Electronic Journal of Statistics
Vol. 6 (2012) 1–9
ISSN: 1935-7524
DOI: 10.1214/12-EJS665

A note on conditional Akaike information for Poisson regression with random effects

Heng Lian

Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University, Singapore 637371, Singapore
e-mail: henglian@ntu.edu.sg

Abstract: A popular model selection approach for generalized linear mixed-effects models is the Akaike information criterion, or AIC. Among others, [7] pointed out the distinction between the marginal and conditional inference depending on the focus of research. The conditional AIC was derived for the linear mixed-effects model which was later generalized by [5]. We show that the similar strategy extends to Poisson regression with random effects, where conditional AIC can be obtained based on our observations. Simulation studies demonstrate the usage of the criterion.

AMS 2000 subject classifications: Primary 62J12.
Keywords and phrases: Akaike information, AIC, model selection, Poisson regression.

Received August 2011.

1. Introduction

Generalized linear models (GLM) are powerful modelling tools that have gained popularity in statistics. It has wide applications in medical studies, pattern classification, sample surveys, etc. The scope of GLM can be greatly expanded by the incorporation of random effects. For example, in typical longitudinal studies, a model with random effects not only models individual characteristics, but attempts to extrapolate to the entire population as well. It takes into account both within cluster and between cluster variations in the study. Model selection in GLM is typically achieved using AIC or BIC combined with step-wise procedures. With fixed-effects models, the definition of AIC is straightforward using the likelihood penalized by a term that depends on the number of parameters. When random effects come into play, it is not entirely clear how the number of parameters in the model should be defined. Based on other previous works such as [1] and [3], [7] made a distinction between marginal and conditional inference and provided a formal definition of conditional Akaike information, cAI, which gives a theoretical justification for some previous approaches. They derived an unbiased estimator of cAI, called conditional AIC or cAIC, when the covariance matrix of random effects is known. [5] derives a more general cAIC that dispenses with such strong assumptions. In [7], the definition of cAI was given for

general mixed-effects models but the unbiased estimator was only derived for linear mixed-effects models. A general approach of getting an unbiased estimator of cAI for generalized linear mixed-effects models (GLMM) seems to be out of reach. In this note, we propose an unbiased estimator of cAI for Poisson regression with random effects. The nature of Poisson regression is very different from the linear model since the responses are discrete. However, it turns out unbiased cAIC exists although it is derived in a different way.

Recently, [2] has extended cAIC to generalized linear and proportional hazards models. Although the class of problems considered by them is much more general, their proposal is based on asymptotic considerations and in particular they show asymptotic unbiasedness of cAIC when both the number of clusters and cluster sizes go to infinity. They also need to assume asymptotic normality of the fixed and random coefficients estimators. Our method is based on finite sample calculations and the cAIC derived is exactly unbiased.

2. Conditional AIC for count data

Suppose we have some count responses $\{y_i\}, i = 1, \ldots, m$ from m clusters that we want to model in relation to covariates X_i and Z_i, with y_i an $n_i \times 1$ vector from cluster i, and X_i,Z_i are $n_i \times p$ and $n_i \times q$ matrices associated with fixed and random effects respectively. We use Poisson GLMM with the canonical link:

$$\begin{aligned} y_i &\sim Pois(\mu_i) \\ \log \mu_i &= X_i\beta + Z_i b_i,\ b_i \sim N(0, G), \end{aligned} \tag{2.1}$$

where β is a $p \times 1$ vector of fixed effects and b_i is a $q \times 1$ vector of random effects following a mean zero Gaussian distribution with unknown covariance matrix G. The total number of observations is thus $N = \sum_{i=1}^{m} n_i$. Let θ be the population parameters in the model, including β and the parameters in G. The marginal likelihood is $g(y|\theta) = \int g(y|b, \theta)g(b|G)db$ where $g(y|b, \theta)$ is the Poisson likelihood conditional on the random effects and $g(b|G)$ is the density of the random effects. Sometimes it is more convenient to represent (2.1) in the condensed form

$$\begin{aligned} y &\sim Pois(\mu) \\ \log \mu &= X\beta + Zb, \end{aligned}$$

where $y = (y_1^T, \ldots, y_m^T)^T$ is an $N \times 1$ vector of count responses, $X = (X_1^T, \ldots, X_m^T)^T$, $Z = \mathrm{diag}(Z_1, \ldots, Z_m)$ and $b = (b_1^T, \ldots, b_m^T)^T$.

