OLSZTYN ECONOMIC JOURNAL
Abbrev.: Olszt. Econ. J., 2011, 6(2)

UNTYPICAL OBSERVATIONS IN LINEAR REGRESSION

Małgorzata Kobylińska

Chair of Quantitative Methods
University of Warmia and Mazury in Olsztyn

K e y w o r d s: linear regression, untypical observations, measure of observation depth in the sample.

A b s t r a c t

In the analysed set of socioeconomic phenomena and processes results differing from the others may occur. Revealing such untypical observations is an important research issue as they may distort the statistical analysis of the investigated phenomenon. The paper discusses the types of untypical observations in two-dimensional sample. The method for detecting untypical observations in linear regression based on the measures of observation depth in the sample was proposed that was illustrated on the base of a numeric example.

OBSERWACJE NIETYPOWE W REGRESJI LINIOWEJ

Małgorzata Kobylińska

Katedra Metod Ilościowych
Uniwersytet Warmińsko-Mazurski w Olsztynie

S ł o w a k l u c z o w e: regresja liniowa, obserwacje nietypowe, miara zanurzania obserwacji w próbie.

A b s t r a k t

W analizowanych zbiorze danych zjawisk i procesów społeczno-ekonomicznych mogą wystąpić wyniki odbiegające od pozostałych. Ujawnienie takich obserwacji nietypowych jest istotnym zagadnieniem badawczym, gdyż mogą one zniekształcać analizę statystyczną badanego zjawiska. W pracy omówiono rodzaje nietypowości obserwacji w próbie dwuwymiarowej. Zaproponowano metodę wykrywania obserwacji nietypowych w regresji liniowej opartą na miarach zanurzania obserwacji w próbie, którą zilustrowano przykładem liczbowym.

Introduction

The sources of statistical data are diversified and depend on the character of the object that is the subject of research. Analysing the general population from the perspective of two or more statistical characteristics the search for and obtaining knowledge on the relations linking individual phenomena is important. Knowledge of correlations occurring between phenomena and processes is helpful in, among others, projecting their development. Analysis of regression is one of the more important and frequently applied statistical methods finding application in, among others, enterprise management and economy. Always, however, the risk exists that observations not matching the other observations will appear in the sets of data. The consequences of using data containing untypical observation for designing the regression function may be expressed by poorer matching of the function to the empirical data.

Currently the methodology of statistical research applicable to analysis of untypical data has developed widely. It has become one of the more important problems in multidimensional statistical analysis. The problem of untypical observations is presented relatively widely in the statistical literature. In the work by BARNETT (1978) the causes for appearance of untypical observations and methods of dealing with them have been described. The untypical observations in the deterministic sense that resulted from specific explainable causes and untypical observations in statistical sense that are inconsistent with the assumed probability distribution were identified. For the purpose of untypical observations identification, e.g. the Dixon's tests based on the quotient of the difference of the observation suspected to be untypical with the preceding or following observation (see, e.g. TRYBUŚ 1983) of Ferguson's skewness and kurtosis tests (FERGUSON 1961) can be applied.

In the subject literature considerations on the here discussed subject can be found, among others, in the works by: BARNETT and LEWIS (1978), CARONI (1993), CZEKAŁA (2001), HUBER (1981), ROUSSEEUW and LEROY (1987), ZELIAŚ (1996). The issue of untypical observations in case of a small sample is discussed in the work by KOWALEWSKI (1994).

In 1975, Tukey introduced the notion of the depth of a point in multidimensional sets. Thanks to allocating to each observation of a value of the depth measure corresponding to it, it is possible to rank statistical units according to their distance from the central cluster. The depth may be used for visualisation of numeric data, both one- and multidimensional and for determination of untypical observations. The notion of depth of the data was developed extensively by many researchers also from the perspective of its suitability for statistical description of one- and multidimensional data. Considerations concerning that issue can be found, among others, in the works by: HE and

WANG (1997), KOSHEVOY (2002), YEh and SINGH (1997), ROUSSEEUW and RUTS (1997), STRUYF and ROUSSEEUW (1998) as well as ZUO and STERFLING (2000).

