33

Discrepancy-Based Model Selection Criteria Using Cross Validation

Joseph E. Cavanaugh, Simon L. Davies, and Andrew A. Neath

Department of Biostatistics, The University of Iowa
Pfizer Global Research and Development, Pfizer, Inc.
Department of Mathematics and Statistics, Southern Illinois University

Abstract:

A model selection criterion is often formulated by constructing an approximately unbiased estimator of an expected discrepancy, a measure that gauges the separation between the true model and a fitted approximating model. The expected discrepancy reflects how well, on average, the fitted approximating model predicts "new" data generated under the true model. A related measure, the estimated discrepancy, reflects how well the fitted approximating model predicts the data at hand.

In general, a model selection criterion consists of a goodness-of-fit term and a penalty term. The natural estimator of the expected discrepancy, the estimated discrepancy, corresponds to the goodness-of-fit term of the criterion. However, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. It therefore serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the penalty term.

Cross validation provides a technique for developing an estimator of an expected discrepancy which need not be adjusted for bias. The basic idea is to construct an empirical discrepancy that evaluates an approximating model by assessing how accurately each case-deleted fitted model predicts the deleted case.

The preceding approach is illustrated in the linear regression framework by formulating estimators of the expected discrepancy based on Kullback's I-divergence and the Gauss (error sum of squares) discrepancy. The traditional criteria that arise by augmenting the estimated discrepancy with a bias adjustment term are the Akaike information criterion and Mallows' conceptual predictive statistic. A simulation study is presented.

Keywords and phrases: AIC, Mallows' Cp, PRESS

33.1 Introduction

A model selection criterion is often formulated by constructing an approximately unbiased estimator of an expected discrepancy, a measure that gauges the separation between the true model and a fitted approximating model. The natural estimator of the expected discrepancy, the estimated discrepancy, corresponds to the goodness-of-fit term of the selection criterion.

The expected discrepancy reflects how well, on average, the fitted approximating model predicts "new" data generated under the true model. On the other hand, the estimated discrepancy reflects how well the fitted approximating model predicts the data at hand. By evaluating the adequacy of the fitted model based on its ability to recover the data used in its own construction, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. Thus, the estimated discrepancy serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the penalty term of the selection criterion.

Cross validation provides a technique for developing an estimator of an expected discrepancy which need not be adjusted for bias. The basic idea involves constructing an empirical discrepancy that evaluates an approximating model by assessing how accurately each case-deleted fitted model predicts the deleted case.

Cross validation facilitates the development of model selection procedures based on predictive principles. In this work, we attempt to establish a more explicit connection between cross validation and traditional discrepancy-based model selection criteria, such as the Akaike (1973) information criterion and Mallows' (1973) conceptual predictive statistic.

In section 2, we outline the framework for discrepancy-based selection criteria. In section 3, we discuss the bias-adjustment approach for developing a model selection criterion, and in section 4, we present the cross-validatory approach. Section 5 features examples of discrepancy-based selection criteria developed using both approaches. The linear regression framework is considered. In section 6, we present simulation results to evaluate the performance of the criteria. Our results show that the cross-validatory criteria compare favorably to their traditional counterparts, offering greater protection from overfitting in small-sample settings.

33.2 Framework for Discrepancy-Based Selection Criteria

Suppose we have an n-dimensional data vector $y = (y_1, \ldots, y_n)'$, where the y_i's may be scalars or vectors and are assumed to be independent. A parametric model is postulated for y. Let θ denote the vector of model parameters.

Let $F(y)$ denote the joint distribution function for y under the generating or "true" model, and let $F_i(y_i)$ denote the marginal distribution for y_i under this model. Let $G(y, \theta)$ denote the joint distribution function for y under the candidate or approximating model.

A *discrepancy* is a measure of disparity between $F(y)$ and $G(y, \theta)$, say $\Delta(F, G)$, which satisfies

$$\Delta(F, G) \geq \Delta(F, F).$$

A discrepancy is not necessarily a formal metric, which would additionally require that $\Delta(F, F) = 0$, that $\Delta(F, G)$ is symmetric in $F(y)$ and $G(y, \theta)$, and that $\Delta(F, G)$ satisfies the triangle inequality. However, the measure $\Delta(F, G)$ serves the same basic role as a distance: i.e., as the dissimilarity between $F(y)$ and $G(y, \theta)$ becomes more pronounced, the size of $\Delta(F, G)$ should increase accordingly.

