ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 164, 2003

Wiesław Wagner, Anna Budka***

RESEARCH OF DISTURBANCE IN MODEL OF LINEAR REGRESSION ESTIMATED ACCORDING TO CRITERIA OF LEAST SQUARES

Abstract. In the paper we undertake a study of numerical approach in estimation of degree perturbations of coefficients in model of linear regression according to classical method of least squares (MLS). Degree of perturbation is examined with reference to random variable (Y) and non-random variable (X) by use of certain disturbance constants, respectively a and c.

In relation to these constants values $SSX(c)$, $b_1(c)$, $b_0(c)$, $r(c)$ (see Chapter 4) and also $SSY(a)$, $b_1(a)$, $b_0(a)$, $r(a)$ (see Chapter 3) are given. For these functions there are drawn graphs which determine monotony of changes produced out by perturbations constants c or a, from considered numerical interval. Also measures of the breakdown points are calculated.

Key words: linear regression, method of least squares, disturbance, disturbed dependent variable, disturbed independent variable

1. INTRODUCTION

When investigating the cause and effect relationships involving two variables (X, Y), the usual regression model used to describe that relationship is lineal regression model. It is typically fitted to empirical data by the method of least of squares (MLS) with the classical assumption about normal distribution of random errors and primary and secondary moments remaining constant for them. It means that for those errors their expected values equal zero and have constant variance σ^2, while the parameter $\sigma^2 > 0$ is unknown.

After acquiring an estimate of the linear regression model from MLS, it is diagnosed with various statistical tests, well described in literature (see

* Professor, Department of Statistics, Academy of Physical Education, Poznań.
** Master of Science, Chair of Mathematical and Statistical Methods, Academy of Agriculture, Poznań.

eg. Belsley *et al.* 1980), using also a whole range of available statistical packages (eg. MINITAB). This diagnosis is especially concerned with the problem of influence points.

The diagnostics of the regression model from MLS may be conducted in a slightly different way than the classical one by using the technique of visualizing the model which takes into account the disturbances of selected observations of bivariate sample. Such disturbances are used separately for the X and Y variables. In the first case, the disturbances used make it possible to perceive the effects of influence points, and in the second case – the effect of detached observations occurrences. For that kind of investigations it suffices to use the disturbances for the first (roughly middle) and the last observation of bivariate sample. Their use leads to a certain transformation of the original bivariate sample, as a result of which the configuration of influence points on the correlation diagram also changes. Comparing the numerical characteristics obtained from the fitted regression model from the disturbed bivariate sample, it is possible to answer the question about their influence on the regression model under investigation.

The present paper presents an attempt to investigate the influence of disturbing observations on some numerical characteristics of the fitted linear regression model from MLS through the technique of visualization. The disturbances were assumed separately and arbitrarily for every X and Y variable from a certain numerical range.

2. NUMERICAL CHARACTERISTICS OF THE ESTIMATED LINEAR REGRESSION MODEL WITHOUT DISTURBANCE

Let the set $P_n = \{(x_i, y_i) : i = 1, 2, ..., n\}$ represent a bivariate sample n independent observations of a variable pair (X, Y), where variable X is constants and Y is random variable. It is assumed that the cause and effect relationship $X \rightarrow Y$ is described with the linear regression model

$$Y = \beta_0 + \beta_1 x + e, \tag{1}$$

where β_0 and β_1 are regression coefficients, while e is a random component, about which it is assumed that it has the expected value $E(e) = 0$ and variance $D^2(e) = \sigma^2$, where $\sigma^2 > 0$ is an unknown parameter. The estimation of model (1) from the sample P_n is expressed as

$$\hat{y} = b_0 + b_1 x, \tag{2}$$

where $b_0 = b_0(P_n)$ i $b_1 = b_1(P_n)$ are estimators of parameters β_0 and β_1 which are certain functions of the sample P_n. For the estimation of model (2) we have numerical characteristics (see eg. Draper and Smith (1966), Krysicki *et al.* (1995)):

(a) sum of product deviation

$$SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \tag{3}$$

(b) sum of squares for independent and dependent variables

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad SSY = \sum_{i-1}^{n} (y_i - \bar{y})^2, \tag{4}$$