The paper discusses the types of untypical observations in the two-dimensional sample. The method for detecting untypical observations in linear regression will be proposed in which the standardised residues (see, e.g. PAWEŁEK and ZELIAŚ 1996) as well as measures of depth of the observation in the sample will be used. At the beginning it will be discussed in detail and next illustrated on the numeric example.

Untypical observations in two-dimensional sample

The observation that does not match the configuration of the entire set of elements is called the untypical observation (ZELIAŚ 1996). Such observations may be a consequence of an error in measurement or in recording, application of inappropriate random sample selection method; they may originate from a different population or result from lack of homogeneity of the statistical sample. If untypical observations appear in the analysed set of data, they can be rejected, their values can be adjusted or they may be accepted and appropriate methods of statistical data analysis can be applied. Untypical observations change and distort the character of correlation between the investigated variables. This is of major importance in case of, e.g. forecasting on the base of the estimated models.

The correlogram of two-dimensional sample may present different configurations of points on the surface. That is why it is reasonable to identify the types of homogeneity in two-dimensional space depending on what compact figure we analyse. As a consequence we identify (JAJUGA 1993):

– ellipsoidal homogenous sets when the set analysed as a set of points in two-dimensional space forms a compact figure similar in shape to the ellipse,

– sets homogenous in the sense of linear regression, if the set analysed as a set of points in two-dimensional space creates within it a compact figure with the shape that allows its approximation by means of linear regression,

– homogenous sets that are none of the above types.

As concerns the ellipsoid homogeneity certain proposals are presented in the work by JAJUGA (1987).

The notion of homogeneity is very important in statistics but unfortunately underappreciated and used in imprecise way by many researchers. It should be highlighted that homogeneity of the set of observations is the condition necessary for usefulness of many statistical methods, including the methods of statistical multidimensional analysis. Homogeneity of the set of observations is linked to closely the notion of the distance of points in two-dimensional sample.

If the set is homogenous the distances of observations from a certain characteristic are small. Individual values in the sample with high values of that distance may be treated as untypical.

The untypical character may appear in:

– marginal distributions,

– joint distribution.

The following types of untypicalness of two-dimensional sample can be identified (WAGNER et al. 1997):

– Linear type with points detached in the direction of the *OY* axis (Fig. 1a)

They are characterised by the following properties:

(w1) min P_y^2 – max P_y^1 >> 0,

(w2) śr P_y^2 – śr P_y^1 >> 0,

(w3) med P_y^2 – med P_y^1 >> 0,

where P_y^2 and P_y^1 are projections of the two-dimensional sample with the population of n on the *OY* axis, to which populations $n_y^2 = \#\, P_y^2$ and $n_y^1 = \#\, P_y^1$ correspond so that $n_y^2 + n_y^1 = n$ and $P_y = P_y^1 \cup P_y^2$, $P_y^1 \cap P_y^2 = \emptyset$, śr P_y^2 and śr P_y^1 represent arithmetic averages from the elements of samples P_y^2 and P_y^1, min P_y^2 and max P_y^1 – minimum and maximum from elements of samples P_y^2 and P_y^1, med P_y^2 and med P_n^1 – medians from elements of samples P_y^2 and P_y^1 while the symbol >> means "much larger than".

