PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1162 2007
Applications of Mathematics and Statistics in Economics

Dagmar Blatná

University of Economics, Prague, Czech Republic

ROBUST REGRESSION

1. Introduction

We consider a simple linear regression model

$$Y = X\beta + \varepsilon, \tag{1}$$

where $Y = (Y_1, ..., Y_n)'$ is a response variable (dependent variable), $X = (x_{ij})$, $i = 1, ..., n$; $j = 1, ..., p$ is known design matrix of measurements (independent variables), $\beta = (\beta_1, ..., \beta_p)'$ is a vector of unknown regression coefficients and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ is a vector of independent identically distributed residuals with a distribution function F, n is the sample size.

Our problem is to estimate β. The least squares method (*LS*) is based on assuming normality of errors. The normal model is never exactly true in the practice, and in the presence of small departures from the normality assumption on the errors *LS* procedures lost efficiency drastically. A bad behaviour of the estimator $\hat{\beta}_j$ may be caused not only by outliers, i.e. the points which are deviated in the response variable, but also by points which are (considerably) far away from the bulk of data in the factor space which are usually called *leverage points*. The leverage points are outliers, but they are not errors – it would be more appropriate to say that the data come from two different populations.

The degree of robustness of an estimate in the presence of outliers may be measured by the concept of the *breakdown-point* which was introduced by Hampel [8]. Donoho [6] and Donoho and Huber [7] gave a finite sample version of this concept. The *breakdown value* of an estimator is defined as the smallest fraction of contamination that can cause the estimator to take on values arbitrarily far from its value on the uncontamined data.

The *LS* estimator (*LSE*) has a finite breakdown point equal to zero. This means that a single point, properly placed, can cause the *LSE* to have virtually any value.

2. Robust estimate in regression

Many methods have been developed in response to problems with outliers in regression. In further text only the simple linear regression problem will be interested.

2.1. *M*-estimators

An *M*-estimate (*ME*) of β is defined by minimizing a sum of functions ρ of residuals

$$\text{minimize} \sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right), \tag{2}$$

where residual $r_i = y_i - \beta x_i$.

After taking derivatives, as a solution of a system of equations

$$\sum_{i=1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right) x_i = 0 \tag{3}$$

with $\psi = \rho'$. If ρ is convex and ψ continuous, the definitions in equations (2) and (3) are equivalent. Various ψ-functions lead to various *ME*s.

The *M*-estimates with monotone ψ-function have breakdown point equal to 0.

2.2. Generalized *M*-estimators (*GME*)

Because of sensitivity of *M* regression to leverage points, generalized *M*-estimators (*GME*) were introduced, with the basic purpose of bounding the influence of outlying x_i by means of some weight function *w*.

*GME*s replace the squaring operator of *ME* by a function that increases less rapidly for large residuals. In the Mallows' version outliers and influential observation automatically receive less weight by relation

$$\sum_{i=1}^{n} w(x_i)\psi\left(\frac{r_i}{\hat{\sigma}}\right) x_i = 0. \tag{4}$$

A disadvantage of this proposal is that it assigns less weight to both "good" and "bad" leverage points.

The Schweppe's form of *GME* only downweights vertical outliers and bad leverage points, but not good leverage points

$$\sum_{i=1}^{n} w(x_i)\psi\left(\frac{r_i}{w(x_i)\hat{\sigma}}\right)x_i = 0. \qquad (5)$$

The breakdown of these estimates is at most $1/(p+1)$ and tends to zero when p increases.

2.3. The estimators with high breakdown point (*HBPE*)

*HBPE*s have simultaneously the breakdown point independent of the dimension of the problem (and near 0.5) and are highly efficient when the errors have a normal distribution.

Siegel [16] was the first who introduced a regression estimator with asymptotic breakdown point equal to 0.5. However, his method of repeated medians has been never used in the practice being too complicated for routine calculations. Rousseeuw and Leroy [14] and others introduced the following high breakdown value estimators for linear regression.