In marginal inference, the focus is on the population parameters and the random effects are just a mechanism for modelling the correlations within the clusters. The standard AIC being used refers to this case and is called marginal AIC, mAIC, by [7], defined by $-2 \log g(y|\hat{\theta}(y)) + 2K$, where K is the dimension of θ. This penalty term is there to correct the bias caused by using the same data to estimate θ as well as to evaluate the marginal likelihood $g(y|\hat{\theta})$. The AIC is designed to approximate the Akaike information, $\mathrm{AI} = -2E_{f(y)}E_{f(y*)} \log g(y^*|\hat{\theta})$,

where y^* is an independent replicate of y coming from the same true distribution $f(y)$, which might not be contained within the family defined by (2.1).

In conditional inference, the focus is on the cluster and the estimation of the random effects is of interest. The prediction in this case refers to new responses with the same clusters. Suppose the true distribution of y is $f(y,u) = f(y|u)p(u)$ where u is the true random effects with density $p(u)$. Following [7], the conditional AI is naturally defined as

$$\text{cAI} = -2E_{f(y,u)}E_{f(y^*|u)} \log g\{y^*|\hat{\theta}(y), \hat{b}(y)),$$

where y^* is independent of y, generated from the same conditional distribution $f(\cdot|u)$. Similar to AI, cAI cannot be directly calculated since the true distribution f is unknown. For linear mixed-effects models, unbiased estimators were derived in [7] and [5]. No unbiased estimator has been proposed for other GLMM to our best knowledge. The following theorem gives an unbiased estimator of cAI for Poisson regression, and the proof is given in the Appendix.

Theorem 2.1. *Assume that the count responses have the true distribution* $y \sim Pois(\mu_0)$*, where* $\mu_0 = (\mu_{01}, \ldots, \mu_{0N})^T$ *is the mean of the Poisson distribution and depends on some covariates as well as the random effects* u*. The data are modelled by (2.1) with conditional likelihood denoted by* $g(y|\theta, b)$*. For any estimator* $\hat{\theta}(y)$ *and* $\hat{b}(y)$*, an unbiased estimator of the cAI is*

$$cAIC = -2\log g(y|\hat{\theta}, \hat{b}) + 2K,$$

where K *is given by*

$$\sum_{j=1}^{N} \left\{ y_j \log[\hat{\mu}_j(y)] - y_j \log[\hat{\mu}_j(y^{y_j-1})] \right\}$$

$\hat{\mu}_j(y)$ *is the fitted value of* μ_j *based on data* y*,* $y^{(y_j-1)}$ *is the same as* y *except its* $j-th$ *component is replaced by* $y_j - 1$*, and* $y_j \log[\hat{\mu}_j(y^{(y_j-1)})] = 0$ *when* $y_j = 0$ *by convention.*

Remark 2.1. Although the derivation of the unbiased estimator for cAI is different from the linear model, with the latter derived by integration by parts [5], the results have some resemblance with each other. For linear models, K is given by $\sum_j \partial\hat{\mu}_j/\partial y_j$ and the partial derivatives are estimated by finite difference. Our K for Poisson regression bares the similarity in that it depends on the difference between $\hat{\mu}_j$ and $\hat{\mu}_j(y^{(y_j-1)})$, the fitted responses after perturbing the original observations.

Remark 2.2. In Theorem 2.1, we only need to assume that the true model is in the Poisson family with means depending on some random effects u, which might also be different from the modelled random effects b. Thus the true model does not have to be included in the candidate model family. Besides, we are not assuming anything about the estimators $\hat{\theta}$ and $\hat{b}$ and they can be any reasonable estimators used in the literature.

TABLE 1
Comparison of bias correction BC with its unbiased estimate, K, based on 500 sets of simulated data

n_i	σ_b	BC	K
5	0.125	6.87	6.53
15	0.125	10.35	10.43
5	0.25	9.18	9.09
15	0.25	11.69	11.45
5	0.5	10.19	10.31
15	0.5	11.59	11.14

3. Simulation study

We conducted some simulations to investigate the properties of our unbiased cAIC estimator and demonstrate the difference between marginal and conditional inference. We simulate data from model (2.1) with a random intercept:

$$\log \mu_{ik} = \beta_0 + \beta_1 x_k + b_i,\ i = 1, \ldots, m = 10,\ k = 0, \ldots, n_i,$$

where $\beta_0 = 1, \beta_1 = 0.2, x_k = k$ and $b_i \sim N(0, \sigma_b^2)$. In our simulation, we consider $n_i = 5$ and $n_i = 15$ with $\sigma_b = 0.125, 0.25$ and 0.5. For each of the six specifications, 500 data sets are generated. We compare the cAIC with the true bias, BC, defined by

$$\text{BC} = E_{f(y,u)} \log g(y|\hat{\theta}, \hat{b}) - E_{f(y,u)} E_{f(y^*|u)} \log g(y^*|\hat{\theta}, \hat{b}),$$

which is estimated by simulation with another independent 500 sets of y^*'s generated from the true conditional distribution $f(\cdot|u)$ that shares common random effects u with current responses.

