Statistics
in Medicine

Research Article

Received 3 February 2010, Accepted 31 May 2011 Published online 29 July 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4323

Bayesian information criterion for longitudinal and clustered data

Richard H. Jones*[†]

When a number of models are fit to the same data set, one method of choosing the 'best' model is to select the model for which Akaike's information criterion (AIC) is lowest. AIC applies when maximum likelihood is used to estimate the unknown parameters in the model. The value of -2 log likelihood for each model fit is penalized by adding twice the number of estimated parameters. The number of estimated parameters includes both the linear parameters and parameters in the covariance structure. Another criterion for model selection is the Bayesian information criterion (BIC). BIC penalizes -2 log likelihood by adding the number of estimated parameters multiplied by the log of the sample size. For large sample sizes, BIC penalizes -2 log likelihood much more than AIC making it harder to enter new parameters into the model. An assumption in BIC is that the observations are independent. In mixed models, the observations are not independent. This paper develops a method for calculating the 'effective sample size' for mixed models based on Fisher's information. The effective sample size replaces the sample size in BIC and can vary from the number of subjects to the number of observations. A number of error models are considered based on a general mixed model including unstructured, compound symmetry (random intercept), first-order autoregression with observational error and random intercept and slope. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: AIC; effective sample size; information criteria; maximum likelihood; mixed models

1. Introduction

An overall, likelihood-based, model selection procedure, which is quite helpful in many applications is Akaike's information criterion (AIC) [1–5]. AIC is based on information and decision theories and attempts to prevent overparameterization of a model. AIC penalizes -2 ln likelihood, (ℓ), for the number of parameters fit to the data to avoid over fitting:

$$\text{AIC} = \ell + 2\ (\text{number of estimated parameters}).$$

The number of estimated parameters included both parameters in the linear model and parameters in the covariance structure. For every model under consideration, AIC is calculated and the model that has the lowest value of AIC is selected as the 'best' model. All models must be fit to the same outcome variables.

Akaike's information criterion has received some criticism in the time series analysis literature because it is not a consistent estimate of the order of an autoregression. If an autoregression has true order p and the number of observations on the time series goes to infinity, the AIC does not always select order p. Sometimes, it chooses too large an order.

Akaike [6] and Schwarz [7] independently developed a Bayesian information criterion for model selection, now referred to as BIC (and sometimes referred to as SC or SIC for Schwarz information criterion). For sample sizes of eight or more, BIC has a higher penalty for overfitting compared with AIC,

$$\text{BIC} = \ell + (\ln n_{\text{T}})\ (\text{number of estimated parameters}), \tag{1}$$

Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 13001 East 17th Place, Mail Stop B119, Aurora, CO 80045, USA

**Correspondence to: Richard H. Jones, Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 13001 East 17th Place, Mail Stop B119, Aurora, CO 80045, USA.*

[†] *E-mail: richard.jones@UCDenver.edu*

where n_T is the total number of observations assumed to be independent. For clustered or longitudinal data, the observations are usually not independent and a modification of n_T is necessary. This modification will be denoted by n_e for ‘effective sample size’.

Today’s computers can and do handle very large data sets. A problem with very large sample sizes using conventional Neyman–Pearson methods is that very small departures from the null hypothesis can be significant, even though these small departures are too small to be of practical importance. In fact, as Kass and Raftery [8] discussed, the null hypothesis can be rejected even when the evidence in the data favors the null hypothesis. They discussed the Bayes factor, which is the posterior odds of the null hypothesis when the prior probability on the null is one-half. They stated that BIC gives a rough approximation to the logarithm of the Bayes factor. BIC should be seriously considered for model selection with very large sample sizes. Even for moderate sample sizes, BIC should be considered because it penalizes -2 log likelihood more than AIC.

Longitudinal and clustered data pose the problem of obtaining the appropriate value of n_T in Equation (1). In SAS (SAS Institute Inc., Cary, NC, USA) PROC MIXED [9], if the model has independent errors (no random or repeated statements), SAS correctly uses for n_T the total number of observations on all subjects. However, if there is a random or repeated statement in the model, SAS uses the number of subjects or clusters, which in the notation here is m instead of n_T. This is a very conservative approach. The effective sample size, n_e, would usually be between these extremes. SAS PROC MIXED has both the maximum likelihood option (ML) and the restricted ML option (REML) with REML being the default. REML is applicable for comparing models only when the fixed effects remain the same. In this paper, I allow the fixed effects to vary during the optimization and use ML. The error models considered here have been discussed by Jennrich and Schlucher [10], Diggle [11], and Jones and Boadi-Boateng [12].

