TEKA. COMMISSION OF MOTORIZATION AND ENERGETICS IN AGRICULTURE – 2012, Vol. 12, No. 2, 111–115

Application of akaike information criterion for the detection of outliers

Andrzej Kornacki, Khachatur Kyureghyan, Szymon Ignaciuk

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 an undesirable effect of "masking" the outliers. This paper presents an alternative method of outlier detection based on the Akaike information criterion. Statistical calculations and comparative analysis for the proposed method were conducted with commonly used statistical tests on the basis of the classical Grubbs experiment and the research into the combustion of biomass with plant composition. The advantages of the method and rationale for the selection of the appropriate statistical model were formulated in the form of conclusions.

Key words: outliers, data entropy, Akaike information criterion, Dixon test, Grubbs test.

1. INTRODUCTION

In the experiments carried out in the field of technical sciences, natural sciences and humanities we are often dealing with a sample, where the numerical values of some 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 considered experiment we have a feature of less common value. So, it should be saved for further statistical analysis, thus increasing its efficiency.

For the detection and final evaluation (inclusion or exclusion from further analysis) of an outlying single observation the appropriate statistical test can be used, described by [24]. The problem with rejecting one outlying observation for the sample taken from a population with normal distribution was investigated by numerous researchers e.g. [8,9,10,12,16,18,25]. In a multivariate normal model rejecting outliers was considered e.g. by [8,13,17,20,21,23,24].

It should be noted that the detection of outliers with a test makes the statistical inference dependent on the level of test significance, which in practice may mean obtaining different conclusions for different levels of the test. Also, statistical conclusions drawn from the performed test often depend on the number of observations considered as outliers (masking outliers). This means that the same "suspicious" observations in one subset of measurements may be recognized as outliers, and in another may not.

The purpose of this paper is to present an alternative method for detecting outliers based on the general criterion of Akaike. This criterion, derived from information theory, was applied to select the best statistical model that describes (in terms of maximum entropy) real experiment data. The following discussion is based on the results of [1,2,21] allowing for the choice from the models describing real data of such a model that maximizes entropy by using the function:

$$AIC = -2\ln(W) + 2K, \qquad (1)$$

where:

W - likelihood calculated for the parameter estimates, obtained by the method of maximum likelihood, K - number of parameters.

As suggested by Sakamoto, it would be best to choose the model for which AIC value is the lowest.

2. THE MODEL OF OUTLIERS

Let us consider observation test n which, when rearranged according to increasing values, creates the set:

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}.$$

So $x_{(k)}$ is the value of the k-th positional statistics $X_{k,n}$. In the rest of the paper 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 r-th positional statistics from the normal population, i.e.:

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

$$\Phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}\exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}dt, \tag{3}$$

$$f_r(x;\mu,\sigma^2) = B(r,n-r+1)^{-1}\Phi(x;\mu,\sigma^2)\times \times\left\{1-\Phi(x;\mu,\sigma^2)^{n-r}\right\}\psi(x;\mu,\sigma^2), \tag{4}$$

where:

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt\,, \quad p>0\,, \quad q>0 \tag{5}$$

denotes the function Beta [David 1979]. It is known that:

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \frac{(p-1)!(q-1)!}{(p+q-1)!} \tag{6}$$

for natural p and q.

The model describing data with possible outliers after taking into account (2) - (6) can be represented by the density function:

$$h_r(x) = \begin{cases} \psi(x;\mu_1,\sigma^2) & 1 \le r \le n_1 \\ f_{r-n_1,n-n_1-n_2}(x;\mu,\sigma^2) & r = n_1+1,\ldots,n-n_2. \\ \psi(x;\mu_{,2},\sigma^2) & r = n-n_2+1,\ldots,n \end{cases} \tag{7}$$

The model described by (7) means that n_1 of initial observations: $x_{(1)},\ldots,x_{(n_1)}$, $n-n_1-n_2$ of the middle observations: $x_{(n_1+1)},\ldots,x_{(n-n_2)}$ and n_2 of the final observations: $x_{(n-n_2+1)},\ldots,x_{(n)}$ are realizations of normal variables with the same variance σ^2, and the means, respectively, μ_1,μ,μ_2. In this model, we consider the results $x_{(1)},\ldots,x_{(n_1)}$ and $x_{(n-n_2+1)},\ldots,x_{(n)}$ as „candidates" for outlying observations.

