Hacettepe Journal of Mathematics and Statistics
Volume 31 (2002), 63–76

BAYESIAN VARIABLE SELECTION IN LINEAR REGRESSION AND A COMPARISON

Atilla Yardımcı* and Aydın Erar†

Received 17. 04. 2002

Abstract

In this study, Bayesian approaches, such as Zellner, Occam's Window and Gibbs sampling, have been compared in terms of selecting the correct subset for the variable selection in a linear regression model. The aim of this comparison is to analyze Bayesian variable selection and the behavior of classical criteria by taking into consideration the different values of β and σ and prior expected levels.

Key Words: Bayesian variable selection, Prior distribution, Gibbs Sampling, Markov Chain Monte Carlo.

1. Introduction

A multiple linear regression equation with n observations and k independent variables can be defined as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $Y_{n\times 1}$ is a response vector; $X_{n\times k'}$ is a non-stochastic input matrix with rank k', where $k' = k + 1$; $\boldsymbol{\beta}_{k'\times 1}$ is an unknown vector of coefficients; $\epsilon_{n\times 1}$ is an error vector with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{E}(\epsilon\epsilon') = \sigma^2\mathbf{I}_n$. Estimation and prediction equations can be defined as $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$ and $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\mathbf{e}$ is the residual vector of the estimated model, and $\hat{\boldsymbol{\beta}}$ is the estimator of the model parameter obtained using the Least Square Error (LSE) method or the Maximum Likelihood (ML) method under the assumption that ϵ is normally distributed. Thus the joint distribution function of the $\mathbf{Y}$'s is obtained by maximizing the likelihood function as follows:

$$\mathrm{L}(\boldsymbol{\beta}, \sigma/\mathbf{Y}) = f(\mathbf{Y}/\boldsymbol{\beta}, \sigma, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right)$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2}$$

$$\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n - k') = \mathrm{SSE}(\hat{\boldsymbol{\beta}})/(n - k') \tag{3}$$

*Başkent University, Faculty of Science and Letters, Department of Statistics and Computer science, Bağlıca, Ankara.

†Hacettepe University, Faculty of Science, Department of Statistics, Beytepe, Ankara.

where $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ and $\hat{\sigma}^2$ is corrected such that $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$ [4]. Variable selection in regression analysis can be used to make estimation with minimum cost, less MSE, leaving out the non-informative independent variables from the model and low standard errors in the presence of variables with a high degree of collinearity.

2. Classical Variable Selection Criteria

The stepwise and all possible subset methods for variable selection are often used in a linear regression model. The all possible subset method is used in relation with mean square error, coefficient of determination, F statistics, Cp and PRESSp [4]. In recent years, the AIC, AICc and BIC criteria have also been used in variable selection procedures having respect to the amount of information. If p is the number of parameters in the subset, AIC and AICc aim to select the subset which minimizes the values of AIC and AICc obtained by the equations

$$\mathrm{AIC} = n\log(\mathrm{AKO} + 1) + 2p \tag{4}$$

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2(p+1)(p+2)}{n-p-2}. \tag{5}$$

If $f(\mathbf{Y}/\hat{\boldsymbol{\beta}}_p, \sigma_p, \mathbf{X})$ is the likelihood function for the subset p, the criterion BIC aims to select the subset which maximizes the equation given below [2]

$$\mathrm{BIC} = \log f(\mathbf{Y}/\hat{\boldsymbol{\beta}}_p, \sigma_p, \mathbf{X}) - \frac{p}{2}\log n.$$

Posterior model probabilities and risk criteria have been also been used occasionally in variable selection. Posterior model probabilities have been assessed for all the subsets using a logarithmic approach given by Wassermann [16] under the assumption of equal prior subset probabilities. The risk criterion has been adapted to variable selection with the help of the equation

$$\mathrm{R}(\hat{\boldsymbol{\beta}}) = \sigma^2\mathrm{tr}(\mathrm{V}(\hat{\boldsymbol{\beta}})) \tag{6}$$

given by Toutenburg [15], where $\mathrm{V}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. The rate (6) found for all the probable subsets corresponds to the risk of the estimations. The aim is to get the model with the smallest risk.

