Identification of Influential Points in a Linear Regression Model

Jan Grosz[a] | *Czech University of Life Sciences in Prague*

Abstract

The article deals with the detection and identification of influential points in the linear regression model. Three methods of detection of outliers and leverage points are described. These procedures can also be used for one-sample (independent) datasets. This paper briefly describes theoretical aspects of several robust methods as well. Robust statistics is a powerful tool to increase the reliability and accuracy of statistical modelling and data analysis. A simulation model of the simple linear regression is presented.

Keywords

leverage points, outliers, method of least squares – LSM, trimmed mean, confidence interval, regression model, data simulation, robust estimators

INTRODUCTION

Regression analysis, along with variance analysis, belongs to such mathematic-statistical methods, which can find a broadest usage in practical applications of various sciences. The main goal of regression analysis is finding of a real function f, which describes the relation of the dependent variable Y and a group of independent variables $X_1, X_2, \dots, X_m$. This function is called the regression function and shall comply with the relation as follows:

$$Y = f\,(X_1, X_2, \dots, X_m) + \varepsilon,$$

where ε is the random variable representing random deviations (errors) of the model.

Let us further limit to the linear class of functions, that is to deal with the model as follows:

$$Y = \beta_1 + \beta_2 \; X_2 + \dots + \beta_m \; X_m + \varepsilon \, .$$

Parameters β_i are called linear regression coefficients and this paper is devoted to their estimators.

It is known that estimators of regression coefficients by means of the classical method of least squares are very sensitive to extreme points that means to the points, which are "standing out of the line" in a certain way. In practice, such data "is created" most frequently by an error when data is entered into the computer, or by potentially erratic filling in of the original source data. Therefore, it is of great importance to identify such points and eliminate them from the dataset because their presence – and there may be the only one such point – would substantially distort or even completely deteriorate the resulting values of regression analysis parameters. Such values are referred to as influential points (observances) and for the sake of simplicity are classified as:

- extreme points, called outliers (type E), occurring at the dependent variable, see Figure 3; and

- outlying leverage points (type V) occurring at the independent variables, see Figure 2.

[a] *Czech University of Life Sciences in Prague, Kamýcká 129, 165 21 Praha 6-Suchdol, e-mail: grosz@pef.czu.cz*

1 GENERIC MODEL OF LINEAR REGRESSION

Let us take the classical model of linear regression:

$$y_i = \sum_{j=1}^{m} x_{ij}\beta_j + \varepsilon_i, \quad i=1,\dots, n, \quad n > m, \qquad (1)$$

where x_{ij} are given values of ith repetition of jth explanatory (independent) variable, εi are independent, random variables of normal distribution with zero mean value and variance σ^2 (so-called "white noise"), β_j means unknown regression coefficients, and y_i is a value of the regressand (or the dependent variable) at ith observation.

The matrix record is in the form

$$Y = X\beta + \varepsilon \qquad (2)$$

where $X = (x_{ij})$ is a matrix of order nxm and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$, $\beta = (\beta_1, \dots, \beta_m)'$, and $Y = (y_1, \dots, y_n)'$ are column vectors.

Therefore, Y is a random vector, which has normal distribution with the mean value (vector) of $X\beta$ and the variance-covariance matrix $\sigma^2. I_n$, where I_n is a unit matrix of order n.

The basic goal of regression analysis is to estimate the vector β by minimising the sum of squares of observed points deviations from the regression line. In mathematic language it is finding of the minimum of the quadratic form of $S(\beta) = (Y - X\beta)'(Y - X\beta)$.

Therefore we seek:

$$\min\,(Y - X\beta)'(Y - X\beta). \qquad (3)$$

Let us say $\hat{\beta}$ is any solution of a linear equations system:

$$X'X\,\beta = X'Y. \qquad (4)$$

The system of (so-called normal) equations (4), which is yielded when solving the task (3) has always one solution, at least, because $L(X) = L(X'X)$. Here $L(X)$ refers to the linear envelope formed of columns of the matrix X – see [7], for instance.