Likelihood function of the model (7) can be written as follows:

$$L(x;n_1,n_2,\mu,\mu_1,\mu_2,\sigma^2) = \prod_{i=1}^{n_1}\psi(x_{(i)},\mu_1,\sigma^2)$$
$$\times\prod_{i=n_1+1}^{n-n_2} f_{i-n_1,n-n_1-n_{23}}(x_{(i)};\mu,\sigma^2)\times\prod_{i=n-n_2+1}^{n}\psi(x_{(i)},\mu_2,\sigma_2) \tag{8}$$

From logarithms of functions (7) we get the relationship:

$$l_1 = -\frac{1}{2}\left\{n\ln 2\pi + n\ln\sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_{(i)}-\mu^i)^2\right\} -$$
$$-\sum_{i=n_1+1)}^{n-n_2}\Big[\ln B(j,k-j+1)-(j-1)\ln\left\{\Phi(x_{(i)}\right\}-(k-j)\ln\left\{1-\psi(x_{(i)})\right\}\Big] \tag{9}$$

where:

$$j = i-n_1\,, \quad k = n-n_1-n_2 \tag{10}$$

and

$$\mu^i = \begin{cases} \mu_1\,, & 1 \le i \le n_1 \\ \mu\,, & n_1 < i \le n-n_2. \\ \mu_2\,, & n-n_2 < i \le n \end{cases} \tag{11}$$

By (8) - (11) the modified Akaike criterion (the minimum value (1)) takes the form:

$$AIC(i,j) = \begin{cases} -2l_1(x;i,j,\hat{\mu},\hat{\sigma}^2)+2\times 2 & (i=j=0) \\ -2l_1(x;i,j,\hat{\mu},\hat{\mu}_1,\hat{\sigma}^2)+2\times 3 & (i\ne 0, j=0) \\ -2l_1(x;i,j,\hat{\mu},\hat{\mu}_2,\hat{\sigma}^2)+2\times 3 & (i=0, j\ne 0) \\ -2l_1(x;i,j,\hat{\mu},\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}^2)+2\times 4 & (i\ne 0, j\ne 0) \end{cases}, \tag{12}$$

where: $\hat{\mu},\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}^2$ denote parameter estimates obtained by the method of maximum likelihood.

3. STATISTICAL CALCULATIONS FOR CLASSICAL TESTS AND INFORMATION CRITERION

Below is a description of the most popular classical tests for detecting one or two outliers.

a) Tests for a single outlying observation

$$\text{(i) } T_1 = \frac{\bar{x}-x_{(1)}}{s}\,, \quad T_n = \frac{x_{(n)}-\bar{x}}{s}, \tag{13}$$

where: s is the sample standard deviation.

b) Dixon tests

$$\text{(ii) } r_{ij}^1 = \frac{x_{(i+1)}-x_{(1)}}{x_{(n-j)}-x_{(1)}} \qquad r_{ij}^n = \frac{x_{(n)}-x_{(n-i)}}{x_{(n)}-x_{(j+1)}}, \tag{14}$$

where:

$$\begin{aligned} &i=1, j=0 \quad for\ n\le 7, \\ &i=j=1 \quad for\ n=8,9,10, \\ &i=2, j=1 \quad for\ n=11,12,13, \\ &i=j=2 \quad for\ n\ge 14. \end{aligned} \tag{15}$$

c) Grubbs tests:

$$\text{(iii) } L_1 = \frac{nS_1^2}{nS^2}\,, \qquad L_n = \frac{nS_n^2}{nS^2}, \tag{16}$$

where:

$$nS^2 = \sum_{i=1}^{n}(x_{(i)} - \overline{x})^2 \quad ,$$

$$nS_1^2 = \sum_{i=2}^{n}(x_{(i)} - \overline{x}_1)^2 \overline{x}_1 = \frac{1}{n-1}\sum_{i=2}^{n} x_{(i)},$$

$$nS_n^2 = \sum_{i=1}^{n-1}(x_{(i)} - \overline{x}_n)^2 \; , \quad \overline{x}_n = \frac{1}{n-1}\sum_{i=1}^{n-1} x_{(i)}. \tag{17}$$

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

$$\text{(iv) } L_k = \frac{nS_k^2}{nS^2} \quad , \quad L_{n-k} = \frac{nS_{n-k}^2}{nS^2}, \tag{18}$$

where:

$$nS_k^2 = \sum_{i=k+1}^{n}(x_{(i)} - \overline{x}_k)^2 \; , nS_k^2 = \sum_{i=1}^{k}(x_{(i)} - \overline{x}_{n-k})^2,$$

$$\overline{x}_k = \frac{1}{n-k}\sum_{i=k+1}^{n} x_{(i)} \; , \quad \overline{x}_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} x_{(i)} \tag{19}$$

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

$$\text{(v) } E_k = \frac{\sum_{i=1}^{n-k}(z_i - \overline{z}_k)^2}{\sum_{i=1}^{n}(z_i - \overline{x})^2}, \tag{20}$$

where: z_i is the value x_i of the i-th smallest distance from the mean $\overline{x}$ and $\overline{z}_k = \frac{1}{n-k}\sum_{i=1}^{n-k} z_i$

Critical values for these statistics for certain significance levels are given in [10,11].

