Biometrical Letters
Vol. 48 (2011), No. 2, 83-99

Investigation of vectors of influential observations in a linear regression model

Anna Budka

Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland; e-mail: budka@up.poznan.pl

SUMMARY

The procedures presented in this work lead to the identification of reasons for the poor fit of a mathematical model. Investigation of vectors of influential observations in a linear regression model leads to the identification of systems of observations regarded as influential. Particularly interesting are those which contain cases which were previously not regarded as influential in the analysis of a full model or 1-cut model.

Key words: *m*-cut linear regression model, vector of influential observations, masking of observations, leverage point

1. Introduction

In the analysis of a linear regression model by the least-squares method, an important issue is the qualitative evaluation of the model. This refers to estimators of the structural parameters and random component, as well as to various test statistics. These may depend directly on one or more vector observations of the dependent variable and the system of independent variables.

In this work we present concepts relating to an *m*-cut linear regression model, as well as a study of a distinguished system of *m* observations influencing the evaluation of the regression model. The concept of a vector of influential observations is defined, and influential observations in an *m*-cut model are discussed, along with the phenomenon of masking. Particular attention is paid to Cook's distance measure as a criterion for diagnostic analysis of the influence of vector observations on the quality of statistics used

to evaluate the regression model. The procedure for discovering influential observations is illustrated with a numerical example.

2. Material and methods

2.1. Concepts and denotations in an *m*-cut linear regression model

Let there be given a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{y}$: $n \times 1$ is a vector of observations of the dependent variable, $\mathbf{X}$: $n \times p$ a matrix of observations of p - 1 independent variables whose first column corresponds to the vector of ones, $\boldsymbol{\beta}$: $p \times 1$ a vector of structural parameters, and e : $n \times 1$ a vector of random errors ($E(\mathbf{e}) = \mathbf{0}$, $D(\mathbf{e}) = \sigma^2\mathbf{I}$).

Let $\hat{\boldsymbol{\beta}} = \mathbf{G}\mathbf{X}'\mathbf{y}$ be an estimator of the vector of structural parameters, obtained by the least-squares method, where $\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}'$: $n \times n$ is an orthogonal projection matrix, and $\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}(n - p)$ is an estimator of the parameter σ^2 of the variance of the components of the vector of random errors.

In order to investigate whether a system of *m* observations is influential, we introduce the symbol *I* to denote a subset of *m* numbers from the set {1,2,…, *n*}, $\mathbf{X}_{(I)}$:(*n* - *m*) × *p* is an *m*-cut matrix, lacking the subset *I* of observations from matrix **X,** $\mathbf{X}_I$: $m \times p$ is the matrix of the distinguished system of *m* observations from matrix **X,** $\mathbf{X} = \begin{bmatrix} \mathbf{X}'_{(I)} & \mathbf{X}'_I \end{bmatrix}'$ is a block division of matrix $\mathbf{X}$ into the indicated submatrices, $\mathbf{y} = \begin{bmatrix} \mathbf{y}'_{(I)} & \mathbf{y}'_I \end{bmatrix}'$ is a division of vector $\mathbf{y}$ into subvectors $\mathbf{y}_{(I)}$:(*n* - *m*) × *1* and $\mathbf{y}_I$ *:* *m* × 1 according to the division of matrix $\mathbf{X}$, { $\mathbf{y}_{(I)}, \mathbf{X}_{(I)}\boldsymbol{\beta}_{(I)}, \sigma^2\mathbf{I}$ } is an m-cut linear regression model, where $\boldsymbol{\beta}_{(I)}$ does not denote a cut of the vector of structural parameters, but emphasizes that the *p*-dimensional vector of parameters $\boldsymbol{\beta}$ is estimated from an *m*-cut model, $\hat{\boldsymbol{\beta}}_{(I)} = \hat{\boldsymbol{\beta}} - \mathbf{G}\mathbf{X}_I(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{r}_I$, : $p \times 1$ is a least-squares estimator of vector $\boldsymbol{\beta}$ from the *m*-cut model, where $\mathbf{H}_I = \mathbf{X}_I\mathbf{G}\mathbf{X}_I : m \times m$ and $\mathbf{r}_I = \mathbf{y}_I - \mathbf{X}_I\hat{\boldsymbol{\beta}}$ (Budka and Wagner, 2007).

