应用概率统计　第二十五卷
第三期　2009年6月

Chinese Journal of Applied Probability
and Statistics Vol.25 No.3 Jun. 2009

Variable Selection in Joint Generalized Linear Models *

Wang Darong[1,2]　　Zhang Zhongzhan[1]

([1] *Department of Applied Mathematics, Beijing University of Technology, Beijing, 100124*)

([2] *The Pilot College of Beijing University of Technology, Beijing, 101101*)

Abstract

We focus on variable selection in this paper, and propose a variable selection criterion based on the extended quasi-likelihood which is for joint generalized linear models with structured dispersions. The new criterion is an extension of Akaike's information criterion. Its performance is investigated through simulation studies and a real data application.

Keywords: AIC, variable selection, joint GLMs, extended quasi-likelihood, Kullback-Leibler information.

AMS Subject Classification: 62B10, 62J12.

《应用概率统计》版权所有

§1. Introduction

Variable (or model) selection is an essential part of any statistics analysis. There is an extensive model selection literature in statistics (e.g. [1] and references therein), but mainly for the classic linear regression. One powerful and widely used model selection criterion is Akaike's Information Criterion(AIC)[2]. Generalized linear models are regression models in a number of cases, including categorical responses, where the classical assumptions are not satisfied, see [3–5], etc.

For generalized linear models, Pregibon[6] and Hosmer et al.[7] proposed two different versions of Mallow's C_p variable selection criterion. Furthermore, Efron[8] and Jin et al.[9] considered two versions of AIC criterion. However, the generalized linear models considered in these papers do not include a dispersion parameter. On the other hand, McCullagh and Nelder[4] (P. 90) suggested that it should be wise to assume that a dispersion parameter is presented in the model unless the data or prior information indicate otherwise. Hurvich and Tsai[10] considered a corrected AIC criterion for a kind of extended quasi-likelihood models in small samples. Pan[11] proposed a modification to AIC based on GEE or quasi-likelihood. However the dispersion parameter was treated as a constant in the two papers.

*The project supported by NSFC (10771010), NSFBeijing (1072003) and PHR (IHLB).
Received August 24, 2006. Revised November 18, 2008.

We focus on joint generalized linear models (JGLMs) for the mean and dispersion, and develop a new variable selection technique for such models using ideas from extended quasi-likelihood (EQL) of Nelder and Pregibon[12]. The rest of this paper is organized as follows. In Section 2, we first present the model structure of JGLMs, and then derive the EAIC. Section 3 provides simulation studies to demonstrate the effectiveness of the proposed criterion, followed by a real data illustration in Section 4. Section 5 gives a brief summary and discussion.

§2. Derivation of EAIC

In this section, we present a heuristic derivation of the proposed criterion.

2.1 Model Structures

Let $y = (y_1, y_2, \cdots, y_n)'$ be an independent response variable vector from a JGLM, X_i and Z_i be the corresponding observations of the explanation variables for the mean and the dispersion of y_i $(i = 1, \cdots, n)$ respectively. The true models considered in this paper are composed of three parts:

(1) a model for the variances $\mathsf{Var}\,(y_i) = \phi_{i0}V(\mu_{i0})$;

(2) a GLM for the means $\eta_{i0} = g(\mu_{i0}) = X'_{i0}\beta_0$, where X_{i0}, $i = 1, \cdots, n$ are the observations of the real mean explanation variables. They are sub-vector of X_i, $i = 1, \cdots, n$ from the same subset of the possible mean explanation variables;

(3) a GLM for the dispersions $\zeta_{i0} = h(\phi_{i0}) = Z'_{i0}\gamma_0$, where Z_{i0}, $i = 1, \cdots, n$ are the real explanation variables for the dispersion, and the functions $V(\cdot)$, $g(\cdot)$ and $h(\cdot)$ are supposed to be known.

Let $\mu_0 = (\mu_{10}, \mu_{20}, \cdots, \mu_{n0})'$ denote the expectation of y evaluated under the true model, $\phi_0 = (\phi_{10}, \phi_{20}, \cdots, \phi_{n0})'$ denote the dispersion vector, and $\theta_0 = (\beta_0, \gamma_0)'$ denote the true parameter vector.