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, you can only initially get interested in outlying observations to the left of the mean, because the very high readings indicate a remarkable plasticity of the material, which is a desired feature. Questionable results here are the two lowest values: 2.02, 2.22. We calculate the values of tests:

Below is the statistical calculation based on the figures from the experience performed by G. Maj in 2011. The experiment tested, among others, [May 2011] the percentage of ash in pellets made from 11 different plant materials, depending on the combustion temperature and moisture levels. Combustion of the tested biomass in the form of test pellets was performed using the muffle furnace Nabertherm L3/B180. 1-2 g test sample of solid fuel was placed in the oven and heated to the temperature of 600^{0}C or 815^{0}C. The ash content in the test sample of solid fuel was calculated using the following formula:

$$\begin{cases} T_1 = \dfrac{3,406 - 2,02}{0,7711} = 1,7975 \\ r_{11}^1 = \dfrac{2,22 - 2,02}{4,11 - 2,02} = 0,0975 \\ L_1 = \dfrac{3,217}{5,351} = 0,6011 \\ L_2 = \dfrac{1,197}{5,351} = 0,224 \end{cases}$$

$$A^a = \frac{m_3 - m_1}{m_2 - m_1} \bullet 100 \tag{21}$$

where:
A^a- ash content of the test sample [%],
m_1- ignited cell mass [g],
m_2- cell mass with the weighed solid fuel [g],
m_3- cell mass with ash [g].

One series of measurements in the context of our discussion seems to be particularly interesting. The giant *Miscanthus* combusted at 815^{0}C rendered the following results:

3,4;3,42;3,45;3,67;3,71;25,93.

The theory presented in Part 2 allows calculation of the Akaike information criterion for various configurations of outliers. The results of calculations are presented in Table 4.1, while the values of classical tests after the calculations are as follows:

$$T_6 = \frac{25,93 - 7,625}{8,3481} = 2,2358,$$

$$r_{10}^6 = \frac{25,93 - 3,71}{25,93 - 3,41} = 0,9867,$$

$$L = \frac{0,0849}{418,14} = 0,002.$$

Table 4.1. Values of Akaike information criterion for the ash content in giant *Miscanthus*

		High outliers		
		None	25,39	25,39 4,25
Low outliers	None	32,6507	**32,0385** *	43,4342
	3,41	38,4383	37,8094	44,4609
	3,41 3,42	41,1633	40,9162	46,7224

4. SUMMARY AND CONCLUSIONS

The presented modified Akaike information criterion allows for the choice of the correct statistical model in the set of models describing a particular experiment and takes into account the maximum value of entropy. At the same time it is independent from the selected different levels of significance of statistical tests used to verify the hypotheses formulated within the study. Simultaneously, it is an analytical indication concerning the exclusion of the optimal number of outliers in the sample, while maintaining a hypothetical probability distribution of the tested characteristic. Unambiguous indication by the criterion of the outliers which need to be removed naturally eliminates the potential for masking effect of outliers in the sample.

Conclusion 1. The classic experiment of Grubbs discussed by many authors is a typical example of the ambiguity of statistical inference based on the classical tests. None of these tests recognizes the outlying of the lowest value (the lowest single observation) in the sample. The resulting test values do not exceed the predicted critical thresholds because $T_1 < 2{,}18;\quad r_{11}^{1} < 0{,}477;\quad L_1 \geq 0{,}418$, while the calculated value $L_2 = 0{,}224 < 0{,}2305$ at the significance level α = 0,05 detects the outlying of the two lowest observations, while for higher values α still does not detect outlying.

The above-mentioned problems are not noticed in the case of the modified Akaike criterion, since the lowest value of the function (1.1) at 16.041 obtained for the two low outliers clearly suggests the rejection of the two lowest observations.

Conclusion 2. The calculated values of the criterion presented in Table 4.1 clearly indicate the correct model configurations (single outlier value to the right of the mean), because the maximum observation value 25.39 (shown in Table 4.1) corresponds to the minimum value of the function (1.1). This conclusion, in this case, is consistent with the conclusions of the classical tests, since at the level of significance α = 0,05 we obtain:

$$T_6 = 2{,}2358 > 1{,}996;\quad r_{10}^{6} = 0{,}9867 > 0{,}56;\quad L_1 = 0{,}002 < 0{,}2032$$

which means that the values of all the classical tests are in the critical area.

