STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 97–110

WINSORIZATION METHODS IN POLISH BUSINESS SURVEY

Grażyna Dehnel[1]

ABSTRACT

One of the major problems involved in estimating information about economic activity across small domains is too small sample size and incompleteness of data sources. For instance, the distribution of enterprises by target variables tends to be considerably right-skewed, with high variation, high kurtosis and outliers. Therefore, it is not obvious that the implementation of traditional estimation methods meets the desired requirements, such as being free from bias or having competitive variance. Furthermore, the pressure to produce accurate estimates at a low level of aggregation or needs to substantially reduce sample size have increased the importance of exploring the possibilities of applying new, more sophisticated methods of estimation. The aim of the study was to test the usefulness of winsorization methods to estimate economic statistics from the DG1 survey.

Key words: domain estimation, business statistics, winsorized estimator.

1. Introduction

Nowadays the growing demand for business information at a low level of aggregation has called for estimation methods that could meet the requirements specified by the user's needs. In practice, business surveys often pose a variety of data problems. For example, target variables tend to be highly skewed and populations can 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. 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 for small domains. In the case of small sample size, outliers can result in estimates greatly diverging from the real value for the population. Even if the sample size is large, the influence of an outlier can significantly increase the variance resulting in a decreased efficiency of estimation.

[1] Poznan University of Economics, Department of Statistics. E-mail: g.dehnel@ue.poznan.pl.

Dealing with outliers has two aspects: the first one involves identifying outlying observations in an objective way, while the second focuses on ways of handling them to reduce their effect on survey estimates.

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

The paper describes implementation of winsorization - one frequently applied estimation method, used to reduce the impact of outlying units. The general idea of winsorization is that if an observation exceeds a preset cutoff value, then the observation is replaced by that cutoff value or by a modified value closer to the cutoff value.

The objective of the referred study was to assess the performance of four various methods use to estimate robust regression parameters, and hence estimate the cutoff values used in the winsorized estimator. The paper presents attempts to estimate basic economic information about small, medium-sized and large businesses at a low level of aggregation (in the joint cross-section of economic activity classification and the territorial division by province).

2. Estimation method

Winsorization is often used for data cleaning in statistical practice. Since outliers are a serious problem in many sample surveys (especially business surveys), an appropriate way of handling them is required. Winsorization involves identifying cutoff values. Sample observations whose values lie outside certain preset cutoff values are transformed in order to make them closer to the cutoff value.

Cutoff values are derived in a way that approximately minimizes the MSE of estimates. All sampled units are divided into two groups. One group contains typical observations which are left unmodified, the other one contains observations regarded as outliers. The classification is made on the basis of two preset cutoff values. Then, values of the study variable outside the cutoff 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 the winsorized estimation, we obtain a „new" 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 winsorized estimator, with GREG estimation, can be expressed as:

$$\hat{Y}_{win} = \sum_{i \in s_d} \tilde{w}_i y_i^* = \sum_{i \in s_d} w_i g_i y_i^* \qquad (1)$$

where, in the presence of outliers, modified values of the study variable y_i^* are calculated in the following manner (Gross, Bode Taylor, Lloyd-Smith, 1986):

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

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

where:

s_d - population parameter for domain *d*

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

$s(s \subseteq N)$ - 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 cutoff value

K_{Li} - lower cutoff value

The cutoff values are calculated to minimize MSE of the winsorized estimator under the model (Preston, 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:

$\mu_i^* = E(Y_i^*)$ - expectation under the assumed 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 $\hat{Y}_{winL}$

$\hat{Y}_{winU}$ - the winsorized estimator of the population total when only upper winsorization is performed

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

When winsorization is mild and reasonably symmetric, being μ_i^* difficult to estimate, we can replace μ_i^* with μ_i . Then, the approximately optimal cutoffs are (Preston, 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, Mackin, 2002) the cutoff values are estimated based on the following formulas:

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

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

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

In order to estimate the bias parameter B_U under winsorization 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, 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 $\hat{D}_i$: $\hat{D}_{(1)} \geq \hat{D}_{(2)} \geq ... \geq 0 \geq ...$. By solving $\psi_U(G) = 0$ one can obtain the value of G.

In order to estimate the cutoff 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, Mackin, 2002) include: *Trimmed least squares (TLS), Trimmed least absolute value (ABS), Sample Splitting (HALF), 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:

$$F = \sum_{i \in s} \left(y_i - \beta^T x_i \right)^2 \tag{12}$$

Then fitted values are calculated, and then residuals. In the second step, units with the largest positive and negative residuals are removed. Finally, a new regression model is fitted to the reduced sample in order to estimate the value of μ_i^*.