In general, the linear envelope of a finite set of elements (vectors) of a vector space is defined as a set of all linear combinations of these vectors.

It holds:

$$(Y - X\beta)'(Y - X\beta) \geq (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

In other words the quadratic form $S(\beta)$ takes its minimum in the point $\beta = \hat{\beta}$.

Here $S(\hat{\beta})$ represents the residual sum of squares of deviations observed from fitted values.

It is easy to show that the following relations are valid:

$$S(\hat{\beta}) = Y'Y - Y'X\hat{\beta} \qquad (5)$$

$$\begin{aligned} &E(S(\hat{\beta})) = (n - h(X)).\sigma^2 \text{ and} \\ &D(Y - X\hat{\beta}) = D(Y) - D(X\hat{\beta}) \end{aligned} \qquad (6)$$

The quantity $S(\hat{\beta})/(n - h(X))$ is therefore an unbiased estimator of the parameter σ^2. The symbol of $h(X)$ means the rank of the matrix X, $E(X)$ is a mean value of the random variable X, and $D(Y)$ is a variance-covariance matrix of the vector Y. Proof can be found in the publication [7] as well. Rather detailed publications dealing with matrix algebra are [6] and [7]. The next section mostly deals with the case m = 2 – that is the most frequently occurring issue of simple linear regression in practice.

2 IDENTIFICATION OF EXTREME POINTS

There are numerous methods, which can identify extreme points. Procedures given here are good to interpret and appropriate characteristics can be easily calculated within the environment of the spreadsheet software Excel – therefore they do not require any special statistical software. In author's experience they are highly effective and sensitive in discovering extreme points of input datasets.

2.1 Identification of leverage points

Let us assume hereinafter that the model (1) is a full rank model, that is $h(X) = m$ is valid.

In this case the solution of the system (4) is determined unambiguously and has the form:

$$\hat{\beta} = (X'X)^{-1}X'Y. \qquad (7)$$

It holds that $E\,\hat{\beta} = \beta$, and so the estimator is unbiased and has the least variance among such estimators. In such case we can call it the best linear unbiased estimator of the vector β.

Now, let us mark $\hat{Y} = X\hat{\beta}$ the "predicted" vector Y. If $\hat{\beta}$ in the equation is replaced with the expression (7) the yield is:

$$X\hat{\beta} = X(X'X)^{-1}X'Y = WY, \quad W = X(X'X)^{-1}X' \tag{8}$$

The matrix W is a square matrix of rank n having properties as follows:

(i) $W' = W$ (symmetry)

(ii) $W_2 = W$ (idempotency)

(iii) $W'X = X$

(iv) W is a hat matrix to $L(X)$

(v) $0 \leq w_{ii} \leq 1, i = 1, \ldots, n$

(vi) $\sum_{i=1}^{n} w_{ii} = \mathrm{m}$

(vii) Let us mark $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)$ *a* $\hat{\boldsymbol{e}} = Y - \hat{Y} = (\hat{e}_1, \ldots, \hat{e}_n)$ – the vector of residuals. Then $var\ (\hat{y}_i) = w_{ii}\sigma^2$ and $var\ (\hat{e}_i) = (1 - w_{ii})\sigma^2$

(viii) $\hat{y}_i = y_i w_{ii} + \sum_{j \neq i} w_{ij} y_j$.

(ix) Diagonal elements of w_{ii} of the hat matrix W represent – roughly – the distance of ith observation from the middle of other points concerning explanatory variables.

(x) Such point x_i can be considered an extreme point, for which:

$$w_{ii} > \frac{2m}{n}, i = 1, \ldots, n. \tag{9}$$

The procedure as follows can be used to explain this boundary. Let us assume that row vectors of the matrix X form multivariate normal distribution. Then, testing the hypothesis that all rows have the mean value constant, the testing statistics

$$F = \frac{n-m}{m-1}\frac{w_{ii} - \frac{1}{n}}{1 - w_{ii}}$$ has Fischer's distribution

with $m - 1$ and $n - m$ degrees of freedom. If critical value of this statistics is roughly equal to 2, then $F > 2$ (which is the critical region to reject the hypothesis) when the relation (9) is approximately valid. Details can be found in [8] or [9].

