STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 137–152

ROBUST REGRESSION IN MONTHLY BUSINESS SURVEY

Grażyna Dehnel[1]

ABSTRACT

There are many sample surveys of populations that contain outliers (extreme values). This is especially true in business, agricultural, household and medicine surveys. Outliers can have a large distorting influence on classical statistical methods that are optimal under the assumption of normality or linearity. As a result, the presence of extreme observations may adversely affect estimation, especially when it is carried out at a low level of aggregation. To deal with this problem, several alternative techniques of estimation, less sensitive to outliers, have been proposed in the statistical literature. In this paper we attempt to apply and assess some robust regression methods (*LTS, M-estimation, S-estimation, MM-estimation*) in the business survey conducted within the framework of official statistics.

Key words*:* robust regression, outlier detection, business statistics.

1. Introduction

One of the main problems involved in estimating population parameters is distributions of enterprises in terms of the variable of interest and auxiliary variables, which are characterised by a high variation, strong asymmetry and kurtosis. This is due to, inter alia, non-response survey errors, a large proportion of zero values for survey variables and extreme values. In this article we focus on the third issue – extreme values. Although some observations are extreme, they need not necessarily be incorrect but may be part of the survey population. The statistical literature refers to these observations as outliers.

Undoubtedly, in many business surveys conducted within the framework of official statistics sample sizes are large enough to compensate for the presence of outliers, which have a relatively small impact on estimates. However, at low levels of unit aggregation the impact of outliers might be significant. The appearance of outlying observations is particularly noticeable in estimates for short-term statistics, where surveys are repeated monthly or quarterly.

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

The aim of outlier treatment is to make estimates as close to the parameters of the population as possible. This is not a simple task in the presence of outliers, since estimators do not retain their properties, such as resistance to bias or efficiency. This means that outlier treatment should provide some kind of a trade-off between variance and bias. And in the case of a population known to comprise outliers (such as populations of enterprises) a robust analysis should be considered in addition to the classical approach. In the statistical literature several robust methods have been proposed. The aim of the present study was to compare the usefulness of four robust regression methods: *LTS, M-estimation, S-estimation, MM-estimation* against LS regression estimation based on data derived from a business survey. An empirical example is conducted in SAS.

2. Robust regression methods

The main objective of robust regression methods is to provide stable results when fundamental assumptions of the least squares regression are not fulfilled due to the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers by reducing the weights of outliers, changing the values of outliers or using robust estimation techniques (Chen, 2007). Many methods have been developed for these problems, but those most commonly used today are Huber *M-estimation, Least Trimmed Squares (LTS) estimation, S-estimation* and *MM-estimation.*

M-estimation

The most widely used general method of robust regression is *M-estimation,* introduced by Huber (1964), which is nearly as efficient as LS (Huber, 1964).

Instead of minimizing the sum of squares of the residuals, a Huber-type M-estimator $\hat{\theta}_M$ of θ minimizes the sum of less rapidly increasing functions of the residuals:

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}(\theta)\right) \tag{1}$$

where $r_i = y_i - X\theta$,

s - scale parameter,

ρ(·) is a loss function, which is even, non-decreasing for positive values and less increasing than the square function.

To guarantee scale equivariance (i.e. independence with respect to the measurement units of the dependent variable), residuals are standardized by a measure of dispersion *s* (Verardi, Croux, 2009).

The estimator is not robust with respect to leverage points, but it is useful in analyzing data for which it can be assumed that the contamination is mainly in the *y*-direction.

Assuming *s* to be known, the M-estimate is found by solving:

$$\sum_{i=1}^{n} \Psi\left(\frac{y_i - \sum_{k=1}^{p} x_{ik}\theta_k}{s}\right) x_i = 0 \tag{2}$$

where Ψ is the first derivative of ρ .

The choice of the Ψ function is based on the preference of how much weight to assign to outliers and this leads to different variants of M-estimators. A monotone Ψ function does not assign weight to large outliers as big as least squares do. A redescending Ψ function increases the weight assigned to an outlier until a specified distance (e.g. 3σ) and then decreases the weight to 0 as the outlying distance gets larger (Alma, 2011).

The choice of the Ψ function is not critical to gaining a good robust estimate, and many choices will give similar results that offer great improvements, in terms of efficiency and bias, over classical estimates in the presence of outliers (Huber, 1981). *M-estimation* has a breakdown point of 1/n.

