Detection of Outliers in Regression Analysis by Information Criteria

Seppo Pynnönen, Department of Mathematics and Statistics, University of Vaasa, BOX 700, 65101 Vaasa, FINLAND,
e-mail sjp@uwasa.fi,
home page: www.uwasa.fi/∼sjp/

Current version: September 1992

Contents

Abstract

Pynnönen, S. (1992). Detection of outliers in regression analysis by information criteria. Proceedings of the University of Vaasa. Discussion Papers 146, 8 p.

In this paper detection of outliers in the usual linear regression model by Akaike's and Schwarz's information criteria is considered. In terms of these criteria the problem can be considered as estimating the number of (dummy) variables in the model. By this method one does not have to concern the definition of underlying distribution of the observed residuals, $\hat{\epsilon}_j$, which in practice has proved to be very complicated (see e.g. Barnett and Lewis 1984, ch. 10).

The method is illustrated by analyzing some well known data sets.

Seppo Pynnönen, School of Business Studies, University of Vaasa, P.O. Box 297, SF-65101 Vaasa, Finland.

1. Introduction

Akaike (1977), Schwarz (1978), and Rissanen (1978) suggested general purpose model selection criteria of the form

$$-\log L_m + g(n,m), \tag{1.1}$$

where $\log L_m$ denotes the maximized log-likelihood function with m estimated parameters, and $g(n,m)$ is a penalty function depending on the sample size, n, and number of estimated parameters, m. In Akaike's criterion (AIC) the penalty function is of the form $g(n,m) = m$, and in Schwarz's and Rissanen's criterion (BIC) it is of the form $g(n,m) = \frac{1}{2}m\log n$. That model is selected, which minimizes (1.1). Hence, we observe that BIC penalizes more the included parameter whenever $n \geq 8$.

Kitagawa and Akaike (1982) and Kitagawa (1984) have applied AIC in detection of outliers by using (quasi) Bayesian approach with predictive likelihood (see e.g. Akaike 1980) in place of the usual likelihood function (see Kitagawa and Akaike 1982, p. 398, and Kitagawa 1984, p. 121). Otherwise, detection of outliers has a long history. The main theme, however, has been around univariate and single outliers. Recently some promising results have obtained in detecting multiple outliers also in multivariate analysis (see, eg Hadi 1992). Paul and Fung (1991) used generalized extreme studentized residuals (GESR), which has improved earlier procedures, like Marasinghe's (1985) procedure. These procedures are based on sequential testing approach. Nevertheless, as mentioned earlier, the main distributional properties of the test statistic remain unknown, which implies that the exact critical levels remain unknown. This problem poses no difficulties in a criterion function approach adopted in this paper.

One more advantage of the criterion function approach is that it provides an easier way for modelling the potential outlier populations. For example GESR quickly fails when there are several outliers from the same populations (see Hadi and Simonoff 1992).

2. The Model

Consider the following linear regression model

$$y = \gamma_k + X\beta + \epsilon, \tag{2.1}$$

where y is a random n-vector, γ_k is an unknown parameter vector with k nonzero coordinates ($k < n-p$), X is an $n \times p$ design matrix of full rank, β is a $p \times 1$ unknown parameter vector, and ϵ is an $N(\mathbf{0}, \sigma^2 I_n)$ residual vector.

Let $\epsilon^* = (\epsilon_1^*, \ldots, \epsilon_n^*)' = y - X\beta$, so that

$$\epsilon^* = \begin{cases} \epsilon_i, & \text{if } \gamma_i = 0 \\ \epsilon_i + \gamma_i, & \text{if } \gamma_i \neq 0, \end{cases} \tag{2.2}$$

and $\epsilon^* \sim N(\gamma_k, \sigma^2 I_n)$.

In order to use criterion (1.1) to specify the outlier model given in (2.1) and (2.2), we have to define its likelihood function. For the purpose, suppose we can order the residuals, ϵ_i^*, such that

$$\epsilon_{(1)}^* \leq \epsilon_{(2)}^* \leq \cdots \leq \epsilon_{(n)}^*, \tag{2.3}$$

and denote the corresponding order statistic as

$$U = (\epsilon_{(1)}^*, \ldots, \epsilon_{(n)}^*)'. \tag{2.4}$$

Then using the results for order statistics (see e.g. David 1981), the joint density of U becomes

$$f_U(u_1, \ldots, u_n) = \begin{cases} \sum^* \prod_{i=1}^n \phi_{j_i}(u_i), & \text{if } u_1 < \cdots < u_n \\ 0, & \text{otherwise,} \end{cases} \tag{2.5}$$

where $\sum^*$ is sum over all permutations $(j_1, \ldots, j_n)$ of $(1, \ldots, n)$, and $\phi_{j_i}(\cdot)$ is the density of the normal distribution $N(\gamma_{j_i}, \sigma^2)$.