2. Effective sample size for Bayesian information criterion

A general linear mixed model with Gaussian errors for subject i is [13, 14]

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where $\mathbf{y}_i$ is a column vector of length n_i of the response variables for subject or cluster i. $\mathbf{X}_i$ is an $n_i \times p$ matrix of known or observed independent variables, and $\boldsymbol{\beta}$ is a column vector of length p of the regression coefficients of these fixed effects to be estimated. $\mathbf{Z}_i$ is an $n_i \times q$ matrix for the random effects, $\boldsymbol{\gamma}_i$, which are assumed to be independently distributed across subjects with distribution $\boldsymbol{\gamma}_i \sim N(0, \mathbf{G})$. The error vector, $\boldsymbol{\epsilon}_i$, is assumed to be independent across subjects with distribution $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{R}_i)$. The covariance matrix of the response variable, $\mathbf{y}_i$, is

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i.$$

I assume that the first column of the $\mathbf{X}_i$ matrices are columns of 1s. This means that the intercept is included in the model. The $\mathbf{X}_i$ matrices can be partitioned into the first column of 1s and the $p-1$ other columns,

$$\mathbf{X}_i\boldsymbol{\beta} = \begin{bmatrix} \mathbf{1}_i & \mathbf{X}_i^* \end{bmatrix} \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}^* \end{bmatrix}.$$

Here, $\mathbf{1}_i$ is a column vector of 1s of length n_i for subject i, $\mathbf{X}_i^*$ is $\mathbf{X}_i$ with the first column removed, and $\boldsymbol{\beta}^*$ is $\boldsymbol{\beta}$ with the intercept removed. The normal equations to be solved to obtain the weighted least squares estimate of $\boldsymbol{\beta}$ are

$$\left(\sum_{i=1}^{m}\begin{bmatrix} \mathbf{1}_i'\mathbf{V}_i^{-1}\mathbf{1}_i & \mathbf{1}_i'\mathbf{V}_i^{-1}\mathbf{X}_i^* \\ (\mathbf{X}_i^*)'\mathbf{V}_i^{-1}\mathbf{1}_i & (\mathbf{X}_i^*)'\mathbf{V}_i^{-1}\mathbf{X}_i^* \end{bmatrix}\right)\begin{bmatrix} b_0 \\ \mathbf{b} \end{bmatrix} = \left(\sum_{i=1}^{m}\begin{bmatrix} \mathbf{1}_i'\mathbf{V}_i^{-1}\mathbf{y}_i \\ (\mathbf{X}_i^*)'\mathbf{V}_i^{-1}\mathbf{y}_i \end{bmatrix}\right).$$

The upper left-hand corner of the matrix on the left is Fisher’s information for the intercept [15, pp. 72 and 327]. This is the sum of the elements of $\mathbf{V}_i^{-1}$ summed over subjects. This information measure depends on the units used for the observations, that is, meters or feet. The use of the correlation matrix corresponding to the $\mathbf{V}$ matrix resolves this problem of different units of measurements. Let $\mathbf{C}_i$ be the correlation matrix corresponding to $\mathbf{V}_i$. Consider the $\mathbf{V}$ matrix for a single subject. Let the square roots of the diagonal elements of $\mathbf{V}$ be

$$\sigma_j = \sqrt{\mathbf{V}_{jj}}.$$

Divide the element j, k of the $\mathbf{V}$ matrix by $\sigma_j \sigma_k$. This produces the correlation matrix, $\mathbf{C}$, corresponding to the covariance matrix $\mathbf{V}$. The unit free measure of information uses the correlation matrix rather than the covariance matrix, and the effective sample size can be defined to be

$$n_{\mathrm{e}} = \sum_{i=1}^{m} \mathbf{1}_i' \mathbf{C}_i^{-1} \mathbf{1}_i. \tag{3}$$

This is the sum of the elements of $\mathbf{C}_i^{-1}$, summed over subjects.