2.2. Determination of vectors of influential observations

The least-squares estimators of the vector of structural parameters $\hat{\boldsymbol{\beta}}$ and the variance of random errors $\hat{\sigma}^2$, or functions of them, may reveal unfavourable properties of the adopted model, whereby it is poorly fitted to the observations. This will be expressed, for example, in high values of the standard deviations which are used to build various test statistics to verify hypotheses in parametric statistical inference, and also to construct confidence intervals (see e.g. Draper and Smith 1973, Oktaba, 1986, Ostasiewicz, 1999). By removing any influential or divergent observations, we improve the fit of the model to the observations. However rejection of observations cannot be done mechanically – it requires critical consideration of the structure of the numerical data in terms of the occurrence of any *atypical observations*, which may appear in the direction of the *X* axis (among the components of the row vectors of matrix **X**) or in the direction of the *Y* axis (among the components of the observable vector **y**).

In both situations the problem relates to a single or multiple vectors of observations. In linear regression analysis this issue is known as *occurrence of influential observations*, since these have a direct influence on the analysed numerical characteristics of the regression model.

To define formally the vector of influential observations, we denote a sample of n $(p+1)$ - dimensional observations in the form of a sequence of row vectors of matrix **X** and components of vector **y:** $\mathbf{Z} = (\mathbf{X}, \mathbf{y}) = ((\mathbf{x}_1', y_1), (\mathbf{x}_2', y_2), ..., (\mathbf{x}_n', y_n))' = ((\mathbf{x}_i', y_i);\ i=1,\ 2,\ ...,\ n)' = P_n^{p+1}$, where $(\mathbf{x}_i', y_i)' \in R^{p+1}$

Definition : In a linear regression model, the vector of observations $((\mathbf{x}_{i_1}', y_{i_1}), ..., (\mathbf{x}_{i_m}', y_{i_m}))'$ in a sample P_n^{p+1}, indexed by a distinguished set of m indicators $\{i_1, i_2, ..., i_m\} \subset \{1,2,...,n\}$, is called a vector of influential observations if its components significantly change the values of the analysed numerical characteristics in that model.

The definition given implies the existence of at most $n![m!(n-m)!]^{-1}$ influential vectors, and this is also the number of m-cut regression models. For each of them one can determine a least-squares estimator for the vector of

parameters and variance of errors. This procedure is inefficient in practice when n and m are large. Moreover the estimators obtained are not always acceptable to the researcher, which fact may indicate the occurrence of influential observations. For such a situation a principle of sequential procedure is proposed, which begins with analysis of 1-cut linear regression models.

2.3. Influential observations in an *m*-cut model

We consider the case $m = 1$. Characteristics of such a model are discussed in, for example, (Besley et al., 1980), (Cook and Weisberg, 1982), (Chatterjee and Hadi, 1988), (Ostasiewicz, 1999), and (Budka and Wagner**,** 2007).

In a 1-cut regression model diagnostic analysis is carried out on the effect of vector observations on the quality of the statistics of evaluation of the regression model. The criterion adopted for this analysis is the Cook distance, as a measure of influence, which can be expressed in the form:

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{ps^2} = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})' (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{ps^2} = \frac{t_i^2 \cdot w_i}{p}, \qquad (1)$$

$i = 1, 2, \ldots, n$, where s^2 is an estimator of the parameter σ^2,

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{r_i}{1 - h_{ii}} \mathbf{G}\mathbf{x}_i$$

an estimator of the vector of structural parameters in a 1-cut model, r_i is the i-th component of the vector of residuals $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, h_{ii} – the i-th diagonal element of matrix $\mathbf{H}$, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ the estimated vector $\mathbf{y}$ from the full model, $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$ the estimated vector $\mathbf{y}$ from the 1-cut model, $t_i^2 = r_i^2 s^{-2} (1 - h_{ii})^{-1}$ the square of the i-th internal studentized residual, and the sensitivity vector $w_i = h_{ii}(1 - h_{ii})^{-1}$ the ratio of the variance of the i-th estimated variable to the residual.

