This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Outlier Treatment

Contents

General section

1. Summary

In business surveys, the distribution of variables is often highly skewed, resulting in sample observations that differ substantially from the majority of observations in the sample. The literature refers to these units as outliers.

Outliers can be *representative* (representing other population units similar in value to the observed outliers) or *non-representative* (unique in the population). Here we will consider only the case of representative outliers, i.e., correct values representing other units in the population. Since representative outliers affect the variability of the standard estimators (such as: Horvitz-Thompson or Generalised regression estimators (GREG)), an appropriate way of handling them is required.

The objective of outlier treatment is to make estimates for the population coherent with the real parameters for the population. This means that outlier treatment should be always a trade-off between variance and bias. For small samples, variance is usually the dominating factor in the MSE. On the other hand, bias dominates when the sample size is large.

The module describes one frequently applied estimation method used to reduce the impact of outlying units: Winsorisation. The general idea of Winsorisation involves modifying the outlying observation so that it has less impact on the estimate of a parameter. The effectiveness of the Winsor estimator in terms of its resistance to unusually large residuals depends on the choice of cut-off values, therefore the methods used to estimate the robust regression parameters and the bias parameters need to estimate cut-off values. The cut-offs are optimal only at the level at which estimates are being conducted. The Winsor estimator is easy to implement, but it performs best under models (used for estimating robust regression parameters) that are only moderately robust. Winsorisation can be applied to a large class of estimators (GREG estimators, model-based regression estimators, ratio estimators) and involves modifying their standard forms. This results in estimates with acceptable bias and a smaller variance than that of standard forms, non-Winsorised estimators. We can observe the bias-variance trade-off at the low level of estimation but aggregated Winsorised estimates have large biases, resulting in less precision compared to standard aggregated estimates.

2. General description of the method

In business surveys target variables tend to be highly skewed and populations may contain a number of extreme values, the so-called outliers. Although outliers are extreme, they need not necessarily be incorrect but are an integral part of each survey population and cannot be dismissed in the analysis.

According to Hawkins, “an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980). In the statistical literature outliers are observations that differ substantially from most of the observations in the sampled and the unsampled parts of the population. Outliers may be extreme big values or extreme small values. We can distinguish *large outliers*, if the values are extremely large than the other values of the “normal” units, or *small outliers*, if the values are extremely smaller than the other values of the “normal” units.

Two distinct types of outliers can be defined: "*y*-outliers" or "outliers in the *y*-direction" and "*x*-outliers" or "outliers in the *x*-direction" (Rousseeuw and Leroy, 2003) where *y-values* and *x-values* are the study variable and an auxiliary variable, respectively.

"*Y*-outliers" denote the *y* values of a few sample units that are very distant from the *y* values of other sample units. Another class of outliers comprises the *x* values of a few sample units that are very distant from the *x*-values of other sample units. These are "*x*-outliers". They can have a substantial impact on the stability of the overall sample estimate because of their so-called "leverage" (Bergdahl et al., 1999).

Some authors (Chambers, 1986; Eltinge and Cantwell, 2006) classify outliers into three groups. The first are *representative outlier* values, which represent other population units similar in value to the observed outliers. These are correctly measured sample values that are outlying relative to the rest of the sample data and we cannot assume that similar values do not exist in the non-sampled part of the survey population. The second group consists of *non-representative outlier* values, which are unique in the population (in the sense that there is no other unit like them) (Chambers, 1986). The third group comprises gross measurement errors, which are outlying observations that are not true values.

Here we will not consider gross errors in the sampled data, caused by deficiencies in the survey processing (e.g., miscoding). Such errors are corrected during the data editing process (Eltinge and Cantwell, 2006).

Since outliers usually have a huge impact on estimates, outlier detection and their treatment are important elements of statistical analysis. This is true especially when estimation is carried out at a low level of aggregation. In the case of small sample sizes, outliers can affect variance. Even if the sample size is large, the influence of an outlier can significantly increase the variance resulting in a decreased efficiency of estimation. Dealing with outliers has two aspects: the first one involves identifying outlying observations in an objective way, and the second one focuses on ways of handling them to reduce their effect on survey estimates.