(c) sum of squares for error

$$SSE = SSY - \frac{(SXY)^2}{SSX}, \tag{5}$$

(d) intercept term and slope coefficient

$$b_0 = \bar{y} - b_1\bar{x}, \quad b_1 = \frac{SXY}{SSX} \tag{6}$$

(e) coefficient of linear correlation

$$r = \frac{SXY}{\sqrt{SSX \cdot SSY}}. \tag{7}$$

3. NUMERICAL CHARACTERISTICS OF THE ESTIMATED REGRESSION MODEL FOR DISTURBED DEPENDENT VARIABLE

The disturbed dependent variable y in model (1) is understood as a transformation of the type $y \rightarrow y + a$, where $a \in \langle a_1^0, a_2^0 \rangle$ and the limits of this interval are defined in such a way that $a_1^0 < a_2^0$. It refers to one bivariate observation, so only one value of the observed dependent variable is disturbed by the constant a. Without losing the generalization it is assumed that observation y_n is thus disturbed, so $y_n(a) = y_n + a$.

Such transformation leads to the following formulas for the numerical characteristics of the estimated linear regression model from MLS and

property related to the disturbance constants a(see eg. B u d k a and W a g n e r 2000):

(a) arithmetic mean of dependent variable

$$\bar{y}(a) = \bar{y} + \frac{a}{n}, \tag{8}$$

(b) sum of product deviation

$$\begin{aligned} SXY(a) &= SXY + a(x_n - \bar{x}) \\ &= SXY, \text{ when } x_n = \bar{x}, \end{aligned} \tag{9}$$

(i) SXY is linear function from a,

(ii) SXY have monotony property in depend of constants a and $x_n - \bar{x}$.

	$x_n - \bar{x} > 0$	$x_n - \bar{x} < 0$
$a > 0$	↗	↘
$a < 0$	↘	↗

(c) intercept term

$$\begin{aligned} b_0(a) &= b_0 + a\left(\frac{1}{n} - \frac{x_n - \bar{x}}{SSX}\bar{x}\right) \\ &= b_0 + \frac{a}{n}, \text{ when } x_n = \bar{x} \end{aligned} \tag{10}$$

where b_0 is given by (6),

(d) slope coefficient

$$\begin{aligned} b_1(a) &= b_1 + \frac{a(x_n - \bar{x})}{SSX} \\ &= b_1, \text{ when } x_n = \bar{x}, \end{aligned} \tag{11}$$

where b_1 is given by (6),

(e) sum of squares for dependent variables

$$SSY(a) = SSY + 2a(y_n - \bar{y}) + \frac{(n-1)a^2}{n} \tag{12}$$

(i) $SSY(a)$ is squared function, which for each a is always plus, $SSY(a) > 0$,

(ii) vertical components $W(a_0, SSY_{min})$ minimum function

$$a_0 = -\frac{n}{n-1}(y_n - \bar{y}), \quad SSY_{min} = SSY - \frac{n}{n-1}(y_n - \bar{y})^2,$$

(f) sum of squares for error

$$SSE(a) = SSY(a) - \frac{(SXY(a))^2}{SSX}, \tag{13}$$

(g) coefficient of linear correlation

$$r(a) = \frac{SXY(a)}{\sqrt{SSX \cdot SSY(a)}}, \tag{14}$$

(h) breakdown point

$$M(a) = |a| \left\{ \left(\frac{1}{n} - \frac{n_n - \bar{x}}{SSX} \bar{x} \right)^2 + \left(\frac{x_n - \bar{x}}{SSX} \right)^2 \right\}^{1/2}.$$

The numerical characteristics of bivariate sample listed here are functions of the disturbance constants a: $\bar{y}(a)$, $SXY(a)$, $b_1(a)$, $b_0(a)$, $M(a)$ – linear, $SSY(a)$, $SSE(a)$ – square, $r(a)$ – rational.

4. NUMERICAL CHARACTERISTICS OF THE ESTIMATED REGRESSION MODEL FOR DISTURBED INDEPENDENT VARIABLE

Now we adapt distributions to n's observation of the bivariate sample for variable X, so that $x_n \rightarrow x_n(c) = x_n + c$, it leads to $\bar{x}(c) = \bar{x} + c/n$. Taking formulas in section 4 it's possible to give formulas for characteristic select statistics in functions dependent on constants c (see eg. B u d k a and W a g n e r 2000):

