ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

Anna Budka[*], *Wiesław Wagner*[**]

ANALYSIS OF LINEAR REGRESSION MODEL AT DIVIDED SYSTEM MATRIX

Abstract. In this study problems connected with the detection of influential observations are investigated in the linear regression model using the least squares estimation of structural parameters. This issue has been presented in three cuts: the complete model, 1-cut model and *m*-cut model.

Key words: linear regression model, influential observations, hat matrix, complete, 1-cut and *m*-cut model.

1. INTRODUCTION

Methods of causality analysis are commonly used in research. Since the moment of its formulation until recently the least squares method has been developed in terms of its theoretical foundations by the so-called diagnostic support. These are advanced statistical methods making it possible to determine which units have a significantly large effect on the quality of estimated parameters of the linear regression model. In consequence this leads to the determination of the so-called influential observations. Frequently they take the form of outliers or leverage observations.

Influential observations may occur separately or in groups. While in the former case numerous methods have been developed to detect them, statistical methods are still being developed for the detection of many such observations occurring simultaneously. Separation of such observations one by one, even at the applied step procedure, does not necessarily provide

[*] Ph.D., Agricultural University of Poznań.

[**] Professor, University of Information Technology and Management in Rzeszów.

correct solutions. This is so due to the fact that in the set of observable explanatory variables the so-called masking effect may occur, i.e. single observations may individually be influential observations, although their cluster may not confirm it.

In this study problems connected with the detection of influential observations are investigated in the linear regression model using the least squares estimation of structural parameters. This issue has been presented in three cuts: the complete model, 1-cut model and *m*-cut model. Detailed methods to investigate influential observations are presented in each case. For this purpose the primary statistics are diagonal elements of the so-called orthogonal projection matrices. Their high values, while all belonging to the (0, 1) interval, and exceeding set threshold values, make it possible to indicate the occurrence of influential observations. Obviously, various possible statistics being to some extent functions of elements of the above mentioned matrix will provide diagnostic information of varying importance concerning influential observations.

2. DETERMINATION OF THE LINEAR REGRESSION MODEL

It is assumed that the investigations covered "causality" $x_1, x_2, ..., x_{p-1} \rightarrow y$ of set $p-1$ of established causes defined by the system $x_1, x_2, ..., x_{p-1}$ "of independent variables" (explanatory, clarifying, regression variables) and effect y being a "dependent variable" (explained or clarified variable, predictor). Moreover, it is assumed that the set effect, apart from the mentioned controlled causes, is affected also by random causes expressed by variable e. In relation to it stochastic assumptions are made:

(E1) $\mathrm{E}(e) = 0$, expected value equals 0,

(E2) $\mathrm{D}(e) = \sigma$, standard deviation equals a certain positive constant $\sigma > 0$,

(E3) $e \sim \mathrm{N}(0, \sigma)$, random variable e has a "normal distribution" (Gaussian distribution) with indicated parameters.

Assumptions (El) and (E2) concern moments of random variable e, whereas assumption (E3) defines its distribution type. This means that random variable e is of the continuous type and its distribution belongs to the class of normal distributions $\mathcal{N}\{(\mu, \sigma): \mu = 0, \sigma > 0\}$. The connection between variable y and variables $x_1, x_2, ..., x_{p-1}$, e is presented by an "additive linear model"

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} + e \qquad (1)$$

also called the "multiple rectilinear regression model". Constants $\beta_0, \beta_1, ..., \beta_{p-1}$ are called "structural parameters or regression coefficients", where paameter β_0 is a free term, whereas $\beta_1, ..., \beta_{p-1}$ are slopes expressing unitary changes of variables $x_1, x_2, ..., x_{p-1}$ on variable y.

Variables y and e are treated as certain "random variables" with set "probability distributions", whereas $x_1, x_2, ..., x_{p-1}$ are set "real variables". Random variable e in model (1) is called "random error" (component).

Random variable y due to the set equation (1) and adopted assumptions (E1), (E2), (E3) for random variable e assumes the following parameters and type of distribution:

$$\text{(Y1)}\,\mathrm{E}(y) = \sum_{j=0}^{p-1} \beta_j x_j,\ \ \text{(Y2)}\,\mathrm{D}^2(y) = \sigma^2,\ \ \text{(Y3)}\, y \sim \mathrm{N}\sum_{j=0}^{p-1} \beta_j x_j, \sigma).$$

Thus assumption (Y3) means that random variable y belongs to the class of normal distributions $\mathcal{N}\left\{(\mu, \sigma)\colon\ \mu = \sum_{j=0}^{p-1} \beta_j x_j \in R, \sigma > 0\right\}$. For the purpose of estimation of unknown structural parameters $\beta_0, \beta_1, ..., \beta_{p-1}$ and standard deviation σ of the random component, statistical testing is conducted on a finite set n of "units" (cases) $J_1, J_2, ..., J_n$. These units constitute a "random sample" selected from a certain "general population" according to a set "sampling pattern". It is assumed that for each unit J_i a $(p-1)$-dimensional "vector of observation" $\mathrm{x}_i' = (x_{i1}, x_{i2}, ..., x_{i,p-1})$ is known and y_i, $i = 1, 2, ..., n$ on independent variables and the dependent variable. This system of n vectors of observation constitutes a "multidimensional sample" with size n. This sample makes it possible to present model (1) in the form of a vector-matrix linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2}$$

where $\mathbf{X} = (\mathbf{x}_0', \mathbf{x}_1', \mathbf{x}_2', ..., \mathbf{x}_n')' : n \times p$ – the system matrix at $\mathbf{x}_0 : n \times 1$ unit vector, $\mathbf{y} : n \times 1$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1}) : p \times 1$ – vector of structural parameters, $\mathbf{e} = (e_1, e_2, ..., e_n) : n \times 1$ – vector of random errors.

