Bias-corrected AIC for selecting variables in Poisson regression models

Ken-ichi Kamo[(a)], Hirokazu Yanagihara[(b)] and Kenichi Satoh[(c)]

[(a)] Corresponding author: Department of Liberal Arts and Sciences,
Sapporo Medical University,
S1 W16, Chuo-ku, Sapporo, Hokkaido, Japan 060-8556.
kamo@sapmed.ac.jp

[(b)]: Department of Mathematics, Graduate School of Science,
Hiroshima University,
1-3-1, Kagamiyama, Higashi-hiroshima, Hiroshima, Japan 739-8521.
yanagi@math.sci.hiroshima-u.ac.jp

[(c)]: Department of Environmentrics and Biometrics, Research Institute for Radiation Biology and Medicine,
Hiroshima University,
Kasumi 1-2-3, Minami-ku, Hiroshima, Japan 734-8551.
kastoh@hiroshima-u.ac.jp

Key Words: Akaike's information criterion; bias correction; generalized linear model; information criterion; model selection; Poisson regression.

Mathematics Subject Classification (2000): primary 62J12; secondary 62H12.

ABSTRACT

In the present paper, we consider the variable selection problem in Poisson regression models. Akaike's information criterion (AIC) is the most commonly applied criterion for selecting variables. However, the bias of the AIC cannot be ignored, especially in small samples. We herein propose a new bias-corrected version of the AIC that is constructed by stochastic expansion of the maximum likelihood estimator. The proposed information criterion can reduce the bias of the AIC from $O(n^{-1})$ to $O(n^{-2})$. The results of numerical

investigations indicate that the proposed criterion is better than the AIC.

1. INTRODUCTION

A Poisson distribution describes the total number of events when it is possible for an event to occur during any of a large number of trials, but the probability of occurrence in any given trial is small. This distribution is regarded as a row of rare events of the binomial distribution. The probability density function of the Poisson distribution is given as

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!},$$

where λ is the intensity parameter. The intensity parameter expresses the mean and variance of the random variable y. Larger values of the intensity parameter will produce a larger number of observations with larger dispersion.

The intensity parameter plays a key role in the Poisson distribution. Then, the Poisson regression model is constructed by allowing the intensity parameter to depend on explanatory variables (Cameron and Trivedi, 1998). Here, the intensity parameter should be restricted to be positive. Therefore, the situation in which the response variables are restricted must be considered. The generalized linear model (GLM) (McCullagh and Nelder, 1989) is widely used in such situations. The GLM is expressed through the link function, which is constructed by a known monotonic function that transforms the expectation of the responses to a scale on which they are unconstrained. The Poisson regression model corresponds to the GLM with log link.

In the present paper, we consider the problem of selecting the optimal subset of variables in Poisson regression models. Generally, in regression analysis, the expectation of the response variable should be expressed by an essential minimum number of variables (principle of parsimony). These optimal variables are determined by minimizing the risk function based on the predicted Kullback-Leibler information (Kullback and Leibler, 1951). Since the Kullback-Leibler information cannot be calculated exactly, Akaike (1973) proposed an information criterion, which has come to be known as Akaike's information criterion, or AIC. Although the AIC is an effective estimator of Kullback-Leibler information, the AIC has a

bias that cannot be ignored, especially in small samples, because the AIC is derived using the asymptotic properties. A number of studies have investigated the development of an information criterion that corrects for such bias (e.g., Bedrick and Tsai (1994), Fujikoshi et al. (2003), Hurvich and Tsai (1989), Satoh et al. (1997), Sugiura (1978), Yanagihara et al. (2003), and Yanagihara (2006)). In the present paper, we propose a new information criterion that is obtained by correcting the bias of the AIC in Poisson regression models.

Another popular regression model with count observation is the logistic regression model, which assumes a binomial distribution. This model is also a GLM having a link function that is given by the logit function. For the logistic regression model, Yanagihara et al. (2003) proposed a bias-corrected AIC derived by stochastic expansion of the maximum likelihood estimator (MLE) of an unknown parameter. In the present study, we apply this procedure to the Poisson regression model.

The remainder of the present paper is organized as follows. In Section 2, we propose a new information criterion by reducing the bias of the AIC in Poisson regression models. In Section 3, we investigate the performance of the proposed information criterion through numerical simulations. The Appendixes present a detailed derivation of the proposed information criterion.