These robust methods have succeeded in staying away from the outlier, and yield a good fit to the majority of the data. Moreover, they lie close to the *LS* estimate applied to the uncontaminated data. It would be wrong to say that robust techniques ignore outliers. On the contrary, the regression with *HBP* fit exposes the presence of such points.

***LMS*-estimator (*LMSE*).** Rousseeuew [12] proposed the least median of squares (*LMS*) procedure. The *LMS* solution for simple regression with intercept is given by

$$\text{minimize } \operatorname*{med}_{i}(y_i - b_0 - b_1 x_i)^2. \qquad (6)$$

Geometrically, it corresponds to finding the narrowest strip covering the half of observations ("half" means $[n/2]+1$, where $[n/2]$ denotes the integer part of $n/2$). The *LMS* line lies exactly at the middle of this band.

The *LMS* method has relatively unsatisfactory asymptotic properties, the convergence to normality is relatively slow, and it is relatively inefficient especially when all the observations satisfy the regression model with normal errors. The *LMSE*s are used primarily as a diagnostic and exploratory tool to detect the regression outliers and leverage points.

This estimator is very robust with respect to outliers in y as well as ones in x. Its breakdown point is 50%. For *LMS* regression the regression parameters are given, together with a corresponding scale estimate. This scale estimate is also defined in a robust way.

The *LMS* also possesses a measure to determine how well the fitted model explains the observed variability in y. In analogy to the classical one, we called it also R^2 or coefficient of determination. In the case of regression with constant term, it is defined by

$$R^2 = 1 - \left(\frac{\operatorname{med}|r_i|}{\operatorname{mad}(y_i)} \right)^2 . \qquad (7)$$

The abbreviation "mad" stands for median absolute deviation, defined as

$$\operatorname{mad}(y_i) = \operatorname*{med}_i \left\{ \left| y_i - \operatorname*{med}_j y_j \right| \right\} . \qquad (8)$$

***LTS*-estimator (*LTSE*).** The least trimmed squares (*LTS*) estimator proposed by Rousseeuw [12] are obtained by

$$\text{minimize} \sum_{i=1}^{h} r_{(i)}^2 , \qquad (9)$$

where $r_{(i)}$ is the i-th order statistic among the squared residuals written in the ascending order, $h = [n/2] + [(p+1/2)]$ and $[x]$ denotes the largest integer which is less or equal to x.

The *LTSE* has a relatively low asymptotic efficiency and plays a role in the class of *MM*-estimators.

In SAS, the initial *LTS* estimate is computes using h done as an integer between $\left[\frac{n}{2}\right]+1$ and $\left[\frac{3n+p+1}{4}\right]$. By default h = $\left[\frac{3n+p+1}{4}\right]$, which corresponds to a breakdown value of around 25%.

LTS is very similar to *LS*, the only difference being that the largest squared residuals are not used in the summation, thereby allowing the fit to stay away from outliers. The best robustness properties are achieved when h is approximately $n/2$, in which case the breakdown point attains 50%.

***S*-estimator (*SE*).** Generalizing the *LMS* and the *LTS* estimators, Rousseeuw and Yohai [15] introduced so called *S*-estimator, corresponding to

$$\text{minimize } S(T), \qquad (10)$$

where $S(T)$ is a certain type of robust M-estimate of the scale of the residuals $r_1(T),...,r_n(T)$. Their breakdown point can also attain 50%. But *SE*s have essentially the same asymptotic performance as regression M-estimators.

Reweighted least squares (*RLS*) regression. In order to improve on the crude *LMS* and *LTS* solutions, and in order to obtain standard quantities (confidence interval, t-values) we can apply a weighted least squares analysis based on the identification of the outliers. This corresponds to minimizing the sum of the squared residuals multiplied by a weight w_i:

$$\text{minimize} \sum_{i=1}^{n} w_i r_i^2, \tag{11}$$

where the weights w_i are determined from the *LMS* solution. The effect of the weights, which can only take values 0 or 1, is the same as deleting the cases for which w_i equals zero.

Therefore, the *RLS* can be seen as ordinary *LS* on a "reduced" data set consisting of only those observations that received a nonzero weight. Because this reduced data set does not contain regression outliers anymore, the statistics and inference are more trustworthy than those associated with *LS* on the whole data set.