If $\mathbf{C}_i$ is the identity matrix, $\mathbf{I}_i$, of size n_i by n_i for subject i,

$$\mathbf{1}_i' \mathbf{C}_i^{-1} \mathbf{1}_i = n_i,$$

that is, if the errors are uncorrelated with constant variances, the effective sample sizes are the actual sample sizes. I show in the following text that Equation (3) is a reasonable definition for a number of commonly used error models.

To compute BIC for a given model, first obtain the ML estimates of the linear parameters and unknown covariance parameters simultaneously. Next, calculate the effective sample size, n_{e} from Equation (3) and BIC from

$$BIC = \ell + (\ln n_{\mathrm{e}})\ (\text{number of estimated parameters}).$$

2.1. *Compound symmetry*

For compound symmetry (random intercept), the correlation matrix for subject i is n_i by n_i

$$\mathbf{C}_i = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & & \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}, \tag{4}$$

where $0 \leqslant \rho < 1$ is the intraclass correlation. The inverse of this matrix has

$$\frac{1}{1-\rho}\left[\frac{1+(n_i-2)\rho}{1+(n_i-1)\rho}\right]$$

on the main diagonal and

$$-\frac{1}{1-\rho}\left[\frac{\rho}{1+(n_i-1)\rho}\right]$$

on the other elements of the matrix [14, p. 17]. The sum of the elements of this inverse matrix is n_i times the diagonal value plus $n_i(n_i-1)$ times the off-diagonal value giving

$$n_{\mathrm{e}} = \sum_{i=1}^{m} \left(\frac{n_i}{1+(n_i-1)\rho}\right). \tag{5}$$

In the balanced case, where every subject or cluster has the same number of observations, n,

$$n_{\mathrm{e}} = \frac{mn}{1+(n-1)\rho}. \tag{6}$$

If $\rho = 0$, there is no correlation between the observations on a subject or cluster. In this case, SAS is correct and $n_{\mathrm{e}} = n_{\mathrm{T}}$ is the total number of observations on all subjects. At the other extreme, if ρ is close to 1, little additional information is gained by having repeated measures on subjects. In this case, SAS is also correct that the n_{e} in BIC is the number of subjects, m. Between these two extremes, the effective sample size to be used for BIC is a function of ρ. In practice, these equations are based on the ML estimate of ρ that is available from the computer output.

2.2. *First-order autoregression within subject errors*

In longitudinal data analysis when subjects are followed over time, there is a natural ordering of the data for each subject. In this situation, a first-order autoregression (AR(1)) is often used to model the within-subject error structure. I will first consider the balanced case with equally spaced observations with m subjects and n observations on each subject. All subjects are observed at the same n time points. The error correlation matrix for each subject is $n \times n$,

$$\mathbf{C} = \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 & \cdots \\ \phi & 1 & \phi & \phi^2 & \\ \phi^2 & \phi & 1 & \phi & \\ \phi^3 & \phi^2 & \phi & 1 & \ddots \\ \vdots & & & \ddots & \ddots \end{bmatrix},$$

where ϕ is the autoregression coefficient in the interval $-1 < \phi < 1$. Siddique [17] showed that $\mathbf{C}^{-1}$ is tridiagonal,

$$\mathbf{C}^{-1} = \frac{1}{1-\phi^2}\begin{bmatrix} 1 & -\phi & 0 & \cdots & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & & 0 & 0 \\ 0 & -\phi & 1+\phi^2 & \ddots & 0 & 0 \\ \vdots & & \ddots & \ddots & \vdots & 0 \\ 0 & 0 & 0 & & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix}.$$

This inverse matrix has a unique lower triangular factorization [14, p. 110], such that

$$\mathbf{C}^{-1} = \mathbf{L}'\mathbf{L},$$

where

$$\mathbf{L} = \frac{1}{\sqrt{1-\phi^2}}\begin{bmatrix} \sqrt{1-\phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & & 0 & 0 \\ 0 & -\phi & 1 & & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix}.$$