Least Trimmed Squares (LTS) Estimation

The *least trimmed squares (LTS) estimate* proposed by Rousseeuw (1984) is given by

$$\hat{\theta}_{LTS} = \arg\min_{\theta} \sum_{i=1}^{h} Q_{LTS}(\theta) \tag{3}$$

where $Q_{LTS}(\theta) = \sum_{i=1}^{h} r_{(i)}^2$,

$r_{(1)}^2 \le r_{(2)}^2 \le \ldots \le r_{(n)}^2$ - are the ordered squared residuals

h – is defined in the range $\frac{n}{2} + 1 \le h \le \frac{3n + p + 1}{4}$ or $h = \frac{n + p + 1}{2}$,

p – number of parameters.

LTS is calculated by minimizing the *h* ordered squares residuals. The largest squared residuals are excluded, which allows those outlier data points to be removed completely.

Depending on the value of *h* and the outlier data configuration, LTS can be very efficient. In fact, if the exact numbers of outliers are trimmed, this method is computationally equivalent to LS (Alma, 2011). However, if there are more extreme values than are trimmed, this method is not as efficient. In turn, if there is more trimming than there are outlying data points, then some good data will be excluded from the estimation. LTS is considered to be a high breakdown method with a breakdown point of 50% (Rousseeuw, Leroy, 1987; Rousseeuw, Driessen, 1998).

S-estimation

S-estimation proposed by Rousseeuw and Yohai (1984) minimizes the dispersion of the residuals. However, it uses a robust measure for the variance. It is defined as

$\hat{\theta}_s = \arg\min_{\theta} \hat{\sigma}(r(\theta))$ where $\hat{\sigma}(r)$ is an M-estimator of scale, found as the solution of

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{Y_i - x_i^{'}\theta}{\hat{\sigma}}\right) = K \tag{4}$$

where $K = const = E[\rho]$.

The final scale estimate, $\hat{\sigma}$, is the standard deviation of the residuals from the fit that minimized the dispersion of the residuals.

Rousseeuw and Yohai (1984) suggest a redescending influence function as:

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4} & \text{for } |x| \le c \\ \\ \dfrac{c^2}{6} & \text{for } |x| > c \end{cases} \tag{5}$$

The parameter *c* is the tuning constant. Trade-offs in breakdown and efficiency are possible based on choices for tuning constant *c* and *K* (Alma, 2011). The usual choice is *c*=1.548 and K=0.1995 for 50% breakdown and about 28% asymptotic efficiency (Rousseeuw, Leroy, 1987). *S*-estimation is a high breakdown value method.

MM-estimation

MM-estimation is a combination of high breakdown value estimation and efficient estimation, which was introduced by Yohai (1987).

The procedure consists of three steps (Alma 2011):

1. Calculation of an *S*-estimate with the influence function
2.

$$\rho(r) = \begin{cases} 3\left(\dfrac{r}{c}\right)^2 - 3\left(\dfrac{r}{c}\right)^4 + \left(\dfrac{r}{c}\right)^6 & \text{for } |r| \le k \\ \\ 1 & \text{for } |r| > k \\ \end{cases} \tag{6}$$

The value of the tuning constant, c, is selected as 1.548.

3. Calculation of the MM parameters that provide the minimum value of $\sum_{i=1}^{n} \rho\left(\frac{Y_i - x_i^{'}\theta_{MM}}{\hat{\sigma}}\right)$ where $\rho(r)$ is the influence function used in the first stage with the tuning constant 4.687 and 0 σˆ is the estimate of scale from the first step (standard deviation of the residuals).
4. Calculation of the MM estimate of scale as the solution to

$$\frac{1}{n-p}\sum_{i=1}^{n} \rho\left(\frac{Y_i - x_i^{'}\theta_{MM}}{\hat{\sigma}}\right) = 0{,}5 \tag{7}$$

MM-estimation is a special type of M-estimation. It is the estimation with a high breakdown point (50%) and high efficiency (70%) under normal error (Stromberg, 1993).

3. Data source

Information for the study came from the DG1 survey conducted by the Statistical Office in Poznan and from tax register. The DG1 survey is a business activity report submitted by large, medium-sized and small enterprises. It is the basic source of short-term information about economic activity of businesses, 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). Tax register served as the auxiliary data source.

