a few household surveys at Statistics Canada and is described in Tremblay (1998) and Alavi and Beaumont (2001). This method requires knowing population totals for a certain number of categories, say  $n_{cat}$ , of an auxiliary variable. This method is very similar to a poststratification with the constraints that the final weight be between one and the original weight, and that only the weight of influential observations be

modified. In the following, this method will be called *constrained poststratification*.

A unit  $l$  is identified as being influential when

$$\frac{w_l^* y_l}{\sum_{k \in s} w_k^* y_k} \geq p\%,$$

where  $p$  is a predetermined cut-off value and  $w_k^*$  is estimation weight. An unweighted version of above can also be used if the interest is in identifying observations that are influential uniquely because of their value.

### 3. ROBUST ESTIMATION METHODS

In this section, methods that are robust and do not need the detection step are considered. Ghangurde (1989) and Pelletier and Rancourt (1998) have studied the determination of the variance structure of the random error  $\varepsilon_k$ . Although these approaches are different, the idea in both papers is to give a reduced weight to outliers (or in our case, influential observations) when estimating  $\beta$ . In fact, this is also equivalent to giving a smaller value of  $a_k$  to influential observations.

An alternative to  $\hat{t}_y^G$  is the following composite estimator:

$$\hat{t}_y^C = \hat{\mathbf{B}}' \mathbf{t}_x + \theta (\hat{t}_y^{HT} - \hat{\mathbf{B}}' \hat{\mathbf{t}}_x^{HT}), \quad (3.1)$$

where  $\theta$  is an unknown parameter to be estimated from the data or to be chosen subjectively and  $\hat{\mathbf{B}}$  is given in (2.2). In fact, estimator (3.1) is a weighted average of the design consistent estimator  $\hat{t}_y$  and the usually more stable (but not necessarily design consistent) synthetic estimator  $\hat{t}_y^S$  ( $\hat{t}_y^C = \theta \hat{t}_y^G + (1-\theta) \hat{t}_y^S$ ). A choice of  $\theta$  between 0 and 1 will usually yield a compromise between the desire to have a low bias and the opposite desire to have a low variance. However, it should be noted that there is no compromise possible when  $a_k = w_k$  and  $c_k = \lambda' \mathbf{x}_k$  since in that case,  $\hat{t}_y^C = \hat{t}_y^G = \hat{t}_y^S$ .

An important issue with the composite estimator (3.1) is the determination of  $a_k$ . The idea is to give

a smaller value of  $a_k$  to influential observations, that is, those having a large design weight, or those having a large regression residual. Several options are possible to achieve this. In this paper, we study the following form for  $a_k$ ,

$$a_k = w_k^\alpha \exp(-\delta z_k),$$

where  $\alpha$  and  $\delta$  are unknown non-negative parameters to be estimated from the data or to be chosen subjectively and  $z$ , a variable to be appropriately chosen. In this paper, we considered, the absolute value of the standardized regression residuals as  $z$ . To calculate the standardized regression residuals, an initial estimate  $\hat{\mathbf{B}}^{(0)}$  of  $\beta$  is required, which can be obtained in replacing  $a_k$  by  $a_k^{(0)} = w_k^\alpha$  in (2.2). The absolute value of the standardized regression residuals can then be given

by  $z_k^{(0)} = \frac{|e_k^{(0)}|}{\hat{\sigma}^{(0)} \sqrt{c_k}}$ , where  $e_k^{(0)} = y_k - \hat{\mathbf{B}}^{(0)} \mathbf{x}_k$  is the regression residual for unit  $k$  and  $\hat{\sigma}^{(0)2} = \frac{\sum_{k \in s} a_k^{(0)} e_k^{(0)2} / c_k}{\sum_{k \in s} a_k^{(0)} - q}$ , where  $q$  is the dimension

of  $\mathbf{x}_k$ . Of course, it would be possible to do an iterative procedure calculating alternately  $a_k$ ,  $\hat{\mathbf{B}}$  and  $z_k$  until some convergence criterion is reached. The estimator  $\hat{\mathbf{B}}$  of  $\beta$  obtained after convergence could be viewed as an M-estimator since it can be shown that it is the solution of the following system of equations:

$$\sum_{k \in s} w_k^\alpha \psi \left( \frac{y_k - \beta' \mathbf{x}_k}{\sigma \sqrt{c_k}} \right) \frac{\mathbf{x}_k}{\sigma \sqrt{c_k}} = 0, \quad (3.2)$$

where  $\psi(t) = t \times \exp(-\delta |t|)$ . This function is known as a redescending  $\psi$  function. For positive values of  $t$  (and  $\delta > 0$ ), the function  $\psi(t)$  is approximately equal to  $t$  for small values of  $t$ , is increasing for  $t < 1/\delta$  and is decreasing toward 0 after that point. The situation is reversed for negatives values of  $t$ . The iterative procedure that has just been described to solve (3.2) is known as the iteratively reweighted least squares (IRLS) algorithm (Beaton and Tukey, 1974). Note that an estimating equation for the unknown parameter  $\sigma^2$  is also required to solve (3.2).

