BAYESIAN ESTIMATION OF LINEAR STATISTICAL MODEL BIAS

Andrew A. Neath[1] and Joseph E. Cavanaugh[2]

[1] Department of Mathematics and Statistics,
Southern Illinois University, Edwardsville, Illinois 62026, USA
e-mail: aneath@siue.edu

[2] Department of Biostatistics,
College of Public Health, The University of Iowa,
200 Hawkins Drive C22-GH, Iowa City, Iowa 52242, USA
e-mail: joe-cavanaugh@uiowa.edu

Abstract: The linear statistical model provides a flexible approach to quantifying the relationship between a set of real-valued input variables and a real-valued response. A scientifically relevant goal is to determine which input variables have only a minor effect, or no effect, on the response. We show how this decision can be framed as an estimation problem by defining a bias parameter for the linear statistical model. A Bayesian approach to estimating the model bias leads us to an easily interpreted quantification of the uncertainty inherent in a statistical decision.

AMS Subject Classification: 62C10, 62F15, 62J05

Key Words: Mallows' Cp, model selection, linear model

1. Introduction to the Linear Statistical Model

A common statistical problem involves the modeling of the relationship between a set of real-valued input variables and a real-valued response variable. For reasons both mathematical and natural, a linear function is often used to describe this relationship. A random error term accounts for any non-deterministic aspects of the association. For k input variables, write

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i,$$

where $x_i' = (1, x_{i1}, \ldots, x_{ik})$ denotes the input vector and y_i denotes the response for the i^{th} observation. We further take the usual assumption that the stochastic component follows the Gaussian, or normal, distribution. That is, $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 denotes the residual variance.

Full data consists of a collection of n independent observations

$$(x_1, y_1), \ldots, (x_n, y_n).$$

In multivariate notation, we write the linear statistical model as

$$y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I), \tag{1}$$

where $y = (y_1, \ldots, y_n)'$ is the $n \times 1$ response vector and $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$ is the $p \times 1$ parameter vector, $p = k + 1$. Let

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

denote the $n \times p$ input matrix. We will assume $\text{rank}(X) = p$. Thus, the k input variables carry no redundant information.

The development of linear statistical model distribution theory proceeds conditional on the choice of an input matrix. So

$$y \sim N_n(X\beta, \sigma^2 I),$$

a multivariate normal distribution with mean vector $X\beta$ and covariance matrix $\sigma^2 I$. Because convenient transformations can often be made on the original variables to achieve linearity and normality, the linear statistical model provides a remarkably flexible approach to quantifying a relationship.

True values for the mean vector $X\beta$ and the variance σ^2 are unknown, and must be estimated from the data. The principle of maximum likelihood yields the estimators

$$\widehat{X\beta} = X(X'X)^{-1}X'y \tag{2}$$

and

$$\widehat{\sigma^2} = \frac{1}{n}\left|y - \widehat{X\beta}\right|^2 \tag{3}$$

(see Christensen [2]). Let $H = X(X'X)^{-1}X'$ denote the projection matrix onto the column space $C(X)$. The estimated mean response vector $\widehat{X\beta} = Hy$ is the vector restricted to $C(X)$ that lies at a minimum distance from the observed response vector y. The estimated variance is based on the squared distance between the observed response and the restricted mean vector space.

Now consider a reduced model represented by

$$y = X_o\beta_o + \epsilon_o, \qquad \epsilon_o \sim N_n(0, \sigma_o^2 I), \tag{4}$$

where X_o is an $n \times p_o$ design matrix such that $C(X_o) \subseteq C(X)$. Let $H_o = X_o(X_o'X_o)^{-1}X_o'$ denote the projection matrix onto the column space $C(X_o)$. Further, let

$$\begin{aligned} \widehat{X_o\beta_o} &= X_o(X_o'X_o)^{-1}X_o'y \\ &= H_o y \end{aligned}$$

and

$$\widehat{\sigma_o^2} = \frac{1}{n}\left|y - \widehat{X_o\beta_o}\right|^2$$

be the maximum likelihood estimators of the parameters for the reduced model.

The reduced model is based on a subset of the available input variables. The motivation is to remove those input variables having only a minor effect, or no effect, on the response. This goal is scientifically relevant in that we are determining those input variables which have a negligible or nonexistent relationship with the response variable. The problem of determining whether or not those input variables removed to form X_o are important can be stated as a decision between the two models under consideration.

The statistical decision between a full model and a reduced model is typically stated as a hypothesis testing problem. In this paper, we show how a decision rule can be framed as an estimation problem. We first show how the overall quality of a model can reasonably be described through a single parameter, called the bias δ. Next, we review some classical approaches to estimating δ. Our current contribution is to introduce the notion of Bayesian analysis within the linear modeling framework for the purpose of estimating the bias parameter. A Bayesian approach allows for an easily interpreted quantification of the uncertainty inherent in a statistical decision.