While non-representative outliers can be treated by post-stratification, representative outliers should be handled in the survey estimation process, by the use of outlier resistant or robust estimation procedures (Ren and Chambers, 2002).

In this module we consider only representative outliers, in other words, any extreme values that represent other true observations in the population.

There are three main methods of dealing with outliers in a finite population, apart from removing them from the dataset (Cox et al., 1995):

1. reducing the weights of outliers (trimming weight),
2. changing the values of outliers (Winsorisation, trimming),
3. using robust estimation techniques such as M-estimation.

Weight trimming reduces large weights to a fixed cut-off value and adjusts weights below this value to maintain the untrimmed weight sum, reducing variability at the cost of introducing some bias (Elliott, 2007). The literature mentions various approaches to determine cut-off points at which to trim weights. Most standard methods are ad hoc in that they do not use the data to optimise bias-variance trade-offs.

According to Potter (1990), a weight should be trimmed at the point where the loss of precision due to a large weight is larger than the bias introduced by trimming the weight. It can be said that the general approach involves reducing the survey weight associated with that observation (Chambers, 1996; Detlefsen, 1992; Elliott and Little, 2000; Potter, 1988, 1993; Theberge, 2000; Zaslavsky et al., 2001). In many business surveys it is a relatively common practice to set the survey weight equal to one. One could say that the identified outlier is a "non-representative" outlier (Eltinge and Cantwell, 2006). In some cases setting a weight equal to one may be viewed as the limiting case of more refined adjustment procedure like Winsorisation or M-estimation (Eltinge and Cantwell, 2006). Winsorisation is frequently used in business surveys, so it is presented below in more detail. The general idea of Winsorisation is that if an observation exceeds a pre-set cut-off value, then the observation is replaced by that cut-off value or by a modified value closer to the cut-off value.

2.1 Winsorisation

In business statistics, which are characterised by skewed distributions, GREG estimation procedures may provide unsatisfactory results. One of the most popular methods suggested in the literature consists in modifying values in the sample so that the estimator becomes robust and is not affected by large residuals (Kokic and Bell, 1994; Chambers, 1996; Chambers et al., 2000; Rivest and Hidiroglou, 2004; Dehnel and Gołata, 2010). This approach is exemplified by Winsor estimation, which was applied for the first time in a survey conducted by Searls (1966). Winsorisation involves identifying cut-off (thresholds) values. Sample observations whose values lie outside certain pre-set cut-off values are transformed in order to make them closer to the cut-off value.

Winsorisation may be one-sided or two-sided. One-sided Winsorisation adjusts influential values deemed to be too large. Two-sided Winsorisation adjusts influential values deemed to be both too large and too small.

Cut-off values are derived in a way that approximately minimises the MSE of estimates. All sampled units are divided into two (or three) groups. One group contains typical observations, which are left unmodified, the other one(s) contain(s) observations regarded as (large or small) outliers. The classification is made on the basis of one (if outliers are not divided into large and small) or two (when, on the contrary, large and small outliers are distinguished) pre-set cut-off values. Then, values of the study variable outside the cut-off values are transformed so that they are no longer regarded as outliers. It should be stressed, however, that the modified values are artificial and may sometimes be unacceptable. As a result of Winsor estimation, we obtain a modified sample, in which untypical observations have been replaced with typical ones. Further calculations are conducted for the modified sample. Any kind of estimation can be used at this stage. Here, GREG estimation is illustrated.

The Winsorised estimator of the population total is defined $\hat{Y}_{win} = \sum_{i \in s} w_i y_i^*$ where y_i^* is the modified value of the study variable.

First, let us consider the case when we have only large outliers.

