BIOMETRICS 57, 529–534
June 2001

Model Selection in Estimating Equations

Wei Pan

Division of Biostatistics, University of Minnesota,
Minneapolis, Minnesota 55455, U.S.A.
email: weip@biostat.umn.edu *and URL address:*
http://www.biostat.umn.edu/~weip

SUMMARY. Model selection is a necessary step in many practical regression analyses. But for methods based on estimating equations, such as the quasi-likelihood and generalized estimating equation (GEE) approaches, there seem to be few well-studied model selection techniques. In this article, we propose a new model selection criterion that minimizes the expected predictive bias (EPB) of estimating equations. A bootstrap smoothed cross-validation (BCV) estimate of EPB is presented and its performance is assessed via simulation for overdispersed generalized linear models. For illustration, the method is applied to a real data set taken from a study of the development of ewe embryos.

KEY WORDS: Akaike information criterion; Bayesian information criterion; Bootstrap; Cross-validation; Generalized estimating equations; Generalized linear models; Quasi-likelihood.

1. Introduction

Given an i.i.d. sample $X = (X_1, \dots, X_n)$ of size n from an unknown distribution F, we are interested in estimating a p-dimensional parameter β. For instance, in linear regression, each X_i contains a pair of a response variable y_i and a covariate vector $Z_i = (Z_{i1}, \dots, Z_{ip})'$ and β is the regression coefficient vector $\mathrm{E}(y_i) = Z_i'\beta$. To estimate β, we may use the least-squares method, and the estimate of β is obtained by solving the normal equations $\Sigma_{i=1}^n Z_i(y_i - Z_i'\beta)/n = 0$. In this article, it is assumed that the parameter of interest can be estimated through such estimating equations. The estimating equations used may be the above normal equations, likelihood score equations, quasi-likelihood score equations for the generalized linear models (GLM) (Wedderburn, 1974), generalized estimating equations (GEE) for dependent responses (Liang and Zeger, 1986), or other forms. Now with a system of estimating equations $S(\cdot \mid \beta)$ satisfying $\mathrm{E}_{X_1} S(X_1 \mid \beta) = 0$, we can estimate β by solving the following p equations:

$$S(X \mid \beta) = \frac{1}{n} \sum_{i=1}^{n} S(X_i \mid \beta) = 0. \qquad (1)$$

Denote the resulting estimate of β as $\hat{\beta}(X)$, which directly depends on the data used (i.e., on X).

Our goal is to discuss model selection when the parameter is estimated through estimating equations. Even though our proposal has potential application to other aspects of model selection, such as determining the link function for GLM, we restrict our discussion to variable selection, i.e., we want to pick up a subset of the given covariates such that the resulting model is closest to the true model. From now on, we assume that each model is indexed by the parameter vector β, which, e.g., in the context of regression, is the regression coefficients of the covariates included in the regression equation.

Model selection has been extensively discussed in the literature (e.g., Miller, 1990; Breiman, 1996a, and references therein). Most of the existing methods are based on the likelihood or squared–error-type loss functions. Akaike's information criterion (AIC) (Akaike, 1973) and its modifications, such as the Bayesian information criterion (BIC) (Schwartz, 1978), are the most widely used methods based on the likelihood function and are