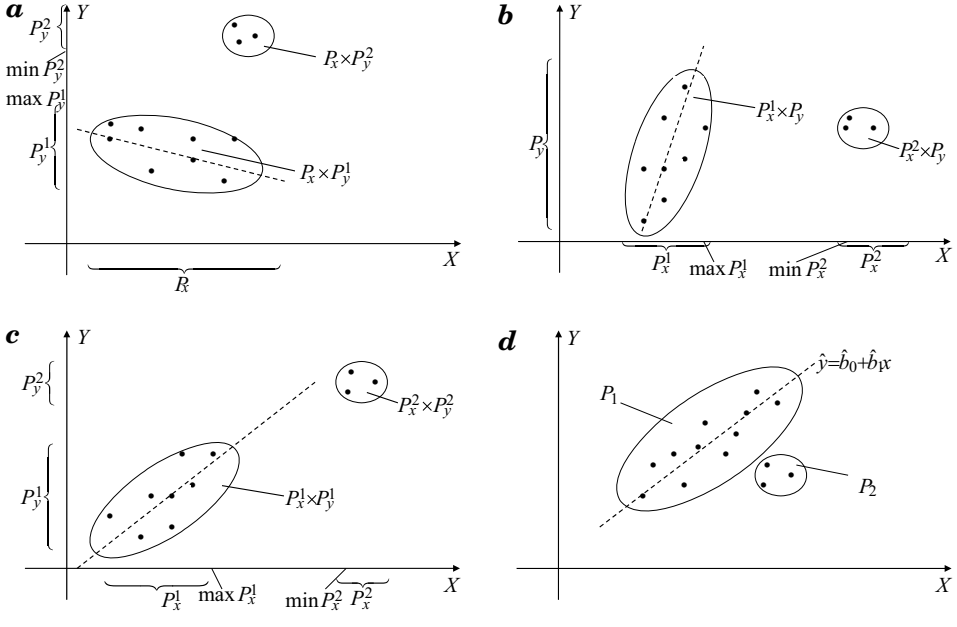

Fig. 1. Linear type with detached points: in the *OY* direction (*a*), in the *OX* direction (*b*), in the *OX* and *OY* direction (*c*), in geometrical sense (*d*)

Source: based on WAGNER et al. 1997

– Linear type with points detached in the direction of the *OX* axis (Fig. 1b) with the properties:

(w1) min P_x^2 – max P_x^1 >> 0,

(w2) śr P_x^2 – śr P_x^1 >> 0,

(w3) med P_x^2 – med P_x^1 >> 0,

where P_x^1 and P_x^2 are projections of two-dimensional sample with the population of n on the *OX* axis, to which populations $n_x^1 = \#\, P_x^1$ and $n_x^2 = \#\, P_x^2$ correspond, so that $n_y^1 + n_x^2 = n$ and $P_x = P_x^1 \cup P_x^2$ and $P_x^1 \cap P_x^2 = \emptyset$, śr P_x^2 and śr P_x^1, present arithmetic averages from the elements of samples P_x^2 and P_x^1, min P_x^2 and max P_x^1 – minimum and maximum from elements of samples P_x^2 and P_x^1, med P_x^2 and med P_x^1 – medians from elements of samples P_x^2 and P_x^1, while the symbol >> means "much larger than".

– Linear type with points detached in the direction of the *OX* and *OY* axes characterised by the following properties (Fig. 1c):

(w1) min P_x^2 – max P_x^1 >> 0, min P_y^2 – max P_y^1 >> 0,

(w2) śr P_x^2 – śr P_x^1 >> 0, śr P_y^2 – śr P_y^1 >> 0,

(w3) med P_x^2 – med P_x^1 >> 0, med P_y^2 – med P_y^1 >> 0,

where P_x^1, P_x^2, P_y^1, P_y^2 are projections of two-dimensional sample on axes *OX* and *OY* with populations of:

$$n_x^k = \#\, P_x^k,\ n_y^k = \#\, P_y^k, \text{ for } k = 1,2, \text{ where } P_x^1 \cup P_x^2 = P_x,\ P_x^1 \cap P_x^2 = \emptyset,$$

$$P_y^1 \cup P_y^2 = P_y,\ P_y^1 \cap P_y^2 = \emptyset \text{ and } n = n_x^1 + n_x^2 = n_y^1 + n_y^2 \text{ and also } n_x^1 = n_y^1,\ n_x^2 = n_y^2,$$

– Linear type with points detached in the geometric sense (fig. 1d)

Existence of separate concentrations P_1 and P_2 such that $P = P_1 \cup P_2$, but without separation of concentrations for samples P_x and P_y is assumed. The following conditions are also satisfied:

(w1) $\bigwedge\limits_{x_i \in P_2} \{x_i \in \langle \min P_x^1,\ \max P_x^1 \rangle\}$,

(w2) $\bigwedge\limits_{y_i \in P_2} \{y_i \in \langle \min P_y^1,\ \max P_y^1 \rangle\}$.