2. BIAS-CORRECTED AIC

2.1. STOCHASTIC EXPANSION OF THE MLE IN THE POISSON REGRESSION MODEL

Let the counted observation y_{ij} be independently distributed according to a Poisson distribution with intensity parameter λ_i $(i = 1, 2, \ldots, m)$, that is,

$$y_{ij} \overset{i.d.}{\sim} Po(\lambda_i).$$

In the present paper, we consider the case in which there are n_i repeated observations under the same parameter λ_i. Therefore, the sub-index j in y_{ij} expresses the number of observations for n_i repeated in each i $(j = 1, \ldots n_i)$. From the reproducing property of the Poisson distribution, $y_i = \sum_{j=1}^{n_i} y_{ij}$ is also independently distributed according to a Poisson

distribution, that is,

$$y_i \overset{i.d.}{\sim} Po(n_i\lambda_i).$$

In what follows, we set y_i to be the response variable. The probability density function for response y_i is then obtained as

$$f(y_i) = \frac{(n_i\lambda_i)^{y_i}\exp(-n_i\lambda_i)}{y_i!}.$$

In a Poisson regression model, $\log\lambda_i$ is expressed as a linear combination of explanatory variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ir})'$, that is,

$$\log\lambda_i = \sum_{j=1}^{r}\beta_j x_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta},$$

where r is the number of explanatory variables and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)'$ is an unknown parameter vector.

Let $\hat{\boldsymbol{\beta}}$ denote the MLE of $\boldsymbol{\beta}$. This is expressed by maximizing the log-likelihood function excluding the constant term as follows:

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{m}\left(y_i\boldsymbol{x}_i'\boldsymbol{\beta} - n_i\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right), \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)'$. In other words, $\hat{\boldsymbol{\beta}}$ satisfies the following equation

$$\sum_{i=1}^{m} n_i\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\boldsymbol{x}_i = \sum_{i=1}^{m} y_i\boldsymbol{x}_i. \tag{2}$$

Equation (2) is a likelihood equation and is derived by differentiating (1) with respect to $\boldsymbol{\beta}$. Using $\hat{\boldsymbol{\beta}}$, the estimator of the i-th intensity parameter and the fitted value of y_i are obtained as $\hat{\lambda}_i = \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$ and $\hat{y}_i = n_i\hat{\lambda}_i$, respectively.

Under the assumption in Appendix 1, $\hat{\boldsymbol{\beta}}$ can be formally expressed as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{1}{\sqrt{n}}\hat{\boldsymbol{\beta}}_1 + \frac{1}{n}\hat{\boldsymbol{\beta}}_2 + \frac{1}{n\sqrt{n}}\hat{\boldsymbol{\beta}}_3 + O(n^{-2}), \tag{3}$$

where $n = \sum_{i=1}^{m} n_i$ is the total number of trials. Let $\boldsymbol{a} = (a_1, \ldots, a_p)'$, and let $\boldsymbol{b}$ be p-dimensional vectors. Then, we define the notations $\boldsymbol{D}_{\boldsymbol{a}} = \mathrm{diag}(a_1, \ldots, a_p)$, $\boldsymbol{D}_{\boldsymbol{ab}} = \boldsymbol{D}_{\boldsymbol{a}}\boldsymbol{D}_{\boldsymbol{b}}$

and $\boldsymbol{D}_{ab^2} = \boldsymbol{D}_a \boldsymbol{D}_b^2$. By using this matrix, we define $\boldsymbol{\Xi}$ such that $\boldsymbol{\Xi} = (\boldsymbol{X}'\boldsymbol{D}_{\boldsymbol{\lambda}\boldsymbol{\rho}}\boldsymbol{X})^{-1}$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)'$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)' = (n_i/n, \ldots, n_m/n)'$, and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)'$. By substituting (3) into (2) and comparing terms of the same order, we obtain the explicit forms of $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\beta}}_3$ as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= \sum_{i=1}^{m} \sqrt{\rho_i} z_i \boldsymbol{\Xi} \boldsymbol{x}_i, \\
\hat{\boldsymbol{\beta}}_2 &= -\frac{1}{2} \sum_{i=1}^{m} \rho_i \lambda_i \left(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_1 \right)^2 \boldsymbol{\Xi} \boldsymbol{x}_i, \\
\hat{\boldsymbol{\beta}}_3 &= -\sum_{i=1}^{m} \rho_i \lambda_i \left\{ \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_1 \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_2 + \frac{1}{6} \left(\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_1 \right)^3 \right\} \boldsymbol{\Xi} \boldsymbol{x}_i,
\end{aligned} \tag{4}$$

where $z_i = (y_i - n_i\lambda_i)/\sqrt{n_i}$. A detailed derivation of (4) is presented in Appendix 2.

2.2. BIAS-CORRECTED AKAIKE INFORMATION CRITERION

In a regression analysis, the optimal subset of variables can be determined by minimizing the risk function based on the predicted Kullback-Leibler information (Kullback and Leibler, 1951), which is defined as follows:

$$R = -2E_{\boldsymbol{y}} E_{\boldsymbol{u}} [\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{u})], \tag{5}$$

where $\boldsymbol{u}$ is a future observation that is independent of $\boldsymbol{y}$ and is distributed according to the same distribution as $\boldsymbol{y}$. This provides a measure of the discrepancy between the true model and candidate models.