$$\mathrm{AIC} = -2L_{\max}\left(X \mid \tilde{\beta}(X)\right) + 2p$$

and

$$\mathrm{BIC} = -2L_{\max}\left(X \mid \tilde{\beta}(X)\right) + p \log n,$$

where $L_{\max}$ and p are, respectively, the maximized log-likelihood and the number of parameters of the model being used. AIC and BIC differ in how to penalize the complexity of a model. If we mimic AIC and construct a model selection criterion as $S(X \mid \hat{\beta}(X))$ + penalty, it will not work since the first term is always zero according to (1). If (1) is a system of quasi-likelihood score equations, one may construct the quasi-likelihood and then use AIC or its modifications (Hurvich and Tsai, 1995; Pan, 2000). However, generally, (1) may not be a system of quasi-likelihood score equations or it may be too difficult to construct the corresponding quasi-likelihood. An important example is GEE with a nondiagonal working correlation matrix being used (McCullagh and Nelder, 1989, Section 9.3.2). On the other hand, Mallows' C_p (Mallows, 1973) is a representative of those based on the predictive mean-squared error in linear regression (see also Breiman and Spector (1992) for related bootstrap and cross-validation estimates). Bootstrapping the predictive mean-squared error in GLM has also been discussed (Shao and Tu, 1995, p. 344; Pan and Le, 2000), which, however, may not be desirable if the estimation method is not based on the least-squares criterion. Here we propose approaching this problem more directly based on the estimation equation (1).

2. New Method

Suppose we have another i.i.d. sample from F, $Y = (Y_1, \ldots, Y_n)$, that is also independent of X. By (1), $S(Y \mid \beta)$ is an unbiased and consistent estimate of $\mathrm{E}_{Y_1} S(Y_1 \mid \beta)$, which is component-wise equal to zeros under the correct model. Furthermore, since $\hat{\beta}(X)$ is also a consistent estimator of the true β under the correct model, we expect that, under the correct model, $S(Y \mid \hat{\beta}(X))$ tends to be close to zeros, at least for a large sample size n. Hence, asymptotically, $\mid S(Y \mid \hat{\beta}(X)) \mid$ is minimized component-wise to zeros under the correct model. Therefore, a model selection criterion can be constructed to minimize the expected predictive bias (EPB) of the estimating equations,

$$EPB = \mathrm{E}_X \mathrm{E}_Y \left| S\left(Y \mid \hat{\beta}(X)\right)\right|. \qquad (2)$$

We will discuss how to combine the components of EPB to form a scalar later.

Since, in practice, we only have one sample X (if we do not want to split it) and we do not know F, we recourse to the bootstrap (Efron, 1979) to estimate EPB in (2). Let F^* be the empirical distribution function of the training sample X; i.e., F^* assigns probability $1/n$ to each X_i. The bootstrap proceeds by pretending F is F^*. Hence, one can draw i.i.d. bootstrap samples X^* and Y^* from F^* and use them to estimate EPB,

$$\begin{aligned}\widehat{\mathrm{EPB}}_{\mathrm{BOOT}} &= \mathrm{E}_{X^*}\mathrm{E}_{Y^*}\left|S\left(Y^* \mid \hat{\beta}(X^*)\right)\right| \\ &= \mathrm{E}_{X^*}\frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\frac{1}{n}\Big|S\left(X_{i_1} \mid \hat{\beta}(X^*)\right) + \cdots \\ &\qquad + S\left(X_{i_n} \mid \hat{\beta}(X^*)\right)\Big|,\end{aligned}$$

which appears to be too complex to calculate even for moderate n. Note that $\widehat{\mathrm{EPB}}_{\mathrm{BOOT}}$ cannot be simplified to a more familiar form,

$$\widehat{\mathrm{EPB}}_{\mathrm{sBOOT}} = \mathrm{E}_{X^*}\left|S\left(X \mid \hat{\beta}(X^*)\right)\right|.$$

In fact, in our simulation study (not shown), we found that $\widehat{\mathrm{EPB}}_{\mathrm{sBOOT}}$ does not work well. Hence, we will not discuss the above two bootstrap estimates in more details.