From the adopted stochastic assumptions for random errors in model (1), we obtain the following assumptions for the vector of random errors $\mathbf{e}$ in model (2):

(WE1) $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, vector of expected values is equal to the zero vector,

(WE2) $\mathrm{D}^2(\mathbf{e}) = \sigma^2\mathbf{I}$, variance-covariance matrix is equal to the scalar diagonal matrix, where $\sigma^2 > 0$ is the variance of random errors. This assumption states also that components of random vector $\mathbf{e}$ are not correlated, i.e. "covariance" $\mathrm{cov}(e_{i,}, e_{i'}) = 0$ for $i \neq i'$; $i, i' = 1, 2, ..., n$,

B.U.Ł

(WE3) $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, random vector $\mathbf{e}$ has an n-dimensional normal distribution with a zero vector of expected value $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\Sigma = \sigma^2\mathbf{I}$, i.e. belongs to the class of normal distributions $\mathcal{N}_n\{(\boldsymbol{\mu}, \Sigma) : \boldsymbol{\mu} = \mathbf{0}, \Sigma = \sigma^2\mathbf{I}\}$.

Assumption (WE3) implies that vector of observation $\mathbf{y}$ has also an n-dimensional normal distribution belonging to class $\mathcal{N}_n\{(\boldsymbol{\mu}, \Sigma) : : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \Sigma = \sigma^2\mathbf{I}\}$, which means that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $D^2(\mathbf{y}) = \sigma^2\mathbf{I}$.

3. SELECTED PROPERTIES OF LINEAR REGRESSION MODEL

For the purposes of investigations of the influence of influential observations on the quality of evaluation of the structural parameter vector, numerous analytical results, connected with a complete linear regression model, are derived. These results may be found in many publications concerning the theory of linear models. Some of them have been presented in new versions using orthogonal projection matrices. Proofs of many of them are available in literature. While presenting results we are using their specification in the set problem groups.

(S1) Estimation using the least squares method:

a) $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$ – the system of standard equations, where $\mathbf{X}'\mathbf{X}$ – matrix of moments on explanatory variables, $\mathbf{X}'\mathbf{y}$ – vector of moments of explanatory variables and the explained variable,

b) $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{G}\mathbf{X}'\mathbf{y} = \mathbf{A}\mathbf{y}$ – estimator of structural parameters vector, where $\mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{A} = \mathbf{G}\mathbf{X}'$ – associated matrix,

c) $E(\hat{\boldsymbol{\beta}}) = \mathbf{A}E(\mathbf{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ – property of unbiasedness,

d) $D^2(\hat{\boldsymbol{\beta}}) = \mathbf{A}D^2(\mathbf{y})\mathbf{A}' = \sigma^2\mathbf{A}\mathbf{A}' = \sigma^2\mathbf{G}$ – a variance-covariance matrix,

e) $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2\mathbf{G})$ – distribution of estimator $\hat{\boldsymbol{\beta}}$.

(S2) Vector of estimated observations:

a) $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{A}\mathbf{y} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$ – evaluation of vector $\hat{\mathbf{y}}$, where $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ – orthogonal projection matrix,

b) $E(\hat{\mathbf{y}}) = \mathbf{H}E(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$ – expected value of vector $\hat{\mathbf{y}}$,

c) $D^2(\hat{\mathbf{y}}) = D^2(\mathbf{H}\mathbf{y}) = \mathbf{H}D^2(\mathbf{y})\mathbf{H}' = \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H} = \sigma^2\mathbf{H}\mathbf{H} = \sigma^2\mathbf{H}$ – variance-covariance matrix.

(S3) Residuls

a) $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}$ – vector of residuals, where $\mathbf{M} = \mathbf{I} - \mathbf{H}$,

b) $E(\mathbf{r}) = \mathbf{M}E(\mathbf{y}) = \mathbf{M}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ – expected value of vector of residuals,

c) $\mathrm{D}^2(\mathbf{r}) = \mathrm{D}(\mathbf{My}) = \mathbf{M}\mathrm{D}(\mathbf{y})\mathbf{M}' = \sigma^2\mathbf{MM} = \sigma^2\mathbf{M}$ – variance-covariance matrix of vector of residuals.

