Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Robust Tools for Statistical Data Editing and Imputation

Prepared by WADA Kazumi[†] and TSUBAKI Hiroe[‡] ,
[†] Tsuda University, [‡] The Institute of Statistical Mathematics (ISM), Japan

I. Introduction

1. This paper aims to introduce two different software prepared and introduced for official statistics production in Japan. Both of them focus on alleviating the influence of outliers on statistical estimation.

2. The first tool is for a multivariate outlier detection method for elliptically distributed data called Modified Stahel-Donoho (MSD) estimators. Statistics Canada adopts these estimators for the Annual Wholesale and Retail Trade Survey (AWRTS), according to Franklin and Brodeur (1997). Béguin and Hulliger (2003) proposed a few improvements to the MSD estimators described by Franklin and Brodeur (1997). Wada (2010) then implemented a function both for the improved and original estimators in R for comparison. It is available at https://github.com/kazwd2008/MSD/ as *msd* function. The improved estimators are adopted for the Unincorporated Enterprise Survey in Japan since 2019. Details are shown in section II.

3. The second tool is for robust estimators for a generalised ratio model. The conventional ratio model is popular for the imputation of official statistics because the model can accommodate heteroscedastic data without data transformation. However, the heteroscedasticity of the model had been the obstacle of robustification. This paper introduces reformulation of the ratio model with homoscedastic error term as same as a regression model, generalisation of the model, and robustification by means of M-estimation based on Wada, Sakashita and Tsubaki (2019). The series of R functions implemented are available at https://github.com/kazwd2008/REGRM. See section III for details. The robust estimators are adopted for imputing major corporate accounting items of the 2016 Economic Census for Business Activity.

4. These tools were prepared during the authors' tenure at National Statistics Centre (NSTAC), Japan.

II. MSD estimators for multivariate outlier detection

5. The MSD estimators are a combination of the Stahel-Donoho (SD) estimators (Stahel, 1981; Donoho, 1983), and projection pursuit (PP) (Patak, 1990). The estimators achieve orthogonal equivariance and sufficient robustness owing to their high breakdown point. The SD estimators are used as the initial robust mean vector and covariance matrix, and principal component analysis of PP, which regards the principal components as "interesting" directions to find outliers, follows. Projection to the principal components also eliminates the correlation between variables. Possible outliers are down weighted in the same manner as the SD estimators, and the final mean vector and covariance matrix are derived. Outliers are decided by the Mahalanobis distance based on them. Please see Wada (2010), Wada and Tsubaki (2013) for the detail of the algorithm.

6. The Euredit project conducted from Mar. 2001 to Feb. 2003 made an evaluation and comparison of various outlier detection methods. A series of reports were published and made available at http://www.cs.york.ac.uk/euredit/, along with five papers published in the Journal of the Royal Statistical Society. In one of the papers, Béguin and Hulliger (2004) describe that NSOs had not used multivariate methods except for the Annual Wholesale and Retail Trade Survey (AWRTS) in Statistics Canada.

7. Wada (2010) implemented the estimators of Franklin and Brodeur (1997) and improved ones suggested by Béguin and Hulliger (2003). The implemented function for both versions is available at https://github.com/kazwd2008/MSD/. The improved version has better performance, while it suffered from the curse of dimensionality. It cannot cope with more than 11 variables with a 32-bit PC.

8. Then, Wada and Tsubaki (2013) implemented another R function by parallel computing so that the function can be applied to higher-dimensional datasets. It is available at https://github.com/kazwd2008/MSD.parallel. The paralleled version could be useful for high dimensional datasets with multi-thread PC, although further tuning may necessary.

9. Wada, Kawano and Tsubaki (2018) compared improved MSD with BACON (Billor et al., 2000), Fast-MCD (Rousseeuw and van Driessen, 1999) and NNVE (Wang and Raftery, 2002) to choose an appropriate one to remove outliers among hot-deck donor candidates for imputing corporate accounting items of the Unincorporated Enterprise Survey. MSD was selected as it showed better performance than others for skewed and long-tailed datasets. See Wada, Kawano and Tsubaki (2018) for more details of comparison and practical application.

III. Robust estimator for a generalized ratio model

10. This section describes the robust estimator prepared for imputation of the major corporate accounting items in the 2016 Economic Census for Business Activity in Japan. See Wada, Sakashita and Tsubaki (2019) for more details, including an application for the Economic Census.

