Bayesian Predictive Simultaneous Variable and Transformation Selection in the Linear Model

Jennifer A. Hoeting
Department of Statistics, Colorado State University

Joseph G. Ibrahim
Department of Biostatistics, Harvard School of Public Health
and Dana-Farber Cancer Institute

Summary

Variable selection and transformation selection are two commonly encountered problems in the linear model. It is often of interest to combine these two procedures in an analysis. Due to recent developments in computing technology, such a procedure is now feasible. In this paper, we propose two variable and transformation selection procedures on the predictor variables in the linear model. The first procedure is a simultaneous variable and transformation selection procedure. For data sets with many predictors, a backward elimination procedure for variables and transformations is also presented. The procedures are based on Bayesian model selection criteria introduced by Ibrahim and Laud (1994) and Laud and Ibrahim (1995). Several examples are given to illustrate the methodology.

Key Words: Bayesian criterion; Box-Cox transformation; Calibration number; Comparison score; Backwards selection.

1 Introduction

In linear regression, both variable subset selection and transformation selection can lead to more accurate predictions and a model that better fits the data. In many problems, transforming the predictors after conducting a variable subset selection will result in a better fitting model. However, since variable selection and transformation selection are typically performed in a particular order, the chosen model often depends on which procedure is done first. In these cases, there is a need to develop methods in which variable and transformation selection can be done simultaneously.

Since variable selection and transformation selection can each be viewed as a model selection problem, we can unify the two procedures into one. With the recent advances in computing technology, such methods are computationally feasible. A recent paper by Hoeting, Raftery, and Madigan (1995) proposes a simultaneous approach to variable and transformation selection based on change point transformations. Smith and Kohn (1996) propose a nonparametric approach to Bayesian variable selection automatically selecting independent variables and a power transformation for the response.

In this article, we address variable selection and transformation selection from a predictive Bayesian viewpoint. We propose two methods to perform variable and transformation selection. The first procedure is a simultaneous variable and transformation selection procedure. A backward elimination procedure for variables and transformations is also presented for problems with many predictors. We use a predictive criterion developed in Ibrahim and Laud (1994) and Laud and Ibrahim (1995) and a comparison score described here to compare models. Several examples involving real data are given to demonstrate the methodology.

The rest of this article is organized as follows. In Section 2, we establish notation, review variable and transformation selection, and discuss the predictive criterion along with the comparison score. In Section 3, we discuss two methods for variable and transformation selection , and in Section 4 we present several examples to illustrate the methodology. We conclude the article with a brief discussion.

2 Preliminaries

2.1 Model and Notation

Consider the usual normal linear regression model

$$Y = X\beta + \epsilon , \tag{2.1}$$

where $Y = (y_1, \ldots, y_n)'$ is an n-vector of responses, X is an $n \times (k+1)$ matrix of fixed predictor variables with ith row $x_i' = (x_{i0}, x_{i1}, \ldots, x_{ik})$, $x_{i0} = 1$, $\beta = (\beta_0, \ldots, \beta_k)'$ is a $(k+1)$-vector of regression coefficients, and ϵ is an n-vector of random errors. The intercept term is included as a matter of convention. The error vector ϵ is assumed to have an n-dimensional multivariate normal distribution with mean 0 and *precision* matrix τI, denoted by $N_n(0, \tau I)$, where τ is a positive scalar parameter, and I is the $n \times n$ identity matrix.

Since we unify the treatment of variable and transformation selection by treating each as a model selection problem, it is more convenient to rewrite equation 2.1 by making explicit the model dependence of the predictor matrix and the regression coefficients. Thus we write

$$Y = X_m \ \beta^{(m)} + \epsilon , \tag{2.2}$$

where X_m and $\beta^{(m)}$ denote the X matrix and β vector under model m, respectively. Further, let $m \in \mathcal{M}$, where $\mathcal{M}$ denotes the appropriate model space. For the variable selection problem, $\mathcal{M}$ is a discrete space, with each point in the space corresponding to a particular predictor subset choice. For example, if there are k original variables $\mathcal{M}$ consists of 2^k elements. In experimental design, we consider models obtained by deleting various main effects and interactions in an analysis of variance.