(S4) Sum of squares for SSE error of the least squares method:

a) $SSE = \mathbf{r}'\mathbf{r} = \mathbf{y}'\mathbf{My}$ – sum of squares for error,

b) $SSE = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Hy} = \dfrac{\begin{vmatrix} \mathbf{y}'\mathbf{y} & \mathbf{y}'\mathbf{X} \\ \mathbf{X}'\mathbf{y} & \mathbf{X}'\mathbf{X} \end{vmatrix}}{|\mathbf{X}'\mathbf{X}|}$ – SSE expression as a quotient of two determinants of an augmented matrix of the $(\mathbf{X}, \mathbf{y})$ system, and the matrix of the $\mathbf{X}$ system, which results from formula $\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}A_{21}|$, where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square matrices and $\mathbf{A}_{22}$ is non-singular,

c) $\mathrm{E}(SSE) = \mathrm{E}(\mathbf{y}'\mathbf{My}) = \mathrm{E}(\mathbf{y}')\mathbf{M}\mathrm{E}(\mathbf{y}) + \mathrm{tr}[\mathbf{M}\mathrm{D}(\mathbf{y})] = \sigma^2\mathrm{tr}(\mathbf{M}) = (n-p)\sigma^2$,

d) $SSE \sim \chi^2_{n-p}$ – the chi-square distribution of sum of squares for error,

e) $\mathrm{E}(SSE) = (n-p)\sigma^2$ and $\mathrm{D}^2(SSE) = 2(n-p)\sigma^4$ – moments of random variable SSE resulting directly from the distribution given in point c),

f) $SSY = SSR + SSE$ – the factorization of sum of squares of deviations into two summands: sum of squares for regression and sum of squares for error, where $SSY = \mathbf{y}'(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{y}$ and $SSR = \mathbf{y}'(\mathbf{H} - \mathbf{1}\mathbf{1}'/n)\mathbf{y}$.

(S5) Estimation of parameter σ^2:

(a) $s^2 = \dfrac{SSE}{n-p} = \dfrac{\mathbf{y}'\mathbf{My}}{r(\mathbf{M})}$ – estimator of parameter σ^2,

(b) $\mathrm{E}(s^2) = \sigma^2$ – property of unbiasedness of estimator s^2,

c) $\mathrm{D}^2(s^2) = \mathrm{D}^2\left(\dfrac{SSE}{n-p}\right) = \dfrac{2(n-p)\sigma^2}{(n-p)^2} = \dfrac{2\sigma^2}{n-p}$ – variance of estimator s^2.

(S6) Augmented matrix of orthogonal projection

a) $\mathbf{H}^* = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, $\mathbf{H}^* : n \times n$, $\mathbf{Z} = (\mathbf{X}, \mathbf{y}) : n \times (p+1)$,

b) $\mathbf{H}^*$ possesses all properties of matrix $\mathbf{H}$,

c) $\mathbf{H}^* = \mathbf{H} + \dfrac{\mathbf{rr}'}{\mathbf{r}'\mathbf{r}} = \mathbf{H} + \dfrac{\mathbf{rr}'}{SSE}$,

d) $h^*_{ij} = h_{ij} + \dfrac{r_i r_j}{SSE}$, $i, j = 1, 2, \ldots, n$,

e) $h^*_{ii} = h_{ii} + \dfrac{r_i^2}{SSE}$,

while properties c) and d) are given after J. B. Gray and R. F. Ling (1984).

4. THE CONCEPT OF INFLUENTIAL OBSERVATIONS

For the purpose of formal determination of influential observations let us introduce the notation of a sample of $n(p+1)$-dimensional observations as a sequence of row vectors of matrix $\mathbf{X}$ and of vector $\mathbf{y}$ in the form of $(\mathbf{X},\mathbf{y}) = ((\mathbf{x}_1', y_1), (\mathbf{x}_2', y_2), \ldots, (\mathbf{x}_n', y_n))' = ((\mathbf{x}_i', y_i);\ i = 1,2,\ldots,n)' = P_n^{p+1}$, while $(\mathbf{x}_i', y_i) \in R^{p+1}$. It is necessary to use the transposition sign, as vectors are always treated as column vectors.

The concept of influential observations is given in the following definition.

Definition. *The system of* $m\,(m \geqslant 1)$ *vector observations* $\{(\mathbf{x}_{i1}', y_{i1}), \ldots, (\mathbf{x}_{im}', y_{im})\}$ *in sample* P_n^{p+1} *indexed by the discriminanted set of* m *indexes* $\{i_i, i_2, \ldots, i_m\} \in \{1, 2, \ldots, n\}$ *is called influential observations (points) if they significantly contribute to changes in the values of analyzed numerical characteristics referring to the investigated model of linear regression.*

The above definition indicates that among numerical data there may be one ($m = 1$) or more observable vectors ($m > 1$), which will constitute outliers in the direction of the x-axis or in the direction of the y-axis, or both at the same time. In the first case such influential points may be detected by inspecting diagonal elements of matrix $\mathbf{H}$, whereas in the second case – by investigating Studentized residuals. There are numerous solutions to this problem.

