Variable Selection and Model Comparison in Regresssion

John Geweke

University of Minnesota and
Federal Reserve Bank of Minneapolis

May 20, 1994
Final revision: November 9, 1994

Abstract

In the specification of linear regression models it is common to indicate a list of candidate variables from which a subset enters the model with nonzero coefficients. This paper interprets this specification as a mixed continuous-discrete prior distribution for coefficient values. It then utilizes a Gibbs sampler to construct posterior moments. It is shown how this method can incorporate sign constraints and provide posterior probabilities for all possible subsets of regressors. The methods are illustrated using some standard data sets.

Forthcoming in J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.), *Proceedings of the Fifth Valencia International Meeting on Bayesian Statistics.* Partial financial support from NSF grant SES-9210070 is gratefully acknowledged. Thanks to Rob McCulloch for providing the data used in this paper. Useful comments by Jennifer Hoeting, Jeremy York and an anonymous referee have improved the paper, but the author is solely responsible for remaining errors or confusion. A more extended version of this paper was presented at the Fifth Valencia International Meeting on Bayesian Statistics, Alicante Spain, June 5-9, 1994. This research was conducted while the author was a visitor at the Federal Reserve Bank of Minneapolis. The views expressed in this paper are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

The purpose of this paper is to propose and illustrate a new technique for an old and recurring problem, that of variable selection in linear regression. Loosely speaking, the task is to find the subsets of a prespecified set of potential covariates that best describe a dependent variable. Model selection and stepwise procedures address this problem; see Miller (1990) for a review and comprehensive bibliography of these procedures. This paper takes an explicitly subjective Bayesian view both of linear regression and the selection problem. The linear regression model is a predictive device -- its parameters are artificial, not real. As in Mitchell and Beauchamp (1988), prior distributions of parameters may be regarded as frequencies within a population of equally credible prediction experts. This explicitly includes the probability that a coefficient is zero, which is the proportion of experts who would omit the corresponding variable from the model.

This subjectivist interpretation carries with it no presumption of conjugacy in the priors. Just the opposite is true: prior information rarely establishes the link between coefficients and disturbance variance that is essential to methods exploiting conjugacy (Poirier, 1985). This paper proposes an independent prior distribution for each coefficient that is a mixture of a point mass at zero and a possibly truncated univariate normal distribution. These distributions are completely subjective, relying on no preprocessing of the data or other methods that destroy the independence of the prior information and the stochastic terms in the model. Through a series of examples, the paper illustrates that elicitation of prior distributions in this family is a natural procedure for a subjective Bayesian.

The problem of Bayesian choice of regressors dates at least to Tierney (1971). The prcedures developed here further develop model choice as described by Stewart (1987) and inequality constraints as approached in Geweke (1986). This work is most closely related to George and McCulloch (1993). It is different from their work in four respects. First, the present paper employs subjective priors, whereas George and McCulloch employ a semiautomatic approach in which the prior incorporates sufficient statistics from the regression. Second, the prior distribution includes the possibility that variables are literally excluded from the model, whereas George and McCulloch for technical reasons utilize absolutely continuous prior cumulative distribution functions. Third, this paper avoids a computational shortcut utilized by George and McCulloch that entails assuming that certain coefficients are known *a priori* to be equal to their least squares estimates. (Both papers use the Gibbs sampling algorithm to carry out the computations. To solve the technical

problems associated with a nonzero probability that coefficients are zero, a different version of the algorithm is employed here.) Finally, the paper introduces methods for more accurate assessment of very small posterior probabilities.

Prior and posterior distributions, and the computational algorithm, are outlined in the next section. Construction of prior distributions and several aspects of the posterior are illustrated in Section 3 through two of the examples used in George and McCulloch; this also affords some comparisons of the performance of the two procedures. The last section summarizes and discusses some possible extensions of this work.

2. Variable selection

This section considers the standard regression variable selection problem, with proper, informative prior distributions for all parameters. In the standard problem, k^* out of k parameters each have a nonzero coefficient with prior probability 1, while there is positive probability that any combination of coefficients of the remaining $k-k^*$ variables have coefficients equal to zero. Thus if by "model" is meant a specific combination of coefficients whose posterior probability of being nonzero is positive, there are 2^{k-k^*} alternative models entertained by the prior distribution.

Here we treat in detail a simple but frequently arising instance of the standard selection problem. First, the regression model is linear in coefficients. Second, disturbances are normally distributed. Third, in the prior distribution all parameters are mutually independent, and for $k-k^*$ of the coefficients there is positive prior probability that the coefficient is zero; even more specifically, we develop the method for coefficient prior distributions that are mixtures of normal or truncated normal distributions, and discrete mass at the point 0. These latter two assumptions may be weakened; discussion of productive directions for weakening is deferred to the final section.