On the other hand, Yardımcı [17] claims that the rates of risk and posterior probability should be evaluated together during the interpretation. Ideally, the aim is to find a subset that has high posterior probability and low risk. Since it is hard to satisfy this condition, preferences are made within acceptable limits [17]. Furthermore, the Risk Inflation Criterion (RIC) has been found for all probable subsets with the help of the equation defined by Foster and George [5] given below

$$\mathrm{RIC} = \mathrm{SSE}_p + 2p\mathrm{MSE}_p\log(k), \tag{7}$$

where SSEp and MSEp are the sum of squares error and mean square error for the subset p, respectively. It is claimed that RIC will produce the lowest values and thus give more consistent deductions when trying to select the correct variables [17].

3. Bayesian Variable Selection Approaches

Bayesian variable selection approaches make use of hypothesis testing and stochastic searching methods in linear regression. Apart from these approaches, there are various methods based on the selection of the subset where the probability of the posterior subset is the highest.

3.1. Posterior Odds and the Zellner Approach

Bayesian approaches to hypothesis testing are quite different from classical probability ratio tests. The difference comes from the use of prior probabilities and distributions. The researcher has to define prior probabilities $P(H_0)$ and $P(H_1)$ for the correctness of the hypothesis. Here H_0 symbolizes the null hypothesis and H_1 stands for the alternative hypothesis. To choose one of the hypotheses, posterior probabilities have to be assessed after defining the prior probabilities. The Bayes Factor of H_1 against H_0 has been given by Lee [11] as follows:

$$\mathrm{BF} = \frac{p(H_0/y)}{p(H_1/y)} \frac{P(H_1)}{P(H_2)}. \tag{8}$$

Thus, the Bayes factor is the ratio of H_0's posterior odds to its prior odds. If the prior probabilities that verify the hypotheses H_0 and H_1 are equal, that is $P(H_0) = P(H_1) = 0.5$, then Bayes factor is equal to the posterior odds of H_0. Although the interpretation of Bayesian hypothesis testing depends on the nature of the research, Kass and Raftery [9] tables are often used. In the Zellner approach, k independent variables in the model are divided into two sets. The variables in the first set are those that are needed in the model and are shown by $\mathbf{X}_1$ and $\boldsymbol{\beta}_1$. The variables in the second set are those that should be removed from the model and are shown by $\mathbf{X}_2$ and $\boldsymbol{\beta}_2$. Here, $\mathbf{X}_1$ and $\mathbf{X}_2 = \mathbf{V}$ are $n \times k_1$ and $n \times k_2$ matrices respectively, where $k = k1 + k2$. The vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have dimensions $k_1 \times 1$ and $k_2 \times 1$. $\mathbf{I}$ denotes the unit vector. The model can be rewritten as follows:

$$\mathbf{Y} = \alpha\mathbf{I} + \mathbf{X}_1\eta + \mathbf{V}\boldsymbol{\beta}_2 + \epsilon,$$

where $\mathbf{V} = \left(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\right)\mathbf{X}_2$, $\eta = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$ and $\mathbf{X}_1'\mathbf{V} = \mathbf{0}$ [19]. The hypotheses that will be tested in the case of the division of variables into two sets may be written as $H_0 : \boldsymbol{\beta}_2 = 0$ and $H_1 : \boldsymbol{\beta}_2 \neq 0$.