While determining influential observations diagonal elements of matrix $\mathbf{H}$ are used. These elements are expressed by vector values referring to explanatory variables. This makes it possible to investigate the behaviour of atypicality of these vectors manifested in their distance from the regression cluster. Excessive concentration is understood as a homogenous set corresponding to the regression dependency characteristic. In case of a regression model with one explanatory variable this is equivalent to the configuration of points on a plane arranged along a certain straight line, whereas in case of a multiple regression model such a configuration constitutes a generalized ellipsoid with an intersecting hyperplane. This means that the vector cases will vary in their number in the estimated linear regression model. In connection with the above remarks, we are going to introduce the following definition.

Definition. *Observations of cases, i.e. row vector of matrix* $\mathbf{X}$ *corresponding to diagonal values of matrix* $\mathbf{H}$ *are called levearage points.*

Values of diagonal elements of matrix $\mathbf{H}$ fall within the interval of $(1/n, 1)$, i.e. they are normalized numbers independent of the number of cases n and the number of properties $p-1$. Thus, assuming the *a priori* set threshold value, let's say h_0 for these diagonal elements, "leverage points" may be distinguished among them, exceeding this value. They are commonly referred to as "high-leverage points". Such points are interesting in terms of their effect on the estimated linear regression model.

5. 1-CUT LINEAR REGRESSION MODEL

In case of investigations of regression models, it is interesting to study the dependency between the complete system matrix and its submatrix divided with the use of vectors. Such a division is most frequently connected with the fact that it is necessary to investigate the submatrix distinguished from the matrix of system $\mathbf{X}$ in the context of estimating model parameters excluding any of the observed vectors of matrix $\mathbf{X}$.

The row division of matrix $\mathbf{X}$ is understood as follows. Let us say set $\{1, 2, ..., n\}$ denotes successive numbers of row vectors of matrix $\mathbf{X}$. Let us assume that in this set an i-th vector was distinguished, which is transposed with the n-th vector. Vectors with numbers $i+1, i+2, ..., n-1, n$ will be transposed by one position to loci with numbers $i, i+1, ..., n-1$. Such a procedure is called the operation of re-numbering and translocation of row vectors of matrix $\mathbf{X}$.

Let us denote with the symbol $\mathbf{X}_{(i)}$ a submatrix formed from matrix $\mathbf{X}$ without the i-th row vector $\mathbf{x}_i'$. According to the operation presented above, this means the division of the system matrix into $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(i)} \\ \mathbf{x}_i' \end{bmatrix}$, where $\mathbf{X}: n \times p$, $\mathbf{X}_{(i)}: (n-1) \times p$ is a cut matrix and $\mathbf{x}_i': 1 \times p$ is a distinguished vector, and moreover $\mathbf{X}'\mathbf{X} = [\mathbf{X}_{(i)}' \, \mathbf{x}_i] \begin{bmatrix} \mathbf{X}_{(i)} \\ \mathbf{x}_i' \end{bmatrix} = \mathbf{X}_{(i)}'\mathbf{X}_{(i)} + \mathbf{x}_i\mathbf{x}_i'$. To the above division we make a permanent assumption of rank $\mathrm{r}(\mathbf{X}_{(i)}) = p$ and we introduce the following denotations:

a) $\mathbf{G}_{(i)} = (\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1}$, $\mathbf{G}_{(i)}: p \times p$,
b) $\mathbf{H}_{(i)} = \mathbf{X}_{(i)}\mathbf{G}_{(i)}\mathbf{X}_{(i)}'$, $\mathbf{H}_{(i)}: (n-1) \times (n-1)$,
c) $\mathbf{v}_{(i)} = \mathbf{X}_{(i)}\mathbf{G}_{(i)}\mathbf{x}_i$, $\mathbf{v}_{(i)}: (n-1) \times 1$,
d) $c_i = \mathbf{x}_i'\mathbf{G}_{(i)}\mathbf{x}_i$,
e) $d_i = (1 + c_i)^{-1}$,
f) $h_{kl(i)} = \mathbf{x}_k'\mathbf{G}_{(i)}\mathbf{x}_l$, $k, \; l = 1, 2, ..., n$.

In correspondence to the given division of matrix **X** with one row vector, the division is conducted of the vector of observable random variables $\mathbf{y} = \begin{bmatrix} \mathbf{y}_{(i)} \\ y_i \end{bmatrix}$, where $\mathbf{y}_{(i)} : (n-1) \times 1$, while y_i is the i-th component of vector **y**.

We will further use the denotation "1-cut" to emphasize that it refers to a linear model investigated for vector **y** and the matrix of system **X** after elimination of the i-th component of vector **y** and the i-th row vector in matrix **X**. Let us denote a 1-cut model in the ternary form

$$\{\mathbf{y}_{(i)}, \mathbf{X}_{(i)}\boldsymbol{\beta}_{(i)}, \sigma^2\mathbf{I}\} \tag{3}$$

where here the identity matrix **I** is of the $(n-1)$-th degree. We will add to the previously given characteristics $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{y}}$, $\mathbf{r}$, SSE their equivalents $\boldsymbol{\beta}_{(i)}$, $\hat{\mathbf{y}}_{(i)}$, $\mathbf{r}_{(i)}$, $SSE_{(i)}$, after the application of the row division of the system matrix **X** and vector **y**.