The results are shown in Table 1. The estimated biases are close to the true value. In general, the bias correction (as well as its estimate K) increases with the variance for the random effects. The same comparison can be made for mAIC and also for fixed-effects models, but we found in our simulations that the estimator K in those cases is very close to the number of population parameters and there appears to be no advantages of using our estimator which only increases the computational burden.

To illustrate the differences between marginal and conditional inference, we use the same setup as before with $(\beta_0, \beta_1) = (1, 0.2), n_i = 5, \sigma_b = 0.125, 0.25$ and 0.5. Laplace approximation is used to approximate the marginal likelihood in the calculation of mAIC, for which the bias is simply estimated by the number of population parameters, 3 in this case. Also, a fixed-effects model $\log \mu_{ik} = \beta_0 + \beta_1 x_k$ is fitted to the data and standard AIC is found. The values of AIC, mAIC and cAIC are shown in Figure 1 for different random effect variances. These values are averages over 500 sets of data simulated from the model. By comparing the information criteria, when $\sigma_b = 0.125$, the fixed-effects model is selected for 395 of the 500 data sets when comparing AIC with mAIC, while it is selected only for 3 data sets when comparing AIC with cAIC. When $\sigma_b = 0.25$,

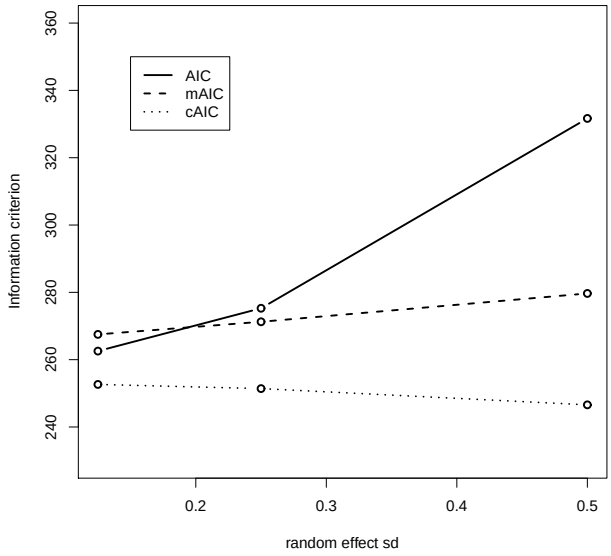

FIG 1. *Comparison of AIC (with fixed effects only), mAIC and cAIC.*

fixed-effects model is selected for 165 of the 500 data sets using mAIC, while it is selected only once when using cAIC.

We use a separate and slightly more complicated simulation to demonstrate the selection of the number of random-effect coefficients using cAIC. The true model is

$$\log \mu_{ik} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})x_{ik1} + (\beta_2 + b_{i2})x_{ik2},\ i = 1, \ldots, m,\ k = 1 \ldots, n,$$

where $(\beta_0, \beta_1, \beta_2) = (1, 0.5, -1), (b_{i0}, b_{i1}) \overset{i.i.d.}{\sim} N(0, 0.5I_2), b_{i2} = 0$, x_{ik1} and x_{ik2} are independently generated from the standard normal distribution. We consider four different models used to fit 100 independently generated data sets. Model 1 is the fixed effects model, Model 2 includes one random effect b_{i0}, Model 3 includes two random effects b_{i0} and b_{i1} and Model 4 includes all three random effects. For each case, we calculated cAI (based on simulation from the true model), cAIC and mAIC. To see the effect of bias correction in cAIC, we also included the values of $-2 \log g(y|\hat{\theta}, \hat{b})$. The average values of different criteria are shown in Table 2 with several different values of (m, n). The bias correction effects for cAIC is obvious. From the Table, for $m = 10$, mAIC will favor fixed effects models over random-effects models with a random intercept (Model 2). When we focus only on the random effects models, both mAIC and cAIC rank the three models (Models 2-4) correctly. Even for $m = 30$ where mAIC can identify the true Model 3 in most cases, values of mAIC for random effects models are much closer to mAIC for fixed effects models than cAIC are. Among the four models, cAIC selected the true model $(72, 68, 89, 90)$