The parameter  $\alpha$  in (3.1) and (3.2) controls the impact of the design weights on  $\hat{\mathbf{B}}$ . The closer to zero is  $\alpha$ , the smaller is the impact of the design weights on  $\hat{\mathbf{B}}$  (and inversely). The extreme case of,  $\alpha=1$  is usually preferred by design-based survey statisticians. In the other extreme case ( $\alpha=0$ ), the design weights are not involved in the estimation of  $\beta$ . This is the case that model-based survey statisticians usually prefer. A value of  $\alpha$  between these two extreme cases can be viewed as an interesting compromise for both types of statisticians. The parameter  $\delta$  controls the impact of the variable  $z$  on  $\hat{\mathbf{B}}$  and must be greater than or equal to 0.

The conditions, under which the design bias of  $\hat{t}_y^C$  should be small, are studied in detail in Alavi and Beaumont (2001). The following simulation study shows that a compromise value for  $\alpha$ , ( $0 < \alpha < 1$ ) gives good results with respect to bias and mean squared error criteria for the synthetic estimator  $\hat{t}_y^S$ .

#### 4. SIMULATION STUDY

For the simulation study, we used data from the Statistics Canada's 1998 Survey of Household Spendings (SHS) to serve as the population. The survey had a stratified multi-stage design and contains information about 15,457 households on several variables. We looked at three key variables from the survey, namely: *Total Expenditure*, *Food* and *Renovation/Repair*.

From the population of households, 1000 samples of expected sample size 300 were selected using Poisson sampling. In estimation, we assumed that only one population total was known, which was the total number of households in the population ( $x_k = c_k = 1$ , for all households  $k$ ). In the following we use the notations WWIN for Weighted Winsorization and CP for Constrained Poststratification. For CP, Income was used as auxiliary variable (with 12 categories) and for the detection step, in order to have a strategy similar to what is used in practice for the SHS. For WWIN, the detection step was carried out separately for each variable of interest. We subjectively specified a cut-off value at 3% level.

For the composite estimator, we used various combinations of  $\theta$ ,  $\alpha$  and  $\delta$  values. Table 1 gives some of the combinations we used in our simulation

study as well as the corresponding notation. Note that, when  $\delta=0.5$ , the variable  $z$  is chosen as the absolute value of the standardized regression residuals obtained in using the key variable Total Expenditure. This variable  $z$  is used for every of the three variables of interest. Also, only one iteration is performed to obtain an M-estimate for  $\beta$  (when  $\delta=0.5$ ).

The estimated relative bias (RB), expressed as a percentage of population mean, was calculated using the formula,

$$\text{est(RB)} = \left[ \left( \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\mu}_i - \mu) \right) \frac{1}{\mu} \right] \times 100\%,$$

where  $\hat{\mu}_i$  is the population mean estimate for the  $i^{\text{th}}$  sample, and  $\mu$  is the population mean. An estimate of the relative root mean squared error (RRMSE) expressed as a percentage can be calculated as,

$$\text{est(RRMSE)} = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{\mu}_i - \mu)^2}{1000}} \times \left( \frac{1}{\mu} \right) \times 100\%.$$

#### 5. RESULTS

In this section, we present the results of the simulation study. A special attention should be given on the Total Expenditure variable. This variable has been used to form the variable  $z$  for the GREG and synthetic estimators.

**Table 1.**  $\theta$ ,  $\alpha$  and  $\delta$  Values and Notation for the Composite Estimator

Notation	$\theta$	$\alpha$	$\delta$
GREG-M	1	0	0
GREGM-M	1	0	0.5
GREG-C	1	0.5	0
GREGM-C	1	0.5	0.5
GREG-D	1	1	0
GREGM-D	1	1	0.5
SYN-M	0	0	0
SYNM-M	0	0	0.5
SYN-C	0	0.5	0
SYNM-C	0	0.5	0.5
SYN-D	0	1	0
SYNM-D	0	1	0.5

For the Renovation/Repair variable, all estimators have larger RRMSE values than the other variables. This is not surprising since it is the most skewed variable. Among the detect-and-treat estimators, CP performed always better than WWIN. This is not surprising since CP uses more auxiliary information (population totals are known for 12 categories of Income). The synthetic estimator with certain parametric combinations for  $\alpha$  and  $\delta$  represents a potential alternative. It is evident from table 2 that all GREG estimators perform similarly. They all have low RBs and relatively large RRMSEs. These estimators provide a basis for comparison since they are asymptotically unbiased and design consistent, but volatile in the presence of influential observations.

