An Alternate Version of the Conceptual Predictive Statistic Based on a Symmetrized Discrepancy Measure

Joseph E. Cavanaugh[1]

Department of Biostatistics, The University of Iowa

Andrew A. Neath

Department of Mathematics and Statistics, Southern Illinois University Edwardsville

Simon L. Davies

Pfizer Global Pharmaceuticals, Inc.

Abstract

The conceptual predictive statistic, C_p, is a widely used criterion for model selection in linear regression. C_p serves as an estimator of a discrepancy, a measure that reflects the disparity between the generating model and a fitted candidate model. This discrepancy, based on scaled squared error loss, is asymmetric: an alternate measure is obtained by reversing the roles of the two models in the definition of the measure. We propose a variant of the C_p statistic based on estimating a symmetrized version of the discrepancy targeted by C_p. We claim that the resulting criterion provides better protection against overfitting than C_p, since the symmetric discrepancy is more sensitive towards detecting overspecification than its asymmetric counterpart. We illustrate our claim by presenting simulation results. Finally, we demonstrate the practical utility of the new criterion by discussing a modeling application based on data collected in a cardiac rehabilitation program at University of Iowa Hospitals and Clinics.

Keywords: discrepancy function, linear models, model selection

[1] Corresponding author: Department of Biostatistics, C22 GH; 200 Hawkins Drive; The University of Iowa; Iowa City, IA, 52242. Phone: (319) 384-5024. Fax: (319) 384-5018. E-mail: joe-cavanaugh@uiowa.edu

1. Introduction

An important component of any linear modeling problem consists of determining an appropriate size and form for the design matrix. Improper specification may substantially impact both estimators of the model parameters and predictors of the response variable. Specifically, underspecification may lead to results which are severely biased, whereas overspecification may lead to results with unnecessarily high variability.

The determination of a suitable design matrix can often be facilitated by the use of a model selection criterion, such as Mallows' (1973) conceptual predictive statistic, C_p. A selection criterion scores every fitted model in a candidate collection in accordance with how effectively the model balances the competing objectives of parsimony and conformity to the data. Ideally, undesirable scores are assigned not only to models which omit essential variables, but also to models which adequately accommodate the data yet involve extraneous or irrelevant variables.

C_p serves as an estimator of a discrepancy, a measure that reflects the disparity between the generating model and a fitted candidate model. This discrepancy, based on scaled squared error loss, is asymmetric: a companion measure is obtained by reversing the roles of the two models in the definition of the measure. A symmetric discrepancy can be formed by adding the discrepancy and its counterpart.

When used to evaluate fitted approximating models which are improperly specified, the discrepancy targeted by C_p is more sensitive towards detecting underfitted models than overfitted models. As a result, C_p has the propensity to select models that are overparameterized. We argue that the symmetrized discrepancy provides a more balanced gauge of model misspecification, and therefore serves as a better basis for the formulation of a selection criterion. With this motivation, we propose a variant of the C_p statistic based on the symmetrized discrepancy. We refer to this variant as *symmetrized* C_p, SC_p.

We claim that SC_p provides better protection against overfitting than C_p. We illustrate this claim through a simulation study. We demonstrate the practical utility of the new criterion by presenting a modeling application based on data collected in a cardiac rehabilitation program at University of Iowa Hospitals and Clinics.

In Section 2, we present the framework for regression model selection. We discuss discrepancy-based model selection criteria in Section 3. In Section 4, we outline the derivation of SC$_p$. Our simulation study is featured in Section 5, and our application in Section 6.

2. Framework for Regression Model Selection

Assume the *generating* or *true model* $\mathcal{M}_o$ corresponds to the normal linear regression model

$$y = X_o\beta_o + e_o, \qquad e_o \sim N_n(0, \sigma_o^2\ I).$$

Here, β_o is a $p_o \times 1$ parameter vector, and X_o is an $n \times p_o$ design matrix of full-column rank.

Consider an *approximating model* $\mathcal{M}_p$ which has the same fundamental structure as $\mathcal{M}_o$, but may be based on a different set of predictor variables:

$$y = X\beta + e, \qquad e \sim N_n(0, \sigma^2\ I),$$

where β is a $p \times 1$ parameter vector, and X is an $n \times p$ design matrix of full-column rank. Let $\widehat{\beta}$, $\widehat{\sigma}^2$ denote the maximum likelihood estimators (MLEs) of β, σ^2 based on this model.

In practice, we consider a *candidate collection* $\mathcal{F}$ of models that consists of different approximating models based on different design matrices of various ranks, say

$$\mathcal{F} = \{\mathcal{M}_{p_1}, \mathcal{M}_{p_2}, \ldots, \mathcal{M}_{p_L}\}.$$

Note that some of the models in this collection may be based on design matrices having the same rank yet different column spaces. For ease of notation, we do not include an index to delineate between such models.

Each approximating model in $\mathcal{F}$ is fit to the data, resulting in a candidate collection of fitted models $\{\widehat{\mathcal{M}}_{p_1}, \widehat{\mathcal{M}}_{p_2}, \ldots, \widehat{\mathcal{M}}_{p_L}\}$. Our objective is to identify the fitted model which is "nearest" to the generating model $\mathcal{M}_o$. To facilitate this objective, we require a measure that provides a suitable reflection of the disparity between the generating model and a fitted model. Such a measure is often called a *discrepancy*.