2.1 Prior and posterior distributions

In standard notation

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right) \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations on a dependent variable and $\mathbf{X}$ is an $n \times k$ matrix of n corresponding observations on k covariates. The likelihood function is

$$\begin{aligned} \mathrm{L}(\beta, \sigma) &= \sigma^{-n} \exp\left[-(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)/2\sigma^2\right] \\ &= \sigma^{-n} \exp\left[-(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})/2\sigma^2\right] \exp\left[-(\beta-\mathbf{b})'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})/2\sigma^2\right] \end{aligned} \tag{2.2}$$

where $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ denotes the solution of the classic least squares problem. The k covariates include all of the regressors considered for inclusion in the model. Excluding a regressor means that the corresponding coefficient is zero in (2.1). It does not entail reducing the dimension of $\mathbf{X}$, which would render the distinction between the models meaningless as discussed by Poirier (1985, p. 712).

The investigator's prior distributions for each of the coefficients and the parameter σ are mutually independent. With prior probability $\underline{p}_i$, $\beta_i = 0$; conditional on $\beta_i \neq 0$ the prior distribution of β_i is $\mathrm{N}(\underline{\beta}_i, \tau_i^2)$, possibly truncated to the interval (λ_i, υ_i):

$$d\Pi_i(\beta_i) = \underline{p}_i d\,\mathrm{H}_i(\beta)$$
$$+(1-\underline{p}_i)(2\pi)^{-1/2}\tau_i^{-1}\left[\Phi\left(\frac{\upsilon_i-\underline{\beta}_i}{\tau_i}\right)-\Phi\left(\frac{\lambda_i-\underline{\beta}_i}{\tau_i}\right)\right]^{-1}\exp\left[\left(\beta_i-\underline{\beta}_i\right)^2/2\tau_i^2\right]\mathrm{I}_{(\lambda_i,\upsilon_i)}(\beta_i) \qquad (2.3)$$

where $\Pi_i(\cdot)$ denotes the prior c.d.f. of β_i; $\mathrm{H}(x) = 0$ if $x < 0$ and $\mathrm{H}(x) = 1$ if $x \geq 0$; $\mathrm{I}_S(x) = 1$ if $x \in S$ and $\mathrm{I}_S(x) = 0$ if $x \notin S$; $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution; $0 < \tau_i < \infty$; $-\infty \leq \lambda_i \leq \upsilon_i < \infty$; and $-\infty < \underline{\beta}_i < \infty$. The prior distribution of σ is of the standard form,

$$\underline{\nu}\underline{\sigma}^2/\sigma^2 \sim \chi^2(\underline{\nu}). \qquad (2.4)$$

The prior distribution is therefore proper and informative but nonconjugate. We choose this form because it is relatively easy to elicit one's subjective prior distribution about the coefficients in this form, yet the computational problem remains fairly simple. (Illustrations of prior construction are provided in Section 3.) The prior distribution is trivially coherent: *i.e.*, the prior distributions of nested models can be obtained as restrictions on each other.

The posterior distribution may be expressed up to a constant by combining (2.2), (2.3) and (2.4) in the usual way, but this expression is not particularly useful either for performing the computations or understanding the relation between the prior and posterior distributions. Instead we move directly to some more informative conditional distributions and the computational method based on them.

2.2 Computation

The computational procedure employed here is a Gibbs sampler with complete blocking. A value for each coefficient β_j is drawn in turn from its distribution conditional on $\beta_l\,(l \neq j)$ and σ, and a value for σ is drawn conditional on β. In the algorithm the Gibbs sampler moves from any point in the support of β and σ to any nondegenerate neighborhood of any other point in the support with positive probability in one step. Convergence of the continuous state Markov chain induced by the Gibbs sampler to the

posterior distribution may therefore be demonstrated following the argument of Tierney (1991).

The conditional distributions involved in the algorithm are simple. Given $\beta_l\,(l \neq j)$ and σ, define $z_i = y_i - \sum_{l \neq j} \beta_l x_{il}$. The conditional distribution of β_j follows from the simplified model

$$z_i = \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{IID}{\sim} \mathrm{N}(0, \sigma^2) \quad (i = 1, \mathrm{K}\ n).$$

The likelihood function kernel is

$$\exp\left[-\sum_{i=1}^{n}\left(z_i - \beta_j x_{ij}\right)^2 \Big/ 2\sigma^2\right].$$