For inference from JGLMs, [12, 13] proposed the use of extended quasi-likelihood. The extended (log) quasi-likelihood function for the true model is

$$-2Q_0^+(y; \mu_0, \phi_0) = \sum_{i=1}^{n} \frac{D_{i0}(y_i; \mu_{i0})}{\phi_{i0}} + \sum_{i=1}^{n} \log\{2\pi\phi_{i0}V(y_i)\},$$

where D_{i0} is the deviance component in the model for the mean,

$$D_{i0}(y_i; \mu_{i0}) = 2\int_{\mu_{i0}}^{y_i} \frac{y_i - t}{V(t)}\mathrm{d}t.$$

《应用概率统计》版权所有

With the specified regression models $\mu_0 = g^{-1}(X_0'\beta_0)$ and $\phi_0 = h^{-1}(Z_0'\gamma_0)$, the extended (log) quasi-likelihood function can be written down as a function of the regression coefficients $\theta_0 = (\beta_0, \gamma_0)'$: $Q_0^+(y; \mu_0, \phi_0) = Q_0^+(y; \beta_0, \gamma_0) = Q_0^+(y|\theta_0)$.

In practice, we do not know the true model, thus we fit the data with a candidate family of JGLMs:

(1) a model for the variances $\mathsf{Var}\,(y_i) = \phi_i V(\mu_i)$;

(2) a GLM for the means $\eta_i = g(\mu_i) = X_i'(a)\beta$;

(3) a GLM for the dispersions $\zeta_i = h(\phi_i) = Z_{i0}'\gamma$,

where $X_i(a)$ is the sub-vector of X_i from a subset a of the mean explanation variables, and X_i is the observation vector of possible explanation variables for the mean of y_i, $i = 1, \cdots, n$, and $\theta = (\beta, \gamma)'$ stands for the unknown parameter vector. Note that we assume the observed dispersion explanation vector to be the same one in the true model, since we focus only on the selection of the mean explanation variables in this paper. We shall select the explanation variables by comparing different subsets according to the criterion derived from the extended quasi-likelihood function. The extended quasi-likelihood function for the candidate model a is given by

$$-2Q_a^+(y; \mu, \phi) = \sum_{i=1}^{n} \frac{D_i(y_i; \mu_i)}{\phi_i} + \sum_{i=1}^{n} \log\{2\pi\phi_i V(y_i)\},$$

where $\mu = (\mu_1, \mu_2, \cdots, \mu_n)'$, $\phi = (\phi_1, \phi_2, \cdots, \phi_n)'$,

$$D_i(y_i; \mu_i) = 2\int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)}\mathrm{d}t.$$

Also, $Q_a^+(y; \mu, \phi)$ can be written down as $Q_a^+(y; \mu, \phi) = Q_a^+(y; \beta, \gamma) = Q_a^+(y|\theta)$.

2.2 Derivation of EAIC

A common measure of the discrepancy between the true model $f(x)$ and the candidate models $g(x)$ is the Kullback-Leibler information:

$$I(f(x), g(x)) = \int f(x) \log \frac{f(x)}{g(x)}\mathrm{d}x. \tag{2.1}$$

AIC was motivated as an asymptotically unbiased estimator of $\mathsf{E}_f[\widehat{I}(f, g)]$, where E_f stands for the expectation under model $f(x)$ with respect to the sample.

We propose to replace the likelihood in (2.1) by the extended quasi-likelihood (EQL) under the working independence model and define a new discrepancy as:

$$I(f, g_a) = \int f(x)[Q_0^+(x|\theta_0) - Q_a^+(x|\theta)]\mathrm{d}x.$$

《应用概率统计》版权所有

The best fitting model g, in the considered models, is simply the model that produces the minimum K-L value. Note that

$$\begin{aligned} I(f, g_a(\cdot|\theta_0)) &= \int f(x)[Q_0^+(x|\theta_0) - Q_a^+(x|\theta_0)]\mathrm{d}x \\ &= \mathsf{E}_f[Q_0^+(x|\theta_0)] - \mathsf{E}_f[Q_a^+(x|\theta_0)]. \end{aligned} \tag{2.2}$$

The first term $\mathsf{E}_f[Q_0^+(x|\theta_0)]$ in (2.2) does not depend on the candidate model and can be viewed as a constant. Ignoring this constant, (2.2) can be expressed as

$$I(f, g_a(\cdot|\theta_0)) = -\mathsf{E}_f[Q_a^+(x|\theta_0)].$$

Given that we have data y as a sample from $f(\cdot)$, $\widehat{\theta} = \widehat{\theta}(y)$ is the maximum extended quasi-likelihood (MEQL) estimate of θ (see, e.g. [13]). Next, we intend to compute $\mathsf{E}_f[Q_a^+(x|\widehat{\theta})]$ based on Taylor series expansions.