This measure makes it possible to distinguish a system of single vectors of influential observations. The threshold value for determination of influential observations based on the Cook distance is taken to be $4/(n\text{-}p\text{-}1)$ (Chatterjee and Hadi, 1988), (Belsley et el., 1980), (Fox, 2005).

We find that this system contains m_1 observations. This serves as a starting point for determining 2-dimensional, and from them 3-dimensional and so on up to *m*-dimensional, vectors of influential observations.

We consider a case with $m > 1$. This is not an ordinary generalization of the 1-cut model. This is because of the phenomenon of *masking* of observations, when the dataset may contain small subsets which are jointly influential, although each observation considered separately is not influential (observations C and D). The reverse may also be the case: observations A and B are individually influential, but are not so when considered jointly (Figure 1).

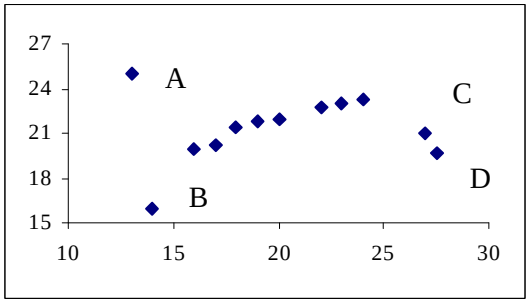

Figure 1. The effect of masking of observations

A symptom of the masking effect may be that after the separation of one or more influential observations, other observations may appear as exceptionally influential, although this was not visible previously. Hence the significance of particular observations may not be revealed when other observations have been previously separated off.

The problem of masking observations has been considered by many authors, including (Cook and Weisberg, 1980), (Gray and Ling, 1984), (Atkinson, 1985), (Rousseeuw and Leroy, 1987), (Chatterjee and Hadi, 1988), (Lawrance, 1995), and (Pena and Yohai, 1995).

In investigating a system of *m* (*p*+1)-dimensional row vectors of influential observations, namely a certain *m*×(*p*+1)-dimensional submatrix separated off from the matrix (**X, y**), attention must be given to the problem – which frequently does not have a definitive answer – of how to establish the value

m=2, 3, ... Should one consider all systems of combinations of *m* observations from an *n*-element set?

It should be noted that the main purpose of investigating distinguished submatrices is to find systems of observations that are considered influential. Important are those which contain cases previously not identified as influential in analysis of the 1-cut model. The reverse situation may occur, where a given system of submatrices was identified as influential, but on addition of new observations to those submatrices it transpires that the new system of enlarged submatrices is not longer considered influential.

The measure of the influence of the Cook distance (1) for an *m*-cut model takes the form:

$$D_I = \frac{(\hat{\boldsymbol{\beta}}_{(I)} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{(I)} - \hat{\boldsymbol{\beta}})}{ps^2} = \frac{1}{ps^2} \mathbf{r}_I' (\mathbf{I} - \mathbf{H}_I)^{-1} \mathbf{H}_I (\mathbf{I} - \mathbf{H}_I)^{-1} \mathbf{r}_I \,, \quad (2)$$

The symbol I denotes m numbers of distinguished vector observations, namely the investigated submatrix of influential observations.

Measure (2) for m=2 and a distinguished pair $I = \{i, j\}$ such that $1 \le i < j \le n$ is expressed by the formula:

$$D_I = D_{\{i,j\}} = B_1 + B_2 + B_3 \,, \quad (3)$$

Where

$$B_1 = (D_i + D_j)\left(1 + \frac{h_{ij}^2}{d_{ij}}\right)^2 ,$$

$$B_2 = \frac{h_{ij}^2}{2s^2 d_{ij}^2}\left\{r_i^2 (2 - h_{jj}) + r_j^2 (2 - h_{ii})\right\}, \; B_3 = \frac{2 r_i r_j h_{ij}}{2s^2 d_{ij}^2}\left\{1 + h_{ij}^2 - h_{ii} h_{jj}\right\},$$

h_{ij} is the (i, j)-th element of matrix **H**, and $d_{ij} = (1 - h_{ii})(1 - h_{jj}) - h_{ij}^2$ (Gray and Ling**,** 1984).