A. Generalisation and robustification of the ratio model

11. The ratio model,

$$y_i = \beta x_i + \epsilon_i, \tag{1}$$

is frequently used for imputation. Missing y_i is replaced by $\hat{y}_i = \hat{\beta} x_i$ with the estimated ratio

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i},$$

where data $i = 1, \ldots, n$ of (x, y) are observed n units in the imputation class (Cochran, 1997, pp. 150-164). The heteroscedastic error term ϵ_i enables to accommodate heteroscedastic datasets without data transformation. It is a significant advantage since data transformation makes the estimation of means and totals unstable. On the other hand, it is an obstacle for robustification employing M-estimation.

12. The model (1) resembles a regression model without an intercept,

$$y_i = \beta x_i + \varepsilon_i. \tag{2}$$

However, their error terms are different. The ratio model (1) has $\epsilon_i \sim N(0, x_i \sigma^2)$ with scale parameter σ, while the regression model (2) has a homoscedastic error, $\varepsilon_i \sim N(0, \sigma^2)$.

13. The model (1) can be re-formulate with homoscedastic error term as follows:

$$y_i = \beta x_i + \varepsilon_i \sqrt{x_i}, \tag{3}$$

because the error terms of the model (1) and (2) have the relation, $\epsilon_i = \sqrt{x_i}\varepsilon_i$. The model can also be generalized as follows:

$$y_i = \beta x_i + x_i^{\gamma} \varepsilon_i. \tag{4}$$

The model (4) is equivalent to (3) when $\gamma = 1/2$. The model can also be expressed as $y_i = \beta x_i + \epsilon_i$, with the heteroscedastic error term $\epsilon_i \sim N(0, x_i^{\gamma}\sigma^2)$. The corresponding estimator for the model (4) is,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i^{1-2\gamma}}{\sum_{i=1}^{n} x_i^{2(1-\gamma)}}. \tag{5}$$

14. The robustification of the estimator (5) is now straight forward since the corresponding model (4) has a homoscedastic error term as same as a regression model. The robustified estimator is,

$$\hat{\beta}_{rob} = \frac{\sum w_i y_i {x_i}^{1-2\gamma}}{\sum w_i {x_i}^{2(1-\gamma)}},$$

where w_i is obtained by a weight function with homoscedastic quasi-residuals

$$\check{r}_i = \frac{y_i - \hat{\beta}_{rob} x_i^{\gamma}}{x_i^{\gamma}},$$

and a scale parameter σ. Table 1 shows the model, estimator, and quasi-residuals with a few different γ values of the robust estimator.

Table 1: Estimators for generalised ratio model with different γs.

γ	Model with homoscedastic quasi-error	Robust estimator	Quasi residuals
$\gamma = 0$	$y_i = \beta x_i + \varepsilon_i$	$\hat{\beta}_{rob} = \frac{\sum w_i y_i x_i}{\sum w_i x_i^2}$	$\check{r}_i = w_i y_i - \hat{\beta}_{rob} w_i x_i$
$\gamma = 1/2$	$y_i = \beta x_i + \varepsilon_i \sqrt{x_i}$	$\hat{\beta}_{rob} = \frac{\sum w_i y_i}{\sum w_i x_i}$	$\check{r}_i = \frac{y_i \sqrt{w_i}}{\sqrt{x_i}} - \hat{\beta}_{rob} \sqrt{w_i x_i}$
$\gamma = 1$	$y_i = \beta x_i + \varepsilon_i x_i$	$\hat{\beta}_{rob} = \frac{\sum w_i (y_i / x_i)}{\sum w_i}$	$\check{r}_i = \frac{w_i y_i}{w_i x_i} - \hat{\beta}_{rob}$

B. Implementation

15. Tukey's biweight function (Beaton and Tukey, 1974)

$$w_i = w\left(\frac{\check{r}_i}{\sigma}\right) = \begin{cases} [1 - (e_i/c)^2]^2 & |e_i| \le c \\ 0 & |e_i| > c, \end{cases}$$

and Huber's weight function (Huber, 1964)

$$w_i = w\left(\frac{\check{r}_i}{\sigma}\right) = \begin{cases} 1 & |e| \le k \\ \frac{k}{|e|} & |e| > k \end{cases}$$

are selected for implementation, as they are popular among a variety of weight functions to obtain robust weight w_i.

16. For scale parameter, average absolute deviation (AAD)

$$\hat{\sigma}_{\mathrm{AAD}} = \mathrm{mean}(|\check{r}_i - \mathrm{mean}(\check{r}_i)|),$$

and median absolute deviation (MAD)

$$\hat{\sigma}_{\mathrm{MAD}} = \mathrm{median}(|\check{r}_i - \mathrm{median}(\check{r}_i)|)$$

are selected.

17. Wada and Noro (2019) standardise the tuning constants for the above-mentioned weight functions, as shown in Table 2. They also propose appropriate values for each setting, as shown in Table 3, based on Bienias et al. (1997), which recommend tuning constant c for Tukey's biweight function with AAD scale from *c*=4 to 8. Please note that R's *mad* function returns adjusted values consistent with standard deviation.