Conditional on $\beta_j = 0$ the value of the kernel is

$$\exp\left[-\sum_{i=1}^{n} z_i^2 \Big/ 2\sigma^2\right]. \tag{2.5}$$

Conditional on $\beta_j \neq 0$ the corresponding kernel density for β_j is

$$\exp\left[-\sum_{i=1}^{n}\left(z_i - \beta_j x_{ij}\right)^2 \Big/ 2\sigma^2\right]$$
$$\cdot (2\pi)^{-1/2}\,\tau_j^{-1}\left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)\Big/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)\Big/\tau_j\right]\right]^{-1} \exp\left[-\left(\beta_j - \underline{\beta}_j\right)^2 \Big/ 2\tau_j^2\right] \mathrm{I}_{(\lambda_j, \upsilon_j)}(\beta_j)$$

$$= \exp\left[-\sum_{i=1}^{n}\left(z_i - b x_{ij}\right)^2 \Big/ 2\sigma^2\right] \exp\left[-\left(\beta_j - b\right)^2 \Big/ 2\omega^2 - \left(\beta_j - \underline{\beta}_j\right)^2 \Big/ 2\tau_j^2\right]$$
$$\cdot (2\pi)^{-1/2}\,\tau_j^{-1}\left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)\Big/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)\Big/\tau_j\right]\right]^{-1} \mathrm{I}_{(\lambda_j, \upsilon_j)}(\beta_j)$$

(where $b = \sum_{i=1}^{n} x_{ij} z_i \Big/ \sum_{i=1}^{n} x_{ij}^2$ and $\omega^2 = \sigma^2 \Big/ \sum_{i=1}^{n} x_{ij}^2$)

$$= \exp\left[-\sum_{i=1}^{n}\left(z_i - b x_{ij}\right)^2 \Big/ 2\sigma^2\right] \exp\left[-\left(\beta_j - \overline{\beta}_j\right)^2 \Big/ 2\sigma_*^2\right] \exp\left[-\left(b^2/2\omega^2 + \underline{\beta}_j^2/2\tau_j^2 - \overline{\beta}_j^2/2\sigma_*^2\right)\right]$$

$$\cdot (2\pi)^{-1/2}\,\tau_j^{-1}\left[\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)\Big/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)\Big/\tau_j\right]\right]^{-1} \mathrm{I}_{(\lambda_j, \upsilon_j)}(\beta_j) \tag{2.6}$$

(where $\sigma_*^2 = \left(\omega^{-2} + \tau_j^{-2}\right)^{-1}$ and $\overline{\beta}_j = \sigma_*^2\left(\omega^{-2} b + \tau_j^{-2} \underline{\beta}_j\right)$).

If the normal prior distribution for β_j is not truncated (*i.e.*, $\lambda_j = -\infty$, $\upsilon_j = +\infty$) then conditional on $\beta_l (l \neq j), \sigma$ and $\beta_j \neq 0$, $\beta_j \sim \mathrm{N}\left(\overline{\beta}_j, \sigma_*^2\right)$ -- the standard result for a normal prior mean when variance is known. If the normal prior distribution is truncated, then conditional distribution is $\beta_j \sim \mathrm{N}\left(\overline{\beta}_j, \sigma_*^2\right)$ truncated to the interval $\left(\lambda_j, \upsilon_j\right)$, or

$$\beta_j \sim \mathrm{TN}_{(\lambda_j, \upsilon_j)}\left(\overline{\beta}_j, \sigma_*^2\right). \tag{2.7}$$

To remove the conditioning on $\beta_j = 0$ or $\beta_j \neq 0$ it is necessary to integrate (2.6) over β_j and compare this expression to (2.5) The integration yields

$$\exp\left[-\sum_{i=1}^{n}\left(z_i - bx_{ij}\right)^2 / 2\sigma^2\right]\exp\left[-\left(b^2/2\omega^2 + \underline{\beta}_j^2/2\tau_j^2 - \overline{\beta}_j^2/2\sigma_*^2\right)\right]\left(\sigma_*/\tau_j\right)$$
$$\cdot\left\{\Phi\left[\left(\upsilon_j - \overline{\beta}_j\right)/\sigma_*\right] - \Phi\left[\left(\lambda_j - \overline{\beta}_j\right)/\sigma_*\right]\right\}\left\{\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)/\tau_j\right]\right\}^{-1}.$$

Thus the conditional Bayes factor in favor of $\beta_j \neq 0$, versus $\beta_j = 0$, is

$$BF = \exp\left[\sum_{i=1}^{n} z_i^2 - \sum_{i=1}^{n}\left(z_i - bx_{ij}\right)^2 / 2\sigma^2\right]\exp\left[-\left(b^2/2\omega^2 + \underline{\beta}_j^2/2\tau_j^2 - \overline{\beta}_j^2/2\sigma_*^2\right)\right]\left(\sigma_*/\tau_j\right)$$
$$\cdot\left\{\Phi\left[\left(\upsilon_j - \overline{\beta}_j\right)/\sigma_*\right] - \Phi\left[\left(\lambda_j - \overline{\beta}_j\right)/\sigma_*\right]\right\}\left\{\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)/\tau_j\right]\right\}^{-1} \mathrm{I}_{(\lambda_j, \upsilon_j)}\left(\beta_j\right)$$