In conclusion, when the prior probabilities for the rejection or acceptance of the hypothesis are defined to be equal and the prior beliefs for the parameters are defined as noninformative, then the odds ratio B_{01} is given by Moulton [13] as follows:

$$\begin{aligned} \mathrm{BF}_{01} &= b(v_1/2)^{k_2/2}(v_1 s_1^2/v_0 s_1^2/v - 0s_0^2)^{(v_1-1)/2} \\ &= b(v_1/2)^{k_2/2}\left[1 + (k_2/v_1)F_{k_2,v_1}\right]^{(v_1-1)/2} \end{aligned} \tag{9}$$

and

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}, \quad \hat{\eta} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y},$$

$$
\begin{aligned}
v_0 s_0^2 &= (y - \bar{y}\mathbf{I} - \mathbf{X}_1\hat{\eta}),\ v_0 = n - k_1 - 1,\\
\hat{\boldsymbol{\beta}}_2 &= (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{Y},\\
v_1 s_1^2 &= (\mathbf{Y} - \bar{y}\mathbf{I} - \mathbf{X}_1\hat{\eta} - \mathbf{V}\hat{\boldsymbol{\beta}}_2)'(\mathbf{Y} - \bar{y}\mathbf{I} - \mathbf{X}_1\hat{\eta} - \mathbf{V}\hat{\boldsymbol{\beta}}_2),\ v_1 = n - k_1 - k_2 - 1,\\
b &= \sqrt{\pi}/\Gamma[(k_2+1)/2],
\end{aligned}
$$

where $F_h = F_{k_2,v_1} = \hat{\boldsymbol{\beta}}_2'\mathbf{V}'\mathbf{V}\hat{\boldsymbol{\beta}}_2/k_2 s_1^2$ is the F-test statistics. Thus the posterior odds ratio given by equation (9) can be used as the Bayesian test for the variables to be included in the model and those to be removed from the model [19].

3.2. Occam's Window Approach

In this approach, if a model is not as strong as the others in terms of their posterior odds, then the model and its sub models are eliminated from consideration. If the posterior probabilities of two models are equal, the simple one is accepted (according to parsimony rule). Hoeting [8] has claimed that if the aim is to interpret the relationship between the variables and to make quick assessments, then Occam Window (OW) approach is useful. If $q = 2^k$ is the number of all possible models, then $M = M1, \ldots, Mq$ will show all possible subsets. The posterior probability of the subset M_m is obtained as

$$p(Y/M_n) = \int p(Y/\boldsymbol{\beta}, M_m)p(\boldsymbol{\beta}/M_m)\, d\boldsymbol{\beta}. \tag{10}$$

Here, $p(Y/\boldsymbol{\beta}, M_m)$ is the marginal likelihood of the subset M_m; $\boldsymbol{\beta}$ is the parameter vector; $p(b/M_m)$ is the prior probability of $\boldsymbol{\beta}$ for the subset M_m. The equation (10) can be rewritten as

$$p(\Delta/Y, H_m) = \int p(\Delta/Y, \boldsymbol{\beta}_m, H_m)p(\boldsymbol{\beta}_m/Y, H_m)\, d\boldsymbol{\beta}_m.$$

with the help of the Bayes theorem [9]. Here $\boldsymbol{\beta}_m$ is the vector of independent variables for the subset M_m. The OW approach has two main characteristics: if the subset estimation is weaker than the subset under consideration, then it is eliminated [12]. Thus, the eliminated subset is defined by Hoeting [8] as follows:

$$A' = \left\{ M_m; \frac{\max(p(M_i/Y))}{p(M_m/Y)} > O_L \right\}.$$

Here, O_L is determined by the researcher. The second characteristic is that it determines the alternative subset with the support of data. For this searching,

$$B = \left\{ M_m; \exists\, M_i \in A,\ M_i \subseteq M_q,\ \frac{p(M_i/Y)}{p(M_m/Y)} > O_R \right\}$$

is used [14]. Note that O_R is also determined by the researcher. Generally, O_R is taken to be 1 [8]. Also, Δ denotes a situation which can be regarded as a future observation or a loss of an event. It will be written as [8]

$$p(\Delta/Y) = \frac{\sum_{M_m \in A} p(\Delta/M_m' Y)p(Y/M_m)p(M_m)}{\sum_{M_m \in A} p(Y/M+m)p(M_m)}.$$

In the OW approach, the rates O_L and O_R are determined by subjective belief. For example, Madigan and Raftery [12] accepted $O_L = 1/20$ and $O_R = 1$. In an application of this approach, model selection is done in two stages: First a logical model from all the $2^k - 1$ possible models is chosen. Then "up or down algorithms" are used to add or remove variables from the models [8].