2. Definition of Bias Parameter

The expected squared error for the estimated mean response vector may be defined as

$$\Delta = E\left\{\left|\widehat{X_o\beta_o} - X\beta\right|^2\right\}.$$

It can be shown (see, for example, Linhart and Zucchini [7]) that

$$\Delta = \sigma^2 (p_o + \delta), \tag{5}$$

where

$$\delta = \frac{|X\beta - H_o X\beta|^2}{\sigma^2}.$$

The bias δ represents the approximation error due to model misspecification from the exclusion of input variables when defining the reduced model. Note that δ is the squared distance between the true response mean vector $X\beta$ and the reduced mean space $C(X_o)$, scaled by the residual variance. As the number of parameters p_o increases with the number of input variables, the dimension of the reduced mean space increases, so bias δ will decrease. If $C(X_o) = C(X)$, then $\delta = 0$, indicating no model specification error.

The term $\sigma^2 p_o$ represents the error due to estimation of the unknown parameters for a specified model. Inversely to approximation error, the estimation error will increase as the dimension of the reduced mean space increases. The goal is to select a model which balances estimation error and approximation error. A full model with many parameters may have negligible bias, but since each parameter must be estimated, estimation error may be large. Therefore, interest centers on the bias parameter δ for a reduced model which includes only a subset of the available input variables.

Let Δ_F and Δ_R represent the expected squared errors for the full model (1) and the reduced model (4), respectively. The bias for the full model is zero because the full model includes all input variables. Then from (5),

$$\Delta_F = \sigma^2 p.$$

The only contribution to the expected squared error is the error from estimating the linear parameters of the full model. On the other hand,

$$\Delta_R = \sigma^2 (p_o + \delta).$$

The expected squared error for the reduced model consists of both estimation error and approximation error. It follows that $\Delta_R < \Delta_F$ if and only if $\delta < p - p_o$. This implies that the reduced model is superior to the full model when the bias is smaller than the difference in dimensions.

The values of p and p_o are known, but δ depends on the unknown parameters $X\beta$ and σ^2. Thus, linear model bias δ is itself an unknown parameter. A rule for deciding between the full model and the reduced model is based on $\widehat{\delta}$, an estimate of δ. We will decide that the reduced model is preferred if $\widehat{\delta} < p - p_o$. Otherwise, we will decide that the full model is preferred.

Mallows [8] was the first to investigate the estimation of δ. Write the estimator of residual variance for the reduced model as

$$\begin{aligned} n\widehat{\sigma_o^2} &= \left| y - \widehat{X_o\beta_o} \right|^2 \\ &= (y - H_o y)' (y - H_o y) \\ &= y'(I - H_o)y, \end{aligned}$$

since H_o is a projection matrix and thus, symmetric and idempotent. One can use the distribution theory for quadratic forms to show

$$E\left\{ n\widehat{\sigma_o^2} \right\} = \sigma^2 \left(n - p_o + \delta \right).$$

We therefore have

$$E\left\{ \frac{n\widehat{\sigma_o^2}}{\sigma^2} - (n - p_o) \right\} = \delta.$$

Mallows' estimator of the bias is then obtained by replacing σ^2 in the preceding with an estimator $\widetilde{\sigma^2}$ that satisfies $E\{\widetilde{\sigma^2}\} = \sigma^2$:

$$\widehat{\delta}_m = \frac{n\widehat{\sigma_o^2}}{\widetilde{\sigma^2}} - (n - p_o). \tag{6}$$

Specifically, $\widetilde{\sigma^2} = [n/(n-p)]\widehat{\sigma^2}$, which reduces the divisor of the maximum likelihood estimator in accordance with the model dimension.

Although $E\{\widetilde{\sigma^2}\} = \sigma^2$, $E\{\widehat{\delta}_m\} \neq \delta$, so the distribution of $\widehat{\delta}_m$ is not centered at the targeted parameter. Fujikoshi and Satoh [4] introduce a corrected estimator

$$\widehat{\delta}_c = \frac{(n - p - 2)\ \widehat{\sigma_o^2}}{\widehat{\sigma^2}} - (n - p_o - 2) \tag{7}$$

so that $E\{\widehat{\delta}_c\} = \delta$. Furthermore, Davies et al [3] prove that $\mathrm{Var}\{\widehat{\delta}_c\}$ is minimized over the class of all estimators of bias δ with the property of $E\{\widehat{\delta}\} = \delta$. Mallows type estimates, such as (6) and (7), are the standard approach to bias estimation.