The formula given makes possible the following interpretation of the distinguished pair (i, j) of vector observations:

if $h_{ij} > 0$,	if $h_{ij} < 0$,
$B_3 > 0$, when $r_i r_j > 0$	$B_3 < 0$, when $r_i r_j > 0$
(jointly influential);	(not influential);
$B_3 < 0$, when $r_i r_j < 0$	$B_3 > 0$, when $r_i r_j < 0$.
(not influential);	(jointly influential).

Next we can consider, for pairs of observations, the *amplification* effect when one or both statistics D_i, D_j are less than the statistic $D_{\{i,j\}}$, the *weakening* effect when one or both Cook statistics for the 1-cut models are greater than the Cook statistics for the 2-cut model, the *conserving* effect when both Cook statistics in the 1-cut models are almost equal to the Cook statistic for the 2-cut model, and the *attracting* effect when one of the Cook statistics for a 1-cut model with large values, and the other with smaller values, lead to large values of the Cook statistic for the 2-cut model.

In determining the vector of influential observations, attention must be paid to their character depending on the values of the diagonal elements of matrix **H**. These elements are expressed by vector values referring to the independent variables. This makes it possible to investigate the atypicality of vectors, expressed by their distance from a regression cluster (understood as a homogeneous set corresponding to a characteristic of the regression relation). In the case of a regression model with one independent variable, this corresponds to a configuration of points on a plane distributed around a certain straight line, whereas in the case of a multiple regression model such a configuration is a generalized ellipsoid with a hyperplane intersecting it.

Definition : Case observations, namely row vectors of matrix **X** corresponding to diagonal values of matrix **H** are called leverage points.

The values of diagonal elements are contained in the interval $\langle 1/n, 1)$. Hence taking the *a priori* set threshold value h_0 for diagonal elements of the matrix **H**, one can distinguish among them the *high-leverage points* which exceed that value. It is these points that will chiefly interest us further, in terms of their

influence on the estimated linear regression model. Observations are considered to be leverage points if $h_0 > (tr[H]/n)$. When $h_{ii} \geq 0.5$ we have a high-leverage point. When $0.2 < h_{ii} < 0.5$ we have medium influence (a medium leverage value) (Belsley et al., 1980).

The considered matrix **H** and the matrix $\mathbf{H}^*$ are matrices of orthogonal projection onto the space of columns of, respectively, the system matrix **X** and the extended system matrix **Z** = [**X**, **y**]. The diagonal elements of matrix $\mathbf{H}^*$ take account of the simultaneous influence of the values of the independent features and the dependent feature in the linear regression model. Hence the leverage points of both matrices give slightly differing information about particular observations. It can be shown that the leverage values h_{ii}^* for the extended matrix of the system are the sum of leverage values h_{ii} and the ratio of the square residuals r_i^2 to the sum of square residuals $SSE = \sum r_i^2$. The residuals r_i will be large for the divergent observations. However it cannot be indicated directly which values h_{ii} and r_i have a direct influence on the values h_{ii}^*. The following situations may occur here:

r_i^2 / h_{ii}	Large	Small
Large	(a) influential	(b) possible influential
Small	(c) possible influential	(d) is not influential

Situation (a) indicates the occurrence of atypical observations among the dependent feature and several independent features. Situations (b) and (c) may lead to the inequality $h_{ii}^* > h_0^*$, where h_0^* is a set threshold value (determined as h_0), and this will indicate an influential case. It should be noted that divergent observations need not be influential, influential observations need not be divergent, high-leverage points generally have small residuals, and influential observations correspond to non-proportional fit. Moreover divergent observations and high-leverage points need not be influential, and influential observations are not necessarily high-leverage points.

2.3. The principle of sequential separation of influential observations

We give here a principle of sequential procedure when separating off influential observations in linear regression models. Let $L=\{1, 2, ..., n\}$ be a set of n consecutive numbers corresponding to observations.

