Detecting and Correcting Influential Values using the Conditional Bias Approach: Application to the Canadian Survey of Household Spending

Christiane Laperrière[1], Aliou Seydi[1]
[1] Statistics Canada, 100 Tunney's Pasture, Ottawa, ON, Canada

Abstract
A classical challenge faced by survey statisticians is how to reduce the impact of certain collected values on the survey estimates. Traditionally, outlier detection methods will focus on the values, or the weighted values, of the variable of interest. However, such approaches ignore the possibility that a typical value may still adversely impact the estimates as a result of the sample design employed, the nature of the parameter to be estimated or the estimator used. Conversely, outlier values may not be influential for a given sample design and estimation method. A more all-encompassing approach was sought, one that would reveal which units have the greatest influence over a given estimate and how exactly it exerts that influence.

To meet the challenge of producing robust estimates, the Survey of Household Spending (SHS) has looked into the notion of conditional bias which has just recently been proposed as a means of gauging a unit's overall influence on estimates. In this paper we describe this innovative approach as well as our results, along with the practical issues one may be facing when implementing this method in a complex survey.

Key Words: Conditional bias, Influential detection, Influential correction, Robust estimators, Tuning constant

1. Context

1.1 The Survey of Household Spending

The Survey of Household Spending (SHS) is an annual and voluntary survey that collects household expenditure data using a personal interview as well as an expenditure diary. The SHS uses a two-stage design and the sample of selected households is spread over 12 monthly collection cycles so that data collection is continuous from January through December. The interview collects regular expenditures (such as rent and electricity) and less frequent expenditures (such as furniture and dwelling repairs) for a reference period that varies in length depending on the type of expenditure. The two-week expenditure diary is used to collect frequent or smaller expenditures, which are difficult for respondents to recall during a retrospective interview. After edits and imputation, all the survey expenditure variables are annualized i.e. they are multiplied by an appropriate factor based on the reference period so that annual expenditure estimates can be produced. The annualization process can inflate expenditure values that are already large and amplify the impact of influential units on the estimates. It is desirable to identify these units and potentially reduce their impact on the estimates. Such corrections will reduce the variance and better allow year-to-year comparisons. This paper addresses the detection and correction of influential units at the estimation step. In other words, it is assumed that the edits and imputation steps are completed and that erroneous values have been corrected.

Therefore, the goal is to evaluate a method to detect and correct influential units that will reduce the variance of the estimates without introducing a large bias. The method described in this paper relies on the concept of conditional bias (Beaumont, Haziza and Ruiz-Gazen, 2013) and it can be used to define robust estimators which have lower variance in the presence of influential units. In this paper, we will apply this new method to the SHS and see how it compares to the usual method that is used in production. We will show how the conditional bias approach reduces manual intervention and is time-efficient compared to the current method. In the first section of this paper, we give some notation and describe the current method that is used to detect and correct the influential values for the SHS. In Section 2, we give an overview of the conditional bias approach for the treatment of influential values, as described in Beaumont et al. (2013). In Section 3, we apply this new method to the context of the SHS and give results related to the bias. Finally, the conclusion gives an overview of the results and describes future work.

1.2 Notation and Current Method to Detect and Correct Influential Values

Let U denote the population of interest and s a sample selected according to a sample design D. Let $I_i = 1$ if $i \epsilon s$, and $I_i = 0$ otherwise, and let π_i denote the probability of selection of unit i. Suppose we want to estimate a population parameter θ, which is a function of a variable of interest y, using the estimator $\hat{\theta}$. A unit will be defined as influential if its exclusion from the population (and therefore from the sample) has a large impact on the sampling error $\hat{\theta} - \theta$. Further, we define the configuration as the following quadruplet:

1. The variable of interest y and its distribution in the population.
2. The population parameter θ that we wish to estimate (and that is a function of the y-values).
3. The sample design and the estimator $\hat{\theta}$.
4. Whether or not unit i is in the sample s.