We will consider discrepancies of the following form:

$$\Delta(F, G) = \Delta(\theta) = \sum_{i=1}^{n} \mathrm{E}_{F_i} \left\{ \delta_i(y_i; \theta) \right\}.$$

In the preceding, $\delta_i(y_i; \theta)$ represents a function that gauges the accuracy with which the i^{th} case y_i is predicted under the approximating model (parameterized by θ).

Let $\widehat{\theta}$ denote an estimator of θ. The *overall discrepancy* results from evaluating the discrepancy between $F(y)$ and $G(y, \theta)$ at $\theta = \widehat{\theta}$:

$$\Delta(\widehat{\theta}) = \sum_{i=1}^{n} \mathrm{E}_{F_i} \left\{ \delta_i(y_i, \theta) \right\} |_{\theta = \widehat{\theta}}.$$

The *expected (overall) discrepancy* results from averaging the overall discrepancy over the sampling distribution of $\widehat{\theta}$:

$$\mathrm{E}_F \left\{ \Delta(\widehat{\theta}) \right\} = \sum_{i=1}^{n} \mathrm{E}_F \left\{ \mathrm{E}_{F_i} \left\{ \delta_i(y_i, \theta) \right\} |_{\theta = \widehat{\theta}} \right\}.$$

The *estimated discrepancy* is given by

$$\widehat{\Delta}(\widehat{\theta}) = \sum_{i=1}^{n} \delta_i(y_i, \widehat{\theta}).$$

Model selection criteria are often constructed by obtaining a statistic that has an expectation which is equal to $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$ (at least approximately). In the next two sections, we explore the bias-adjustment and cross-validatory approaches to obtaining such statistics.

33.3 The Bias-Adjustment Approach to Developing a Criterion

The overall discrepancy $\Delta(\widehat{\theta})$ is not a statistic since its evaluation requires knowledge of the true distribution $F(y)$. The estimated discrepancy $\widehat{\Delta}(\widehat{\theta})$ is a statistic and can be used to estimate the expected discrepancy $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$. However, $\widehat{\Delta}(\widehat{\theta})$ serves as a biased estimator.

Consider writing $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$ as follows:

$$\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\} = \mathrm{E}_F\left\{\widehat{\Delta}(\widehat{\theta})\right\} + \left[\mathrm{E}_F\left\{\Delta(\widehat{\theta}) - \widehat{\Delta}(\widehat{\theta})\right\}\right].$$

The bracketed quantity on the right is often referred to as the *expected optimism* in judging the fit of a model using the same data as that which was used to construct the fit. The expected optimism is positive, implying that $\widehat{\Delta}(\widehat{\theta})$ is a negatively biased estimator of $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$. In order to correct for the negative bias, we must evaluate or approximate the bias adjustment represented by the expected optimism.

There are numerous approaches for contending with the bias adjustment. These approaches include deriving an asymptotic approximation for the adjustment (e.g., Akaike, 1973), deriving an exact expression (e.g., Hurvich and Tsai, 1989), or obtaining an approximation using Monte Carlo simulation (e.g., Bengtsson and Cavanaugh, 2006).

We will now introduce a general cross-validatory estimate of the expected discrepancy that need not be adjusted for bias. As a model selection criterion, such an estimate has several advantages over a bias-adjusted counterpart.

First, the form of a cross-validatory criterion facilitates a convenient interpretation of the statistic as a measure of predictive efficacy. Broadly speaking, such a criterion evaluates an approximating model by gauging how accurately each case-deleted fitted model predicts a "new" datum, represented by the deleted case. The criterion provides a composite measure of accuracy resulting from the systematic deletion and prediction of each case. In contrast, the form of a bias-adjusted criterion is more esoteric, consisting of an additive combination of a goodness-of-fit term and a penalty term. These terms work in opposition to balance the competing modeling objectives of conformity to the

data and parsimony. However, the connection between achieving such a balance and predictive efficacy is not transparent.