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

$$F = \sum_{i \in s} \left| y_i - \beta^T x_i \right| \tag{13}$$

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 *ABS* 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 (HALF)*** based on Ordinary Least Squares (OLS). It is applied to data 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 *HALF* technique is expected to be more robust than TLS because the residuals used to remove the 'outlier' units are not calculated from the 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 (Preston, Mackin, 2002).

3. Data source

Information for the study came from the DG1 survey conducted by the Statistical Office in Poznan. The survey is conducted in the form of monthly reports submitted by all large and medium-sized enterprises and a 10% sample of small enterprises. Its objective is to collect up-to-date information about basic indicators of economic activity of enterprises, such as *revenue from sales (of products and services), number of employees, gross wages, volume of wholesale trade and retail sales, excise tax, specific subsidies.* The sample frame includes 98,000 units, of which 19,000 are medium-sized and large enterprises (with over 49 employees), 80,000 are small enterprises (from 10 to 49 employees). In effect, about 30,000 units participate in the survey every month.

4. Description of the study

The study was limited to enterprises that were active in August of 2012. *Gross wages* were the target variable, while *revenue from sales of products (goods and services)* was the auxiliary variable.

The general population included all enterprises that participated in the DG1 survey. This choice enabled access to detailed information about the target and auxiliary variables. With the general population defined in this way, it was possible to conduct a simulation study, which was then used to evaluate estimation precision.

The level of aggregation adopted for the study was a combination of economic activity classification (NACE Rev.2) and the territorial division by province.

5. Precision assessment methods

The precision of estimators analysed in the study was evaluated using the bootstrap method. 1000 iterations of drawing 20% samples were made, which were then used to calculate:

- Relative estimation error (REE)

$$CV\left(\hat{Y}_d\right)=\frac{\sqrt{Var\left(\hat{Y}_d\right)}}{E\left(\hat{Y}_d\right)} \tag{14}$$

where: $$Var\left(\hat{Y}_d\right)=\frac{1}{999}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d}-\hat{Y}_d\right)^2 \tag{15}$$

- Mean absolute relative bias (ARB)

$$ARB\left(\hat{Y}_d\right)=\frac{1}{1000}\left|\sum_{b=1}^{1000}\frac{\hat{Y}_{b,d}-Y_d}{Y_d}\right| \tag{16}$$

- Relative root mean square error (RMSE)

$$RMSE\left(\hat{Y}_d\right)=\frac{\sqrt{\frac{1}{1000}\sum_{b=1}^{1000}\left(\hat{Y}_{b,d}-Y_d\right)^2}}{Y_d} \tag{17}$$

- To describe the general precision of combined estimates, for all small areas, mean values of Relative root mean square error applied to particular domains were calculated. The mean values were calculated as arithmetic means used in empirical studies and as weighted means

$$\overline{RMSE}=\frac{1}{D}\sum_{d=1}^{D}RMSE(\hat{Y}_d)$$

Owing to the large volume of estimation results, the following presentation is limited to estimates for the variable of *gross wages* for two PKD categories: *manufacturing, construction* and *trade.*

6. Estimation results and assessment of their precision

The effect of different winsorized estimation techniques on the value of the study variable is shown on a scatterplot (see Fig. 1). To illustrate the shift in values as a result of modification, only domains for *manufacturing* have been selected. Empirical values of the study variable in domains are marked by a black cross. Each domain is represented by five points: the real value and values modified as a result of each of the four robust regression techniques. The degree of modification depends on the type of robust regression technique. It is also worth noting that in nearly all the cases the HT estimates were significantly different from the winsorized estimates.

