A Note on Conditional AIC for Linear Mixed-Effects Models
Author(s): Hua Liang, Hulin Wu and Guohua Zou
Source: *Biometrika*, Vol. 95, No. 3 (Sep., 2008), pp. 773-778
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/20441501
Accessed: 22-05-2018 14:17 UTC

Biometrika (2008), **95**, 3, *pp.* 773–778
© 2008 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asn023
Advance Access publication 12 August 2008

Miscellanea

A note on conditional AIC for linear mixed-effects models

BY HUA LIANG, HULIN WU

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, U.S.A.

hliang@bst.rochester.edu hwu@bst.rochester.edu

AND GUOHUA ZOU

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China

guohua_zou@urmc.rochester.edu

SUMMARY

The conventional model selection criterion, the Akaike information criterion, AIC, has been applied to choose candidate models in mixed-effects models by the consideration of marginal likelihood. Vaida & Blanchard (2005) demonstrated that such a marginal AIC and its small sample correction are inappropriate when the research focus is on clusters. Correspondingly, these authors suggested the use of conditional AIC. Their conditional AIC is derived under the assumption that the variance-covariance matrix or scaled variance-covariance matrix of random effects is known. This note provides a general conditional AIC but without these strong assumptions. Simulation studies show that the proposed method is promising.

Some key words: Akaike information criterion; Conditional likelihood; Longitudinal data; Marginal likelihood; Mixed-effects model; Model selection.

1. INTRODUCTION

Linear mixed-effects models (Laird & Ware, 1982) provide a powerful tool for the analysis of longitudinal data, because they can incorporate within-cluster and between-cluster variation. Statistical estimation and inference for these models have been widely studied (Vonesh & Chinchilli, 1996; Pinheiro & Bates, 2000; Verbeke & Molenberghs, 2000). The fundamental question of model selection seems to have been disregarded, however. Traditional selection criteria such as the Akaike information criterion (AIC) (Akaike, 1973) and the Bayes information criterion (BIC) (Schwarz, 1978) for cross-sectional data have been applied without justification in the selection of linear mixed-effects models (Pinheiro & Bates, 2000; Ngo & Brand, 2002). This deficiency was recently noticed by Vaida & Blanchard (2005), who explained why, when the researchers' focus is on clusters instead of the population, AIC and its small sample correction AIC_c are not appropriate, and suggested the conditional Akaike information and the corresponding model selection criterion, conditional AIC. However, in deriving the conditional AIC, they required that the variance-covariance matrix of random effects should be known when the variance of the measurement error term is known, or the scaled variance-covariance matrix of random effects should be known when the variance of the measurement error term is unknown. These requirements may limit the use of the conditional AIC. The objective of this note is to remove Vaida and Blanchard's assumptions and to provide more general conditional AICs. This note reports the results for the case of known error variance. A discussion of the

case of unknown error variance can be found in a University of Rochester technical report by the authors, available from the first author upon request.

2. GENERAL CONDITIONAL AIC FOR LINEAR MIXED-EFFECTS MODELS

We assume the data from m clusters to be modelled by the linear mixed-effects model

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad i = 1, \ldots, m, \tag{1}$$

where y_i is an $n_i \times 1$ vector of observations for cluster i, β is a $p \times 1$ vector of fixed effects, b_i is a $q \times 1$ vector of random effects for cluster i, X_i and Z_i are the $n_i \times p$ and $n_i \times q$ design matrices of full column rank for the fixed and random effects, respectively, and ε_i is the disturbance. We assume that b_i and ε_i are independent and normally distributed with zero means and covariance matrices G_0 and $\sigma^2 I_{n_i}$, respectively, where I_{n_i} is the $n_i \times n_i$ identity matrix. Let $N = \sum_{i=1}^{m} n_i$ be the total number of observations, and let θ be the vector of parameters in the model, including β, σ^2 and the parameters in G_0. Model (1) can be written as