To define a discrepancy, let $\mathcal{M}_o$ denote the generating model and let $\widehat{\mathcal{M}}$ denote a fitted candidate model. Assume that $\mathcal{M}_o$ and $\widehat{\mathcal{M}}$ belong to the class $\mathcal{C}$. Consider a mapping Δ from $\mathcal{C} \times \mathcal{C}$ to $\Re$ that has the property

$$\Delta(\widehat{\mathcal{M}}, \mathcal{M}_o) \geq \Delta(\mathcal{M}_o, \mathcal{M}_o). \tag{2.1}$$

A discrepancy is defined as a mapping Δ that satisfies the property (2.1). (See Linhart and Zucchini, 1986, p. 11.) Such a mapping need not be a formal metric, since the range of the mapping need not be necessarily positive, the mapping need not be symmetric, and the mapping need not satisfy the triangle inequality. However, for a discrepancy to be useful in the present context, the measure Δ should have the same utility as a distance: the magnitude of $\Delta(\widehat{\mathcal{M}}, \mathcal{M}_o)$ should increase in accordance with the disparity between $\mathcal{M}_o$ and $\widehat{\mathcal{M}}$. The definition of a discrepancy is inherently vague so as to accommodate a wide variety of separation measures.

Model selection criteria are often derived by constructing approximately unbiased estimators of expected discrepancies. For instance, the Akaike (1973, 1974) information criterion (AIC) estimates an expected discrepancy based on Kullback-Leibler information (Kullback, 1968). Akaike's (1969) final prediction error, FPE, and the predictive sum of squares (Allen, 1974), PRESS, estimate an expected discrepancy based on squared prediction error. Mallows' (1973) C_p estimates an expected discrepancy based on scaled squared error loss.

The selection patterns of discrepancy-based selection criteria can often be linked to the characteristics of the underlying discrepancy. For instance, certain model selection criteria exhibit a propensity towards selecting underfitted or overfitted models. This propensity may be exceptionally strong in smaller-sample settings. If the targeted discrepancy is asymmetric, a more balanced criterion can often be obtained by formulating an estimator of a symmetrized version of the discrepancy (e.g., Cavanaugh, 1999, 2004; Kim and Cavanaugh, 2005). A key advantage of this approach is that the new criterion often retains the fundamental large-sample properties as well as some of the interpretive characteristics of the original. In the next two sections, we explore these conceptual and methodological ideas in the context of the conceptual predictive statistic.

3. The C_p Discrepancy and Its Counterpart

Mallows' conceptual predictive statistic is traditionally defined in terms of variance estimators based on both the fitted candidate model of interest and the largest fitted model in the candidate collection. Let $\mathcal{M}_*$ denote the largest model in $\mathcal{F}$, with design matrix X_* of rank p_*. Let $\widehat{\sigma}^2_*$ and MSE_* respectively denote the MLE of the variance and the mean square error based on the fitted model $\widehat{\mathcal{M}}_*$. Recall that $\widehat{\sigma}_*^2 = \left[(n - p*)MSE_*\right]/n$. The conceptual predictive statistic is then defined as

$$\mathrm{C}_p = \frac{n\,\widehat{\sigma}^2}{MSE_*} + 2p - n.$$

The discrepancy targeted by C_p is given by

$$\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right) = \frac{\parallel X\widehat{\beta} - X_o\beta_o \parallel^2}{\sigma_o^2}.$$

This discrepancy is asymmetric in that an alternate measure may be obtained by reversing the roles of the two models in the definition. Take

$$\begin{aligned}\Delta_2\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right) &= \frac{\parallel X_o\beta_o - X\widehat{\beta}\|^2}{\widehat{\sigma}^2} \\ &= \Delta_1\left(\mathcal{M}_o, \widehat{\mathcal{M}}_p\right).\end{aligned}$$

C_p provides an approximately unbiased estimator of $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$, where E represents expectation taken with respect to the generating model. That is, as we argue more formally in the next section, $E\left(\mathrm{C}_p\right) \approx E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. Thus, in choosing the fitted model corresponding to the minimum value of C_p, one is hoping to identify the fitted model which is "nearest" on average to $\mathcal{M}_o$, where proximity is gauged by the Δ_1 discrepancy.

Let $H = X\left(X'X\right)^{-1}X'$ be the projection matrix onto $C\left(X\right)$, the column space of X. Then

$$E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\} = E\left\{\frac{\parallel X\widehat{\beta} - HX_o\beta_o\|^2}{\sigma_o^2}\right\} + \frac{\parallel X_o\beta_o - HX_o\beta_o \parallel^2}{\sigma_o^2}. \tag{3.1}$$

Linhart and Zucchini (1986) refer to the two terms on the right-hand side of (3.1) as the *discrepancy due to estimation* and the *discrepancy due to approximation*, respectively. Consider the former. We have

$$\begin{aligned}\frac{\parallel X\widehat{\beta} - HX_o\beta_o\|^2}{\sigma_o^2} &= \frac{\parallel H\left(y - X_o\beta_o\right)\parallel^2}{\sigma_o^2} \\ &= \frac{\left(y - X_o\beta_o\right)' H \left(y - X_o\beta_o\right)}{\sigma_o^2}.\end{aligned}$$

The preceding quadratic form is a random variable distributed as χ^2 with degrees of freedom $df = r\left(H\right) = r\left(X\right) = p$, where $r\left(\cdot\right)$ denotes the rank. Thus,

$$E\left\{\frac{\parallel X\widehat{\beta} - HX_o\beta_o \parallel^2}{\sigma_o^2}\right\} = p.$$

The discrepancy due to estimation thereby increases with the complexity of the model, characterized by the design matrix X and its rank $r\left(X\right)$.