Second, a cross-validatory criterion serves as an exactly unbiased estimator of a cross-validatory expected discrepancy that may be viewed as a natural analogue of the expected discrepancy $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$. This unbiasedness holds without imposing conditions that may restrict the applicability of the resulting criterion, conditions which are routinely required for the justifications of bias corrections.

Third, the difference between the cross-validatory expected discrepancy and its traditional counterpart converges to zero. Thus, in large sample settings, the cross-validatory criterion estimates the traditional expected discrepancy with negligible bias.

The key assumption for establishing the asymptotic equivalence of the cross-validatory and traditional expected discrepancies is that the difference in expectation between the full-data estimator and any case-deleted estimator is $o(n^{-1})$. The proof is provided in the appendix. For settings where the method of estimation is maximum likelihood and the approximating model is correctly specified or overspecified, the asymptotic condition on the estimators is verified.

33.4 The Cross-Validatory Approach to Developing a Criterion

Let $y[i]$ denote the data set y with the i^{th} case y_i excluded. Let $\widehat{\theta}[i]$ denote an estimator of θ based on $y[i]$.

Recall that the overall discrepancy is defined as

$$\Delta(\widehat{\theta}) = \sum_{i=1}^{n} \mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}}. \tag{33.1}$$

Now consider the following variant of the overall discrepancy:

$$\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n]) = \sum_{i=1}^{n} \mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}[i]}. \tag{33.2}$$

The expected (overall) discrepancy corresponding to (33.1) is given by

$$\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\} = \sum_{i=1}^{n} \mathrm{E}_F\left\{\mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}}\right\}; \tag{33.3}$$

the expected (overall) discrepancy corresponding to (33.2) is given by

$$\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\} = \sum_{i=1}^{n} \mathrm{E}_F\left\{\mathrm{E}_{F_i}\left\{\delta_i(y_i,\theta)\right\}|_{\theta=\widehat{\theta}[i]}\right\}. \tag{33.4}$$

Under the assumption that the difference in expectation between the full-data estimator $\widehat{\theta}$ and any case-deleted estimator $\widehat{\theta}[i]$ is $o(n^{-1})$, it can be established that the difference between $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$ and $\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\}$ is $o(1)$. (The proof is outlined in the appendix.) Hence, an unbiased estimator of (33.4) is approximately unbiased for (33.3).

Now the estimated discrepancy

$$\widehat{\Delta}(\widehat{\theta}) = \sum_{i=1}^{n} \delta_i(y_i, \widehat{\theta})$$

is *negatively biased* for (33.3). However, the empirical discrepancy defined as

$$\widehat{\Delta}^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n]) = \sum_{i=1}^{n} \delta_i(y_i, \widehat{\theta}[i]) \tag{33.5}$$

is *exactly unbiased* for (33.4). The justification of this fact is straightforward.

Since $\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\} \approx \mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$, it follows that $\widehat{\Delta}^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])$ is *approximately unbiased* for $\mathrm{E}_F\left\{\widehat{\Delta}(\widehat{\theta})\right\}$. Thus, the empirical discrepancy $\widehat{\Delta}^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])$

(a) estimates $\mathrm{E}_F\left\{\Delta^*(\widehat{\theta}[1],\ldots,\widehat{\theta}[n])\right\}$ without bias,

(b) estimates $\mathrm{E}_F\left\{\Delta(\widehat{\theta})\right\}$ with negligible bias for large n.

33.5 Examples in the Linear Regression Setting

Consider a setting where a continuous response variable is to be modeled using a linear regression model.

Under the approximating model, assume the y_i are independent with mean $x_i^{'}\beta$ and variance σ^2. Let $\theta = (\beta^{'}\sigma^2)^{'}$. Further, let $g(y,\theta)$ denote the approximating density for y, and let $g_i(y_i,\theta)$ denote the approximating density for y_i.