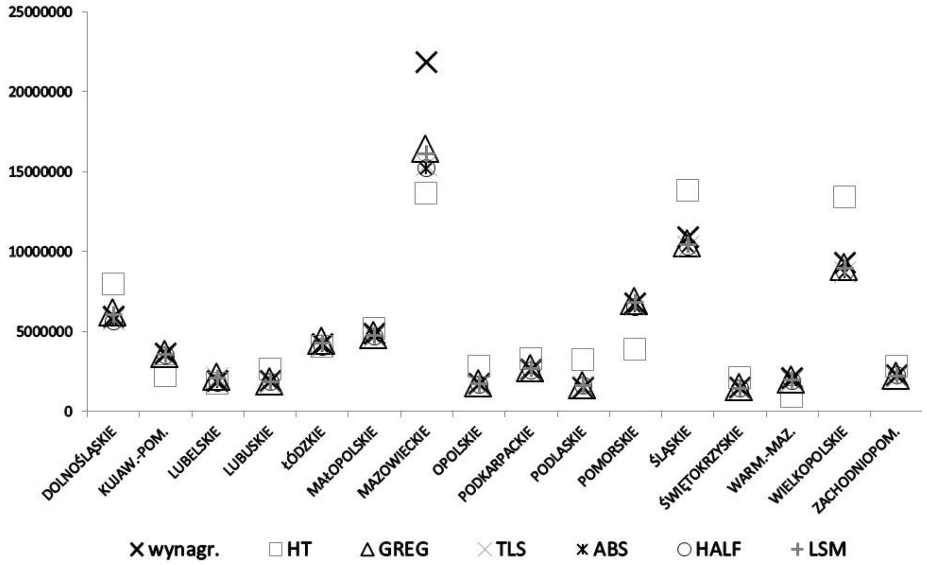


Figure 1. Real values (*Y*- Gross Wage) and values estimated by winsorization
Source: Own calculations based on DG1 survey, data from August 2012.

The scatterplot shows both the direction and the degree of modification of the study variable. In the case of units classified as *x-outliers*, namely for small values of wages paid by businesses with high revenue, the modification involved increasing the value of the study variable. The study variable was decreased in the case of outliers corresponding to businesses paying high wages but reporting low revenue.

Figures 2-7 present the distribution of three performance criteria: relative estimation error, mean absolute relative bias and relative root mean square error for two analyzed sections: *construction* and *trade*. From the results in Fig. 2 and 3 we can see that in most cases the winsorized estimator has considerably less REE than the HT and GREG estimators.

The amount of bias induced by winsorizing is for most cases almost insignificant except in the case of province characterised by high variation of the auxiliary variable (see Fig. 4 and 5). In terms of RMSE, the performance of the winsorized estimators is considerably better than the HT and GREG estimator (see Fig. 6 and 7).

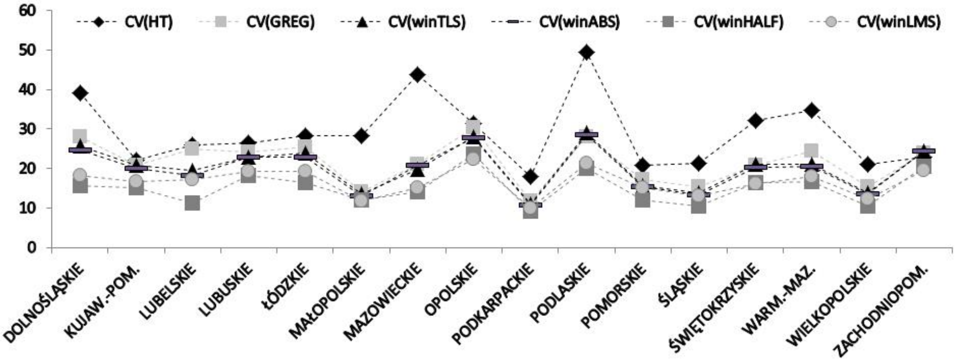


Figure 2. Relative estimation error for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

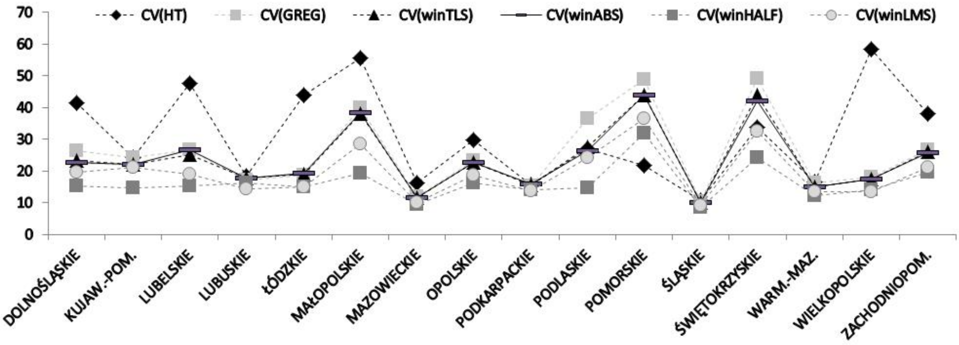


Figure 3. Relative estimation error for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