2.2 Identification of extreme values – outliers

First, let us introduce the term of so-called trimmed mean α $(0 < \alpha < 0.25)$. This is an arithmetic average, which remains after $100{*}\alpha$ % of the smallest and largest values are eliminated.

That means more precisely:
let us mark $y_{(1)} \leq y_{(2)} \leq \ldots\ldots \leq y_{(n)}$ ordered original values of the dependent variable, $n_1 = \{\alpha{*}n\}$, $n_2 = n - n_1$, (symbol $\{x\}$ means the nearest natural number higher or equal to x). Thus in total n_1 of the least and largest values are eliminated and the sample then contains $k = n_2 - n_1$ values. Then let us define α – the trimmed mean and α – trimmed variance as follows:

$$\bar{y}_{\alpha} = \frac{1}{k}\sum_{i=n_{1+1}}^{n_2} y_{(i)} \tag{10}$$

$$\sigma_{\alpha}^{2} = \frac{1}{(k-1)}\sum_{i=n_{1+1}}^{n_2} (y_{(i)} - \bar{y}_{\alpha})^2 \cdot \tag{11}$$

In practice it is selected to be $0.05 \leq \alpha \leq 0.1$, ie. ca 10% – 20% of sample values are eliminated and the aforementioned characteristics of the mean value and variance are calculated using the rest of the sample values.

Further procedure is based on the modification to the known three sigma rule, which holds for normal distribution. Let us choose the confidence interval

$$(\bar{y}_{\alpha} - 3\sigma_{\alpha}, \bar{y}_{\alpha} + 3\sigma_{\alpha})$$

and detect such points y_i, lying outside this interval, i.e. such y_i, which meet $| y_i - \bar{y}_\alpha| > 3\sigma_\alpha$, $i=1, \ldots, n$.

This way identified points can be considered extreme ones. These values need to be subject to further assessment. Verification, if these are erratic data (which is quite common case in data entering), or these are really extreme values, has to be carried out. In the first case the points are corrected, of course, in the second case such values may be either eliminated from the dataset and then to calculate the vector β estimator using the common method of least squares; or we can chose some other method. This provides a certain guarantee that the dataset got cleaned of "suspicious values".

2.3 Identification of influential points

Ali S Hadi (1992) proposed the following (additive) statistics, which tests influential points in the model of linear regression as follows:

$$H_i = \frac{w_{ii}}{1 - w_{ii}} + \frac{m}{1 - w_{ii}} \frac{d_i^2}{1 - d_i^2}, \text{ where} \tag{12}$$

$$d_i = \hat{e}_i / \sqrt{S(\hat{\beta})}$$

is so-called normalized residual. $i = 1, 2, \ldots, n$.

The first summand in (12) represents a portion of influence of the explanatory variable, the second addend then represents influence of the dependent variable. The test therefore consists in the fact an influential point is either of E or V type, respectively. High values of H_i prove that ith observance represents an influential point while there is no exact limit determined in this case. Recommendation is to set preliminary critical value to 1.

3 DATA SIMULATIONS – ILLUSTRATIVE EXAMPLES

For the purpose of quality verification of the aforementioned methods a simulation experiment was carried out by means of a random number generator for the model of simple linear regression with parameters as follows:

$n = 30$ (sample size), $m = 2$ (number of parameters), $\beta = (3,7)'$, $\sigma^2 = 4$

Therefore the model (1) has the shape:

$$y_i = 3x_i + 7 + \varepsilon_i\,, \;\; i = 1, \ldots, 30. \tag{13}$$

The explanatory variable xi was generated from uniform distribution $R(20, 30)$ and ε_i has normal distribution with zero mean value and standard deviation 2.