(a) sum of product deviation

$$SXY(c) = SXY + c(y_n - \bar{y}), \tag{15}$$

(b) sum of squares for independent variables

$$SSX(c) = SSX + 2c(x_n - \bar{x}) + \frac{(n-1)c^2}{n}, \tag{16}$$

(c) intercept term

$$b_0(c) = \bar{y} - b_1(c)\bar{x}(c), \tag{17}$$

(d) slope coefficient

$$b_1(c) = \frac{SXY(c)}{SSX(c)}, \tag{18}$$

(e) coefficient of linear correlation

$$r(c) = \frac{SXY(c)}{\sqrt{SSX(c) \cdot SSY}}, \tag{19}$$

(f) breakdown point

$$M(c) = \left\{[b_1(c)\bar{x}(c) - b_1\bar{x}]^2 + \left[b_1 - \frac{SXY(c)}{SSX(c)}\right]^2\right\}^{1/2} \tag{20}$$

The numerical characteristics of bivariate sample listed here are functions of the disturbance constants c: $SXY(c)$ – linear, $SSX(c)$ – square and for $b_1(c)$, $b_0(c)$, $r(c)$, $M(c)$ – rational.

5. NUMERICAL EXAMPLE

For illustrated effect disturbance in linear regression model, we given data: X – soil quality index and Y – four-crop yield [dt/ha]:

No	x	y	No	x	y	No	x	y	No	x	y
1	1.29	16.1	5	1.47	21.0	9	1.53	20.0	13	1.66	22.1
2	1.34	18.8	6	1.49	18.9	10	1.61	20.1	14	1.66	27.4
3	1.35	16.3	7	1.50	17.5	11	1.61	25.1	15	1.74	25.6
4	1.35	19.6	8	1.51	22.1	12	1.65	22.5	16	1.84	28.7

The estimated linear regression model give numerical characteristics with MLS

(i) $\hat{y} = -10.539 + 20.749x$,
(ii) $\bar{x} = 1.538$, $\bar{y} = 21.363$,
(iii) $SSX = 0.365$, $SSY = 212.158$, $SXY = 7.580$,
(iv) $SSE = 54.893$, $\hat{\sigma} = \sqrt{54.893/14} = 1.98$,
(v) $r = 0.861$, $R^2 = 0.741$ (74.1%),
(vi) plot estimate model and plot index residuals (Figure 1).

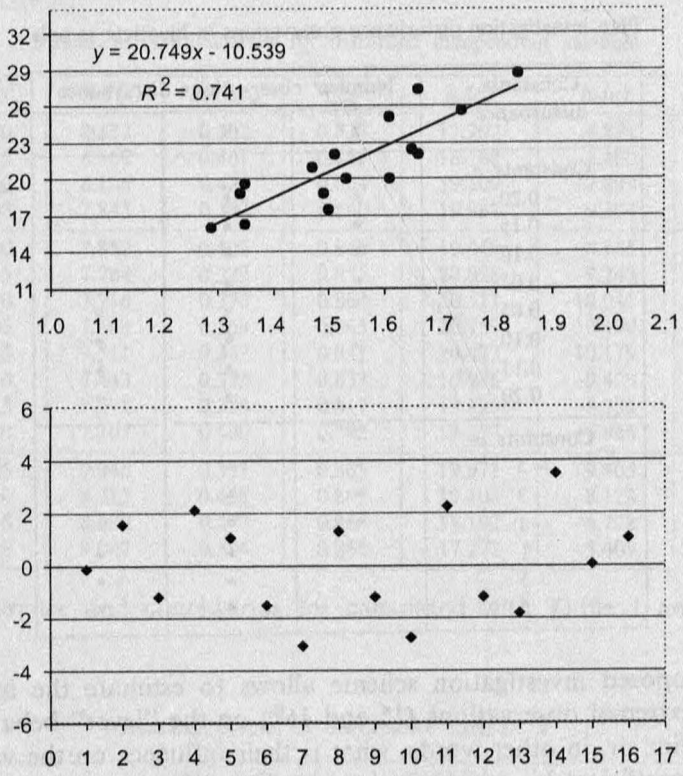

Fig. 1. Scatter plot and index plot residuals

The influence of disturbances on the estimated linear regression model was investigated with the following disturbance constants for the variables:

– independent $c = -0.2, -0.15, -0.10, -0.05, 0.05, 0.10, 0.15, 0.2,$

– dependent $a = -5, -3, -1, 1, 3, 5.$

The 1st, 9th and 16th observations in bivariate sample were disturbed according to Schema 1, where an asterisk * denote case research.

