Predictive Model Selection

By
Purushottam W. Laud - Medical College of Wisconsin
and
Joseph G. Ibrahim - Harvard University

SUMMARY

We consider the problem of selecting one model from a large class of plausible models. A predictive Bayesian viewpoint is advocated to avoid the specification of prior probabilities for the candidate models and the detailed interpretation of the parameters in each model. Using criteria derived from a certain predictive density and a prior specification that emphasizes the observables, we implement the proposed methodology for three common problems arising in normal linear models : (i) variable subset selection (ii) selection of a transformation of predictor variables, and (iii) estimation of a parametric variance function. Interpretation of the relative magnitudes of the criterion values for various models is facilitated by a calibration of the criteria. Relationships between the proposed criteria and other well known criteria are examined.

Keywords: Bayesian Linear Model; Predictive Distribution; Replicated Experiment; Transformations; Variable Selection

1 Predictive Methodology

Selecting a suitable model from a large class of plausible models is an important problem in statistics. A classic example is the variable selection problem in linear regression analysis. The literature contains many Bayesian and nonBayesian techniques advanced to address this problem. See for example Lindley (1968), Lempers (1971), Mallows (1973), Hocking (1976), and Mitchell and Beauchamp (1988). Choosing suitable transformations of the predictor and/or the response variable in linear regression is another major instance of model selection. Box and Tidwell (1962), Box and Cox (1964), Cook and Weisberg (1982), and Carroll and Ruppert (1988) define and discuss this at length. Selecting appropriate variance functions in the heteroscedastic linear model (see Carroll and Ruppert (1988)) can also be looked upon as a model selection problem. All of these problems arise again in generalized linear models developed in McCullagh and Nelder (1989). See, for example, Christensen (1990), Hosmer and Lemeshow (1989), and Cox and Snell (1989). Model selection in time series analysis (see Box and Jenkins (1981),

West and Harrison (1989)) and nonlinear models (see Bates and Watts (1988)) has also been addressed in the literature.

Among the several criteria that have been proposed for model selection, Akaike's AIC (Akaike, 1973) and Schwarz's BIC (Schwarz, 1978) are widely accepted. An inherent problem with these criteria is that they do not allow prior input for model choice. Moreover, their definitions and/or calibrations rely heavily on asymptotic considerations. On the other hand, the "fully" Bayesian approach to model selection requires the daunting specification of prior probabilities over the large class of models under consideration. Then, one must also specify appropriate priors for the parameters of each model. In the case of selecting between two models, often one can reasonably carry out these specifications and use Bayes factors or posterior model probabilities to make the final model choice. With a large number of models, the fully Bayesian solution is difficult to implement.

In this article we propose three criteria that can be used to address model selection. These emphasize observables rather than parameters and are based on a certain Bayesian predictive density. They have a unifying basis that is simple and interpretable, are free of asymptotic definitions, and allow the incorporation of prior information. Moreover, two of these criteria are readily calibrated.

To fix ideas, consider first the variable selection problem in linear regression. Starting with a full predictor matrix consisting of a column of ones for the intercept term followed by k columns, each representing a predictor, we can write the full model as

$$Y = X\beta + \epsilon \ , \tag{1.1}$$

where Y is an n-vector of responses, β is a $(k+1)$-vector of regression coefficients, and ϵ is an n-vector of random errors. The distribution of ϵ is usually taken to be multivariate normal with mean 0 and *precision* matrix τI, where τ is a positive scalar and I is the $n \times n$ identity matrix. Following the notation of Aitchison and Dunsmore (1975), we write $\epsilon|\tau \sim No_n(0, \tau I)$. In selecting variables, we are interested in considering the 2^k possible models obtained from (1.1) by retaining various subsets of the last k columns of the matrix X, and modifying the length of β accordingly. To be specific, let m be a subset of the integers $\{0, \ldots, k\}$ containing 0, and let k_m denote the number of elements of m. Thus m identifies a model with an intercept and a specific choice of $k_m - 1$ predictor variables. With $\mathcal{M}$ denoting the set of all 2^k models under consideration, we can write these as

$$Y = X_m \ \beta^{(m)} + \epsilon, \qquad m \in \mathcal{M} \ , \tag{1.2}$$

where X_m denotes the $n \times k_m$ full rank predictor matrix under model m, and $\beta^{(m)}$ is the corresponding coefficient vector. Choosing one of the models in (1.2) is the goal of variable selection methods.

For the model selection problem in general, one can replace (1.2) and the accompanying descriptions of distributions for various quantities by considering probability models for the observable Y conditioned on each model m and the attendant parameter vector $\theta^{(m)}$. Thus we write

$$p(y|m,\theta^{(m)}) \ , \qquad\qquad m \in \mathcal{M}, \quad \theta^{(m)} \in \Theta^{(m)} \ ,$$

where $\mathcal{M}$ is the model space and $\Theta^{(m)}$ is the parameter space for model m. To address the task of selecting an m, we adopt a predictive Bayesian viewpoint that allows one to de-emphasize the parameters and focus on the observables. Model selection is an ideal setting for such an approach since the parameter $\theta^{(m)}$ does not carry much physical meaning at the outset. The predictive philosophy is implemented below in two ways. First, whenever possible, the priors for $\theta^{(m)}|m$ are constructed in an automated fashion from a prior prediction for Y and a number quantifying one's belief in this guess relative to the information contained in the experiment. In Section 2, we show how this can be done for the normal linear model. Secondly, we do not use a prior distribution on the model space $\mathcal{M}$. Instead, for each model $m \in \mathcal{M}$, we calculate the criterion and choose the model with a suitably small value. We also provide calibrations of the criteria.