Generally, (5) cannot be calculated exactly, because it includes unknown parameters. Although the rough estimator of R is easily obtained as $-2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y})$, this has a constant bias. Akaike (1973) evaluated such a constant bias as "2×(the number of parameters)" and proposed the AIC by adding this bias to the rough estimator. In the case of a Poisson regression, the AIC is defined as follows:

$$\mathrm{AIC} = -2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) + 2r.$$

In this case, the AIC is interpreted as the sum of the "goodness of fit to the model" and the "model complexity penalty". The former corresponds to $-2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y})$ and the latter corresponds to $2r$.

The AIC has a non-negligible bias, especially in small samples. Therefore, in the present paper, we propose a new information criterion, called the CAIC (Bias-corrected AIC), which is defined as follows:

$$\mathrm{CAIC} = -2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) + 2r + \frac{1}{n}\mathbf{1}'_m \boldsymbol{D}_{\hat{\boldsymbol{\lambda}}\boldsymbol{\rho}} \left\{ \left(\hat{\boldsymbol{W}}_{(3)} + \boldsymbol{D}_{\hat{\boldsymbol{w}}} \hat{\boldsymbol{W}} \boldsymbol{D}_{\hat{\boldsymbol{w}}} \right) \boldsymbol{D}_{\hat{\boldsymbol{\lambda}}\boldsymbol{\rho}} - \boldsymbol{D}^2_{\hat{\boldsymbol{w}}} \right\} \mathbf{1}_m, \tag{6}$$

where $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_m)'$, $\hat{\boldsymbol{W}} = [\hat{w}_{ij}] = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}\boldsymbol{\rho}}\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\hat{\boldsymbol{w}} = (\hat{w}_{11}, \ldots, \hat{w}_{mm})'$, $\hat{\boldsymbol{W}}_{(3)} = [\hat{w}^3_{ij}]$, and $\mathbf{1}_m$ is a $m \times 1$ vector in which all elements are 1. This criterion is constructed by adding the bias correcting term of $O(n^{-1})$ to the AIC. Although the bias correcting term of the (6) is complicated, the order of the bias is improved to $O(n^{-2})$.

The bias correcting term in the (6) is derived as follows. By using the stochastic expansion (3), the bias of $-2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y})$ to R is expanded as

$$\begin{aligned}
B &= E_{\boldsymbol{y}} E_{\boldsymbol{u}} \left[2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) - 2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{u}) \right] \\
&= 2\sum_{i=1}^{m} \sqrt{n_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}} \right] \\
&= 2\sqrt{n} \sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}}[z_i] \boldsymbol{x}'_i \boldsymbol{\beta} + 2\sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_1 \right] \\
&\quad + \frac{2}{\sqrt{n}} \sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_2 \right] + \frac{2}{n} \sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_3 \right] + O(n^{-2}).
\end{aligned}$$

Note that the first term becomes 0, because $E_{\boldsymbol{y}}[z_i] = 0$. From Appendix 3, we see that

$$\begin{aligned}
&\sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_1 \right] = r, \\
&\sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_2 \right] = -\frac{1}{2\sqrt{n}} \boldsymbol{\lambda}' \boldsymbol{D}_{\boldsymbol{\rho}} \boldsymbol{W}_{(3)} \boldsymbol{D}_{\boldsymbol{\rho}} \boldsymbol{\lambda}, \\
&\sum_{i=1}^{m} \sqrt{\rho_i} E_{\boldsymbol{y}} \left[z_i \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_3 \right] = -\boldsymbol{\lambda}' \boldsymbol{D}_{\boldsymbol{\rho}} (\boldsymbol{D}_{\boldsymbol{w}} \boldsymbol{W}_{(3)} \boldsymbol{D}_{\boldsymbol{w}} - 2\boldsymbol{W}_{(3)}) \boldsymbol{D}_{\boldsymbol{\rho}} \boldsymbol{\lambda} - \frac{1}{2} \boldsymbol{\lambda}' \boldsymbol{D}_{\boldsymbol{\rho}\boldsymbol{w}^2} \mathbf{1}_m + O(n^{-1}),
\end{aligned} \tag{7}$$

where $\boldsymbol{W} = [w_{ij}] = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{D}_{\boldsymbol{\lambda}\boldsymbol{\rho}}\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{w} = (w_{11}, \ldots, w_{mm})'$ and $\boldsymbol{W}_{(3)} = [w^3_{ij}]$. Therefore, we obtain an expansion of B as

$$B = 2r + \frac{1}{n}\mathbf{1}'_m \boldsymbol{D}_{\boldsymbol{\lambda}\boldsymbol{\rho}} \left\{ \left(\boldsymbol{W}_{(3)} + \boldsymbol{D}_{\boldsymbol{w}} \boldsymbol{W} \boldsymbol{D}_{\boldsymbol{w}} \right) \boldsymbol{D}_{\boldsymbol{\lambda}\boldsymbol{\rho}} - \boldsymbol{D}^2_{\boldsymbol{w}} \right\} \mathbf{1}_m + O(n^{-2}). \tag{8}$$

By substituting $\hat{\boldsymbol{\beta}}$ into $\boldsymbol{\beta}$ in (8) and omitting the $O(n^{-2})$ term, the bias correcting term in the (6) is obtained.