The measure of detachment of the set P_2 from the nucleus P_1 of two-dimensional sample is expressed by the geometrical distance of point $(x_0, y_0) \in P_2$ from the regression line $\hat{y} = b_0 + b_1 x$ estimated on the base of the data contained in sample P_1 with the form where $d = \dfrac{|\hat{b}_1 x_0 + \hat{b}_0 - y_0|}{\sqrt{\hat{b}_1^2 + 1}} \geq d_0$, where d_0 represents the distance set arbitrarily.

Also the residues from the estimated linear regression function are used for detection of untypical observations (PAWEŁEK and ZELIAŚ 1996). In the theory of linear regression, next to the typical observations also observations that are:

– untypical (Fig. 2a),
– influential (Fig. 2b),
– distant from the other observations (Fig. 2c)

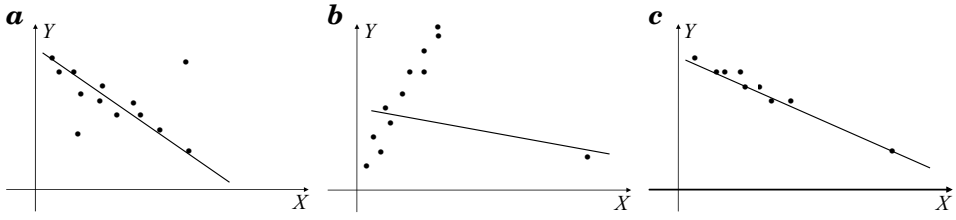

Fig. 2. Observations typical and untypical in relation to the linear regression (*a*), typical observations and influential observation (*b*), typical observations and observation distant from the other observations (*c*)
Source: based on PAWEŁEK and ZELIAŚ 1996.

Standardised residues can be used for determination of untypical observations in linear regression

$$\tilde{e}_i = \frac{e_i}{S_e},\ 1,2,\ldots,\ n \tag{1}$$

where
$\tilde{e}_i$ – standardised residue for observation i,
e_i – residue i of regression,
n – number of observations,
S_e – standard deviation of the regression residue determined according to the formula

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k}} \tag{2}$$

where k is the number of estimated parameters.

Method for determination of untypical and influential observations in linear regression

Let $P_n^2 = \{x_a, x_2,\ldots, x_n\}$ be a system of observable vectors expressing the two-dimensional sample with population n originating from a certain two-dimensional distribution defined by the distribution function F_2 and let $\theta \in R^2$ represent a certain point from the real space R^2. In particular, it may belong to the system of points from sample P_n^2. It is assumed that at least

$h = [n/2] + 1$ observations from sample P_n^2 are not positioned on any straight line. If no more than two observations belong to any straight line then sample P_n^2 is called the generally positive set of points according to the nomenclature introduced by DONOHO and GASKO (1992). The criterion for determination of the Mahalanobis depth measure in case of two-dimensional case assumes the following form:

The function

$$\text{Mzan}_2\ (\theta;\ P_n^2) = [1 + Q(\theta,\ P_n^2)]^{-1} \tag{3}$$

$$\text{where } Q(\theta,P_n^2) = (\theta_1 - \bar{x}_1)^2\ s^{11} + 2(\theta_1 - \bar{x}_1)(\theta_2 - \bar{x}_2)\ s^{12} + (\theta_2 - \bar{x}_2)^2\ s^{22},$$

while

$$\theta = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix},\ \bar{x} = \begin{bmatrix}\bar{x}_1\\ \bar{x}_2\end{bmatrix},\ \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j,\ S = \frac{1}{n-1}\sum_{i=1}^{n}(x_j - \bar{x})(x_j - \bar{x})^T,\ S^{-1} = \begin{bmatrix}s^{11} & s^{12}\\ s^{21} & s^{22}\end{bmatrix},$$

we call the Mahalanobis depth measure Mzan_2 for point θ in sample P_n^2.