In the transformation selection problem, $\mathcal{M}$ is often a continuous model space whose elements correspond to specific members of a given transformation family. For example, for a single predictor X_1 the model space $\mathcal{M}$ might be $(\alpha : \alpha > 0)$ where X_1^α is the transformation for X_1. We note here that the dimension of $\beta^{(m)}$ changes with m for the subset selection problem as different models have different numbers of predictors, but it does not change for the transformation selection problem as one transformation is selected for a particular predictor. However, the rank of X_m changes, in general, with m in either problem.

2.2 Variable Selection

In selecting variables in linear regression, we are interested in considering the 2^k possible models that can be obtained from equation 2.1 by retaining various subsets of the last k columns of the full matrix X, and modifying the length of β accordingly. In this case, m is a subset of the integers $\{0, \ldots, k\}$ containing 0, and k_m denotes the number of elements of m. Thus m identifies a model with an intercept and a specific choice of $k_m - 1$ predictor variables. Choosing one of the models in equation 2.2 is the goal of variable selection methods. This problem has received much attention in the literature. See, for example, Lindley (1968), Mallows (1973), Hocking (1976), and Lempers (1971). Additional references appear in an article by Mitchell and Beauchamp (1988).

From the Bayesian viewpoint, the approach to the variable selection problem is, in principle, straightforward. The researcher needs to specify the prior probability of each model, a prior distribution for all of the parameters in each model, and compute the posterior probability of each model given the data. As pointed out by Mitchell and Beauchamp (1988) among others, specifying priors for all these parameters is a monumental task seldom undertaken in practice. They have therefore aimed their article at providing what Berger (1988) calls "a semiautomatic Bayesian method" in his discussion of their article.

The Akaike information criterion (AIC) (Akaike, 1973) and the Bayes information criterion (BIC) (Schwarz, 1978) are two widely accepted criterion methods for variable selection. Hurvich and Tsai (1989) suggested an adjustment to AIC to be used when the ratio of sample size to the number of parameters is small (AICC). Two drawbacks of these criteria are that they do not allow for prior input and their definitions and/or calibrations depend upon asymptotics.

Ibrahim and Laud (1994) and Laud and Ibrahim (1995) propose a different approach to variable selection by adopting a predictive Bayesian viewpoint in principle. Geisser and Eddy (1979), and Clayton, Geisser, and Jennings (1986) also advocate the predictive view with the argument that a researcher's ultimate goal is often prediction rather than parameter estimation. Our approach here is based on the philosophy of Geisser (1971), where he encourages the use of predictive distributions for inferential purposes. We deemphasize the parameters and focus on the observables in our prior elicitation and variable selection. Variable selection is an ideal setting for such an approach since the parameter

vector β does not carry much physical meaning at the outset. One is uncertain even about its length. More importantly, any given component of this vector has an entirely different physical meaning that depends on the model it appears in. We implement the predictive philosophy in two ways. First, the prior for $\beta^{(m)}|\tau$ is constructed in an automated fashion from a prior prediction (or guess) for Y along with a number quantifying one's belief in this guess relative to the information contained in the experiment. Secondly, the variable selection is based on how well each model would predict the observed $Y = y$ for a replicate experiment with the same covariates. The necessary quantification is accomplished via a criterion computed from certain predictive distributions. The main motivation of the predictive criterion is that it identifies a set covariates that when measured will yield predictions as close as possible to the observed data for a replicate experiment.

In a similar fashion, Geisser and Eddy (1979) introduce the *predictive sample reuse* (PSR) method for model selection, and Johnson and Geisser (1983) define predictive criteria for detecting influential observations. The criterion we present here is easily interpreted and has explicit closed form expressions.

2.3   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. Many different transformation families have been suggested in the literature, along with a few methods for selecting a specific member of such a family. A non-Bayesian method was proposed in Box and Tidwell (1962), while 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. For a recent general treatment of transformations in regression, see Cook and Weisberg (1982), Carroll and Ruppert (1988), and the many references therein.

Laud and Ibrahim (1995) use a predictive criterion for selecting a specific member of a suitably chosen parametric transformation family. For this problem, a model $m \in \mathcal{M}$ consists of a specific member of a given transformation family and is indexed by a vector of parameters $\alpha = (\alpha_1, \ldots, \alpha_p)'$. Thus X_m in equation 2.2 denotes a matrix of transformed predictors, and $\beta^{(m)}$ is the vector of regression coefficients corresponding to X_m.

