Linear Model Evaluation
Based on Estimation of Model Bias

Andrew A. Neath
Department of Mathematics and Statistics
Southern Illinois University Edwardsville

Joseph E. Cavanaugh
Department of Biostatistics
The University of Iowa

Abstract

Model selection criteria often arise by constructing estimators of measures known as expected overall discrepancies. Such measures provide an evaluation of a candidate model by quantifying the disparity between the true model which generated the observed data and the candidate model. However, attention is seldom paid to the problem of accounting for discrepancy estimator variability, or to the companion problem of establishing discrepancy estimators with certain optimality properties. The expected overall Gauss (error sum of squares) discrepancy for a linear model can be decomposed into a term representing the estimation error, due to unknown model coefficients, and a term representing the approximation error, or bias, due to model misspecification. Since the first error term is seen to depend only on model dimension, a known quantity, the problem of estimating the expected overall Gauss discrepancy reduces to the problem of estimating a bias parameter. In this paper, we derive estimators of model bias with frequentist optimality properties and consider how confidence interval estimation can be used to quantify the uncertainty inherent to the problem of bias parameter estimation. We also show how the problem of estimating model bias can be approached from a Bayesian perspective. To illustrate our methodology, we present a modeling application based on data from a cardiac rehabilitation program at The University of Iowa Hospitals and Clinics.

Key Words: Gauss discrepancy; Mallows' Cp; model selection

1. Introduction

An important topic in statistical modeling is the problem of determining which input variables are needed for estimating a response function. Such a decision is facilitated by the use of a model selection criterion. A model selection criterion is often formulated by constructing an estimator of a measure known as an expected overall discrepancy. Such a measure provides an evaluation of a candidate model by quantifying the disparity between the true model (i.e., the model which generated the observed data) and the candidate model.

In selecting a discrepancy, one must consider which aspect of the candidate model is required to conform with the true model. A reasonable goal is to evaluate a model with the idea that the fitted values should be near to the response means on the input space. The Gauss discrepancy, defined through the error sum of squares, is an appropriate discrepancy for model selection in this context. In linear model problems, it can be shown that the expected overall Gauss discrepancy is the sum of a term which represents the "estimation error," due to unknown regression coefficients, and a term which represents the "approximation error," due to candidate model misspecification. Furthermore, the estimation error depends only on the dimension of the approximating model, a known quantity. We will define the approximation error through a parameter we call the model bias. Thus, the problem of estimating the expected overall Gauss discrepancy reduces to the problem of estimating the bias parameter.

Much of the model evaluation literature is dedicated to the construction of estimators of expected overall discrepancies. (See, for instance, Linhart and Zucchini, 1986, and McQuarrie and Tsai, 1998.) However, attention is seldom paid to the problem of how to account for the variability inherent to an estimator of an expected overall discrepancy. A first step in addressing this problem is to derive discrepancy function estimators with optimality properties. In this paper, we develop estimators of the expected overall Gauss discrepancy by developing estimators of the linear model bias parameter. Furthermore, we investigate the optimality properties of these estimators. We consider both frequentist and Bayesian approaches to quantifying the variability of estimators for model bias.

The next section provides an outline of model evaluation based on the Gauss discrepancy in the linear model setting. In addition, we provide some comments comparing the philosophies of model evaluation within the Gauss discrepancy framework to hypothesis testing using the general linear test statistic. In Section 3, we discuss estimation of model bias. The well known conceptual predictive statistic C_p (Mallows, 1973) is an estimator for the expected overall Gauss discrepancy. The C_p estimator for model bias is derived, as well as estimators for model bias having the frequentist optimality properties of minimum variance unbiasedness (MVUE) and maximum likelihood (MLE). In Section 4, a confidence interval with frequentist properties is developed for the model bias parameter. In Section 5, we take a Bayesian approach to the problem of quantifying the variability inherent to estimation of model bias. Sections 6 and 7 close the paper with an application and some concluding remarks.