In the mathematical sense the depth measure Mzan_2 obtained according to the formula (3) is based on the distance between the point of the space R^2 and the vector of averages $\bar{x}$ according to metrics determined by the inverse S matrix. It should be pointed out that for determination of the Mahalanobis distance in case when matrix S is not positively defined the so-called generalised Mahalanobis distance can be determined (see, e.g. BARTKOWIAK 1988). Other criteria for determination of the measures of depth of the observation in a sample are presented, among others, in the works by WAGNER and KOBYLIŃSKA (2000, 2002).

The algorithm for determination of untypical observations in linear regression using the measure of depth of the observation in the sample will be proposed. It involves the following steps:

Step 1. We estimate the linear regression equation for the values of analysed variables observed in the sample. The linear regression model of y in relation to x is represented by the equation

$$\hat{y} = a_1 x + a_0$$

where:

$\hat{y}$ – theoretical values of the regression function $\hat{y} = f(x)$ corresponding to the given level of performance of variable X,

a_0, a_1 – estimates of the parameters of regression function Y to X, where a_1 is the estimate of the linear regression coefficient of variable Y in relation to X, a_0 – the estimate of the free expression.

Step 2. Determination of the values of standardised residues $\tilde{e}_i$ according to formula 1,

Step 3. Determination of the Mahalanobis depth measure $\text{Mzan}_2\ (x_1, P_n^2)$ of observations in two-dimensional sample according to formula 3,

Step 4. Each observation x_i of the two-dimensional sample is represented by the vector $[\text{Mzan}_2\ (x_i, P_n^2); \tilde{e}_i]$. For the purpose of determining untypical and influential observations in linear regression we conduct a review of observations P_n^2 relative to the determined values of the depth measures and the standardised deviations. Observations represented by the lowest values of the depth measure and relatively high or low values of the standardised residues can be considered untypical in relation to the estimated linear regression. Observations represented by the lowest depth measures and the values of standardised residues close to zero can be considered distant from the others.

Numeric example. The two-dimensional sample P_{34}^2 is the set of 34 pairs. On the correlogram (Fig. 3) significant concentration of observations in the area of low and mean values of both variables and presence of untypical observations can be noticed. The Pearson's linear correlation coefficient value is $r = 0{,}303$. This does not indicate strong correlation of the analysed variables.

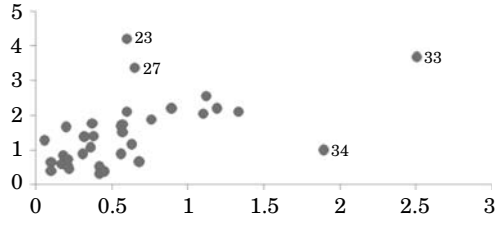

Fig. 3. Correlogram of the two-dimensional sample

Source: own work.

Table 1, next to the values of observations X and Y, presents values of depth measures and the standardised residues computed for them. Vectors $[\text{Mzan}_2\ (x_i, P_n^2); \tilde{e}_i]$ were organised according to the values of Mahalanobis depth measures for observations in two-dimensional sample. Considering the values of standardised residues and depth measures it can be noticed that observations 23, 34 and 27 are represented by the lowest depth measure values and relatively high values of standardised residues. They can be considered untypical in relation to the linear regression. The low value of the depth measure and relatively low value of the standardised residues correspond to observation 33. It can be considered influential.