A useful example of a family of transformations on a predictor x is the Box-Cox power family of transformations given by

$$g(x; \alpha) = \begin{cases} \frac{(x^\alpha - 1)}{\alpha} & \alpha \neq 0 \\ log(x) & \alpha = 0 \end{cases} . \tag{2.3}$$

If one were to choose this family for each of the predictors, then for a given α, the ith row of X_m would be $(g(x_{i1}; \alpha_1), \ldots, g(x_{ip}; \alpha_p))$. One can also choose different families for different predictors, leading to X_m having ith row of the form $(g_1(x_{i1}; \alpha_1), \ldots, g_p(x_{ip}; \alpha_p))$. 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_m having ith row $(cos(\alpha t_i), sin(\alpha t_i))$. The predictive methods discussed here allow a great deal of flexibility in principle. Further possibilities and more examples including variance stabilization are given in Laud and Ibrahim (1995).

2.4   Predictive Criterion

In this section we briefly review the predictive criterion proposed by Ibrahim and Laud (1994) and Laud and Ibrahim (1995). Let $\theta^{(m)}$ be the vector of parameters for model m. 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

$$p(z|m, y) = \int p(z|m, \theta^{(m)}) \; \pi(\theta^{(m)}|y, m) \; d\theta^{(m)} . \tag{2.4}$$

The density in (2.4) is called the *Predictive Density of a Replicate Experiment*(PDRE). In equation 2.2 for example, the PDRE depends on the design matrix X_m for model m and would be given by

5

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

For a given model m, consider

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

where the expectation is taken with respect to the PDRE in (2.4). The measure L_m^2 has the decomposition

$$L_m^2 = \sum_{i=1}^{n} \left\{ [E(Z_i) - y_i]^2 + Var(Z_i) \right\} \; , \tag{2.5}$$

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. The statistic L_m is called the *L criterion*. We see from (2.5) that the predictive criterion identifies covariates that would yield predictions as close as possible to the observed data for a replicate experiment. In (2.5), a given value of i corresponds to a given covariate profile. The weighting of the covariate profiles will determine the magnitude of the overall criteria, and hence the manner in which the data are collected. Thus different "best" models will emerge with different weightings of the covariate profiles.

2.5   Prior Distributions

In this section, we suggest appropriate informative and noninformative prior distributions and discuss the interpretation of the L_m criteria under these prior distributions as described in Ibrahim and Laud (1994) and Laud and Ibrahim (1995).

For noninformative priors for $(\beta^{(m)}, \tau)$, we consider Jeffreys' modified prior which given by

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

The resulting predictive distribution in this case is

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

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.

Ibrahim and Laud (1994) and Laud and Ibrahim (1995) discuss informative prior distributions for $(\beta^{(m)}, \tau)$ which are generated from observables. We briefly describe these here. The investigator incorporates all prior knowledge into a prior prediction for Y, denoted by η_0. Under model m with design matrix X_m, the prior mean of $\beta^{(m)} | \tau$ is taken to be

$$\mu^{(m)} = (X_m' X_m)^{-1} X_m' \ \eta_0 \ . \tag{2.6}$$

The vector η_0 is a fixed vector regardless of the model under consideration. Its specification may be made in one of several ways as discussed by Ibrahim and Laud (1994) and Laud and Ibrahim (1995). The prior precision matrix of $\beta^{(m)} | \tau$ is taken to be of the form $\tau \ T_m$, where

$$T_m = c \ (X_m' X_m),$$

with $c \geq 0$ quantifying, in multiples of the present experiment, the importance one wishes to attach to the prior guess η_0. Finally, $\beta^{(m)} | \tau$ is taken to be normally distributed, i.e.,

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

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. Similarly, the prior on the precision matrix of $\beta | \tau$ changes with the model as the prior depends upon $(X_m' X_m)$.

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

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

With this prior and the likelihood implied by equation 2.2 for each m, a straightforward derivation yields

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

where $S_n(\nu, \mu, \Sigma)$ denotes the n dimensional multivariate t-distribution with ν degrees of freedom, location parameter μ, and dispersion matrix Σ. Here $\gamma = c/(1+c)$, $P_m = X_m(X_m'X_m)^{-1}X_m'$ $\eta_m = P_m\left(\gamma\eta_0 + (1-\gamma)y\right)$, $s_m^2 = (n+\delta_0)^{-1}(q_m + \gamma p_m + \gamma_0)$, $q_m = y'(I-P_m)y$, and $p_m = (y-\eta_0)'P_m(y-\eta_0)$.