2. Gauss Discrepancy for Linear Regression

Consider a collection of data y, generated according to the linear model

$$y = X_o\beta_o + \varepsilon_o, \quad \varepsilon_o \sim N_n\left(0, \sigma_o^2 I\right). \tag{2.1}$$

We assume that the response vector y is $(n \times 1)$, the design matrix X_o is $(n \times p_o)$ of full column rank, and the coefficient vector β_o is $(p_o \times 1)$. Let $C\left(X_o\right)$ denote the column space of X_o, with projection matrix given by $H_o = X_o\left(X_o'X_o\right)^{-1}X_o'$. Let $\theta_o = \left(\beta_o', \sigma_o^2\right)'$ represent the $(p_o + 1)-$dimensional parameter vector. We refer to (2.1) as the *true model.*

The goal of model selection is to evaluate the models from a candidate class to determine which provides the "best" approximation to (2.1). Consider a candidate model of the form

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N_n\left(0, \sigma^2 I\right). \tag{2.2}$$

Here, we assume X is $(n \times p)$ of full column rank, and β is $(p \times 1)$. Denote the corresponding column space as $C\left(X\right)$, with projection matrix $H = X\left(X'X\right)^{-1}X'$. Let $\widehat{\beta} = \left(X'X\right)^{-1}X'y$ denote the least squares estimator for β computed under a candidate model, so that the fitted response vector $X\widehat{\beta} = Hy$ serves as an estimator for the true mean response vector $X_o\beta_o$.

The collection of all candidate models of interest is known as the *candidate family.* We assume that the full model (i.e., the largest candidate model) is of the same form as the true model. Under this assumption, for any specific candidate model (2.2), $C\left(X\right) \subseteq C\left(X_o\right)$.

To evaluate a model in the candidate family, we require a measure which provides a suitable reflection of the disparity between the true model and the candidate model. The *overall Gauss discrepancy* reflects the squared distance between the fitted response vector under the candidate model and the true mean response vector. In the linear model context, the measure is given by

$$\left|X\widehat{\beta} - X_o\beta_o\right|^2 = \left(X\widehat{\beta} - X_o\beta_o\right)'\left(X\widehat{\beta} - X_o\beta_o\right).$$

(See Linhart and Zucchini, 1986, pp. 11, 18–19, 118.) For our purposes, we scale the preceding measure by the true error variance σ_o^2, thereby obtaining

$$d_G\left(\widehat{\theta}, \theta_o\right) = \frac{1}{\sigma_o^2}\left|X\widehat{\beta} - X_o\beta_o\right|^2.$$

Let E_o denote the expectation under the true model. The expected overall Gauss discrepancy is therefore defined as

$$\begin{aligned}\Delta_G\left(\theta_o, p\right) &= E_o\left\{d_G\left(\widehat{\theta}, \theta_o\right)\right\} \\ &= E_o\left\{\frac{1}{\sigma_o^2}\left|X\widehat{\beta} - X_o\beta_o\right|^2\right\}. \qquad (2.3)\end{aligned}$$

Under the true model, $E_o\left\{X\widehat{\beta}\right\} = HX_o\beta_o$. Then

$$\left|X\widehat{\beta} - X_o\beta_o\right|^2 = \left|Hy - HX_o\beta_o\right|^2 + \left|X_o\beta_o - HX_o\beta_o\right|^2. \qquad (2.4)$$

The first term on the right hand side of (2.4) can be written as a quadratic form

$$\left(y - X_o\beta_o\right)' H \left(y - X_o\beta_o\right),$$

distributed as $\sigma_o^2\,\chi^2$ with degrees of freedom given as $rank(H) = \dim\left(C\left(X\right)\right) = p$. Thus

$$E_o\left\{\left|Hy - HX_o\beta_o\right|^2\right\} = p\,\sigma_o^2. \qquad (2.5)$$

Using (2.4) and (2.5) with (2.3), we obtain

$$\Delta_G\left(\theta_o, p\right) = p + \delta \qquad (2.6)$$

where

$$\delta = \frac{\left|X_o\beta_o - HX_o\beta_o\right|^2}{\sigma_o^2}$$

is defined as the model bias.

