ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Dorota Pruska**

DISPERSION OF ESTIMATES OF LINEAR REGRESSION PARAMETERS IN CASE OF THE DEEPEST REGRESSION METHOD

ABSTRACT. The deepest regression method is such a method of estimation of regression parameters that the maximal regression depth characterises the obtained model.

In this paper the deeepest regression method is presented and the simulation analysis (Monte Carlo experiments) of dispersion of linear regression parameter estimates is conducted in case of data sets with different numbers of outliers. On the basis of the results of Monte Carlo experiments the characteristics of distribution of regression parameter estimates are determined and compared with the results of analogous experiments conducted with the use of the least square method.

Key words: the deepest regression method, outliers, dispersion, breakdown value.

I. INTRODUCTION

In the paper we analyse and compare the results of Monte Carlo experiments dealing with dispersion of estimates of linear regression parameters for two methods of estimation: the deepest regression method (DRM) and the least squares method (LSM) for data sets with outliers. An observation is called an outlier in a set of observations, if its distance from majority of other observations is significantly greater than the distance between the majority of pairs of other observations. In regression analysis the residuals can be used for detecting the outliers (see Zeliaś (1996), Ostasiewicz (1998), Domański, Pruska, (2000)).

II. THE DEEPEST REGRESSION METHOD

The deepest regression method was proposed by P. J. Rousseeuw and M. Hubert (see Rousseeuw, Hubert (1999)). The idea of DRM is to estimate regression parameters in such a way that the regression depth is maximal for the obtained

* MSc, Chair of Statistical Methods, University of Łódź.

model. A median which is robust for outliers is used in the estimation algorithm (see Van Aelst et al. (2000)). The DRM is a nonparametric method of estimation.

The regression depth (*rdepth*) gives information how well the model fits the data set. Consider the data set $Z_n = \{(x_{i1}, ..., x_{i,p-1}, y_i); i = 1,...,n\} \subset R^p$ and the following linear model:

$$y = \theta_1 x_1 + ... + \theta_{p-1} x_{p-1} + \theta_p \tag{1}$$

describing dependence between variables Y and X, whose realizations are respectively: y_i and $(x_{i1}, ..., x_{i,p-1})$, $i=1,..., n$. Let θ be a vector of model parameters and $\theta=[\theta_1,..., \theta_p]^T$. The regression depth for the model (1) with vector parameters θ for data set Z_n is defined as follows:

$$rdepth(\theta,Z_n)=\min_{u,v} \{\text{card}\{i: r_i \geq 0 \text{ and } x_i^T u<v\} + \text{card}\{i: r_i \leq 0 \text{ and } x_i^T u>v\}\}, \tag{2}$$

where $r_i = y_i - (\theta_1 x_{i1} + ... + \theta_{p-1} x_{i,p-1} + \theta_p)$ and u is versor in R^{p-1}, $v \in R$, $x_i^T u \neq v$, $(x_i^T, y_i) \in Z_n$.

In the DRM we assume that the model (1) is the best fitted to the data set Z_n for such a vector of parameters θ for which the regression depth is maximal. The estimator obtained by DRM for Z_n is described by the following equation (see Rousseeuw, Hubert, (1999)):

$$T_r^*(Z_n) = \arg\max_{\theta} rdepth(\theta, Z_n). \tag{3}$$

In applications of DRM the algorithm MEDSWEEP can be used. It was presented in the paper written by Van Aelst et al. (2000).

In the deepest regression method the finite-sample addition breakdown value ε_n^* of an estimator T_n is defined as the smallest fraction of outliers which, added to the data set Z_n, make the estimator unrobust (see Van Aelst et al. (2000)). Let Z_{n+m} be the data set obtained by adding m outliers to Z_n. The breakdown value is of the form:

$$\varepsilon_n^*(T_n, Z_n) = \min\left\{\frac{m}{m+n} : \sup_{Z_{n+m}} \left\| T_{n+m}(Z_{n+m}) - T_n(Z_n) \right\| = \infty\right\}. \tag{4}$$

For the dimension $p \geq 2$ the breakdown value has the following property:

$$\varepsilon_n^*(T_n, Z_n) \xrightarrow[n\to\infty]{a.s.} \frac{1}{3}. \tag{5}$$

III. MONTE CARLO ANALYSIS OF DISPERSION OF LINEAR REGRESSION PARAMETERS ESTIMATES FOR THE DRM

The dispersion of DRM-estimates of regression parameters was analysed on the basis of Monte Carlo experiments. We compared dispersion of model parameter estimates for DRM and LSM for different cases of number of outliers in data set.