The L criterion under model m is now given by

$$L_m = \left\{(1+\lambda_m)q_m + \gamma(\gamma+\lambda_m)p_m + \lambda_m\gamma_0\right\}^{1/2} \ , \tag{2.7}$$

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 noninformative priors, equation 2.7 reduces to $L_m = (2(n-1)(n-k_m-2)^{-1}q_m)^{1/2}$. In this case, L_m is similar to the root mean square criterion. We see that (2.7) depends on η_0 through p_m. Potentially different orderings of the models may result by varying the choices of η_0. Laud and Ibrahim (1995) and Ibrahim (1997) show that model choice may be sensitive to the choice of η_0.

2.6 Calibrating the L_m criteria

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

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

Ibrahim and Laud (1994) and Laud and Ibrahim (1995) refer to $S_{L_{m^*}}$ 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 (2.1) and (2.2), the marginal distribution of Y is multivariate t. To calculate the calibration number $S_{L_{m^*}}$, 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 deviation of these values provides a Monte Carlo approximation to $S_{L_{m^*}}$. If one is using the noninformative priors in equation 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 $Y|\tau \sim N_n(0, \tau(I - P_{m^*}))$ 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 $[Var(L_{m^*}|\tau = \tilde{\tau})]^{1/2}$.

Here, we use the calibration number to compute a standardized score for each model. The *comparison score* is a distance measure which gives the number of calibration units that a given model is from the model with the smallest criterion value. The comparison score for a model m is given by

$$\psi_m = \frac{L_m - L_{m^*}}{\hat{S}_{L_{m^*}}},$$

where m^* is the model with the smallest L_m value.

When selecting the "best" model, one can either choose the model with the smallest L_m value or apply Occam's razor and choose the most parsimonious model with a small comparison score. Under this second scenario, we generally choose the most parsimonious model with a comparison score of 2 or less. This somewhat arbitrary cut-off for ψ corresponds to selecting a model within 2 standard deviations of the model with the smallest comparison score (m^*), where the standard deviation is computed with respect to m^*.

3 Two Methods for Variable and Transformation Selection

3.1 Predictive Simultaneous Variable and Transformation Selection

The first method we present is a predictive simultaneous variable and transformation selection. In this method, we first construct the 2^k possible subset models. Then, for each model, we find the optimal set of transformations by minimizing the L criterion given a parametric transformation family for the predictors. This procedure leads to 2^k minimizations of the L criterion, and thus may be quite intensive when the number of predictors is large. This algorithm can be summarized as follows.

1. Construct the 2^k subset models arising from all possible subsets of the k predictors.

2. For each model, apply a parametric transformation method such as Box-Cox and find the set of transformations which minimizes the L criterion. Thus, there are 2^k criterion values based on 2^k transformed models. We carry out the minimization numerically as described in Section 5.

3. Compute the calibration number based on the criterion minimizing model and compute the comparison score for each model.

4. Select a model based on the criterion value and the comparison score.

We construct the 2^k possible sets as in regular subset selection, but then we transform each subset so that the criterion value is based on the transformed predictors. This procedure can be computationally intensive if k is large, say $k \geq 15$. However, for small to moderate k, this procedure works quite well in identifying good models. For large k, modifications of the procedure are needed, such as a backward elimination procedure that does not require minimization of L_m for all the 2^k possible models. One such procedure is described in Section 3.2.

In the two examples discussed below, we adopt the Box-Cox approach to transformations, where the components of α can theoretically take on any value from $-\infty$ to ∞.

To some, this may be not be a completely satisfactory approach since a power transformation of $-.3956$, for example, is not easy to interpret. An alternative to the continuous transformation approach is to consider a discrete set of transformations. For example, the set $\alpha = (-1, 0, 0.5, 1, 2)$ which corresponds to $(X^{-1}, log(X), \sqrt{X}, X, X^2)$ could be considered as the set of possible transformations for each predictor. One advantage of the discrete approach is that the transformations are easier to interpret than the continuous transformations. For some data sets, the discrete approach also allows for faster computation as the minimization step required for continuous transformations is avoided. We demonstrate both the continuous and discrete approach to transformations below.

