VERSITA
DOI:10.2478/bile-2013-0022

Biometrical Letters
Vol. 50 (2013), No. 2, 117-126

Detection of outlying observations using the Akaike information criterion

Andrzej Kornacki

Department of Applied Mathematics and Computer Science, University of Life Sciences in Lublin, Akademicka 13, 20-950 Lublin, Poland, e-mail: andrzej.kornacki@up.lublin.pl

SUMMARY

For the detection of outliers (observations which are seemingly different from the others) the method of testing hypotheses is most often used. This approach, however, depends on the level of significance adopted by the investigator. Moreover, it can lead to the undesirable effect of "masking" of the outliers. This paper presents an alternative method of outlier detection based on the Akaike information criterion. The theory presented is applied to analysis of the results of beet leaf mass determination.

Keywords: outliers, entropy, Akaike information criterion, Dixon test, Grubbs test

1. Introduction

In experiments carried out in the technical sciences, natural sciences and humanities, we often deal with a sample where the numerical values of certain observations differ significantly from the others. The presence of such an observation in a sample (i.e. an outlier) may be due to various types of measurement errors, equipment failures, etc. In other words these observations should be regarded as undesirable, derived from a different population and ultimately excluded from statistical analysis.

However, outliers with apparently large or small values can be accepted by the probability distribution of the characteristic, which would mean that in the experiment we simply have a feature of less common value. Thus it should be saved for the further statistical analysis, thereby increasing the efficiency of that analysis.

For the detection and final evaluation (inclusion or exclusion from further analysis) of an outlying observation, appropriate statistical tests can be used. The problem of the rejection of one outlying observation for a sample taken from a population with normal distribution has been investigated by numerous researchers, e.g. Breuning et al. (2000), Ferguson (1961), Galpin and Hawkins (1981), Grubbs (1950), Grubbs (1969), Joshi (1972), Rosseuw and Leroy (2000), Sakamoto et al. (1986), Stefansky (1972), Tietjen and Moore (1972).

In a multivariate normal model, rejection of outliers has been considered for example by Ferguson (1961), Karlin and Traux (1960), Pan and Fang (1995), Ramaswamy et al. (2000), Schwager and Margolin (1982), Srivastava and Von Rosen (1998), and Wilks (1963).

It should be noted that the detection of outliers using a test makes the statistical inference dependent on the level of significance of test, which in practice may mean obtaining different conclusions for different levels of significance. Moreover, statistical conclusions drawn from the performed test often depend on the number of observations considered as outliers, and the effect of masking of outliers may appear. This means that the same “suspicious” observations may be recognized as outliers in one subset of measurements, while in another they may not.

The purpose of this paper is to present an alternative method for detecting outliers, based on the general criterion of Akaike (AIC). This criterion, derived from information theory, was applied to select the best statistical model that describes (in terms of maximum entropy) real experimental data. The following discussion is based on the results of Akaike, (Akaike, 1973, 1977; Sakamoto et al., 1986) allowing the selection, from the models describing real data, of such a model that maximizes entropy using the function:

$$AIC = -2\ln(W) + 2K\,, \tag{1}$$

where W denotes the likelihood calculated for the maximum likelihood estimators of the parameters, and K is the number of parameters.

As suggested by Sakamoto (Sakamoto et al.) it would be best to choose the model for which the value of the AIC is the lowest.

The proposed method is an objective decision procedure. It is independent of the choice of significance level, the number of outliers and whether the outliers are highest or lowest values.

2. Camouflage of outlying observations

We will now quote an example of camouflage of outlying observations, given by Grubbs (1969). Before doing that, we recall some classical tests for detecting one or more outliers (Grubbs, 1969; Tietjen and Moore, 1972).

a) Tests for a single outlying observation:

$$\text{(i) } T_1 = \frac{\bar{x} - x_{(1)}}{s}, \quad T_n = \frac{x_{(n)} - \bar{x}}{s} \quad , \tag{2}$$

where s is the sample standard deviation of the form:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2}{n}} \tag{3}$$

b) Dixon tests:

$$\text{(ii)} \quad \begin{aligned} r_{ij}^1 &= \frac{x_{(i+1)} - x_{(1)}}{x_{(n-j)} - x_{(1)}} \\ r_{ij}^n &= \frac{x_{(n)} - x_{(n-i)}}{x_{(n)} - x_{(j+1)}}, \end{aligned} \tag{4}$$

where:

$$\begin{aligned} &i = 1,\ j = 0 \quad \textit{for } n \le 7 \ , \\ &i = j = 1 \quad \textit{for } n = 8,9,10 \ , \\ &i = 2,\ j = 1 \quad \textit{for } n = 11,12,13 \ , \\ &i = j = 2 \quad \textit{for } n \ge 14 \ . \end{aligned} \tag{5}$$