Stochastic properties for the 1-cut model will be noted according to the same principle as for the complete model, but denotation "-1" is additionally placed for emphasize the fact that the cut model is used.

(S1-1) Least squares estimation:

a) $\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{G}_{(i)}\mathbf{X}_{(i)}'\mathbf{y}_{(i)} = \mathbf{A}_{(i)}\mathbf{y}_{(i)}$ – estimated vector of structural parameters,

b) $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \mathbf{G}_{(i)}\mathbf{x}_i r_i = \dfrac{r_i}{m_{ii}}\mathbf{G}\mathbf{x}_i$ – the difference between estimated vectors of structural parameters of the complete and 1-cut models, where r_i denote residuals,

(S2-1) Forecasting vector **y** in the 1-cut model:

a) $\hat{\mathbf{y}}_{(i)} = \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)} = \hat{\theta}_{(i)}$ – estimation (prediction) of vector **y**,

b) $\hat{y}_{j(i)} = \mathbf{x}_j'\hat{\boldsymbol{\beta}}_{(i)} = \hat{\theta}_{j(i)}$ – j-th component of vector $\hat{\mathbf{y}}_{(i)}$, i.e. forecast of the j-th observation after the elimination of the i-th observation,

c) $\hat{y}_j - \hat{y}_{j(i)} = \dfrac{r_i}{m_{ii}}\mathbf{x}_j'\mathbf{G}\mathbf{x}_i$ – the difference of the j-th component estimated in the complete and 1-cut models,

d) $\hat{y}_i = h_{ii}y_i + m_{ii}\theta_{(i)}$ – results from (S1-1) b) – linear combination of observations y_i and $\hat{\theta}_{(i)} = \mathbf{x}_i'\hat{\beta}_{(i)}$ at weights expressed by the i-th diagonal element of matrix **X**,

(S3-1) Residuals and sum squares for error:

a) $r_{(i)} = y_i - \mathbf{x}_i'\hat{\beta}_{(i)} = y_i - \hat{\theta}_{(i)} = \dfrac{r_i}{m_i}$ – residuals in the 1-cut model,

b) $SSE_{(i)} = \mathbf{y}_{(i)}'(\mathbf{I} - \mathbf{H}_{(i)})\mathbf{y}_{(i)}$ – sum of squares for error in the 1-cut model, where $\mathbf{H}_{(i)}$ is the submatrix of matrix $\mathbf{H}$,

c) $SSE = SSE_{(i)} + d_i r_{(i)}^2$ – connection between sum of squares for error in the complete and 1-cut models,

d) $s_{(i)}^2 = \frac{1}{\nu - 1} \sum_{j=1}^{n} (y_j - \mathbf{x}_j'\hat{\boldsymbol{\beta}}_{(i)})^2 = \frac{SSE_{(i)}}{\nu - 1} = \frac{\nu}{\nu - 1} s^2 - \frac{r_i^2}{(\nu - 1)m_{j \neq i}}$ – estimator of parameter σ^2 in the 1-cut model,

e) $\mathrm{D}^2(r_{(ii)}) = s_{(i)}[1 + \mathbf{x}_i' \mathbf{G}_{(i)} \mathbf{x}_i]^{1/2} = \frac{s_{(i)}}{\sqrt{m_{ii}}}$ – standard error for residuals in the 1-cut model (Haoglin, Welsch 1978),

f) $\nu s^2 = (\nu - 1)s_{(i)}^2 + w_{ii} r_{(i)}^2$ – equation b) expressed by estimators s^2 and $s_{(i)}^2$ of parameter σ^2 of the complete and 1-cut models, where $w_{ii} = \frac{h_{ii}}{m_{ii}}$,

g) $s_{(i)}^2 \sim \frac{\sigma^2}{\nu - 1} \chi_{\nu - 1}$ – chi-square distribution of estimator $s_{(i)}^2$,

h) squares of residuals r_i^2 and estimates $s_{(i)}^2$ are statistically independent (La Motte 1994),

(S4-1) Studentized residuals:

a) $t_{(i)} = \frac{r_{(i)}}{D(r_{(i)})} = \frac{r_i}{s_{(i)}\sqrt{m_{ii}}}$ – expression of parameter σ^2 by estimator $s_{(i)}^2$ in the 1-cut model (the so-called "external Studentized residuals of least squares"),

b) $t_{(i)} = t_i \sqrt{\frac{\nu - 1}{\nu - t_i^2}}$ – expression by standardized residuals and t-Student distribution with $\nu - 1$ degrees of freedom.