4. Description of the study

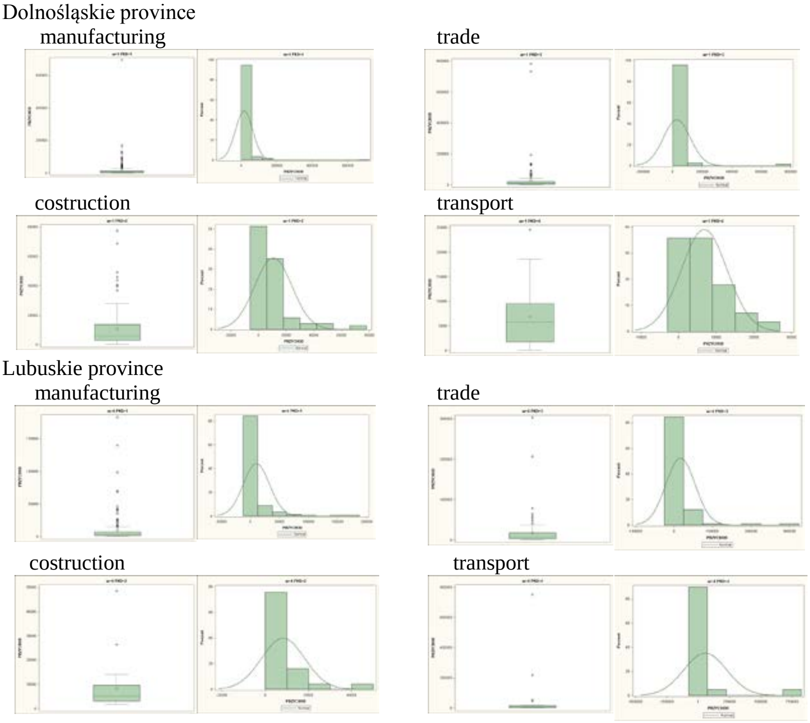
The study was limited to small enterprises (from 10 to 49 employed persons) that were active in December 2011. We took into consideration two models with a different number of auxiliary variables. There was one auxiliary variable (cost) in the first model and there were three auxiliary variables (income, cost and *revenue)* in the second model. *Revenue from sales of products (goods and services)* was the target variable. The general population included all small and medium-sized enterprises that participated in the DG1 survey. This choice enabled access to detailed information about the target and auxiliary variables. The level of aggregation adopted for the study was a combination of the territorial division by province and the biggest four sections of economic activity (NACE Rev.2) - *manufacturing, construction, trade* and *transport*. The population was thus broken down into 64 domains (16 provinces x 4 NACE sections). Owing to the large volume of study results, the following presentation is limited to two provinces - Dolnośląskie, Lubuskie and the 4 NACE sections. The selection of provinces was made on the basis of information about the goodness of fit of the model. The main objective of choice was to include domains with a high variation of the coefficient of determination (from 0.041 to 0.999), see Table 1.

Table 1. Coefficient of determination for the regression models of *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

AUXILIARY VARIABLES	*cost*	*income, cost, revenue*	*cost*	*income, cost, revenue*
NACE / Provinces	*Dolnośląskie*		*Lubuskie*	
Manufacturing	0.975	0.996	0.985	0.991
Construction	0.742	0.815	0.980	0.995
Trade	0.989	0.991	0.999	0.999
Transport	0.498	0.510	0.041	0.050

Source: Own calculations based on DG1 survey and tax register from December 2011.

The analysis started with the assessment of the distributions of the variable of interest coming from DG1 survey. A look at the distributions of *revenue* shows that a relatively large percentage of economic entities display zero values (or close to zero) in this respect. Moreover, there are very long right-hand tails in the histograms, as expected, see Fig. 1. This is the justification for the use of statistical techniques that are able to cope with or to detect outlying observations.



Dolnośląskie province
manufacturing
trade
costruction
transport
Lubuskie province
manufacturing
trade
costruction
transport

Figure 1. Distribution of enterprises by revenue for *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces
Source: Own calculations based on DG1 survey and tax register from December 2011.

Based on the distributions, the analysis of outliers was divided into two stages: the first one involved detecting outlying observations, the second focused on ways of handling them to reduce their effect on survey estimates by applying robust regression methods.

To identify outlying observations *RSTUDENT* was applied. This is one of the most widely used measures to identify outliers. The value of the RSTUDENT for each observation is the difference between the observed y_i and the predicted value of $\hat{y}$ excluding this observation from the regression and can be calculated using the following formula:

$$r_i * = \frac{e_i}{\sqrt{MSE_i} \cdot \sqrt{1 - h_i}} \tag{8}$$

where: $r_i *$ - RSTUDENT,

$e_i = y_i - \hat{y}$ - the *i*-th residual,

MSE$_i$ - the error variance estimated without the *i*-th observation

$h_i = \mathbf{x}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i$ − the *i*-th diagonal of the *hat matrix* (projection matrix).