To introduce the criteria, suppose that a prior $\pi(\theta^{(m)}|m)$ has been specified for each $\theta^{(m)}$, $m \in \mathcal{M}$. The posterior for $\theta^{(m)}$ under each model m, given data $Y = y$, is given by

$$\pi(\theta^{(m)}|y,m) = \frac{\pi(\theta^{(m)}|m)\ p(y|m,\theta^{(m)})}{\int \pi(\theta^{(m)}|m)\ p(y|m,\theta^{(m)})\ d\theta^{(m)}} \ .$$

Now envision replicating the entire experiment and denote by Z the vector of responses that might result. In the variable selection problem, for instance, $\theta^{(m)} = (\beta^{(m)}, \tau)$, and each m specifies a predictor matrix X_m. The conceptual replicate experiment has the same design matrix X as the current experiment. Moreover, under any model $m \in \mathcal{M}$, we again have the same future design matrix X_m. The predictive density for Z under model m is

$$p(z|m,y) = \int p(z|m,\theta^{(m)})\ \pi(\theta^{(m)}|y,m)\ d\theta^{(m)} \ .$$

We call this density the *Predictive Density of a Replicate Experiment* and abbreviate it PDRE. In (1.2) for example, the PDRE depends on the design matrix X_m for model m and would be given by

$$p(z|X_m,y) = \int\int p(z|X_m,\beta^{(m)},\tau)\ p(\beta^{(m)},\tau|X_m,y)\ d\beta^{(m)}\ d\tau \ .$$

For notational ease, we denote the PDRE by f_m.

Although the PDRE is central to what follows, we neither expect the current experiment to be repeated nor center our interest on predictions at the current (or any other particular) predictor matrix. The replicate experiment is an imaginary device that puts the predictive

density to inferential use, adapting the philosophy advocated in Geisser (1971). The imagined replication makes y and Z comparable; in fact, exchangeable a priori. Moreover, the parameters in the model play a minimal role under replication. It seems clear that good models, among those under consideration, should make predictions close to what has been observed for an identical experiment. The criteria below are defined with this motivation.

For a given model m, consider

$$L_m^2 = E[(Z-y)'(Z-y)] \ ,$$

where the expectation is taken with respect to the PDRE f_m. The measure L_m^2 has the decomposition

$$L_m^2 = \sum_{i=1}^{n} \left\{ [E(Z_i) - y_i]^2 + Var(Z_i) \right\} \ ,$$

as a sum of two components, one involving the means of the predictive distribution, and the other involving the variances. Thus a model's performance is measured by a combination of how close its predictions are to the observed data and the variability of the predictions. Good models will have small values of L_m^2. It is often more convenient to use the measure

$$L_m = \sqrt{L_m^2}$$

since it is a distance on the response axis, measured in the same units as the response variable. We refer to L_m as the *L criterion*.

To define the second criterion, consider

$$M_m^* = f_m(y) \ .$$

This is the PDRE under model m, evaluated at the observed response y. A good model will have a large value of M_m^*. A ratio of M_m^*'s for two different models is an instance of what Aitkin (1991) calls the *posterior Bayes factor*. Again, to facilitate interpretation, let

$$M_m = (M_m^*)^{-1/n}$$

which is in the units of the response variable, and small values of it indicate good models. We refer to M_m as the *M criterion*.

The third criterion we introduce for model selection is the Kullback-Leibler (KL) divergence between two predictive densities. Suppose f_1 and f_2 are two densities with respect to Lebesgue measure. Then, the KL divergence between f_1 and f_2 is defined by

$$K(f_1, f_2) = \int \ log\left[f_1(x)/f_2(x)\right] \ f_1(x) \ dx \ .$$

In general, $K(f_1, f_2) \neq K(f_2, f_1)$, and $K(a, b) \geq 0$ with equality occurring only if $a = b$. The KL divergence has been used in the literature for a wide variety of statistical problems, and in connection with the Bayesian predictive distribution. For example, with its use Aitchison (1975) shows that the predictive distribution best approximates the sampling distribution, Johnson and Geisser (1983) detect influential observations in linear regression, and McCulloch (1989) assesses the influence of model assumptions. Bhattacharjee and Dunsmore (1991) use the KL directed divergence to select variables in logistic regression.

For our purposes, suppose m_0 is a fixed model in $\mathcal{M}$ from which we measure other models. In variable selection, for instance, a natural choice for m_0 might be the full model (1.1) with all of the k predictors. Using PDRE's of m_0 and m, we define

$$K_m = K(m_0, m) + K(m, m_0) \ .$$

The criteria L_m and M_m measure how "close" the data vector is to the PDRE f_m for each m, whereas K_m is a measure of how far apart the PDRE's of the two models are. If K_m is small, then m and m_0 provide nearly the same information. We see that this criterion is best used when the focus is on a comparison between two specific models. However, one can also consider using it to compare all possible models by fixing m_0 and computing K_m for all possible $m \in \mathcal{M}$. In the variable selection problem, taking the full model as m_0, K_m can be interpreted as the amount of information lost in omitting some predictors from m_0 to get m. Thus, small values of K_m imply that the subset model is nearly as good as the full model, implying that the removed variables may not be important. Another reasonable choice of m_0 is the model with just an intercept. In this case, K_m can be interpreted as the amount of information gained in including the predictors from model m in comparison to the model with just a mean. We note that $K_{m_0} = 0$, whereas the L and M criteria are not minimized at any model fixed in advance. Also, desirable values of K_m may be either large or small depending on the choice of m_0. In this sense, model choice based on K_m is not as straightforward as with L_m and M_m. We refer to K_m as the K *criterion.*

Although most criterion based methods do not quantify the uncertainty inherent in the criterion values, it is desirable to do so. Using the model m^* with the smallest criterion value, one can calculate the standard deviation of the criterion, viewed as a function of the observable Y, with respect to the marginal distribution of Y. In particular, for the L criterion one would compute

$$S_L \ = \ [Var(L_{m^*}(Y))]^{1/2}.$$

We refer to S_L as the *calibration number* for the L criterion. In most instances, its calculation can be effected by obtaining Monte Carlo samples of Y using one of the many techniques now available. In the context of the variable selection problem (1.1) and (1.2), the marginal distribution of Y is multivariate t as outlined in Section 2. We illustrate the use of S_L and the analogous calibration number S_M via examples in Section 3.