An alternative resampling method is cross-validation (CV) (Allen, 1974; Stone, 1974; Geisser, 1975). A standard leave-one-out CV estimate of EPB is

$$\widehat{\mathrm{EPB}}_{1CV} = \sum_{i=1}^{n}\left|S\left(X_i \mid \hat{\beta}(X_{-i})\right)\right|/n,$$

where $X_{-i} = X - \{X_i\}$ is obtained by deleting the ith observation from the original sample. Generally, CV is almost unbiased but has a large variance. Using the bootstrap to smooth unstable estimators can reduce their variability (Breiman, 1996b; Efron and Tibshirani, 1997). Hence, we may use a bootstrap smoothed CV estimate (BCV),

$$\widehat{\mathrm{EPB}}_{\mathrm{BCV}} = \mathrm{E}_{X^*}\left|S\left(X^{*-} \mid \hat{\beta}(X^*)\right)\right|, \qquad (3)$$

where $X^{*-} = X - X^*$ contains the observations in X but not in X^*. It is confirmed in our simulation study (not shown) that BCV has a better performance than does the usual leave-one-out CV, which we will not discuss further.

In the current context, in addition to easier computation when compared with the bootstrap, there is another advantage of using BCV. Since X is finite, it is expected that X^* and Y^* (or X) are largely overlapped. Hence, some observations that appear in both X^* and Y^* (or X) are used twice in calculating $\widehat{\mathrm{EPB}}_{\mathrm{BOOT}}$ (or $\widehat{\mathrm{EPB}}_{\mathrm{sBOOT}}$), first in estimating β and then in predicting the bias of the estimating equations. This may influence the performance of the bootstrap estimates. In contrast, no observations are used twice in BCV (for any given X^*).

In general, there is no closed-form solution to (3). As usual, we use Monte Carlo simulations to approximate it. First, draw B i.i.d. bootstrap samples X^{*b} from F^*, $b = 1, \ldots, B$. Then

$$\widehat{\mathrm{EPB}}_{\mathrm{BCV}} = \sum_{b=1}^{B}\left|S\left(X^{*b-} \mid \hat{\beta}\left(X^{*b}\right)\right)\right|/B,$$

where $X^{*b-} = X - X^{*b}$ contains observations in X but not in X^{*b}. In particular, under this approximation, the BCV estimate is equivalent to the leave-one-out bootstrap estimate of Efron and Tibshirani (1997). In our simulation studies, we found that the balanced bootstrap (Davison, Hinkley, and Schechtman, 1986) is much more efficient than the above usual bootstrap. In the balanced bootstrap, each observation X_i in X is guaranteed to appear exactly a total of B times in the B bootstrap samples $X^{*1}, \ldots, X^{*B}$. Empirically, it is found that a relatively small bootstrap replication number, $B = 100$, suffices in the balanced bootstrap. We use the balanced bootstrap with $B = 100$ throughout.

Finally, we need to combine the evidence embedded in each component of $\widehat{\mathrm{EPB}}_{\mathrm{BCV}}$ of supporting the current candidate model. Ideally, if each component of the $\widehat{\mathrm{EPB}}_{\mathrm{BCV}}$ of a model is the minimum when compared with those of the other models, it is an easy decision to choose that model. However, in practice, there is a possibility that no such a model exists for the given data. In addition, the components of $\widehat{\mathrm{EPB}}_{\mathrm{BCV}}$ may not be in the same scale. For instance, the jth component of the normal equations is essentially a weighted sum of residuals and is in the scale of the jth covariate. Hence, some standardization on the covariates, e.g., centering and scaling the covariates such that they all have a zero sample mean and a unit sample standard deviation, prior to model fitting is helpful. In general, a weighted sum of the components of $\widehat{\mathrm{EPB}}_{\mathrm{BCV}}$ can be used as a summary statistic/criterion. The weights can be chosen to be inversely proportional to the variances of the components of $\widehat{\mathrm{EPB}}_{\mathrm{BCV}}$, which can be estimated directly based on the bootstrap samples. We note that the choice of such weights may influence the efficiency of the resulting method, but it is asymptotically valid to use any weights because each component of EPB is asymptotically minimized to zero under the correct model. This point is also evidenced in our simulation study (see BCV_1 below).