Table 1

Values of observations in two-dimensional sample, values of depth measures and standardised residues

Value of observation of the two-dimensional sample			Value of observation of the two-dimensional sample			Observations of the two-dimensional sample and values of standardised residues organised according to the values of the depth measure			Observations of the two-dimensional sample and values of standardised residues organised according to the values of the depth measure		
no of obser-vations	X	Y	no of obser-vations	X	Y	no of obser-vations	$Mzan_2$	standard residues	no of obser-vations	$Mzan_2$	standard residues
1	0.06	1.28	18	0.45	0.39	**33**	**0.068**	**0.375**	4	0.495	-0.524
2	0.1	0.65	19	0.56	0.9	**23**	**0.077**	**3.367**	15	0.519	0.101
3	0.1	0.4	20	0.56	1.69	**34**	**0.082**	**-2.161**	8	0.564	-0.402
4	0.17	0.6	21	0.57	1.74	**27**	**0.151**	**2.296**	5	0.573	-0.217
5	0.18	0.86	22	0.57	1.5	32	0.334	-0.124	13	0.593	0.665
6	0.2	1.66	**23**	**0.6**	**4.2**	17	0.383	-1.18	24	0.604	0.785
7	0.21	0.55	24	0.6	2.1	25	0.397	0.703	30	0.618	0.547
8	0.21	0.74	25	1.12	2.56	3	0.404	-0.683	10	0.688	-0.33
9	0.22	0.45	26	0.63	1.16	18	0.407	-1.131	19	0.692	-0.641
10	0.31	0.9	**27**	**0.65**	**3.38**	1	0.425	0.449	11	0.724	0.248
11	0.32	1.38	28	0.68	0.67	31	0.432	0.174	12	0.793	-0.158
12	0.36	1.09	29	0.76	1.87	28	0.446	-1.073	14	0.81	0.186
13	0.37	1.77	30	0.89	2.2	6	0.451	0.742	29	0.844	0.303
14	0.38	1.39	31	1.19	2.2	9	0.454	-0.771	26	0.848	-0.408
15	1.1	2.05	32	1.33	2.1	2	0.47	-0.376	21	0.863	0.38
16	0.42	0.53	**33**	**2.5**	**3.69**	16	0.491	-0.922	20	0.889	0.331
17	0.42	0.32	**34**	**1.89**	**1.01**	7	0.492	-0.636	22	0.987	0.084

Source: own work based on the conventional data.

Tables 2 and 3 present the results of the linear regression equation estimation. The linear regression equation was estimated for all 34 observations of P_n^2 (Tab. 2) and after elimination of untypical observations (Tab. 3). The determination coefficient is 0,303 and 0,662 respectively. It is significantly higher for the estimation after elimination of observations 23, 34 and 27.

Table 2

Estimations of regression equation parameters for 34 observations in two-dimensional sample

a_0	t_{a_0}	a_1	t_{a_1}	R^2
0.854 (0.216)	3.954	1.012 (0.272)	3.721	0.303

Source: own work.

Table 3
Estimations of regression equation parameters for 31 observations in two-dimensional sample

a_0	t_{a_0}	a_1	t_{a_1}	R^2
0.601 (0.128)	4.695	1.300 (0.172)	7.558	0.662

Source: own work.

Conclusion

The paper discusses the types of untypical observations in two-dimensional sample and proposes a method for elimination of untypical observations using the measure of depth of the observation in the sample. The considerations presented lead to the conclusion that the problem of appearance of untypical observations is a major limitation encountered during estimation of statistical population parameters. Detecting them is the first stage followed by elimination of them and application of the appropriate data analysis method.

In the STATISTICA package many tools (statistics and graphs) exist that facilitate detection of diverging observations. For that purpose the model residues are used. Next to the observed values, the values of the residue and their standardised values we find numerous statistics intended for residue analysis. The Mahalanobis distance and Cook distance are popular and frequently applied.