Let

$$\lambda = \frac{\| X_o\beta_o - HX_o\beta_o \|^2}{\sigma_o^2} \tag{3.2}$$

denote the discrepancy due to approximation. If $C\left(X_o\right) \subset C\left(X\right)$, then the model $\mathcal{M}_p$ is *overspecified*, since the model includes all of the regressors in the true model in addition to certain false regressors. In this case, $\lambda = 0$. If $C\left(X_o\right) \nsubseteq C\left(X\right)$, then we call the model $\mathcal{M}_p$ *underspecified*, since the model excludes at least some of the regressors in the true model. In this case, the size of λ reflects the extent of the underspecification. Therefore, the expected value of the discrepancy targeted by C$_p$ (3.1) is given as

$$E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\} = p + \lambda. \tag{3.3}$$

Based on our convention, the class of underspecified models is comprised of two characteristically different types of candidate models: those for which $C\left(X\right) \subset C\left(X_o\right)$, where no false regressors are included, and models for which $C(X) \nsubseteq C(X_o)$ and $C(X) \nsupseteq C(X_o)$, where at least some false predictors are included. Models in the first subclass are necessarily smaller than the true model, whereas models in the second subclass may be larger than, smaller than, or the same size as the true model.

The concepts of underspecification and overspecification are pertinent in considering the behavior of C$_p$ as a model selection criterion. A well documented problem with the use of C$_p$ is its tendency to select overfitted models. (See, for instance, McQuarrie and Tsai, 1998, pp. 35–45.) This propensity can be at least partly attributed to the behavior of Δ_1. Relative to its counterpart Δ_2, the discrepancy Δ_1 is more sensitive towards detecting underspecification, and less sensitive towards detecting overspecification.

We can gain insight into the preceding property by comparing the expected values of the two measures. From linear model theory, we have that $X\widehat{\beta}$ is independent of $\widehat{\sigma}^2$, with

$$E\left\{\| X_o\beta_o - X\widehat{\beta}\|^2\right\} = \sigma_o^2\left(p + \lambda\right),$$

as derived for equation (3.3). Now,

$$\begin{aligned}\frac{\| \, y - X\widehat{\beta} \, \|^2}{\sigma_o^2} &= \frac{\| \, (I-H) \, y \|^2}{\sigma_o^2} \\ &= \frac{y' \, (I-H) \, y}{\sigma_o^2}\end{aligned}$$

is a random variable distributed as χ^2 with degrees of freedom $df = r\,(I-H) = n-p$ and noncentrality parameter

$$\begin{aligned}\frac{(X_o\beta_o)' \, (I-H) \, (X_o\beta_o)}{\sigma_o^2} &= \frac{\| \, X_o\beta_o - HX_o\beta_o \, \|^2}{\sigma_o^2} \\ &= \lambda.\end{aligned}$$

Thus,

$$\widehat{\sigma}^2 \sim \frac{\sigma_o^2}{n} \cdot \chi^2 \, (n-p, \lambda) \, .$$

To find $E\left\{\Delta_2\left(\widehat{\mathcal{M}_p}, \mathcal{M}_o\right)\right\}$, we need $E\,\{1/\widehat{\sigma}^2\}$. With the intention of providing better mathematical tractability, we will use the second-order approximation

$$E\left\{\frac{1}{W}\right\} \approx \frac{1}{E\,(W)}\left(1 + \frac{Var\,(W)}{E^2\,(W)}\right).$$

(The preceding is derived by taking a second-order Taylor series expansion of $1/W$ in the argument W about its mean $E(W)$, and then taking the expected values of both sides of the resulting expression.) Since $E\,(\widehat{\sigma}^2) = (\sigma_o^2/n)(n-p+\lambda)$ and $Var\,(\widehat{\sigma}^2) = (\sigma_o^2/n)^2\,(2(n-p)+4\lambda)$, we have

$$E\left\{\frac{1}{\widehat{\sigma}^2}\right\} \approx \frac{n}{\sigma_o^2}\left(\frac{1}{n-p+\lambda}\right)\left(1 + \frac{2(n-p)+4\lambda}{(n-p+\lambda)^2}\right).$$

We thereby obtain

$$\begin{aligned}E\left\{\Delta_2\left(\widehat{\mathcal{M}_p}, \mathcal{M}_o\right)\right\} &= E\left\{\frac{1}{\widehat{\sigma}^2}\right\} \cdot E\left\{\| \, X_o\beta_o - X\widehat{\beta}\|^2\right\} \\ &\approx \frac{n(p+\lambda)}{n-p+\lambda}\left(1 + \frac{2(n-p)+4\lambda}{(n-p+\lambda)^2}\right). \qquad (3.4)\end{aligned}$$

We will later use approximation (3.4) in the development of our criterion. For now, let's simplify the comparison between Δ_1 and Δ_2 by comparing (3.3) and

$$E\left\{\Delta_2\left(\widehat{\mathcal{M}_p}, \mathcal{M}_o\right)\right\} \approx \frac{n(p+\lambda)}{n-p+\lambda}. \qquad (3.5)$$

As the degree of underspecification becomes more pronounced, λ grows. One can see from the partial derivatives of (3.3) and (3.5) that, on average, as λ grows, the growth in Δ_1 exceeds that of Δ_2. For overspecified models, $\lambda = 0$. As the degree of overspecification becomes more pronounced, p grows. Again from partial differentiation, on average, as p grows, the growth in Δ_2 exceeds that of Δ_1. Hence, Δ_1 is more sensitive than Δ_2 towards detecting underspecification, and Δ_2 is more sensitive than Δ_1 towards detecting overspecification.