If $|RSTUDENT| \geq 2$ then the observation is identified as an outlier, see Table 2.

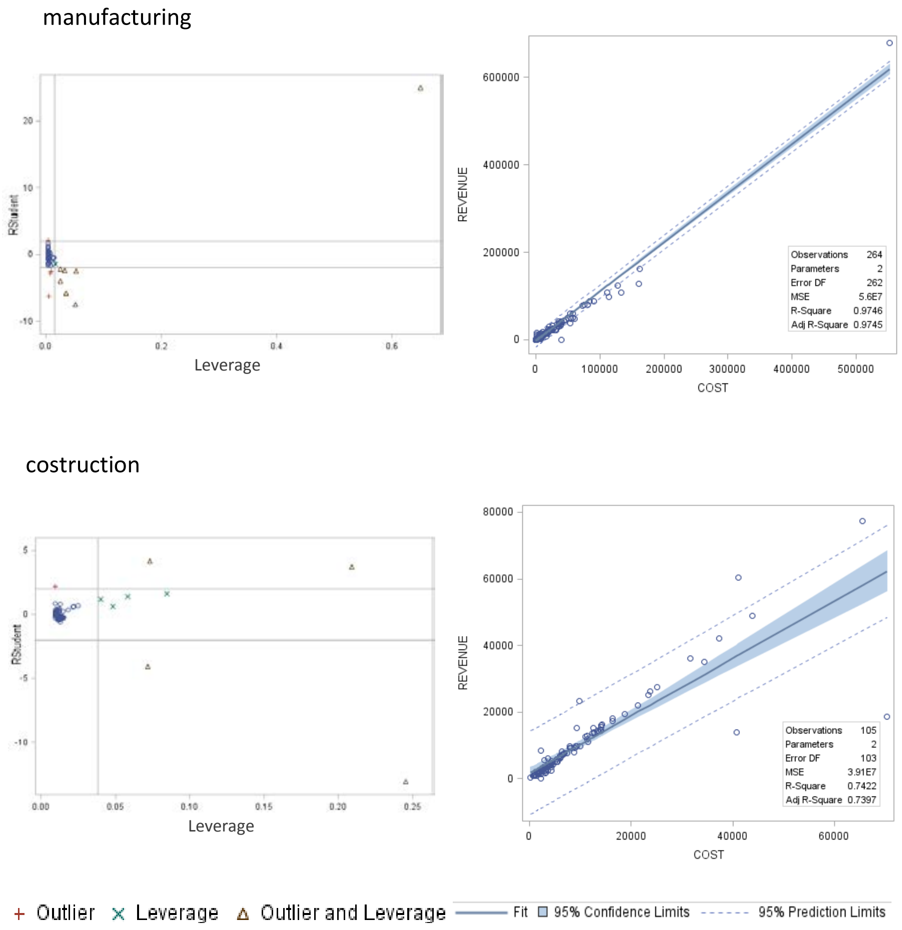
Table 2. Number of enterprises and percentage of outliers for *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

NACE / Provinces	Number of observations	Percentage of outliers	Number of observations	Percentage of outliers
	Dolnośląskie		*Lubuskie*	
manufacturing	772	1,6	368	3.3
construction	207	4,3	56	3.6
trade	315	1,3	149	4.7
transport	80	3,8	46	4.3

Source: Own calculations based on DG1 survey and tax register from December 2011.

It is important not only to identify outliers but also to classify them into types. This can be achieved by calling on the graphical tool − scatter plot of the RSTUDENT versus Leverage (the leverage statistic measures how far the observation is from the centroid of the *x*-space), see Fig. 2, 3. In order to better illustrate the relationship between the study variable *revenue* and auxiliary variable *cost*, scatter plots (with 95% confidence intervals) are presented, see Fig. 2, 3. We can distinguish tree types of extreme values (Rousseeuw, Leroy, 1987): *outliers in the y-direction* (in Fig. 2,3 denoted as *outliers), outliers in the* *x-direction* (in Fig. 2, 3 denoted as *Leverage)*, and *good leverage points*. Graphing relationships among variables reveal exceptions to general rules. The picture in the Figure 2 shows *outliers in the y-direction, in the x-direction and good leverage points.*