Kullback's I-divergence and the Gauss (error sum of squares) discrepancy have applicability to many modeling frameworks, including linear regression. In the context of model selection, the I-divergence may be defined as

$$\Delta_{\mathrm{I}}(\theta) = \mathrm{E}_F\left\{-2\ln g(y,\theta)\right\} = \sum_{i=1}^{n} \mathrm{E}_{F_i}\left\{\delta_i^{\mathrm{I}}(y_i;\theta)\right\}, \tag{33.6}$$

where $\delta_i^{\mathrm{I}}(y_i;\theta) = -2\ln g_i(y_i,\theta)$. (See Linhart and Zucchini, 1986, p. 18; Hurvich and Tsai, 1989, p. 299.) For the linear regression framework, the Gauss discrepancy may be expressed as

$$\Delta_{\mathrm{G}}(\theta) = \mathrm{E}_F\left\{\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right\} = \sum_{i=1}^{n}\mathrm{E}_{F_i}\left\{\delta_i^{\mathrm{G}}(y_i;\theta)\right\}, \tag{33.7}$$

where $\delta_i^{\mathrm{G}}(y_i;\theta) = (y_i - x_i'\beta)^2$. (See Linhart and Zucchini, 1986, p. 118.)

Provided that the approximating model of interest is correctly specified or overspecified, the Akaike information criterion provides an asymptotically unbiased estimator of the expected discrepancy corresponding to (33.6). In the present setting, AIC is given by

$$\mathrm{AIC} = -2\ln g(y,\widehat{\theta}) + 2(p+1),$$

where p denotes the number of regression parameters, and $\widehat{\theta}$ denotes the maximum likelihood estimator (MLE) of θ. Under the additional assumption that the errors are normally distributed, the "corrected" Akaike information criterion, AICc, provides an *exactly* unbiased estimator of the expected discrepancy (Hurvich and Tsai, 1989). AICc is given by

$$\mathrm{AICc} = -2\ln g(y,\widehat{\theta}) + \frac{2(p+1)n}{n-p-2}.$$

Provided that the largest approximating model in the candidate collection is correctly specified or overspecified, a simple variant of Mallows' conceptual predictive statistic (with identical selection properties) provides an exactly unbiased estimator of the expected discrepancy corresponding to (33.7). Mallows' statistic is given by

$$\mathrm{C}_p = \frac{\mathrm{SSE}}{\mathrm{MSE}_L} + (2p - n),$$

where SSE denotes the error sum of squares. The aforementioned variant is given by $(\mathrm{C}_p + n)\mathrm{MSE}_L$, where MSE_L denotes the error mean square for the largest approximating model.

The cross-validatory criterion (33.5) based on the I-divergence (33.6) is given by

$$\sum_{i=1}^{n} -2\ln g_i(y_i,\widehat{\theta}[i]),$$

where $\widehat{\theta}[i]$ represents the case-deleted MLE of θ. Assuming normal errors, the preceding reduces to

$$\sum_{i=1}^{n}\ln\hat{\sigma}_{-i}^2 + \sum_{i=1}^{n}\frac{(y_i - \hat{y}_{i,-i})^2}{\hat{\sigma}_{-i}^2},$$

where $\hat{y}_{i,-i}$ denotes the fitted value for y_i based on the case-deleted data set $y[i]$, and $\hat{\sigma}^2_{-i}$ denotes the case-deleted MLE for σ^2. Davies, Neath, and Cavanaugh (2005) refer to the preceding criterion as the *predictive divergence criterion*, PDC. (See also Stone, 1977.)

The cross-validatory criterion (33.5) based on the Gauss discrepancy (33.7) is given by

$$\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2,$$

the well known PRESS (predictive sum of squares) statistic (Allen, 1974).

The preceding development indicates that PDC and PRESS may be respectively viewed as the cross-validatory analogues of AIC and C_p. In simulation studies, such cross-validatory criteria compare favorably to their traditional counterparts. In settings where the generating model is among the collection of candidate models under consideration, the cross-validatory criteria tend to select the correctly specified model more frequently and to select overspecified models less frequently than their bias-adjusted analogues. In the next section, we present representative sets from the simulation studies we have conducted.