The bias parameter δ represents the approximation error due to model misspecification from the exclusion of input variables when defining the candidate model. Note that δ is the squared distance between the true mean response vector $X_o\beta_o$ and the mean space $C\left(X\right)$, scaled by the true error variance. As the number of parameters p in a candidate model increases, the dimension of $C\left(X\right)$ increases. So as parameters are added to a candidate model, the approximation error as measured by the model bias δ will decrease.

Expression (2.5) represents the error due to estimation of the unknown parameters in the candidate model. Estimation error appears in (2.6) only through the number of parameters p. Inversely to approximation error, estimation error increases as parameters are added to a candidate model. Model selection based on the Gauss discrepancy seeks a model which balances estimation error and approximation error. Since p is known, the evaluation of a candidate model centers on the model bias δ.

In the next section, we consider approaches to estimating model bias. But first we will compare the Gauss discrepancy approach to model evaluation with the general linear hypothesis test. Suppose that one is testing a full model described as (2.1) versus a candidate model described as (2.2), with $C(X) \subset C(X_o)$. As in our framework, one is assuming here that the full model is true. A candidate model is created by placing a set of linear restrictions on the coefficient vector. The null hypothesis of the test is that the reduced model is correct, or equivalently, that the linear restrictions are precisely true.

From the theory of linear statistical models, the statistic

$$F^* = \frac{y'(H_o - H)\, y \,/\, (p_o - p)}{y'(I - H_o)\, y \,/\, (n - p_o)} \tag{2.7}$$

follows an F distribution with $p_o - p$ and $n - p_o$ degrees of freedom, and noncentrality parameter

$$ncp = \frac{(X_o\beta_o)'(H_o - H)(X_o\beta_o)}{\sigma_o^2}.$$

Since $H_oX_o\beta_o = X_o\beta_o$, then the noncentrality parameter is simply the model bias δ. When the null hypothesis is true, then $HX_o\beta_o = X_o\beta_o$ as well, and $\delta = 0$. Thus, the general linear hypothesis test is for whether the reduced model bias parameter is zero or not.

A Bayesian approach to the problem proceeds from the same viewpoint. (See George, 2000, and Clyde and George, 2004, for a review of the Bayesian variable selection literature.) The posterior probability on the candidate model represents the probability that its model bias is zero.

The evaluation of a candidate model based on the Gauss discrepancy takes on a different philosophy. From the assumption that the full model has the same form as the true model, the full model has no bias. Thus, we will prefer the candidate model to the full model if $\delta + p < 0 + p_o$, or $\delta < (p_o - p)$. Hence, one is interested in determining whether or not the model bias is smaller than the increase in estimation error for the added parameters in the full model. An explanation for the different philosophy lies in the view that under the Gauss discrepancy, one is looking to evaluate a model based on mean squared error (i.e., mean squared distance between the fitted response vector and the true mean response vector). Consider a case where the linear restrictions on the parameter vector defining the candidate model are nearly satisfied, but not satisfied exactly. The fitted response vector under the candidate model may be closer (on average) to the true mean response vector by taking the linear restrictions as true than by selecting the full model and adding an extra source of estimation error. It is thus possible for the candidate model to be better than the full model under the Gauss discrepancy even though the full model is correctly specified and the candidate model is not.