The determination coefficient for *RLS* is defined in an analogous way as for *LS*, but all terms are now multiplied by their weight w_i. The resulting estimator still possesses the high breakdown point, but is more efficient in a statistical sense.

***MM*-estimators (*MME*).** In terms of simultaneously achieving a high breakdown point, a bounded influence function, and a high asymptotic efficiency under the normal model, some other methods were suggested, among them those belonging to the class of *MM*-estimates of Yohai [20].

The *MM*-estimates are defined by a three-stage procedure. At the first stage an initial regression estimate is computed which is consistent, robust, has high breakdown-point but is not necessarily efficient. At the second stage an M-estimate of the errors scale is computed using residuals based on the initial estimate. Finally, at the third stage, a (final) M-estimate of *MM* estimates is a combination of high breakdown value estimation and efficient estimate of the regression parameters based on a proper redescending ψ-function is computed.

The robust version of R-squared for *MM* estimate is defined as

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum \rho\left(\frac{y_i - x_i\hat{\beta}}{\hat{s}}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)}, \qquad (12)$$

where $\rho' = \psi$, $\hat{\beta}$ is the *MM* estimator of β, $\hat{\mu}$ is the *MM* estimator of location, and $\hat{s}$ is the *MM* estimator of the scale parameter in the full model.

The asymptotic efficiency of the *MM*-estimates with respect to the *LSE* is independent of the distribution of the explanatory variables.

3. Illustration example

We can present the usefulness of robust regression methods on a simple example of simple regression analysis. We generated 50 values according to

$$y = 1 + 3x + \varepsilon_i, \quad x = 1, \ldots, 50,$$

where ε_i is random number from normal distribution $N(0;4)$.

Nine different sets were analysed:

- set *A* consists of generated uncontaminated data,
- sets *B*, *C*, *D* and *E* contain 2%, 10%, 20% or 30% outliers in *y* (*p*% values y_i were increased by 100),
- sets *F*, *G*, *H*, *I* contain 2%, 10%, 20%, 30% outliers in *x* (*p*% values x_i were increased by100).

In Tables 1, 2, 3 results of estimates of regression parameters (the intercept b_0 and the slope b_1) for the nine sets of data obtained by *LS*-method and by some robust methods are presented. The results of analyses were obtained using the SAS 9.1.3 system.

The following methods were applied:

- least squares methods (*LS*),
- Huber-type *M* – regression (*M*),
- least trimmed squares regression with breakdown point value 26% (*LTS*26), which corresponds to default value of $h = [(3n + p + 1)/4]$,
- reweighted least squares regression connected with *LTS*26 (*RWLS*26),
- least trimmed squares regression with break down point 42% (*LTS*42),
- reweighted least squares regression connected with *LTS*42 (*RWLS*42),
- *S*-regression (*S*),
- *MM* – regression based on initial *LTSE* and at the second step the Tukey function ψ(*MM*).

Table 1. Regression parameters for the sets *A*-*C*

Method	Set *A*			Set *B*			Set *C*		
	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2
LS	1.382	2.950	0.992	–2.618	3.185	0.929	–16.169	4.029	0.839
M	1.299	2.957	0.873	0.966	2.974	0.856	1.100	2.969	0.787
LTS(26)	0.981	2.989	0.954	0.981	2.300	0.994	0.372	3.003	0.994
FWLS	1.382	2.950	0.992	0.980	2.973	0.993	1.084	2.967	0.996
LTS(42)	–	–	–	3.392	2.906	0.996	2.772	2.927	0.915
FWLS	1.382	2.950	0.992	0.980	2.973	0.993	1.084	2.967	0.996
S	1.018	2.975	0.993	0.955	2.978	0.993	1.109	2.969	0.992
MM	1.168	2.966	0.793	0.980	2.976	0.791	1.103	2.969	0.769