The sequential principle includes the following steps:

- identification of influential observations in a 1-cut model
  a. (a_j) – sequence of distinguished single influential observations from the set of all observations,
  b. $j = 1, 2, ..., q_1$,
  c. L_1 – set of q_1 indices for distinguished observations;
- identification of pairs of influential observations in a 2-cut model
  a. (a_j, a_i) – sequence of investigated pairs, where $j \in L_1$, $i \in L$, and $i \neq j$,
  b. in (i) for each a_j there are identified m_{a_j} influential pairs, influential $q_2 = \sum_{a_j} m_{a_j}$ pairs are established in total,
  c. (b_j, b_i) – selected influential pairs indexed by the set L_2 of cardinality q_2,
  d. $L_2 = \{(j_1, i_1),(j_2, i_2),....,(j_{q_2}, i_{q_2})$ – set of new indices of distinguished influential pairs,
- identification of triples of influential observations in a 3-cut model
  a. (b_j, b_i, a_k) – sequence of investigated triples, where $k \in L$, $k \neq i, j$,
  b. for each pair (b_j, b_i) there are distinguished m_{bjb_i} influential triples,
  c. $q_3 = \sum m_{b_j b_i}$ influential triples are determined in total,
  d. $(c_j, c_{i,}^{(b_j, b_i)} c_k) \in I_3$ – selected influential triples indexed by the set L_3 of cardinality $q_{3,}$
  e. $L_3 = \{(j_1, i_1, k_1),(j_2, i_2, k_2),....,(j_{q_3}, i_{q_3}, k_{q_3})\}$.
- the process can be continued for quadruples, quintuples etc. of influential observations.

In each situation when specifying pairs, triples, quadruples etc. it is necessary to omit those which have already been identified at an earlier stage.

3. Research problem

The discussed procedure for determining submatrices of influential observations will be illustrated using data taken from a meteorological experiment. The experiment was performed in 1975 in the United States under the name FACE (Florida Area Cumulus Experiment) (Woodley et al., 1977). It studied the relations between a dependent variable *Ln_Y* – the logarithm of the value of rainfall in the experimental area in a 6-hour interval after the first cloud seeding – for each of selected days in the 80-day duration of the experiment, and selected factors affecting the occurrence of rain:

- *A* – binary feature (1 – introduced, 0 – not introduced) for introduction of volatile compound of silver iodide into clouds,
- *T* – numbers of selected days in the period from 16/06–15/09 (1 – 16/06, 2 – 17/06, …, 92 – 15/09),
- *S-Ne* – seeding capability adjustment – difference between the height (km) of a rain cloud before and after seeding with silver iodide,
- *C* – range echo – percentage coverage of the 3000-square-mile experimental area by clouds,
- *Ln_P* – natural logarithm of value of total precipitation on the experimental area an hour before seeding [m^3 x 10^7],
- *E* – qualitative feature – radar echo (1 – mobile, 2 – stationary radar) (Cook and Weisberg, 1980).

The data include $n = 24$ vector observations for $p+1 = 7$ features. The estimated linear regression model and calculated statistics – multiple correlation coefficient, adjusted multiple determination coefficient, standard deviation of random error evaluation, Fisher-Snedecor test statistic, probability of rejection of zero hypothesis – took the form:

$$\hat{y} = 12.898 + 0.726x_1 - 0.0154x_2 - 0.223x_3 - 0.013x_4 + 0.311x_5 + 0.627x_6,$$
$$R = 0.717, \quad \widetilde{D} = 34.2\%, \quad s = 0.769, \quad F = 2.991, \quad \widetilde{p}_F = 0.035,$$

The standard deviations of evaluations of structural parameters and the probability of rejection of zero hypotheses concerning structural parameters are recorded in Table 1.

Table 1. The standard deviations of evaluations of structural parameters and $\tilde{p}_j$-value

j	0	1	2	3	4	5	6
$D(\hat{\beta}_i)$	3.392	0.340	0.008	0.219	0.028	0.228	0.430
$\tilde{p}_j$	0.001	0.047	0.070	0.323	0.636	0.190	0.164

The calculations show that:

- the multiple correlation coefficient is moderately high (0.717),
- the standard error evaluation value is high (0.769),
- the estimated linear regression model is significant for the adopted significance level of 0.05
- significance at level 0.05 was found only for feature A, while at a significance level of 0.1 only the two features A and T can be considered significant,
- features A, ln_P and E have a positive influence on the dependent variable, while the others have a negative influence,
- the residuals, in increasing order, together with the observation numbers, are given in Table 2.
- high absolute values of residuals are found for observations 7 (-2.1808) and 15 (1.0642).