3.2 Backward Elimination Procedure for Variables and Transformations

As an alternative to the computer intensive simultaneous procedure, we suggest a methodology for backwards elimination. The procedure, which alternates between variable selection and transformation selection, is useful when there are a large number of predictors.

In the backwards elimination procedure, we start with the full model with k predictors, and then construct the transformed model. For the second step, we find the optimal model among all possible combinations of $k - 1$ *transformed* predictors. Using these $k - 1$ predictors, we refit the transformations to find the optimal set of transformations for the selected $k - 1$ predictors. We repeat this process until all predictors have been omitted, alternating between variable selection using the most recent set of transformations and transformation selection for the selected predictors. We then compare models at each step via the comparison score.

The backwards elimination procedure can be summarized as follows:

1. Find the set of transformations which minimizes the L-criterion for the full model with k predictors.

2. From the set of all possible predictors, find the model with $k - 1$ predictors which minimizes the L-criterion using k *transformed* predictors from Step 1.

3. Once the best $k - 1$ predictor model is identified in Step 2, find the set of transformations which minimizes the L-criterion for the $k - 1$ selected predictors.

4. Using the $k - 1$ transformed predictors from Step 3, find the model with $k - 2$ predictors which minimizes the L-criterion.

5. Repeat this process until all predictors are omitted.

6. Select a model based on the criterion value and the comparison score.

The main difference between this approach and more traditional backwards elimination is the addition of transformations in the procedure.

4   Examples

4.1   Example 1: Hald Cement Data

4.1.1   Variable and Transformation Selection

The Hald (1952) cement data have been analyzed by many researchers. A description of the data can be found in Draper and Smith (1981). There are four predictors, each measuring the percentage composition of a particular ingredient in samples of cement concrete. The response is the heat evolved in calories per gram of cement. To demonstrate our methodology we will present results from the simultaneous all subsets and backwards elimination procedures under both noninformative and informative priors for the Hald data.

For the informative prior case, we adopt (and describe here) the prior parameters presented in Laud and Ibrahim (1995). Based upon 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., $P(\tau < 0.5) \approx 1$. These conditions lead to $\delta_0 = 25$, and $\gamma_0 = 125$.

Table 1 reports the L_m values for the top eight models from the simultaneous all subsets procedure under informative prior distributions along with the comparison scores. The full model with 4 predictors yields the smallest L_m value. Applying Occam's razor, the most parsimonious model within 2 calibration units of the best model is the model with predictors 1 and 2.

In Table 1 we also include AIC, AICC, and BIC values for comparison with our methods. The number of estimable parameters used to compute the complexity penalty for these three techniques included the intercept, the coefficients for each predictor, the variance, and the transformation parameters. In spite of the penalties for overparameterization, AIC and BIC indicate the four predictor model as the best model. The small sample version of AIC (AICC) indicates the more parsimonious two predictor model. While the AICC method indicated the more parsimonious model, it is important to note that the AIC, BIC, and AICC methods are typically used for either variable selection *or* transformation selection, but not the simultaneous application of variable and transformation selection. The models compared here were chosen using the Bayesian predictive method.

Table 2 gives the five models chosen by the backwards elimination procedure under informative prior distributions along with the comparison scores. The same 4 predictor model indicated by the simultaneous all subsets procedure was identified by the stepwise procedure. Note that line 2 of Table 2 does not agree with line 2 of Table 1 (the best 3 predictor model) because the best 3 predictor model in the backwards selection procedure is chosen based on the transformations for the 4 predictor model. This is a short-coming of the backwards selection procedure.

Table 3 reports the results for the top eight models from the all subsets procedure under noninformative prior distributions along with the comparison scores. With the exception of an alternating in order of the second best and third best model indicated under informative prior distributions, the informative and noninformative prior distributions produce very similar results for the simultaneous all subsets procedure.