TABLE 2
Comparison of cAIC and mAIC for model selection, based on 100 *sets of simulated data*

Model	cAI	cAIC	$-2\log g(y\|\hat{\theta},\hat{b})$	mAIC
$(m,n)=(10,10)$				
Model 1	540.6	541.7	537.3	543.6
Model 2	435.5	433.5	411.4	557.2
Model 3	377.2	380.9	346.3	438.7
Model 4	379.3	382.4	344.4	442.6
$(m,n)=(10,30)$				
Model 1	2441.9	2445.8	2438.4	2446.4
Model 2	1558.4	1554.8	1530.9	2458.5
Model 3	1168.1	1167.3	1127.0	2033.1
Model 4	1171.4	1170.0	1125.0	2042.7
$(m,n)=(30,10)$				
Model 1	2372.3	2372.9	2360.5	2372.5
Model 2	1510.3	1512.2	1452.2	2177.4
Model 3	1156.6	1156.7	1051.9	1824.8
Model 4	1159.1	1158.2	1048.6	1825.1
$(m,n)=(30,30)$				
Model 1	6904.2	6917.2	6898.9	6916.9
Model 2	4405.1	4418.2	4385.3	6421.7
Model 3	3381.7	3390.7	3277.0	5026.9
Model 4	3384.7	3393.3	3271.9	5031.0

TABLE 3
Computation time of cAIC for the last simulation example in minutes (m) and seconds (s)

	Model 1	Model 2	Model 3	Model 4
(m,n)=(10,10)	1s	51s	1m7s	1m43s
(m,n)=(10,30)	3s	3m2s	4m7s	6m37s
(m,n)=(30,10)	4s	3m54s	3m30s	5m29s
(m,n)=(30,30)	15s	9m23s	13m20s	18m1s

times under the four settings of (m,n) respectively, while mAIC selected the true model $(59,62,88,92)$ times respectively. Thus the advantage of cAIC in model selection is seen at $m=10$ while cAIC and mAIC perform similarly for $m=30$.

Finally, we note the computation of cAIC is relatively slow, which obviously is due to that we need to fit the model for mn times in total. As an indication of computation speed with implementation in R, the computation times on a single simulated dataset for the last simulation above are recorded in Table 3. The entire simulation on the 100 datasets takes a few days on our HP workstation 6800 running Windows 7, with 3GB memory and Intel(R) Core(TM)2 Quad CPU Q9300 @2.50GHz.

4. Concluding remarks

Previous study of unbiased conditional AIC is only limited to the linear mixed-effects models. We provided the corresponding cAIC for Poisson regression. Since

the derivation of the estimator does not depend on either the normality of random effects or specific estimators used for the fixed and random effects, the same formula works in more general contexts such as when using the approach of hierarchical likelihood [4] which has become very popular in recent years. Based on asymptotic arguments, [2] has considered cAIC for general linear and proportional hazards mixed models, which is asymptotically unbiased under some regularity assumptions, although currently unbiased cAIC is lacking in such generality.

Obviously the cAIC derived in the text is still unbiased for more complex random effects structure, for example for hierarchical or crossed designs, since the proof in Theorem 2.1 only used simple property of Poisson distribution. It is also applicable to semiparametric regression using splines due to its equivalent formulation in terms of mixed-effects models although in this context there seems to be no need to perform random effects selection [6]. However, for these more complex structures which usually concerns larger datasets, we expect the computation to be a bottleneck for the application of the method.

Acknowledgement

The author thanks the Editor, the Associate Editor, and two referees for their helpful comments and suggestions that improve the manuscript. This research is support by Singapore MOE Tier 1 Grant.