The discrepancy Δ_1 is scaled by the true error variance, σ_o^2. Since σ_o^2 is constant across the candidate collection of fitted models, candidate models are judged according to Δ_1 exclusively by whether they yield an accurate estimate of the mean vector. The discrepancy Δ_2 is scaled using the approximating model variance $\widehat{\sigma}^2$, so candidate models are additionally judged according to Δ_2 by whether they yield a maximum likelihood estimate that provides an appropriate accounting of the variance. The inadequacy of an underspecified model with large model variance will be better reflected by Δ_1; the redundancy of an overspecified model with small model variance will be better reflected by Δ_2.

Since C$_p$ targets Δ_1, we see that while C$_p$ provides relatively strong protection from choosing an underfitted model, it provides relatively weak protection from choosing an overfitted model. We can further illustrate the model selection properties of C$_p$ by exploring its target discrepancy Δ_1 and its counterpart Δ_2 in a simple example. Suppose that samples of size n are generated from the true model $\mathcal{M}_o$:

$$y_i = 1 + x_{i1} + x_{i2} + x_{i3} + 0.5\, x_{i4} + e_{oi},$$

where $e_{oi} \sim iid\ N\,(0,4)$, and all covariates x_{ij} are distributed as independent replicates from $N\,(0,12)$. For each sample, we fit three approximating models to the data:

$$\begin{aligned}
\mathcal{M}_4 \quad &: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i, \\
\mathcal{M}_5 \quad &: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, \\
\mathcal{M}_{11} \quad &: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \ldots + \beta_{10} x_{i,10} + e_i,
\end{aligned}$$

where $e_i \sim iid\ N\,(0,\sigma^2)$. Candidate model $\mathcal{M}_5$ is correctly specified. Candidate model $\mathcal{M}_4$ is underspecified whereas candidate model $\mathcal{M}_{11}$ is overspecified. We generate 1000 samples from $\mathcal{M}_o$ in order to approximate $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ and $E\left\{\Delta_2\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ for each of

$\mathcal{M}_4$, $\mathcal{M}_5$, and $\mathcal{M}_{11}$. We consider sample sizes $n = 15$, $n = 30$, and $n = 100$. The simulated averages for $\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right) = \parallel X\widehat{\beta} - X_o\beta_o \parallel^2/\sigma_o^2$ and $\Delta_2\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right) = \parallel X\widehat{\beta} - X_o\beta_o \parallel^2/\hat{\sigma}^2$ are given in Table 1. The table also features the model selections obtained by C$_p$ for each set of 1000 samples.

Table 1. C$_p$ selections and simulated expected values of Δ_1, Δ_2 for three candidate models.

n	Model	C$_p$ Selections	$\overline{\Delta}_1$	$\overline{\Delta}_2$
15	$\mathcal{M}_4$	117	12.16	11.04
	$\mathcal{M}_5$	**644**	**4.94**	**9.14**
	$\mathcal{M}_{11}$	239	10.93	73.57
30	$\mathcal{M}_4$	6	23.72	16.31
	$\mathcal{M}_5$	**872**	**4.99**	**6.57**
	$\mathcal{M}_{11}$	122	10.94	19.64
100	$\mathcal{M}_4$	0	75.80	45.76
	$\mathcal{M}_5$	**924**	**4.92**	**5.29**
	$\mathcal{M}_{11}$	76	10.94	12.57

In all three settings, we see that Δ_2 penalizes the overspecified model $\mathcal{M}_{11}$ to a greater extent than Δ_1. However, Δ_2 is not a uniformly more discriminating measure than Δ_1, since Δ_1 penalizes the underspecified model $\mathcal{M}_4$ to a greater extent than Δ_2. Among the incorrect model selections, C$_p$ chooses the overspecified model $\mathcal{M}_{11}$ much more frequently than the underspecified model $\mathcal{M}_4$. This propensity is entirely consistent with the behavior of the discrepancy Δ_1 that is targeted by C$_p$.

4. The Symmetrized Conceptual Predictive Statistic

To attenuate the overfitting tendencies of C$_p$, we propose a model selection criterion based on the symmetrized discrepancy

$$\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right) = \Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right) + \Delta_2\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right). \qquad (4.1)$$

The symmetrized conceptual predictive statistic, SC$_p$, will serve as an approximately unbiased estimator of $E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. To derive this criterion, recall that $\mathcal{M}_*$ denotes the largest model in the candidate collection $\mathcal{F}$, with design matrix X_* of rank p_*. Henceforth, we impose the assumption that the largest candidate model subsumes the true model. That is, $C\left(X_o\right) \subseteq C\left(X_*\right)$. This assumption is also required for the derivation of C$_p$.