c) Grubbs tests:

$$\text{(iii)}\ L_1 = \frac{nS_1^2}{nS^2} \quad , \qquad L_n = \frac{nS_n^2}{nS^2}, \tag{6}$$

where:

$$\begin{aligned} nS^2 &= \sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 \quad , \\ nS_1^2 &= \sum_{i=2}^{n} (x_{(i)} - \bar{x}_1)^2, \quad \bar{x}_1 = \frac{1}{n-1}\sum_{i=2}^{n} x_{(i)}, \\ nS_n^2 &= \sum_{i=1}^{n-1} (x_{(i)} - \bar{x}_n)^2 \ , \\ \bar{x}_n &= \frac{1}{n-1}\sum_{i=1}^{n-1} x_{(i)}. \end{aligned} \tag{7}$$

d) Tests for multiple outliers (single-sided case):

$$\text{(iv)}\ L_k = \frac{nS_k^2}{nS^2} \quad , \quad L_{n-k} = \frac{nS_{n-k}^2}{nS^2} \ , \tag{8}$$

where:

$$\begin{aligned} nS_k^2 &= \sum_{i=k+1}^{n} (x_{(i)} - \bar{x}_k)^2 \ , \\ nS_{n-k}^2 &= \sum_{i=1}^{n-k} (x_{(i)} - \bar{x}_{n-k})^2 \ , \\ \bar{x}_k &= \frac{1}{n-k} \sum_{i=k+1}^{n} x_{(i)} \\ \bar{x}_{n-k} &= \frac{1}{n-k} \sum_{i=1}^{n-k} x_{(i)}. \end{aligned} \tag{9}$$

e) Tests for multiple outliers (double-sided case):

$$\text{(v)}\ E_k = \frac{\sum_{i=1}^{n-k} (z_i - \bar{z}_k)^2}{\sum_{i=1}^{n} (z_i - \bar{x})^2}, \tag{10}$$

where z_i is the value x_i of the *i*th smallest distance from the mean $\bar{x}$ and

$$\bar{z}_k = \frac{1}{n-k}\sum_{i=1}^{n-k} z_i .$$

Critical values for these statistics for certain significance levels are given in Grubbs (1969) and Tietjen and Moore (1972).

Grubbs (1969) cites the following data on the percentage elongation at break of selected synthetic materials (after ordering):

2.02; 2.22; 3.04; 3.23; 3.59; 3.73; 3.94; 4.05; 4.11; 4.13

In this case, one may initially be interested only in outlying observations to the left of the mean, because very high readings indicate a remarkable plasticity of the material, which is a desired feature. The questionable results here are the two lowest values: 2.02, 2.22. We calculate the test values:

$$\begin{cases} T_1 = \dfrac{3.406-2.02}{0.7711} = 1.7975 \\ r_{11}^1 = \dfrac{2.22-2.02}{4.11-2.02} = 0.0975 \\ L_1 = \dfrac{3.217}{5.351} = 0.6011 \\ L_2 = \dfrac{1.197}{5.351} = 0.224 \end{cases} \tag{11}$$

None of these tests recognizes the lowest value (the lowest single observation) in the sample as an outlier. The resulting test values do not exceed the critical value, because $T_1 \le 2.18$; $r_{11}^1 < 0.477$; $L_1 \le 0.0418$, while the calculated value $L_2 = 0.224 < 0.2305$ at the significance level $\alpha = 0.05$ detects as outliers the two lowest observations. This example shows that the tests applied sequentially (first to one and then to two observations) are not always useful, because of the effect of masking of outlying observations.

The above-mentioned problems do not arise in the case of the Akaike information criterion. The unambiguous indication by this criterion of outliers which need to be removed naturally eliminates the masking effect of outlier observations in the sample.

3. Model of outlying observations

Let us consider a sample of n observations, arranged in increasing order of value $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Here $x_{(k)}$ is the value of the kth positional statistic $X_{k,n}$. Henceforth we use the following notation: $\psi(x;\mu,\sigma^2)$ is the density of a normal distribution with mean μ and variance σ^2, $\Phi(x;\mu,\sigma^2)$ is the distribution function of this distribution, and $f_r(x;\mu,\sigma^2)$ is the density of the rth positional statistics from the normal population, i.e.:

$$\psi(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{12}$$

$$\Phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt ,$$

(13)

and (see David, 1979)

$$f_r(x;\mu,\sigma^2) = B(r,n-r+1)^{-1} \Phi(x;\mu,\sigma^2)^{r-1} \times \left\{1-\Phi(x;\mu,\sigma^2)^{n-r}\right\} \psi(x;\mu,\sigma^2), \tag{14}$$

where:

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt,\ p>0, q>0 \tag{15}$$

denotes the beta function. It is known that:

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \frac{(p-1)!(q-1)!}{(p+q-1)!} \tag{16}$$

for natural p and q. Moreover, Γ denotes the Γ function of the form:

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad \operatorname{Re} z > 0 \tag{17}$$

The model describing data with possible outliers after taking account of (12)–(16) can be represented by the density function:

$$h_r(x)=\begin{cases}\psi(x;\mu,\tau^2) & r=1,\ldots,n_1\\ f_{r-n_1,n-n_1-n_2}(x;\mu,\sigma^2) & r=n_1+1,\ldots,n-n_2\\ \psi(x;\mu,\tau^2) & r=n-n_2+1,\ldots,n\end{cases}. \tag{18}$$

The model described by (18) means that the $n–n_1–n_2$ middle observations $x_{(n_1+1)},\ldots,x_{(n-n_2)}$ are realizations of normal variables with mean μ and variance σ^2, while the n_1 initial observations $x_{(1)},\ldots,x_{(n_1)}$ and the n_2 final observations $x_{(n-n_2+1)},\ldots,x_{(n)}$ are drawn from a normal population with the same mean μ and variance τ^2. In this model, we consider the results $x_{(1)},\ldots,x_{(n_1)}$ and $x_{(n-n_2+1)},\ldots,x_{(n)}$ as "candidate" outlying observations.

The model without outliers is certainly given by:

$$g(x)=\psi(x,\mu,\sigma^2) \tag{19}$$

If for experimental data, the AIC value for model (18) is less than for model (1), we state that observations $x_{(1)},\ldots,x_{(n_1)}$ and $x_{(n-n_2+1)},\ldots,x_{(n)}$ are outliers. Otherwise none of these observations are outliers.

Now we derive the formula for the AIC for model (18). The likelihood function of this model can be written as follows:

$$\begin{aligned}&L(x;n_1,n_2,\mu,\sigma^2,\tau^2)=\\ &=\prod_{i=1}^{n_1}\psi(x_{(i)},\mu,\tau^2)\times\prod_{i=n_1+1}^{n-n_2}f_{i-n_1,n-n_1-n_{23}}(x_{(i)};\mu,\sigma^2)\times\prod_{i=n-n_2+1}^{n}\psi(x_{(i)},\mu,\tau^2).\end{aligned} \tag{20}$$

Taking logarithms of the functions (8) we obtain the relationship:

$$\begin{aligned}&l=-\frac{1}{2}\left\{n\ln 2\pi+\sum_{i=1}^{n}\ln\sigma^2(i)+\sum_{i=1}^{n}\frac{(x_{(i)}-\mu)^2}{\sigma^2(i)}\right\}-\\ &-\sum_{i=n_1+1)}^{n-n_2}\left[\ln B(j,k-j+1)-(j-1)\ln\{\Phi(x_{(i)}\}-(k-j)\ln\{1-\Phi(x_{(i)})\}\right],\end{aligned} \tag{21}$$

where $j=i-n_1$, $k=n-n_1-n_2$, (22)

and

$$\sigma^2(i)=\begin{cases}\sigma^2, & n_1+1\le i\le n-n_2,\quad n_1<i\le n-n_2\\ \tau^2, & 1<i\le n_1\ or\ n-n_2<i\le n\end{cases} \tag{23}$$

By (21)–(23) the Akaike information criterion (the minimum value (1)) takes the form:

$$AIC(i,j)=\begin{cases}-2l(x;i,j,\hat{\mu},\hat{\sigma}^2)+2\times 2 & (i=j=0)\\ -2l(x;i,j,\hat{\mu},\hat{\sigma}^2,\hat{\tau}^2)+2\times 3 & (otherwise)\end{cases},$$

where $\hat{\mu},\hat{\sigma}^2,\hat{\tau}^2$ denote the maximum likelihood estimators of the parameters.

4. Numerical example of use of the Akaike criterion for detecting outliers

The theory presented above will now be applied to the results of an experiment with *Beta vulgaris*. An experiment with *Beta vulgaris* L. "Vulcan" was conducted in the spring of 2006 (29 March to 22 May) and 2007 (17 March to 14 May) in a greenhouse. The experiment was set up by the method of complete randomization. The purpose of the research was to determine the effect of the type of potassium fertilizer on the yield and chemical composition of *Beta vulgaris* L. (Dzida et al., 2011). For one of the series (6 repetitions) the following results were obtained for beet mass: 32; 61; 106; 140; 199; 203 (g). The theory presented in Chapter 3 allows us to calculate the value of the Akaike information criterion for different configurations of outliers. The results of the calculations are presented in Table 1.