$$= \exp\left[\overline{\beta}_j^2/2\sigma_*^2 - \underline{\beta}_j^2/2\tau_j^2\right]\left(\sigma_*/\tau_j\right)\left\{\Phi\left[\left(\upsilon_j - \overline{\beta}_j\right)/\sigma_*\right] - \Phi\left[\left(\lambda_j - \overline{\beta}_j\right)/\sigma_*\right]\right\}$$
$$\cdot\left\{\Phi\left[\left(\upsilon_j - \underline{\beta}_j\right)/\tau_j\right] - \Phi\left[\left(\lambda_j - \underline{\beta}_j\right)/\tau_j\right]\right\}^{-1}. \qquad (2.8)$$

To draw β_j from its conditional distribution the conditional posterior probability that $\beta_j = 0$ is computed from the conditional Bayes factor (2.8):

$$\overline{p}_j = \frac{\underline{p}_j}{\underline{p}_j + \left(1 - \underline{p}_j\right)BF}. \qquad (2.9)$$

Based on a comparison of this probability with a drawing from the uniform distribution on [0, 1], the choice $\beta_j = 0$ or $\beta_j \neq 0$ is made. If $\beta_j \neq 0$ then β_j is drawn from (2.7).

Conditional on all β_j,

$$\left[\underline{\upsilon}\underline{\sigma}^2 + \left(\mathbf{y} - \mathbf{X}\beta\right)'\left(\mathbf{y} - \mathbf{X}\beta\right)\right]/\sigma^2 \sim \chi^2\left(\underline{\nu} + n\right).$$

The Gibbs sampling computational algorithm proceeds in the usual way. After an initial value for (β, σ) is drawn from the prior distribution, the parameters $\beta_1, \beta_2, \mathrm{K}, \beta_k, \sigma$ are drawn in succession from their respective conditional posterior distributions. In most applications a key objective is determination of the posterior probability of each of the $2^{(k-k^*)}$ models. This could be done in the obvious way, by recording an indicator variable for the model corresponding to the non-zero β_j at the end of each iteration. More accurate approximations may be based on (2.9), however. In each *step* of each iteration, record the value $\overline{p}_j$ for the model corresponding to $\beta_j = 0$ and $\beta_l (l \neq j)$ either zero or non-zero as is

the case in the conditional distribution for β_j. Similarly record the value $\left(1-\bar{p}_j\right)$ for the model corresponding to $\beta_j \neq 0$ and the values of $\beta_l (l \neq j)$. In similar fashion, posterior expectations of functions of interest of the parameter vector (β, σ) may be approximated more efficiently by drawing a value for $\beta_j \neq 0$ whether β_j is set to zero in the Markov chain or not, and weighting by the probability $\bar{p}_j$ for the function with $\beta_j = 0$ and by $\left(1-\bar{p}_j\right)$ for the function with $\beta_j \neq 0$. (One could pursue this strategy for additional steps to achieve even more accurate assessment of small model probabilities, but the number of computations required increases exponentially with the number of steps.)

2.3 Computational efficiency

Since the Gibbs sampling algorithm described here employs complete blocking, the degree of serial correlation in the Monte Carlo Markov chain generated by the Gibbs sampler will depend on the degree of multicollinearity in the correlation matrix of the regressors (Geweke, 1992). If all sample correlation coefficients were zero, then the draw from the Gibbs sampler would be serially uncorrelated. However such a situation would be exceptional. Indeed, the most interesting and difficult cases -- the ones for which proceeding to a formal analysis of the kind described here is most compelling -- are precisely those in which there is a high degree of collinearity among regressors. Experience with the algorithm indicates that the higher the ratio of the largest to smallest eigenvalues of the sample correlation matrix of the regressors, the more iterations will be required to achieve the same degree of numerical accuracy. For small regression problems the algorithm is fast: using code for which little optimization has been undertaken, 10^6 iterations for a 5-regressor model requires about 40 seconds on a Sun 10/51 with untruncated normal priors and about 75 seconds with truncated normal priors. Execution time appears to be roughly proportional to the cube of the number of regressors, so computation time for larger models can be much longer.

3. Examples

We now take up two specific examples of noncontingent variable selection in regression. The examples are also considered in George and McCulloch (1993). The objectives in these examples are to demonstrate a convenient method for the formulation of subjective priors, illustrate the numerical accuracy of the procedure, and study the relation between prior and posterior distributions in this model.