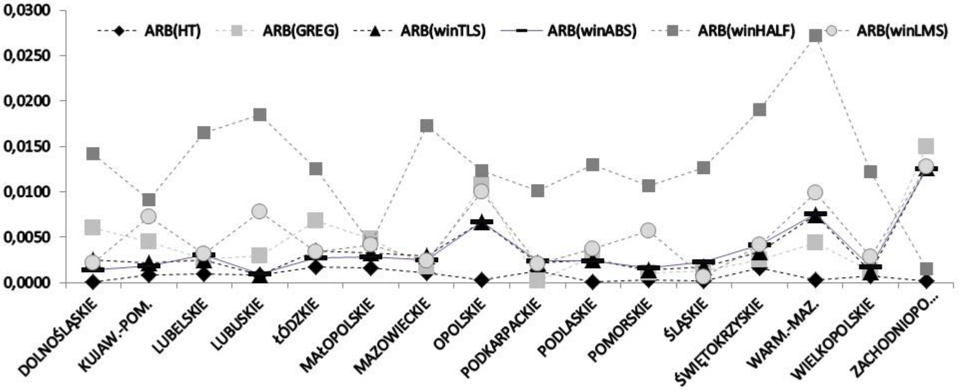


Figure 4. Mean absolute relative bias for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

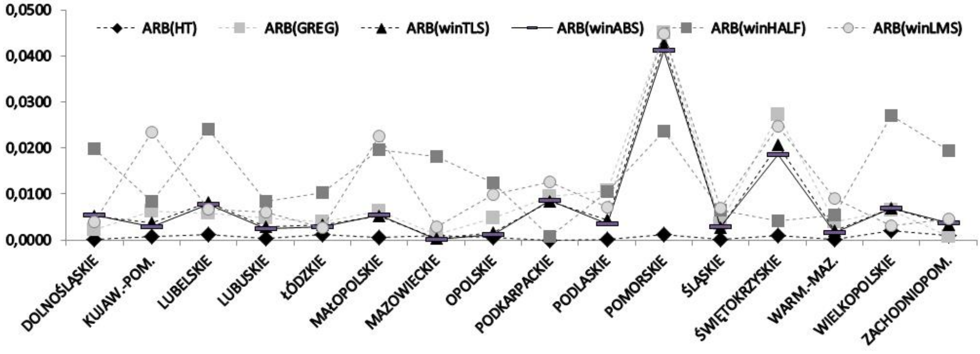


Figure 5. Mean absolute relative bias for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

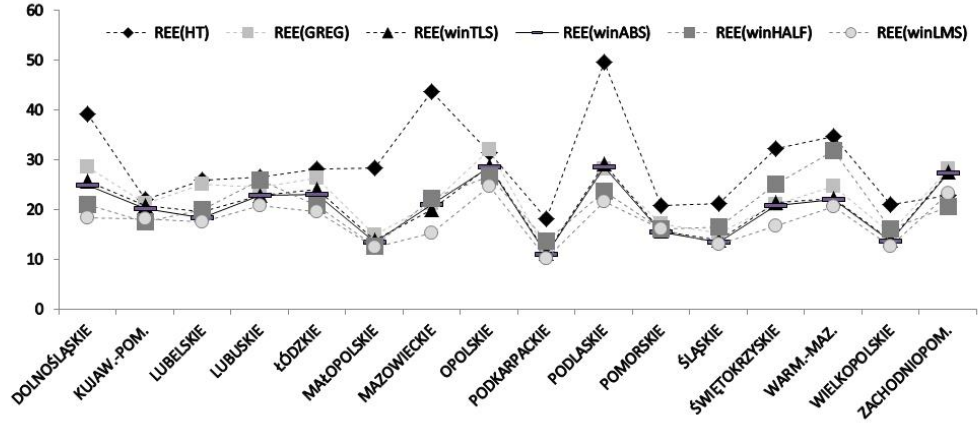


Figure 6. Relative root mean square error for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