Table 2. Regression parameters for the sets *D*-*F*

Method	Set *D*			Set *E*			Set *F*		
	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2
LS	–27.598	4.870	0.853	–32.904	5.470	0.886	33.758	1.559	0.404
M	–26.704	4.821	0.722	–32.904	5.494	0.823	1.196	2.960	0.859
LTS(26)	1.167	2.960	0.990	–27.164	5.759	0.871	–1.647	3.064	0.994
FWLS	1.395	2.947	0.995	–32.904	5.470	0.886	1.309	2.952	0.996
LTS(42)	1.039	2.992	0.992	0.603	2.978	0.989	6.026	2.832	0.995
FWLS	1.395	2.947	0.995	1.478	2.940	0.993	1.309	2.952	0.996
S	1.407	2.948	0.985	–32.882	5.588	0.885	0.820	2.979	0.992
MM	1.403	2.947	0.713	–32.916	5.533	0.610	1.059	2.968	0.797

Table 3. Regression parameters for the sets *G*-*I*

Method	Set *G*			Set *H*			Set *I*		
	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2
LS	76.669	0.003	0.0001	101.785	–0.567	0.156	120.514	–0.830	0.378
M	75.637	0.028	0.0002	102.101	–0.599	0.153	119.970	–0.845	0.362
LTS(26)	–1.996	3.073	0.993	0.775	2.983	0.990	98.043	–0.706	0.608
FWLS	–0.676	3.022	–	0.599	2.973	0.994	120.514	–0.830	0.378
LTS(42)	3.649	2.900	0.994	5.672	2.282	0.993	2.198	2.964	0.990
FWLS	–0.037	2.991	0.996	0.600	2.973	0.994	2.772	2.916	0.993
S	–0.303	3.009	0.989	0.607	2.975	0.958	117.441	–0.850	0.418
MM	–0.207	3.003	0.770	0.605	2.974	0.710	119.285	–0.851	0.354

As an illustration, results for sets *B*, *F*, *G* are presented in the following figures.