3. Frequentist Point Estimation of the Bias Parameter

The numerator and denominator quadratic forms in the general linear test statistic F^* given in (2.7) are distributed as

$$y'\left(H_o - H\right)y \sim \sigma_o^2\,\chi^2,$$

with $p_o - p$ degrees of freedom and noncentrality parameter δ, independent of

$$y'\left(I - H_o\right)y \sim \sigma_o^2\,\chi^2,$$

with $n - p_o$ degrees of freedom. Thus,

$$E_o\left\{\frac{y'\left(H_o - H\right)y}{p_o - p}\right\} = \frac{\sigma_o^2\left(p_o - p + \delta\right)}{p_o - p} \tag{3.1}$$

and

$$E_o\left\{\frac{y'\left(I - H_o\right)y}{n - p_o}\right\} = \sigma_o^2. \tag{3.2}$$

Using (3.1) and (3.2) with (2.7), let n tend to infinity so that

$$E_o\left\{F^*\right\} = E_o\left\{\frac{y'\left(H_o - H\right)y\,/\,\left(p_o - p\right)}{y'\left(I - H_o\right)y\,/\,\left(n - p_o\right)}\right\}$$

approaches

$$\frac{E_o\left\{y'\left(H_o - H\right)y\,/\,\left(p_o - p\right)\right\}}{E_o\left\{y'\left(I - H_o\right)y\,/\,\left(n - p_o\right)\right\}} = 1 + \frac{\delta}{p_o - p}. \tag{3.3}$$

Then for large n,

$$E_o\left\{\left(p_o - p\right)\left(F^* - 1\right)\right\} \approx \delta.$$

Define

$$\widehat{\delta}_c = \left(p_o - p\right)\left(F^* - 1\right). \tag{3.4}$$

The well known C_p statistic (Mallows, 1973) is an estimator of $\Delta_G\left(\theta_o, p\right) = p + \delta$. It is easy to derive that $\mathrm{C}_p = p + \widehat{\delta}_c$. Thus, we name the C_p bias estimator $\widehat{\delta}_c$ after its companion statistic.

The expected value for F^* in (3.3) is only an approximation. To obtain an exact result, note

$$\frac{n - p_o}{y'\left(I - H_o\right)y} \sim \frac{n - p_o}{\sigma_o^2\,\chi^2},$$

so that

$$\begin{aligned} E_o\{F^*\} &= E_o\left\{\frac{y'(H_o-H)y}{p_o-p}\right\} E_o\left\{\frac{n-p_o}{y'(I-H_o)y}\right\} \\ &= \left(\frac{p_o-p+\delta}{p_o-p}\right)\left(\frac{n-p_o}{n-p_o-2}\right). \end{aligned} \tag{3.5}$$

Then from (3.5), one can derive the modified estimator of model bias as

$$\widehat{\delta}_m = (p_o-p)\left[\left(\frac{n-p_o-2}{n-p_o}\right)F^*-1\right] \tag{3.6}$$

by requiring $E_o\left\{\widehat{\delta}_m\right\}=\delta$. The modified conceptual predictive statistic (Fujikoshi and Satoh, 1997) is derived to be an unbiased estimator of $\Delta_G(\theta_o,p)$. Then $\mathrm{MC}_p = p+\widehat{\delta}_m$.

Since the support set for the statistic F^* is the interval of positive real numbers, $\widehat{\delta}_c$ and $\widehat{\delta}_m$ both have nonzero probability of taking on negative values. This is an unfortunate consequence since the parameter space for δ precludes the possibility of a negative value. One may wish to define an adjusted estimator as

$$\widetilde{\delta} = \begin{cases} \widehat{\delta} & \text{, if } \widehat{\delta}>0 \\ 0 & \text{, if } \widehat{\delta}<0 \end{cases}.$$

At this point, we take up the problem of establishing bias parameter estimators having the optimality properties of minimum variance unbiasedness and maximum likelihood. The likelihood function for data y is written based on the full model as

$$\begin{aligned} L(\theta_o|y) &= \left(2\pi\sigma_o^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma_o^2}(y-X_o\beta_o)'(y-X_o\beta_o)\right\} \\ &= c(\theta_o)\exp\left\{-\frac{1}{2\sigma_o^2}(y'y)+\left(\frac{1}{\sigma_o^2}\beta_o'\right)(X_o'y)\right\}, \end{aligned}$$

where

$$c(\theta_o) = \left(2\pi\sigma_o^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma_o^2}\beta_o'(X_o'X_o)\beta_o\right\}.$$