$$y = X\beta + Zb + \varepsilon, \quad b \sim N(0, G), \tag{2}$$

where $y = (y_1^{\mathrm{T}}, \ldots, y_m^{\mathrm{T}})^{\mathrm{T}}$ is an $N \times 1$ vector of observations, $X = (X_1^{\mathrm{T}}, \ldots, X_m^{\mathrm{T}})^{\mathrm{T}}$ is an $N \times p$ matrix of rank p, $Z = \mathrm{diag}(Z_1, \ldots, Z_m)$ is an $N \times r$ block-diagonal matrix of rank $r = mq$, $b = (b_1^{\mathrm{T}}, \ldots, b_m^{\mathrm{T}})^{\mathrm{T}}$, $\varepsilon = (\varepsilon_1^{\mathrm{T}}, \ldots, \varepsilon_m^{\mathrm{T}})^{\mathrm{T}}$ and $G = \mathrm{diag}(G_0, \ldots, G_0)$ is an $r \times r$ block-diagonal matrix. Denote the joint density function of y and b under model (2) by $g(y, b \mid \theta)$. Thus, given b, the conditional likelihood is $g(y \mid \theta, b)$ and the marginal likelihood is $g(y \mid \theta) = \int g(y, b \mid \theta) db$. Let the true conditional distribution of y be $f(y \mid u)$, where u is the true random-effects vector with distribution $p(u)$, and let $f(y, u)$ be the joint density of y and u. Then Vaida & Blanchard (2005) defined the conditional Akaike information as follows.

DEFINITION. *The conditional Akaike information is defined to be*

$$\begin{aligned} \mathrm{cAI} &= -2E_{f(y,u)}\big(E_{f(y^*|u)}\big[\log g\{y^* \mid \hat\theta(y), \hat b(y)\}\big]\big) \\ &= -2\int \log g\{y^* \mid \hat\theta(y), \hat b(y)\} f(y^* \mid u) f(y, u) dy^* dy du, \end{aligned} \tag{3}$$

where y^* is the prediction dataset which is independent of y conditional on u, and from the same distribution $f(\cdot \mid u)$ as y, and $\hat\theta(y)$ and $\hat b(y)$ are the estimators of θ and b, respectively.

The following theorem derives an unbiased estimator of cAI when the variance σ^2 is known. The proof is given in the Appendix. Let $\hat\theta(y)$ and $\hat b(y)$ be the maximum likelihood and the empirical Bayes estimators of θ and b, respectively.

THEOREM 1. *Assume that the data y have true density $f(y \mid u) = g(y \mid \theta_0, u)$ for some θ_0 and some random effect u with distribution $p(u)$. Let the data be modelled by (2), with densities denoted by $g(y \mid \theta, b)$ and $p(b)$. If σ^2 is known, then an unbiased estimator of* cAI *in* (3) *is*

$$\mathrm{cAIC} = -2\log g\{y \mid \hat\theta(y), \hat b(y)\} + 2\Phi_0(y), \tag{4}$$

where $\Phi_0(y) = \sum_{i=1}^{N} \partial\hat y_i / \partial y_i = tr(\partial \hat y^{\mathrm{T}} / \partial y)$, and y_i and $\hat y_i$ are the ith components of y and the fitted vector $\hat y = X\hat\beta + Z\hat b$, respectively.

The expectation of $\Phi_0(y)$ in (4), conditional on u, is the generalized degrees of freedom defined by Ye (1998) for linear mixed-effects models. Ye (1998) used the generalized degrees of freedom as a penalty term which, unlike $\Phi_0(y)$, generally depends on the unknown parameters. From (4), we see that, unlike for linear fixed-effects models, the penalty term depends on the observed data y for linear mixed-effects models. The calculation of the penalty function $\Phi_0(y)$ involves the first partial derivatives $\partial\hat y_i / \partial y_i$, $i = 1, \ldots, N$, that can be calculated directly or approximated numerically by

$\{\hat{y}_i(y + he_i) - \hat{y}_i(y)\}/h$, where h is a small number and e_i is the $N \times 1$ vector, with ith component equal to 1 and other components equal to 0.

Remark 1. Assuming that σ^2 is known, Vaida & Blanchard (2005) developed in their Theorem 1 a result for the case of known G. Our Theorem 1 generalizes their result and provides an unbiased estimator of cAI for the case of unknown G, thereby solving a problem proposed by Vaida & Blanchard (2005).

COROLLARY 1 (Vaida & Blanchard, 2005). *Under the assumptions of Theorem* 1, *further assume that G is known. Then an unbiased estimator of* cAI *is*

$$\text{cAIC} = -2 \log g\{y \mid \hat{\theta}(y), \hat{b}(y)\} + 2\rho,$$

where $\rho = \text{tr}(H_1)$ is the 'effective degrees of freedom' of Hodges & Sargent (2001), *and H_1 is the 'hat' matrix that maps the observed data vector y into the fitted vector $\hat{y}$, that is, $\hat{y} = H_1 y$.*

Proof. See the Appendix.