Table 4 reports the top 8 models from the simultaneous all subsets procedure with a discrete set of power transformations $\alpha = (-1, 0, 0.5, 1, 2)$ under noninformative priors. The model which minimizes L_m for the discrete set of transformations is quite similar to the model selected under continuous transformations (Table 3). There is an increase in

the minimum L_m value with the discrete transformations, but the trade-off is that the discrete transformation model is more interpretable than the continuous transformation model.

In Tables 1–4, AIC and BIC were minimized for the model with the minimum L_m. AICC tended to select models with fewer parameters.

4.1.2 Sensitivity Analyses

We performed sensitivity analyses to investigate the sensitivity of the results to the prior hyperparameter values for the informative prior distributions. The hyperparameter for the prior mean of $\beta^{(m)}|\tau$ in equation 2.6 (η_0) is of particular interest as η_0 is the prior predictor for the response, Y. In addition to the results shown in Table 1, we explored two alternative scenarios for η_0. First, we set η_0 equal to the least squares estimate for Y when all untransformed predictors were in the model. Second, we set η_0 equal to the least squares estimate for Y when only the first predictor (X_1) was in the model. In both cases, the ranking of the models via the L criterion was identical to the results shown in Table 1 *except* that the order of the second and third models were reversed for the η_0 values based on the X_1 model. The selected transformations under these three scenarios were quite similar.

The results for the informative prior in the cement example (Table 1) are quite similar to the results for the non-informative prior (Table 3). This is partly due to the small value of γ (γ=.1) used for this example (see equation 2.7) and also due to the particular values of η_0 that were selected. For larger values of γ, the model ordering remains the same. For other values of η_0, for example when η_0 is set to equal the least squares estimate for Y when only the first predictor (X_1) was in the model, the ordering of the models does change as the value of γ increases. One might argue that a preferred approach is to use non-informative priors to avoid this difficulty. However, if an informative prior is selected by an expert with prior knowledge about the data, the expert judgment might be useful for the model selection process. As pointed out by a referee, if the sample size is large, the non-informative priors will probably be more practical.

The results are quite insensitive to the choice of the remaining hyperparameters, δ_0 and γ_0. Setting η_0 and γ as in the informative prior example for Table 1, we used several

14

combinations of δ_0 and γ_0 so the shape of the distribution of τ was quite different and the results were very similar to those shown in Table 1, with the same model ordering and very similar transformations.

4.2   Example 2: Weisberg Highway Data

To demonstrate the backwards elimination procedure for a larger data set, we consider the highway accident data from Weisberg (1985). The dependent variable is the automobile accident rate on 39 highway sections, and there are 13 potential predictor variables, listed in Table 5. For transformation selection, the predictors 1–4, 6, and 9 were considered to be continuous. The other predictors were not considered for transformation, but were included in the variable selection component of the backwards elimination procedure.

With the exception of the null and one predictor models, all models have very similar L_m values (Table 6). The most parsimonious model that is less than 2 calibration units from the best model is the 2 predictor model which includes predictors 1 (length of the highway segment) and 9 (number of access points per mile). This same model is selected by AICC and BIC while AIC indicates the four predictor model. In previous analyses with untransformed predictors, the model with predictors 1, 4, and 9 has been indicated (Weisberg 1985).

5   Discussion

The analyses of the Hald data (Tables 1–4) demonstrate that there is considerable model uncertainty. While the same model minimizes L_m under informative and noninformative priors and using both the simultaneous all subsets and backwards selection methods, there are many models that have quite similar L_m values. One way to address the issue of model uncertainty is to average over the selected models. Bayesian model averaging, which has been shown to improve predictive performance, has been discussed in many contexts (see, for example, Kass and Raftery 1995). One method for computing the posterior model probabilities required for Bayesian model averaging for linear models is presented in Raftery, Hoeting, and Madigan (1997). An approximation to the posterior model probability which is based on BIC is suggested by Raftery (1995).

For the continuous Box-Cox transformations, the minimizations of the L_m for transformations were carried out numerically, since analytic methods are not readily available. The LISP-STAT (Tierney, 1990) functions NEWTONMAX and NELMEADMAX were used to perform these minimizations. For the starting values of the transformation parameter, $\alpha = (1, \ldots, 1)'$ worked well. Other starting values were also used.