III. NUMERICAL STUDIES

In this section, we verify the performance of the CAIC through numerical simulations. We simulate 10,000 realizations of $\boldsymbol{y}$. Throughout the examination, we also consider Takeuchi's information criterion (TIC), which was proposed by Takeuchi (1976) and is defined as

$$\mathrm{TIC} = -2\mathcal{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{y}) + 2\mathrm{tr}(\boldsymbol{I}(\hat{\boldsymbol{\beta}})\boldsymbol{J}(\hat{\boldsymbol{\beta}})^{-1}),$$

where $\boldsymbol{I}(\boldsymbol{\beta}) = (\partial\mathcal{L}/\partial\boldsymbol{\beta})(\partial\mathcal{L}/\partial\boldsymbol{\beta}')$ and $\boldsymbol{J}(\boldsymbol{\beta}) = -\partial^2\mathcal{L}/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}')$. In the Poisson regression model, $\boldsymbol{I}$ and $\boldsymbol{J}$ are expressed as

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{m}(y_i - n_i\hat{\lambda}_i)^2\boldsymbol{x}_i\boldsymbol{x}_i' \quad \text{and} \quad \boldsymbol{J}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{m} n_i\hat{\lambda}_i\boldsymbol{x}_i\boldsymbol{x}_i'.$$

These matrixes are obtained by recalculating the bias correcting term of the AIC when the true model is not always included among the candidate models.

We prepare two situations with $n = 10$ (examinations 1 and 2 are set as $(n_i, m) = (1, 10)$ and $(n_i, m) = (2, 5)$, respectively) as the candidate models. The $m \times 4$ explanatory variable matrix $\boldsymbol{X}_F$ is constructed such that the first column is $\mathbf{1}_m$ and the remainder elements are generated by the uniform distribution $U(-1, 1)$. The relationship between the true intensity parameter and the true regressor variables is set by $(\log\lambda_1, \ldots, \log\lambda_m)' = \boldsymbol{X}_F\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0 = (1, 1, 0, 0)'$. Hence, the true model is constructed from the first and second columns only. In our examinations, the true $\boldsymbol{\lambda}$s in examinations 1 and 2 are
$(4.52, 1.77, 1.22, 6.74, 2.30, 2.48, 6.97, 3.22, 6.85, 4.59)'$ and $(1.68, 4.26, 6.12, 6.67, 1.16)'$, respectively.

In order to clearly illustrate the results, we index the models as shown in Table I. The total number of candidates is $2^3 = 8$ in both examinations. Indexes 1 through 4 are assigned to the under-specified models in order of higher risk, and indexes 5 through 8 are assigned to the over-specified models in order of lower risk. In both examinations, the model with the lowest risk corresponds to the true model, which is assigned index 5. The horizontal axes in Figures I and II denote the index.

Table I: Model indexes.

	under specified				over specified			
	1	2	3	4	5	6	7	8
Examination 1	{1,4}	{1}	{1,3,4}	{1,3}	{1,2}	{1,2,4}	{1,2,3}	{1,2,3,4}
Examination 2	{1,3}	{1,3,4}	{1}	{1,4}	{1,2}	{1,2,4}	{1,2,3}	{1,2,3,4}

The number in {} means one of column in the explanatory variable matrix $\boldsymbol{X}_F$. For example, $\{1, 2\}$ denotes the model with first and second columns of $\boldsymbol{X}_F$, which is the true model.

Figure I shows the target risk and the average values of the three criteria, i.e., AIC, TIC, and CAIC, in each examination. In both examinations, the curves of risk, the AIC, and the CAIC have minima at index 5, which agrees with the true model. All of the curves are nearly flat in the intervals between indexes 1 through 4 and indexes 5 through 8, which correspond to under-specified and over-specified models, respectively. There is a gap between the under-specified and over-specified models. In the over-specified models, the average of the CAIC is in approximate agreement with the risk. This means that the CAIC has the smallest bias, which is the primary objective of the present study. All criteria tend to underestimate the risk in the over-specified models, and the TIC has an especially large bias.

Figure I: Risk and average values of AIC, TIC, and CAIC.