33.6 Linear Regression Simulations

Consider a setting where samples of size n are generated from a true linear regression model of the form $y_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + x_{i6} + \epsilon_i$, where $\epsilon_i \sim iid\ N(0, 4)$. For every sample, nested candidate models with an intercept and k regressor variables ($k = 1, \ldots, 12$) are fit to the data. (Note that $p = k + 1$.) Specifically, the first model fit to each sample is based on only the covariate x_{i1}, the second is based on the covariates x_{i1} and x_{i2}, etc. The sixth fitted model ($k = 6,\ p = 7$) is correctly specified. Subsequent fitted models are overspecified, since they contain the regressor variables for the generating model (x_{i1} through x_{i6}) in addition to extraneous covariates ($x_{i7}, \ldots, x_{i,12}$). All regressor variables are generated as *iid* replicates from a uniform distribution over the interval $(0, 10)$.

Suppose our objective is to search the candidate collection for the fitted model which serves as the best approximation to the truth. The strength of the approximation is reflected via the expected discrepancy, either (33.3) or (33.4).

We present six simulation sets based on the preceding setting. In the first three sets, we examine the effectiveness of AIC, AICc, and PDC at selecting the correctly specified model. In the next three sets, we examine the effectiveness of C_p and PRESS at achieving the same objective. We group the criterion selections into three categories: underfit (UF), correctly specified (CS), and overfit (OF).

The results of sets 1–3 are presented in Table 1. These sets feature sample sizes of $n = 25$, 50, and 75, respectively. In each set, PDC obtains the most correct selections, followed by AICc. The performance of AIC is relatively poor. In general, AIC favors overspecified models in settings where the sample size is insufficient to ensure the adequacy of the criterion's bias correction.

Table 1. Selection results for AIC, AICc, PDC.

Set	n	Selections	Criterion		
			AIC	AICc	PDC
1	25	UF	0	1	18
		CS	418	913	929
		OF	582	86	53
2	50	UF	0	0	0
		CS	606	815	870
		OF	394	185	130
3	75	UF	0	0	0
		CS	685	789	833
		OF	315	211	167

For the results from set 3, the figure on page 10 features a plot of criterion averages versus k. The expected overall discrepancies (33.3) and (33.4) are also plotted. The plot illustrates the exact unbiasedness of PDC for (33.4) and AICc for (33.3), yet also indicates the negative bias of AIC for (33.3) resulting from the poor bias approximation. This negative bias creates the criterion's propensity to favor over parameterized models.

The figure also reflects the similarity of the curves for the expected overall discrepancies (33.3) and (33.4). As the sample size increases, the difference between these curves becomes negligible. Thus, in large sample settings, the selections of PDC, AICc, and AIC should agree. However, in smaller sample settings, where the predictive accuracy of the selected model may be greatly diminished by the inclusion of unnecessary covariates, PDC and its target discrepancy favor more parsimonious models.

The results of sets 4–6 are presented in Table 2. These sets feature sample sizes of $n = 15$, 20, and 25, respectively. In sets 4 and 5, PRESS obtains more correct selections than C_p. This is mainly due to the difference in the behaviors of the targeted discrepancies: in smaller sample settings, (33.4) penalizes more heavily than (33.3) to protect against the inflation in predictive variability that accompanies the incorporation of extraneous regressors. However, in this setting, the asymptotic equivalence of (33.3) and (33.4) takes effect for relatively small n: in the third set, where n is 25, the selection patterns are the same for the two criteria.