Traditionally, outlier detection methods will focus on the y-values, on the survey weights or the weighted values of the variable of interest. However, such approaches do not fully consider all the elements of the configuration. This will be illustrated with examples in Section 2.1. In fact, we will show that a unit can be influential according to a given configuration and yet not influential for another. This will illustrate the importance of using a detection and correction method that fully accounts for all characteristics of the configuration. As mentioned earlier, it is assumed that the edits and imputation steps are completed and that erroneous values have been corrected. The focus is therefore on true reported y-values that have a large impact on the sampling error.

The current production method that is used to detect influential values for the SHS relies on the contribution of a unit, which is defined as the product of the final calibrated weight and the expenditure (y-variable). More precisely, we first compute the contribution $w_i y_i$ for all $i \epsilon s_r$, where s_r is the set of respondent households, then all the contribution values are ordered and the ratio between each consecutive contribution is computed. If a given unit has a ratio exceeding a certain threshold δ, then the unit will be considered influential. Once a unit is influential, all other units with higher contributions will also be considered influential, regardless of the value of their ratios. As the weights are considered final at this stage, a unit i that is identified as being influential will have its y-value reduced so that it is no longer influential. To do this, the ratio of the closest non-influential unit, let's say unit j, is used to correct y_i. In particular, the new y-value for unit i is defined as follows:

$$\tilde{y}_i = \frac{ratio_j w_j y_j}{w_i y_i} * y_i$$

After correction, the new contribution of unit i is therefore

$$w_i \tilde{y}_i = ratio_j w_j y_j$$

The new ratio of unit i will be equal to $ratio_j$, which by choice of unit j is smaller than the threshold δ. This method relies on parameters that must be specified for every year of the survey. Typically, we have $\delta = 1.85$, and only the top 4% of units (in terms of contribution) are eligible to be corrected. Furthermore, if fewer than 25 households reported an expenditure for the variable y, no correction is applied. The idea behind this last constraint is that the method needs a sufficient sample size to be reliable. Each year, these parameters must be confirmed by ensuring that the method is correcting appropriate values. This type of verification is time consuming, and as mentioned above, such a method does not fully consider all aspects of the configuration. For these reasons, an alternative method was considered, one that relies on the concept of conditional bias and that will be described in the next section.

2. The Conditional Bias as a Measure of Influence

2.1 Definition of the Conditional Bias and Examples

In this section, we give an overview of the conditional bias approach for the treatment of influential values, as described in Beaumont et al. (2013). The conditional bias of a sampled unit i with respect to the estimator $\hat{\theta}$ is defined as:

$$B_{1i} = E_D\left[\,\hat{\theta} - \theta \,|\, I_i = 1\,\right]$$

It can be seen as the average of the sampling error over all samples containing i. The idea is to go through all possible samples containing i and taking the average of all the estimates produced by each sample. If this average is far away from the true population value, then unit i will have a large conditional bias. In practice, we only have access to one sample which implies that the conditional bias is unknown and must be estimated. From the definition of the conditional bias, one can see that all the features of the configuration come into play.

As an example, we can apply the conditional bias formula to the case where the Horvitz-Thompson estimator $\hat{\theta}^{HT} = \sum_{j\in s}\frac{1}{\pi_j}y_j$ is used to estimate the population total $\theta = \sum_{j\in U} y_j$.

$$B_{1i}^{HT} = E_D\left[\,\hat{\theta}^{HT} - \theta \,|\, I_i = 1\,\right] = E_D\left[\sum_{j\in s}\frac{1}{\pi_j}y_j - \sum_{j\in U} y_j \,|I_i = 1\right]$$

$$= E_D\left[\sum_{j\in U}\frac{1}{\pi_j}y_jI_j - \sum_{j\in U}y_j|I_i = 1\right]$$
$$= E_D\left[\sum_{j\in U}\left(\frac{1}{\pi_j}I_j - 1\right)y_j|I_i = 1\right]$$
$$= \sum_{j\in U}\left(\frac{1}{\pi_j}E_D[I_j|I_i = 1] - 1\right)y_j$$
$$= \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)y_j$$

where π_{ij} is the second-order inclusion probability of units i and j. From this example, we see that the sum is taken over all units in the population, and therefore the conditional bias is unknown in practice. A conditionally unbiased estimator for B_{1i}^{HT} can be given by

$$\hat{B}_{1i}^{HT} = \sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}}y_j$$