The presence of o*utliers in the y-direction* affects the estimated intercept of LS. *Good leverage points* that are outlying in the space of explanatory variables and are located close to the regression line do not affect the LS estimation. Their presence has an impact on statistical inference by deflating the estimated standard errors. The presence of *outliers in the x-direction* that are both outlying in the space of explanatory variables and located far from the regression line influences the LS estimation of both the intercept and the slope (Verardi, Croux, 2009).


manufacturing

costruction

Figure 2. Outlier and Leverage diagnostic for *manufacturing and construction in* Dolnośląskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

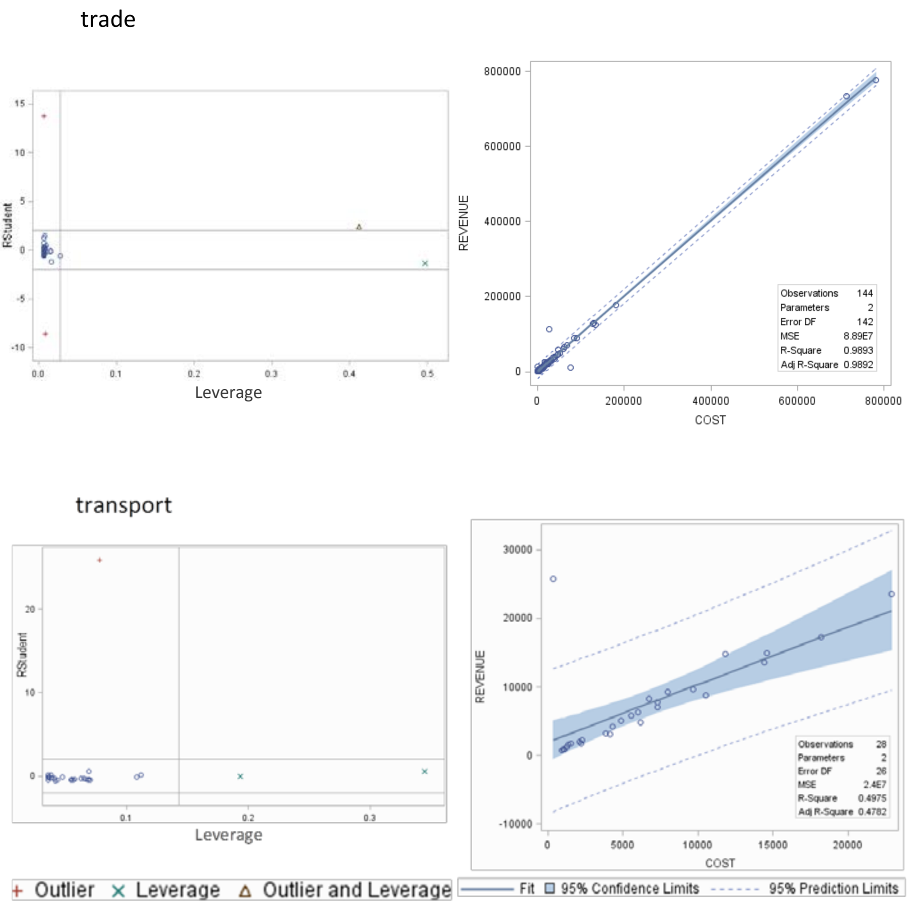
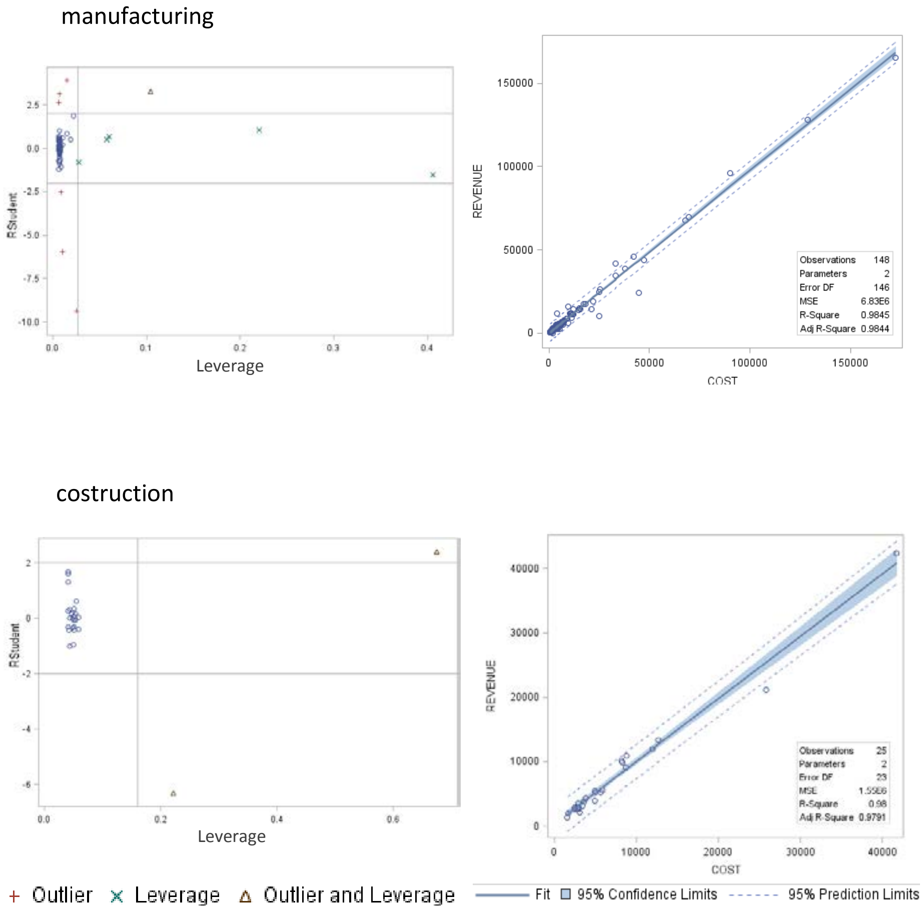