REFERENCES

1. **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, 267-281. Budapest; Akademiai Kiado.
2. **Akaike H., 1977:** On entropy maximization principle. Proc Symposium on Applications of Statistics, ed. P.R. Krishnaiah, 27-47, Amsterdam: North Holland.
3. **David H.A., 1956a:** On the application of an elementary theorem in probability. *Biometrika* 43 85-91.
4. **David H.A., 1956b:** Revised upper percentage points of the extreme deviate from the sample mean. *Biometrika* 43 449-451.
5. **David 1979.** *Pariadkowyje statistiki.*Mockba Nauka
6. **Ellenberg, J.H., 1973:** The joint distribution of the standardized least squares residuals from a general linear regression.*J.Amer.Statist.Assoc.* 68 941-943.
7. **Ellenberg J.H., 1976:** Testing for a single outlier from a general linear regression. *Biometrics* 32 637-645.
8. **Ferguson T.S., 1961:** On the rejection of outliers. In *Proc.Fourth Berkeley Symposium Math.Statist.Prob.1*, 253-287.
9. **Galpin J.S., and Hawkins D.M. 1981:** Rejection of a single outlier in two or three-way layouts. *Technometrics* 23 65-70.
10. **Grubbs F.E., 1950:** Sample criteria for testing outlying observations. *Ann.Math.Statist.* 21 27-58.
11. **Grubbs F.E., 1969:** Procedures for detecting outlying observations in samples. *Technometrics.*11 1-21.
12. **Joshi P.C., 1972:** Some slippage tests of mean for a single outlier in linear regression. *Biometrika* **59** 109-120.
13. **Karlin S., and Traux D., 1960:** Slippage problems. *Ann.Math.Statist* 31 296-324.
14. **Kudô A., 1956:** On the testing of outlying observations. *Sankhya* 17 67-76.
15. **Nair K.R., 1948:** The distribution of the extreme deviate of the sample mean and its studentized form. *Biometrika* 35 118-134.
16. **Niedziółka I., Szymanek M., 2010:** An estimation of phisycal properties briquettes produced from plant biomass. TEKA commission of motorization and energetics in agriculture Vol. 10, No. 2, 301-307.
17. **Pan J.X., and Fang K.T., 1995:** Multiple outlier detection in growth curve model with unstructured covariance matrix.*Ann.Inst.Statist.Math.*47. 137-153.
18. **Queensberry C.P., and David H.A., 1961:** Some tests of outliers *Biometrika* 48 370-390.
19. **Schwager S.J and Margolin B.H., 1982:** Detection of multivariate normal outliers. *Ann.Statist.* 10. 943-954.
20. **Siotani M., 1959:** The extreme value of generalized distances of the individual points in the multivariate normal sample. *Ann.Inst.Statist.Math* 10 183-208
21. **Srikantan K.S., 1961:** Testing for the single outlier in regression model. *Sankhya A* 23 251-260.
22. **Srivastava M.S., 1997:** Slippage tests of mean for a single outlier in multivariate normal data *Amer.J.Manage. Sci.*
23. **Srivastava M.S., and Von Rosen D., 1998:** Outliers in Multivariate Regression Models.*J.Mult.Anal.* 65. 195-208.
24. **Stefansky W., 1972:** Rejecting outliers in factorial designs. *Technometrics.* 14 469-479.
25. **Thompson W.R., 1935:** On a criterion for the rejection of observations and the distribution of the ratio of the deviation of the sample standard deviation. *Ann.Math. Statist.* 6 214-219.
26. **Wilks S.S., 1963:** Multivariate statistical outliers. *Sankhya A.* 25 406-427.

ZASTOSOWANIE KRYTERIUM INFORMACYJNEGO AKAIKE DO WYKRYWANIA OBSERWACJI ODSTAJĄCYCH

S t r e s z c z e n i e . Do wykrywania obserwacji odstających (pozornie odbiegających od pozostałych) najczęściej stosuje się metody testowania hipotez. Podejście takie zależy jednak od przyjętego przez badacza poziomu istotności. Ponadto może ono prowadzić do niepożądanego efektu „maskowania” ostających obserwacji. W niniejszej pracy przedstawiono alternatywną metodę wykrywania odstających obserwacji bazującą na kryterium informacyjnym Akaike. Kalkulacje statystyczne oraz analizę porównawczą, proponowanej metody z powszechnie stosowanymi testami statystycznymi, przeprowadzono na podstawie klasycznego eksperymentu Grubbsa oraz badań dotyczących spalania biomasy o składzie roślinnym. Zalety metody oraz uzasadnienie wyboru odpowiedniego modelu statystycznego sformułowano w postaci wniosków końcowych.

Słowa kluczowe: obserwacje odstające, entropia danych, kryterium informacyjne Akaike, test Dixona, test Grubbsa.