Table 2. Tuning constants for 95% asymptotic efficiency.

Tuning constant	95% asymptotic efficiency for normal distribution		
	σ_{SD}	σ_{MAD}	σ_{AAD}
c for Tukey	4.685	3.160	3.738
k for Huber	1.345	0.907	1.073

Table 3. Standardised tuning constants.

Weight function	Scale parameter	Very robust	...	Less robust	Default
Tukey	AAD	4	6	8	8
Tukey	MAD(SD)	5.01	7.52	10.03	10.03
Huber	AAD	1.15	1.72	2.30	2.30
Huber	MAD(SD)	1.44	2.16	2.88	2.88

18. The following R functions are implemented for the robust estimator of the generalised ratio model by the iteratively re-weighted least squares (IRLS) algorithm (Holland and Welsch, 1977) in accordance with Bienias et al. (1997). Two weight functions, two scale parameters, and three choices of the gamma value are selected, as shown in Table 2. They are available as REGRM package at https://github.com/kazwd2008/REGRM, together with an integrated function *REGRM,* which calls an appropriate child function among those shown in Table 2.

Table 2: Implemented functions

Weight function	Scale parameter	$\gamma = 1$	$\gamma = 1/2$	$\gamma = 0$
Tukey	AAD	RrTa.aad	RrTb.aad	RrTc.aad
Tukey	MAD	RrTa.mad	RrTb.mad	RrTc.mad
Huber	AAD	RrHa.aad	RrHb.aad	RrHc.aad
Huber	MAD	RrHa.mad	RrHb.mad	RrHc.mad

19. An attempt of simultaneous robust estimation of β and γ of the generalised ratio model is also underway based on the two-stage least squares (2SLS). See Wada, Takata and Tsubaki (2019) for the progress.

References

Beaton, A. E. & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics*, 16, 147-185.

Béguin, C. & Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. *EUREDIT Deliverable,* D4/5.2.1/2 Part C. [http://www.cs.york.ac.uk/euredit/]

Béguin, C. & Hulliger, B. (2004). Multivariate outlier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations, *Journal of the Royal Statistical Association*, Series A, 167, Part 2, 275-294.

Billor, N., Hadi, A. S., Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34, 279-298.

Cochran WG (1977). *Sampling Techniques*, 3rd ed. John Wiley & Sons.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. Qualifying paper, Harvard University.

Franklin, S. & Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 186-191.

Huber, P. J. (1964). Robust Estimation of a Location Parameter, *Annals of Mathematical Statistics,* **35**(1), 73-101.

Patak Z. (1990). Robust Principal Component Analysis via Projection Pursuit," M. Sc. Thesis, University of British Columbia, Canada.

Rousseeuw, P. J. and Van Driessen, K., (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212-223.

Stahel, W. A. (1981). Breakdown of covariance estimators. Research Report 31, *Fachgruppe für Statistik*, E.T.H. Zürich.

Wada K. (2010). Detection of Multivariate Outliers: Modified Stahel-Donoho Estimators (in Japanese). *Research Memoir of Official Statistics*, 67, 89-157.
URL http://www.stat.go.jp/training/2kenkyu/pdf/ihou/67/wada1.pdf

Wada, K., Kawano, M. & Tsubaki, H. (2018). Comparison of multivariate outlier detection methods for nearly elliptical distributions. *Austrian Journal of Statistics, **49(2),*** 1-17. URL https://doi.org/10.17713/ajs.v49i2.872.

Wada, K. & Noro, T. (2019). Consideration on the Influence of Weight Functions and the Scale for Robust Regression Estimator (in Japanese). *Research Memoir of Official Statistics*, 76, 101-114.
https://www.stat.go.jp/training/2kenkyu/ihou/76/pdf/2-2-767.pdf

Wada, K., Takata, S. & Tsubaki, H. (2019) An algorithm of generalized Robust ratio model estimation for imputation. *JSM Proceedings*, Government Statistics Session, Denver, CO: American Statistical Association.

Wada, K., Sakashita, K. & Tsubaki, H. (2019) Robust Estimation for a Generalised Ratio Model. to be appeared in *Austrian Journal of Statistics* of the special issue for uRos2019.

Wada, K. & Tsubaki, H. (2013). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection. *Proceedings of 2013 IEEE International Conference on Cloud Computing and Big Data* (CloudCom-Asia), 304-311, 16 -19, Dec. 2013, Fuzhou, China.

Wang, N. and Raftery, A. E. (2002). "Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning." *Journal of the American Statistical Association*, **97**(460), 994-1019.