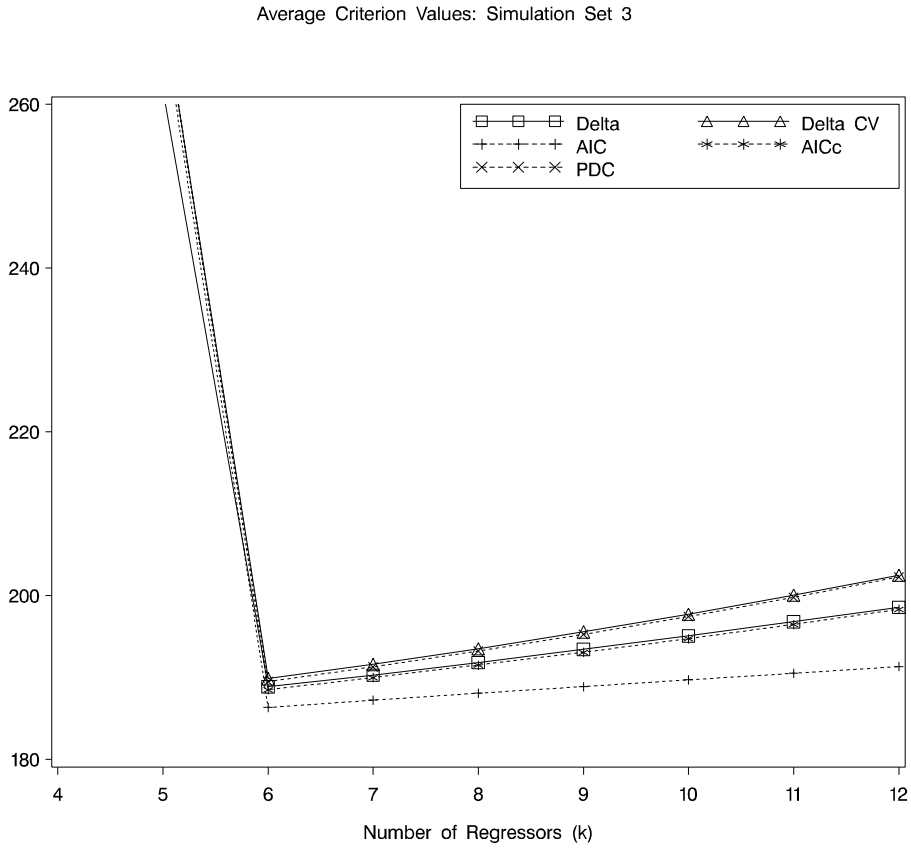

Table 2. Selection results for C_p and PRESS.

Set	n	Selections	Criterion	
			C_p	PRESS
4	15	UF	22	19
		CS	491	587
		OF	487	394
5	20	UF	4	1
		CS	634	671
		OF	362	328
6	25	UF	0	0
		CS	668	668
		OF	332	332

The simulation results presented constitute a small yet representative sample from a larger simulation study. In general, our results show that cross-validatory criteria perform well relative to their traditional counterparts, offering greater protection from overfitting in smaller-sample settings, and exhibiting similar behavioral tendencies in larger-sample settings.

In conclusion, cross-validatory model selection criteria provide an appealing alternative to traditional bias-adjusted selection criteria (such as AIC and C_p). For many traditional expected discrepancies, a cross-validatory criterion may be easily formulated. Such a criterion is approximately unbiased for the traditional expected discrepancy, and exactly unbiased for an analogous expected discrepancy based on cross validation. The preceding unbiasedness properties hold without requiring stringent conditions which may limit applicability. Moreover, the form of a cross-validatory criterion facilitates a convenient, intuitive interpretation of the statistic as a measure of predictive efficacy.

Acknowledgment

The authors wish to express their appreciation to an anonymous referee for valuable feedback which helped to improve the original version of this paper.

Appendix

In what follows, we establish the asymptotic equivalence of (33.3) and (33.4); specifically

$$\mathrm{E}_F\{\Delta(\widehat{\theta})\} - \mathrm{E}_F\{\Delta^*(\widehat{\theta}[1], \ldots, \widehat{\theta}[n])\} = o(1). \tag{33.8}$$

We assume that the estimator $\widehat{\theta}$ converges weakly to some interior point θ_* of the parameter space Θ: i.e., $\widehat{\theta} = \theta_* + o_p(1)$. Thus, we should also have $\widehat{\theta}[i] = \theta_* + o_p(1)$ for each $i = 1, \ldots, n$.

Let $\Delta_i(\theta) = \mathrm{E}_{F_i}\left\{\delta_i(y_i, \theta)\right\}$, so that

$$\Delta(\widehat{\theta}) = \sum_{i=1}^{n} \Delta_i(\widehat{\theta}), \qquad \Delta^*(\widehat{\theta}[1], \ldots, \widehat{\theta}[n]) = \sum_{i=1}^{n} \Delta_i(\widehat{\theta}[i]),$$

and

$$\Delta(\widehat{\theta}) - \Delta^*(\widehat{\theta}[1], \ldots, \widehat{\theta}[n]) = \sum_{i=1}^{n} \left[\Delta_i(\widehat{\theta}) - \Delta_i(\widehat{\theta}[i])\right]. \tag{33.9}$$

Our approach is to show that for each i,

$$\mathrm{E}_F\{\Delta_i(\widehat{\theta}) - \Delta_i(\widehat{\theta}[i])\} = o(n^{-1}). \tag{33.10}$$

Clearly, (33.10) in conjunction with (33.9) will establish (33.8).