Examination 1 ($n_i = 1$ and $m = 10$) Examination 2 ($n_i = 2$ and $m = 5$)

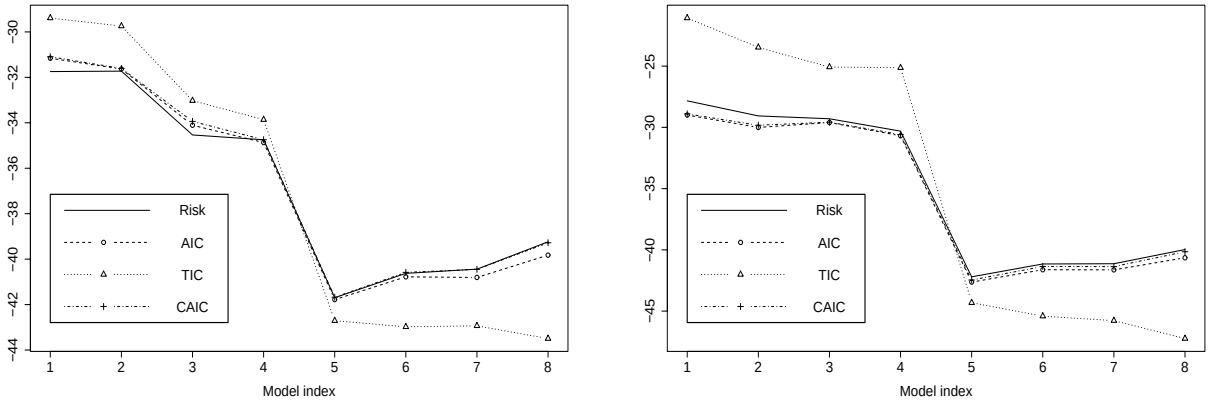
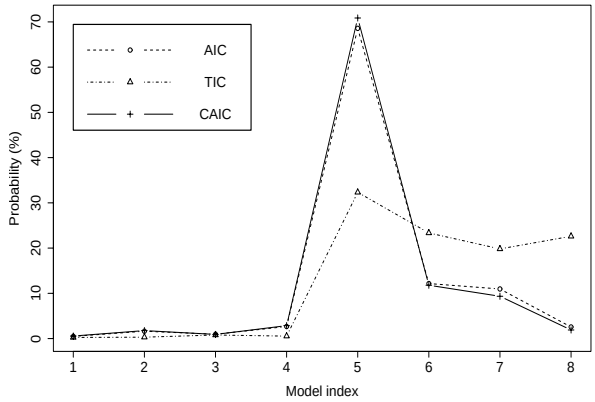


Figure II shows the frequencies of the models selected by the three criteria (AIC, TIC, and CAIC). The probability that the model with the smallest risk will be selected as the best model is the highest for the CAIC (70.86% and 73.35% in examinations 1 and 2, respectively) in both examinations (see Table II). If the true model is not selected, all of the criteria tend to select the model with larger number of variables. The under-specified models are rarely selected, especially by the TIC (1.89% and 0.18% in examinations 1 and 2, respectively). This may be due to the fact that the criteria tend to underestimate the risk in the over-specified models. In examination 2, the TIC most often selects the model with index 8 (58.73%), which is the most complicated model. This may be due to the fact that the average value of the TIC attains minimum at this index and the TIC tends to selects this model. In both examinations, the improvement for CAIC in the selection probability of the model with the smallest risk as compared with the AIC is slight (2% in both examinations) .

Figure II: Frequencies of the selected models.

Examination 1 ($n_i = 1$ and $m = 10$)



Examination 2 ($n_i = 2$ and $m = 5$)

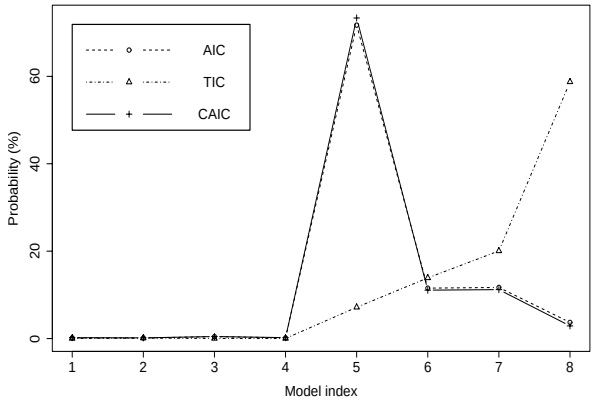


Table II: Probability of selecting the model with the smallest risk and the estimated risk.

	Examination 1		Examination 2	
	Probability (%)	Estimated risk	Probability (%)	Estimated risk
AIC	68.57	−39.50	71.72	−34.84
TIC	32.33	−39.41	7.16	−34.54
CAIC	70.86	−39.51	73.35	−34.85

Table II also shows the estimated risk, which is obtained by substituting the MLE of the estimated best model for the risk (5). It is preferable when this value is small, because the model with the smallest risk is selected. The smallest estimated risk is obtained by the CAIC in both examinations.