Some procedures from program MEDSWEEP (presented on a web site www.agoras.ua.ac.be) and procedures of pseudo-random numbers generating (see Zieliński (1979), Brandt (1999)) were used in simulations.

Experiments consist in 1000 repetitions of estimation of model parameters on the basis of samples generated according to two-dimensional normal distribution. Parameters of this distrbution were determined on the basis of data dealing with variables: gross domestic expenditure on research and development activity (y), and employment in research and development activity (x) in voivodships in Poland in 2004, except for mazowieckie and małopolskie i.e. for 14 voivodships. Parameters of the distribution were determined with the use of data presented in table 1. Mazowieckie and małopolskie voivodships are outliers according to the measures presented in the paper edited by Ostasiewicz (1998, p. 249–274) and according to the Dixon test (see Domański (1990)) conducted to each variable x and y separately with an assumption that data from table 1. create random sample.

Data dealing with variables x and y for all voivodships, except for mazowieckie and małopolskie i.e. for 14 voivodships, were treated as a realization of random sample and on the basis of them the hypothesis of normality of two-dimensional distribution was verified. The Shapiro-Wilk test did not reject the hypothesis, so it was assumed in the experiments that the population has a two-dimensional normal distribution $N(\mu_0, \Sigma_0)$, where

$$\mu_0 = \begin{bmatrix} 5403 \\ 161 \end{bmatrix}, \quad \Sigma_0 = \begin{bmatrix} 15686689 & 527382 \\ 527382 & 18502 \end{bmatrix}$$

and elements of vector μ_0 and matrix Σ_0 obtained as the result of estimation on the basis of data on 14 voivodships. Linear regression of variables y with respect to x is of the form:

$$y = 0{,}0336x - 21{,}074. \tag{6}$$

Table 1

Gross domestic expenditure[1] on research and development activity (y) and employment in research and development activity (x) in voivodships in Poland in 2004 (in PLN m)

Voivodship	x	y
Dolnośląskie	9620	289,80
Kujawsko-pomorskie	4718	120,50
Lubelskie	6896	168,00
Lubuskie	1326	23,20
Łódzkie	7748	299,90
Małopolskie*	***17007***	***645,50***
Mazowieckie*	***34702***	***2261,70***
Opolskie	1545	29,40
Podkarpackie	2975	104,00
Podlaskie	2408	51,50
Pomorskie	6646	247,60
Śląskie	12692	402,80
Świętokrzyskie	1124	18,40
Warmińsko-mazurskie	2277	56,30
Wielkopolskie	12136	372,60
Zachodniopomorskie	3536	64,20

*) Małopolskie and Mazowieckie voivodships are treated as outliers.

Source: *Statistical Yearbook of Voivodships 2005*.

According to the distribution $N(\mu_0, \Sigma_0)$ one thousand samples of 14 elements each were generated. Next, to the generated 14-element samples we added two elements generated according to distributions for which expected values were the observed values for małopolskie and mazowieckie voivodships and 6 cases of covariance matrix were considered.

Two cases of correlation coefficients were taken into consideration:

- $r_{x,y} = 0{,}98$,
- $r_{x,y} = 0{,}70$

and for each of them three cases of coefficients of variation were considered:

- $V_x = 0{,}1$ and $V_y = 0{,}1$
- $V_x = 0{,}3$ and $V_y = 0{,}3$
- $V_x = 0{,}5$ and $V_y = 0{,}5$.

On the basis of the above covariance matrixes were determined.

Next, for DRM and LSM one thousand estimates of model parameters

$$y = ax + b, \tag{7}$$

[1] Excluding depreciation of fixed assets.

were calculated. On the basis of 1000 estimates of parameter for each method the mean, minimal and maximal values were determined.