6. *M*-CUT LINEAR REGRESSION MODEL

Apart from the investigations of influential observations in the 1-cut model, the problem of the m-cut model is also studied. Let us denote submatrix $\mathbf{X}_{(I)}$ formed without the subset of I row vectors of matrix $\mathbf{X}$ and $\mathbf{X}_I$ the system of such discriminated vectors. Matrix $\mathbf{X}$ will be denoted by the form: $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(I)} \\ \mathbf{X}_I \end{bmatrix}$, where $\mathbf{X} : n \times p$, $\mathbf{X}_{(I)} : (n - m) \times p$ – m-cut matrix and $\mathbf{X}_I' : m \times p$ – discriminated matrix (discriminated system of m row vectors). The following ratio occurs for it $\mathbf{X}'\mathbf{X} = [\mathbf{X}_{(I)}' \, \mathbf{X}_I] \begin{bmatrix} \mathbf{X}_{(I)} \\ \mathbf{X}_I \end{bmatrix} = \mathbf{X}_{(I)}' X_{(I)} + \mathbf{X}_I \mathbf{X}_I'$, and moreover for the given division we adopt a permanent assumption with rank $\mathbf{r}(\mathbf{X}_{(I)}) = p$ and we introduce the following denotations:

a) $\mathbf{G}_{(I)} = (\mathbf{X}_{(I)}'\mathbf{X}_{(I)})^{-1}$, $\mathbf{G}_{(I)} : p \times p$,
b) $\mathbf{H}_{(I)} = \mathbf{X}_{(I)}\mathbf{G}_{(I)}\mathbf{X}_{(I)}'$, $\mathbf{H}_{(I)} : (n-m) \times (n-m)$,
c) $\mathbf{F}_I = \mathbf{X}_I'\mathbf{G}_{(I)}\mathbf{X}_I$, $\mathbf{F}_I : m \times m$,
d) $\mathbf{V}_{(I)} = \mathbf{X}_{(I)}\mathbf{G}_{(I)}\mathbf{X}_I$, $\mathbf{V}_{(I)} : (n-m) \times m$,
e) $\mathbf{E}_I = [\mathbf{I} + \mathbf{F}_I]^{-1}$, $\mathbf{E}_I : m \times m$.

The given division of matrix of system $\mathbf{X}$ is applied to the vector of observable random variables $\mathbf{y} = \begin{bmatrix} \mathbf{y}_{(I)} \\ \mathbf{y}_I \end{bmatrix}$, where $\mathbf{y}_{(I)} : (n-m) \times 1$, while $\mathbf{y}_I : m \times 1$ is a column vector containing, without the loss of generality, the last m components of vector $\mathbf{y}$. Further we apply the "m-cut" denotation to emphasize that we are considering a linear model of vector $\mathbf{y}$ and system matrix $\mathbf{X}$ after the elimination from the above the indicated subvector $\mathbf{y}_I$ and matrix $\mathbf{X}_I$. Let us present the m-cut model in the ternary form

$$\{\mathbf{y}_{(I)}, \mathbf{X}_{(I)}\boldsymbol{\beta}_{(I)}, \sigma^2\mathbf{I}\} \tag{4}$$

where here the identity matrix $\mathbf{I}$ is a matrix of $(n-m)$-th degree. Let us also stress that symbol $\boldsymbol{\beta}_{(I)}$ does not refer to the cutting of the vector of structural parameters in model (4), but that we apply such a denotation to emphasize that a p-dimensional vector of parameters $\boldsymbol{\beta}$ is estimated from the m-cut model.

To stochastic characteristics $\hat{\boldsymbol{\beta}}, \hat{\mathbf{y}}, \mathbf{r}, SSE$ for the complete model, we will give their equivalents $\boldsymbol{\beta}_{(I)}, \hat{\mathbf{y}}_{(I)}, \mathbf{r}_{(I)}, SSE_{(I)}$ after the application of row division of the system matrix $\mathbf{X}$ and vector $\mathbf{y}$. Stochastic properties of the m-cut model will be denoted according to the same principle as for the 1-cut model, but additionally the denotation "-m" will be placed to emphasize the application of the m-cut model. Earlier we will give several denotations:

- $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(I)} \\ \mathbf{X}_I' \end{bmatrix}$, $\mathbf{X} : n \times p$, $\mathbf{X}_{(I)} : (n-m) \times p$, $\mathbf{X}_I' : m \times p$,
- $\mathrm{r}(\mathbf{X}_{(I)}'\mathbf{X}_{(I)}) = p$, $\mathbf{G}_{(I)} = (\mathbf{X}_{(I)}'\mathbf{X}_{(I)})^{-1}$.

(S1-m) Estimation of least squares:

a) $\mathbf{X}_{(I)}'\mathbf{X}_{(I)}\hat{\boldsymbol{\beta}}_{(I)} = \mathbf{X}_{(I)}'\mathbf{y}_{(I)}$ – system of normal equations,
b) $\hat{\boldsymbol{\beta}}_{(I)} = \mathbf{G}_{(I)}\mathbf{X}_{(I)}'\mathbf{y}_{(I)}$ – estimator of vector of structural parameters $\boldsymbol{\beta}$,
c) $\mathrm{E}(\hat{\boldsymbol{\beta}}_{(I)}) = \boldsymbol{\beta}_{(I)}$ and $\mathrm{D}^2(\hat{\boldsymbol{\beta}}_{(I)}) = \sigma^2\mathbf{G}_{(I)}$ at the assumptions of model (4),
d) $\hat{\boldsymbol{\beta}}_{(I)} = \hat{\boldsymbol{\beta}} + \mathbf{G}\mathbf{X}_I(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{r}_I$, where matrix $\mathbf{H}_I$ is the submatrix from