Here we discuss two simple ways of forming a summary statistic using the components of $\widehat{\text{EPB}}_{\text{BCV}}$. In the first one, we only consider one component of $\widehat{\text{EPB}}_{\text{BCV}}$. Which component to use may depend on which covariate is believed to be the most important or most of interest. Another intuitive choice is the one corresponding to a covariate that is included in every or most candidate models. For instance, suppose we are using the least-squares method to compare three linear regression models: $\mathrm{E}(y) = \beta_1 + \beta_2 Z_2$, $\mathrm{E}(y) = \beta_1 + \beta_3 Z_3$, and $\mathrm{E}(y) = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3$. Then all three have the intercept term. Hence, we may compare the first component of the $\widehat{\text{EPB}}_{\text{BCV}}$, $\widehat{\text{EPB}}_{BCV,1}$, of each model and pick up the one with the smallest $\widehat{\text{EPB}}_{BCV,1}$.

The second way is to consider the largest candidate model that covers all (or most) of the candidate models. In the above example, the third model is such a model. Hence, for each model, we calculate a possibly expanded $\widehat{\text{EPB}}_{\text{BCV}}$ with three components as for model 3, even though the third or second component of the normal equations S is not used in estimating β in model 1 or model 2 (since it does not have parameter β_3 or β_2). Then we compare equally weighted averages of the three components of the $\widehat{\text{EPB}}_{\text{BCV}}$'s for the models. A more specific example is discussed in Section 3.2.

The above two approaches can be considered as two special cases of choosing the weights to calculate a weighted average of the components of $\widehat{\text{EPB}}_{\text{BCV}}$, which is expanded to include all the components as for the largest model. The first approach uses the weights as a vector of all zeros except a one for one component, whereas the second has a weight vector with all elements equal to $1/p$, where p is the number of the components of $\widehat{\text{EPB}}_{\text{BCV}}$ based on the largest model. In this article, we investigate the performance of the above two implementations, denoted as BCV_1 and BCV_a, respectively, leaving it open to investigate other more general alternatives in the future.

3. Application to Overdispersed Generalized Linear Models

3.1 *Overdispersion, Underdispersion, and Beta-Binomial Distribution*

Suppose our observations are possibly an i.i.d. sample of binomial data X with $X_i = (y_i, m_i, Z_i)$, $i = 1, \ldots, n$, where the response y_i usually is assumed to have a binomial distribution $y_i \sim \text{bin}(m_i, \pi_i)$ and Z_i is a vector of covariates. We consider fitting a linear logistic regression model, $\text{logit}(\pi_i) = Z_i'\beta$. Our goal is to select a model that only uses a subset of given covariates Z_i and has the best predictive performance. If the distributional assumption on y_i is (approximately) correct, we can use AIC and its modifications to do model selection as usual.

However, in practice, it is common that overdispersion or underdispersion happens (McCullagh and Nelder, 1989, p. 124). In other words, the variance of y_i is greater or smaller than the nominal variance $m_i\pi_i(1 - \pi_i)$ under the binomial assumption, in which case, obviously, y_i does not have a binomial distribution. In the quasi-likelihood approach, we do not need to specify the distribution of y_i but do its first two moments, and it still yields consistent estimates. The quasi-likelihood estimate of β is obtained from a system of estimation equations, which is the same as the likelihood score equations in the usual logistic regression.

Overdispersion can be modeled, among others, by the beta-binomial distribution (Williams, 1975; Crowder, 1978). It is assumed that, conditional on $P_i = \pi_i$, y_i is binomial $\text{bin}(m_i, \pi_i)$ and P_i has a beta distribution $\text{beta}(\alpha_i, \delta_i)$. Hence, marginally, the distribution of y_i is not binomial but rather is beta binomial, i.e.,

$$\Pr(y_i = y) = \binom{m_i}{y} \frac{B(\alpha_i + y, m_i + \delta_i - y)}{B(\alpha_i, \delta_i)}, \qquad y = 1, 2, \ldots, m_i,$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ and Γ is the gamma function. Then

$$\mathrm{E}(y_i) = m_i\mu_i$$
$$\mathrm{var}(y_i) = m_i\mu_i(1 - \mu_i)(1 + (m_i - 1)\phi_i),$$

where $\mu_i = \alpha_i/(\alpha_i + \delta_i)$ and $\phi_i = 1/(\alpha_i + \delta_i + 1)$.