In other words, we have that $E_D\left[\hat{B}_{1i}^{HT}|I_i = 1\right] = B_{1i}^{HT}$. The expressions of B_{1i}^{HT} and $\hat{B}_{1i}^{HT}$ also illustrate the fact that probabilities of inclusion of order one and two are needed; this can become a complicated exercise for complex survey designs, such as the SHS. However, for simpler designs, the expression of the conditional bias can easily be found. For instance, here is the expression of the conditional bias of unit i for the estimator $\hat{\theta}^{HT}$ in the cases of a simple random sample without replacement (SRSWOR) and Poisson sampling, respectively:

SRSWOR: $B_{1i}^{HT} = \frac{N}{N-1}\left(\frac{N}{n} - 1\right)(y_i - \bar{Y})$

Poisson: $B_{1i}^{HT} = (w_i - 1)y_i$

Notice that the two expressions differ slightly, a fact highlighting the importance of considering the sample design when detecting influential values. Under a SRSWOR design, a unit will have a large conditional bias if its y-value is far away from the population average. Under a Poisson design, a unit will have a large conditional bias if its y-value is large or if its survey weight is large. Favre-Martinoz, Haziza and Beaumont (2016) extended the concept of conditional bias to calibration estimators. Calibration estimators can be seen as complex functions of estimated totals, and the authors used first-order Taylor expansions to approximate the conditional bias. This leads to an expression of the conditional bias that is very similar to B_{1i}^{HT} except that the y-values are replaced by calibration residuals. We will come back to this later, in Section 3.1, when we derive the expression of the conditional bias for the SHS.

2.2 Using the Conditional Bias as a Measure of Influence and to Define a Robust Estimator

Beaumont et al. (2013) demonstrated that under certain conditions, the sampling error of the Horvitz-Thompson estimator can be written as a sum of conditional biases:

$$\hat{\theta}^{HT} - \theta \cong \sum_{i\in s} B_{1i}^{HT} + \sum_{i\in U-s} B_{0i}^{HT}$$

where $B_{0i} = E_D\left[\, \hat{\theta} - \theta |\, I_i = 0 \,\right]$ is simply the conditional bias when unit i is excluded from the sample. This relationship can be extended to most estimators using linearization. Moreover, this relationship illustrates how the conditional bias is a natural measure of influence; indeed, the conditional bias of a unit can be seen as its contribution to the sampling error. It is therefore intuitive to want to reduce the conditional bias of a unit, in order to reduce the sampling error.

Furthermore, the sampling variance of $\hat{\theta}^{HT}$ can be written as a weighted sum of the conditional bias of all units in the population. According to Beaumont et al. (2013), we have the following result:

$$V_D\left(\hat{\theta}^{HT}\right) = \sum_{i\in U} y_i B_{1i}^{HT}$$

It is interesting to notice that there is a direct link between the concept of conditional bias (i.e. influence) and the variance. A sampled unit with a large conditional bias will contribute to inflate the variance of the estimator, and consequently will make $\hat{\theta}^{HT}$ unstable.

Moreover, the conditional bias can be used to define a robust estimator, as described in Beaumont et al. (2013). The authors show that a robust estimator, denoted $\hat{\theta}^{R}$, can be defined as $\hat{\theta}^R(c) = \hat{\theta} + \Delta_c$, where $\hat{\theta}$ is the non-robust estimator (e.g. Horvitz-Thompson or a calibration estimator) and Δ_c is a term that will depend on a tuning constant c and that will reduce the conditional bias of the most influential units for the estimator $\hat{\theta}$. More specifically, $\Delta_c = \sum_{i\in s}\{\psi_c(\hat{B}_{1i}) - \hat{B}_{1i}\}$where $\hat{B}_{1i}$ is the estimated conditional bias of unit i based on the estimator $\hat{\theta}$, and ψ_c is the Huber function of parameter c defined as $\psi_c(x) = sign(x) * \min(|x|, c)$. The function $\psi_c(x)$ is equal to x unless $|x|$ exceeds the threshold c. For instance, here is a graphic for $\psi_4(x)$:

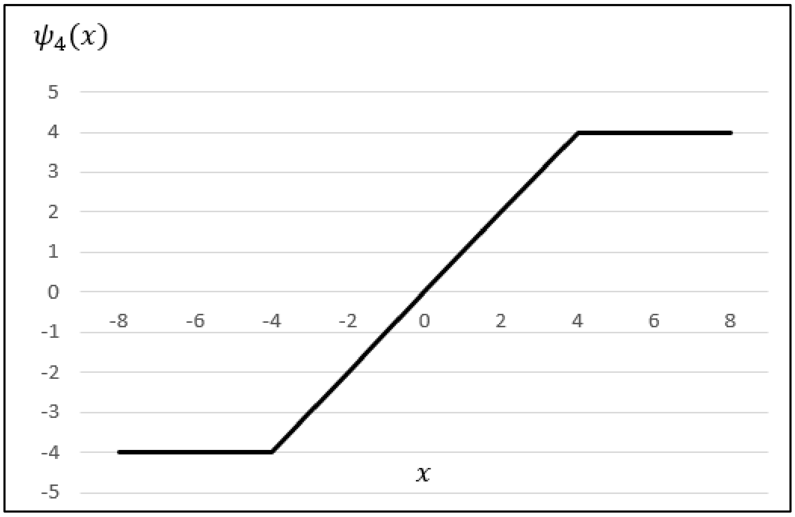

Figure 1: Huber function $\psi_4(x) = sign(x) * \min(|x|, 4)$.

The constant c will be the tuning parameter on which the robust estimator will depend. The value of this parameter will be chosen as to minimise the maximum value of the conditional bias of the robust estimator. Formally, we choose c so as to minimize $max\{|\hat{B}_{1i}^{R}(c)|; i \in s\}$ which ensures that the robust estimator $\hat{\theta}^{R}(c)$ is less sensitive to influential values than $\hat{\theta}$. Beaumont et al. (2013) showed that the optimal value of c which satisfies this constraint is such that $\Delta_{C_{opt}} = -\frac{1}{2}(\hat{B}_{min} + \hat{B}_{max})$, where $\hat{B}_{min}$ and $\hat{B}_{max}$ are the minimum and maximum values of the conditional bias of $\hat{\theta}$, respectively, over all units in the sample.

Beaumont et al. (2013) showed that the robust estimator can be written as the weighted sum of the modified y-values. More precisely, we have that $\hat{\theta}^{R}(c) = \sum_{i \in s} w_i \tilde{y}_i$ where $\tilde{y}_i = y_i - \phi_i \frac{\hat{B}_{1i}}{w_i}$ and $\phi_i = 1 - \frac{\psi_c(\hat{B}_{1i})}{\hat{B}_{1i}}$. Since $0 \leq \phi_i \leq 1$, we have that when the conditional bias of unit i is small (for a given value of c), then $\psi_c(\hat{B}_{1i}) = \hat{B}_{1i}$ and so $\phi_i = 0$, which implies that $\tilde{y}_i = y_i$.

Favre-Martinoz, Haziza and Beaumont (2015) showed through simulations that the efficiency of the robust estimator $\hat{\theta}^{R}$ as defined above is equal or superior to that of winzorisation estimators of orders 1, 2 and 3. It was therefore considered an interesting option for the SHS. In order to apply this theory to the SHS context, the conditional bias definition had to be applied to the complex two-stage design and the calibration estimator of the SHS. This will be described in the following section.