Figure 2a. Outlier and Leverage diagnostic for *trade and transport in* Dolnośląskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.



manufacturing

costruction

Figure 3. Outlier and Leverage diagnostic for *manufacturing and construction in* Lubuskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

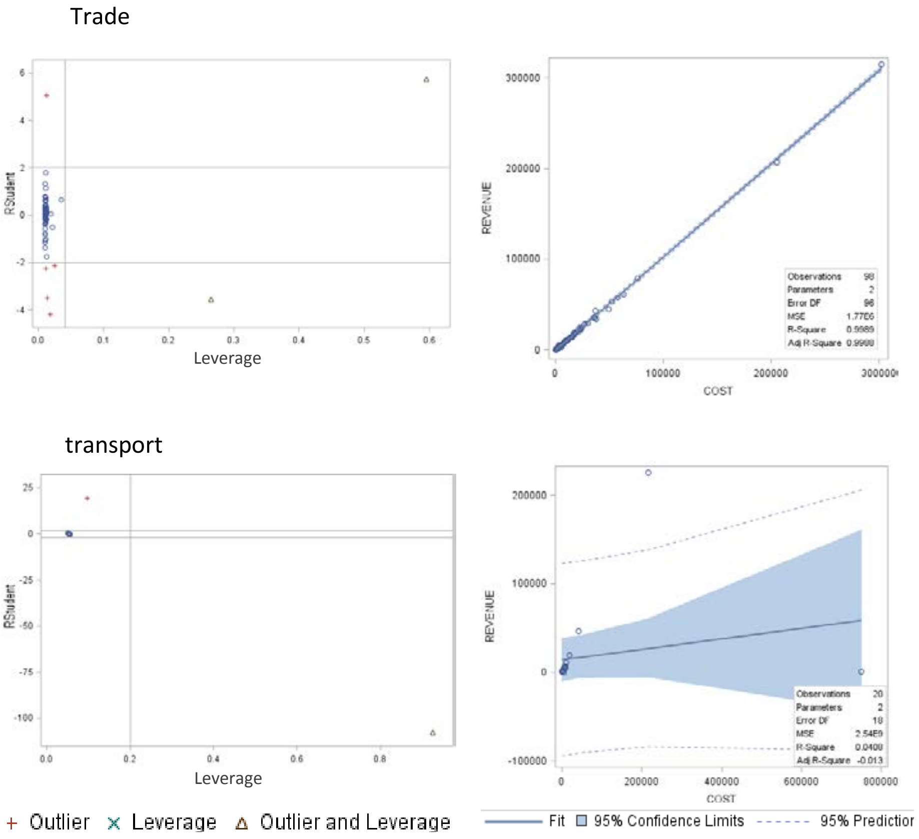


Figure 3. Outlier and Leverage diagnostic for *trade and transport in* Lubuskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

Outliers violate the assumption of normally distributed residuals in the least squares regression. It means that the estimation of model parameters using classical LS is not credible. Therefore, four robust regression methods were applied. The objective of this study was to compare *M-estimation*, *LTS*, *S-estimation* and *MM-estimation* against the LS regression estimation method in terms of the goodness of fit of the model that is represented by the coefficient of determination. The robust version of the coefficient of determination is defined as:

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum \rho\left(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} \tag{9}$$

where ρ is the objective function for the robust estimate, $\hat{\mu}$ is the robust location estimator, and $\hat{s}$ is the robust scale estimator in the full model.