(Recall that the full model is of the same form as the true model.) Following the work of Davies, Neath, and Cavanaugh (2006), we will show that $\widehat{\delta}_m$ is the minimum variance unbiased estimator of δ by making use of the Lehmann-Scheffe$'$ Theorem. By Theorem 2.5.3 of Christensen (2002, pp. 31–32), $T(y)=(y'y,\ X_o'y)$ is a complete sufficient statistic. We have seen that $E_o\left\{\widehat{\delta}_m\right\}=\delta$. It remains to be seen that $\widehat{\delta}_m$ is a function of the complete sufficient statistic $T(y)$.

The unbiased estimator $\widehat{\delta}_m$ is a function of the data y only through the statistic F^*, which in turn is a function of the data through the quadratic forms $y'(H_o - H)\,y$ and $y'(I - H_o)\,y$. Write

$$y'(I - H_o)\,y = y'y - (H_o y)'(H_o y)\,.$$

Noting $C(X) \subseteq C(X_o)$ so that $H\,H_o = H$, we can write

$$\begin{aligned} y'(H_o - H)\,y &= y'H_o y - y'Hy \\ &= (H_o y)'(H_o y) - (Hy)'(Hy) \\ &= (H_o y)'(H_o y) - (HH_o y)'(HH_o y)\,. \end{aligned}$$

Since $y'y$ and $H_o y = X_o\,(X_o'X_o)^{-1}\,(X_o'y)$ are functions of $T(y)$, then $\widehat{\delta}_m$ is a function of $T(y)$. Hence, we can conclude $\widehat{\delta}_m$ is the minimum variance unbiased estimator for δ.

The maximum likelihood estimator for δ is derived from the maximum likelihood estimators $\widehat{\beta}_o = (X_o'X_o)^{-1}\,X_o'y$ and $\widehat{\sigma}_o^2 = \left|y - X_o\widehat{\beta}_o\right|^2 / \,n$. By applying the invariance principle, we obtain

$$\widehat{\delta}_{MLE} = \frac{\left|X_o\widehat{\beta}_o - HX_o\widehat{\beta}_o\right|^2}{\widehat{\sigma}_o^2}.$$

Since $X_o\widehat{\beta}_o = H_o y$ and $HX_o\widehat{\beta}_o = HH_o y = Hy$, the numerator in $\widehat{\delta}_{MLE}$ is the quadratic form $y'(H_o - H)\,y$. The denominator in $\widehat{\delta}_{MLE}$ depends on the quadratic form $y'(I - H_o)\,y$. To see how the maximum likelihood estimator for δ can also be written as a function of F^*, we have

$$\begin{aligned} \widehat{\delta}_{MLE} &= \frac{y'(H_o - H)\,y}{y'(I - H_o)\,y\,/\,n} \\ &= \frac{n}{n - p_o}\,(p_o - p)\,F^*. \end{aligned} \qquad (3.7)$$

The optimal point estimators for δ derived in this section can be written as functions of F^*. Therefore, we can account for estimator variability by developing a common distribution theory using the distribution for F^* in (2.7). We do so in the next section in the development of an interval estimator for model bias δ.