We assume that $\Delta_i(\theta)$ has continuous first-order derivatives with respect to θ. Let $D_i(\theta) = \partial\Delta_i(\theta)/\partial\theta$. Using a first-order Taylor series expansion, we have

$$\Delta_i(\widehat{\theta}) = \Delta_i(\widehat{\theta}[i]) + D_i(\xi)'(\widehat{\theta} - \widehat{\theta}[i]), \tag{33.11}$$

where $\xi = \widehat{\theta}[i] + \lambda(\widehat{\theta} - \widehat{\theta}[i])$ for some $0 \leq \lambda \leq 1$. Thus, ξ converges to θ_*, and $D_i(\xi)$ converges to $D_i(\theta_*)$. From (33.11), we therefore have

$$\Delta_i(\widehat{\theta}) - \Delta_i(\widehat{\theta}[i]) = [D_i(\theta_*) + o_p(1)]'(\widehat{\theta} - \widehat{\theta}[i]). \tag{33.12}$$

Now assume that

$$\mathrm{E}_F\{(\widehat{\theta} - \widehat{\theta}[i])\} = o(n^{-1}). \tag{33.13}$$

Since $D_i(\theta_*) = O(1)$, (33.13) together with (33.12) implies (33.10).

We now verify that (33.13) holds in a specific setting, namely one in which the method of estimation is maximum likelihood, and the approximating model is correctly specified or overspecified.

The latter assumption implies that the joint distribution function $F(y)$ under the generating model belongs to the same class as the joint distribution function $G(y, \theta)$ under the approximating model. We may therefore write $F(y)$ as $F(y, \theta_o)$, where θ_o is an interior point of Θ. Thus, θ_o defines the "true" parameter vector.

Let $L(\theta|y) = \prod_{i=1}^{n} g_i(y_i, \theta)$ denote the likelihood function for θ based on y. Assume that each of the likelihood contributions $g_i(y_i, \theta)$ is differentiable and suitably bounded: specifically, that for some function $h(\cdot)$ with $\int h(u)\, du < \infty$, we have

$$\left|\frac{\partial g_i(u, \theta)}{\partial\theta}\right| < h(u) \quad \text{for all} \quad (u, \theta). \tag{33.14}$$

For the overall likelihood $L(\theta|y)$, assume that $\ln L(\theta|y)$ has first- and second-order derivatives which are continuous and bounded over Θ. Let

$$V_n(\theta) = -\frac{1}{n}\ln L(\theta|y), \quad V_n^{(1)}(\theta) = \frac{\partial V_n(\theta)}{\partial\theta}, \quad \text{and} \quad V_n^{(2)}(\theta) = \frac{\partial^2 V_n(\theta)}{\partial\theta\partial\theta'}.$$

Here, $\widehat{\theta} = \mathrm{argmin}_{\theta\in\Theta} V_n(\theta)$; i.e., $\widehat{\theta}$ is the maximum likelihood estimator of θ.

Let $W_n(\theta) = \mathrm{E}_F\{V_n(\theta)\}$. Assume that as $n \to \infty$, $W_n(\theta)$ converges to a function $W(\theta)$ uniformly in θ over Θ, and that $W(\theta)$ has a unique global minimum at θ_o. Further, suppose that $W(\theta)$ has first- and second-order derivatives which are continuous and bounded over Θ. Let $W^{(2)}(\theta) = (\partial^2 W(\theta))/(\partial\theta\partial\theta')$. Assume that $W^{(2)}(\theta)$ is positive definite in a neighborhood of θ_o.

Finally, assume that $V_n(\theta)$ converges to $W(\theta)$, that $V_n^{(2)}(\theta)$ converges to $W^{(2)}(\theta)$, and that the convergence is uniform in θ over Θ.