3.2 *Simulations*

To investigate the behavior of our proposed methods, we did some simulation studies where the data were generated from the binomial model or the beta-binomial model. The covariate matrix Z is an $n \times 5$ matrix with elements i.i.d. from a uniform distribution $\mathrm{U}(-1, 1)$. Data were generated from the beta-binomial model, $\text{logit}(\mu_i) = Z_i'\beta$ with $\beta = (1, 2, 3, 0, 0)$. Hence, the true model consists of the first three columns of Z. For simplicity, the candidate models are sequentially nested: model p consists of the first $1, 2, \ldots, p$ columns of Z. In the beta-binomial model, we take $\phi_i \equiv \phi$ as a constant, here as $\phi = 0$ (corresponding to the binomial data), 0.2, 0.5, and 0.8, respectively. The sample size is $n = 40$ and cluster sizes m_i are randomly generated from a binomial distribution $\text{bin}(20, 0.65)$. The regression coefficients were estimated via the quasi-likelihood method. The quasi-likelihood score equation for model k is $S_{(k)} = (S_1(y, Z), \ldots, S_k(y, Z))' = 0$ with

$$S_j(y, Z) = \sum_{i=1}^{n} Z_{ij}(y_i - m_i\mu_i),$$

where $j = 1, \ldots, k$ and $k = 1, \ldots, 5$. For each model, BCV_1 is the $\widehat{\text{EPB}}_{\text{BCV}}$ in (3) with S replaced by S_1. To obtain BCV_a, we do the following for each model. First, we calculate $\widehat{\text{EPB}}_{\text{BCV}}$ in (3) by replacing S with $S_{(5)}$. Then BCV_a is an equally weighted average of the five components of $\widehat{\text{EPB}}_{\text{BCV}}$. Finally, we select the model with the smallest BCV_1 or BCV_a, depending on which implementation is being used.

The results are shown in Table 1. AIC and BIC were calculated under the binomial distribution, which is correct only when $\phi = 0$. It is observed that, in general, BCV is the best. AIC and BIC work well only when $\phi = 0$ but become worse and worse when their distributional assumption (i.e., $\phi = 0$) is more and more seriously violated. In contrast, BCV is not much influenced by the increasing values of ϕ. Between the two implementations of BCV, BCV_a performs better than BCV_1, though their difference is not huge. That means that using only a subset of the components of $\widehat{\text{EPB}}_{\text{BCV}}$ loses some information and is less effective; however, the resulting loss of information is not substantial.

Table 1

Frequency of the models selected by different criteria from 100 *independent replications. The correct beta-binomial model is the one with* $p = 3$.

	$\phi = 0$					$\phi = 0.2$					$\phi = 0.5$					$\phi = 0.8$				
Criterion/p	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
AIC	0	0	71	15	14	0	0	36	20	44	0	0	19	20	61	0	0	14	17	69
BIC	0	0	91	8	1	0	0	55	20	25	0	0	34	19	47	0	0	23	18	59
BCV_1	0	0	68	19	13	0	0	75	17	8	0	2	67	14	17	0	1	74	16	19
BCV_a	0	0	79	14	7	0	0	83	13	4	0	0	76	13	11	0	1	86	9	4

In practice, it is likely that some important covariates are missing. To mimic this, we did a simulation study by using $\beta = (1, 2, 3, 0, 0, 1)$, but the sixth column of the covariate matrix Z is omitted. All the other aspects are similar to the previous ones. The results are shown in Table 2. The conclusion is the same as before. BCV is the best. AIC and BIC do not work unless $\phi = 0$. Between the two versions of BCV, BCV_a is better. In addition, BCV does not appear to be much influenced by the missing covariate.