In conclusion, the bias of the AIC is reduced by the proposed CAIC. As a result, the CAIC has several advantageous properties, which include a high probability of selecting the model with the smallest risk and minimum risk estimation.

APPENDIXES

Appendix 1 (Assumptions).

In order to guarantee the validity of stochastic expansion in (3), we have the following assumption, as introduced in Nordberg (1980):

Assumption. *Let* $M_n = n^{-1}\sum_{i=1}^m n_i \boldsymbol{x}_i'\boldsymbol{x}_i$. *We assume that*

- $\{M_n\}_{n=1}^{\infty}$ *is uniformly positive definite,*
- *There exists an* N *such that* $|x_{ij}| \leq N < \infty$.

If $\rho_i^{-1} = O(1)$, then this assumption is satisfied when $\mathrm{rank}(\boldsymbol{X}) = r$.

Appendix 2 (Derivation of equation (4)).

Using z_i, we rewrite the likelihood equation of (2) as follows:

$$\sum_{i=1}^m n_i\hat{\lambda}_i\boldsymbol{x}_i = \sum_{i=1}^m(\sqrt{n_i}z_i + n_i\lambda_i)\boldsymbol{x}_i,$$

which is equivalent to

$$\sqrt{n}\sum_{i=1}^m \rho_i(\hat{\lambda}_i - \lambda_i)\boldsymbol{x}_i = \sum_{i=1}^m \sqrt{\rho_i}z_i\boldsymbol{x}_i. \tag{9}$$

The stochastic expansion (3) yields the following expression of $\hat{\lambda}_i - \lambda_i$:

$$\begin{aligned}\hat{\lambda}_i - \lambda_i = \frac{\lambda_i}{\sqrt{n}}\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1 + \frac{\lambda_i}{n}\left\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_2 + \frac{1}{2}\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\right)^2\right\} \\ + \frac{\lambda_i}{n\sqrt{n}}\left\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_3 + \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_2 + \frac{1}{6}\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\right)^3\right\} + O(n^{-2})\end{aligned}$$

Substituting this into the left-hand side of (9) and comparing terms of the same order to n, we obtain the following three relations:

$$\begin{aligned}&\sum_{i=1}^m \rho_i\lambda_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\boldsymbol{x}_i = \sum_{i=1}^m \sqrt{\rho_i}z_i\boldsymbol{x}_i, \\ &\sum_{i=1}^m \rho_i\lambda_i\left\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_2 + \frac{1}{2}\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\right)^2\right\}\boldsymbol{x}_i = \boldsymbol{0}, \\ &\sum_{i=1}^m \rho_i\lambda_i\left\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_3 + \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_2 + \frac{1}{6}\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1\right)^3\right\}\boldsymbol{x}_i = \boldsymbol{0}.\end{aligned}$$

Using the relation $\sum_{i=1}^m \rho_i\lambda_i\boldsymbol{x}_i\boldsymbol{x}_i' = \boldsymbol{\Xi}^{-1}$, we simplify the above equation and then obtain the explicit forms of $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ in (4).

Appendix 3 (Derivation of equation (7)).

We first consider the first through fourth moments of z_i. Since y_i is distributed according to the Poisson distribution with intensity parameter $n_i\lambda_i$, these moments are calculated as

$$E_{\boldsymbol{y}}[z_i] = 0, \quad E_{\boldsymbol{y}}[z_i^2] = \lambda_i, \quad E_{\boldsymbol{y}}[z_i^3] = \frac{\lambda_i}{\sqrt{n_i}}, \quad \text{and} \quad E_{\boldsymbol{y}}[z_i^4] = \frac{3n_i\lambda_i^2 + \lambda_i}{n_i}. \tag{10}$$

Let us now consider the first equation in (7). Substituting the explicit form of $\hat{\boldsymbol{\beta}}_1$ obtained by (4) into the first equation of (7), we have

$$\begin{aligned}
\sum_{i=1}^m \sqrt{\rho_i} E_{\boldsymbol{y}}\left[z_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_1\right] &= \sum_{i=1}^m \rho_i E_{\boldsymbol{y}}[z_i^2] \boldsymbol{x}_i' \boldsymbol{\Xi} \boldsymbol{x}_i \\
&= \sum_{i=1}^m \rho_i \lambda_i \boldsymbol{x}_i' \boldsymbol{\Xi} \boldsymbol{x}_i \\
&= \mathrm{tr}(\boldsymbol{D}_{\boldsymbol{\lambda}\boldsymbol{\rho}} \boldsymbol{X} \boldsymbol{\Xi} \boldsymbol{X}') \\
&= \mathrm{tr}(\boldsymbol{I}_r) = r.
\end{aligned}$$

