Astronomical and Astrophysical Transactions (*AApTr*), 2012, Vol. 27, Issue 3, pp. 537–538,
ISSN 1055-6796, Photocopying permitted by license only, ©Cambridge Scientific Publishers

Exclusion of measurements with excessive residuals (blunders) in estimating model parameters

I.I. Nikiforov*

Sobolev Astronomical Institute of St. Petersburg State University, 28 Universitetskij Prospekt, Staryj Peterhof, St. Petersburg 198504, Russia

Received 20 November, 2008

An adjustable algorithm of exclusion of conditional equations with excessive residuals is proposed. The criteria applied in the algorithm use variable exclusion limits which decrease as the number of equations goes down. The algorithm is easy to use, it possesses rapid convergence, minimal subjectivity, and high degree of generality.

Keywords: Estimation of model parameters; Conditional equations; Large residuals; Criteria of exclusion

1 Introduction

In many astronomical (and not only astronomical) problems of estimation of model parameters, it is important to exclude unreliable data which produce large residuals, i.e., deviations, ε, of measurements from the accepted model: $|\varepsilon_j|/\sigma_j \gg 1$, where σ_j is the standard deviation for j-th measurement, $j = 1, \ldots, N$, N is the number of measurements, i.e., of conditional equations. The occurrence of large residuals ("blunders") contradicts the basic assumption of least-squares fitting on the normal distribution of measurement errors and can cause strong biases of parameter estimates. The common "3σ" criterion to exclude blunders

$$\frac{|\varepsilon_j|}{\sigma_j} > k = 3 \tag{1}$$

does not allow for the probability of accidental occurrence of residual (1) to increase with N and become not negligible already at N of the order of several tens.

In this paper, a more adjustable algorithm of exclusion of equations with excessive residuals on the basis of a variable criterion limit is elaborated.

2 Algorithm of excluding measurements with excessive residuals

1. For a given N, a value of κ which satisfies the equation

$$[1 - \psi(\kappa)]\, N = 1, \qquad \psi(z) \equiv \sqrt{\frac{2}{\pi}} \int_0^z e^{-\frac{1}{2}t^2}\, dt, \tag{2}$$

*Email: nii@astro.spbu.ru

where $\psi(z)$ is the probability integral, is found. The expectation value for the number of conditional equations with residuals

$$|\varepsilon_j|/\sigma_j > \kappa, \tag{3}$$

equals one, if residuals are normally distributed. A larger number of equations with such residuals may be considered as probably excessive.

2. The number L of equations satisfying the criterion (3) is determined.

3. If $L > 1$, $L - L'$ equations with the largest values of $|\varepsilon_j|/\sigma_j$ are excluded from consideration. Here, $L' \geq 1$ is a parameter of the algorithm.

4. The criterion (1) with k depending on N is applied to the remaining equations, in particular if $L = 1$:

$$|\varepsilon_j|/\sigma_j > k_\gamma(N), \tag{4}$$

where k_γ is the root of the equation

$$1 - [\psi(k_\gamma)]^N = \gamma. \tag{5}$$

Here, γ is an accepted confidence level. For low γ, i.e., for low $1 - \psi(k_\gamma)$, in lieu of (5) an approximate equation can be used:

$$[1 - \psi(k_\gamma)]N = \gamma. \tag{6}$$

5. Following the exclusion of equations with excessive residuals, a new solution of the problem is found from the remaining equations. Thereupon points 1–4 of this algorithm are applied again with new estimates of parameters and σ_j. The iterations are interrupted if no further exclusion happens.

The probability $\mathcal{P}(L)$ of accidental occurrence of L residuals satisfying (3) can be approximately evaluated with the Poisson distribution, which is $\mathcal{P}(L) = e^{-1}/L!$ in this case. This approach gives

$$\mathcal{P}(L \geq 2) \approx 0.264, \qquad \mathcal{P}(L \geq 3) \approx 0.080, \qquad \mathcal{P}(L \geq 4) \approx 0.019.$$

Thus numbers of $L = 3$ and 4 can be considered as excessive, i.e., $L' = 2$ or 3 can be correspondingly accepted. However, if unbiased parameters are more important than an unbiased residual variance, $L' = 1$ is also allowed.

Point 4 of the algorithm is essential in the case of only a single (or few) very large blunder(s), when point 3 cannot come into action. A level of $\gamma = 0.05$, being the standard one in many statistical criteria, can be accepted.

Acknowledgments

The work is partly supported by the Russian Foundation for Basic Research grant 08-02-00361 and the Russian President Grant for State Support of Leading Scientific Schools of Russia no. NSh-1323.2008.2.