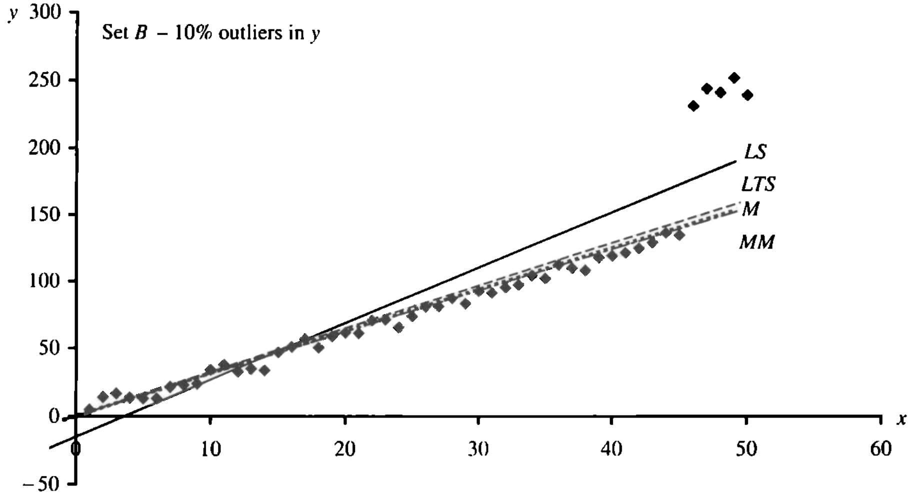

Fig. 1. Results of estimation of regression parameters for the set *B*

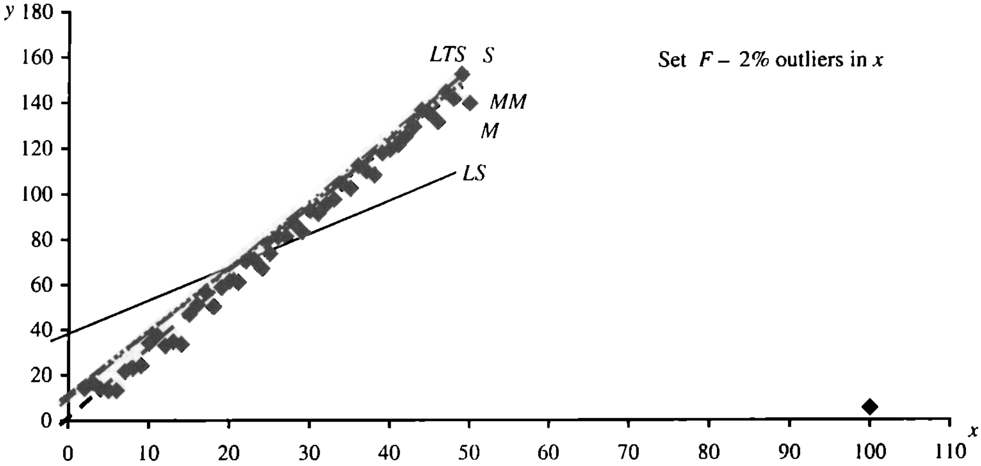

Fig. 2. Results of estimation of regression parameters for the set *F*

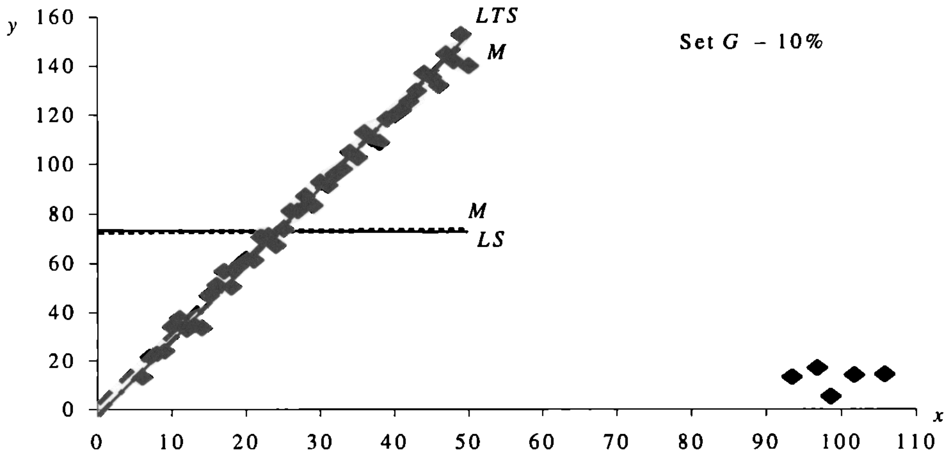

Fig. 3. Results of estimation of regression parameters for the set *G*

It is obvious from the results of this illustration example that all robust regression methods give relatively stable estimates of regression parameters in the case of uncontaminated sample (set *A*) – they are close to the "right" value 3.0.

LSE for the underlying model are getting worse with increasing percentage of contamination, in case of higher contamination in *x*-direction give even reverse direction of dependence. Several unusual *x's* completely dominated the *LSE* of the regression parameters.

Huber *MEs* reduced influence of small percentage of outliers in the *y*-direction, with increasing stage these estimations are getting worse. But *M*-estimations are not robust with respect to leverage points (outliers in *x*-space, sets *F, G, H, I*), that is, like in case *LS*, outlier *x*'s completely dominated the estimate of the regression parameters.

Robust methods with high breakdown point (*LTS, S, MM*) eliminate influence of outliers in both directions and provide resistant results in presence of outlier in the response variable *Y* (sets *B, C, D*), as well as in presence of outlier in independent variable *X* (sets *F, G, H*).

As to higher contamination (in our example 30%, sets *E, I*), the estimates with default values of parameters fail but *LTS*s with breakdown value 42% give values of regression coefficient close the "right" value as well.

4. Conclusions

The choice of acceptable method depends on type of contamination and data quality. In contrary to the classical *LS* regression, the results of robust methods need much more careful interpretation. No method can cover all problems, especially when processing real data. No method dominates all others in all situations. Moreover, any method might yield substantially different conclusions from many of the other methods that might be used. It seems that at some point, we should consider two or more methods and, if discrepancies arise, try to understand why.

References

[1] Antoch J., Vorlíčková D., *Vybrané metody statistické analýzy dat*, Academia, Praha 1992.