Two types of Winsorisation can be applied in the treatment of outliers. Winsorised Type I estimator is based on an arbitrary assumption whereby any outliers exceeding a pre-set cut-off value *K* are always replaced by that value *K*:

$$y_i^* = K \text{ if } y_i > K \text{ and } y_i^* = y_i \text{ otherwise.}$$

On the contrary, with a Type II estimator, as the GREG weight $\tilde{w}_i$ decreases, the contribution of the observed values of the outliers increases – that is the modified value of the study variable "approaches" the value of the outlier, i.e., the real value of the variable.

Under Type II Winsorisation:

$$y_i^* = \left(\frac{1}{w_i}\right) y_i + \left(1 - \frac{1}{w_i}\right) K \text{ if } y_i > K \text{ and } y_i^* = y_i \text{ otherwise.}$$

The use of Winsor estimation reduces estimator variance, while, at the same time, it may introduce bias. However, if cut-off values are chosen appropriately, the decline in variance is big enough to offset the bias of MSE (Hedlin, 2004).

The main difficulty then lies in the choice of cut-off values for dividing observations in the sample. The optimum selection has a strong effect on estimation quality.

The Winsor estimator, with GREG estimation, can be expressed as:

$$\hat{Y}_{win} = \sum_{i \in s} \tilde{w}_i y_i^* = \sum_{i \in s} w_i g_i y_i^* \tag{1}$$

where, in the presence of outliers, modified values of the study variable y_i^* are calculated in the following manner (Gross et al., 1986):

$$y_i^* = \begin{cases} \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{Ui} & \text{if} \quad y_i > K_{Ui} \\ y_i & \text{if} \quad K_{Li} \le y_i \le K_{Ui} \\ \left(\frac{1}{\tilde{w}_i}\right) y_i + \left(1 - \frac{1}{\tilde{w}_i}\right) K_{Li} & \text{if} \quad y_i < K_{Li} \end{cases} \tag{2}$$

$$g_i = \left(1 + x_i' \left(\sum_{i \in s} w_i x_i x_i'\right)^{-1} \left(t_x - \sum_{i \in s} w_i x_i\right)'\right) \tag{3}$$

where:

$U = \{1, \ldots. i, \ldots. N\}$ - target population of size *N*;

$s\,(s \subseteq U)$ - sample;

$\tilde{w}_i = w_i g_i$;

$w_i = 1/\pi_i$ - sampling weights;

g_i - weights dependent on the value of a vector of auxiliary variables for sampled units;

$x_i = (x_{1i}, \ldots, x_{ki}, \ldots, x_{Ki})'$ - vector of auxiliary variables;

$t_x = \sum_{i \notin U} x_i$ - population total;

K_{Ui} - upper cut-off value; K_{Li} - lower cut-off value.

Based on formula (1) it can be assumed that a unit drawn into the sample is regarded as an element representing $(\tilde{w}_i - 1)$ non-sampled units. Hence, according to formula (2), an observation regarded as an outlier contributes its unweighted values, while the non-sampled units, represented by the remainder of the weight $(\tilde{w}_i - 1)$, contribute pre-set upper or lower cut-off values.

Cut-off values are calculated to minimise MSE of Winsorised estimator under the model (Preston and Mackin, 2002):

$$K_{Ui} = \mu_i^* - \frac{B_U}{(\tilde{w}_i - 1)} \tag{4}$$

$$K_{Li} = \mu_i^* - \frac{B_L}{(\tilde{w}_i - 1)} \tag{5}$$

where:

$\mu_i^* = E(Y_i^* \mid x_i)$ - conditional expectation under the assumed regression model;

$B_U = E\left[\hat{Y}_{winU} - \hat{Y}_{DIR}\right]$ - bias of $\hat{Y}_{winU}$;

$B_L = E\left[\hat{Y}_{winL} - \hat{Y}_{DIR}\right]$ - bias of $\hat{Y}_{winL}$;

$\hat{Y}_{winU}$ - Winsor estimator of the population total when only upper Winsorisation is performed;

$\hat{Y}_{winL}$ - Winsor estimator of the population total when only lower Winsorisation is performed.