For the synthetic estimators, we have very interesting results. As mentioned in section 3, synthetic estimation has been proposed as a way to reduce variance (while keeping the bias reasonably low). However, the strictly model-based estimator, SYN-M, can be too much biased. On the other hand, the design consistent estimator SYN-D (which is equivalent to GREG-D) has a low RB, but relatively high RRMSE. It seems that SYN-C, which is a compromise between the model-based estimator (SYN-M) and the design consistent estimator (GREG-D), is an interesting alternative with respect to all criteria considered. It resulted in

RB values for all variables that were less than 5%. Although other estimators (for example, CP) had this property too, the RRMSEs were higher than those obtained with SYN-C. In general, the RB of the synthetic estimators has a lower bound equal to that of the GREG estimators, and the RRMSE of the synthetic estimators have upper bounds equal to those of the GREG estimators. When  $\delta = 0.5$ , the RB was still in general reasonably close (but slightly larger) to the case  $\delta = 0$ , which indicates that the value of this parameter was not set too high. For example, SYN-M-C performed comparably to SYN-C for all three variables.

## 6. CONCLUSION

In this paper, a number of estimators have been proposed and discussed to deal with the problem of influential observations, which occur because of extreme values, large design weights or the combination of both. All these estimators can take advantage of useful auxiliary information at the design stage and at the estimation stage. In fact, alternatives to the GREG estimator may not be needed when such useful auxiliary variables are available. However, in household surveys, such information is often not available, especially at the design stage. Therefore, more robust estimators are needed.

Table 2. Results from the Simulation Study

Estimator	Total Expenditure		Food		Renovation/Repair	
	RB	RRMSE	RB	RRMSE	RB	RRMSE
WWIN	-8.64	10.58	-6.29	8.11	-38.47	40.81
CP	-2.91	6.98	-1.03	5.83	-1.06	32.91
GREG-M	0.32	8.28	0.34	6.31	2.80	39.04
GREGM-M	0.28	8.31	0.37	6.42	2.81	39.00
GREG-C	0.26	8.29	0.27	6.17	2.79	38.89
GREGM-C	0.24	8.33	0.28	6.18	2.79	38.87
GREG-D	0.23	8.31	0.22	6.16	2.77	38.91
GREGM-D	0.22	8.32	0.22	6.17	2.77	38.91
SYN-M	4.74	6.33	12.84	13.39	-10.20	17.87
SYNM-M	-2.86	4.81	17.05	17.53	-6.62	18.04
SYN-C	-0.11	5.14	3.36	5.20	-3.26	22.55
SYNM-C	-4.37	6.44	4.72	6.21	-1.42	24.28
SYN-D	0.23	8.31	0.22	6.16	2.77	38.91
SYNM-D	-1.50	8.08	0.63	6.24	3.55	40.14

If the presence of influential observations can be justified, at least in part, by the presence of large design weights, then a model-based estimator may be useful under some conditions. In this paper, a compromise (SYN-C) between a strictly model-based estimator (SYN-M) and the design consistent estimator (GREG-D), has been shown to be very attractive through a simulation study using real life survey data. Finally, whatever estimator is chosen, it is always a good idea to verify empirically the relationships between the variables of interest and the auxiliary variables used at the estimation stage, even if a design consistent estimator is preferred.

## REFERENCES

- Alavi, A, and Beaumont, J.F. (2001), On the treatment of influential observations in household surveys. Working Paper: HSMD-2001-002E, Methodology Branch, Statistics Canada
- Beaton, A.E., and Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-185.
- Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- Ganghurde, P.D. (1989). Outliers in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 736-739.
- Gwet, J.P., and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- Hidirolou, M.A., and Srinath, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.
- Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.
- Lee, H. (1991). Model-based estimators that are robust to outliers. *Proceedings of the Annual Research Conference*, Washington, DC, U.S. Bureau of the Census, 178-202.
- Lee, H. (1995). Outliers in business surveys. In *Business Survey Methods*, Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J., and Kott, P.S. (editors), Chapter 26, New-York, John Wiley & Sons, Inc.
- Pelletier, E., and Rancourt, E. (1998). Spécification du paramètre de la structure de variance du modèle dans l'estimateur de régression généralisé. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 159-164.
- Tambay, J.L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 229-234.
- Tremblay, J. (1998). Détection des observations influentes pour l'Enquête sur les finances des consommateurs (EFC) et l'Enquête sur l'dynamique du travail et du revenu (EDTR). Internal report, Statistics Canada, Household Survey Method Division.