An intuitive explanation of ρ, the penalty term when both σ^2 and G are known, can be provided as follows. From the definition of H_1, see the proof of Corollary 1, it can be shown that

$$\rho = p + \sum_{i=1}^{r_0} \frac{\lambda_i}{1 + \lambda_i}, \tag{5}$$

where $\lambda_1, \ldots, \lambda_{r_0}$ are the nonzero eigenvalues of the matrix $D^{1/2} Z^{\mathrm{T}} (I - P_X) Z D^{1/2}$ with $D = \sigma^{-2} G$ and $P_X = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. In Corollary 1 only β is unknown, so the first term on the right-hand side of (5) is the total number of parameters in the linear mixed-effects model. Therefore, unlike for the usual linear fixed-effects model, the penalty term is not just the number of unknown parameters for the model. The second term in the expression for ρ is the extra penalty due to random effects. This term is smaller than the number of random effects, r, showing that the extra penalty is not the number of random-effects terms, although the random effects may be independent; the covariate matrix Z in model (2) need not be block diagonal. Furthermore, when G is unknown, Vaida & Blanchard (2005) suggested the use of the observed $\hat{\rho} = \text{tr}\{H_1(\hat{G})\}$, where $\hat{G}$ is the maximum likelihood estimator of G. Observe that, when G is unknown, we have $\hat{y} = H_1(\hat{G})y$. From Theorem 1, therefore, the exact penalty term when G is unknown is

$$\Phi_0(y) = \hat{\rho} + 1_N^{\mathrm{T}} H(\hat{G}) y, \tag{6}$$

where 1_N is the $N \times 1$ vector of ones, and

$$H(\hat{G}) = \begin{pmatrix} \frac{\partial h_{11}(\hat{G})}{\partial y_1} & \cdots & \frac{\partial h_{1N}(\hat{G})}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{N1}(\hat{G})}{\partial y_N} & \cdots & \frac{\partial h_{NN}(\hat{G})}{\partial y_N} \end{pmatrix},$$

in which $h_{ij}(\hat{G})$ is the (i, j)th element of the matrix $H_1(\hat{G})$; here we write H as a function of $\hat{G}$, but it may depend on y not only through $\hat{G}$. The second term in (6) is the additional penalty due to the variability of estimating unknown G.

Remark 2. In Theorem 1 and Corollary 1, the assumption that $f(y \mid u) = g(y \mid \theta_0, u)$ means that the true model is included in the candidate model family. This is a traditional assumption in deriving model selection criteria (Akaike, 1973; Hurvich & Tsai, 1989; Burnham & Anderson, 1998; Hurvich et al., 1998). The further assumption in Corollary 1 that G is known implies that the covariate matrices for random effects under the true and candidate models are equal. The removal of this further assumption shows that the covariate matrices for random effects under the true and candidate models can be different. Furthermore, in the proof of Theorem 1, the expression $X\beta_0 + Zu$ for μ, where β_0 is the true parameter for fixed effects, is not useful. This means that even the assumption that the candidate models include the true model can be removed. For example, Theorem 1 and Corollary 1 still hold if the data y come from a

Table 1. *Simulation study. Comparison of bias correction,* BC, *and its two estimators,* $\hat{\rho}$ *and* $\Phi_0(y)$ *based on* 500 *runs*

n_i	σ	BC	$\hat{\rho}$	$\Phi_0(y)$
6	0·0705	19·549	19·994	19·380
26	0·0705	19·875	19·999	19·837
51	0·0705	19·926	19·999	19·891
6	0·141	17·638	19·731	18·253
26	0·141	19·339	19·976	19·355
51	0·141	19·547	19·986	19·597
6	0·282	15·818	16·944	15·436
26	0·282	17·832	19·265	17·927
51	0·282	18·723	19·763	18·648

linear mixed-effects model mentioned in Vaida & Blanchard (2005), defined by $y = P\alpha + Qv + e$ with $v \sim N(0, S)$, $e \sim N(0, \sigma_0^2 I_N)$ and P and Q containing covariates different from X and Z.