When Winsorisation is mild and reasonably symmetric, being μ_i^* difficult to estimate, we can replace μ_i^* with μ_i . Then the approximately optimal cut-offs are (Preston and Mackin, 2002):

$$K_{Ui} = \mu_i - \frac{B_U}{(\tilde{w}_i - 1)} = \mu_i + \frac{G}{(\tilde{w}_i - 1)} \tag{6}$$

$$K_{Li} = \mu_i - \frac{B_L}{(\tilde{w}_i - 1)} = \mu_i + \frac{H}{(\tilde{w}_i - 1)} \tag{7}$$

Under the assumption $\mu_i = \hat{\mu}_i = \hat{\beta} x_i$ (Preston and Mackin, 2002), cut-off values are estimated based on the following formulas:

$$\hat{K}_{Ui} = \hat{\mu}_i - \frac{B_U}{(\tilde{w}_i - 1)} = \hat{\mu}_i + \frac{G}{(\tilde{w}_i - 1)} \qquad \text{where } G = -B_U \tag{8}$$

$$\hat{K}_{Li} = \hat{\mu}_i - \frac{B_L}{(\tilde{w}_i - 1)} = \hat{\mu}_i + \frac{H}{(\tilde{w}_i - 1)} \qquad \text{where } H = -B_L \tag{9}$$

where $\hat{\mu}_i = \hat{\beta} x_i$ - a robust estimate of regression parameter μ_i (see below).

In order to estimate the bias B_U under Winsorisation we can use the Kokic and Bell approach (1994). According to that approach, the value of B_U can be calculated by solving the equation:

$$G - E\left[\sum_{i \in s} \max\{D_i - G, 0\}\right] = 0 \tag{10}$$

where $D_i = (Y_i - \mu_i^*)(\tilde{w}_i - 1)$ are weighted residuals. Assuming $\hat{\mu}_i$ is a robust estimate of parameter μ_i, we obtain $\hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1)$.

We can write the function $\psi_U\left(\hat{D}_{(k)}\right)$ (Kokic and Bell, 1994):

$$\psi_U\left(\hat{D}_{(k)}\right) = \hat{D}_{(k)} - \sum_{i \in s} \max\left\{\hat{D}_i - \hat{D}_{(k)}, 0\right\} = (k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)} \tag{11}$$

where:

(k) - a number assigned to the unit drawn into the sample after ordering all units in the sample according to non-ascending estimated residuals $\hat{D}_i$: $\hat{D}_{(1)} \geq \hat{D}_{(2)} \geq \ldots \geq 0 \geq \ldots$.

By solving $\psi_U(G) = 0$ one can obtain the value of G. In practice, since it is difficult to find the right solution of the equation, two methods are proposed. According to the first one, G is estimated using the formula:

$$\hat{G} = \frac{1}{\left(k^* + 1\right)} \sum_{j=1}^{k^*} \hat{D}_{(j)} \tag{12}$$

where k^* is the last value of k for which the value of $\psi_U\left(\hat{D}_{(k)}\right)$ is non-negative.

The second approach involves using linear interpolation between $\frac{1}{\left(k^*+1\right)} \sum_{j=1}^{k^*} \hat{D}_{(j)}$ and $\frac{1}{\left(k^*+2\right)} \sum_{j=1}^{k^*+1} \hat{D}_{(j)}$.