4. Frequentist Interval Estimation of the Bias Parameter

The accuracy of an estimator of model bias may be quantified by an interval estimator. A common approach to developing an interval estimator is to include all parameter values which are accepted by a corresponding two-sided hypothesis test. In particular, a $(1 - 2\alpha)\,100\%$

confidence interval for δ includes all values δ_0 such that a level 2α test for $H_0 : \delta = \delta_0$ accepts the null hypothesis. Under $H_0 : \delta = \delta_0$, the general linear test statistic is distributed as

$$F^* \sim F\left(p_o - p, n - p_o, \delta_0\right).$$

The upper-tailed p-value is given as a function of δ_0 by

$$pvalue1\left(\delta_0\right) = P\left[F\left(p_o - p, n - p_o, \delta_0\right) > F^*\right].$$

The lower-tailed p-value is given as a function of δ_0 by

$$pvalue2\left(\delta_0\right) = P\left[F\left(p_o - p, n - p_o, \delta_0\right) < F^*\right].$$

Define the function v as

$$v\left(\delta_0\right) = pvalue1\left(\delta_0\right)$$

so that

$$1 - v\left(\delta_0\right) = pvalue2\left(\delta_0\right).$$

A level 2α test of $H_0 : \delta = \delta_0$ against $H_A : \delta \neq \delta_0$ accepts the null hypothesis if and only if

$$\min\left\{pvalue1\left(\delta_0\right), pvalue2\left(\delta_0\right)\right\} \geq \alpha. \tag{4.1}$$

Now, $pvalue1\left(\delta_0\right) \geq \alpha$ is equivalent to $v\left(\delta_0\right) \geq \alpha$ and $pvalue2\left(\delta_0\right) \geq \alpha$ is equivalent to $v\left(\delta_0\right) \leq 1 - \alpha$. So condition (4.1) is equivalent to the condition

$$\alpha \leq v\left(\delta_0\right) \leq 1 - \alpha. \tag{4.2}$$

Since $v\left(\delta_0\right)$ is an upper-tail probability for an F distribution with noncentrality parameter δ_0, $v\left(\delta_0\right)$ is an increasing function of δ_0. This result, combined with (4.2), gives a clear description of those values of model bias for which $H_0 : \delta = \delta_0$ is accepted. Therefore, we have that a $\left(1 - 2\alpha\right) 100\%$ confidence interval for δ is given by $\left[\delta_L, \delta_U\right]$ where

$$v\left(\delta_L\right) = \alpha \quad \text{and} \quad v\left(\delta_U\right) = 1 - \alpha. \tag{4.3}$$

Solutions to the equations in (4.3) are not available in closed form, but solutions are readily available using a numerical computation of the cumulative distribution function for a noncentral F statistic.

We discussed in Section 2 how a test of a candidate model in the general linear hypothesis is equivalent to a test of $H_0 : \delta = 0$ against $H_A : \delta > 0$. The null hypothesis is accepted here at level α when the upper-tailed p-value for $H_0 : \delta = 0$ exceeds α. Then $v\left(0\right) > \alpha$ and there is no solution to the first equation in (4.3). In this case, we will take $\delta_L = 0$ as the lower bound for plausible values of the model bias. If $v(0) > 1-\alpha$, then there is no solution to either equation in (4.3). The hypothesis $H_0 : \delta = \delta_0$ is rejected in favor of $\delta < \delta_0$ for all $\delta_0 > 0$. We define the resulting confidence interval in the limit as consisting of only the single point $\{0\}$. Fortunately, this situation is unlikely to occur. A degenerate confidence interval of a single point $\{0\}$ for δ occurs when the F^* statistic is less than the $100\alpha^{\text{th}}$ percentile for an F distribution with noncentrality parameter equal to zero. So even in the most extreme case, the probability of observing an F^* which is not compatible to any bias parameter value $\delta_0 > 0$ is less than α.