Next, we consider the second equation in (7). Substituting the explicit form of $\hat{\boldsymbol{\beta}}_2$ obtained by (4) into the second equation of (7) yields

$$\begin{aligned}
\sum_{i=1}^m \sqrt{\rho_i} E_{\boldsymbol{y}}\left[z_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_2\right] &= -\frac{1}{2} \sum_{ij}^m \sqrt{\rho_i} \rho_j \lambda_j \boldsymbol{x}_i' \boldsymbol{\Xi} \boldsymbol{x}_j E_{\boldsymbol{y}}\left[z_i (\boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_1)^2\right] \\
&= -\frac{1}{2} \sum_{ijk\ell}^m \sqrt{\rho_i \rho_k \rho_\ell} \rho_j \lambda_j \boldsymbol{x}_i' \boldsymbol{\Xi} \boldsymbol{x}_j \boldsymbol{x}_j' \boldsymbol{\Xi} \boldsymbol{x}_k \boldsymbol{x}_j' \boldsymbol{\Xi} \boldsymbol{x}_\ell E_{\boldsymbol{y}}[z_i z_k z_\ell] \\
&= -\frac{1}{2} \sum_{ij}^m \rho_i^{3/2} \rho_j \lambda_j (\boldsymbol{x}_i' \boldsymbol{\Xi} \boldsymbol{x}_j)^3 E_{\boldsymbol{y}}[z_i^3] \\
&= -\frac{1}{2\sqrt{n}} \sum_{ij}^m \rho_i \rho_j \lambda_i \lambda_j w_{ij}^3 \\
&= -\frac{1}{2\sqrt{n}} \boldsymbol{\lambda}' \boldsymbol{D}_{\boldsymbol{\rho}} \boldsymbol{W}_{(3)} \boldsymbol{D}_{\boldsymbol{\rho}} \boldsymbol{\lambda},
\end{aligned}$$

where the notation $\sum_{a_1 a_2 \cdots}^m$ means $\sum_{a_1=1}^m \sum_{a_2=1}^m \cdots$.

Finally, we consider the last equation in (7). Substituting the explicit form of $\hat{\boldsymbol{\beta}}_3$ obtained by (4) into the last equation of (7), we have

$$\sum_{i=1}^m \sqrt{\rho_i} E_{\boldsymbol{y}}\left[z_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_3\right] = -\sum_{ij}^m \sqrt{\rho_i} \rho_j \lambda_j w_{ij} E_{\boldsymbol{y}}\left[z_i \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_1 \boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_2\right] - \frac{1}{6} \sum_{ij}^m \sqrt{\rho_i} \rho_j \lambda_j w_{ij} E_{\boldsymbol{y}}\left[z_i \left(\boldsymbol{x}_j' \hat{\boldsymbol{\beta}}_1\right)^3\right].$$

Let the first term of the right-hand side be η_1, and let the last term of the right-hand side be η_2. These terms can then be rewritten as

$$\eta_1 = \frac{1}{2}\sum_{ab}^{m} \rho_a\rho_b\lambda_a\lambda_b w_{ab} \sum_{ijk\ell}^{m} \sqrt{\rho_i\rho_j\rho_k\rho_\ell} E_{\boldsymbol{y}}[z_i z_j z_k z_\ell] w_{ai}w_{aj}w_{bk}w_{b\ell},$$
$$\eta_2 = -\frac{1}{6}\sum_{a}^{m} \rho_a\lambda_a \sum_{ijk\ell}^{m} \sqrt{\rho_i\rho_j\rho_k\rho_\ell} E_{\boldsymbol{y}}[z_i z_j z_k z_\ell] w_{ai}w_{aj}w_{ak}w_{a\ell}.$$

The expectation $E_{\boldsymbol{z}}[z_i z_j z_k z_\ell]$ remains only in the following four combinations for sub-indexes: i) $i=j=k=\ell$, ii) $i=j$, $k=\ell$, and $i\neq k$, iii) $i=k$, $j=\ell$, and $i\neq j$, and iv) $i=\ell$, $j=k$, and $i\neq j$, because z_i and z_j are independent when $i\neq j$ and $E_{\boldsymbol{y}}[z_i]=0$. Since $\sum_i^m \rho_i\lambda_i w_{ai}w_{bj} = w_{ab}$, the parts concerning i,j,k and ℓ in c_1 and c_2 are calculated as