3. Applying the Conditional Bias to the Survey of Household Spending

3.1 Conditional Bias for a Two-stage Design followed by Unit Nonresponse

The goal of this paper is to show how the conditional bias approach was used to detect and correct influential values in the context of the SHS. In this section, we expand the conditional bias formula in the case of a two-stage design followed by unit nonresponse, which is the design of the SHS. The nonresponse must be considered as it is an additional phase of sampling. To start, we use the Horvitz-Thompson estimator of a total as an example, and later we will extend this to a calibration estimator.

First, if we assume 100% response rates, then we have that $\hat{\theta}^{HT} = \sum_{l \in s} \sum_{j \in s_l} \frac{1}{\pi_l \pi_{j|l}} y_{lj}$, where the first sum goes through all the selected primary sampling units (PSUs) and the second sum goes over all the second stage units (SSUs). In the case of the SHS, the PSUs are geographical areas, and SSUs are dwellings within the selected PSUs. For SSU i belonging to PSU g, we have that

$$B_{1i|g}^{HT} = \sum_{l \in U} \left(\frac{\pi_{gl}}{\pi_g \pi_l} - 1 \right) Y_l + \frac{1}{\pi_g} \sum_{j \in U_g} \left(\frac{\pi_{ij|g}}{\pi_{i|g} \pi_{j|g}} - 1 \right) y_{gj}$$

where π_l is the inclusion probability of PSU l , π_{gl} is the second order inclusion probability of PSUs g and l, Y_l is the total in PSU l (i.e. $Y_l = \sum_{j \in U_l} y_{lj}$), and y_{lj} is the y-value of SSU j in PSU l. The term $\pi_{j|g}$ is the first order inclusion probability of SSU j

conditional on PSU g being in the sample and $\pi_{ij|g}$ is the second order inclusion probability of SSUs i and j given that PSU g is selected in the first stage of the sample.

If we further assume that we are in the presence of unit nonresponse, we can use the following unbiased estimator of the total $\hat{\theta}_{NR} = \sum_{l\in s}\sum_{j\in s_{lr}} \frac{1}{\pi_l \pi_{j|l} p_{j|l}} y_{lj}$ where $p_{j|l}$ is the probability of response of SSU j in PSU l and s_{lr} is the set of responding SSUs from PSU l. Then the conditional bias of responding SSU i belonging to PSU g is given by:

$$B_{1i|g}^{NR} = \sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g \pi_l} - 1\right) Y_l + \frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right) y_{gj} + \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right) y_{gi}$$

We notice that the expression of $B_{1i|g}^{NR}$ can be obtained from $B_{1i|g}^{HT}$ by adding to it a third term which takes into account the nonresponse as an additional sampling phase. If instead of the Horvitz-Thompson estimator, we had used a calibration estimator, the expression of the conditional bias would change, just as it was mentioned in Section 2.1. In particular, if we calibrate the survey weights adjusted for nonresponse, we obtain the calibration estimator $\hat{\theta}_{Cal} = \hat{\theta}_{NR} + \left(X - \hat{X}_{NR}\right)^T \hat{B}$, where X is the matrix of calibration variables. Using this calibration estimator, the conditional bias will now depend on the calibration residuals instead of the y -values.

$$B_{1i|g}^{Cal} = \sum_{l\in U}\left(\frac{\pi_{gl}}{\pi_g \pi_l} - 1\right) E_l + \frac{1}{\pi_g}\sum_{j\in U_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right) e_{gj} + \frac{1}{\pi_g}\frac{1}{\pi_{i|g}}\left(\frac{1}{p_{i|g}} - 1\right) e_{gi}$$

where $e_{gi} = y_{gi} - x_{gi}^T B$, $E_l = \sum_{j\in U_l} e_{lj}$, and B is the population parameter estimated by $\hat{B}$. An unbiased estimator for the conditional bias is given by:

$$\hat{B}_{1i|g}^{Cal} = \sum_{l\in s}\frac{\pi_{gl} - \pi_g \pi_l}{\pi_{gl}\pi_l}\sum_{j\in s_{lr}}\frac{1}{\pi_{j|l} p_{j|l}}\hat{e}_{lj} + \sum_{j\in s_{gr}}\frac{1}{p_{j|g}}\frac{\pi_{ij|g} - \pi_{i|g}\pi_{j|g}}{\pi_{ij|g}\pi_{j|g}}\hat{e}_{gj}$$
$$+ \left(\frac{1}{p_{i|g}} - 1\right)\hat{e}_{gi}$$

where $\hat{e}_{gi} = y_{gi} - x_{gi}^T \hat{B}$.