Appendix: Technical details

Proof of Theorem 2.1. Suppose that the true conditional likelihood is $f(y|u) = \prod_{j=1}^{N} e^{-\mu_{0j}} \mu_{0j}^{y_j}/y_j!$, where $\mu_0 = (\mu_{01}, \ldots, \mu_{0N})$ depends on the random effects u. Let $\hat{\mu}$ be the fitted responses from the mixed-effects model. The conditional Akaike information is

$$\begin{aligned}
\text{cAI} &= -2E_{f(y,u)}E_{f(y^*|u)} \log g(y^*|\hat{\theta}, \hat{b}) \\
&= -2E_{f(y,u)}E_{f(y^*|u)} \left[\sum_j (-\hat{\mu}_j + y_j^* \log \hat{\mu}_j - \log y_j^*!) \right] \\
&= -2E_{f(y,u)} \left[\sum_j (-\hat{\mu}_j + \mu_{0j} \log \hat{\mu}_j - E_{f(y^*|u)} \log y_j^*!) \right].
\end{aligned}$$

Meanwhile,

$$-2E_{f(y,u)} \log g(y|\hat{\theta}, \hat{b}) = -2E_{f(y,u)} \left[\sum_j (-\hat{\mu}_i + y_j \log \hat{\mu}_j - \log y_j!) \right].$$

Thus

$$\text{cAI} - (-2E_{f(y,u)} \log g(y|\hat{\theta}, \hat{b})) = 2E_{f(y,u)} \sum_j (y_j - \mu_{0j}) \log \hat{\mu}_j.$$

In addition, we have that

$$\begin{aligned}
E_{f(y|u)}[\mu_{0j} \log \hat{\mu}_j] &= \sum_{y_{j'}, j' \neq j} \sum_{y_j=0}^{\infty} \left\{ \log[\hat{\mu}_j(y)] \mu_{0j} \frac{e^{-\mu_{0j}} \mu_{0j}^{y_j}}{y_j!} \prod_{j' \neq j} \frac{e^{-\mu_{0j'}} \mu_{0j'}^{y_{j'}}}{y_{j'}!} \right\} \\
&= \sum_{y_{j'}, j' \neq j} \sum_{y_j=0}^{\infty} \left\{ (y_j + 1) \log[\hat{\mu}_j(y)] \frac{e^{-\mu_{0j}} \mu_{0j}^{y_j+1}}{(y_j+1)!} \prod_{j' \neq j} \frac{e^{-\mu_{0j'}} \mu_{0j'}^{y_{j'}}}{y_{j'}!} \right\} \\
&= \sum_{y_{j'}, j' \neq j} \sum_{z_j=1}^{\infty} \left\{ z_j \log[\hat{\mu}_j(y^{(z_j-1)})] \frac{e^{-\mu_{0j}} \mu_{0j}^{z_j}}{z_j!} \prod_{j' \neq j} \frac{e^{-\mu_{0j'}} \mu_{0j'}^{y_{j'}}}{y_{j'}!} \right\} \\
&= \sum_{y_{j'}, j' \neq j} \sum_{z_j=0}^{\infty} \left\{ z_j \log[\hat{\mu}_j(y^{(z_j-1)})] \frac{e^{-\mu_{0j}} \mu_{0j}^{z_j}}{z_j!} \prod_{j' \neq j} \frac{e^{-\mu_{0j'}} \mu_{0j'}^{y_{j'}}}{y_{j'}!} \right\} \\
&= E_{f(y|u)} \left\{ y_j \log[\hat{\mu}_j(y^{(y_j-1)})] \right\},
\end{aligned}$$

where $y^{(z_j-1)}$ is the vector y whose $j-$th component has been replaced by $z_j - 1$, and similarly for $y^{(y_j-1)}$.

Therefore,

$$\begin{aligned}
\text{cAI} - (-2E_{f(y,u)} \log g(y|\hat{\theta}, \hat{b})) &= 2E_{p(u)} E_{f(y|u)} \left\{ \sum_{j=1}^{N} (y_j - \mu_{0j}) \log \hat{\mu}_j \right\} \\
&= 2E_{f(y,u)} \left\{ \sum_{j=1}^{N} y_j \log \hat{\mu}_j - y_j \log[\hat{\mu}_j(y^{(y_j-1)})] \right\}.
\end{aligned}$$

□

References

[1] Burnham, K. P. and Anderson, D. P. (1998). *Model selection and inference: a practical information-theoretical approach*. Springer, New York. MR1919620

[2] Donohue, M. C., Overholser, R., Xu, R. and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika* **98** 685–700.

[3] Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **88** 367–379. MR1844837

[4] Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B-Methodological* **58** 619–656. MR1410182

[5] Liang, H., Wu, H. L. and Zou, G. H. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika* **95** 773–778. MR2443190

[6] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression* **12**. Cambridge University Press. MR1998720

[7] Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92** 351–370. MR2201364