2 Prior Distributions and Expressions for the Linear Model

Before proposing an informative prior for the linear model (1.2), we note that Jeffreys's modified prior in this case is given by

$$\pi(\beta^{(m)}, \tau)\ d\beta^{(m)}\ d\tau \propto \tau^{-1}\ d\beta^{(m)}\ d\tau\ . \tag{2.1}$$

The resulting predictive distribution is

$$Z \sim S_n(n - k_m, P_m y, s_m^2(I + P_m))\ , \tag{2.2}$$

where

$$s_m^2 = (n - k_m)^{-1} y'(I - P_m)y\ ,$$

and $S_n(\nu, \mu, \Sigma)$ denotes the n dimensional multivariate t distribution with ν degrees of freedom, location parameter μ, and dispersion matrix Σ (see Box and Tiao, 1973). Here, $P_m = X_m(X_m'X_m)^{-1}X_m'$ is the orthogonal projection operator onto the column space of X_m.

Selecting meaningful informative priors for $\beta^{(m)}$ even for a fixed model m is not an easy task. For a collection of models, it is generally not feasible to interpret each component of $\beta^{(m)}$ for each possible model. One can only hope to use reasonable priors that lead to useful results rather than hope to quantify precisely any real subjective information. Viewing the model mainly as a predictive device, we focus on the response variable when specifying a prior distribution. Incorporating prior knowledge into a guess at the value of the $n \times 1$ response vector Y to be observed at the design matrix X, we denote this guess by η_0. Under model m with design matrix X_m, the prior mean of $\beta^{(m)}|\tau$ is now recommended to be

$$\mu^{(m)} = (X_m'X_m)^{-1}X_m'\ \eta_0\ . \tag{2.3}$$

Clearly $\mu^{(m)}$ is the least squares solution to normal equations written with the design matrix of the model under consideration and the prior guess for Y. If X_m is less than full rank, then $\mu^{(m)}$ is the orthogonal projection of such a least squares solution onto the column space of $X_m'X_m$.

The vector η_0 is a fixed vector regardless of the model under consideration. Its specification may be made in one of several ways. For example, if in previous analyses, the researcher has used a particular submodel m of (1.1) with predictor matrix X_m and estimated the coefficient vector as $\tilde{\beta}^{(m)}$, he/she may choose η_0 to be $X_m\tilde{\beta}^{(m)}$. If the researcher has used previous data to fit a nonlinear model of the form $E(Y) = f(X_m, \beta^{(m)})$, where f and β are estimated, then η_0 may be taken to be $\tilde{f}(X_m, \tilde{\beta}^{(m)})$. Certainly there are many other ways of specifying η_0 that distill all prior information into a guess at Y.

Next, we choose the prior precision matrix of $\beta^{(m)}|\tau$ to be of the form $\tau\ T_m$, where

$$T_m = c\ (X_m'X_m), \tag{2.4}$$

with $c \geq 0$ quantifying, in multiples of the present experiment, the importance one wishes to attach to the prior guess η_0. Thus under model m, T_m is a scalar multiple of the Fisher information matrix for $\beta^{(m)}$. Zellner's g-priors (Zellner, 1986) also have this structure for the precision matrix. It has the advantage of leading to analytically tractable and computationally feasible solutions.

Now, we take $\beta^{(m)}|\tau$ to be normally distributed, i.e.,

$$\beta^{(m)}|\tau \sim No_{k_m}(\mu^{(m)}, \tau\, T_m) \ . \tag{2.5}$$

As a result of focusing on the observables, only a few easily interpreted quantities are needed to specify the prior. In particular, the prediction η_0 is turned into a prior for $\beta^{(m)}|\tau$ for each m in an automated fashion.

Finally, the prior distribution for τ is taken to be a gamma distribution with parameters $(\delta_0/2, \gamma_0/2)$, i.e., with density

$$\pi(\tau)\ d\tau \ \propto\ \tau^{\delta_0/2-1}\ exp\{-\gamma_0\tau/2\}\, d\tau \ . \tag{2.6}$$

For a fixed model m, (2.5) and (2.6) result in the conjugate normal-gamma prior.

With this prior and the likelihood implied by (1.2) for each m, a straightforward derivation yields

$$Z \sim S_n\left(n+\delta_0, \eta_m, s_m^2(I+(1-\gamma)P_m)\right) \ , \tag{2.7}$$

where $\gamma = c/(1+c)$, $\eta_m = P_m\,(\gamma\eta_0 + (1-\gamma)y)$, $s_m^2 = (n+\delta_0)^{-1}(q_m + \gamma p_m + \gamma_0)$, $q_m = y'(I-P_m)y$, and $p_m = (y-\eta_0)'P_m(y-\eta_0)$. The PDRE in (2.2) for noninformative priors can be obtained from (2.7) by formally setting $\gamma = 0$, $\delta_0 = -k_m$, and $\gamma_0 = 0$. Moreover if X_m has rank $r_m < k_m$, replace k_m by r_m in (2.7) above. For brevity, any relevant expressions are given only for the case of conjugate priors in the remainder of this article.

The L criterion under model m is now given by

$$L_m = \{(1+\lambda_m)q_m + \gamma(\gamma+\lambda_m)p_m + \lambda_m\gamma_0\}^{1/2} \ , \tag{2.8}$$

where $\lambda_m = \frac{n+(1-\gamma)k_m}{n+\delta_0-2}$. We see that L_m^2 above is a linear function of q_m and p_m. The quantity q_m is the squared length of the projection of the data onto the error space of model m, i.e., the error sum of squares for model m. The quantity p_m represents a penalty for a bad prior guess at Y. It is the squared length of the projection of the "guessing error" onto the model's column space. Under reference priors, (2.8) reduces to $L_m = \left(2(n-1)(n-k_m-2)^{-1}q_m\right)^{1/2}$. In this case, L_m is similar to the root mean square criterion.