[2] Antoch J., Ekblom H., Víšek J.A., *Robust Estimation in Linear Model*, XploRe Macros: http://www.quantlet.de/codes/rob/ROB.htlm 1999.

[3] Blatná D., *Practical Reasons for Using Robust Regression*, [in:] *Aplimat 2005, Part I*, Slovak University of Technology, Bratislava 2005, pp. 255–260.

[4] Blatná D., *Robust Approach in Regression*, [in:] *Applications of Mathematics and Statistics in Economy*, Professional Publishing, Praha 2004, pp. 48–53.

[5] Bryndák M., *Některé aspekty robustní regrese*, Diplomová práce, VŠE, Praha 2001.

[6] Donoho D.L., *Breakdown properties of multivariate location estimators*, Ph.D. qualifying paper, Department of Statistics, Harvard University, Cambridge, Mass. 1982.

[7] Donoho D.L., Huber P.J., *The notion of breakdown point*, [in:] *A Festschrift for Erich L. Lehmann*, P.J. Bickel, K.A. Doksum, J.L. Hodges Jr. (eds.), Wadsworth, Belmont, Calif. 1983, pp. 157-184.

[8] Hampel F.R., "A general qualitative definition of robustness", *Annals of Mathematical Statistics* 1971, 42, 1887-1896.

[9] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A., *Robust Statistics. The Approach Based on Influence Functions*, John Wiley, New York 1986.

[10] Huber P.J., *Robust Statistics*, John Wiley, New York 1981.

[11] Jurečková J., *Robustní statistické metody*, Skripta MFF UK, Karolinum, Praha 2001, 132 pp.

[12] Rousseeuw P.J., "Least median of squares regression", *Journal of American Statistical Association* 1984, 79, 871-880.

[13] Rousseeuw P.J., Croux Ch., "Alternatives to the Median Absolute Deviation", *Journal of American Statistical Association* 1973, Vol. 88, 1273-1283.

[14] Rousseeuw P.J., Leroy A.M., *Robust Regression and Outlier Detection*. John Wiley, New York 1987.

[15] Rousseeuw P.J., Yohai V., *Robust regression by means of S-estimators*, [in:] *Robust and Nonlinear Time Series Analysis*, J. Franke, W. Haerdle and R.D. Martin (eds.), Lecture Notes in Statistics 26, Springer, New York 1984, pp. 256-272.

[16] Siegel A.F., "Robust regression using repeated medians", *Biometrika* 1982, 69, 242-244.

[17] Welsh A.H., "On M-Processes and M-Estimation", *The Annals of Statistics* 1989, Vol. 17, No. 1, 337-381.

[18] Wilcox R.R., *Fundamentals of Modern Statistical Methods*, Springer, New York 2001.

[19] Wilcox R.R., *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, London 1999.

[20] Yohai V.J., "High Breakdown-Point and High Efficiency Robust Estimates for Regression", *The Annals of Statistics* 1987, Vol. 15, 642-656.

REGRESJA ODPORNA

Streszczenie

Regresja należy do najczęściej używanych metod statystycznych. Klasyczne podejście statystyczne – estymacja parametrów za pomocą metody najmniejszych kwadratów *LSM* (*least squares methods*) jest oparta na założeniu normalności błędów. *LSM* może być bardzo niezadowalająca w przypadku istnienia obserwacji nietypowych. Regresja odporna jest ważnym narzędziem analizy danych zanieczyszczonych obserwacjami nietypowymi. Gdy zanieczyszczenie występuje głównie w kierunku *y* *M*-estymatory mogą być wykorzystane, ale ta metoda nie chroni przed punktami wpływowymi (obserwacje nietypowe w przestrzeni *x*). Metody odporne z wysokim punktem załamania (*LMS*, *LTS*, *S*) eliminują wpływ obserwacji nietypowych w obu kierunkach. Estymacja MM łączy estymację z wysokim punktem załamania i *M*-estymację.

Słowa kluczowe: regresja liniowa, estymatory odporne, *M*-regresja, metody z wysokim punktem załamania.