Then, $\hat{G}$ can be expressed as (Preston and Mackin, 2002):

$$\hat{G} = \frac{\psi_U\left(\hat{D}_{(k^*+1)}\right)\left[\frac{1}{\left(k^*+1\right)} \sum_{j=1}^{k^*} \hat{D}_{(j)}\right] - \psi_U\left(\hat{D}_{(k^*)}\right)\left[\frac{1}{\left(k^*+2\right)} \sum_{j=1}^{k^*+1} \hat{D}_{(j)}\right]}{\left(\psi_U\left(\hat{D}_{(k^*+1)}\right) - \psi_U\left(\hat{D}_{(k^*)}\right)\right)} \tag{13}$$

The value of H can be computed similarly. Estimates of weighted residuals $\hat{D}_i = (Y_i - \hat{\mu}_i)(\tilde{w}_i - 1)$ are arranged in ascending order $\hat{D}_{[1]} \leq \hat{D}_{[2]} \leq \ldots \leq 0 \leq \ldots$. Function $\psi_L\left(\hat{D}_{[m]}\right)$ can be written as:

$$\psi_L\left(\hat{D}_{[m]}\right) = \hat{D}_{[m]} - \sum_{i \in s} \min\left\{\hat{D}_i - \hat{D}_{[m]}, 0\right\} = (m+1)\hat{D}_{[m]} - \sum_{l=1}^{m} \hat{D}_{[l]} \tag{14}$$

where:

$[l] = [1]..[m]$ - a number assigned to the unit drawn into the sample after ordering all units in the sample by estimated residuals $\hat{D}_i$.

The value $\hat{H}$ can thus be evaluated as (cf. formula (15)) (Preston and Mackin, 2002):

$$\hat{H} = \frac{\psi_L\left(\hat{D}_{[m^{**}+1]}\right)\left[\frac{1}{[m^{**}+1]}\sum_{l=1}^{m^{**}}\hat{D}_{[l]}\right] - \psi_L\left(\hat{D}_{[m^{**}]}\right)\left[\frac{1}{[m^{**}+2]}\sum_{l=1}^{m^{**}+1}\hat{D}_{[l]}\right]}{\left(\psi_L\left(\hat{D}_{[m^{**}+1]}\right) - \psi_L\left(\hat{D}_{[m^{**}]}\right)\right)} \tag{15}$$

where m^{**} is the last value of m for which the value of $\psi_U\left(\hat{D}_{[m]}\right)$ is non-positive.

In order to estimate cut-off values $\hat{K}_{Ui}$ and $\hat{K}_{Li}$, in addition to the above bias parameters $G = -B_U$ and $H = -B_L$ it is necessary to compute $\hat{\mu}_i = \hat{\beta} x_i$ which is an estimate of μ_i^*. For this purpose, robust regression methods can be used. Those recommended in the literature (Preston and Mackin, 2002) include: *Trimmed least squares (TLS), Trimmed least absolute value (LAV), Sample Splitting, Least median of squares (LMS).*

The method of *Trimmed least squares (TLS)* involves first fitting an Ordinary Least Squares (OLS) regression model to minimise the function:

$$F = \sum_{i \in s}\left(y_i - \beta^T x_i\right)^2 \tag{16}$$

Then fitted values are calculated, and then residuals. In the second step, units with the largest positive and negative residuals are removed. As a rule, the sample is reduced by about 5%. Finally, a new regression model is fitted to the reduced sample in order to estimate the value of μ_i^*. One advantage of the TLS is that it is quick to run and simple.

Another method used in robust regression is *Trimmed least absolute value (LAV).* It consists in fitting a regression model to minimise the function:

$$F = \sum_{i \in s}\left|y_i - \beta^T x_i\right| \tag{17}$$

After evaluating fitted values and residuals, as is the case in the TLS method, units with the largest positive and negative residuals are removed. A new regression model is fitted to the reduced sample. It is expected that the *LAV* method is a more robust regression model than the *TLS* technique because large residuals which are not squared have less influence on the regression parameters.

Another example of robust regression is *Sample Splitting Technique* based on Ordinary Least Squares (OLS). It is applied to a dataset that has been randomly split into two halves. A regression model is fitted to each half of the data while the residuals are calculated using the model applied to the half of the data that was not used to fit the model. Then, after merging the data, units with the largest positive and negative residuals are removed. The process is repeated until a certain percentage of data has been deleted. The *SS* technique is expected to be more robust than TLS because the residuals used to remove the 'outlier' units are not calculated from a regression model that has been generated using these 'outlier' units.