5. Bayesian Interval Estimation of the Bias Parameter

A Bayesian approach to inference on model bias is straightforward to implement. Our goal, in general terms, is to quantify our information on model bias δ. The data, conditional on the parameter vector θ_o, follows model (2.1). We can write this as

$$y \,|\, \beta_o, \sigma_o^2 \sim N_n\left(X_o\beta_o, \sigma_o^2 I\right).$$

In an effort to stay objective, we will take a noninformative prior on the parameters $\theta_o = (\beta_o', \sigma_o^2)'$, although it is not necessary to follow this convention if good prior information is available. The posterior distribution updates easily (see Gelman et al., 2003, for example) to become

$$\beta_o \,|\, \sigma_o^2, y \sim N_{p_o}\left(\widehat{\beta_o}, \sigma_o^2 \left(X_o' X_o\right)^{-1}\right) \tag{5.1}$$

$$\sigma_o^2 \,|\, y \sim \frac{n\,\widehat{\sigma_o^2}}{\chi^2\left(n-p_o\right)}, \tag{5.2}$$

where $\widehat{\beta_o}$ and $\widehat{\sigma_o^2}$ are the maximum likelihood estimators of β_o and σ_o^2, respectively.

For any candidate model, the bias parameter

$$\delta = \frac{\left|X_o\beta_o - HX_o\beta_o\right|^2}{\sigma_o^2}$$

is a function of the parameters β_o and σ_o^2, so the posterior distribution on δ is induced from the distributions in (5.1) and (5.2). Since it is easy to generate outcomes from the multivariate normal and chi-square distributions, the posterior distribution on model bias δ can be described via simulation.

A $(1-2\alpha)\,100\%$ Bayesian confidence interval for model bias can be defined by taking the lower and upper $100\alpha^{\text{th}}$ percentiles from the posterior distribution on δ as the lower and upper limits of the interval. Recall that the candidate model is preferred within the Gauss discrepancy framework when the model bias is such that $\delta < (p_o - p)$. One can use the posterior distribution on δ in deciding between a candidate model and the full model by calculating $P\left[\delta < (p_o - p) \mid y\right]$.

6. An Application

In this section, we illustrate the use of bias estimation for model evaluation. We consider a modeling application based on data from a cardiac rehabilitation program at The University of Iowa Hospitals and Clinics. The data consist of measurements on $n = 35$ patients who have had a myocardial infarction and have completed the rehabilitation program. 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 input variables include the patient's initial score on this test. Additional input variables are the patient's age, the patient's gender, an interaction for age and gender, and the patient's baseline body mass index (BMI), dichotomized based on whether BMI is greater than or less than 30. (A BMI of 30 is the standard cutoff for obesity.) The research question is should BMI be included as a predictor of a patient's rehabilitation. The research question translates into the statistical question of whether the input variable BMI carries enough information on the response variable to warrant nonzero estimation of its regression coefficient.

The dimension of the full model is $p_o = 6$. A candidate model is formed by excluding body mass index as an input variable. Then $p = 5$. According to the argument from Section 2, the candidate model is preferred according to the overall Gauss discrepancy if the bias introduced by the exclusion of input variable BMI is less than $p_0 - p = 1$. That is, the candidate model is preferred over the full model if $\delta < 1$, where δ is the model bias. The general linear statistic for testing the full model with input variable BMI against the candidate model without input variable BMI is computed to be $F^* = 2.319$. We compute the C$_p$ estimator, the modified estimator, and the maximum likelihood estimator of the bias parameter using (3.4), (3.6), and (3.7), respectively, as $\widehat{\delta}_c = 1.319$, $\widehat{\delta}_m = 1.159$, and $\widehat{\delta}_{MLE} = 2.799$. Point estimates of δ suggest that the model bias is nonnegligible. A decision between the full model and the candidate model based on the decision rule $\widehat{\delta} > 1$ leads one to accept the model which includes input variable BMI. However, such a decision rule makes no attempt to account for the variability inherent to an estimate of the bias parameter. We investigate the uncertainty involved with this decision using the techniques developed in Sections 4 and 5.

The function v in expression (4.2) is defined for this problem as

$$v(\delta_0) = P\{F(1, 29, \delta_0) > 2.319\}.$$

Since $v(0) = .139$ exceeds $\alpha = .05$, we set $\delta_L = 0$. The frequentist confidence interval includes zero bias as a plausible setting. Solving $v(\delta_U) = .95$ results in $\delta_U = 10.16$. Thus, the 90% frequentist confidence interval for δ is given by $[0, 10.16]$. The decision between the candidate model ($\delta < 1$) and the full model ($\delta > 1$) within the Gauss discrepancy framework is not decisive at the 90% confidence level as values for δ both smaller than 1 and larger than 1 are plausible.