$$\begin{aligned}
&\sum_{ijk\ell}^{m} \sqrt{\rho_i\rho_j\rho_k\rho_\ell} E_{\boldsymbol{y}}[z_i z_j z_k z_\ell] w_{ai}w_{aj}w_{bk}w_{b\ell} \\
&= \sum_{i=1}^{m} \rho_i^2 E_{\boldsymbol{y}}[z_i^4] w_{ai}^2 w_{bi}^2 + \sum_{j\neq i}^{m} \rho_i\rho_j E_{\boldsymbol{y}}[z_i^2]E_{\boldsymbol{y}}[z_j^2] w_{ai}^2 w_{bj}^2 + 2\sum_{j\neq i}^{m} \rho_i\rho_j E_{\boldsymbol{y}}[z_i^2]E_{\boldsymbol{y}}[z_j^2] w_{ai}w_{aj}w_{bi}w_{bj} \\
&= \sum_{i=1}^{m} \rho_i^2 w_{ai}^2 w_{bi}^2 \left(3\lambda_i^2 + \frac{\lambda_i}{n_i}\right) + \sum_{j\neq i}^{m} \rho_i\rho_j\lambda_i\lambda_j w_{ai}^2 w_{bj}^2 + 2\sum_{j\neq i}^{m} \rho_i\rho_j\lambda_i\lambda_j w_{ai}w_{aj}w_{bi}w_{bj} \\
&= \sum_{ij}^{m} \rho_i\rho_j\lambda_i\lambda_j w_{ai}^2 w_{bj}^2 + 2\sum_{ij}^{m} \rho_i\rho_j\lambda_i\lambda_j w_{ai}w_{aj}w_{bi}w_{bj} + O(n^{-1}) \\
&= w_{aa}w_{bb} + 2w_{ab}^2 + O(n^{-1})
\end{aligned}$$

and

$$\begin{aligned}
\sum_{ijkl}^{m} \sqrt{\rho_i\rho_j\rho_k\rho_\ell} E_{\boldsymbol{y}}[z_i z_j z_k z_\ell] w_{ai}w_{aj}w_{ak}w_{a\ell} &= \sum_{i=1}^{m} \rho_i^2 E_{\boldsymbol{y}}[z_i^4] w_{ai}^4 + 3\sum_{j\neq i}^{m} \rho_i\rho_j E_{\boldsymbol{y}}[z_i^2]E_{\boldsymbol{y}}[z_j^2] w_{ai}^2 w_{aj}^2 \\
&= \sum_{i=1}^{m} \rho_i^2 w_{ai}^4 \left(3\lambda_i^2 + \frac{\lambda_i}{n_i}\right) + 3\sum_{j\neq i}^{m} \rho_i\rho_j\lambda_i\lambda_j w_{ai}^2 w_{aj}^2 \\
&= 3\sum_{i=1}^{m} \rho_i\lambda_i w_{ai}^2 \sum_{j=1}^{m} \rho_j\lambda_j w_{aj}^2 + O(n^{-1}) \\
&= 3w_{aa}^2 + O(n^{-1}),
\end{aligned}$$

where $\sum_{a_1\neq a_2}^m$ denotes $\sum_{a_2=1}^m \sum_{a_1=1, a_1\neq a_2}^m$. Using these results, η_1 and η_2 are expressed as

$$\begin{aligned}
\eta_1 &= \frac{1}{2}\sum_{a}^{m} \rho_a\lambda_a \sum_{b}^{m} \rho_b\lambda_b w_{ab}(w_{aa}w_{bb} + 2w_{ab}^2) + O(n^{-1}) \\
&= \boldsymbol{\lambda}'\boldsymbol{D}_{\boldsymbol{\rho}}\boldsymbol{W}_{(3)}\boldsymbol{D}_{\boldsymbol{\rho}}\boldsymbol{\lambda} + \frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{D}_{\boldsymbol{\rho w}}\boldsymbol{W}_{(3)}\boldsymbol{D}_{\boldsymbol{\rho w}}\boldsymbol{\lambda} + O(n^{-1})
\end{aligned}$$

and

$$\eta_2 = -\frac{1}{2}\sum_{a}^{m}\rho_a\lambda_a w_{aa}^2 + O(n^{-1})$$
$$= -\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{D}_{\boldsymbol{\rho}\boldsymbol{w}^2}\mathbf{1}_m + O(n^{-1}),$$

respectively. Hence, we obtain the desired result.

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In 2nd. International Symposium on Information Theory (B. N. Petrov & F. Csáki eds.), Akadémiai Kiadó, Budapest*, 267–281.

Bedrick, E. J. and Tsai, C. L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression analysis of count data.* Cambridge; New York.

Fujikoshi, Y., Noguchi, T., Ohtaki, M. and Yanagihara, H. (2003). Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models. *Ann. Inst. Statist. Math.*, **55**, 537–553.

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

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models, 2nd edition.* Chapman & Hall: London.

Nordberg, L. (1980). Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models. *Scand. J. Statist.*, **7**, 27–32.

Satoh, K., Kobayashi, M. and Fujikoshi, Y. (1997). Variable selection for the growth curve model. *J. Multivariate Analysis*, **60**, 277–292.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. -Theory Meth.*, **7**, 13–26.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*. **153**, 12–18 (in Japanese).

Yanagihara, H., Sekiguchi, R. and Fujikoshi, Y. (2003). Bias correction of AIC in logistic regression models. *J. Statist. Plann. Inference*, **115**, 349–360.