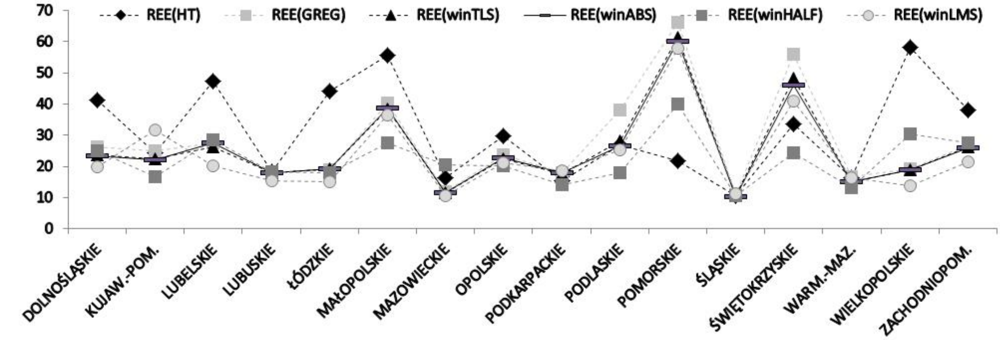


Figure 7. Relative root mean square error for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

For most values of the estimated cutoffs (calculated according to the four various methods of estimate robust regression parameters), the winsorized estimator significantly outperformed the expansion estimators (see Tab. 1). There are very few cases when the HT and GREG estimation is better than the winsorized estimator. The winsorized estimator nearly always had considerably smaller RMSE than the expansion estimators.

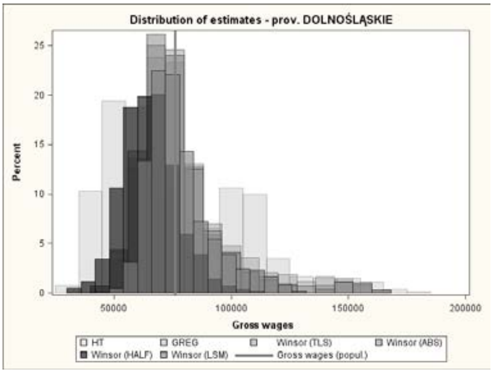
The results described above indicate that winsorizing optimize trade-off between variance and bias. The improvement in the general performance of the estimator that is obtained against extremely large errors from winsorizing is usually at the price of introducing a small amount of bias in estimation.

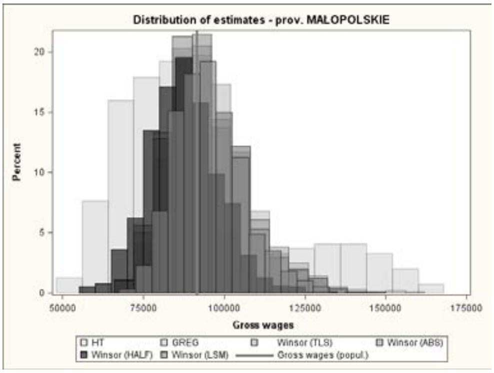
Table 1. Relative root mean square error for *construction* and *trade*

$\overline{RMSE}$	HT	GREG	winTLS	winABS	winHALF	winLMS
Construction	29.1	22.2	20.7	20.3	20.6	17.6
	RMSE<RMSE$_{HT}$ (%)	88	94	94	100	94
Trade	31.1	27.7	25.5	25.2	22.1	23.5
	RMSE<RMSE$_{HT}$ (%)	56	75	81	88	69

Source: Own calculations based on DG1 survey, data from August 2012.

Figures 8-9 present the distribution of estimates for selected provinces for *construction* and *trade*. The use of the winsorized estimation reduces estimator variance compared to direct estimation. The distribution of the winsorized estimates is significantly more leptokurtic than DIRECT or GREG estimates. In many cases it follows the normal distribution while the distribution of DIRECT or GREG estimators is sometimes multimodal or highly skewed. It is very difficult to point out which type of the winsorized estimators has better properties based on the presented figures.





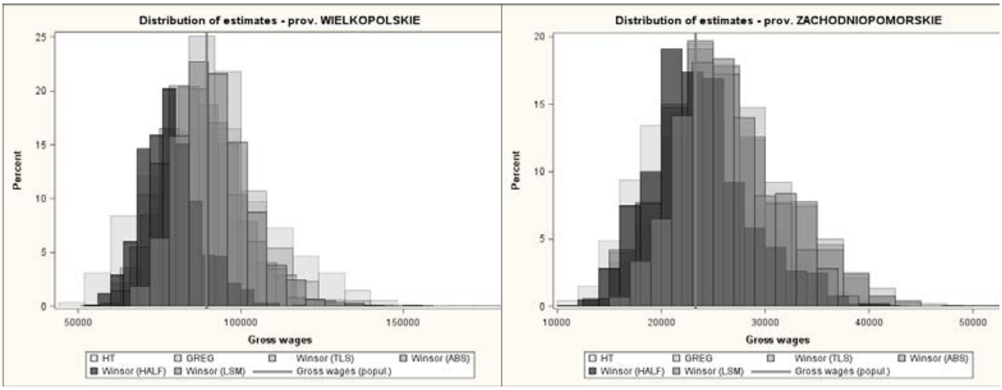


Figure 8. Distribution of estimates for selected provinces for *construction*

Source: Own calculations based on DG1 survey, data from August 2012.