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(I)} \\ \mathbf{X}_I' \end{bmatrix} \mathbf{G}[\mathbf{X}_{(I)}' \, \mathbf{X}_I] = \begin{bmatrix} \mathbf{X}_{(I)}\mathbf{G}\mathbf{X}_{(I)}' & \mathbf{X}_{(I)}\mathbf{G}\mathbf{X}_I \\ \mathbf{X}_I'\mathbf{G}\mathbf{X}_{(I)}' & \mathbf{X}_I'\mathbf{G}\mathbf{X}_I \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{(I)} & \mathbf{H}_{I(I)} \\ \mathbf{H}_{I(I)} & \mathbf{H}_I \end{bmatrix},$$

and $\mathbf{r}_I = \mathbf{y}_I - \mathbf{X}_I'\hat{\boldsymbol{\beta}}$,

e) $\hat{\mathbf{y}}_{(I)} = \mathbf{H}_{(I)}\mathbf{y}_{(I)}$ – estimated vector $\mathbf{y}_{(I)}$ from model (4),

f) $\mathbf{X}_I'\hat{\boldsymbol{\beta}}_{(I)} = \mathbf{X}_I'\hat{\boldsymbol{\beta}} + \mathrm{H}_{(I)}\mathbf{r}_I$, which results from expression d) and formula

$$\mathbf{G}_{(I)}\mathbf{X}_I = \mathbf{G}\mathbf{X}_I + \mathbf{G}\mathbf{X}_I(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{H}_I = \mathbf{G}\mathbf{X}_I(\mathbf{I} - \mathbf{H}_I)^{-1}.$$

(S2-*m*) Vectors of residuals $\mathbf{r}_{(I)}$ and $\mathbf{r}_I$:

a) $\mathbf{r}_{(I)} - \mathbf{y}_{(I)} - \mathbf{X}_{(I)}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H}_{(I)})\mathbf{y}_{(I)} - \mathbf{H}_{(I)I}\mathbf{y}_I$,

b) $\mathbf{r}_I - \mathbf{y}_I - \mathbf{X}_I'\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H}_I)\mathbf{y}_I - \mathbf{H}_{I(I)}\mathbf{y}_{(I)}$,

c) $\mathrm{E}(\mathbf{r}_{(I)}) = \mathbf{0}$, $\mathrm{D}^2(\mathbf{r}_{(I)}) = \sigma^2(\mathbf{I} - \mathbf{H}_{(I)})$,

$$\mathrm{E}(\mathbf{r}_{(I)}) = (\mathbf{I} - \mathbf{H}_{(II)})\mathrm{E}(\mathbf{y}_{(I)}) - \mathbf{H}_{(I)I}\mathrm{E}(\mathbf{y}_I) = [(\mathbf{I} - \mathbf{H}_{(II)})\mathbf{X}_{(I)} - \mathbf{H}_{(I)I}\mathbf{X}_I']\boldsymbol{\beta} =$$
$$= [\mathbf{H}_{(I)I}\mathbf{X}_I' - \mathbf{H}_{(I)I}\mathbf{X}_I']\boldsymbol{\beta} = \mathbf{0},$$

$$\mathrm{D}^2(\mathbf{r}_{(I)}) = (\mathbf{I} - \mathbf{H}_{(II)})\mathrm{D}^2(\mathbf{y}_{(I)})(\mathbf{I} - \mathbf{H}_{(I)}) + \mathbf{H}_{(I)I}\mathrm{D}^2(\mathbf{y}_I)\mathbf{H}_{(I)I}' =$$
$$= \sigma^2[(\mathbf{I} - \mathbf{H}_{(I)})(\mathbf{I} - \mathbf{H}_{(I)}) + \mathbf{H}_{(I)I}\mathbf{H}_{I(I)}] =$$
$$= \sigma^2[\mathbf{I} - \mathbf{H}_{(I)} - \mathbf{H}_{(I)I}\mathbf{H}_{I(I)} + \mathbf{H}_{(I)I}\mathbf{H}_{I(I)}] = \sigma^2(\mathbf{I} - \mathbf{H}_{(I)}),$$

d) $\mathrm{E}(\mathbf{r}_I) = \mathbf{0}$, $\mathrm{D}^2(\mathbf{r}_I) - = \sigma^2(\mathbf{I} - \mathbf{H}_I)$, as is shown analogically as in c),

e) $\mathbf{r}_{(I)} = [(\mathbf{I} - \mathbf{H}_{(II)}) - \mathbf{H}_{I(I)}(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{H}_{I(I)}]\mathbf{y}_{(I)}$ – vector of residuals expressed by the vector of observations in the *m*-cut model,