Table 2. The residuals together with the observation numbers

Obs.	Res.	Obs.	Res.	Obs.	Res.	Obs.	Res.	Obs.	Res.	Obs.	Res.
7	-2.181	17	-0.408	23	-0.141	6	0.071	16	0.274	9	0.726
24	-0.725	2	-0.404	19	-0.075	21	0.117	20	0.283	8	0.803
5	-0.686	22	-0.213	12	-0.046	10	0.236	14	0.306	13	0.876
3	-0.419	11	-0.200	18	0.023	4	0.244	1	0.472	15	1.064

Based on the descriptive analysis of continuous variables (with particular attention to the smallest and largest residuals) the first set of potential influential observations W_1={1, 2, 3, 5, 6, 7, 11, 13, 15, 24} was identified. Alongside the given set W_1, we consider a new set of influential observations determined using analysis of leverage points, namely the diagonal elements of matrix **H.** Their values, in non-increasing order, are recorded in Table 3.

Table 3. The diagonal elements of matrix **H**

No.	Obs.	h_{ii}	*No.*	Obs.	h_{ii}	*No.*	Obs.	h_{ii}
1	2	0.852	9	7	0.327	17	21	0.199
2	6	0.596	10	20	0.322	18	22	0.199
3	3	0.438	11	23	0.258	19	9	0.179
4	24	0.418	12	4	0.241	20	8	0.165
5	17	0.359	13	10	0.228	21	19	0.165
6	1	0.348	14	13	0.223	22	14	0.140
7	18	0.337	15	15	0.218	23	12	0.129
8	5	0.334	16	11	0.207	24	16	0.117

Interpretation of leverage points:

- leverage values are contained in the interval $\langle 0.117, 0.852 \rangle$,
- the highest leverage value is found for case 2 (0.852),
- high-leverage points exceeding the threshold value 0.292 (=7/24) determine a new set of influential observations: W_2={ 1, 2, 3, 5, 6, 7, 17, 18, 20, 24}.

On combining the two sets we obtain the set $W_{1,2} = W_1 \cup W_2$ = {1, 2, 3, 5, 6, 7, 11, 13, 15, 17, 18, 20, 24}. The obtained set of observations is reduced by analysing the matrix $\mathbf{H}^*$. The diagonal points of that matrix are given in Table 4.

Based on the values of the diagonal elements of matrix $\mathbf{H}^*$ ($h_0^* = 0.33$) and the differences $|\mathbf{H} - \mathbf{H}^*|$ the set $W_{1,2}$ was reduced to W_3 = {1, 2, 3, 5, 6, 7, 15, 17, 18, 20, 24}. From the preliminary analysis 11 observations were finally distinguished.

Reduction of the given set W_3 was performed by diagnostic analysis of 1-cut models. Detailed analyses are contained in (Budka, 2005). For our further needs, we give the calculated values of the Cook statistic.

Table 4. The diagonal elements of matrix $\mathbf{H}^*$

No.	*i*	h_{ii}^*	*No.*	*i*	h_{ii}^*	*No.*	*i*	h_{ii}^*
1	2	0.853	9	7	0.327	17	21	0.200
2	6	0.596	10	20	0.322	18	22	0.199
3	3	0.438	11	23	0.258	19	9	0.179
4	24	0.418	12	4	0.241	20	8	0.165
5	17	0.359	13	10	0.228	21	19	0.165
6	1	0.348	14	13	0.223	22	14	0.141
7	18	0.337	15	15	0.218	23	12	0.129
8	5	0.334	16	11	0.207	24	16	0.117

Table 5. The values of the Cook statistic

No.	D_i	*No.*	D_i	*No.*	D_i	*No.*	D_i	*No.*	D_i	*No.*	D_i
12	0.000	23	0.002	14	0.004	20	0.014	1	0.044	15	0.097
18	0.000	16	0.003	6	0.004	9	0.034	3	0.059	24	0.157
19	0.000	11	0.003	10	0.005	17	0.035	13	0.069	7	0.830
21	0.001	22	0.003	4	0.006	8	0.037	5	0.086	2	1.544

It was finally determined, based on analysis of the structural parameters of the regression model and the Cook measure (threshold value 0.1) that the influential cases are contained in the set $W_4 = \{2, 7, 15, 24\}$.