Table 1 Generated data (13)

I	y_i	x_i	i	y_i	x_i
1	77.03	23.20	16	81.18	25.03
2	92.16	28.48	17	74.86	22.85
3	86.23	26.00	18	74.17	22.29
4	84.13	25.26	19	82.46	24.15
5	78.98	22.86	20	95.61	29.21
6	89.43	27.20	21	70.16	22.33
7	84.13	26.34	22	70.41	20.83
8	71.25	21.41	23	75.99	23.28
9	70.44	21.58	24	83.18	25.22
10	89.07	26.78	25	67.65	20.35
11	78.88	24.29	26	89.90	28.08
12	71.55	21.17	27	77.91	23.84
13	88.01	25.98	28	69.66	20.45
14	75.70	22.52	29	77.06	24.11
15	90.54	27.71	30	95.50	29.63

Source: own research

The estimates of the regression coefficient parameter β and standard deviation σ of the model (13) obtained by the method of least squares were $\hat{\beta} = (3.017; 6.769)'$ and $\hat{\sigma} = 2.6$.

Data (13) was subsequently "contaminated" with influential values of E and V types this way:

V: $x_1' = 40$ and $x_3' = 4$ (original values were $x_1 = 23.2$ and $x_3 = 26$).

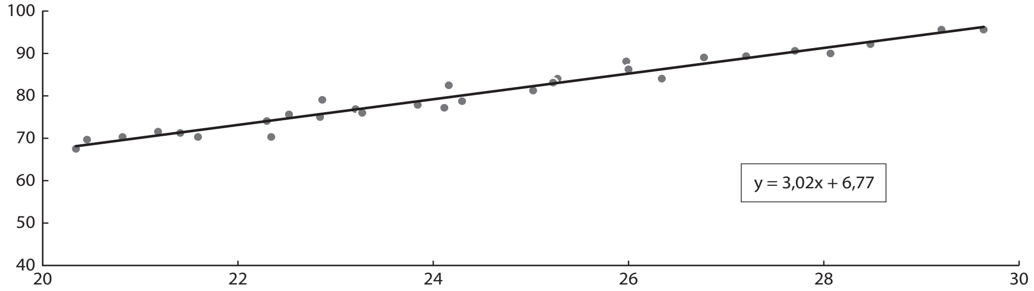
Diagonal elements of the matrix W are used for the detection of extreme values, as derived above; if

$$w_{ii} > \frac{2m}{n},$$

then such a point may be considered extreme value. In our case this critical value is 0.13, while $w_{11} = 0.31$ and $w_{33} = 0.51$, and for the other $w_{ii} < 0.07$, so x_1', x_3' can be considered extreme values.

In the calculation of the hat matrix W in the Excel environment functions are used as follows:

Figure 1 Linear regression of a dataset (13)



Source: own research

TRANSPOSITION *(A)* – carries out transposition of the given matrix *A´*;

MATRIX.PRODUCT*(A; B)* – result is a product of matrixes AB; and

INVERSION *(A)* – calculates the inversion matrix A^{-1} (if there is any).

The inversion matrix calculation is of sufficient accuracy; nevertheless this function has its limitations (especially for matrixes of higher orders, for instance with n > 50).

Table 2 Values of diagonal elements of the hat matrix *W*

I	w_{ii}	i	w_{ii}
1	0.321	16	0.034
2	0.054	17	0.036
3	0.508	18	0.038
4	0.035	19	0.033
5	0.036	20	0.062
6	0.044	21	0.038
7	0.038	22	0.047
8	0.043	23	0.034
9	0.042	24	0.034
10	0.041	25	0.051
11	0.033	26	0.050
12	0.044	27	0.034
13	0.037	28	0.050
14	0.037	29	0.033
15	0.047	30	0.067

Source: own research

Further two values were replaced with extreme points this way:

E: $y_1' = 50$ and $y_2' = 140$
(original values were $y_1 = 77.02$ and $y_2 = 92.2$).