The effective sample size from m subjects is

$$n_{\mathrm{e}} = m\mathbf{1}'\mathbf{C}^{-1}\mathbf{1} = m\mathbf{1}'\mathbf{L}'\mathbf{L}\mathbf{1} = m(\mathbf{L}\mathbf{1})'(\mathbf{L}\mathbf{1}).$$

Calculating $\mathbf{L1}$ and the sum of squares of the elements of the resulting vector gives the effective sample size

$$n_{\mathrm{e}} = m\left[1 + (n-1)\frac{1-\phi}{1+\phi}\right]. \tag{7}$$

If $\phi = 0$, $n_{\mathrm{e}} = mn$, the total number of observations. If ϕ approaches 1, n_{e} approaches m, the number of subjects or clusters. In practice, the ML estimate of ϕ from the computer output will be used in this equation.

2.3. *Unequally spaced first-order autoregression error models*

Unbalanced designs can be caused by missing observations or data that are collected unequally spaced. For truly unequally spaced data, not equally spaced observations with missing values, a continuous time AR(1) model must be used [14]). A continuous time AR(1) has a negative exponential correlation function that depends on the time between the observations. Suppose the observation j on subject i is taken at time t_{ij}. Then, the correlation between $\epsilon(t_{ij})$ and $\epsilon(t_{ij'})$ is

$$\phi(|t_{ij} - t_{ij'}|) = \exp\{-\alpha(|t_{ij} - t_{ij'}|)\},$$

where $\alpha > 0$ is the continuous time autoregression coefficient. Let the time interval between observations for subject i be

$$\delta_{ij} = t_{ij} - t_{i,j-1}, \qquad \text{for} \quad j = 2, 3 \cdots n_i,$$

where n_i is the number of observations on subject i. Now

$$\phi(\delta_{ij}) = e^{-\alpha\delta_{ij}}.$$

The factored form for the inverse of the correlation matrix for subject i is [14]

$$\mathbf{L}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \frac{-\phi(\delta_{i2})}{\sqrt{1-\phi^2(\delta_{i2})}} & \frac{1}{\sqrt{1-\phi^2(\delta_{i2})}} & 0 & & 0 & 0 \\ 0 & \frac{-\phi(\delta_{i3})}{\sqrt{1-\phi^2(\delta_{i3})}} & \frac{1}{\sqrt{1-\phi^2(\delta_{i3})}} & & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & & \frac{1}{\sqrt{1-\phi^2(\delta_{i,n-1})}} & 0 \\ 0 & 0 & 0 & \cdots & \frac{-\phi(\delta_{i,n})}{\sqrt{1-\phi^2(\delta_{i,n})}} & \frac{1}{\sqrt{1-\phi^2(\delta_{i,n})}} \end{bmatrix}.$$

Calculating $\mathbf{L}_i\mathbf{1}$ and the sum of squares of the elements of the resulting vector and summing over subjects gives the effective sample size

$$n_{\mathrm{e}} = \sum_{i=1}^{m}\left[1 + \sum_{j=2}^{n_i}\frac{1-\phi(\delta_{ij})}{1+\phi(\delta_{ij})}\right]. \tag{8}$$

This can be calculated using the ML estimate of α. For balanced equally spaced data, this reduces to Equation (7).

2.4. *Including observational error*

Random observational error adds a positive constant, the variance of the observational error, to the diagonal elements of the covariance matrices for AR(1) errors. In geophysics, this is often called the nugget effect. With the addition of the observational error variance to the diagonal elements of the covariance matrix, the inverse is no longer tridiagonal. The effective sample size can be calculated by the direct evaluation of Equation (3). One way of doing this is to augment each $\mathbf{C}_i$ by a column of ones, $\mathbf{1}$, and do a Cholesky factorization on the augmented matrix [18]. Let

$$\mathbf{C}_i = \mathbf{T}_i'\mathbf{T}_i$$

be the factored correlation matrix for subject i. The augmented correlation matrix is

$$\begin{bmatrix} \mathbf{C}_i & \mathbf{1} \end{bmatrix}.$$

After the in-place factorization, the matrix becomes

$$\begin{bmatrix} \mathbf{T}_i & (\mathbf{T}_i')^{-1}\mathbf{1} \end{bmatrix}.$$

Now, n_e for subject i is the sum of squares of the right most column of this matrix. If the data are balanced, that is, every subject is observed at the same times, n_e can be multiplied by the number of subjects. If the data are unbalanced, n_e needs to be calculated for each $\mathbf{C}_i$ and summed across subjects. This is a general method for calculating the effective sample size from the correlation matrices of each subject or cluster.