Schema 1

Plan investigation disturbance observations in bivariate sample

Constants disturbance	Number observations disturbance		
	1	9	16
Constants c			
−0.20	*	*	
−0.15	*	*	
−0.10	*	*	
−0.05	*	*	
0.05		*	*
0.10		*	*
0.15		*	*
0.20		*	*
Constants a			
−5	*	*	
−3	*	*	
−1	*	*	
1		*	*
3		*	*
5		*	*

The proposed investigation scheme allows to estimate the influence of disturbed extremal observations (1^{st} and 16^{th}) on the "lever" behavior of the regression line or, in other words, what is their influence on the value of the directional coefficient b_1, which is represented by different angle of the line on the correlation diagram. The 9^{th} observation, which gives $x = 1.53$, almost equal to the mean $x = 1.538$, makes it possible to see how its disturbance changes the trajectory of the bivariate sample gravity center. At the same time it can be seen how an unusual value of the middle observation in a bivariate sample influences the estimation of linear regression model, especially as the data were ordered according to non-decreasing values of the variable X.

Constructing the description of the features of disturbances model in relation to the model without disturbances, the following data were given: a) the number of disturbance observation, b) disturbing constants, c) transformed bivariate observation, d) estimated linear regression, e) coefficient determination R^2 with interval (0, 1), f) scatter plot correlation with two fitting lines; continuous line is for the fitting model for disturbance data; interrupted line for output data, g) intersection of two lines.

All information about point a), ..., g) we present in annex on selection 14 pairs scatter plots and residual index plots for disturbances of variable x and y conformable to the Schema 1.

Numerical results for disturbance variables x are present in Table 1.

Table 1

Numerical characteristics for disturbed independent variable

No	c	$SXY(c)$	$SSX(c)$	$r(c)$	$b_1(c)$	$b_0(c)$	$SSE(c)$
1	-0.20	8.632	0.502	0.837	17.202	-4.871	63.669
	-0.15	8.369	0.461	0.847	18.168	-6.400	60.114
	-0.10	8.106	0.424	0.854	19.109	-7.899	57.261
	-0.05	7.843	0.392	0.860	19.987	-9.304	55.410
9	-0.20	7.852	0.406	0.846	19.349	-8.145	60.226
	-0.15	7.784	0.389	0.857	20.028	-9.243	56.260
	-0.10	7.716	0.376	0.864	20.511	-10.045	53.899
	-0.05	7.648	0.368	0.865	20.759	-10.490	53.398
	0.05	7.511	0.367	0.851	20.473	-10.179	58.378
	0.10	7.443	0.373	0.837	19.946	-9.429	63.696
	0.15	7.375	0.384	0.817	19.199	-8.336	70.563
	0.20	7.307	0.400	0.793	18.277	-6.966	78.610
16	0.05	7.946	0.398	0.865	19.971	-9.405	53.460
	0.10	8.313	0.435	0.865	19.103	-8.128	53.348
	0.15	8.680	0.477	0.863	18.192	-6.778	54.250
	0.20	9.047	0.524	0.858	17.272	-5.409	55.899

The inference and conclusions for calculated with Table 1 are given in Table 2.

Table 2

Research results of disturbance variable x

No	Inferences and conclusions
1	(i) with reduction of constant value c decrease R^2 follows from 0.739 for $c = -0.05$ to 0.700 for $c = -0.2$ (ii) the decrease coefficient of inclination regression line from 19.987 for $c = -0.05$ to 17.202 for $c = -0.2$, (iii) intersection point of both regression lines displace from central concentration of data with increase disturbance constant from 1.598 for $c = -0.2$ to 1.62 for $c = -0.05$, (iv) residues for the first observation changes from 2 for $c = -0.2$ to about 0.5 for $c = -0.05$ while the others residues are subordinated not enough significant changes.
9	(i) change of the constant c from -0.2 to 0.2 causes significant change of residues for the ninth observation, (ii) value of the determination coefficient R^2 changes from $c = -0.2$, 0.748 $c = -0.05$ to 0.629 for $c = 0.20$, (iii) intersection point of the regression lines displaces from higher values for $c < 0$ and for lower values for $c < 0$.
16	(i) for $c = 0.05$, 0.1 and 0.15 we obtained next to similar value R^2 and its significant change takes place only for $c = 0.2$, (ii) displacement of the point x_0 of intersection of both lines take place to higher values, it shows an influence of influential values on estimated linear regression model.