The list of robust regression techniques cannot be complete without the *Least median of squares (LMS)* technique. It was described by Rousseeuw and Leroy (2003). It resembles the bootstrap method. It involves drawing subsamples of size $n - 1$ from a sample of size n using simple random sampling with replacement. For each subsample trial regression model parameters are calculated and then their squared residuals, which are used to calculate the median. The model with the smallest median of squared residuals is selected. The *LMS* technique should be more robust than TLS because an OLS regression model is fitted in the absence of "outlier" units, without totally removing these ‘outlier’ units.

3. Preparatory phase

4. Examples – not tool specific

5. Examples – tool specific

6. Glossary

For definitions of terms used in this module, please refer to the separate “Glossary” provided as part of the handbook.

7. References

Bergdahl, M., Black, O., Bowater, R., Chambers, R., Davies, P., Draper, D., Elvers, E., Full, S., Holmes, D., Lundqvist, P., Lundström, S., Nordberg, L., Perry, J., Pont, M., Prestwood, M., Richardson, I., Skinner, C., Smith, P., Underwood, C., and Williams, M. (1999), *Model Quality Report in Business Statistics, Volume I, Theory and Methods for Quality Evaluation.* http://users.soe.ucsc.edu/~draper/bergdahl-etal-1999-v1.pdf

Chambers, R. L. (1986), Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association* **81**, 1063–1069.

Chambers, R. L. (1996), Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3–32.

Chambers, R., Dorfman, A. H., and Wehrly, T. E. (1993), Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. *Journal of the American Statistical Association* **88**, 268–277.

Chambers, R., Kokic, P., Smith, P., and Cruddas, M. (2000), Winsorization for Identifying and Treating Outliers in Business Surveys. *Proceedings of the Second International Conference on Establishment Surveys (ICES II)*, 687–696.

Chambers, R., Brown, G., Heady, P., and Heasman, D. (2001), Evaluation of Small Area Estimation Methods – an Application to Unemployment Estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: a Methodological Perspective.*

Chambers, R. L, Falvey, H., Hedlin, D., and Kokic P. (2001a), Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics* **17**, 527–544.

Cox, B. G., Binder, A., Chinnappa, N. B., Christianson, A., Colledge, M. J., and Kott P. S. (eds.) (1995), *Business Survey Methods*. John Wiley & Sons.

Dehnel, G. and Gołata, E. (2010), On some robust estimators for Polish Business Survey. *Statistics in Transition* - new series **11**, number 2, Warszawa 2010. s. 287-312 (Central Statistical Office and Polish Statistical Association), 58–71, Summ. - Bibliogr. ISBN 978-83-7027-431-3.

Detlefsen, R. (1992), A Weight Adjustment Technique. Internal Memorandum, Bureau of the Census.

Elliott, M. (2007), Bayesian weight trimming for generalized linear regression models. *Survey Methodology* **33**, 23–34.

Elliott, M. R. and Little, R. J. A. (2000), Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics* **16**, 191–209.

Eltinge, J. and Cantwell, P. (2006), Outliers and Influential Observations in Establishment Surveys. Paper prepared for presentation to the Federal Economic Statistics Advisory Committee (FESAC), 09-06-2006.

Gross, W. F., Bode, G., Taylor, J. M., and Lloyd-Smith, C. W. (1986), Some finite population estimators which reduce the contribution of outliers. *Proceedings of the Pacific Statistical Conference, 20–24 May 1985, Auckland, New Zealand.*

Hawkins, D. (1980), *Identification of Outliers*. Chapman and Hall.

Hedlin, D. (2004), Business Survey Estimation. R&D, Sweden.

Kokic, P. N. and Bell, P. A. (1994), Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* **10**, 419–435.

Potter, F. J. (1988), Survey of Procedures to Control Extreme Sampling Weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453–458.