The Taylor expansion of $Q_a^+(x|\widehat{\theta})$ around θ_0 for any given x is:

$$\begin{aligned} Q_a^+(x|\widehat{\theta}) &\approx Q_a^+(x|\theta_0) + \Big[\frac{\partial Q_a^+(x|\theta_0)}{\partial\theta}\Big]'(\widehat{\theta} - \theta_0) \\ &\quad + \frac{1}{2}(\widehat{\theta} - \theta_0)'\frac{\partial^2 Q_a^+(x|\theta_0)}{\partial\theta\partial\theta'}(\widehat{\theta} - \theta_0). \\ \mathsf{E}_f[Q_a^+(x|\widehat{\theta})] &\approx \mathsf{E}_f[Q_a^+(x|\theta_0)] + \mathsf{E}_f\Big[\frac{\partial Q_a^+(x|\theta_0)}{\partial\theta}\Big]'(\widehat{\theta} - \theta_0) \\ &\quad + \frac{1}{2}(\widehat{\theta} - \theta_0)'\Big[\mathsf{E}_f\frac{\partial^2 Q_a^+(x|\theta_0)}{\partial\theta\partial\theta'}\Big](\widehat{\theta} - \theta_0). \end{aligned} \tag{2.3}$$

The second term in (2.3) vanishes given that $\mathsf{E}_f[\partial Q_a^+(x|\theta_0)/\partial\theta] = 0$, which is at least approximately satisfied.

We define

$$I(\theta_0) = \mathsf{E}_f\Big[-\frac{\partial^2 Q_a^+(x|\theta_0)}{\partial\theta\partial\theta'}\Big],$$

then (2.3) leads to

$$\mathsf{E}_f[Q_a^+(x|\widehat{\theta})] \approx \mathsf{E}_f[Q_a^+(x|\theta_0)] - \frac{1}{2}(\widehat{\theta} - \theta_0)'I(\theta_0)(\widehat{\theta} - \theta_0). \tag{2.4}$$

Take the expectation of (2.4) with respect to $\widehat{\theta}$ (i.e., y; in fact, both with respect to truth f) and get

$$\mathsf{E}_{\widehat{\theta}}\mathsf{E}_f[Q_a^+(x|\widehat{\theta})] \approx \mathsf{E}_f[Q_a^+(x|\theta_0)] - \frac{1}{2}\mathrm{tr}\big[I(\theta_0)\mathsf{E}_{\widehat{\theta}}[(\widehat{\theta} - \theta_0)(\widehat{\theta} - \theta_0)']\big].$$

Let $\mathsf{E}_{\widehat{\theta}}\mathsf{E}_f[Q_a^+(x|\widehat{\theta})] = T$, $\mathsf{E}_{\widehat{\theta}}[(\widehat{\theta} - \theta_0)(\widehat{\theta} - \theta_0)'] = \Sigma$. Thus we have

$$T \approx \mathsf{E}_f[Q_a^+(x|\theta_0)] - \frac{1}{2}\mathrm{tr}[I(\theta_0)\Sigma]. \tag{2.5}$$

《应用概率统计》版权所有

Next, we do the expansion of $Q_a^+(x|\theta_0)$ about $\widehat{\theta}(x)$, taking account of the relationship between T and $\mathsf{E}_f[Q_a^+(x|\widehat{\theta}(x))]$. We treat x as the sample data, hence get MEQL estimate of θ for this x. It does not matter what notation we use for these sample points: x or y, because all we are interested in is an expected value, which means taking an integral over all possible points in the sample space. And note that here, $\widehat{\theta} = \widehat{\theta}(x)$, the Taylor expansion is

$$Q_a^+(x|\theta_0) \approx Q_a^+(x|\widehat{\theta}) + \Big[\frac{\partial Q_a^+(x|\widehat{\theta})}{\partial\theta}\Big]'(\theta_0 - \widehat{\theta}) + \frac{1}{2}(\theta_0 - \widehat{\theta})'\frac{\partial^2 Q_a^+(x|\widehat{\theta})}{\partial\theta\partial\theta'}(\theta_0 - \widehat{\theta}). \tag{2.6}$$