Numerical results for disturbance variables y are present in Table 3.

Table 3

Numerical characteristics for disturbed dependent variable

No	a	$SXY(a)$	$SSY(a)$	$r(a)$	$b_1(a)$	$b_0(a)$	$SSE(a)$
1	−5	8.817	288.220	0.859	24.136	−16.060	75.410
	−3	8.322	252.170	0.867	22.781	−13.851	62.584
	−1	7.827	223.620	0.866	21.426	−11.643	55.917
9	−5	7.617	249.220	0.798	20.851	−11.009	90.395
	−3	7.602	228.770	0.832	20.810	−10.821	70.570
	−1	7.587	215.820	0.854	20.769	−10.633	58.244
	1	7.572	210.370	0.864	20.728	−10.445	53.416
	3	7.557	212.420	0.858	20.687	−10.256	56.088
	5	7.542	221.970	0.838	20.646	−10.068	66.258
16	1	7.882	227.770	0.864	21.577	−11.749	57.702
	3	8.487	264.620	0.863	23.233	−14.171	67.442
	5	9.092	308.970	0.856	24.889	−16.592	82.678

The inference and conclusions calculated with Table 3 are given in Table 4.

Table 4

Research results of disturbance variable y

No	Inferences and conclusions
1	(i) the influence of untypical observation $y = 11.1$ for $a = -5$ causes significant increase in coefficient of inclination regression line, it may be significant in its essential interpretation, (ii) intersection points of both lines are next to similar.
9	(i) coefficient of inclination regression line is nearly without any changes and takes values from interval 20.646 for $a = 5$ to 20.851 for $a = -5$, (ii) both lines are nearly parallel and their intersection point are significantly outlying from gravity point two-dimensional trial, (iii) when only $a = -5$, it can be shown disturbance observation (1. 53, 15) is untypical, (iv) with increase of the constant a change of type of value residues follows for the ninth observation with high negative (minus) values to high positive (plus) values.
16	(i) with increase of constant a values R^2 decrease insignificantly and value of estimated coefficient direction of regression line increase significantly, (ii) intersection point of both regression lines is next to constant and amounts to near 1.46, (iii) residue for the sixteenth observation increases with increase constant a.

For research of divergence between estimated model from data without disturbance and models obtained after disturbance of data, it used measure breakdown point, which is expressed by formulas (15) and (20). Table 5 shows appropriate calculations breakdown points.

With calculation from Table 5 we have two implications:

$$[(c<0) \wedge (c \to -\infty)] \vee [(c>0) \wedge (c \to \infty)] \Rightarrow M(c) \to \infty$$

$$[(a<0) \wedge (a \to -\infty)] \vee [(a>0) \wedge (a \to \infty)] \Rightarrow M(a) \to \infty$$

Table 5

Breakdown point (b.p.) for disturbed variables

c	1	9	16	a	1	9	16
–0.20	6.686	2.772		–5	1.296	0.481	
–0.15	4.877	1.482		–3	3.886	0.289	
–0.10	3.107	0.548		–1	6.477	0.096	
–0.05	1.450	0.050		1		0.096	1.467
0.05		0.454	1.374	3		0.289	4.400
0.10		1.370	2.919	5		0.481	7.334
0.15		2.693	4.547				
0.20		4.344	6.197				

REFERENCES

Belsley D. A., Kuh E., Welsch R. E., (1980), *Regression Diagnostics*. Wiley, New York.

Budka A., Wagner W., (2000), *Badanie wpływu zaburzeń obserwacji próby dwuwymiarowej na szacowanie modelu regresji liniowej według kryterium sumy i mediany kwadratów reszt*, „Wyzwania i Dylematy Statystyki XXI wieku", 17–41.

Draper N. R., Smith H., (1966), *Applied Regression Analysis*, Wiley, New York (tłum. pol. Analiza regresji stosowana, Warszawa 1973).

Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M. (1995), *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach*, cz. II, *Statystyka matematyczna*, PWN, Warszawa.

Wiesław Wagner, Anna Budka

ANALIZA ZAKŁÓCEŃ W MODELU REGRESJI LINIOWEJ W PORÓWNANIU DO KLASYCZNEJ METODY NAJMNIEJSZYCH KWADRATÓW

(Streszczenie)

W pracy przedstawiono wyniki analiz numerycznych związanych z estymacją stopnia zakłócenia współczynników występujących w równaniu prostej regresji.