Table 6. The influential cases

No.	Date	**A**	**T**	**S**	**C**	**ln_P**	**E**	**ln_Y**
		x_1	x_2	x_3	x_4	x_5	x_6	y
2	22/06	1	1	2.7	37.9	16.355	1	17.826
7	09/07	0	18	1.3	4.6	14.937	1	15.363
15	29/07	1	38	2.05	7	14.18	1	18.591
24	12/09	0	83	4.65	7.4	14.334	1	14.845

Analysis of the 1-cut models distinguished the single influential cases 2, 7, 15, 24. For possible detection of masking, weakening, amplifying and conserving effects, the sequence of observations given earlier was extended to the 9 observations with the numbers 1, 2, 3, 5, 7, 9, 15, 17 and 24 based on

detailed analysis of 1-cut models taking account of t-Student statistics with values higher than for the full model. For each of 36 pairs of observations the Cook statistics were determined in a 2-cut model. Examples of their values for the first case with the remaining ones are given in Table 7.

Table 7. The Cook statistics determined in a 2-cut model for the first case with the remaining ones are given

	1	2	3	4	5	6	7	8
i	1	1	1	1	1	1	1	1
J	2	3	5	7	9	15	17	24
h_{ii}	0.138	0.126	-0.074	0.138	0.120	-0.003	0.157	-0.043
h_{jj}	0.348	0.348	0.348	0.348	0.348	0.348	0.348	0.348
h_{ij}	0.852	0.438	0.334	0.327	0.179	0.218	0.359	0.418
r_i	0.472	0.472	0.472	0.472	0.472	0.472	0.472	0.472
r_j	-0.404	-0.419	-0.686	-2.181	0.726	1.064	0.165	-0.725
D_i	0.044	0.044	0.044	0.044	0.044	0.044	0.044	0.044
D_j	1.544	0.059	0.086	0.830	0.034	0.097	0.035	0.157
d_{ij}	0.077	0.350	0.429	0.420	0.521	0.510	0.393	0.378
B_1	2.472	0.113	0.133	0.955	0.082	0.142	0.090	0.203
B_2	0.408	0.020	0.008	0.215	0.016	0.000	0.016	0.004
B_3	-1.548	-0.084	0.056	-0.353	0.070	-0.002	0.034	0.043
$D_{\{i,j\}}$	1.332	0.048	0.197	0.817	0.168	0.139	0.140	0.250
r_i*r_j	-0.191	-0.198	-0.324	-1.030	0.343	0.503	0.078	-0.342

Depending on the values of the Cook statistics for the 1-cut model and 2-cut models, one can distinguish four types of observations:

- *amplifying* – one or both of the statistics D_i, D_j are less than $D_{\{i, j\}}$: (1, 5), (1, 9), (1, 15), (1, 17), (1, 24), (2, 5), (2, 17), (2, 24), (3, 5), (3, 9), (3, 15), (5, 9), (5, 15), (5, 17), (5, 24), (9, 24), (15, 17), (15, 24) and (17, 24),
- *weakening* – one or both of the Cook statistics for 1-cut models are greater than the Cook statistic for the 2-cut model: (1, 2), (2, 7), (2, 9), (5, 7), (7, 5), (7, 24) and (9, 24),
- *conserving* – both Cook statistics for 1-cut models are almost equal to the Cook statistic for the 2-cut model: (1,3), (2,3), (2, 15), (3, 7), (3, 17), (3, 24), (7, 17) and (9, 17),

- *attracting* – one of the Cook statistics for the 1-cut model with large values and the other with smaller values lead to large values of the Cook statistics for the 2-cut model, e.g. pair (1,7).

In each of the situations listed, the changes that take place between the Cook statistic values for the 1-cut and 2-cut models are of a different order. In this meaning all pairs with small values for the statistic $D_{\{i,\ j\}}$ should be omitted. Adopting an arbitrary threshold value for that statistic, equal to 0.5, there will remain for consideration the system of pairs (1, 2), (1, 7), (2, 3), (2, 5), (2, 7), (2, 9), (2, 15), (2, 17), (2, 24), (3, 5), (3, 7), (5, 7), (7, 9), (7, 15), (7, 17) and (7, 24). Among these, based on the principle for investigating pairs of influential observations as given after formula (3), five pairs can be identified as influential: (2, 5), (2, 17), (2, 24), (3, 5) and (7, 17). Among these pairs, cases 2, 3, 5, 7, 17 and 24 are distinguished.