3.1 The happiness data

These data were collected from 39 employed MBA students in a class at the University of Chicago Graduate School of Business. Five variables were recorded: y_i = Happiness, recorded on a 10-point scale with 1 representing a suicidal state, 5 a feeling of "just muddling along" and 10 a euphoric state; x_{i1} = Money, measured by family income in thousands of dollars; x_{i2} = Sex, measured by 0 or 1 with 1 being a satisfactory level of sexual activity; x_{i3} = Love, with 1 indicating loneliness and isolation, 2 a set of secure relationships, and 3 a deep feeling of belonging and caring in a family or community; x_{i4} = Work, recorded on a 5-point scale with 1 indicating that the individual is seeking other employment, 3 that the individual's job is "OK", and 5 indicating that the job is enjoyable. The linear regression model is

$$y_i = \beta_1 + \sum_{j=1}^{4} \beta_{j+1} x_{ij} + \varepsilon_i,$$

a specific instance of (2.1).

For this example, consider specification of a prior distribution that reflects the belief that each regressor may be a substantively significant determinant of Happiness, or it may not enter the model at all. We assume that each regressor coefficient is nonnegative, and employ a half-normal prior $\left(\lambda_j = 0, \upsilon_j = +\infty\right)$ with $\underline{\beta}_j = 0$. We interpret "substantially significant determinant" to mean that a major change in the regressor in question, Δx_{ij}, ought to bring about a major change in the dependent variable, Δy_i. We then set the parameter τ_j of the half-normal distributions, $\tau_j = \tau_j^* = \Delta y_i / \Delta x_{ij}$. For the results here major change in Happiness was set at $\Delta y_i = 4$; in Money at $\Delta x_{i1} = 50$; in Sex at $\Delta x_{i2} = .5$; in Love at $\Delta x_{i3} = 1$; and in Work at $\Delta x_{i4} = 2$. The mapping from "substantially significant determinant" to the τ_j is itself subjective, and we present results for $\tau_j = .5\tau_j^*$ and $\tau_j = 2\tau_j^*$ as well as for $\tau_j = \tau_j^*$. For the intercept term we choose a full normal prior with $\underline{\beta}_1 = 0$ and $\tau_1 = 9$, reflecting uncertainty about the inclusion of various combinations of x_{ij} in the model. For the standard deviation of ε_i, $\underline{\sigma} = 2.5$, based on a prior mean of 5 for the standard deviation of y_i and a prior mean of .75 for the multiple correlation coefficient; $\underline{\nu} = .01$.

Different prior beliefs will, of course, lead to other choices for the τ_j. For instance, if it is thought that a certain regressor may either be a substantially insignificant determinant of Happiness or it may not enter the model at all, then the corresponding τ_j would be smaller and its value may be set employing the same kind of reasoning about marginal effects. As in George and McCulloch (1993, p. 883) the idea is to support β_j that are different from 0, but not so large as to dilute support for realistic values with support for unrealistically large

values. The scaling procedures of Mitchell and Beauchamp (1988) accomplish much the same objective. The procedure followed here is that suggested by Berger (1988).

For illustrative purposes we take as a base prior probability that each variable is excluded from the model, $\underline{p}_j = .5 \; (j = 2,\mathsf{K}\;,5)$; the intercept β_1 is always included ($\underline{p}_1 = 0$). To study the relation between the prior and posterior distributions, we also consider $\underline{p}_j = .2$ $(j = 2,\mathsf{K}\;,5)$ and $\underline{p}_j = .8 \; (j = 2,\mathsf{K}\;5)$, always maintaining $\underline{p}_1 = 0$.

Summaries of the happiness data and prior distribution are provided in Table 1. Collinearity among regressors is modest. Posterior probabilities of alternative models are presented in Table 2. These results are obtained using the methods described in Section 2, with $m = 10^5$ iterations of the Gibbs sampling algorithm. The numerical accuracy of these results was assessed in two ways. The first uses the numerical standard error and relative numerical efficiency discussed in Geweke (1992) for the Gibbs sampling algorithm. Relative numerical efficiency, in turn, may be expressed in terms of an i.i.d.-equivalent number of iterations, m^*: that is, the numerical accuracy achieved for m iterations of the Gibbs sampling algorithm is the same as that for m^* hypothetical i.i.d. drawings from the posterior. Thus, for a posterior probability $\overline{p}_j$ presented in Table 2 the numerical standard error is $\left[\overline{p}_j\left(1-\overline{p}_j\right)\big/m^*\right]^{1/2}$. In Table 2 m^* ranges from 1,250 to 25,000 for the $.5\tau_j^*$ priors; from 960 to 24,000 for the τ_j^* priors; and from 570 to 24,000 for the $2\tau_j^*$ priors. The lower bound on numerical accuracy is higher for small values of the τ_j, because the smaller values reduce the collinearity in the posterior precision matrix for β, thereby reducing serial correlation in the Gibbs sampler and increasing computational efficiency.