The MEQL estimate $\widehat{\theta}$ satisfies

$$\frac{\partial Q_a^+(x|\widehat{\theta})}{\partial\theta} = 0.$$

Then,

$$\mathsf{E}_f[Q_a^+(x|\theta_0)] \approx \mathsf{E}_f[Q_a^+(x|\widehat{\theta})] - \frac{1}{2}\mathrm{tr}\big[\mathsf{E}_f\{[\widehat{I}(\widehat{\theta})](\theta_0 - \widehat{\theta})(\theta_0 - \widehat{\theta})'\}\big], \tag{2.7}$$

where

$$\widehat{I}(\widehat{\theta}) = -\frac{\partial^2 Q_a^+(x|\widehat{\theta})}{\partial\theta\partial\theta'}.$$

The notation $\widehat{I}(\theta_0)$ is

$$\widehat{I}(\theta_0) = -\frac{\partial^2 Q_a^+(x|\theta_0)}{\partial\theta\partial\theta'}.$$

It is obvious that $\mathsf{E}_f[\widehat{I}(\theta_0)] = I(\theta_0)$, and $\widehat{I}(\theta_0)$ converges to $I(\theta_0)$ as $n \to \infty$.

Assume that the MEQL estimate $\widehat{\theta}$ converges to θ_0 as $n \to \infty$, thus, $\widehat{I}(\widehat{\theta}) \approx I(\theta_0)$. And we have a result

$$T(a) \approx \mathsf{E}_f[Q_a^+(x|\widehat{\theta})] - \mathrm{tr}[I(\theta_0)\Sigma].$$

Thus, we can infer that a criterion for JGLMs selection is structurally of the form

$$\widehat{T}(a) = Q_a^+(x|\widehat{\theta}) - \widehat{\mathrm{tr}}[I(\theta_0)\Sigma],$$

or

$$\mathrm{EAIC}(a) = -2Q_a^+(x|\widehat{\theta}) + 2\widehat{\mathrm{tr}}[I(\theta_0)\Sigma], \tag{2.8}$$

where

$$I(\theta_0) = \mathsf{E}_f\Big[-\frac{\partial^2 Q_a^+(x|\theta_0)}{\partial\theta\partial\theta'}\Big], \qquad \Sigma = \mathsf{E}_{\widehat{\theta}}[(\widehat{\theta} - \theta_0)(\widehat{\theta} - \theta_0)'].$$

Note that (2.8) would be exact AIC when Q_a^+ is replaced by log-likelihood function, and the second term on right hand side would be $2K_a$, where K_a is the number of parameters in the candidate model a. After Akaike's innovative derivation of AIC, people noticed

《应用概率统计》版权所有

a heuristic interpretation that the second term in AIC, $\mathrm{AIC} = -2\log(L_a(x)) + 2K_a$, can be interpreted as a penalty for increasing the size of the model, and be generalized into a function of K_a. Therefore, we employ a function of K_a, $\mathrm{F}(K_a)$, to substitute the penalty term in (2.8) in the following. The criterion got by estimating the penalty term will be presented in more general framework later. The obtained criterion is

$$\mathrm{EAIC}(a) = -2Q_a^+(x|\widehat{\theta}) + \mathrm{F}(K_a).$$

For x is treated as the sample data above, traditionally, the criterion is presented in the following way

$$\mathrm{EAIC}(a) = -2Q_a^+(y|\widehat{\theta}) + \mathrm{F}(K_a). \tag{2.9}$$

One needs to calculate the EAIC for all candidate models, and then pick up the one with the smallest EAIC.

§3. Simulation Studies

We make some simulation studies to evaluate the performance of EAIC.

3.1 Two Overdispersed Models

In practice, it is common that overdispersion or underdispersion happens[4]. In other words, the variance of y_i may be either greater or smaller than the nominal variance. Two models with overdispersion are used in the simulations.