2.5. *Random intercept and slope models*

Let the $\mathbf{X}_i$ matrix for subject i be

$$\mathbf{X}_i = \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix},$$

where the x_{ij} are values of a longitudinal variable such as age or time that are ordered from smallest to largest. The form of the general linear mixed model (2) for this application is

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{X}_i \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i.$$

Here, the error vector, $\boldsymbol{\epsilon}_i$, is assumed to be independent across subjects with distribution $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \mathbf{I}_i)$. The covariance matrix of $\mathbf{y}_i$ is

$$\mathbf{V}_i = \mathbf{X}_i \mathbf{G} \mathbf{X}_i' + \sigma^2 \mathbf{I}_i$$

where $\mathbf{G}$ is a general 2×2 covariance matrix.

Standard software can be used to obtain the ML estimates of σ^2 and the three distinct elements of $\mathbf{G}$. $\mathbf{V}_i$ can then be calculated and reduced to its correlation matrix, $\mathbf{C}_i$, and the effective sample size, n_e, calculate using Equation (3).

3. Discussion

Bayesian information criterion has a very important role to play in model selection in data sets with a large sample size. In conventional Neyman–Pearson statistics with very large sample sizes, small changes that are too small to be of practical importance may be significant. BIC penalizes the value of -2 log likelihood by the log of the sample size multiplied by the number of estimated parameters. The number of estimated parameters includes both the fixed effect regression coefficients and the error model parameters. The assumption is that the observations are independent.

In mixed models with subjects or clusters, the observations within subjects or clusters are usually correlated. This paper develops an effective sample size, n_e based on Fisher's information correlation matrix. For compound symmetry, the result is well known and unequal cluster sizes are easily handled by calculating n_e for each subject and summing over subjects. Formulas are derived for AR(1) errors, balanced (every subject observed at the same times) or unbalanced designs including missing and unequally spaced observation. For more complicated models, such as AR(1) with observational error and models with a random intercept and slope, a general method is used based on the sum of the elements of the inverse of each subject's correlation matrix.

References

1. Akaike H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Petrov BN, Csaki F (eds). Akademia Kaido: Budapest, 1973; 267–281.
2. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723.
3. Jones RH. Identification and autoregressive spectrum estimation. *IEEE Transactions on Automatic Control* 1974; **19**:894–897.
4. Sakamoto Y, Ishiguro M, Kitagawa G. *Akaike Information Criterion Statistics*. Kluwer Academic Publishers: Dordrecht, Holland, 1986.
5. Bozdogan H. Model selection and Akaike's information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 1987; **52**:345–370.
6. Akaike H. A new look at the Bayes procedure. *Biometrika* 1978; **65**:53–59.

7. Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 1978; **6**:461–464.
8. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistics Association* 1995; **90**:773–795.
9. SAS Institute Inc. *SAS/STAT 9.1 User's Guide*, Cary, NC 2004.
10. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariances matrices. *Biometrics* 1986; **42**:805–820.
11. Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics* 1988; **44**:959–971.
12. Jones RH, Boadi-Boateng F. Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* 1991; **47**:161–175.
13. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
14. Jones RH. *Longitudinal Data with Serial Correlation: A State-Space Approach*. Chapman & Hall/CRC: London and Boca Raton, Florida, 1993.
15. McCullagh P, Nelder JA. *Generalized Linear Models*, 2nd edition. London and Boca Raton, Florida: Chapman & Hall/CRC, 1989.
16. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* 1981; **114**:906–914.
17. Siddiqui MM. On the inversion of the sample covariance matrix in a stationary autoregressive process. *Annals of Mathematical Statistics* 1958; **29**:585–588.
18. Graybill FA. *Theory and Application of the Linear Model*. Duxbury Press: North Scituate, Massachusetts, 1976.