In order to recognize the consequences of the existence of the listed observations in 2-cut models, regression analysis was performed for 15 distinguished 2-cut models containing the pairs (2, 3), (2, 5), (2, 7), …, (17, 24).

In the analysis of 2-cut models, large values were obtained for the statistic F after cutting of the following pairs: (2, 7) – F = 14.599, (3, 7) – F = 11.301, (5, 7) – F = 12.668. To the selected pair with the highest F there was added one case at a time from the others listed, leading to consideration of three 3-cut models with the cases (2, 3, 7), (2, 5, 7) and (2, 7, 24). For each of these three triples, regression analysis of the 3-cut model was carried out.

- The greatest value of F (17.0432) was obtained for the triple of observations 2, 5 and 7, and the smallest (1.808) for the triple of observations 2, 3 and 24.
- Small values of the statistic F were obtained for the triples of observations (2, 3, 5), (2, 3, 24), (2, 5, 24) and (3, 5, 24).
- Apart from the situations listed in (ii), in the remaining 3-cut models significance at the level 0.05 was found for the features A, T, S, ln_P.
- Only for one triple (3, 5, 7) was feature C found to be significant alongside A, T, S and ln P.

- In all of the given 3-cut models feature E proved to be insignificant.

In summary, it can be concluded that:

- the single influential observations are found to be observations 2, 7 and 24,
- the influential pairs of observations are found to be (2, 7), (3, 7), (5, 7) and (7, 24),
- for distinguished 3-cut models, the best fit is found to be the linear regression model excluding observations 2, 5 and 7.

4. Conclusions

This work has presented diagnostic procedures used for investigating vectors of influential observations in a linear regression model. They lead to the identification of systems of observations which are considered to be influential. Of particular interest are those which contain cases which were not previously considered influential in analysis of the 1-cut model. The reverse situation may occur, where a given system of submatrices is regarded as influential, although after the addition of new observations to those submatrices it transpires that the new system of enlarged submatrices is no longer regarded as influential.

The method presented for sequential determination of m-element vectors of influential observations is a proposed procedure when considering the problem of investigation of influential observations.

REFERENCES

Atkinson A. C. (1985): Plots, Transformations and Regressions. An Introduction to Graphical Methods of Diagnostic Regression Analysis. Clarendon Press, Oxford.

Belsley D. A., Kuh E., Welsch R. E. (1980): Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley and Sons, New York.

Budka A., Wagner W. (2007): Analysis of Linear Regression Model at Divided System Matrix. Acta Universitatis Lodziensis. Folia Oeconomica 206: 31–43.

Chatterjee S., Hadi A. S. (1988): Sensitivity Analysis in Linear Regression. Wiley and Sons, New York.

Cook R. D., Weisberg S. (1980): Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. Technometrics 22: 495–507.

Cook R. D., Weisberg S. (1982): Residuals and Influence in Regression. Chapman and Hall, New York.
Draper N.R., Smith H. (1973): Analiza Regresji Stosowana. PWN, Warsaw.
Fox J.(2002): An R and S-PLUS Companion to Applied Regression. Sage, California.
Gray J. B., Ling R. F. (1984): K-Clustering as a Detection Tool for Influential Subsets in Regression. Technometrics 26: 305–318.
Lawrance J. (1995): Deletion Influence and Masking in Regression. J. R. Statist. Soc. B, 57: 181–189.
Oktaba W. (1986): Metody Matematyczne w Doświadczalnictwie. PWN, Warsaw.
Ostasiewicz W. (ed.) (1999): Statystyczne Metody Analizy Danych. PWN, Warsaw.
Pena D., Yohai V. (1995): The Detection of Influential Subsets in Linear Regression Using an Influence. J. R. Statis. Soc. B, 57: 145–156.
Rousseeuw P. J , Leroy A.M. (1987): Robust Regression and Outlier Detection. Wiley and Sons, New York.