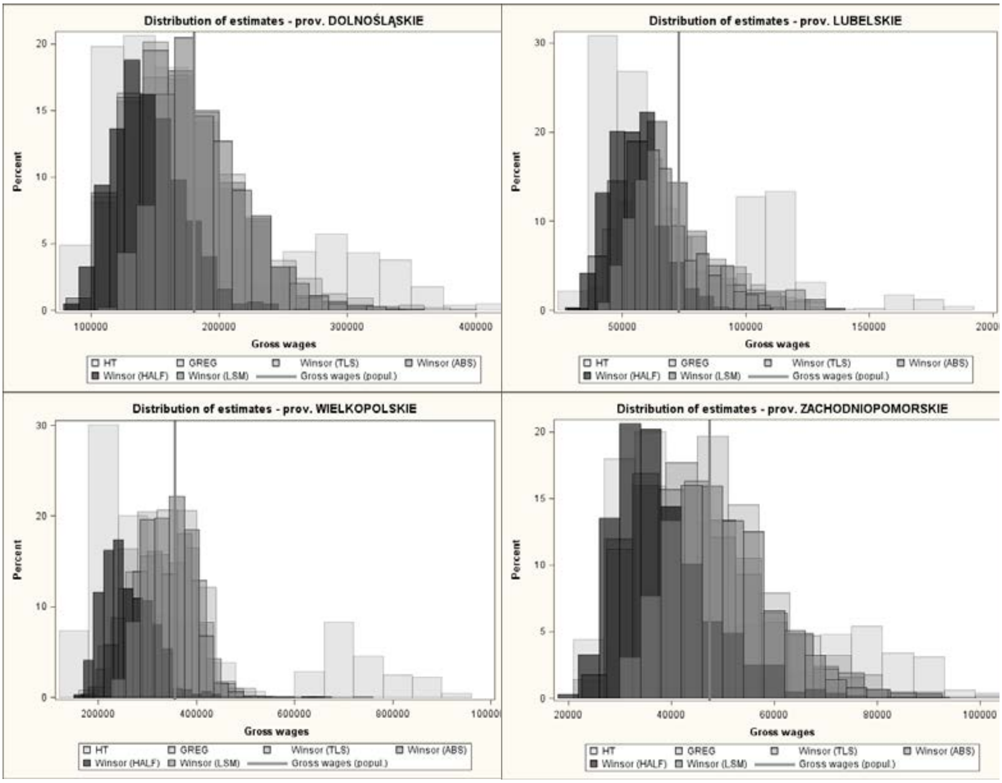


Figure 9. Distribution of estimates for selected provinces for *trade*

Source: Own calculations based on DG1 survey, data from August 2012.

7. Conclusion

- Simulation research demonstrated the relation between efficiency of estimation and a type of robust regression technique used.
- The effectiveness of the winsorized estimator in terms of its resistance to unusually large residuals depends on the choice of cutoff values - in other words, on methods of estimating bias parameters and regression parameters. The more robust regression technique was applied, the more efficient estimates were produced.
- The use of the winsorized estimation reduces estimator variance.
- Winsorization reduces outliers values, producing an insignificant estimated bias in the characteristic estimates.
- If cutoff values are chosen appropriately, the decline in variance is big enough to offset the bias of MSE. The winsorized estimator nearly always outperforms the expansion estimator in terms of MSE.

REFERENCES

CHAMBERS, R., KOKIC, P., SMITH, P., 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.

COX, B. G., BINDER, A., CHINNAPPA, N. B., CHRISTIANSON, A., COLLEDGE, M. J., KOTT, P. S., (1995). Business Survey Methods, John Wiley & Sons.

GROSS, W. F., BODE, G., TAYLOR, J. M., LLOYD–SMITH, C. W., (1986). Some finite population estimators which reduce the contribution of outliers, [in:] Proceedings of the Pacific Statistical Conference, 20–24 May 1985, Auckland, New Zealand.

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

PRESTON, J., MACKIN, C., (2002). Winsorization for Generalised Regression Estimation, Australian Bureau of Statistics.

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