3. SIMULATION STUDY

In this section, we present simulation results to study the behaviour of the proposed method under small and moderate sample sizes. To make a comparison, we generate data from the model used by Vaida & Blanchard (2005); that is,

$$y_{ij} = (\beta_0 + \beta_1 t_j) + (b_{0i} + b_{1i} t_j) + \varepsilon_{ij}, \quad i = 1, \ldots, m = 10, \quad j = 1, \ldots, n_i,$$

where $\beta_0 = -2{\cdot}78$, $\beta_1 = -0{\cdot}186$, $t_j = 5j$, $(b_{0i}, b_{1i})^{\mathrm{T}}$ follows a normal distribution with the mean zero and variance-covariance matrix

$$\begin{pmatrix} 0{\cdot}0367 & -0{\cdot}00126 \\ -0{\cdot}00126 & 0{\cdot}00279 \end{pmatrix},$$

and ε_{ij} are independent and identically distributed with the distribution $N(0, \sigma^2)$. In our simulation experiments, as in Vaida & Blanchard (2005), we consider $\sigma = 0{\cdot}0705$, $0{\cdot}141$ and $0{\cdot}282$ and implement the following three scenarios: (i) $j = 0, 1, \ldots, 5$, giving $n_i = 6$; (ii) $j = 0, 1, \ldots, 25$, giving $n_i = 26$; and (iii) $j = 0, 1, \ldots, 50$, giving $n_i = 51$. For each of the nine configurations, 500 independent sets of data are generated. Assuming that the covariance matrix G is unstructured, we mainly compare the estimates of the bias correction, BC, which is defined through $\mathrm{cAI} = E_{f(y,u)}\{-2 \log g(y \mid \hat{\theta}, \hat{b})\} + 2\mathrm{BC}$, based on our proposed method, $\Phi_0(y)$, and Vaida & Blanchard's (2005) method, $\hat{\rho}$, with the true bias-correction values. We take $h = 0{\cdot}0001$ in the calculation of $\Phi_0(y)$.

The results, shown in Table 1, are in accordance with the theory. The estimated values based on the proposed method and Vaida & Blanchard's (2005) method are both close to the true bias-correction values, especially for the small values of error variance. In general, the larger is the sample size, the closer are the values. However, the estimated values based on the new method are consistently closer to the true values than those based on Vaida & Blanchard's (2005) method, especially if the error variance is large or n_i is small.

4. CONCLUDING REMARKS

The derivation of cAIC does not require the assumption that the candidate models include the true model. This means that, when the error variance σ^2 is known, this traditional assumption is not necessary for deriving a reasonable model selection criterion. Further analysis shows that this conclusion is still true even if the error variances under the true and candidate models are unknown but equal. Also, the assumption that the true model is included in the candidate model family is needed only in the derivation of the estimator of the true error variance, as shown in our technical report. Given that the error variance is a nuisance

parameter, this explains in part why the commonly used AIC and AIC_c in fixed-effects models often perform well even though the candidate model family does not include the true model, although these selection criteria were derived under the above assumption.

In contrast to derivations in the model selection literature, we made use of the integration-by-parts technique, which has been used to obtain risk-unbiased estimators before (Stein, 1981; Lu & Berger, 1989) in the derivation of selection criterion for linear mixed-effects models. This method can also be applied to obtain marginal AIC based on the marginal likelihood and overall AIC based on the joint likelihood for linear mixed-effects models, as well as AIC_c for nonparametric regression models (Hurvich et al., 1998) and single-index models (Naik & Tsai, 2001). Furthermore, the principle of this note may be extended to deal with generalized mixed-effects models, and can be applied to select smoothing parameters in semiparametric regression.

ACKNOWLEDGEMENT

The authors thank Professor D. M. Titterington and a referee for their constructive comments which greatly improved the original manuscript. The research of each author was partially supported by grants from the U.S. National Institute of Allergy and Infectious Diseases. Zou's research was also partially supported by a grant from the National Natural Science Foundation of China.