An estimate of the expected symmetrized discrepancy function $E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ will be developed by first defining an estimate of the parameter λ.

Theorem 4.1: Under the conditions set forth previously,

$$\widehat{\lambda} = \left(n - p_* - 2\right)\frac{\widehat{\sigma}^2}{\widehat{\sigma}_*^2} - \left(n - p - 2\right) \tag{4.2}$$

is an unbiased estimator of λ, as defined in (3.2).

Proof: We begin by considering the quadratic forms $y'\left(H_* - H\right)y$ and $y'(I - H_*)y$, where H_* is the projection matrix onto $C\left(X_*\right)$. These quadratic forms, which appear in the F statistic for testing model $\mathcal{M}_p$ against model $\mathcal{M}_*$, are independently distributed chi-square random variables. We have

$$\frac{1}{\sigma_o^2}y'\left(H_* - H\right)y \sim \chi^2,$$

with degrees of freedom $df_1 = r\left(H_* - H\right) = p_* - p$ and noncentrality parameter

$$ncp_1 = \frac{1}{\sigma_o^2}\| \, H_* X_o \beta_o - H X_o \beta_o \, \|^2.$$

Also,

$$\frac{1}{\sigma_o^2}y'\left(I - H_*\right)y \sim \chi^2,$$

with degrees of freedom $df_2 = r\left(I - H_*\right) = n - p_*$ and noncentrality parameter

$$ncp_2 = \frac{1}{\sigma_o^2}\| \, X_o \beta_o - H_* X_o \beta_o \, \|^2.$$

Since $H_* X_o \beta_o = X_o \beta_o$ from the condition $C\left(X_o\right) \subseteq C\left(X_*\right)$, it follows that $ncp_1 = \lambda$ and $ncp_2 = 0$. So

$$E\left\{\frac{y'(H_* - H)y}{y'(I - H_*)y}\right\} = \frac{p_* - p + \lambda}{n - p_* - 2}.$$

Now, $n\,\widehat{\sigma}^2 = y'(I - H)y = y'(I - H_*)y + y'(H_* - H)y$. Thus,

$$\widehat{\sigma}^2 = \widehat{\sigma}_*^2 + \frac{y'(H_* - H)y}{n},$$

$$E\left\{\frac{y'(H_* - H)y}{y'(I - H_*)y}\right\} = E\left\{\frac{\widehat{\sigma}^2 - \widehat{\sigma}_*^2}{\widehat{\sigma}_*^2}\right\},$$

and

$$E\left\{\frac{\widehat{\sigma}^2}{\widehat{\sigma}_*^2}\right\} = \frac{p_* - p + \lambda}{n - p_* - 2} + 1.$$

Algebraic manipulation yields the stated result for $\widehat{\lambda}$. □

The derivation of C_p initially assumes that σ_o^2 is known. It is then established that

$$E\left\{\frac{n\,\widehat{\sigma}^2}{\sigma_o^2} + 2p - n\right\} = E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}.$$

In the preceding, σ_o^2 is subsequently estimated using the mean square error from the largest candidate model, MSE_*. Under the condition $C\left(X_o\right) \subseteq C\left(X_*\right)$, MSE_* is unbiased for σ_o^2, yet

$$\mathrm{C}_p = \frac{n\,\widehat{\sigma}^2}{MSE_*} + 2p - n$$

is not unbiased for $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. From (3.3), the statistic $p + \widehat{\lambda}$ is exactly unbiased for $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. The statistic C_p is not defined as $p + \widehat{\lambda}$. Rather, using (4.2), one can show that

$$\mathrm{C}_p = \left(p + \widehat{\lambda}\right) + 2\left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_*^2} - 1\right),$$

implying that the bias of C_p is in the positive direction. Fujikoshi and Satoh (1997) discuss correcting for the bias of C_p induced when σ_o^2 is estimated by MSE_*. Their correction amounts to using $p + \widehat{\lambda}$ to estimate $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. Following Fujikoshi and Satoh, we define the modified conceptual predictive statistic, MC_p, as

$$\mathrm{MC}_p = p + \widehat{\lambda}.$$

Under suitable conditions, Davies, Neath, and Cavanaugh (2006) establish that MC_p serves as the minimum variance unbiased estimator of $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. Our proposal in the current paper is to estimate the expected symmetric discrepancy $E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. To this end, we define the *symmetrized conceptual predictive statistic*, SC_p, from (4.1), (3.3), and (3.4) by replacing λ with $\widehat{\lambda}$ from (4.2). The resulting criterion,

$$\mathrm{SC}_p = (p + \widehat{\lambda}) + \frac{n(p + \widehat{\lambda})}{n - p + \widehat{\lambda}}\left(1 + \frac{2(n - p) + 4\widehat{\lambda}}{(n - p + \widehat{\lambda})^2}\right),$$

provides an approximately unbiased estimator of $E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$. Thus, in large-sample settings, $E\left\{\mathrm{SC}_p\right\} \approx E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$.