The first one is beta-binomial model[14, 15]. It is assumed that conditionally on π_i, the response variable y_i is binomial $y_i|\pi_i \sim \mathrm{bin}(m_i, \pi_i)$, and π_i has a beta distribution $\mathrm{Beta}(\alpha_i, \delta_i)$. Hence, marginally, the distribution of y_i is not binomial but beta-binomial. Then

$$\mathsf{E}(y_i) = m_i\lambda_i, \qquad \mathsf{Var}\,(y_i) = m_i\lambda_i(1-\lambda_i)(1+(m_i-1)\theta_i),$$

where $\lambda_i = \alpha_i/(\alpha_i+\delta_i)$ and $\theta_i = 1/(\alpha_i+\delta_i+1)$. Thus the dispersion parameter is $\phi_i = 1+(m_i-1)\theta_i$.

Another overdispersed model is extra-Poisson model[16]. Conditionally on m_i, the response $y_i|m_i$ is Poisson with parameter m_i, and m_i itself is $\mathrm{Gamma}(\nu_i, \alpha_i)$, so that

$$\mathsf{E}(y_i) = \mu_i = \nu_i\alpha_i, \qquad \mathsf{Var}\,(y_i) = \nu_i\alpha_i + \nu_i\alpha_i^2 = \mu_i(1+\alpha_i).$$

And the dispersion parameter is $\phi_i = 1+\alpha_i$.

《应用概率统计》版权所有

3.2 Simulation for Beta-Binomial Distribution

For each observation (y_i, X_i), the covariate X_i is a 5×1 vector with elements generated randomly from a uniform distribution $\mathrm{U}(-1,1)$, and the y_i is generated from the Beta-Binomial model: $\mathrm{logit}(\mu_i) = X_i'\beta_0$ with $\beta_0 = (1,2,3,0,0)$. Hence, the true model contains only the first 3 components of X_i. The cluster size m_i is randomly generated from a Binomial distribution $\mathrm{bin}(40, 0.65)$. For simplicity we only consider five nested candidate models which use the first K components of X_i as explanation variables, $K = 1,2,3,4,5$, and thus we drop out the subset notation a. The structure of the dispersion that we adopt is $\phi_i = 1+(m_i-1)\theta_i = \exp(Z_{i0}'\gamma_0)$ with $\gamma_0 = (1,1)$ and Z_{i0} is a 2×1 vector with elements generated randomly from a uniform distribution $\mathrm{U}(0,1)$.

Table 1 Frequencies of models selected in 1000 replications for Beta-Binomial distribution

	$n=50$					$n=100$				
Criterion	$K=1$	2	3	4	5	$K=1$	2	3	4	5
AIC	0	0	457	256	287	0	0	400	256	344
$2K$	0	1	678	181	140	0	1	677	150	172
$2K\log(\log n)$	0	1	714	165	120	0	1	743	132	124
$2K\log n$	16	0	887	74	23	1	5	914	54	26
$2K(\log n+1)$	23	0	918	49	10	3	6	940	43	8

We respectively take $\mathrm{F}(K) = 2K$, $2K\log(\log n)$, $2K\log n$, $2K(\log n+1)$, for each sample size $n = 50, 100$. In these computations, we employ the adjusted EQL, in which the second term of $-2Q_a^+(y|\theta)$ is multiplied by $(n-K)/n$. In the simulations, for comparison, we also compute the MLE and thus AIC in which the overdispersion isn't considered. Table 1 presents the frequencies of the candidate models selected by the various criteria in 1000 replications, where $K = 3$ stands for the true model. As may be expected, the new criterion derived from EQL is much better than the naive AIC, although no exact likelihood function is involved in EQL. Note also that the larger the penalty $\mathrm{F}(K,n)$ is, the larger the probability of correct selection is, and the smaller the probability of overfitting is. This is due to the fact that the true model is included in the candidate models. In fact, using the techniques of Shao[17], it may be shown under some regular conditions that when $\mathrm{F}(K,n)\to\infty$ and $\mathrm{F}(K,n)/n\to0$ as $n\to\infty$, the probability that the criterion selects the true model tends to 1. However, this property doesn't hold when the true model is not a candidate model, and in this case, the criterion with $\mathrm{F}(K) = 2K$ may have some optimal

《应用概率统计》版权所有

properties[17].