To calculate the calibration number S_L, one can sample from the marginal distribution

$$Y \sim S_n\left(\delta_0, \eta_{m^*}, \gamma_0{\delta_0}^{-1}(I+\gamma^{-1}(1-\gamma)P_{m^*})\right) \ ,$$

and calculate L_{m^*} with each sample. (Here m^* is the model that minimizes L_m.) The standard deviations of these values provides a Monte Carlo approximation to S_L. If one is using the reference priors in (2.1), however, it is well known that the marginal distribution of Y is improper. In this case, one could sample from the conditional distribution $Y|\tau \sim No_n(0, \tau(I - P_{m^*}))$ with τ replaced by $\tilde{\tau}$, the mode of the posterior distribution of τ using m^*. The standard deviation of the resulting samples of L_{m^*} can be viewed as an approximation to $D_L = [Var(L_{m^*}|\tau = \tilde{\tau})]^{1/2}$. For large n one can obtain the analytic approximation

$$D_L \approx \frac{\tilde{\tau}^{-1/2}}{2}(1 - \frac{k_{m^*}}{n})^{1/2}[1 - \frac{1}{n}\{1 + \frac{1}{32}(1 - \frac{k_{m^*}}{n})(1 - \frac{2}{n})\}]^{1/2} \tag{2.9}$$

by computing $\mathrm{E}[L^2_{m^*}|\tau]$ and using a Taylor series approximation for $\mathrm{E}[L_{m^*}|\tau]$. We use D_L and the similarly defined D_M in some of the examples in Section 3. Expressions for D_M as well as M_m and K_m are given in the Appendix.

3 Applications

3.1 Variable Selection

For this problem stated in (1.2), several criteria have been proposed in the literature. A widely accepted nonBayesian criterion is Mallows's C_p, which is equivalent to Akaike's AIC for linear regression models. A standard Bayesian criterion is Schwarz's BIC. Criteria based on a predictive Bayesian distribution include those of Geisser and Eddy (1979) and San Martini and Spezzaferri (1984, 1986). In this section, we discuss the issue of possible overfitting, and illustrate our procedures with data.

When selecting models, an important concern is whether the criterion has a tendency to prefer the larger of two nested models. To address this issue, Smith and Spiegelhalter (1980) present a general form for variable selection criteria. For two nested models $m \subset m_0$, this form is given by

$$\Lambda(a) = \lambda - a\ (k_{m_0} - k_m)\ ,$$

where λ denotes the likelihood ratio statistic and a quantifies a penalty for overfitting. They further point out that if $a \geq 1$, then smaller models are favored over more complex models. Clayton, Geisser, and Jennings (1986) also mention that this is a sensible property, especially for prediction problems. Under noninformative priors, it can be shown that

$$2n\ log\left(\frac{L_m}{L_{m_0}}\right) = \lambda - a_L(k_{m_0} - k_m)\ ,$$

where

$$a_L = \frac{n}{k_{m_0} - k_m}\ log\left(\frac{n - k_m - 2}{n - k_{m_0} - 2}\right)\ .$$

Similarly, we have

$$2n\ log\left(\frac{M_m}{M_{m_0}}\right) = \lambda - a_M(k_{m_0} - k_m)\ ,$$

where

$$a_M = \frac{2}{k_{m_0} - k_m}\ log\ \left(\frac{,\ (n - \frac{k_m}{2}),\ (\frac{n-k_{m_0}}{2})}{,\ (n - \frac{k_{m_0}}{2}),\ (\frac{n-k_m}{2})} \right)\ .$$

Thus $a_L > 1$ for all n, and decreases to 1 as $n \to \infty$. On the other hand, $a_M > 1$ for small to moderate n and decreases to $log(2)$. In this context we note that $a = 1$ for the criterion given by Box and Kanemasu (1973) while $a = 3/2$ yields the local Bayes factor discussed in Smith and Spiegelhalter (1980). The AIC criterion and the asymptotic version of the PSR criterion (Geisser and Eddy, 1979) correspond to $a = 2$, while Schwarz's criterion is equivalent to $a = log(n)$. San Martini and Spezzaferri (1984) propose $a = log\left(n\bar{c}^b\right)$, where $\bar{c} = 2n\lambda^{-1}(e^{\lambda/n} - 1) - 1$ and $b = 2/(k_{m_0} - k_m)$.

Overfitting properties of the criteria under informative priors depend on the prior parameters. Writing $r(m, m_0)$ to denote the ratio $L_m^2/L_{m_0}^2$ for two nested models $m \subset m_0$, we obtain

$$r(m, m_0) = \frac{\lambda_m\gamma_0 + \gamma(\gamma + \lambda_m)p_m + (1 + \lambda_m)q_m}{\lambda_{m_0}\gamma_0 + \gamma(\gamma + \lambda_{m_0})p_{m_0} + (1 + \lambda_{m_0})q_{m_0}}. \tag{3.1}$$

First, $r(m, m_0)$ is a decreasing function of γ, $(0 \leq \gamma \leq 1)$. Thus smaller models are favored more when the precision in the prior for the regression coefficients is increased. Moreover, if η_0 is a linear combination of the columns of X_m, and we let $\gamma \to 1$, then $r(m, m_0) \to 1$, thus favoring the smaller model. For any η_0 in general and $\gamma \to 1$, m will be preferred over m_0 if

$$(y - \eta_0)'(P_{m_0} - P_m)(y - \eta_0) \geq y'(P_{m_0} - P_m)y$$

Setting $\gamma = 0$ in (3.1) corresponds to using a noninformative prior on $\beta^{(m)}|\tau$, but a gamma prior on τ with parameters δ_0 and γ_0. Now $r(m, m_0)$ is a decreasing function of γ_0. Moreover, it can be shown that

$$r(m, m_0) \leq \left(\frac{n - k_{m_0} - 2}{n - k_m - 2} \right) \frac{q_m}{q_{m_0}} \tag{3.2}$$

for all γ_0 larger than some γ^*. The quantity on the right in (3.2) equals $r(m, m_0)$ under joint noninformative priors on (β, τ). Noting that the prior variance of τ is $2\delta_0/\gamma_0^2$, we can interpret this to mean that, as long as the prior variance of τ is not too large (i.e., γ_0 not too small), we are less susceptible to overfitting as compared to the noninformative case. Finally, as n increases, γ^* decreases to zero so that the qualification in the last statement becomes inoperative. Overall, in the context of the L criterion, the informative priors protect the user against overfitting better than the noninformative priors. Moreover, as can be seen in the examples below, the use of calibration numbers can mitigate the overfitting problem.