3.2 Necessary Assumptions to Evaluate the Conditional Bias for SHS

In the previous section, we developed the expression of the estimator of the conditional bias in the case of a two-stage design, followed by nonresponse, and for which a calibration estimator is used to estimate the total. As can be seen from this expression, inclusion probabilities of order one and two are required for both stages, as well as the probabilities of responding to the survey and the calibration residuals. In order to obtain these quantities in the context of the SHS, some assumptions were necessary.

For the first stage, geographical areas (PSUs) are selected according to the Rao-Hartley-Cochran proportional to size sample design (Rao, Hartley and Cochran, 1962). This design consists in randomly allocating each PSU of a given stratum into groups. As the SHS design is coordinated with the Labour Force Survey design, the same random groups are used for both surveys and six groups are used per stratum. Once the PSUs are allocated into such groups, one PSU is selected within each group according to a proportional to size design where the number of dwellings is the size measure.

For the Rao-Hartley-Cochran scheme hereby defined, it can be very difficult to compute the inclusion probability of order two. For simplicity, we assumed that a Poisson design had been used instead to select the PSUs. In particular, we assumed that the PSUs within the same stratum were selected independently from one another. This is not quite accurate because of the random allocation into groups, but conditional on the group allocation, it is accurate to consider the selection of PSUs as independent. Therefore, for PSUs g and$l$ belonging to the same stratum, we assumed that $\pi_{gl} = \pi_g \pi_l$ (if $g \neq l$) and $\pi_{gl} = \pi_g$ (if $g = l$).

In the second stage, a systematic sample of dwellings (SSUs) is selected within each of the selected PSU. Since it is possible for two dwellings i and j belonging to the same PSU to have second-order selection probability equal to zero, we had to assume that a simple random sample without replacement had been used instead of a systematic sample. Therefore, for SSUs i and j belonging to the same PSU g, we assumed that

$$\pi_{ij|g} = \begin{cases} \dfrac{n_g}{N_g} & if \quad i = j \\ \dfrac{n_g}{N_g} \dfrac{n_g - 1}{N_g - 1} & if \ i \neq j \end{cases}$$

where n_g is the number of dwellings selected in PSU g and N_g is the total number of dwellings in PSU g.

Finally, the nonresponse phase must be considered as a subsampling step, and hence selection probabilities (i.e. response probabilities) appear in the conditional bias expression. In the weighting process of the SHS, nonresponse adjustments are computed using a logistic regression. Auxiliary variables are used in the logistic regression model to predict the response probability of a given household. Then, households of similar response probabilities are grouped together and a nonresponse adjustment is computed within each nonresponse adjustment group. For simplicity, we assumed that the inverse of these nonresponse adjustments corresponded to the true probability of response. Alternatively, we could have considered using the estimated response probabilities, which would have required a linearization exercise to obtain the expression of the conditional bias (as shown in Favre-Martinoz et al. (2016)). Furthermore, we assumed that the nonresponse mechanism was independent across households (i.e. $p_{ij} = p_i p_j, i \neq j$).

3.3 Results

Data from the SHS 2015 was used to test both methods; the current method used in production and the conditional bias method yielding robust estimators. To compare the performance of both methods, the absolute relative bias was computed for major expenditure categories; shelter and transportation expenditures, as well as total consumption. Both methods correct the y-value of influential units and create a corrected value $\tilde{y}$ to be used for weighted estimates. We consider the original total estimate $\sum_{i \epsilon s_r} w_i y_i$ to be the unbiased benchmark, where w_i is the weight after calibration. In other words, the absolute relative bias is given by the following formula:

$$\frac{\left|\sum_{i\epsilon s_r}\tilde{y}_i w_i - \sum_{i\epsilon s_r} y_i w_i\right|}{\left|\sum_{i\epsilon s_r} y_i w_i\right|} \times 100.$$

Results are provided in table 1 below for estimates at the national level (10 provinces combined).