A Bayesian analysis of the problem leads to a similar conclusion. A 90% Bayesian confidence interval computed using the simulated posterior is given as $[0.04, 9.97]$. The posterior probability that the candidate model is preferred to the full model within the Gauss discrepancy framework is computed using the simulated posterior to be

$$P[\delta < 1 \mid y] = .2994.$$

The width of the confidence intervals is an indication of the variability present in this model evaluation problem and illustrates how model selection based on a point estimator, even one with optimality properties, can potentially be misleading. Mallows (1973) provides a similar sentiment in a warning against the strict use of C_p as a model selection criterion. Quoting from Mallows (1973):

"The device (C_p) cannot be expected to provide a single best equation when the data are intrinsically inadequate to support such a strong inference. The greatest value of the device is that it helps the statistician recognize the ambiguities that confront him (her)."

As we have seen in the application, a confidence interval which quantifies the uncertainty involved in the estimation of model bias allows for a quantification of those ambiguities which confront a statistician in this model evaluation problem. Although we still tend to favor the full model in our application, the decision must be tempered in light of the present uncertainty. Such a conclusion is not stated as a simple decision between the candidate model and the full model, but is more appropriate in light of the information from the data.

7. Concluding Remarks

Alternate techniques have been proposed for quantifying the uncertainty involved with the use of the C_p statistic for the purpose of model selection. Mallows (1973) considered an approach based on simultaneous inference for all regression coefficients. The region for accepting that a candidate model has zero bias, i.e., $E(\mathrm{C}_p) = p$, is based on whether or not there exists a coefficient vector β contained within the Scheffé confidence ellipsoid that satisfies the linear restriction for that candidate model. The decision rule can be expressed in terms of a cut-off point on the magnitude of $\mathrm{C}_p - p$. Gilmour (1996) presents an approach for testing the bias of a candidate model simultaneously against all models with one additional parameter. The test is based on a null distribution on the C_p statistic defined through the maximum of independent F random variables. This null distribution is appropriate when all inputs with nonzero coefficients have been included in the candidate model. In other words, the null distribution is based on the candidate model having zero bias. Although these approaches use the C_p statistic directly, the hypothesis being tested is the same as that of the general linear hypothesis test discussed earlier. Namely, the tests of Mallows and Gilmour are for whether a candidate model bias parameter is zero or not. These tests do not address the issue raised in this paper, where we are interested in determining whether a candidate model is better than a larger model based on its evaluation under the Gauss discrepancy.

Model selection based on the Gauss discrepancy seeks to evaluate models from among a candidate class according to $\Delta_G(\theta_o, p) = p + \delta$. Inference on $\Delta_G(\theta_o, p)$ for a candidate model reduces to inference on its bias parameter δ. In the current paper, we have derived point estimators of δ having the optimality properties of minimum variance unbiasedness and maximum likelihood. We have also developed interval estimators for δ using both frequentist and Bayesian approaches, allowing one to provide a measure of uncertainty in an estimate of model bias. By reducing the model selection problem to the problem of estimating an unknown bias parameter, we have developed approaches which are optimal in some sense, and which allow for quantifying the uncertainty inherent to model selection.

Acknowledgements

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

References

Christensen, R. (2002). *Plane Answers to Complex Questions*, Third Edition. New York: Springer.

Clyde, M. and George, E.I. (2004). Model uncertainty. *Statistical Science,* 19, 81-94.

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

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

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*, Second Edition. London: Chapman and Hall.

George, E.I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95, 1304-1308.

Gilmour, S.G. (1996). The interpretation of Mallows' Cp statistic. *The Statistician,* 45, 49-56.

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

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

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