3.3. The Gibbs Sampling Approach

Bayesian approaches concentrate on the determination of posterior probabilities or distributions. However, in some cases, the analytic solution of the integrals needed for the assessment of the posterior moments are impossible or hard to achieve. In such cases, approaches involving the creation of Markov chains and the use of convergence characteristics to obtain the posterior distributions are used [17]. Thus, by using Markov Chain and stochastic simulation techniques together, Markov Chain Monte Carlo (MCMC) approaches have been developed.

The aim of the MCMC approach is to create a random walk which converges to the target posterior distributions [1]. In the MCMC approach, a sequence θ_j converging to the posterior distribution is produced in which θ^{j+1} depends on the previous rate θ^j. With this in mind, sampling is done using the Markov transition distribution $g(\theta/\theta^j)$ [7]. For a given posterior distribution, there are various methods of taking samples from the transition distribution and processing this information. Metropolis algorithms and Gibbs sampling approaches are used to show how these samples taken from a posterior distribution achieve the characteristic of a Markov chain.

In Gibbs sampling, initially starting points $(X_1^0, X_2^0, \ldots, X_k^0)'$ are determined when $j = 0$, and then points $\mathbf{X}^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_k^{(j)})'$ are produced in the state $x^{(j-1)}$. Switching to the state $j = j + 1$, the process continues until convergence is obtained [6]. When convergence is achieved the points $x^{(j)}$ correspond to the rates $p(x_1, \ldots, x_k/y)$. There are various Bayesian variable selection approaches that depend on Gibbs sampling. Here the approach of Kuo and Mallick [10], that is used in this study, will be explained.

In Gibbs sampling, a dummy variable given by $\gamma = (\gamma_1, \ldots, \gamma_k)'$ is used to define the position of the 2^{k-1} independent variable subsets in the model. If the variable considered is in the model, then $\gamma_j = 1$, if not $\gamma_j = 0$. Since the γ rate is not known, it has to be added to the Bayes process. Thus, the joint prior that will be used in the variable selection process is defined as

$$p(\boldsymbol{\beta}, \sigma, \gamma) = p(\boldsymbol{\beta}/\sigma, \gamma)p(\sigma/\gamma)p(\gamma).$$

Here, the prior defined by Kuo and Mallick [10] for the β coefficients is assumed to be

$$\boldsymbol{\beta}/\sigma, \quad \gamma \sim \mathrm{N}(\boldsymbol{\beta}_0, D_0 I).$$

In addition, a noninformative prior for σ is defined by

$$p(\sigma) \propto \frac{1}{\sigma}.$$

The prior distribution of σ is obtained with the help of the inverted gamma distribution and the assumptions $\alpha \to 0$ and $\eta \to 0$. The noninformative prior for γ is given as

$$\gamma_j \sim \text{Bernoulli}(1, p).$$

The deductions concentrate on these distributions since the marginal posterior distribution $p(\gamma/\mathbf{Y})$ contains all the information that is needed for variable selection. In Kuo and Mallick [10], it has been shown that $p(\gamma/\mathbf{Y})$ can be estimated using Gibbs sampling. Moreover, the least square estimation for the full model can be used when the starting values of $\boldsymbol{\beta}^0$ and σ^0, that are needed for this process, are $\gamma_j = 1$, $(j = 1, \ldots, k)$. Initially we may take $\gamma^0 = (1, 1, \ldots, 1)'$. In this way, the marginal posterior density of γ_j can be given as

$$p(\gamma_j/\gamma_{-j}, \boldsymbol{\beta}, \sigma, \mathbf{Y}) = \text{B}(1, \tilde{p_j}),\ j = 1, 2, \ldots, k,$$

where, if $\gamma_{-j} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_k)$, then

$$\begin{aligned}
\tilde{p_j} &= \frac{c_j}{(c_j + d_j)}, \\
c_j &= p_j \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{v}_j^*)'(\mathbf{Y} - \mathbf{X}\mathbf{v}_j^*)\right), \\
d_j &= (1 - p_j) \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{v}_j^{**})'(\mathbf{Y} - \mathbf{X}\mathbf{v}_j^{**})\right).
\end{aligned}$$

Here, the vector $\mathbf{v}_j^*$ is the column vector of ν with the j^{th} entry replaced by β_j, similarly, $\mathbf{v}_j^{**}$ is the column vector of ν with the j^{th} entry replaced by 0. Thus, c_j is a probability for the acceptance of the j^{th} variable and d_j a probability for its rejection.