Suppose next that there are k outliers that are clustered into $m < n-p$ clusters such that the jth cluster includes k_j observations, $j = 1, \ldots, m$, $k_1 + \cdots + k_m = k < n-p$. Hence, we actually have $m+1$ populations one of which is the main, non-outlier, population. In each of the outlier populations $\gamma_i \neq 0$ and in the main body of data $\gamma_i = 0$. Henceforth we assume that the m_1 ($1 \leq m_1 \leq m$) first populations are lower outliers and the rest are upper outliers, i.e. $\gamma_1 < \cdots < \gamma_{m_1} < 0 < \gamma_{m_1+1} < \cdots < \gamma_m$. Hence, the main data is between with $n-k$ observations.

From (2.5) we then get straightforwardly

$$\begin{aligned} f_U(u_1, \ldots, u_n) &= k_1! \cdots k_m! (n-k)! \\ &\quad \times \sum{}^+ \left\{ \prod_{i=1}^{k_1} \phi_{j_i}(u_{j_i}) \prod_{i=k_1+1}^{k_1+k_2} \phi_{j_i}(u_i) \cdots \prod_{i=n-k_m+1}^{n} \phi_{j_i}(u_{j_i}) \right\}, \end{aligned} \tag{2.6}$$

if $u_1 < u_2 < \cdots < u_n$

Where the summation, $\sum^+$, extends over all distinguishable permutations of n individuals, which are grouped into $m+1$ groups. The number of such permutations is $n!/[k_1! \cdots k_m!(n-k)!]$.

Generally model (2.6) is too complicated for practical use. Hence, we have to simplify the model with suitable assumptions. We assume that the populations, $N(\gamma_i, \sigma^2)$, are well separated from each other, such that $|\gamma_j - \gamma_l|/\sigma$, $j \neq l$ are sufficiently large. Then $\phi_j(u) \approx 0$, if the observation comes from some other population than $N(\gamma_j, \sigma^2)$. Taking this into account simplifies (2.6) to

$$f_U(u_1, \ldots, u_n) = k_1! \ldots k_m!(n-k)! \prod_{i=1}^{k_1} \phi_1(u_i) \cdots \prod_{i=n-k_m+1}^{n} \phi_m(u_i), \quad \text{if } u_1 < \cdots < u_n \tag{2.7}$$

Letting $\epsilon^*_{(i)} = u_i$, the log-likelihood becomes then from (2.7)

$$\begin{aligned} -\frac{n}{2}\log(2\pi\sigma^2) \;-\; & \frac{1}{2\sigma^2}\Big\{\sum_{i=1}^{k_1}(\epsilon^*_{(i)} - \gamma_i)^2 + \cdots \\ & + \sum_{i=k_1+\cdots+k_{m_1}+1}^{n-k} \epsilon^{*2}_{(i)} + \cdots + \sum_{i=n-k_m+1}^{n}(\epsilon^*_{(i)} - \gamma_i)^2\Big\} \\ & + \log k_1! + \cdots + \log k_m! + \log(n-k)!. \end{aligned} \tag{2.8}$$

Recalling that $\epsilon^* = y - X\beta$, we can write (2.8) in the form

$$\begin{aligned} -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - \gamma_k - X\beta)'(y - \gamma_k - X\beta) \\ + \log k_1! + \cdots + \log k_m! + \log(n-k)!. \end{aligned} \tag{2.9}$$

Replacing the unknown parameters by the maximum likelihood estimators, the maximum of (2.9) becomes, dropping some obvious constants, as

$$-\frac{n}{2}\log\hat{\sigma}^2_k + \log k_1! + \cdots + \log k_m! + \log(n-k)!. \tag{2.10}$$

where $\hat{\sigma}^2_k = (y - \hat{\gamma}_k - X\hat{\beta})'(y - \hat{\gamma}_k - X\hat{\beta})/n$ is the maximum likelihood estimator of σ^2, and $\hat{\gamma}_k$ and $\hat{\beta}$ are the maximum likelihood estimators of γ_k and β, respectively.

In the special case, where each outlier emerges from different population, i.e., $k_j = 1, j = 1, \ldots, m = k$, then (2.11) simplifies to

$$-\frac{n}{2}\log\hat{\sigma}^2_k + \log(n-k)!. \tag{2.11}$$

Hence, finally, we can write the formula for the criterion (1.1), call it C, as

$$C(k) = \frac{1}{2}n\log\hat{\sigma}^2_k - \log k_1 - \cdots - \log k_m - \log(n-k)! + g(n, m+p). \tag{2.12}$$

3. Examples

As illustrations we consider here three numerical data sets. The first one, given in Table 3.1, is from Barnett (1983).