Similar experiments for the data set with four and six outliers were conducted (in the experiments with six outliers their fraction exceeds the breakdown value). According to the normal distribution $N(\mu_0, \Sigma_0)$ 12-element and 10-element samples, instead of 14-element samples, were generated respectively. From the other two distributions 2-element and 3-element samples from each, instead of 1-element samples, were generated. The results obtained for the three groups of experiments are presented in tables 2–4.

In all cases of the experiments with data set containing two outliers the means of estimates of slope for DRM model are about 0,035, which is very close to the real slope given in formula (6). The mean values estimated by LSM are about 0,06 and their range is larger than in case of DRM. Estimates of slope for LSM decrease while coefficients of variation increase.

In each case the mean of estimates of absolute term for DRM belongs to the interval [–26,54; –23,07] and it is close to the parameter given in formula (6). The mean values of free term estimated by LSM are included in the interval (–169; –121). The minimal and maximal values are more differentiated in case of LSM. While coefficients of variation increase, the range of absolute term becomes larger.

For DRM in the experiments with four outliers the mean of the slope estimates is about 0,038 and the mean of free term estimates belongs to the interval (–39; –28). Both estimates are slightly further from the values given in formula (6) than in case of experiment with two outliers. The means of slope estimates and absolute terms estimates obtained by LSM belong to intervals [0,059; 0,065] and [–203; –138], respectively.

In the experiments for data set with six outliers the DRM-estimates of model parameters differ from the ones given in formula (6). In all cases the means of slope estimates exceed 0,04 and the means of absolute term estimates take the value from the interval (–66;–49). All the means of slope estimates obtained by LSM are about 0,06 and the means of absolute term estimates belong to the interval (–235;–193). The model parameter estimates obtained by DRM are closer to the parameters from formula (6) than LSM-estimation of parameters, in spite of the fact that DRM is not robust for such an amount of outliers.

Table 2

Mean, minimal and maximal values from 1000 estimates of parameters of model (1) for two outliers

Method of estimation	Model parameters	Characteristics of parameter estimates	Correlation coefficient					
			$r_{x,y} = 0,7$			$r_{x,y} = 0,98$		
			Coefficients of variation V_x and V_y					
			0,1	0,3	0,5	0,1	0,3	0,5
DRM	*a*	mean	0,035	0,035	0,035	0,036	0,036	0,035
		min	0,024	0,023	0,024	0,024	0,024	0,024
		max	0,051	0,054	0,050	0,051	0,051	0,051
	b	mean	–26,087	–25,014	–23,070	–26,540	–26,447	–25,391
		min	–98,078	–118,165	–123,065	–98,078	–115,162	–123,065
		max	35,239	60,218	50,572	35,239	35,239	48,224
LSM	*a*	mean	0,061	0,059	0,055	0,062	0,060	0,057
		min	0,045	0,020	–0,038	0,051	0,034	0,028
		max	0,079	0,103	0,130	0,071	0,076	0,082
	b	mean	–168,100	–151,245	–121,810	–168,643	–155,588	–140,463
		min	–393,790	–536,158	–548,925	–348,107	–388,532	–432,145
		max	–53,046	40,744	514,395	–64,264	14,104	52,247

Source: own calculations.

Table 3

Mean, minimal and maximal values from 1000 estimates of parameters of model (1) for four outliers

Method of estimation	Model parameters	Characteristics of parameter estimates	Correlation coefficient					
			$r_{x,y} = 0,7$			$r_{x,y} = 0,98$		
			Coefficients of variation V_x and V_y					
			0,1	0,3	0,5	0,1	0,3	0,5
DRM	*a*	mean	0,039	0,038	0,036	0,039	0,038	0,038
		min	0,027	0,018	0,011	0,027	0,027	0,026
		max	0,053	0,068	0,065	0,050	0,062	0,066
	b	mean	–37,895	–35,029	–28,504	–38,888	–37,399	–34,526
		min	–127,093	–187,831	–237,442	–116,273	–187,831	–175,054
		max	40,966	69,390	121,125	40,966	40,966	41,054
LSM	*a*	mean	0,065	0,063	0,059	0,065	0,064	0,062
		min	0,053	0,028	–0,001	0,058	0,044	0,036
		max	0,078	0,098	0,117	0,073	0,076	0,082
	b	mean	–201,202	–177,659	–138,914	–202,641	–190,367	–173,483
		min	–421,784	–500,182	–513,481	–402,726	–396,160	–388,851
		max	–80,314	81,631	388,245	–85,102	–13,428	14,109

Source: own calculations.