5. A Simulation Study

We look to compare the performance of the conceptual predictive statistics. In each of four simulation sets, one thousand samples of size n are generated from a true regression model having an $n \times p_o$ design matrix, a parameter vector of the form $\beta_o = (1, 1, \ldots, 1)'$, and an error variance of $\sigma_o^2 = 4$. For every sample, candidate models with design matrices of ranks $p = 2, 3, \ldots, p_*$ are fit to the data. The first column of each design matrix is a vector of ones. All other covariates are generated as independent replicates from a $N(0, 8)$ distribution. In each of the simulation sets, the condition $C\left(X_o\right) \subseteq C\left(X_*\right)$ is met. Our examination focuses on the effectiveness of C_p, MC_p, and SC_p at selecting p_o, the true order of the generating model.

In the first two sets, the candidate models are nested. Simulation set I features a sample size of $n = 16$, a largest candidate model of order $p_* = 11$, and a true order of $p_o = 5$. Simulation set II has $n = 20$, $p_* = 16$, and $p_o = 7$. The results are displayed in Tables 2 and 3. The MC_p criterion greatly outperforms C_p in the number of correct selections. The modification of C_p to an exactly unbiased estimator of $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ has led to improved model selection. It is also seen that SC_p provides further improvement over MC_p by increasing the number of correct selections. This improvement can be attributed to a decrease in the propensity to choose an overspecified model.

In simulation sets I and II, the expectations of the target discrepancies $E\left\{\Delta_1\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ and $E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ are computed for each of the candidate model orders. These expectations are displayed graphically in Figures 1 and 2. To maintain comparable scaling for the two discrepancy curves, values of $E\left\{\Delta_S\left(\widehat{\mathcal{M}}_p, \mathcal{M}_o\right)\right\}$ are divided by two.

In these figures, we see the reason for the strong performance of SC_p in the simulations. The target of SC_p provides a greater degree of delineation among the overspecified models than the target of MC_p and C_p. For the Δ_S curve, note that the rate of change increases in accordance with the extent of the overparameterization; for the Δ_1 curve, the rate of change is constant. Thus, SC_p has fewer errors in selecting overspecified models.

Table 2. Order selections for simulation set I.

p	2	3	4	**5**	6	7	8	9	10	11
C_p	1	0	7	**586**	104	54	45	57	53	93
MC_p	1	6	16	**749**	74	32	30	28	24	40
SC_p	1	9	20	**887**	53	15	8	3	3	1

Table 3. Order selections for simulation set II.

p	2–5	6	**7**	8	9	10	11	12	13	14	15	16
C_p	0	2	**541**	79	49	39	27	31	36	33	51	112
MC_p	9	15	**772**	58	32	13	7	13	12	11	16	42
SC_p	8	18	**893**	37	21	9	4	4	2	3	1	0

In the next two sets, candidate models are formulated using all possible subsets of the $(p_* - 1)$ regressor variables. In simulation set III, $n = 16$, $p_* = 6$, and $p_o = 4$; in simulation set IV, $n = 16$, $p_* = 6$, and $p_o = 3$. Tables 4 and 5 summarize the number of selections of underspecified models, correctly specified models, and overspecified models, for each of C_p, MC_p, and SC_p. The all possible subsets case is a harder model selection problem than the nested candidate models case because of the increase in the size of the candidate collection in general, and the increase of the number of overspecified models in the candidate collection in particular. But the behaviors of the selection criteria here are similar to the sets with nested candidate models, with SC_p having more correct selections than MC_p and C_p by providing better protection against overfitting. Specifically, note that in Table 4, C_p is over 20 times more likely to select an overspecified model as opposed to an underspecified model, while MC_p is 10 times, and SC_p is only 5 times. Similarly, in Table 5, C_p is over 30 times more likely to select an overspecified model as opposed to an underspecified model, while MC_p is 22 times, and SC_p is only 16 times. In both scenarios, C_p and MC_p exhibit a much more extreme propensity than SC_p to select overspecified models.

Table 4. Model selections for simulation set III.

	Underspecified		Correctly Specified	Overspecified
	$C(X) \subset C(X_o)$	$C(X_o) \not\subseteq C(X)$ and $C(X) \not\subseteq C(X_o)$	$C(X_o) = C(X)$	$C(X_o) \subset C(X)$
C_p	11	4	**663**	322
MC_p	20	5	**716**	259
SC_p	24	9	**805**	162

Table 5. Model selections for simulation set IV.

	Underspecified		Correctly Specified	Overspecified
	$C(X) \subset C(X_o)$	$C(X_o) \not\subseteq C(X)$ and $C(X) \not\subseteq C(X_o)$	$C(X_o) = C(X)$	$C(X_o) \subset C(X)$
C_p	5	8	**579**	408
MC_p	8	7	**648**	337
SC_p	8	9	**708**	275

To illustrate how other popular model selection criteria perform in these simulation sets, we summarize the selection results in sets I through IV for the Akaike (1973) information criterion, AIC, the corrected Akaike information criterion (Hurvich and Tsai, 1989), AICc, the final prediction error (Akaike, 1969), FPE, and the Bayesian or Schwarz (1978) information criterion, BIC. Table 6 presents the number of correct and overspecified model selections in each set. Due to the small sample sizes considered in the simulation sets, AIC, FPE, and BIC all exhibit strong tendencies to choose overfitted models. AICc performs well, and obtains more correct selections than the conceptual predictive criteria. These results are consistent with those reported in similar simulation studies: see, for instance, Hurvich and Tsai, 1989, pp. 300-301.