Subsequently, α – trimmed mean
and variance for $\alpha = 0.05$ were calculated:

$n_1 = \{30*0.05\} = 2,$
$n_2 = 30 - 2 = 28,$
$k = 26,$
$\bar{y}_\alpha = 80.03, \sqrt{\sigma_\alpha^{\,2}} = 7.39 .$

There were 4 points eliminated and the confidence interval was calculated in the form:

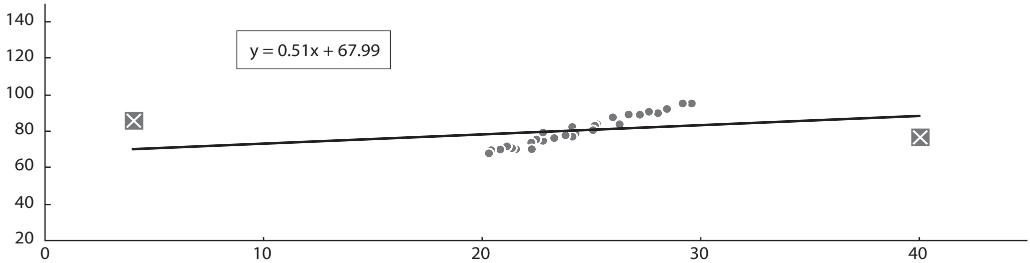
$$(57.8;\ 102.2) \qquad (14)$$

There are solely points y_1' and y_2' out of the interval (14) (see Figure 3 below).

What can also be seen in Figures 2 and 3 is the presence of outliers lead to significantly worse results than the presence of leverage points. Both the methods described can be employed for the detection of extreme (erratic) values for one-sample datasets.

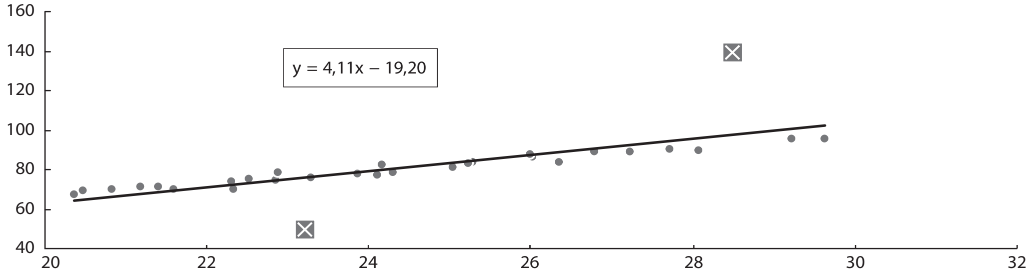
In order to verify the Hadi measure the original dataset was “contaminated” with extreme values of E and V types simultaneously: $x_1' = 40$ and $x_3' = 4$ and $y_1' = 50$ and $y_2' = 140$ (the way the first three pairs of the original values were replaced (13)). It is clear in Figure below how this modification deteriorated “proper” parameters of the regression function.

Figure 2 Linear regression of a dataset containing two extreme points



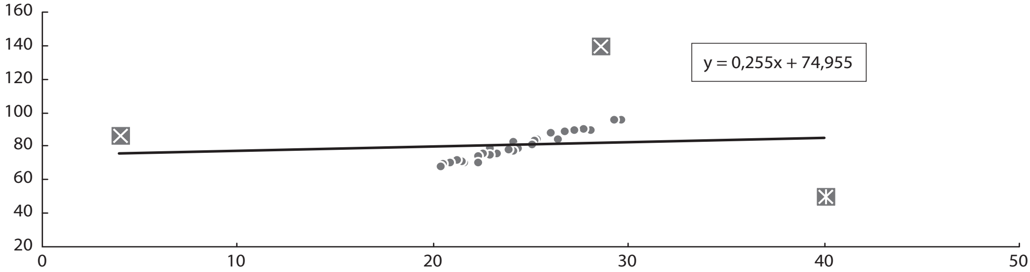
Source: own research

Figure 3 Linear regression of a dataset including two influential points



Source: own research

Figure 4 Linear regression of a dataset containing three influential points



Source: own research

It can be seen from Table 3 that the Hadi measure attains the highest value for H_3 and H_1, and further then for H_2, which indicates the presence of influential points. Other H_i are by an order of magnitude lower.