Table 1. Absolute Relative Bias of Robust Method and Production Method for Three Major Expenditure Categories (all 10 provinces combined, SHS 2015)

Expenditure Category	Robust Method	Production Method
Shelter	1.16%	0.97%
Transportation	1.20%	0.16%
Total Consumption	1.62%	1.23%

First, we note that the absolute relative bias for both methods is small (less than 2% for all three expenditure categories). This shows that both methods are not excessively correcting the y-values. Second, the results show that the production method yields smaller absolute relative bias than the robust method. Note, however, that the production method relies on subjective parameters which can change from year to year; therefore, one can imagine a scenario where different parameters would yield different values of absolute relative bias. Another important detail about the production method is that it is only applied when there is a sufficient number of reporting households (this threshold is normally set to 25 reporting households in the estimation domain, which is the province). This constraint was not applied to the robust method, and this might explain why the robust method is correcting the y-values to a greater extent.

The impact of both methods on the variance of the expenditure estimates was not measured. This is because it is currently unknown how to estimate the variance of the robust estimator. Indeed, as mentioned in Beaumont et al. (2013), it is not obvious how to properly bootstrap $\hat{B}_{min}$ and $\hat{B}_{max}$ which appear in the $\Delta_{C_{opt}}$ term of the robust estimator. More research will be necessary in order to estimate the mean squared error of the robust estimator. A simplified approach could be to apply the correction to the y-values, as defined by the conditional bias approach, to all bootstrap replicates. The impact of this simplification on the variance estimates would need to be determined, perhaps through simulations.

4. Conclusion

In this paper, we described how the conditional bias approach can be used to define robust estimators (as shown in Beaumont et al. (2013)) and we derived the expression of the estimator of the conditional bias in the context of a two-stage design followed by nonresponse where calibration estimators are used. This corresponds to the sample design of the Survey of Household Spending, a survey in which the issue of influential values must be addressed. The current production method to detect and correct influential values was compared to the new method that relies on robust estimators. It was found that both methods yield small absolute relative biases. The advantage of the robust method is that it is transparent and it leads to a streamlined process; the parameter of the method (the tuning constant) is automatically determined, which improves the timeliness compared to the production method which relies on several manual adjustments and verifications. As described in this paper, the robust estimator can be difficult to compute in practice, especially when the sample design is complex as is the case with the SHS. Assumptions were necessary to estimate the conditional bias for the SHS, and these assumptions may have prevented us from fully capitalizing on the benefits of this new method. Another drawback of using the robust estimator is that its variance is currently unknown. Future work could be done in this area of research in order to determine a way to estimate the variance of the robust estimator and hence enable the comparison of its efficiency to the estimator used in production.

Acknowledgements

The authors would like to thank Jean-François Beaumont and David Haziza for their assistance and helpful advice during this project.

References

Beaumont, J.-F., Haziza, D., Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. Biometrika, 555-569.

Favre-Martinoz, C., Haziza, D., Beaumont, J.-F. (2015). A method of determining the winzorisation threshold, with an application to domain estimation. Survey Methodology, Statistics Canada, Volume 41, Number 1, June 2015.

Favre-Martinoz, C., Haziza, D., Beaumont, J.-F. (2016). Robust Inference in Two-Phase Sampling Designs with Application to Unit Nonresponse. Scandinavian Journal of Statistics, Theory and Applications. DOI: 10.1111/sjos.12226

Rao, J.N.K, Hartley, H.O., Cochran, W.G. (1962). On a Simple Procedure of Unequal Probability Sampling without Replacement. Journal of the Royal Statistical Society. Series B (Methodological). Vol. 24, No. 2, p.482-491.