The checking of model assumptions is an important component of any model selection exercise. An ideal procedure would simultaneously check model assumptions while selecting predictors and transformations. However, a satisfactory approach to this problem has not been proposed. Cook and Weisberg (1982) recommend that diagnostic checks on the full or initial model should be performed before any variable selection. Once a model is selected, the investigator should also do a diagnostic check to determine whether the model assumptions hold. Additional work needs to be done to investigate the robustness of the L_m criteria to violations in the assumptions of linear regression. We also note that the powers of the covariates can change dramatically when different models are considered as seen in Table 1. This implies that the choice of metric is poorly determined, a result which is consistent with Box and Tiao (1973).

We have presented here two new procedures for variable and transformation selection. There are several benefits to using these procedures. First the simultaneous approach to variable and transformation selection avoids the problem that the chosen model depends on the order of the methods. Second, the comparison score described here operationalizes the calibration of Laud and Ibrahim's L_m criteria. AIC, AICC, and BIC do not have calibrations and thus model selection is based on the minimum criterion value. The calibration of the L_m criterion allows for more meaningful comparisons between models and allows the user to choose the model based on both parsimony and quality of predictions.

6   Appendix: Software for Implementing These Procedures

SIMSEL is a set of XLISP-STAT functions which can be obtained free of charge from Statlib via the World Wide Web at http://lib.stat.cmu.edu/xlispstat/simsel, or by sending the e-mail message "send simsel from xlispstat" to *statlib@stat.cmu.edu*. SIMSEL includes a procedure to compute the L-criterion for a specified model and also an implementation

of the simultaneous all subsets procedure described in the paper. Information on how to obtain XLISP-STAT, a Lisp based statistical computing environment, is also available from Statlib.

Acknowledgements

References

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.

Berger, J. O. (1988), Comments on "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1033-1034.

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.

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

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

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

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.

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

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

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.

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.

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

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

Hoeting, J.A., Raftery, A.E., and Madigan, D. (1995), "Simultaneous Variable and Transformation Selection in Linear Regression," Technical Report 9506, Department of Statistics, Colorado State University.

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

Ibrahim, J. G., and Laud, P. W. (1994), "A Predictive Approach to the Analysis of Designed Experiments,". *Journal of the American Statistical Association*, 89, 309-319.

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.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Laud, P.W. and Ibrahim, J.G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society - B*, 57, 247-262.

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

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

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

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.

Raftery, A.E. (1995), "Bayesian model selection in social research," *Sociological Methodology 1995* (ed. Peter V. Marsden), Oxford, U.K.: Blackwells, 111-196.

Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997), "Bayesian Model Averaging for Linear Regression Models," to appear in *Journal of the American Statistical Association*.

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

Smith, M. and Kohn, R. (1996), "Nonparametrics Regression using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317-367.

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

Weisberg, S. (1985), *Applied Linear Regression*, New York: Wiley.

Table 1: Cement data: Top 8 models for all subsets under informative prior distributions. Calibration number =1.69.

Predictors (power)				L_m	ψ	AIC	AICC	BIC
X_1	X_2	X_3	X_4					
1.11	−0.75	−0.06	1.88	8.3		38	148	69
1.16	0.24	−1.52		10.5	1.3	66	102	91
1.14	0.23			10.6	1.4	63	77	82
1.09	0.05		0.47	10.7	1.5	67	103	92
1.24		1.34	1.19	11.3	1.8	69	105	95
0.71			1.09	12.2	2.3	70	84	88
	1.91	0.61	0.86	12.4	2.4	73	109	99
		−0.06	1.35	14.6	3.7	76	90	95

Table 2: Cement data: Backwards selection under informative prior distributions. Calibration number = 1.69.

# of predictors	Predictors (power)				L_m	ψ	AIC	AICC	BIC
	X_1	X_2	X_3	X_4					
4	1.11	−0.75	−0.06	1.88	8.3		38	148	69
3	1.09	0.05		0.47	10.7	1.5	67	103	92
2	1.14	0.23			10.6	1.4	63	77	82
1		−0.72			34.2	15.4	98	103	111
0	null model				61.8	31.7	110	112	117

Table 3: Cement data: Top 8 models for all subsets under noninformative prior distributions. Calibration number = 0.85.

| Predictors (power) | | | | L_m | ψ | AIC | AICC | BIC |
X_1	X_2	X_3	X_4					
1.11	-0.75	-0.06	1.88	3.5		38	148	69
1.14	0.23			10.7	8.5	63	77	82
1.17	0.25	-1.67		10.9	8.7	66	102	91
1.09	0.08		0.55	11.4	9.3	67	103	92
1.25		1.34	1.19	12.6	10.8	69	105	95
0.71			1.09	13.9	12.2	70	84	88
	1.91	0.60	0.86	14.7	13.3	73	109	99
		-0.07	1.34	18.0	17.2	76	90	95

Table 4: Cement data: Top 8 models for all subsets, discrete transformations, under noninformative prior distributions. Calibration number = 1.34.