f) $\mathbf{Q}_m = \mathbf{r}_I'(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{r}_I$ – "outlier sum of squares" connected with the *m*-dimensional vector of parameters γ in the extended model $\begin{bmatrix} \mathrm{E}(\mathbf{y}_{(I)}) \\ \mathrm{E}(\mathbf{y}_I) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(I)} & \mathbf{0} \\ \mathbf{X}_I & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$ in comparison to model $\begin{bmatrix} \mathrm{E}(\mathbf{y}_{(I)}) \\ \mathrm{E}(\mathbf{y}_I) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(I)} \\ \mathbf{X}_I \end{bmatrix} \boldsymbol{\beta}$ (Gentleman, Wilk 1975; Draper, John 1981),

g) $\mathbf{Q}_m = \mathbf{r}_I'\mathbf{r}_I + \mathbf{r}_{(I)}'\mathbf{H}_{(I)}\mathbf{r}_{(I)}$ – outlier sum of squares split into sum of squares of direct residuals indexed with the vector of observations $\mathbf{y}_I$ of residuals from the *m*-cut model (4).

(S3-*m*) Sum of squares for error and estimation of variance:

a) $SSE_{(I)} = \mathbf{y}_{(I)}'(\mathbf{I} - \mathbf{H}_{(I)})\mathbf{y}_{(I)} = SSE - \mathbf{r}_I'(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{r}_{\mathrm{I}} = SSE - Q_m$, where $\mathbf{H}_{(I)}$ is the matrix of orthogonal projection for the system matrix in model (4), matrix $\mathbf{H}_I$ was given and vector of residuals $\mathbf{r}_I$ is given by formula S2-*m* b), whereas Q_m was given in S2-*m* f),

b) $s^2_{(I)} = \frac{I}{\nu - m}(\nu s^2 - Q_m)$ – estimation of variance for error expressed by the variance of the complete model s^2 corrected by the outlier sum of squares,

c) $F_{(I)} = \frac{\nu - m}{m} \cdot \frac{Q_m}{SSE_{(I)}}$ – the *F* distribution with $\nu - m$ and *m* degrees of freedom at stochastic assumptions of model (4).

In connection with these stochastic considerations for the *m*-model of linear regression, let us supply two more important results (Cook, Weisberg 1980):

a) $\mathbf{r}_I'(\mathbf{I} - \mathbf{H}_I)^{-1}\mathbf{r}_I = \sum_{i \in I} h^2_{ii}$ – generalized Studentized residual for the set of cases *I*, i.e. it is directly expressed by the sum of squares of diagonal elements of matrix **H**, indexed by set *I*,

b) $\mathrm{tr}[\mathbf{H}_I(\mathbf{I} - \mathbf{H}_I)^{-1}]$ – generalized leverage point for set *I* of discriminated cases.

7. CONCLUSIONS

The paper presents theoretical results referring to the complete, 1-cut and *m*-cut models in linear regression. Individual observations or the *m*-system are investigated in the above mentioned cut models. As it was shown in the last two chapters, this investigation may be conducted on units from the complete model. This makes it possible to considerably simplify numerical operations. Various testing statistics are derived from the data determined for the 1-cut and *m*-cut models in order to analyze the occurrence of influential observations. They are used for the purpose of practical identification of influential observations found in the linear regression model. A list of such statistics will be supplied y the authors in another study.

REFERENCES

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., Eisberg S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, New York.

Draper N. R., John J. A. (1981), *Influential Observations and Outliers in Regression*, "Technometrics", 23, 21–26.

Gentleman J. F., Wilk M. B. (1975), *Detecting Outliers II: Supplementing the Direct Analysis of Residuals*, "Biometrics", 31, 387–410.

Gray J. B., Ling R. F. (1984), *K – Clustering as a Detection Tool for Influential Subsets in Regression*, "Technometrics", 26, 305–318.

Haoglin D. C., Welsch R. E. (1978), *The Hat Matrix in Regression and ANOVA*, "American Statistician", 32, 17–22.

La Motte L. (1994), *A Note on the Role of Independence in T Statistics Constructed from Linear Statistics in Regression Model*, Amer. Statist., 48, 238–240.

Anna Budka, Wiesław Wagner

ANALIZA MODELU REGRESJI LINIOWEJ PRZY PODZIELONEJ MACIERZY UKŁADU

W pracy przedstawiono zagadnienia związane z wykrywaniem obserwacji wpływowych w modelu regresji liniowej przy zastosowaniu estymacji parametrów strukturalnych za pomocą MNK. Temat ten jest ujęty w trzech przekrojach: model pełny, 1-ucięty oraz model *m*-ucięty. W każdym przypadku prezentowane są szczegółowe metody badania obserwacji wpływowych. Podstawowymi dla tych celów statystykami są elementy diagonalne tzw. macierzy ortogonalnego rzutu. Ich duże wartości, przy czym wszystkie należą do przedziału (0, 1), przekraczające zadane wartości progowe pozwalają na wskazanie istnienia obserwacji wpływowych. Oczywiście różne możliwe statystyki będące w jakimś stopniu funkcjami elementów wspomnianej macierzy będą dostarczały informacji diagnostycznych o różnym znaczeniu, dotyczącym obserwacji wpływowych.