4 ROBUST METHODS

Methods, which reduce sensitivity to extreme values and simultaneously give high quality regression coefficient estimators, are called robust methods. A whole number of such methods were proposed due to fast progress in computer technology. These methods are theoretically described in a very detailed manner in the today already classical monograph [2], and newly can also be found in [5].

The most often applied procedure in the estimating of regression coefficients is M-estimators (maximum likelihood). It is such an estimator of $\hat{\beta}$, which minimises the sum of residuals using a suitable way chosen function ρ, which is convex, and there is a derivation ρ´. That means it is a certain generalisation of the method of least squares where

Table 3
The Hadi measure values

I	H_i	i	H_i
1	2.837	16	0.035
2	0.314	17	0.038
3	4.072	18	0.041
4	0.036	19	0.037
5	0.042	20	0.067
6	0.046	21	0.039
7	0.041	22	0.053
8	0.047	23	0.036
9	0.044	24	0.036
10	0.043	25	0.057
11	0.035	26	0.054
12	0.050	27	0.035
13	0.039	28	0.058
14	0.040	29	0.035
15	0.050	30	0.074

Source: own research

$\rho(x) = x^2$. The M-estimator therefore depends on the selection of the function ρ. Its drawback is the M-estimator eliminates solely effects of outliers and not those of leverage points.

Other applied estimator is the LTS-estimator (least trimmed squares estimator). This estimator is calculated by omitting a certain number of the smallest and largest residuals (similar way as in 3.2).

The LMS-estimators or LMedS-estimators (least median of squares estimators) are based on the idea of minimising the median of squared residuals. Generalisation of LMS and LTS estimators give birth to the S-estimator.

Statistical software SAS ver. 9.2 has, in its routine ROBUSTREG, four methods of estimators, including testing for the presence of outliers and leverage points: M-, LTS-, S-, and MM-estimators. Yet the identification of outliers is based on other methods than those mentioned here above.

CONCLUSION

In real applications one can often face the issue of identification and detection of extreme (and/or leverage) points, which are such points that in principal manner affect the dataset analysis. Such points are classified as outliers of values of the dependent variable, leverage points of the independent variable, or influential points of both the variables. It is right the presence of such points that results in often completely worthless regression parameters estimators using the method of least squares. Therefore the type of the analysed data contamination must be identified first. Three methods were chosen out of a number of existing methods as follows: detection by means of a projection matrix, "robust" confidence interval, and the Hadi measure. In author's experience these methods have worked very well in practice, namely in accuracy checking of PC entered data.

Some of the methods of the regression coefficient calculation by means of so-called robust methods are briefly described in section 5. These methods are implemented in the SAS system.

Remark: All necessary numeric calculations were carried out in the spreadsheet software EXCEL.

References

[1] BARNETT V., LEWIS T. *Outliers in Statistical data.* Wiley: New York, 1994.

[2] HUBER, P. J. *Robust Statistics.* John Wiley: New York, 1981.

[3] CHAJDIAK, J. *Štatistika v Exceli.* Statis: Bratislava, 2002.

[4] CHATTERJEE, S., HADI, A.S. *Regression analysis by example.* Wiley-Interscience: Hoboken, 2006.

[5] MARONNA R., MARTIN D.,YOHAI V. *Robust Statistics: Theory and Methods.* Wiley, 2006.

[6] NERING, D.E. *Linear algebra and matrix Tudory.* Wiley: New York, 1970.

[7] RAO, C. R. *Linear Statistical Inference and Its Applications.* Wiley: New York, 1965.

[8] RYAN, T.P. *Modern regression methods.* Wiley: Hoboken, 2009.

[9] ZVÁRA, K. *Regresní analýza.* Academia: Praha, 1989.