The second method for assessing numerical accuracy is based on the observation that for the same values of the prior standard deviations τ_j, changing common values of the $\underline{p}_j$ $(j = 2,\mathsf{K}\;5)$ from one common value to another will not affect the relative posterior probability of those models with the same number of regressors. For example, the posterior probability of all 2-variable models conditional on the model containing two variables and $\tau_j = \tau_j^*$ should be the same whether $\underline{p}_j = .2$, $\underline{p}_j = .5$ or $\underline{p}_j = .8$. These conditional probabilities -- simply the ratio of entries for specific models to the entry in the "2 regressors" row of the same columns of Table 2 -- are indeed equal to within the tolerance indicated by the i.i.d.-equivalent number of iterations m^*.

Regardless of the particular prior distribution the models with regressors Love alone (x_{i3}), Love and Work (x_{i3} and x_{i4}) or Love, Work and Money (x_{i3}, x_{i4}, and x_{i1}) have total posterior probability at least two-thirds and often much more. Three systematic effects of the prior distributions on the posterior probabilities of the alternative models are evident. First, increases in $\underline{p}_j$, the prior probability that $\beta_j = 0$, favor smaller models. Second,

increases in τ_j also favor smaller models. This is due to the fact that the values of the τ_j are all fairly large compared to the mass of the likelihood function (compare the τ_j^* and least squares coefficients in Table 1). As the τ_j increase over this range, the Bayes factors corresponding to models with large numbers of regressors decrease relative to those for models with smaller numbers of regressors, and the magnitude of the prior density decreases in the region of the mass of the likelihood function. In the limit, as all $\tau_j \to \infty$, all posterior probability becomes concentrated on the model with no regressors, consistent with Lindley's paradox (Bartlett, 1957; Lindley, 1957). For these data and the range of τ_j's employed the effect is to move the model probability from the Love-Work-Money model to the Love-Work model to the Love model.

Our results are only indirectly comparable with those of George and McCulloch (1993). Even the likelihood function is not the same, since (as discussed in the introduction) they do not account for uncertainty about the intercept. However, results are broadly consistent. Both their results and ours favor the Love model among all one-variable models, the Love-Work model among all two-variable models, and the Love-Work-Money model among all three-variable models. The approach taken here has some tendency to favor smaller models than does the George-McCulloch approach, but differences are not great.

3.2 The Hald data

These data are presented in Draper and Smith (1981) and are often used to illustrate techniques for selective regressors. The five variables are y_i = Heat produced in the hardening of cement (in calories per gram), x_{i1} = percentage of input composed of tricalcium aluminate, x_{i2} = percentage of input composed of tricalcium silicate, x_{i3} =percentage of input composed of tetracalcium alumino ferrite, and x_{i4} = percentage of input composed of dicalcium silicate. There are only 13 observations in the data set.

For this example full normal priors centered at $\underline{\beta}_j = 0$ were loyed for all coefficients. The parameters τ_j^* were constructed in the same way as in the previous example, taking $\Delta y_i = 20°$ as a major change in the dependent variable. Major changes in regressor variables were set to one-half their range in the data set: $\Delta x_{i1} = 10$, $\Delta x_{i2} = 22.5$, $\Delta x_{i3} = 8.5$, and $\Delta x_{i4} = 27$. Alternative values of the τ_j were set accordingly and once again either $\underline{p}_j = .2$ or $\underline{p}_j = .5$ or $\underline{p}_j = .8$ for all $j = 2, \ldots, 5$. A summary of the τ_j^* and of the data is provided in Table 3. Collinearity in the regressors is quite high, with the ratio of largest to smallest eigenvalues of the correlation matrix exceeding 10^3.

Posterior probabilities for alternative models are presented in Table 4 in the same format as were the results for the Happiness data in Table 2. In view of the ill-conditioned data, computations employed 10^6 iterations of the Gibbs sampling algorithm, 10 times more than for the Happiness data. The number of i.i.d.-equivalent iterations m^*, for purposes of judging numerical accuracy, ranged from $m^* = 2,000$ to $m^* = 15,000$ for priors with $\tau_j = .5\tau_j^*$, from $m^* = 580$ to $m^* = 13,000$ for priors with $\tau_j = \tau_j^*$, and from $m^* = 30$ to $m^* = 6,000$ when $\tau_j = 2\tau_j^*$. The explanation for the effect of τ_j^* on m^* is the same as for the happiness data, but the magnitude is larger because of the greater collinearity of the covariates.

Overall the favored models incorporate x_{i1} and x_{i2}; x_{i1}, x_{i2} and x_{i3}; x_{i1}, x_{i2}, and x_{i4}; or all four regressors. Between them these four models always account for at least 80% of the posterior probability, and in some cases close to 100%. The qualitative effects of changing the $\underline{p}_j$ or the τ_j are the same as in the Happiness data for the same reasons. However the sensitivity of the posterior model probabilities to changes in the prior is much greater, as one would expect both from the conditioning of the regressor moment matrix and the small sample size.