In the next section, we explore Bayesian methods for estimating the bias δ.

3. Bayesian Approach

We first introduce some basic ideas behind Bayesian estimation. In general, the purpose of statistical inference is to draw conclusions about an unknown parameter θ. Bayes rule states that from a prior distribution, $p(\theta)$, and a distribution for the data, $p(y|\theta)$, we can calculate a posterior distribution as

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{p(y)},$$

where $p(y) = \int p(\theta)\, p(y|\theta)\, d\theta$. The prior distribution models subjective information about θ prior to data observation, while the posterior distribution models all information about θ, both subjective and observed.

Since data y is observable and constant within the posterior distribution, we can write Bayes rule as a proportionality

$$p(\theta|y) \propto p(\theta)\, p(y|\theta),$$

or

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

A criticism of the Bayesian approach is that the use of prior information destroys the scientific integrity of the analysis. One can answer this criticism by using what is called a noninformative prior. Essentially, the proportionality then becomes

$$\text{posterior} \propto \text{likelihood},$$

so Bayesian analysis is comparable to likelihood based inference.

We now focus on Bayesian methods for the general linear model. The data is distributed as

$$y|X\beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I\right).$$

It can be shown (Gelman et al [5]) that under a noninformative prior on the unknown parameters $X\beta$ and σ^2, the posterior distribution can be described as

$$X\beta|\sigma^2, y \sim N_n\left(\widehat{X\beta}, \sigma^2 H\right) \tag{8}$$

and

$$\sigma^2|y \sim \frac{n\widehat{\sigma^2}}{\chi^2_{n-p}} \tag{9}$$

where $\widehat{X\beta}$ and $\widehat{\sigma^2}$ are defined in (2) and (3). The bias

$$\delta = \left|X\beta - H_o X\beta\right|^2 / \sigma^2$$

is a function of $X\beta$ and σ^2, so its posterior distribution can be induced from the distributions in (8) and (9).

Simulating values from the posterior distribution is a simple and logical approach to describing what is known about the bias δ. It is straightforward to generate random variates from the multivariate normal and chi-square distributions (see, for example, Bickel and Doksum [1]). An algorithm for generating $\left\{\delta_{(k)} : k = 1, \ldots, N\right\}$ is given as follows.

Algorithm A

(1) Generate $(X\beta)_{(k)}$ *from* (8).

(2) Generate $(\sigma^2)_{(k)}$ *from* (9).

(3) Set $\delta_{(k)} = \left|(X\beta)_{(k)} - H_o\,(X\beta)_{(k)}\right|^2 / (\sigma^2)_{(k)}$.
Repeat for $k = 1, \ldots, N$.

A clear advantage of the Bayesian method over the Mallows type estimators described in Section 2 is that Bayes provides a well-defined measure of uncertainty. Rather than merely a point estimate, the posterior distribution for δ represents a multitude of possibilities. Such a description is imperative when accounting for the uncertainty in a statistical decision.

4. An Application

In this section, we illustrate the use of bias estimation for model selection. Kutner et al [6] present data on survival in patients undergoing a particular type of liver operation. The pool of input variables include

X_1	a blood clotting score,
X_2	a prognostic index,
X_3	an enzyme function test score,
X_4	a liver function test score.

The response variable Y is survival time adjusted for age and gender. Input variables X_1, X_2, X_3 can be measured without excess discomfort to the patient, whereas input variable X_4 requires a more invasive procedure. Focus is on the need for inclusion of input X_4, the liver function score, for predicting patient survival time after surgery. The full model is given as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i.$$

The reduced model is then

$$y_i = \beta_{o0} + \beta_{o1} x_{i1} + \beta_{o2} x_{i2} + \beta_{o3} x_{i3} + \varepsilon_{oi}.$$

The sample consists of $n = 54$ patients. Each of the input variables shows a pairwise correlation with response. But since the input variables are intercorrelated as well, our decision is whether or not the liver function score carries significant information beyond that of the three other input variables.

The dimensions of the models are $p = 5$ and $p_o = 4$. According to the argument from Section 2, the reduced model is better than the full model if the bias introduced by the exclusion of input X_4 is less than $p - p_o = 1$. That is, the reduced model is better than the full model if $\delta < 1$.

Summary calculations yield

$$\left|y - \widehat{X\beta}\right|^2 = 3.084\ ,\ \left|y - \widehat{X_o\beta_o}\right|^2 = 3.109.$$

From equation (7), we calculate the bias estimate as $\widehat{\delta}_c = -0.619$. This illustrates an unfortunate consequence of Mallows type estimators of bias. Although δ is necessarily nonnegative, there is a possibility that its estimate will

be negative. It is unclear how one should proceed in this situation. A common recommendation is to set $\widehat{\delta} = 0$ and conclude that the reduced model exhibits zero bias.