Table 4

Mean, minimal and maximal values from 1000 estimates of parameters of model (1) for six outliers

Method of estimation	Model parameters	Characteristics of parameter estimates	Correlation coefficient					
			$r_{x,y} = 0,7$			$r_{x,y} = 0,98$		
			Coefficients of variation V_x and V_y					
			0,1	0,3	0,5	0,1	0,3	0,5
DRM	*a*	mean	0,043	0,044	0,041	0,042	0,043	0,042
		min	0,034	0,029	0,023	0,038	0,034	0,028
		max	0,063	0,075	0,091	0,065	0,070	0,072
	b	mean	–63,516	–65,402	–49,474	–58,963	–60,951	–54,340
		min	–235,496	–324,347	–306,811	–213,002	–325,321	–218,513
		max	9,240	25,146	77,903	6,045	4,323	31,707
LSM	*a*	mean	0,067	0,064	0,060	0,067	0,066	0,064
		min	0,057	0,035	0,014	0,061	0,054	0,043
		max	0,079	0,097	0,112	0,074	0,076	0,079
	b	mean	–213,133	–193,776	–138,854	–234,022	–216,618	–193,106
		min	–466,357	–504,677	–563,365	–438,599	–428,587	–410,251
		max	–88,881	62,752	330,524	–105,107	–69,063	–17,659

Source: own calculations.

IV. CONCLUSIONS

Monte Carlo experiments conducted on the considered distributions and 16-element sample confirm that the deepest regression method is more robust for outliers than the least square method. In all cases of introduced outliers, DRM-estimates of parameters were closer to their values obtained for data set without outliers than LSM-estimates, according to the comparison of obtained mean, minimal and maximal values from 1000 repetitions of parameter estimation for each method. For data set containing so many outliers that both methods were not robust, the deepest regression method gives more precise estimates of model parameters than the least square method. In case of the considered small samples we observe, that if fraction of outliers in data set is near 1/3 the DRM is low robust.

REFERENCES

Brandt S. (1999), *Analiza danych. Metody statystyczne i obliczeniowe*, PWN, Warszawa.
Domański Cz. (1990), *Testy statystyczne*, PWE, Warszawa.
Domański Cz., Pruska K. (2000), *Nieklasyczne metody statystyczne*, PWE, Warszawa.
Ostasiewicz W. (ed.), (1998), *Statystyczne metody analizy danych*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langegeo we Wrocławiu, Wrocław.
Rousseeuw P. J., Hubert M. (1999) Regression Depth, *JASA*, 94, 388–402.
Statistical Yearbook of Voivodships 2005.
Van Aelst S., Rousseeuw P. J., Hubert M., Struyf A. (2000), The Deepest Regression Method, web site www.agoras.ua.ac.be.
Zeliaś A. (1996), Metody wykrywania obserwacji nietypowych w badaniach ekonomicznych, *Wiadomości Statystyczne* 8, 16–27.
Zieliński R. (1979), *Generatory liczb losowych*, WNT, Warszawa.

Dorota Pruska

ZRÓŻNICOWANIE OCEN PARAMETRÓW REGRESJI LINIOWEJ UZYSKANYCH METODĄ NAJGŁĘBSZEJ REGRESJI

Metoda najgłębszej regresji polega na oszacowaniu parametrów liniowej funkcji regresji w taki sposób, aby uzyskanemu modelowi odpowiadała największa głębia regresyjna.

W pracy przedstawiono charakterystykę metody najgłębszej regresji i przeprowadzono symulacyjną analizę (metodami Monte Carlo) zróżnicowania ocen parametrów modelu regresji liniowej uzyskanych tą metodą dla zbiorów danych zawierających różną liczbę obserwacji nietypowych. Na podstawie przeprowadzonych eksperymentów Monte Carlo wyznaczono charakterystyki rozkładu ocen parametrów i dokonano porównania otrzymanych wyników z wynikami analogicznych eksperymentów, w których do estymacji parametrów wykorzystano metodę najmniejszych kwadratów.