The results are presented in Table 3. The variation between the robust regression methods reflects their inherent sensitivity to the presence of outliers, see Fig. 2, 3. Moreover, their performance (and the resulting estimates) also depends on the type of outliers and their distance from the bulk of the data. The use of *M-estimation* improves the goodness of fit of the model only if the contamination is mainly in the *y*-direction. This limitation also applies to *MM-estimation*, which is determined by *M-estimation*.

Table 3. Coefficient of determination regression methods for *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

auxiliary variables: cost

Regression methods	**LS**	**M**	**LTS**	**S**	**MM**
NACE / PROVINCE	**Dolnośląskie**				
manufacturing	0.975	0.648	0.980	0.978	0.760
construction	0.742	0.740	0.979	0.984	0.699
trade	0.989	0.703	0.994	0.994	0.772
transport	0.498	0.717	0.979	0.977	0.707
Lubuskie					
manufacturing	0.985	0.673	0.980	0.975	0.662
construction	0.980	0.718	0.979	0.970	0.773
trade	0.999	0.704	0.997	0.996	0.775
transport	0.041	0.016	0.934	0.851	0.791

auxiliary variables: income. cost. *revenue*

Regression methods	**LS**	**M**	**LTS**	**S**	**MM**
NACE / PROVINCE	**Dolnośląskie**				
manufacturing	0.996	0.677	0.998	0.998	0.768
construction	0.815	0.695	0.999	0.997	0.766
trade	0.991	0.703	0.999	0.998	0.770
transport	0.510	0.800	0.999	0.998	0.810
Lubuskie					
manufacturing	0.991	0.680	0.997	0.992	0.778
construction	0.995	0.854	0.986	0.981	0.823
trade	0.999	0.744	0.999	0.999	0.780
transport	0.050	0.070	0.966	0.862	0.845

Source: Own calculations based on DG1 survey and tax register from December 2011.

A closer look at section *transport* in Lubuskie province reveals a low value of the coefficient of determination for LS. The characteristics of the regression models for this domain are presented in Table 3. As can be seen, the use of *M-estimation*, in the presence of leverage points, had an adverse effect on the goodness of fit of the model. The value of the coefficient of determination decreased from 0.041 to 0.016. Both in the case of LS and *M-estimation*, the slope estimate equals 0.06 and -0.004 respectively, which means a lack of correlation between variables, see. Fig. 4. *Cost* is an insignificant auxiliary variable (its *p-value* equals 0.4), see Table 4. A considerable improvement in the goodness of fit can be observed for *S-estimation*, which is characterized by a high breakdown point. The *S-estimation* method performs better than *M-estimation* and *MM-estimation* because the data contains high leverage points.

The best goodness of fit was obtained for LTS (R^2 = 0.934), where the model accounts for 0.82% observations – observations with the largest squared residuals are excluded ($h = 36$, n=*46*).

As we can notice, the percentage of outliers for this domain equals 4.3%, see Table 2.

This means that the number of removed observations is higher than the number of outliers. None of the outlying observations affects the parameter of estimation. Unfortunately, some good data was excluded from the estimation, as should be expected.

Table 4. Characteristics of the regression models for *revenue* based on cost, transport, Lubuskie province

METHOD	Parameter	Estimate	Standard error	95% Confidence interval		p-value
LS	**intercept**	14256.241	11874.01	7.20	28504.80	0.2455
	COST	0.059	0.07	0.00	0.12	0.3933
M	**intercept**	3621.988	822.48	2009.95	5234.03	<.0001
	COST	-0.004	0.01	-0.01	0.01	0.4183
LTS	**intercept**	-590.616	207.13	-996.58	-184.65	0.0044
	COST	1.044	0.01	1.04	1.05	<.0001
S	**intercept**	-571.268	292.29	-1144.14	1.61	0.0506
	COST	1.045	0.01	1.03	1.06	<.0001
MM	**intercept**	-960.423	487.64	-1916.18	-4.67	0.0489
	COST	1.049	0.01	1.03	1.07	<.0001

Source: Own calculations based on DG1 survey and tax register from December 2011.