However, despite the strong performance of AICc, conceptual predictive statistics are often preferred to Akaike information criteria since values of the latter cannot be compared to benchmarks for detecting model misspecification. As evident from (3.3), (3.4), and (4.1), p serves as a benchmark for C_p and MC_p, whereas $2p$ serves as a benchmark for SC_p. C_p and MC_p should be close to p when $\lambda \approx 0$ (reflecting low approximation error), whereas SC_p should be close to $2p$ when $\lambda \approx 0$ and n is much larger than p (reflecting low approximation error and low estimation error). Such references do not exist for the criteria considered in Table 6; thus, criterion values are only meaningful when the differences between them are considered.

Table 6. Correct model selections for AIC, AICc, FPE, and BIC
in simulation sets I through IV.
(Number of overspecified selections in parentheses.)

Criterion	Set			
	I	II	III	IV
AIC	300 (697)	146 (853)	540 (451)	432 (559)
AICc	918 (51)	933 (53)	853 (103)	776 (205)
FPE	420 (577)	317 (682)	571 (420)	461 (530)
BIC	467 (529)	366 (633)	660 (323)	592 (395)

6. An Application

We consider a regression modeling application based on data from a cardiac rehabilitation program at University of Iowa Hospitals and Clinics. The data consist of measurements based on 35 patients who have had a myocardial infarction and have completed the program. The data may be obtained by either downloading a text file from the first author's website (http://myweb.uiowa.edu/cavaaugh/) or by e-mailing the first author (joe-cavanaugh@uiowa.edu).

The response variable is the final score on a test that reflects the capability of the patient to physically exert himself / herself. The score is in units of metabolic equivalents (METs). One MET corresponds to the rate of oxygen consumption for an average person at rest. The covariates include the patient's initial score on this test. Additional covariates include the patient's age, the patient's gender, and the patient's baseline body mass index (BMI) dichotomized based on whether BMI is less than 30. (A BMI of 30 is the standard cutoff for obesity.) Interactions under consideration include (a) initial score and gender, (b) initial score and BMI, (c) age and gender, (d) age and BMI. In defining the candidate collection, we consider models corresponding to all possible regressor subsets that satisfy the following criteria: (1) the initial test score is included, (2) if an interaction is included, both of the covariates represented in the interaction are also included.

The conceptual predictive statistics C_p, MC_p, and SC_p are computed for each of the models in the candidate collection. C_p and MC_p choose a model that includes initial score, age, gender, BMI, and the age/gender interaction (model $\mathcal{M}_6$, $p = 6$). SC_p chooses a model that includes initial score, age, gender, the age/gender interaction, but not BMI (model $\mathcal{M}_5$, $p = 5$). Model $\mathcal{M}_6$ is ranked second by SC_p; model $\mathcal{M}_5$ is ranked third by C_p and second by MC_p.

The criteria differ on the inclusion of BMI in the final candidate model. From a clinical standpoint, BMI might initially seem like a relevant predictor. However, much of the information conveyed by this dichotomous variable is represented by the other explanatory variables, and the marginal relationship between this variable and the response is relatively weak. All 12 obese patients, with BMI ≥ 30, were male. Furthermore, the obese patients were considerably younger: the mean age for the obese group was 53.3 years versus 66.9 years for the non-obese group. The obese patients had slightly higher initial test scores, most likely because they were younger: the mean initial score was 6.2 METs for the obese group versus 5.2 METs for the non-obese group. This difference still persists in the final test scores: the mean final score was 9.1 METs for the obese group versus 8.3 METs for the non-obese group. In the multivariable candidate models, it is plausible that the difference in final test scores between the weight groups is explained by the variation in age, gender, and initial test scores.

We have seen in our theoretical development and simulations that SC_p provides better protection against overfitting than C_p and MC_p. It is not surprising to see in practice that the SC_p criterion is more stringent about including input variables which have a questionable effect on response.

7. Conclusion

By formulating an estimator of a symmetrized version of the discrepancy targeted by traditional C_p, we obtain an alternate version of C_p that provides better protection from overfitting and tends to favor more parsimonious models. The new criterion, SC_p, is developed in the same spirit as C_p, although the interpretability of the symmetric discrepancy targeted by SC_p is less transparent than that of the asymmetric discrepancy targeted by C_p.

Because of the common basis for SC_p and C_p, the criteria share certain key characteristics and may be used in a similar manner. In particular, as previously mentioned, both criteria may be compared to benchmarks for assessing model misspecification: C_p to p, and SC_p to $2p$. Based on these benchmarks, C_p–type plots constructed using either set of criterion values may be utilized in screening candidate models. However, since SC_p is more sensitive towards detecting overspecification than its traditional counterpart, plots based on SC_p may have greater utility in identifying models that include extraneous or irrelevant regressors. In interpreting such plots, fitted models for which $\mathrm{SC}_p \approx 2p$ should be viewed as viable candidates. Values of SC_p exceeding $2p$ represent models that are potentially undesirable, due to either (1) the exclusion of important variables, or (2) the inclusion of unnecessary variables in settings where the sample size is insufficient to permit accurate estimation of the corresponding regression parameters.

Acknowledgements

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

Figure 1. Expected discrepancy comparison for simulation set I.