TABLE 3.1. Barnett's (1983) data.

No. of days (x)	4	5	7	9	11	14	17	20	23	26	30	35
Measurement (z)	110	81	90	74	20	30	37	22	38	25	18	9
$y = \log z$	4.7	4.4	4.5	4.3	3.0	3.4	3.6	3.1	3.6	3.2	2.9	2.2

Next we fit the model

$$\mathrm{E}\, y = \beta_0 + \beta x \tag{3.1}$$

for different number of potential outliers. The best fitting models in terms of AIC and BIC are given in Table 3.2. We find that both methods suggest the observation with x-value equal to 11 (i.e., observation number five) being an outlier, for both of the criteria assume minima when this observation is eliminated. If one plots a scatter diagram of the data set, one sees that this observation [i.e., observation $(11, 3.0)$] is clearly out of pattern, too.

TABLE 3.2. AIC and BIC values for model (3.1).

Number of outliers ($k = m$)	Observation number	R^2	AIC*)	BIC*)
0	-	0.7379	-56.04	-56.04
1	11	0.8974	**-60.33**	**-59.84**
2	11, 14	0.9338	-58.41	-57.82
3	11, 14, 20	0.9606	-58.41	-56.96

*) $\mathrm{AIC} = n \log(1 - R^2) - 2 \log(n - k)! + 2m$
$\mathrm{BIC} = n \log(1 - R^2) - 2 \log(n - k)! + m \log n$

Data for our second example is from Guttman *et al.* (1978).

TABLE 3.3 Guttman's *et al.* (1978) data. For each (x_1, x_2) pair there are two observations on y ($n = 20$).

x_1	x_2	y	
1	0	7.2	6.9
-0.5	0.866	9.3	9.6
-0.5	-0.866	10.4	9.8
0	0	12.3	11.7
0	0	12.2	12.1
-1	0	7.7	7.8
0.5	0.866	6.2	5.8
0.5	-0.866	11.3	11.6
0	0	11.8	12.4
0	0	12.7	12.0

The model considered here is the following second order regression equation

$$E\,y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2. \tag{3.2}$$

In Table 3.4. we have reported R-squares with AIC and BIC-values. It is found that the model with no outliers is the best fitting one. This is also the conclusion obtained by Guttman *et al.* (1978).

TABLE 3.4. AIC and BIC values for model (3.2).

Number of outliers ($k = m$)	Observation number	R^2	AIC*)	BIC*)
0	-	0.9825	**-165.58**	**-165.58**
1	5	0.9873	-164.00	-163.01
2	5, 19	0.9906	-162.13	-160.14
3	5, 6, 19	0.9940	-163.33	-160.34
4	5, 6, 8, 19	0.9956	-161.87	-157.88

*) See Table 3.2

Data for our last example is from Draper and John (1981).

TABLE 3.5. Data from Draper and John (1981) ($n = 21$).

Observation number	x	y	Observation number	x	y
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100

As Draper and John, we fit the simple linear model

$$\mathrm{E}\,y = \beta_0 + \beta x. \tag{3.3}$$

If we look at a scatterplot of the data, we find that observations 18 and 19 are out of the body of other observations. Hence, these are potential outliers.

Table 3.6. gives AIC and BIC-values for some outlier models. In terms of both criteria the best fitting model is obtained when observation number 19 is discarded. Hence, observation 18 is not an outlier. Nevertheless, it can be considered as an *influential* observation as deduced in Draper and John (1981, p. 23).

TABLE 3.6. AIC and BIC values for model (3.3).

Number of outliers ($k = m$)	Observation number	R^2	AIC*)	BIC*)
0	-	0.4100	-101.84	-101.84
1	19	0.6575	**-105.17**	**-104.13**
2	3, 19	0.7139	-100.96	-98.70
3	3, 13, 19	0.7787	-98.46	-95.33
4	3, 13, 14, 19	0.8339	-96.71	-92.53
5	3, 13, 14, 19, 20	0.8819	-96.20	-90.98

*) See Table 3.2

Our last example considers clustered outliers. The data is an artificial set given in Hadi and Simonoff (1992). Here we, however, do not perform an exhaustive search as in the above example, but instead compare some potential outlier specifications. In this data there are 25 pairs of observations (x, y). The first 18 observations were generated by the regression model $y = x + \epsilon$, $\epsilon \sim N(0, 1)$. The x values were generated as $U(0, 15)$. Cases 19–25 are outliers planted near $x = 15$, resulting to leverage point outliers. For more details, see Hadi and Simonoff (1992).