We also compute the mean of $\widehat{\beta}$ and its mean squared error (MSE) with the 1000 replications for each model. The results for the sample size of 50 are listed in Table 2. And the MSE is computed by the formula $1000^{-1}\sum_{i=1}^{1000}(\widehat{\beta}^{(i)}-\beta_0)'(\widehat{\beta}^{(i)}-\beta_0)$, where $\widehat{\beta}^{(i)}$ is the ith estimate of β in the 1000 replications. Table 2 shows that for Model 3, the true model, the average estimates of components for β are near the true values and its MSE is the smallest.

Table 2 Estimates of coefficients and the MSEs for Beta-Binomial distribution, sample size $n = 50$

	$\overline{\widehat{\beta}_1}$	$\overline{\widehat{\beta}_2}$	$\overline{\widehat{\beta}_3}$	$\overline{\widehat{\beta}_4}$	$\overline{\widehat{\beta}_5}$	MSE($\widehat{\beta}$)	$\overline{\widehat{\gamma}_1}$	$\overline{\widehat{\gamma}_2}$	MSE($\widehat{\gamma}$)
Model 1	0.2179	–	–	–	–	13.6296	-0.4895	-0.3371	0.0038
Model 2	0.6943	2.4122	–	–	–	9.6146	-0.7108	0.1123	0.0031
Model 3	0.8852	2.0362	3.0719	–	–	0.7557	-1.9364	-1.8552	0.0102
Model 4	0.8988	2.0736	3.1049	-0.1184	–	0.9916	-1.9817	-1.7998	0.0105
Model 5	0.9017	2.0810	3.1213	-0.1245	0.0276	1.0467	-1.9783	-1.8042	0.0097

3.3 Simulation for Extra-Poisson Distribution

Table 3 Frequencies of models selected in 1000 replications for extra-Poisson distribution

	$n = 50$					$n = 100$				
Criterion	$K = 1$	2	3	4	5	$K = 1$	2	3	4	5
AIC	0	0	466	239	295	0	0	436	255	309
$2K$	0	0	650	199	151	0	0	712	184	104
$2K\log(\log n)$	0	0	742	172	86	0	0	834	126	40
$2K\log n$	0	0	957	41	2	0	0	994	6	0
$2K(\log n+1)$	0	0	980	20	0	0	0	997	3	0

In this simulation, the structure of the mean model is $\mathsf{E}(y_i)=\mu_i=\nu_i\alpha_i=\exp(X_i'\beta_0)$, where X_i is a 5×1 vector with elements independently sampled from a uniform distribution $\mathrm{U}(0,1)$, and $\beta_0=(1,2,3,0,0)$. The structure of the dispersion parameter is $\phi_i=1+\alpha_i=\exp(Z_{i0}'\gamma_0)$ with $\gamma_0=(1,1)$ and Z_{i0} is a 2×1 vector with elements generated randomly from a uniform distribution $\mathrm{U}(0,1)$. The sample sizes are $n=50,100$.

《应用概率统计》版权所有

《应用概率统计》版权所有

The simulation results on the selection frequencies in 1000 replications are presented in Table 3, and those on the estimates and their MSEs are listed in Table 4. The performances are similar to those shown in Table 1 and Table 2.

Table 4 Estimates of coefficients and the MSEs for extra-Poisson distribution, $n = 50$

	$\overline{\widehat{\beta}_1}$	$\overline{\widehat{\beta}_2}$	$\overline{\widehat{\beta}_3}$	$\overline{\widehat{\beta}_4}$	$\overline{\widehat{\beta}_5}$	MSE($\widehat{\beta}$)	$\overline{\widehat{\gamma}_1}$	$\overline{\widehat{\gamma}_2}$	MSE($\widehat{\gamma}$)
Model 1	5.0089	–	–	–	–	29.0758	3.7640	2.2871	0.0113
Model 2	2.7930	2.9702	–	–	–	13.2161	3.6775	1.2912	0.0097
Model 3	1.0030	1.9981	2.9943	–	–	0.0487	0.9983	0.9591	0.0002
Model 4	1.0022	1.9986	2.9946	-0.0016	–	0.0627	0.9895	0.9468	0.0004
Model 5	1.0022	1.9997	2.9974	-0.0021	-0.0070	0.0854	0.9769	0.9306	0.0005

To investigate the stability of EAIC, we do some simulations when some of β_0 are adjusted to be near zero. The results from 100 independent replications appear to be promising in some situations. We list the results in Table 5, 6 when $\beta_0 = (1, 1, 1, 0, 0)$, $\beta_0 = (0.1, 0.2, 3, 0, 0)$ for extra-Poisson model when $n = 50$. And the results show that the EAIC is stable for change of the coefficient.