George and McCulloch also find posterior model probabilities sensitive to their prior distribution, but in most other respects their results differ from ours. In general, their methods produce models with small numbers of regressors, producing probability .44 of no regressors in one case, and never producing a probability greater than .02 for the model with all four regressors. Their most probable two-variable model is the same as ours. They are unable to discriminate among the three-regressor models (x_{i1}, x_{i2}, x_{i3}), (x_{i1}, x_{i2}, x_{i4}) and (x_{i1}, x_{i2}, x_{i4}), whereas the results in Table 4place relatively less posterior probability on (x_{i1}, x_{i2}, x_{i4}).

4. Summary and Extensions

This paper has proposed a family of nonconjugate priors for subjective Bayesian treatment of variable selection and model comparison in linear regression. Since priors for different coefficients are independent the investigator can consider one coefficient at a time. This investigator is, however, forced to think explicitly about plausible magnitudes for each coefficient conditional on the corresponding regressor appearing in the model. Sign restrictions and other limitations on support are easily accommodated.

In experiments with two models having five regressors each, it was found that priors interact with data in an understandable way: *e.g.*, increased probability of variable exclusion leads to smaller models, as do increasingly diffuse priors for coefficients of included

variables. The computational efficiency of the Gibbs sampling computational algorithm is largely a function of collinearity in the posterior distribution of the coefficients: *e.g.*, the more ill-conditioned the covariate sample correlation matrix the less efficient the algorithm; the greater the prior precision of coefficients of included variable the more efficient the algorithm.

In all the examples considered the posterior distribution changes in important ways in response to reasonable changes in the prior distribution. This fact underscores the importance of choosing prior distributions carefully, and should make subjective Bayesians even more wary of "automatic" procedures that seek to avoid explicit specification of priors. Only in small models with many observations -- many more than employed in the examples taken up here -- will the posterior be robust to reasonable changes in the prior. But given many observations investigators generally enlarge the number of models considered, thus perpetuating sensitivity to the prior distribution.

Several extensions to these developments are natural and would involve no problems beyond normal technical difficulties in implementation. Following West (1984), Geweke (1993), Diebolt and Robert (1994), and others, the assumption that disturbances are normal may be weakened through appropriate use of mixture models. Nor is the procedure limited to truncated normal prior distributions for coefficients of included variables: since the Gibbs sampling algorithm is fully blocked essentially arbitrary prior distributions may be specified for each coefficient with no serious technical impediment. Extension of the methods proposed here to multivariate regression models in general and vector autoregressions in particular is also straightforward. Second, the procedures developed here can be extended readily to the problem of contingent variable selectrion in which a regresor enters the model only if other regressors also enter. This includes the choice of one model from an ordered or semi-ordered sequence of models, for example choosing the length(s) of distributed lag(s) in time series regression. Finally, through modest modification of the fully blocked Gibbs sampling algorithm introduced in Geweke (1994) one could adapt the algorithm of this paper to variable selection and sign restrictions in essentially any model for which the posterior distribution can be expressed in closed form; whether the algorithm would still be fast enough to be practical is an open question.

References

Bartlett, M.S., 1957, "A Comment on D.V. Lindley's Statistical Paradox," *Biometrika* **44**: 533-534.

Berger, J.O. 1988, "Comment" (on Mitchell and Beauchamp, 1988), *Journal of the American Statistical Association* **83**: 1033-1034.

Diebolt, J., and C.P. Robert, 1994, "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society Series B* **56**: 363-376.

Draper, N., and H. Smith, 1981, *Applied Regression Analysis* (Second edition). New York: John Wiley.

George, E.I. and R.E. McCulloch, 1993, "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association* **88**: 881-889.

Geweke, J., 1986, "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics* **1**: 127-142.

Geweke, J., 1992, "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.), *Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics*, 169-194. Oxford: Oxford University Press, 1992.

Geweke, J., 1993, "Bayesian Treatment of the Student-*t* Linear Model," *Journal of Applied Econometrics* **8**: S19-S40.

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

Lindley, D.V., 1957, "A Statistical Paradox," *Biometrika* **44**: 187-192.

Madigan, D. and A.E. Raftery, 1994, "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, forthcoming.

Miller, A.J., 1990, *Subset Selection in Regression.* New York: Chapman and Hall.

Mitchell, T.J. and J.J. Beauchamp, 1988, "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association* **83**: 1023-1032.