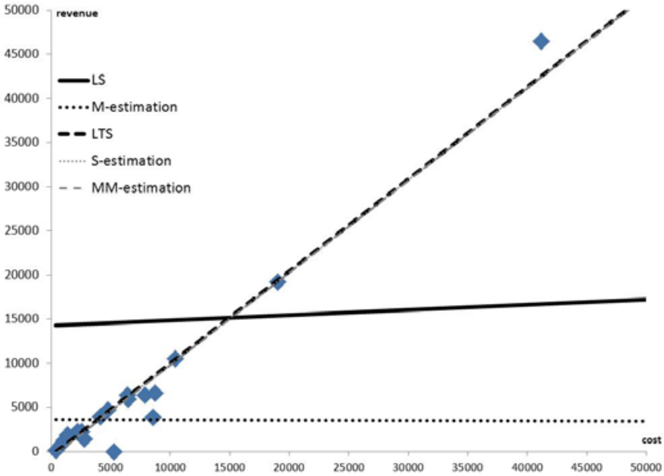


Figure 4. Regression lines for the data using LS, M-estimation, LTS, S-estimation and MM-estimation Methods, TRANSPORT, Lubuskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

Figure 4 presents information contained in Tables 3 and 4. The scatterplot displays a roughly linear relationship between two variables *revenue* and *cost*. The LS and *M-estimation* are immediately affected by the leverage points, so the estimated slope is close to zero – in this case *cost* turns out to be insignificant. After applying LTS, *S-estimation* and *MM-estimation* auxiliary variable *cost* became significant, see Table 4. Regression lines for *LTS, S-estimation* and *MM-estimation* coincide.

7. Conclusion

Conclusions about the assessment of robust regression methods drawn on the basis of the study overlap with those published in the literature of the subject (concerning the properties of the estimators in question):

- in general, the use of the M-estimator in the presence of outliers tends to improve the efficiency and reduce the bias compared to the classical methods of estimation,
- the M-estimator is not robust with respect to high leverage points, so it should be used in situations where high leverage points do not occur,
- the LTS method can be very efficient, but only under specific circumstances – when the number of trimmed observations is equal to the number of outliers. If there are more outliers than trimmed observations, the efficiency of LTS

method is low. In turn, if there is more trimming than there are outlying data points, then some good data will be excluded from the estimation,

- the MM-estimator is affected by the M-estimation, so it should be used with caution given the presence of high leverage points,
- robust regression methods can considerably improve estimation precision, but they should not be automatically applied instead of the classical methods.

REFERENCES

ALMA, Ö., G., (2011). Comparison of Robust Regression Methods in Linear Regression, [in:] Int. J. Contemp. Math. Sciences, Vol. 6, no. 9, pp. 409−421.

CHEN, C., (2007). Robust Regression and Outlier Detection with the ROBUSTREG Procedure, SUGI, http://www2.sas.com/proceedings/sugi27/pp.265-27.pdf.

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

HUBER, P. H., (1964). Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35, pp.7−101.

HUBER, P. H., (1981). *Robust Statistics*, New York: John Wiley and Sons.

ROUSSEEUW, P. J., (1984). Least Median of Squares Regression, [in:] *Journal of the American Statistical Association*, 79, pp. 871−880.

ROUSSEEUW, P. J., YOHAI, V., (1984). Robust regression by means of S-estimators, [in:] W. H. J. Franke and D. Martin (Editors.), Robust and Nonlinear Time Series Analysis, Springer-Verlag, New-York, pp. 256−272.

ROUSSEEUW, P. J., LEROY, A. M., (1987). Robust Regression and Outlier Detection. Wiley-Interscience, New York.

ROUSSEEUW, P. J., DRIESSEN, K., (1998). Computing LTS regression for large data sets, Technical Report, University of Antwerp.

STROMBERG, A. J., (1993). Computation of high breakdown nonlinear regression parameters, [in:] Journal of the American Statistical Association, 88 (421).

VERARDI, V., CROUX, C., (2009). Robust regression in Stata, [in:] The Stata Journal, 9, Number 3, pp. 439−453.

YOHAI,V.J., (1987). High breakdown-point and high efficiency robust estimates for regression, The Annals of Statistics, 15, pp. 642−656.