3.3 *An Example*

We consider a data set taken from a study of the development of embryos for 119 Texel ewes (Engel and te Brake, 1993; the data set is listed in their Appendix A.1). The study goal was to determine how the embryonic development is influenced by the following covariates. The first is the treatment (T): almost half of the 119 ewes were treated with Fecundin and the remaining served as a control group. The ewes were classified into four age groups (A): ≤ 0.5, 0.5–1.5, 1.5–2.5, and > 2.5 years. There were two mating periods (M), which started on October 1 and October 22, 1986, respectively.

As before, let m_i and y_i denote the numbers of ova and fetuses for the ith ewe, respectively. Engel and te Brake (1993) also introduced a factor N representing the two levels of the number of ova: $N = 1$ if $m_i > 2$; $N = 0$ otherwise. The goal was to model the probability π_i of an ovum's developing into a fetus using the variables T, A, M, and N. A logistic regression approach that takes account of overdispersion or underdispersion was proposed by Engel and te Brake. After fitting several models, they found that the interaction terms were not needed. Hence, the following main-effects model was fitted:

$$\text{Model 1:} \qquad \text{logit}(\pi) = T + A + M + N,$$

in a simplified notation. In particular, the existence of underdispersion relative to binomial variation was verified. The estimated ϕ is -0.1175 when the variance of y_i is modeled as $m_i\pi_i(1 - \pi_i)(1 + (m_i - 1)\phi)$. Using the quasi-likelihood approach, the results of fitting several reasonable candidate models are listed in Table 3.

Table 2

Frequency of the models selected by different criteria from 100 *independent replications. The correct beta-binomial model is the one with* $p = 3$ *after one relevant covariate is omitted.*

	$\phi = 0$					$\phi = 0.5$				
Criterion/p	1	2	3	4	5	1	2	3	4	5
AIC	0	0	60	17	23	0	0	19	19	62
BIC	0	0	81	11	8	0	0	33	19	48
BCV_1	0	0	67	17	16	1	4	62	18	15
BCV_a	0	0	78	13	9	0	0	81	9	10

Based on the result for the above main-effects model (model 1 in Table 3), it is obvious that the mating period (M) is not important and age (A) appears to be important. On the other hand, the effects of the treatment and the variable N are not so clear cut. They are at borderline significant with p-values slightly larger than 0.05. If we use the commonly adopted cut-off point of keeping a variable at the significance level 0.1 (or 0.15 or 0.2), as in many automatic sequential variable selection procedures for linear regression or standard logistic regression in statistical packages (e.g., SAS Proc Reg and Proc Logistic), then we will choose model 2 by deleting the variable M from model 1. However, as Engel and te Brake (1993) observed, if the variable N is thrown out (to form model 3), then the treatment will, but age will not, be significant (see Table 3). Engel and te Brake offered a possible explanation for this phenomenon.

An interesting question is what will happen if we delete the age variable rather than N in model 2 (to form model 4). Then, as shown in Table 3, the treatment is still significant and N is not. Hence, either way, we reach a further reduced model (model 5) that includes the treatment as the only covariate, which turns out to be highly significant.

Each of the five candidate models somewhat appears to be reasonable. Now we want to rank them in terms of their predictive ability. Using our proposed model selection criterion $\widehat{\text{EPB}}_{\text{BCV}}$, model 5 is ranked as the best, followed by model 4. The full main-effects model (model 1) is the worst. This result seems to be reasonable by considering the effects of the covariates as shown in Table 3. Also, the rankings are the same when various implementations of the proposal, such as only the first component (BCV_1), only the second component (BCV_2), an equally weighted sum of all the components except the first one BCV_{a-1}), and an equally weighted sum of all components (BCV_a) of $\widehat{\text{EPB}}_{\text{BCV}}$, are used (Table 3).