The preceding regularity conditions are typical of those used to ensure the consistency and the asymptotic normality of the maximum likelihood estimator

of $\widehat{\theta}$. (See, for instance, section 3 of Cavanaugh and Neath, 1999.) In the setting at hand, the point of convergence θ_* for the estimator $\widehat{\theta}$ corresponds to the true parameter vector θ_o.

Expand $V_n^{(1)}(\widehat{\theta})$ about θ_o to obtain

$$\begin{aligned} 0 &= V_n^{(1)}(\widehat{\theta}) \\ &= V_n^{(1)}(\theta_o) + V_n^{(2)}(\widetilde{\theta})(\widehat{\theta} - \theta_o), \end{aligned}$$

where $\widetilde{\theta} = \theta_o + \gamma(\widehat{\theta} - \theta_o)$ for some $0 \leq \gamma \leq 1$. Then,

$$\widehat{\theta} = \theta_o - [V_n^{(2)}(\widetilde{\theta})]^{-1} V_n^{(1)}(\theta_o).$$

The preceding relation along with the assumed regularity conditions and the consistency of $\widehat{\theta}$ leads to

$$\widehat{\theta} = \theta_o - [W^{(2)}(\theta_o) + o_p(1)]^{-1} V_n^{(1)}(\theta_o).$$

Now without loss of generality, take $\widehat{\theta}[i] = \widehat{\theta}[1]$. Then we have

$$\widehat{\theta} - \widehat{\theta}[1] = -[W^{(2)}(\theta_o) + o_p(1)]^{-1} [V_n^{(1)}(\theta_o) - V_{n-1}^{(1)}(\theta_o)], \tag{33.15}$$

where

$$V_n^{(1)}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln g_i(y_i, \theta)}{\partial \theta} \quad \text{and} \quad V_{n-1}^{(1)}(\theta) = -\frac{1}{n-1} \sum_{i=2}^{n} \frac{\partial \ln g_i(y_i, \theta)}{\partial \theta}.$$

Note that

$$V_n^{(1)}(\theta_o) - V_{n-1}^{(1)}(\theta_o) = -\frac{1}{n} \frac{\partial \ln g_1(y_1, \theta_o)}{\partial \theta} - \frac{1}{n} V_{n-1}^{(1)}(\theta_o). \tag{33.16}$$

Using (33.16) in conjunction with (33.15), we obtain

$$n(\widehat{\theta} - \widehat{\theta}[1]) = [W^{(2)}(\theta_o) + o_p(1)]^{-1} \left[\frac{\partial \ln g_1(y_1, \theta_o)}{\partial \theta} + V_{n-1}^{(1)}(\theta_o) \right]. \tag{33.17}$$

Now the assumed regularity conditions along with the consistency of the maximum likelihood estimator allow us to conclude that the difference between $V_{n-1}^{(1)}(\theta_o)$ and $V_{n-1}^{(1)}(\widehat{\theta}[1]) = 0$ is $o_p(1)$, which implies that $V_{n-1}^{(1)}(\theta_o) = o_p(1)$. Moreover, one can argue that $\mathrm{E}_F \{(\partial \ln g_1(y_1, \theta_o))/(\partial \theta)\} = 0$. This result is established by exchanging the order of differentiation and integration, which is permissible via the Lebesgue Dominated Convergence Theorem under the imposed assumption (33.14). The preceding results along with (33.17) allow us to argue

$$\mathrm{E}_F\{n(\widehat{\theta} - \widehat{\theta}[1])\} = o(1).$$

Thus, (33.13) is established.

References

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

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

3. Bengtsson, T. and Cavanaugh, J. E. (2006). An improved Akaike information criterion for state–space model selection, *Computational Statistics and Data Analysis*, **50**, 2635–2654.

4. Cavanaugh, J. E. and Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion, *Communications in Statistics – Theory and Methods*, **28**, 49–66.

5. Davies, S. L., Neath, A. A. and Cavanaugh J. E. (2005). Cross validation model selection criteria for linear regression based on the Kullback–Leibler discrepancy, *Statistical Methodology*, **2**, 249–266.

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

7. Linhart, H. and Zucchini, W. (1986). *Model Selection*, Wiley, New York.

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

9. Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B*, **39**, 44–47.