Table 5 Frequencies of models selected in 100 replications when $\beta_0 = (1, 1, 1, 0, 0)$ for extra-Poisson distribution, sample size $n = 50$

criterion	$K = 1$	2	3	4	5
$2K$	0	2	75	16	7
$2K\log(\log n)$	0	2	85	11	2
$2K\log n$	3	11	85	1	0
$2K(\log n + 1)$	4	6	90	0	0

Table 6 Frequencies of models selected in 100 replications when $\beta_0 = (0.1, 0.2, 3, 0, 0)$ for extra-Poisson distribution, sample size $n = 50$

criterion	$K = 1$	2	3	4	5
$2K$	0	0	63	19	18
$2K\log(\log n)$	0	0	73	15	12
$2K\log n$	0	0	97	3	0
$2K(\log n + 1)$	0	0	97	3	0

Remark 1 In principle, one could use $K\log(\log n)$, $K\log n$ and $K(\log n + 1)$ as penalty. A constant before them does not change the consistency of the identified order

of the model, see [18]. We, in fact, compute the simulations without multiplier 2 in the penalties, and the result is slightly worse than those given above.

§4. An Example

We consider a data set taken from the Pignatiello and Ramberg[19], discussing an experiment related to the production of leaf springs for trucks. The same data have been studied by McCullagh and Nelder[4]. The response variable y is the free height of leaf springs and five controllable factors are B (furnace temperature), C (heating time), D (transfer time), E (hold-down time), O (quench oil temperature). The experiment was conducted following an orthogonal design. Note that the defining contrast for this design is $I = BCDE$. Thus the aliased pairs of two-factor interactions are $BD \equiv CE$, $CD \equiv BE$, $DE \equiv BC$, and the all two-factor interactions may be expressed as $BC, BD, BE, BO, CO, DO, EO$. Hence, there are twelve explanation variables in the full model. For the illustration of the proposed method, we require a model for the mean assuming the dispersion model has been known in advance. We take the variance function as $V(\mu) = 1$ and the full mean model as

$$\eta = \mu = \beta + \beta_B B + \beta_C C + \cdots + \beta_{EO} EO,$$

and the dispersion model as

$$\log \phi = \gamma + \gamma_B B + \gamma_C C.$$

All candidate models 4095 (i.e. $2^{12} - 1$) are reduced to 311 models by requiring that the main effects are included in the model if their interactions are included.

Table 7 The mean models selected and the corresponding values of criteria

Model	factors selected	$2K$	$2K\log(\log n)$	$2K\log n$	$2K(\log n + 1)$
1	$B, C, D, E, O, BO, CO, DO, EO$	**-57.99**	**-50.21**	5.17	27.17
2	B, C, E, O, CO	-48.27	-43.32	**-8.07**	5.93
3	B, C, O, CO	-41.64	-37.40	-7.19	**4.81**
4	B, C, E, O, BO, CO	-50.99	-45.33	-5.05	10.95

Table 7 presents four models followed by their corresponding EAIC values. Model 1 is selected by the criteria with $\mathrm{F}(K) = 2K$ and $2K\log(\log n)$. When $\mathrm{F}(K) = 2K\log n$, Model 2 is selected, and $\mathrm{F}(K) = 2K(\log n + 1)$ selects Model 3. Model 4 is given by McCullagh and Nelder[4].

《应用概率统计》版权所有

The parameter estimators and their corresponding standard errors and p-values of all models are investigated, but we don't present them here since the space is limited. It clearly indicates that coefficients in Model 2 and Model 3 are highly significant at the 1% level. The p-value of the coefficient BO in Model 4 is 0.0264, and not significant at the 1% level. While for Model 1, the p-values of the coefficients DO, BO, EO, D are 0.0027, 0.0166, 0.0291, 0.2614 respectively, and the other coefficients are highly significant at the 1% level. We can see that BO, EO are not significant at the 1% level, and DO is significant, however D is clearly not significant. Maybe Model 1 implies some information which isn't revealed by Model 4, and whether DO and D should be included in the mean model would need further experiments or investigations based on the background.