We note that an advantage of a flexible model selection criterion, such as the proposed $\widehat{\text{EPB}}_{\text{BCV}}$ here, is that it can be used to compare two nonnested models, such as model 3 and model 4, which, however, cannot be accomplished by the usual hypothesis testing.

Table 3
P-values for each covariate in the five candidate models and their associated values of various versions of $\widehat{\mathrm{EPB}}_{\mathrm{BCV}}$ *for the ewe data*

Covariate	Model 1	Model 2	Model 3	Model 4	Model 5
Fecundin	0.0568	0.0583	0.0059	0.0208	0.0048
Age ≤ 0.5	0.0399	0.0383	0.1262	—	—
Age 0.5–1.5	0.0561	0.0539	0.0876	—	—
Age 1.5–2.5	0.3280	0.3249	0.2884	—	—
Mating Oct. 1	0.7788	—	—	—	—
Ova ≥ 3	0.0772	0.0763	—	0.2828	—
BCV_1	0.1883	0.0930	0.0932	0.0886	0.0874
BCV_2	0.0913	0.0891	0.0899	0.0873	0.0856
BCV_{a-1}	0.0579	0.0551	0.0555	0.0515	0.0489
BCV_a	0.0765	0.0605	0.0609	0.0568	0.0544

4. Discussion

For likelihood-based methods, there are many well-studied model selection criteria such as AIC and BIC. But for non-likelihood-based methods, such as the quasi-likelihood and GEE approaches for the generalized linear models, there is relatively a lack of literature on model selection. In this article, we propose a general model selection criterion that minimizes the expected predictive bias (EPB), naturally based on the estimation method in the class of estimating equation approaches. The bootstrap smoothed cross-validation is used to estimate the EPB. Through simulation studies, we found that our proposal works well for overdispersed generalized linear models. Further applications to model selection in other contexts of using estimating equations are conceptually straightforward and warrant future studies.

Acknowledgement

The author is grateful to two referees, an associate editor, and the editor for helpful suggestions. In particular, the stimulating comments from the associate editor helped motivate the presentation of EPB and discussion of its various implementations.

Résumé

La sélection de modèles est une étape nécessaire dans beaucoup d'analyses de régression. Mais, pour les méthodes fondées sur les équations d'estimation, comme la quasi-vraisemblance ou les équations d'estimation généralisées (GEE), il semble y avoir peu de méthodes de sélection bien étudiées. Dans ce papier, nous proposons un nouveau critère de sélection, qui minimise le biais de prédiction attendu (EPB) des équations d'estimation. Un estimateur de bootstrap lissé avec validation croisé (BCV) de EPB est présenté et sa performance évaluée par simulation de modèles linéaires généralisés sur-dispersés. A titre d'illustration, la méthode est appliquée aux données d'une étude du développement des embryons de brebis.

References

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

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16,** 125–127.

Breiman, L. (1996a). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24,** 2350–2383.

Breiman, L. (1996b). Bagging predictors. *Machine Learning* **26,** 123–140.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression: The X random case. *International Statistical Review* **60,** 291–319.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics* **27,** 34–37.

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1986). Efficient bootstrap simulations. *Biometrika* **75,** 417–431.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7,** 1–26.

Efron, B. and Tibshirani, R. J. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* **92,** 548–560.

Engel, B. and te Brake, J. (1993). Analysis of embryonic development with a model for under- or overdispersion relative to binomial variation. *Biometrics* **49,** 269–279.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70,** 320–328.

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

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15,** 661–675.

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

Miller, A. J. (1990). *Subset Selection in Regression.* London: Chapman and Hall.

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

Pan, W. and Le, C. T. (2001). Bootstrap model selection in generalized linear models. *Journal of Agricultural, Biological, and Environmental Statistics* **6,** 49–61.

Schartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics* **6,** 461–464.

Shao, J. and Tu, D. S. (1995). *The Jackknife and Bootstrap.* New York: Springer.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* **36,** 111–147.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.

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

Received December 1998. *Revised September* 2000.
Accepted October 2000.