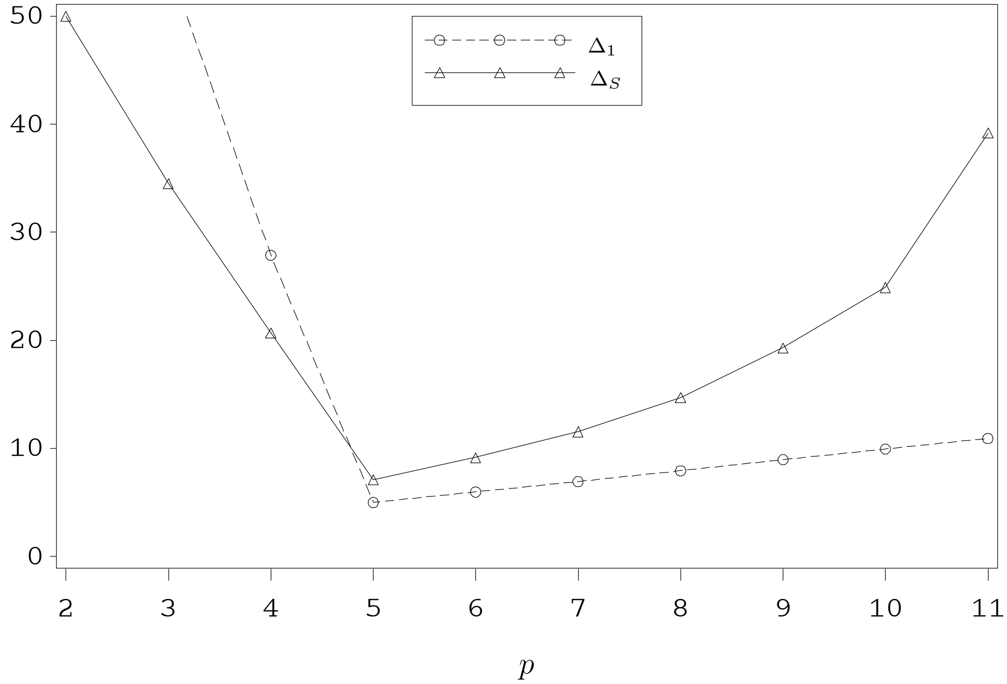
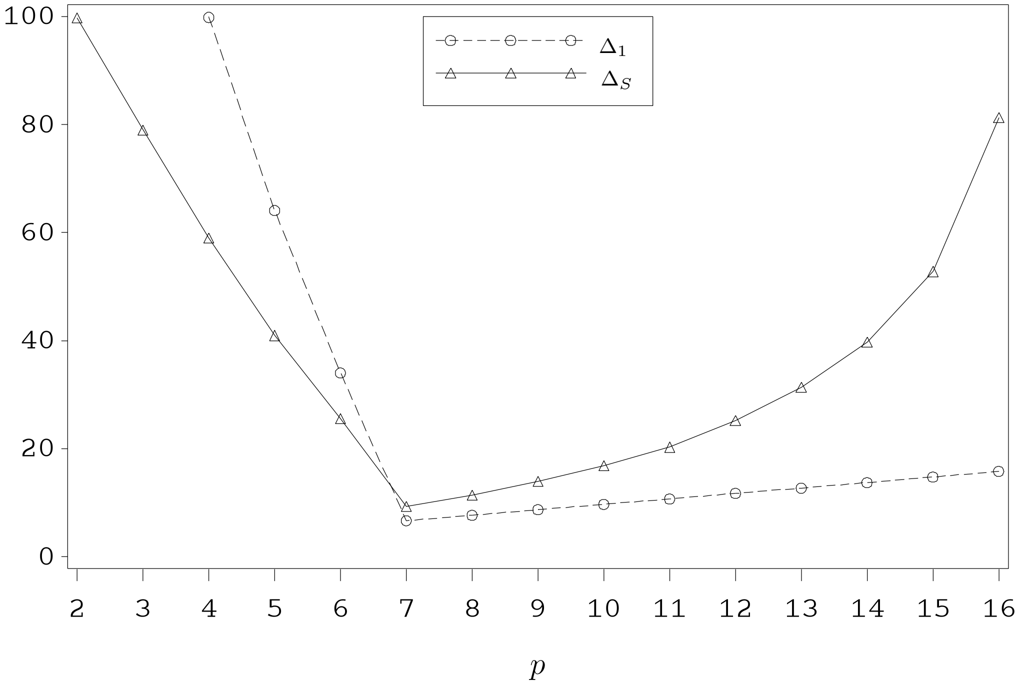


Figure 2. Expected discrepancy comparison for simulation set II.



References

Akaike, H. (1969). Fitting autoregressive models for prediction. *The Annals of the Institute of Statistical Mathematics* **21**, 243–247.

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

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

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters* **44**, 333–344.

Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. *Australian and New Zealand Journal of Statistics*, **46**, 257–274.

Davies, S. L., Neath, A. A. and Cavanaugh, J. E. (2006). Estimation optimality of corrected AIC and modified C_p in linear regression. *International Statistical Review*, **74**, 161–168.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707–716.

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

Kim, H.–J. and Cavanaugh, J. E. (2005). Model selection criteria based on Kullback information measures for nonlinear regression. *Journal of Statistical Planning and Inference*, **134**, 332–349.

Kullback, S. (1968). *Information Theory and Statistics.* Dover, New York.

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

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

McQuarrie A. D. R. and Tsai, C.–L. (1998). *Regression and Time Series Model Selection.* World Scientific, River Edge, New Jersey.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.