§5. Summary and Discussion

For likelihood-based methods, there are many well-studied model selection criteria, such as AIC. But for non-likelihood-based methods, such as the extended quasi-likelihood approaches for the joint generalized linear models, there is relatively a lack of literature on model selection. In this article, we have proposed a new criterion EAIC that works for joint generalized linear models. Using the extended quasi-likelihood approach, we only need to know the first two moments of y_i without specifying it's distribution. In simulation studies we found that the EAIC works well and stably in variable selection. The penalty term in AIC or EAIC, however, is not arbitrary; rather, it should be the asymptotic bias-correction term. We might use the bootstrap to compute $\widehat{\mathrm{tr}}[I(\theta_0)\Sigma]$. Furthermore, in principle, our proposed criterion can also be used to select variables for the dispersion model. Further applications warrant further studies.

《应用概率统计》版权所有

References

[1] Miller, A., *Subset Selection in Regression* (2nd ed.), Chapman & Hall, London, 2002.

[2] Akaike, H., Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, Eds B.N. Petrov and F.Csaki, Budapest, 1973, 267–281.

[3] Fahrmeir, L. and Kaufmann, H., Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *Ann. Statist.*, **13(1)**(1985), 342–368.

[4] McCullagh, P. and Nelder, J.A., *Generalized Linear Models* (2nd ed.), Chapman & Hall, London, 1989.

[5] Nelder, J.A. and Wedderburn, R.W.M., Generalized linear models, *J. Roy. Statist. Soc. Ser. A*, **135(3)**(1972), 370–384.

[6] Pregibon, D., Data analytic methods for generalized linear models, Ph.D thesis, University of Toronto, 1979.

[7] Hosmer, D.W., Jovanovic, B. and Lemeshow, S., Best subsets logistic regression, *Biometrics*, **45(4)** (1989), 1265–1270.

[8] Efron, B., How biased is the apparent error rate of a prediction rule? *Journal of The American Statistical Association*, **81(394)**(1986), 461–470.

[9] Jin M., Fang Y.-X. and Zhao L.-C., Variable selection in generalized linear models with canonical link functions, *Statistics and Probability Letters*, **71(4)**(2005), 371–382.

[10] Hurvich, C.M. and Tsai, C.L., Model selection for extended quasi-likelihood models in small samples, *Biometrics*, **51(3)**(1995), 1077–1084.

[11] Pan, W., Akaike's information criterion in generalized estimating equations, *Biometrics*, **57(1)**(2001), 120–125.

[12] Nelder, J.A. and Pregibon, D., An extended quasi-likelihood function, *Biometrika*, **74(2)**(1987), 221–232.

[13] Lee, Y. and Nelder, J.A., Robust design via generalized linear models, *Journal of Quality Technology*, **35(1)**(2003), 2–12.

[14] Crowder, M.J., Beta-binomial anova for proportions, *Applied Statistics*, **27(1)**(1978), 34–37.

[15] Williams, D.A., The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, **31(4)**(1975), 949–952.

[16] Lee, Y., Nelder, J.A. and Pawitan, Y., *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman & Hall/CRC, London, 2006.

[17] Shao, J., An asymptotic theory for linear model selection, *Statistica Sinica*, **7**(1997), 221–264.

[18] Nishii R., Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.*, **12(2)**1984, 758–765.

[19] Pignatiello, J.J. and Ramberg, J.S., Contribution to discussion of off-line quality-control parameter design, and the Taguchi method, *Journal of Quality Technology*, **17**(1985), 198–206.

联合广义线性模型中的变量选择

王大荣[1,2] 张忠占[1]

([1]北京工业大学应用数理学院, 北京, 100124; [2]北京工业大学实验学院, 北京, 101101)

在联合广义线性模型中, 均值和散度参数都被赋予了广义线性模型的结构, 本文主要考虑该模型的变量选择问题. 文章利用扩展拟似然函数, 提出了一个适用于联合广义线性模型的新的变量选择准则(EAIC), 该准则是Akaike信息准则的推广. 论文通过模拟研究和一个实例分析验证了该准则的效果.

关键词: Akaike信息准则, 变量选择, 联合广义线性模型, 扩展拟似然, Kullback-Leibler信息量.

学科分类号: O212.1.

《应用概率统计》版权所有