Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Statistical Planning and Inference 135 (2005) 477–486

journal of statistical planning and inference

www.elsevier.com/locate/jspi

The probability to select the correct model using likelihood-ratio based criteria in choosing between two nested models of which the more extended one is true

Nelly van der Hoeven[a, b, *]

[a] *Department of Theoretical Biology, Institute of Biology, Leiden University, P.O. Box 9516, 2300 RA Leiden, The Netherlands*
[b] *ECOSTAT, Vondellaan 23, 2332 AA Leiden, The Netherlands*

Received 24 March 2003; accepted 27 April 2004
Available online 23 July 2004

Abstract

The probability to select the correct model is calculated for likelihood-ratio-based criteria to compare two nested models. If the more extended of the two models is true, the difference between twice the maximised log-likelihoods is approximately noncentral chi-square distributed with d.f. the difference in the number of parameters. The noncentrality parameter of this noncentral chi-square distribution can be approximated by twice the minimum Kullback–Leibler divergence (MKLD) of the best-fitting simple model to the true version of the extended model.

The MKLD, and therefore the probability to select the correct model increases approximately proportionally to the number of observations if all observations are performed under the same conditions. If a new set of observations can only be performed under different conditions, the model parameters may depend on the conditions, and therefore have to be estimated for each set of observations separately. An increase in observations will then go together with an increase in the number of model parameters. In this case, the power of the likelihood-ratio test will increase with an increasing number of observations. However, the probability to choose the correct model with the AIC will only increase if for each set of observations the MKLD is more than 0.5. If the MKLD is less than 0.5, that probability will decrease. The probability to choose the correct model with the BIC will always decrease,

*Corresponding address. ECOSTAT, Vondellaan 23, 2332 AA Leiden, The Netherlands. Tel.: +31-71-5315011; fax: +31-842-116988.

E-mail address: nvdh@ecostat.nl (N. van der Hoeven).

0378-3758/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.jspi.2004.04.025

sometimes after an initial increase for a small number of observation sets. The results are illustrated by a simulation study with a set of five nested nonlinear models for binary data.
© 2004 Published by Elsevier B.V.

MSC: 62J02; 62F07; 62B10

Keywords: AIC; BIC; Kullback–Leibler divergence; Noncentral Chi-square; Power

1. Introduction

Several methods have been developed to determine which of two nested models, f_1 and f_2 with f_1 nested in f_2, conform better to a set of data. Many of the general applicable methods such as likelihood-ratio (LR) testing, Akaikean information criterion (AIC) and Bayesian information criterion (BIC) all use the same test statistic

$$T = 2 \times (\ln(L_2) - \ln(L_1)), \tag{1}$$

where L_1 and L_2 are the maximum of the likelihood functions of f_1 and f_2, respectively. If the simpler nested model f_1 is true, the distribution of T converges for a large number of observations, n, to the chi-square-distribution with v degrees of freedom (d.f.) (see for instance, Cox and Hinkley, 1974) with $v = k_2 - k_1$, where k_1 and k_2 are the number of parameters of functions f_1 and f_2, respectively. This holds as long as a set of not very restrictive regularity conditions are met.

For power analysis, the distribution of T should be known if the extended model (f_2) is true. In this paper, a general solution for the distribution of T will be given. For this general solution, it is assumed that the functions f_1 and f_2 are locally linear in their parameters and that the deviance between the true model (f_2) and the simpler model (f_1) is small.

In many biological problems, the number of observations which can be obtained under constant conditions is limited. Therefore, experiments are often repeated several times within a study, each experiment having its own conditions. If models are fitted to such sets of experiments, it will often be necessary to estimate the model parameters for each experiment separately. The consequence of the distribution of T for the LR test, the AIC and the BIC are explored. The statistical results are illustrated by a simulation study on some nonlinear models for binary data.

2. Three methods to compare two nested models

2.1. The LR test

The LR test is a general hypothesis testing method to evaluate a simpler model (f_1) against a more complex alternative (f_2). This test uses the ratio of the maximum likelihood (ML) for the simple model and the ML for the more complex model. If f_1 is true the difference between twice the log of the ML's, T, is approximately χ^2 distributed. This means that for large numbers of observations the α-critical value for T is approximately $\chi^2_{\alpha,v}$. The d.f. v is

Download English Version:

https://daneshyari.com/en/article/10524929

Download Persian Version:

https://daneshyari.com/article/10524929

Daneshyari.com