The data in Table 1 show that there are no outliers, because it is for such a configuration that the Akaike information criterion value is the lowest (marked with an asterisk). This conclusion is confirmed by the tests described in section 3. In the present example, we have:

$$\begin{cases}T_1=\dfrac{123.5-32}{58.8102}=1.5558\\ r_{10}^1=\dfrac{61-32}{203-32}=0.17\\ l_1=\dfrac{14830.8}{24877.5}=0.59\end{cases}$$

Table 1. Values of Akaike information criterion for mass of *Beta vulgaris*. L. 'Vulcan'

		High outliers			
		None	203	203; 199
Low outliers	None	**59.5793***	64.4871	69.9154
	31	65.1081	65.4880	68.3227
	31; 62	70.5428	69.0551	67.1289

Since T_1 and r_{10}^1 are less than the critical values 1.996 and 0.56 respectively, and L is greater than the critical value 0.2032, all of the tests confirm that observation **31** is not an outlier.

Similarly both the single observation **203** and the two pairs of observations **(31;62)** and (**199;203**) are not outliers according to the value of the statistics:

$$\begin{cases} L_6 = \dfrac{17293}{24877.5} = 0.69 \\ L_2 = \dfrac{6670}{24877.5} = 0.067 \\ L_{6-2} = \dfrac{6850.75}{24877.5} = 0.2754 \end{cases}$$

and the critical values: $L^{(1)}$=0.2032 and $L^{(2)}$=0.0565.

5. Conclusions

This paper considers the problem of experimental results that are "strikingly" different from others – so-called outliers. Classical statistical methods for detecting and rejecting outliers are based on testing of hypotheses. These methods, however, depend on the adopted level of significance, and in some cases may lead to the effect of masking of outlying observations. In this paper it is proposed to use an alternative outlier detection method based on the Akaike information criterion. This is an objective procedure without the abovementioned drawbacks of inference based on hypothesis testing.

REFERENCES

Akaike H. (1973): Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, eds. B.N. Petrv and F. Csaki, Budapest; Akademiai Kiado: 267-281.

Akaike H. (1977): On entropy maximization principle. Proc Symposium on Applications of Statistics, ed. P.R. Krishnaiah, Amsterdam: North Holland: 27-47.

Barnett V., Lewis T. (1993): Outliers in Statistical Data. John Wiley & Sons.

Breuning M., Kriegel H.P, Sander J. (2000): LOF: Identifying Density-Based Local Outliers. In: Proceedings of the ACM SIGMOND Conference: 93-104.

David H.A. (1979): Pariadkowyje statistiki. Mockba Nauka.

Dzida K., Jarosz Z., Michałojć Z. (2011): The effect of diversified potassium fertilization on the field and chemical composition on Beta Vulgaris L. Acta Sci. Pol. Hortus. Cultus 10(40): 263-274.

Ellenberg J.H. (1976): Testing for a single outlier from a general linear regression. Biometrics 32: 637-645.

Ferguson T.S. (1961): On the rejection of outliers. In Proc. Fourth Berkeley Symposium Math. Statist. Prob.1: 253-287.

Galpin J.S., Hawkins D.M. (1981): Rejection of a single outlier in two or three-way layouts. Technometrics 23: 65-70.

Grubbs F.E. (1950): Sample criteria for testing outlying observations. Ann. Math. Statist. 21: 27-58.

Grubbs F.E. (1969): Procedures for detecting outlying observations in samples. Technometrics 11: 1-21.

Joshi P.C. (1972): Some slippage tests of mean for a single outlier in linear regression. Biometrika 59: 109-120.

Karlin S., Traux D. (1960): Slippage problems. Ann. Math. Statist 31**:** 296-324.

Pan J.X., Fang K.T. (1995): Multiple outlier detection in growth curve model with unstructured covariance matrix. Ann. Inst. Statist. Math. 47: 137-153.

Ramaswamy S., Rastogi R., Shim K. (2000): Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM SIGMOND Conference on Management of data. Dallas: 427-438.

Rosseuw P., Leroy A. (2000): Robust Regression and Outlier Detection. John Wiley & Sons.

Sakamoto Y., Ishigura M. (1986): Akaike Information Criterion Statistics. Tokyo Reidel Publishing Company.

Schwager S.J., Margolin B.H. (1982): Detection of multivariate normal outliers. Ann. Statist. 10: 943-954.

Srivastava M.S., Von Rosen D. (1998): Outliers in Multivariate Regression Models. J. Mult. Anal. 65: 195-208.

Stefansky W. (1972): Rejecting outliers in factorial designs. Technometrics 14: 469-479

Tietjen G.L., Moore R.H. (1972): Some Grubbs-type statistics for the detection of several outliers. Technometrics 14: 583-597.

Wilks S.S. (1963): Multivariate statistical outliers. Sankhya *A.* 25: 406-427.