Predictors				L_m	ψ	AIC	AICC	BIC
X_1	X_2^{-1}	$\log X_3$	X_4^2	5.4		49	159	81
X_1	X_2^{-1}	$\sqrt{X_3}$	X_4^2	10.0	3.4	65	175	97
X_1	$\log X_2$	$\log X_3$	X_4^2	10.5	3.8	67	177	98
X_1	$\log X_2$	$\sqrt{X_3}$	X_4^2	10.7	3.9	67	177	98
X_1	$\sqrt{X_2}$	$\log X_3$	X_4	10.9	4.2	68	178	99
X_1	$\log X_2$			11.0	4.2	64	78	82
X_1	$\sqrt{X_2}$	$\sqrt{X_3}$	X_4	11.1	4.2	68	178	99
X_1	$\log X_2$	X_3^{-1}		11.2	4.4	66	102	91

Table 5: Highway data: Predictors of highway accident rate.

1	length of the segment in miles
2	average daily traffic count in thousands
3	truck volume as a percent of the total volume
4	speed limit (in 1973, before the 55 mpg limits)
5	lane width in feet
6	width in feet of outer shoulder on the roadway
7	number of freeway-type interchanges per mile in the segment
8	number of signalized interchanges per mile in the segment
9	number of access points per mile in the segment
10	total number of lanes of traffic in both directions
11	1 if federal aid interstate highway, 0 otherwise
12	1 if principal arterial highway, 0 otherwise
13	1 if major arterial highway, 0 otherwise

Table 6: Highway data: Backwards elimination under noninformative priors. The top table shows the models selected for each model size. The model rank given in the table is the model order, where model 1 has the lowest L_m value. The transformations for predictors 1, 2, 3, 4, 6, and 9 are given in the next table. Calibration number = 1.09.

p	Predictors	L_m	ψ	Rank	AIC	AICC	BIC
13	1 2 3 4 5 6 7 8 9 10 11 12 13	9.4	.624	10	152	314	301
12	1 2 3 4 5 6 7 8 9 10 11 12	9.2	.444	9	148	265	286
11	1 2 3 4 6 7 8 9 10 11 12	9.0	.289	7	144	230	272
10	1 2 3 4 6 7 8 9 11 12	8.9	.149	6	140	203	257
9	1 2 3 4 6 8 9 11 12	8.7	.047	4	137	183	243
8	1 2 3 4 6 8 9 12	8.7		1	134	168	229
7	1 2 4 6 8 9 12	8.7	.013	3	131	156	216
6	1 4 6 8 9 12	8.7	.001	2	128	146	203
5	1 4 8 9 12	8.8	.139	5	127	139	191
4	1 8 9 12	9.2	.425	8	127	135	180
3	1 8 9	9.7	.929	11	129	133	171
2	1 9	9.9	1.141	12	128	130	160
1	9	11.9	2.892	13	138	140	160
0	null model	17.8	8.332	14	167	168	178

	Predictors (power)					
p	X_1	X_2	X_3	X_4	X_6	X_9
13	-0.59	-2.57	-0.83	2.14	-4.86	0.40
12	-0.58	-2.54	-0.81	2.14	-4.87	0.39
11	-0.56	-2.35	-0.83	2.13	-5.20	0.40
10	-0.51	-2.23	-1.15	1.88	-5.34	0.39
9	-0.50	-1.93	-1.77	1.32	-5.77	0.35
8	-0.41	-3.16	-1.56	1.03	-5.40	0.43
7	-0.34	-3.06		1.14	-5.43	0.41
6	-0.35			1.65	-5.77	0.37
5	-0.45			1.34		0.44
4	-0.55					0.25
3	-0.72					0.53
2	-0.62					0.64
1						0.86