One relevant aspect of L_m is that, under proper priors, it will not equal zero even if the coefficient of multiple determination, R^2, equals one for some model. Although this can be deduced easily from (2.8), the basic reason for it lies more generally in the averaging operation over the parameter space implied in computing the PDRE. In addition, under any nondeterministic model, the observable y is not a constant a priori. This results in a nonzero calibration

number S_L. Under the improper reference priors, however, one is not guaranteed such automatic protection and hence must be careful to not include in $\mathcal{M}$ any model that can yield a PDRE concentrated at y.

Example 1

Data from Hald (1952) are described in Draper and Smith (1981). These contain four predictors, each measuring the percent composition of each of four ingredients in samples of cement concrete. The response variable measures the heat evolved in calories per gram of cement. These data have been analyzed by many in the literature. Here we illustrate how the priors proposed in (2.3)-(2.6) might be employed in this context. Using experience with similar past experiments, previous models, rows of the current X matrix and other case specific information, suppose the investigator makes the prediction
$\eta_0 = (79, 77, 104, 90, 99, 108, 105, 73, 93, 111, 88, 115, 113)'$.
Putting a relatively small weight on this guess, he assigns $\gamma = 0.1$. Also suppose that previous analyses indicate a prior mean of 0.2 for the precision parameter τ so that $\delta_0/\gamma_0 = 0.2$ and that he is fairly certain that the precision will not exceed 0.5, i.e., $P(\tau < 0.5) \approx 0.95$. These conditions lead to $\delta_0 = 25$, and $\gamma_0 = 125$.

Table 1 reports the results for the top eight models along with the calibration numbers for L_m and M_m. For comparison, values of AIC, BIC and Mallows's C_p are also included.

Table 1 - K, L, and M for the Hald data

Model	K_m	L_m	M_m	C_p	p	AIC	$-2\ BIC$
$x_1\ x_2\ x_4$	0.243	11.43	8.98	3.02	4	61.87	64.12
$x_1\ x_2\ x_3$	.229	11.44	8.99	3.04	4	61.90	64.16
$x_1\ x_3\ x_4$	0.761	11.61	9.21	3.50	4	62.62	64.88
$x_1\ x_2\ x_3\ x_4$	0	11.63	9.19	5.0	5	63.84	66.66
$x_1\ x_2$	1.62	11.84	9.54	2.68	3	62.31	64.00
$x_1\ x_4$	4.64	12.82	10.82	5.49	3	65.63	67.33
$x_2\ x_3\ x_4$	3.94	12.99	11.02	7.34	4	67.47	69.72
$x_3\ x_4$	14.81	17.56	17.51	22.37	3	76.74	78.43
Calibration number		1.74	2.15				

The model (x_1, x_2, x_4) yields the smallest L_m (and M_m) value of 11.43 while (x_1, x_2) has an L_m of 11.84. Although the latter is larger by 0.41 calories per gram, the value of S_L suggests that this difference can be considered small, being about one-quarter calibration unit. Occam's razor then justifies choosing (x_1, x_2). Any further parsimony is not advisable since the best single-predictor-variable model shows $L_m = 36.05$. This is a substantial increase, more than 13 calibration units. On the other hand, the two-variable model (x_1, x_4) appears almost as good as (x_1, x_2). The M criterion also yields the same conclusions.

3.2 Transformation Selection

In linear regression, transformations of the predictor variables can often lead to more accurate predictions and a model that better fits the data. Box and Cox (1964) discuss transformations with an emphasis on transforming the response variable. They also mention briefly a possible Bayesian approach. It appears, however, that the literature on Bayesian transformation methods is sparse at best.

Here, we show how two of the predictive criteria can be used to select a specific member of a suitably chosen parametric transformation family. The K criterion as defined in this article is not applicable to this problem. Consider equation (1.2) where a single model $m \in \mathcal{M}$ consists of a specific member of a given transformation family, and is indexed by a vector of parameters $\alpha = (\alpha_1, \ldots, \alpha_k)'$. Thus X_m in (1.2) denotes a matrix of transformed predictors, and $\beta^{(m)}$ is the vector of regression coefficients corresponding to X_m. The task is to select or estimate α. The criteria L_m and M_m now are functions of α and can be written alternatively as $L(\alpha)$ and $M(\alpha)$. Also X_m may be written as X_α.

The widely used Box-Cox family of power transformations is given by

$$g(x;\alpha) = \begin{cases} \frac{(x^\alpha - 1)}{\alpha} & \alpha \neq 0 \\ log(x) & \alpha = 0 \end{cases} .$$

With this family for each of the predictors, the ith row of X_α would be $(g(x_{i1};\alpha_1), \ldots, g(x_{ik};\alpha_k))$. One can also choose different families for different predictors. Yet another possibility is to transform two predictors with a common parameter value from the same or different families. For instance, one may consider two transformations, $cos(\alpha t)$ and $sin(\alpha t)$, of the same variable t with a common parameter α, leading to X_α having ith row $(cos(\alpha t_i), sin(\alpha t_i))$. It may also be meaningful to consider repeating the same transformation family with the same predictor, but with different parameter values. For example, suppose we have a single predictor x whose value for the ith observation is x_i, and we take $\alpha = (\alpha_1, \alpha_2)'$. Thus X_α would have its ith row of the form $(g(x_i;\alpha_1), g(x_i;\alpha_2))$. These types of transformations may be suitable when one believes the regression function to be linear in two different powers of a variable. Finally, one may choose not to transform some of the variables at all. One ordinary instance of this is the inclusion of an intercept term. The methods presented here allow a great degree of flexibility, in principle. Thus, α can have an effective dimension that is different from the number of columns of X_α, which in turn can be different from the number of physically meaningful variables in the problem.