The Bayesian approach will not suffer from the deficiency of estimates outside of the parameter space. As described in Section 3, we are able to simulate a multitude of possibilities for the true bias δ. Figure 1 contains a histogram for $\left\{\delta_{(k)} : k = 1, \ldots, 5000\right\}$ generated according to Algorithm A. Note that although the greatest likelihood is for δ near zero, there is still a non-negligible chance that the bias is much larger. We can use the simulated values to quantify inferential claims using probability. For example, define a 90% credible interval (L, U) for δ as

$$P\left[L < \delta_{(k)} < U\right] = .90.$$

We calculate $L = 0.0045$ and $U = 4.0113$. Instead of a single estimate of exactly zero bias, a range of plausible outcomes is provided.

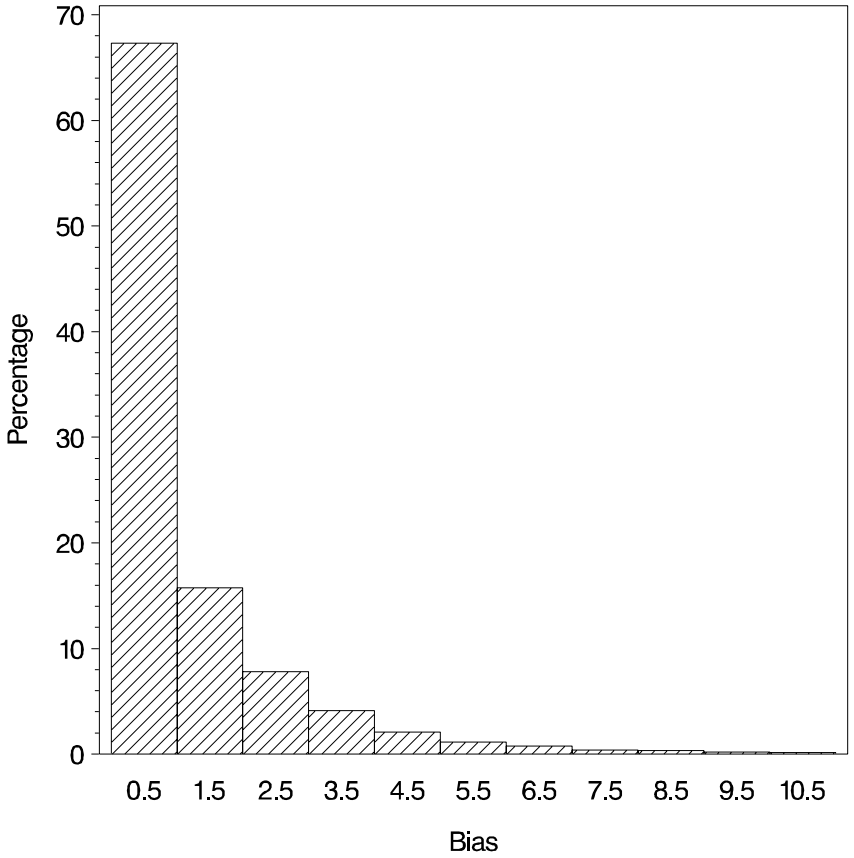

Figure 1. Histogram for $\left\{\delta_{(k)} : k = 1, \ldots, 5000\right\}$.

The key to our decision between the reduced model and the full model is the event

$$[\Delta_R < \Delta_F] \Longleftrightarrow [\delta < 1]\,.$$

From the simulation, we calculate

$$P\left[\delta_{(k)} < 1\right] = .6732.$$

The reduced model is selected over the full model (the liver function score is not needed as an additional input variable), but the selection is more tempered than the Mallows' point estimate $\widehat{\delta} = 0$ would lead us to believe.

Bayesian inference provides a useful method for quantifying the uncertainty about the true value of the unknown bias parameter. Hence, we are better able to quantify our model selection decision.

References

[1] P.J. Bickel, K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics* (2nd Ed.), Prentice-Hall, Upper Saddle River, New Jersey (2000).

[2] R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models* (3rd Ed.), Springer, New York (2002).

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

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

[5] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, *Bayesian Data Analysis* (2nd Ed.), Chapman and Hall / CRC, Boca Raton, Florida (2003).

[6] M.H. Kutner, C.J. Nachtsheim, J. Neter, *Applied Linear Regression Models* (Rev. Ed.), McGraw-Hill, New York (2004).

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

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