Potter, F. (1990), A Study of Procedures to Identify and Trim Extreme Sample Weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 225–230.

Potter, F. J. (1993), The Effect of Weight Trimming on Nonlinear Survey Estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758–763.

Preston, J. and Mackin, C. (2002), Winsorization for Generalised Regression Estimation. Paper for the Methodological Advisory Committee, November 2002, Australian Bureau of Statistics.

Ren, R. and Chambers, R. L. (2002), Outlier robust imputation of survey data via reverse calibration. S3RI Methodology Working Paper M03/19.

Ren, R. and Chambers, R. L. (2002a), Outlier Robust Methods: Outlier Robust Estimation and Outlier Robust Imputation By Reverse Calibration. Report for Euredit, http://www.cs.york.ac.uk/euredit/.

Rivest, L.-P. and Hidiroglou, M. (2004), Outlier treatment for disaggregated estimates. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Rousseeuw, P. and Leroy, A. (2003), *Robust Regression and Outlier Detection*. John Wiley & Sons.

Searls, D. T. (1966), An estimator which reduces large true observations. *Journal of the American Statistical Association* **61**, 1200–1204.

Theberge, A. (2000), Calibration and Restricted Weights. *Survey Methodology* **26**, 99–107.

Zaslavsky, A. M., Schenker, N. and Belin, T. R. (2001), Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey.

Specific section

8. Purpose of the method

The module describes one estimation method frequently applied in business surveys, used to identify and handle outliers: Winsorisation. The general idea of Winsorisation is that if an observation exceeds a pre-set cut-off value, then the observation is replaced by that cut-off value or by a modified value closer to the cut-off value. As a result of Winsor estimation, we obtain a modified sample, in which untypical observations have been replaced with typical ones. The impact of outlying units is reduced. Further calculations are conducted for the modified sample. Any kind of estimation can be used at this stage.

9. Recommended use of the method

1. The method presented in this module is recommended for use in the case when the study variable(s) are highly skewed and several auxiliary variables that can be used to improve estimation including outliers. Such a situation is common in business statistics. The growing use of auxiliary information from administrative registers and the need to substantially reduce sample sizes or to produce more effective estimates has increased the importance of recognising and dealing with the data problem.
2. It is particularly suited to sample survey estimation. It can be used for various estimators (here, GREG estimation is illustrated) and sampling schemes.
3. In the case of stratified random sampling, the use of Winsor estimator can reduce the impact of outliers on stratum estimates while the population estimates remain unchanged (Rivest and Hidiroglou, 2004).
4. The method is flexible because the cut-offs can be chosen to suit the situation. It is simple to implement for applications with multiple variables and estimates.

10. Possible disadvantages of the method

1. The one-sided Winsorisation can introduces a negative bias, which can result in inconsistent estimates.
2. The values modified by Winsor estimator are artificial and may sometimes be unacceptable.

11. Variants of the method

1. There are three main methods of dealing with outliers in a finite population (Cox et al., 1995):
   - reducing the weights of outliers (trimming weight);
   - changing the values of outliers (Winsorisation, trimming);
   - using robust estimation techniques such as M-estimation.

   Winsorisation is most frequently used in business surveys. Two types of Winsorisation can be distinguished. The difference between them consists in the treatment of outliers.

Winsorised Type I estimator is based on an arbitrary assumption whereby any outliers exceeding a preset cut-off value K are always replaced by that value K.

In the case of Type II estimator, as weight $\tilde{w}_i$ decreases, the contribution of outliers increases – the modified value of the study variable "approaches" the value of the outlier, i.e., the real value of the variable.

Winsorisation cut-offs can be chosen on different levels, e.g.:

- specifying a cut-off value for observations by stratum;
- specifying an individual cut-off value for each observation.

12. Input data

1. The input data set has to contain individual information for all units in the sample. The input data set can contain information coming from auxiliary sources, e.g., administrative register. Specific software (e.g., SAS) may be based on different structures of the input data set in the procedure of robust estimation.