Example 2

Tukey (1977, p.188) considers data obtained to study the relationship between vapor pressure and temperature of water. The response variable, Y, is log(vapor pressure), and the single predictor x, is water temperature, in degrees Kelvin. Tukey suggests an inverse power transformation on x. To illustrate the proposed procedures, we consider the model in (1.2) with

an intercept and the Box-Cox transformation on x. Again, to denote the dependence of the criteria on α, we write $M_m \equiv M(\alpha)$ and $L_m \equiv L(\alpha)$. Under the noninformative prior (2.1), $M(\alpha)$ and $L(\alpha)$ are equivalent and we get the minimizer $\hat{\alpha} = -1.325$, with $L(\hat{\alpha}) = 0.160$ and $M(\hat{\alpha}) = 0.099$. Results for the coefficient of determination (R^2), residual sums of squares (RSS), and the criterion functions are given for three regression models in Table 2.

Table 2 - Comparison of Models, Vapor Pressure Data

Model	R^2	RSS	$L(\alpha)$	$M(\alpha)$
Y on x	0.907	5.858	3.713	2.299
Y on x^{-1}	0.998	0.130	0.554	0.343
Y on $x_{\hat{\alpha}}$	0.999	0.011	0.160	0.099

Clearly, Y regressed on $x_{\hat{\alpha}}$ gives a larger R^2, a smaller RSS, and a smaller $L(\alpha)$ (and $(M(\alpha))$ than Y regressed on x^{-1}. To compare the two transformations, we have $D_L = .0130$ and $D_M = .0162$. The transformation x^{-1} has an L criterion value that is about 30 calibration units larger than the transformation $x_{\hat{\alpha}}$. A similar large discrepancy is observed under the M criterion. If one wishes to round off the power -1.325 to a more convenient number, one can calculate $L(-1.3) = 0.165$ and note that this is only a small fraction of the calibration number away from $L(-1.325) = 0.160$. Thus -1.3 could be judged an adequate rounded estimate, whereas further rounding to -1 cannot be justified.

3.3 Estimation of a Parametric Variance Function

The above methodology for estimating a transformation on the predictors can be adapted to the problem of estimating a parametric variance function arising from a heteroscedastic model. Consider the regression model

$$Y = X\beta + \epsilon,$$

where

$$\epsilon|\tau \sim No_n(0, \tau\ W)\ ,$$

and W is an unknown diagonal matrix with ith diagonal element $w_i > 0$. Thus, the variance of Y_i is $\tau^{-1}w_i^{-1}$. The w_i's are often modeled as functions of the predictors and its reciprocal is called the variance function. A common form is $w_i^{-1} = exp(x_i'\alpha)$. For a comprehensive nonBayesian treatment of variance function estimation in regression, see Carroll and Ruppert (1988). To emphasize the dependence of W on α, we write $W_m \equiv W_\alpha$ and demonstrate the use of the criteria $L(\alpha)$ and $M(\alpha)$ to estimate α.

With the above precision structure on ϵ, it is convenient to replace X_m and η_0 in (2.3) and (2.4) by $W_m^{1/2}X_m$ and $W_m^{1/2}\eta_0$, respectively. The PDRE now takes the form

$$Z \sim S_n\left(n + \delta_0, W_m^{-1/2}\ \tilde{\eta}_m, \tilde{s}_m^2\ W_m^{-1/2}\ (I + (1-\gamma)\tilde{P}_m)\ W_m^{-1/2}\right)\ , \tag{3.3}$$

where $\tilde{\eta}_m = \tilde{P}_m((1-\gamma)\tilde{Y}+\gamma\tilde{\eta}_0)$, $\tilde{P}_m$ is the orthogonal projection operator onto the column space of $W_m^{1/2}X_m$, $\tilde{Y} = W_m^{1/2}Y$, and $\tilde{\eta}_0 = W_m^{1/2}\eta_0$. Furthermore, $\tilde{s}_m^2 = (n+\delta_0)^{-1}(\tilde{q}_m + \gamma\tilde{p}_m + \gamma_0)$, where $\tilde{q}_m$, $\tilde{p}_m$ are like q_m , p_m with P_m, Y, and η_0 replaced by $\tilde{P}_m$, $\tilde{Y}$, and $\tilde{\eta}_0$. The expressions for the criteria follow from (3.3) in a manner similar to that in Section 2, as do the calibration numbers.

Example 3

Box and Meyer (1986) give data from a fractional-factorial experiment concerning the tensile strength of welds (Y) in an off-line welding experiment performed by the National Railway Corporation of Japan (Taguchi and Wu, 1980). An analysis with regard to variance function estimation is given in Carroll and Ruppert (1988). From the plots they give, there is clear evidence of non-constant variance, and they discuss two parametric models for the variance function. These are

$$var(Y_i) = \tau^{-1}\ exp\,(\alpha\ C_i)\ , \tag{3.4}$$

and

$$var(Y_i) = \tau^{-1}\ exp\,(\alpha_1 B_i + \alpha_2 C_i)\ , \tag{3.5}$$

where B_i and C_i denote the levels of factors B and C, respectively, each taking the values of $+1$ or -1 for each observation. Results using reference priors are summarized in Table 3 below where $\hat{\alpha}_L$ and $\hat{\alpha}_M$ denote the estimates of α based on the L and M criteria. The variance functions (3.4) and (3.5) are denoted by (C) and (B,C), respectively.