APPENDIX

Technical details

Proof of Theorem 1. Let $\mu = X\beta_0 + Zu$, where β_0 is the true parameter for fixed effects. Then it is readily seen that

$$\begin{aligned}\text{cAI} &= -2E_{f(y,u)}\left(E_{f(y^*|u)}\left[\log g\{y^* \mid \hat{\theta}(y), \hat{b}(y)\}\right]\right)\\ &= E_{f(y,u)}\left\{N\log(2\pi\sigma^2) + N + \frac{1}{\sigma^2}(\hat{y}-\mu)^{\mathrm{T}}(\hat{y}-\mu)\right\}.\end{aligned}$$

Also, we have

$$E_{f(y,u)}\{-2\log g(y \mid \hat{\theta}, \hat{b})\} = E_{f(y,u)}\left\{N\log(2\pi\sigma^2) + \frac{1}{\sigma^2}(y-\hat{y})^{\mathrm{T}}(y-\hat{y})\right\}.$$

Thus, after some calculations, we obtain

$$\begin{aligned}\text{cAI} - E_{f(y,u)}\{-2\log g(y \mid \hat{\theta}, \hat{b})\} &= \frac{2}{\sigma^2}E_{f(y,u)}\{(y-\mu)^{\mathrm{T}}\hat{y}\}\\ &= \frac{2}{\sigma^2}E_{p(u)}E_{f(y|u)}\left\{\sum_{i=1}^{N}(y_i-\mu_i)\hat{y}_i\right\}, \qquad \text{(A1)}\end{aligned}$$

where μ_i is the ith component of μ.

Under the true model, for given u, $y \sim N(\mu, \sigma^2 I_N)$. If we assume that $\hat{y}_i$ is a continuous function with piecewise-continuous partial derivatives with respect to y, it can be shown by integration by parts that

$$E_{f(y|u)}\left\{\sum_{i=1}^{N}(y_i-\mu_i)\hat{y}_i\right\} = \sigma^2 E_{f(y|u)}\left(\sum_{i=1}^{N}\frac{\partial\hat{y}_i}{\partial y_i}\right),$$

provided that each expectation on the right exists; see also Stein (1981) and Lu & Berger (1989). Therefore, (A1) becomes

$$\begin{aligned}\text{cAI} - E_{f(y,u)}\{-2\log g(y \mid \hat{\theta}, \hat{b})\} &= 2E_{p(u)}\left\{E_{f(y|u)}\left(\sum_{i=1}^{N}\frac{\partial\hat{y}_i}{\partial y_i}\right)\right\}\\ &= 2E_{f(y,u)}\{\Phi_0(y)\}.\end{aligned}$$

Thus, an unbiased estimator of cAI is given by cAIC in (4), and this completes the proof of Theorem 1. □

Proof of Corollary 1. From Hodges & Sargent (2001) or Vaida & Blanchard (2005), when σ^2 and G are known, the fitted vector is $\hat{y} = X\hat{\beta} + Z\hat{b} = H_1 y$, where $H_1 = (X\ Z)(M^{\mathrm{T}}M)^{-1}(X\ Z)^{\mathrm{T}}$ with

$$M = \begin{pmatrix} X & Z \\ O & -\Delta \end{pmatrix},$$

in which Δ is some $r \times r$ matrix such that $\sigma^{-2}G = (\Delta^{\mathrm{T}}\Delta)^{-1}$. Thus

$$\Phi_0(y) = \mathrm{tr}\left(\frac{\partial \hat{y}^{\mathrm{T}}}{\partial y}\right) = \mathrm{tr}(H_1) = \rho. \qquad \square$$

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Ed. B. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado.

BURNHAM, K. P. & ANDERSON, D. P. (1998). *Model Selection and Inference: A Practical Information-Theoretical Approach*. New York: Springer.

HODGES, J. S. & SARGENT, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367–79.

HURVICH, C. M., SIMONOFF, J. S. & TSAI, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B **60**, 271–93.

HURVICH, C. M. & TSAI, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

LAIRD, N. M. & WARE, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–74.

LU, K. L. & BERGER, J. O. (1989). Estimation of normal means: Frequentist estimation of loss. *Ann. Statist.* **17**, 890–906.

NAIK, P. A. & TSAI, C. L. (2001). Single-index model selections. *Biometrika* **88**, 821–32.

NGO, L. & BRAND, R. (2002). Model selection in linear mixed effects models using SAS Proc Mixed. SAS Global Forum 22.

PINHEIRO, J. C. & BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–51.

VAIDA, F. & BLANCHARD, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351–70.

VERBEKE, G. & MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

VONESH, E. F. & CHINCHILLI, V. M. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

YE, J. M. (1998). On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Assoc.* **93**, 120–31.

[*Received September* 2006. *Revised January* 2008]