13. Logical preconditions

1. Missing values
    1.
2. Erroneous values
    1.
3. Other quality related preconditions
    1.
4. Other types of preconditions
    1.

14. Tuning parameters

1. Cut-offs.

15. Recommended use of the individual variants of the method

1. Surveys of very skewed populations which contain a few extreme values: surveys of business, agriculture, personal income and fortune.

16. Output data

1. Estimates of desired levels (target variable values after Winsorisation), quality measures for the estimates (e.g., variances, MSE).

17. Properties of the output data

1. The user should check the quality of estimates based on their knowledge of the investigated phenomenon, MSE, variance, bias of estimates.

18. Unit of input data suitable for the method

Information about variable of interest and auxiliary variables should be available for all units in the sample (sample level).

19. User interaction - not tool specific

The countermeasures against outliers can be divided into:

1. The detection of outliers – quantitative judgment, which requires an indicator of the degree of divergence of each data. Various methods of computing such indicators have been developed.
2. Outlier treatment:
    - “weight modification,” under which the weight of the sample unit is modified;
    - “value modification,” under which the value reported by the sample unit is modified;
    - the combination of the two, under which both the weight and the value reported by the sample are modified;
    - robust estimation techniques.

20. Logging indicators

1. Run time of the application.
2. Characteristics of the input data.
3. The number of units for which Winsorisation changed the values.

21. Quality indicators of the output data

1. MSE
2. Variance
3. Bias

22. Actual use of the method

1. *Survey of Employment, Payrolls, and Hours (SEPH)*, Statistics Canada: weight modification.
2. *State and Metro Area Employment, Hours, and Earnings*, Bureau of Labor Statistics America: reduces the impact of outliers through “weight reduction”.
3. *Consumer Price Index*, Australian Bureau of Statistics: Modifies the value of the outlier to the value next in size to the outlier through Winsorisation.

Interconnections with other modules

23. Themes that refer explicitly to this module

1. Statistical Data Editing – Main Module
2. Weighting and Estimation – Main Module

24. Related methods described in other modules

1. Weighting and Estimation – Calibration
2. Weighting and Estimation – Generalised Regression Estimator

25. Mathematical techniques used by the method described in this module

1. Regression
2. Ordinary Least Squares (OLS)
3. Trimmed least squares (TLS)
4. Trimmed least absolute value (LAV)
5. Sample Splitting
6. Least median of squares (LMS)

26. GSBPM phases where the method described in this module is used

1. 5.3 Review, validate, edit
2. 5.4 Impute
3. 5.6 Calculate weights
4. 5.7 Calculate aggregates

27. Tools that implement the method described in this module

1. Several popular statistical packages – including SAS, R, STATA, S-PLUS, LIMDEP, and E-Views – have procedures for robust regression analysis.
2. Least absolute deviations – SAS users call this procedure with the *LAV* command within the IML library. In STATA, median regression is performed with the quantile regression (qreg) procedure.
3. Least median of squares – SAS users can call least median of squares with the *LMS* call in *PROC IML*, S-Plus users can execute this algorithm with *lmsreg*.
4. Weighted least squares – SAS users call this procedure with the *LTS* command within the IML library.

28. Process step performed by the method

Estimation

Administrative section

29. Module code

Weighting and Estimation-M-Outlier Treatment

30. Version history

Version	Date	Description of changes	Author	Institute
0.1	29-02-2012	first version	Grazyna Dehnel	GUS
0.2	31-05-2012	second version	Grazyna Dehnel	GUS
0.3	10-12-2012	third version	Grazyna Dehnel	GUS
0.4	31-05-2013	fourth version	Grazyna Dehnel	GUS
0.5	09-09-2013	fifth version	Grazyna Dehnel	GUS
0.6	25-02-2014	sixth version	Grazyna Dehnel	GUS
0.6.1	11-03-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

31. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	26-3-2014 13:32