Using GESR procedure the due to Paul and Fung (1991) Hadi and Simonoff report that the procedure nominates cases 6, 1, 2, 14, 7, 16, and 4 as potential outliers. They also use a procedure called Least Median Squares (LMS). This puts 1, 2, 6, 7, 10, 11, and 14 on the suspect list. A Two-Phase (T-P) method suggested by Paul and Fung (1991) as an extension to their GESR procedure nominates 2, 14, 1, 6, 11, 7, and 9 as outliers. None of these, however, emerge from the outlier population.

In the next table we have compared the outlier models in the above section in addition to the no-outlier alternative with AIC and BIC. The model used here is that in each case it is assumed that all of the outliers stem from the same outlier model. Hence the number of population identification parameters, k, is in this case one.

TABLE 3.7. AIC and BIC values Hadi's artificial data.

Number of outliers ($k, m = 1$)	Observation number	R^2	AIC*)	BIC*)
0	-	0.9561	-194.15	-194.15
7	19–25	0.9890	**-200.59**	**-199.37**
7	1,2,4,6,7,14,16	0.9625	-169.93	-168.71
7	1,2,6,7,10,11,16	0.9565	-166.22	-165.00
7	1,2,6,7,9,11,14	0.9568	-168.39	-165.17

*) AIC $= n \log(1 - R^2) - 2[\log k! + \log(n - k)!] + 2m$
BIC $= n \log(1 - R^2) - 2[logk! + \log(n - k)!] + m \log n$

We observe from Table 3.4 that the model with outliers 19–25 (the correct model) is best supported by both of the criterion functions, whereas the other three outlier models are not at all supported. Even the no-outlier alternative is considered better fitting. Of course in the last three cases a better alternative would be to allow also

the intercept and slope parameters to be different from the main body of data. This kind of outlier seeking, however, quickly leads to finding spurious outliers and too complicated models, as it actually would also be in this data set. Thus we do not continue reparametrizing the model any further.

4. Conclusions

In this paper we have considered detection of outliers in multiple regression analysis by information. The main advantage of these methods is that on does not have to bother the distribution of the observed residuals, which has proved to be complicated for the simple reason that the estimated residuals do not have a constant variance. One possibility solve this problem is to take so called Studentized residuals which have constant variances. Nevertheless, exact distribution for appropriate test statistics based on these adjusted residuals become intractable (see e.g. Barnett 1983, Ch. 10).

Detection of multiple outliers is fairly demanding procedure if performed as an exhaustive search procedure. That is why for future research it might be useful to develope some suboptimal procedures, like stepwise approach, for identifying multiple outliers.

REFERENCES

Akaike, H. (1977). On entropy maximization principle. In *Applications of Statistics,* ed. P.R. Krishnaiah, pp. 27–41. North-Holland, Amsterdam.

Akaike, H. (1980). On the use of the predictive likelihood of a Gaussian model. *Ann. Inst. Statist. Math.,* **32**, Part **A**, pp. 311–324.

Barnett, V. (1983). Principles and methods for handling outliers in data sets. *Statistical Methods and The Improvement of Data Quality,* pp. 131–166.

Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data,* 2nd ed. Wiley: New York.

David, H.A. (1981). *Order Statistics,* 2nd ed. Wiley: New York.

Draper, N.R. and John, J.A. (1981). Influential observations and outliers in regression. *Technometrics,* **23**, pp. 21–26.

Hadi, A. (1992). identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society*, B, **54**, pp. 761–771.

Hadi, A, and Simonoff, J (1992). Comment to Paul and Fung (1991). *Technometrics*, **34**, pp. 373-374.

Kitagawa, G. (1979). On the use of AIC for the detection of outliers. *Technometrics,* **21**, pp. 193–199. Corrigenda: *Technometrics,* **23**, pp. 320–321.

Kitagawa, G. (1984). Bayesian analysis of outliers via Akaike's predictive likelihood of a model. *Commun. Statist. -Simula. Computa.,* **13**, pp. 107–126.

Kitagawa, G. and Akaike, H. (1982). A quasi Bayesian approach to outlier detection. *Ann. Inst. Stat. Math.,* **34**, pp. 389–398.

Paul, S.R., and Fung, K.Y. (1991). A generalized extreme Studentized residual multiple-outlier-detection procedure in linear regression. *Technometrics*, **33**, pp. 339–348.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica,* **14**, pp. 465–471.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* **6**, pp. 461–464.