Table 3 - Variance Function Parameter Estimates, Tensile Strength Data

Variance Function	$\hat{\alpha}_L$	$\hat{\alpha}_M$	MLE	$L(\alpha)$	$M(\alpha)$
(C)	0.128	0.257	0.257	7.440	5.153
(B,C)	(-.121, .824)	(-.815, 2.001)	(-.815, 2.001)	2.910	1.295

The two criteria give different but comparable estimates. The MLE's of the parameters of each variance function are the same as the estimates obtained by the M criterion since we are using the reference prior (2.1). The calibration numbers D_L and D_M can now be used to compare the two variance functions. Using the larger model with both factors B and C, we obtained $D_L = .762$ and $D_M = .421$. The model with only the factor C has an L criterion value that is about 5.9 calibration units larger. A similar discrepancy is observed under the M criterion. Since the L and M criteria yield different estimates of α, it is interesting to compute L at the minimizer of M and vice-versa. Here, under the model with both factors, $L(\hat{\alpha}_M) = 3.783$ and $M(\hat{\alpha}_L) = 1.600$. Thus 2.910 and 3.783 are about 1.1 calibration units apart under the L criterion, and 1.295 and 1.60 are about 0.7 calibration units apart under the M criterion. The two criteria yield essentially equivalent estimates.

4 Discussion

The minimizations of the criterion functions L and M for the transformation problems were carried out numerically since analytic methods are not readily available. The computations were greatly facilitated by LISP-STAT (Tierney, 1990), which made it possible to carry out the calculations with relatively few lines of code. The functions NEWTONMAX and NELMEADMAX were used with good success. For the examples of this paper, the calculations proceeded quite fast on a SUN SPARC station. Starting values of $\alpha = (1, \ldots, 1)'$ worked well. Other starting values were also used.

An important issue in any model selection procedure is that of model assumptions. It is well known that violations of the same can result in the addition or omission of variables in a variable selection procedure. AIC and BIC, for instance, are not robust to outliers or influential points. The criteria proposed in this article likely suffer from the same problems. Simultaneously checking and selecting models is difficult, and there are no definitive solutions to this problem. However, Cook and Weisberg (1982) recommend that diagnostic checking should precede any variable selection. They advise that the initial or full model be used in the former step. Bayesian techniques available for this purpose include those in Pettit and Smith (1985), Johnson and Geisser (1983), Geisser (1980), Bhattacharjee and Dunsmore (1991) and Bernardo (1985). In the presence of outliers and/or influential points, computation of the criterion with and without the suspect cases will shed light on their effects. Once a model is selected, we recommend the investigator do another diagnostic check with this selection. More work is needed in the area of investigating the robustness of the proposed criteria.

There are several advantages to using the proposed criteria over other well accepted existing model selection criteria in the literature such as AIC and BIC. First, they allow prior input whereas criteria such as AIC and BIC do not. Moreover, our criteria stem from a unified predictive philosophy and the simple notion of a replicate experiment. Justification of BIC for example, is based on an asymptotic argument. Another major advantage is the available calibrations of the L and M criteria. AIC and BIC for example do not have calibrations, and model selection is based on the minimum value. Asymptotic considerations are required for a formal comparison of the AIC values of competing models. The proposed criteria are quite general and, in principle, may be applied in various types of model selection situations. For example, this methodology has been successfully implemented by the authors for the class of generalized linear models.

Appendix

The expression for the M criterion, under model m in (1.2) and priors as in (2.3)-(2.6), is

given by

$$M_m = \pi^{1/2} \left(\frac{, \left(\frac{n+\delta_0}{2}\right)}{, (n+\delta_0/2)} \ (2-\gamma)^{k_m/2} \right)^{1/n} \ a_m^{1/2} \left(1 + \frac{b_m}{a_m}\right)^{1+\frac{\delta_0}{2n}} \ ,$$

where $a_m = q_m + \gamma p_m + \gamma_0$ and $b_m = q_m + \frac{\gamma^2}{2-\gamma} p_m$. Again, both a_m and b_m are linear combinations of the residual sum of squares and the "guessing error".

An exact expression for the K criterion is not available since the necessary integral is not tractable. However, for large n, we can approximate the distribution in (2.7) by a $No_n \left(\eta_m, \left(\frac{n+\delta_0}{n+\delta_0-2}\right)^{-1} s_m^{-2} (I + (1-\gamma)P_m)^{-1} \right)$ distribution. Taking m_0 to be the full model, define

$$v = \frac{(n+\delta_m)(n+\delta_{m_0}-2)}{(n+\delta_{m_0})(n+\delta_m-2)} \ ,$$

where $\delta_m = \delta_{m_0} = \delta_0$ for the normal-gamma priors, and $\delta_m = -k_m$, $\delta_{m_0} = -k_{m_0}$ under noninformative priors. With

$$\zeta = \frac{n+\delta_{m_0}-2}{2s_{m_0}^2(n+\delta_{m_0})(2-\gamma)} + \ \frac{n+\delta_m-2}{2s_m^2(n+\delta_m)},$$

and η_{m,m_0} denoting η_m as in (2.7) with P_m replaced by $P_{m_0} - P_m$, we get

$$K_m \approx \zeta \ \eta'_{m,m_0} \ \eta_{m,m_0} + \ \frac{n}{2} \left(\frac{v s_m^2}{s_{m_0}^2} + \frac{s_{m_0}^2}{v s_m^2} - 2 \right) + \ \frac{k_{m_0} - k_m}{2} \left(\frac{(1-\gamma)s_{m_0}^2}{v s_m^2} - \frac{(1-\gamma)v s_m^2}{(2-\gamma)s_{m_0}^2} \right) \ .$$

If one takes m_0 to be the intercept model, a similar expression for K_m follows.