The presented paper presents the usefulness of measures of depth of the observations in the sample for detecting and elimination of untypical observations in linear regression. Elimination of those observations improves matching of linear regression to the empirical data. Using the values of the measures of depth in the sample and considering at the same time the values of standardised residues (according to formula 1) the observations that are distant from the other ones can be determined.

Translated by Jerzy Gozdek

Accepted for print 24.08.2011

References

Bartkowiak A. 1988. *An Algorithm for Repeated Calculations of the Generalized Mahalanobis Distance.* AMSE Review, 8(3): 9–18.

Barnett V., Lewis T. 1978. *Outliers in statistical data.* Wiley and Sons, New York.

Caroni C., Prescott P. 1993. *Union-Intersection Testing for Outliers in Multivariate Normal Data,* J. Statist. Comput. Simulation, to appear.

Czekała M. 2001. *Statystyki pozycyjne w modelowaniu ekonometrycznym.* AE, Wrocław.

Donoho D.L., Gasko M. 1992. Breakdown *Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness.* The Annals of Statistics, 20: 1803–1827.

HE X., WANG G. 1997. *Convergence of Depth Contours for Multivariate Datasets.* The Annals of Statistics, 25: 495–504.

HUBER P.J. 1981. *Robust Statistics*. Wiley & Sons, New York.

JAJUGA K. 1987. *Statystyka ekonomicznych zjawisk złożonych – wykrywanie i analiza niejednorodnych rozkładów wielowymiarowych*. Prace naukowe AE, 371, Wrocław.

JAJUGA K. 1993. *Statystyczna analiza wielowymiarowa.* PWN, Warszawa.

KOWALEWSKI G. 1994. *Obserwacje nietypowe w regresji liniowej* (maszynopis rozprawy doktorskiej). AE, Wrocław.

KOSHEVOY G.A. 2002. *The Tukey Depth Characterizes the Atomic Measure.* Journal of Multivariate Analysis, 83: 360–364.

LIU R.Y., PARELIUS J.M., SINGH K. 1999. *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference.* The Annals of Statistics, 27: 783–858.

PAWEŁEK B., ZELIAŚ A. 1996. *Obserwacje nietypowe w badaniach ekonometrycznych.* Badania operacyjne i decyzje, 2: 59–86,

ROUSSEEUW P.J., LEROY A. 1987. *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW P.J., RUTS I. 1997. *The Bagplot: a Bivariate Box- and-Whiskers Plot.* Technical Report, University of Antwerp.

STRUYF A., ROUSSEEUW P.J. 1998. *Halfspace Depth and Regression Depth Characterize the Empirical Distribution.* Journal of Multivariate Analysis, 69: 135–153.

TUKEY J.W. 1975. *Mathematics and the Picturing of Data*. Proceedings of the International Congress of Mathematicians, pp. 523–531.

WAGNER W., LIRA J., BŁAŻCZAK P. 1997. *Typy obszarów rozrzutu i obserwacji odstających przy szacowaniu regresji liniowej*. XXVII Colloquium Biometryczne, pp. 330–341.

WAGNER W., KOBYLIŃSKA M. 2000. *Miary i kontury zanurzania w opisie próby dwuwymiarowej*. Wyzwania i Dylematy Statystyki XXI wieku, AE Wrocław, pp. 201–216.

WAGNER W., KOBYLIŃSKA M. 2002. *Przegląd metod wyznaczania miar i konturów zanurzania w próbie dwuwymiarowej*. Przegląd Statystyczny, 49(4): 119–131.

YEH B.A., SINGH K. 1997. *Balanced Confidence Regions Baser on Tukey's Depth and the Bootstrap*. Journal Royal Statistical Society, 59: 639–652.

ZELIAŚ A. 1996. *Metody wykrywania obserwacji nietypowych badaniach ekonomicznych*. Wiadomości Statystyczne, 8: 16–27.

ZUO Y., SERFLING R. 2000. *General Notations of Statistical Depth Function.* Annals Statistics, 28: 461–482.