Poirier, D.J., 1985, "Bayesian Hypothesis Testing in Linear Models with Continuously Induced Conjugate Priors Across Hypotheses," in J.M. Bernardo et al. (eds.), *Bayesian Statistics 2*, pp. 711-722. Amsterdam: North-Holland.

Raftery, A., D. Madigan and . Hoeting, 1993, "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," University of Washington Department of Statistics Technical Report No. 262.

Stewart, L., 1987, "Hierarchical Bayesian Analysis using Monte Carlo Integration: Computing Posterior Distributions when there are Many Possible Models," *The Statistician* **36**: 211-219.

Tierney, L., 1991. "Markov Chains for Exploring Posterior Distributions." University of Minnesota School of Statistics Technical Report No. 560.

West, M., 1984. "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society Series B* **46**: 431-439.

Table 1

Happiness data

Variable	Definition	Least squares coefficient	Least squares standard error	Prior standard deviation τ_j^*
1	Money	.00958	.00521	.08
2	Sex	-.149	.419	8.00
3	Love	1.92	.295	4.00
4	Work	.476	.199	2.00

Covariate sample correlation matrix,

1.000	.307	.126	.068
	1.000	.047	-.316
		1.000	.386
			1.000

Eigenvalues of covariate sample correlation matrix

.4405	.7356	1.3468	1.477

Table 2

Model posterior probabilities, Happiness data with half-normal priors

$\underline{p}_j = P(\beta_j = 0)$	---------- 0.2 ----------			---------- 0.5 ----------			---------- 0.8 ----------		
τ_j	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$
0 regressors									
1									
2									
3	.0025	.0098	.0317	.0271	.0893	.1867	.1763	.3394	.5368
4									
1 regressor	.0025	.0098	.0317	.0271	.0893	.1867	.1763	.3394	.5368
1,2									
1,3	.0013	.0247	.0352	.0315	.0552	.0574	.0554	.0455	.0377
1,4									
2,3	.0007	.0013	.0021	.0018	.0027	.0033	.0033	.0027	.0021
2,4									
3,4	.1230	.2561	.4318	.4130	.5556	.6212	.6309	.5693	.4296
2 regressors	.1367	.2561	.4138	.4130	.5556	.6212	.6309	.5693	.4296
1,2,3	.0024	.0024	.0015	.0014	.0013	.0007	.0007	.0003	.0001
1,3,4	.6079	.5815	.4691	.4709	.3109	.1706	.1708	.0812	.0302
1,2,4									
2,3,4	.0617	.0591	.0048	.0494	.0309	.0173	.0178	.0089	.0032
3 regressors	.6721	.6431	.5175	.5217	.3431	.1886	.1892	.0905	.0334
1,2,3,4	.1887	.0911	.0370	.0381	.0120	.0034	.0036	.0008	.0001

All models have strictly positive posterior probability. Empty cells indicate posterior probability less than 10^{-4}.

Table 3

Hald data

Variable	Definition	Least squares coefficient	Least squares standard error	Prior standard deviation τ_j^*
1	% Tricalcium aluminate	1.55	.745	2.0
2	% Tricalcium silicate	.510	.724	.89
3	% Tetracalcium alumino ferrite	.102	.755	2.1
4	% dicalcium silicate	-.144	.709	.74

Covariate sample correlation matrix

1.000	.229	-.824	-.245
	1.000	-.139	-.973
		1.000	.030
			1.000

Eigenvalues of covariate sample correlation matrix

.001624	.1866	1.576	2.236

Table 4

Model posterior probabilities, Hald data with full normal priors

$\underline{p}_j = P(\beta_j = 0)$	---------- 0.2 ----------			---------- 0.5 ----------			---------- 0.8 ----------		
τ_j	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$	$.5\tau_j^*$	τ_j^*	$2\tau_j^*$
0 regressors									
1									
2									
3									
4									
1 regressor									.0001
1,2	.0936	.2018	.2882	.4468	.5881	.5834	.8007	.8589	.7469
1,3									
1,4		.0043	.0497		.0146	.1317	.0001	.0084	.1660
2,3									
2,4									
3,4			.0004			.0009			.0008
2 regressors	.0936	.2061	.3383	.4468	.6028	.7161	.8008	.8673	.9137
1,2,3	.1969	.1873	.1301	.2306	.1359	.0062	.1079	.0505	.0212
1,3,4	.0001	.0065	.0052		.0057	.0381	.	.0008	.0119
1,2,4	.1504	.2748	.3097	.1746	.1967	.1605	.0749	.0755	.0514
2,3,4			.0004			.0001			
3 regressors	.3474	.4687	.4954	.4052	.3383	.2648	.1828	.1268	.0845
1,2,3,4	.5590	.3252	.1664	.1480	.0589	.0191	.0165	.0059	.0017

All models have strictly positive posterior probability. Empty cells indicate posterior probability less than 10^{-4}.