Analogous to (2.9), we have

$$D_M = (2\pi)^{1/2} \ \tau^{-1/2} \ \ 2^{\frac{k_{m^*}}{2n}} \left(\left(\frac{n}{n-2} \right)^{(n-k_{m^*})/2} - \left(\frac{n}{n-1} \right)^{(n-k_{m^*})} \right)^{1/2} \ .$$

ACKNOWLEDGEMENTS

We thank the Editor, the Associate Editor and the referees for their patient and encouraging reviews of related previous manuscripts. In particular, one referee's comments brought into focus the scope of the proposed criteria while another referee's pointed questions helped in refining the calibration method.

References

[1] Aitchison, J. (1975), "Goodness of Fit Prediction," *Biometrika*, 62, 547-554.

[2] Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, New York : Cambridge University Press.

[3] Aitkin, M. (1991), "Posterior Bayes Factors," (with discussion) , *Journal of the Royal Statistical Society*, Ser. B, 53, 111-142.

[4] Akaike, H. (1973), "Information Theory and an Extension of the Maximum likelihood Principle," *International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, pp.267-281. Budapest: Akademia Kiado.

[5] Bates, D. M., and Watts, D. G. (1988) *Nonlinear Regression Analysis and its Applications*, New York : John Wiley.

[6] Bernardo, J. M. (1985), Comment on "Outliers and Influential Observations in Linear Models", (with discussion), in *Bayesian Statistics 2*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., Amsterdam : North-Holland, p.492.

[7] Bhattacharjee, S. K., and Dunsmore, I. R. (1991) "The Influence of Variables in Logistic Regression", *Biometrika*, 78, 851-856.

[8] Box, G. E. P., and Cox, D. R. (1964), "The Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 211-252.

[9] Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis : Forecasting and Control*, (2nd Ed.), San Francisco: Holden-Day.

[10] Box, G. E. P., and Kanemasu, H. (1973), "Posterior Probabilities of Candidate Models in Model Discrimination," Technical Report 322, University of Wisconsin.

[11] Box, G.E. P., and Meyer, D. R. (1986), "Dispersion Effects From Fractional Designs," *Technometrics*, 28, 19-27.

[12] Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

[13] Box, G. E. P. , and Tidwell, P. W. (1962), "Transformation of the Independent Variables," *Technometrics*, 4, 531-550.

[14] Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London : Chapman and Hall.

[15] Christensen, R. (1990), *Log-Linear Models*, New York : Springer-Verlag.

[16] Clayton, M. K., Geisser, S., and Jennings, D. E. (1986), "A comparison of Several Model Selection Procedures," in *Studies in Bayesian Econometrics and Statistics*, eds. P. K. Goel and A. Zellner, New York : Elsevier.

[17] Cook, R. D., and Weisberg S. (1982), *Residuals and Influence in Regression*, London : Chapman and Hall.

[18] Cox, D. R., and Snell, E. J. (1989), *Analysis of Binary Data*, London : Chapman and Hall.

[19] Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis*, (2nd ed.), New York : John Wiley.

[20] Geisser, S. (1971), "The inferential Use of Predictive Distributions," in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto : Holt, Rinehart and Winston, pp. 456-469.

[21] Geisser, S. (1980), "Discussion: Sampling and Bayes' Inference in Scintific Modelling and Robustness, by G. E. P. Box. *Journal of the Royal Statistical Society*, Ser. A, 30, 143, 416-417.

[22] Geisser, S., and Eddy, W. F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association* , 74, 153-160; Correction, 75, 765.

[23] Hald, A. (1952), *Statistical Theory With Engineering Applications*, New York : John Wiley.

[24] Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-51.

[25] Hosmer, D. W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York : John Wiley.

[26] Johnson, W., and Geisser, S. (1983), "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis," *Journal of the American Statistical Association*, 78, 137-144.

[27] Lempers, F. B. (1971), *Posterior Probabilities of Alternative Linear Models*, Rotterdam : Rotterdam University Press.

[28] Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 30, 31-66.

[29] Mallows, C. L. (1973), "Some Comments on C_p, " *Technometrics*, 15, 661-675.

[30] McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

[31] McCulloch, R. E. (1989), "Local Model Influence," *Journal of the American Statistical Association*, 84, 473-478.

[32] Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," (with discussion), *Journal of the American Statistical Association* , 83, 1023-1036.

[33] Pettit, L. I., and Smith, A. F. M. (1985), "Outliers and Influential Observations in Linear Models", (with discussion), in *Bayesian Statistics 2*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., Amsterdam : North-Holland, p.492.

[34] San Martini, A., and Spezzaferri, F. (1984), "A Predictive Model Selection Criterion," *Journal of the Royal Society*, Ser. B, 46, 296-303.

[35] San Martini, A., and Spezzaferri, F. (1986) "Selection of Variables in Multiple Regression for Prediction and Control, " *Statistica*, 118-127.

[36] Schwarz, G. (1978), "Estimating the Dimension of a Model", *Annals of Statistics*, 6, 461-464.

[37] Smith, A. F. M., and Spiegelhalter, D. J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B., 42, 213-220.

[38] Spiegelhalter, D. J., and Smith, A. F. M. (1982), "Bayes Factors for Linear and Log-linear Models with Vague Prior Information", *Journal of the Royal Statistical Society*, Ser. B., 44, 377-387.

[39] Taguchi, G., and Wu, Y. (1980), *Introduction to Off-Line Quality Control*, Nagoya, Japan : Central Japan Quality Control Association.

[40] Tierney, L. (1990), *Lisp-Stat : An Object-Oriented Environment for Statistical Computing and Dynamic Graphics* , New York: John Wiley.

[41] Tukey, J. W. (1977), *Exploratory Data Analysis,* Reading, MA: Addison-Wesley.

[42] West, M., and Harrison, J. (1989), *Bayesian Forecasting and Dynamic Models*, New York : Springer-Verlag.

[43] Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions", in *Studies in Bayesian Econometrics and Statistics*, eds. P. K. Goel and A. Zellner, New York : Elsevier, pp.233-243.