4. Comparison of the Selection Criteria with Artificial Data

For a comparison of selection criteria in parallel to the study done by Kuo and Mallick [10] and George and McCulloch, artificial data was produced in this study for $n = 20$ and $k = 5$. The distribution of the variables X_j has been assumed to be normal with $\text{N}(0, 1)$. Four types of β structure, (Beta1,...,Beta4), have been determined to observe the effects of the significance level of variables in the model that will be used in selection:

Beta 1 : $(3, 2, 1, 0, 0)$	Beta 2 : $(1, 1, 1, 0, 0)$
Beta 3 : $(3, -2, 1, 0, 0)$	Beta 4 : $(1, -1, 1, 0, 0)$

On the other hand, the levels $\sigma^2 = 0.01, 1.0, 10$ are chosen to see the effects of σ^2 on the results. However, the results for $\sigma^2 = 1.0$ are not given in the tables, but they are mentioned in interpretations. In this study, to observe the effects of prior information levels on the result, the prior assumptions given in Table 4.1 are used. Furthermore, the assumption $e_i \sim \text{N}(0, \sigma^2)$ is made for the errors. The list of all possible subsets for 5 independent variables is given in Appendix 1.

Table 4.1: Prior Assumptions

Bayesian Approach	**Prior Assumption**
Zellner	Equal subset prior probability
Occam's Window	a. Equal subset prior probability. b. For the 2^{nd} and 5^{th} subsets, prior probability is 0.1. c. For the 26^{th} subset, prior probability is 0.10.
Gibbs Sampling	a. $D_0 = 16, 1$ b. Equal prior probability for variables c. Different prior probability for variables

During the comparison of these approaches, the BARVAS (Bayesian Regression-Variable Selection) computer program, developed especially for the Bayesian variable selection process by Yardımcı [17] is used. The BARVAS program can be used for Bayesian regression and variable selection in linear regression, and it is a packet program where the classical variable selection criteria for variable selection could be also used together [18].

4.1. Classical Variable Selection

In the comparison of classical variable selection, the well-known criteria (Cp, AICc, BIC, Press) will be referred to as Group I, and posterior model probability, Risk, RIC as Group II. The results of classical variable selection for the data are summarized in Table 4.2. In this table, the criteria are classified among themselves and the best three subsets are given. It has been observed that the selection criteria in Group I have not been affected by β and σ^2. For the β structures under consideration, all the criteria in Group I find the best subset, which is subset number 5. Also, subsets 10 and 26 are found in all criteria as the best selection subsets. The AICc criterion is more appropriate for a model which has fewer variables because of its structure. Group II criteria show parallel results with other criteria. Subset posterior model probability (P) gave the same classification as Cp in all structural cases. The risk criterion gives better results for subsets with fewer variables. Although subsets with lower risk accord with other criteria, these subsets have lower dimension for the condition $\sigma^2 = 0.01$. RIC gives similar results to Group I. As a result of the analysis of different classifications it is observed that the RIC rates are very close to one another.

4.2. Bayesian Approaches

The Bayesian variable selection approaches, which are calculated using the BARVAS program, are given in Table 4.2. The numbers in the table show the numbers of the selected models. The detailed results for the Bayesian approaches are given as follows.

Table 4.2. Number of Selected Models through the Classical and Bayesian Approaches

(See the Appendix for these numbers).

		Classical Criteria							Bayesian Approaches				
									Zellner		OW	Gibbs	
σ^2	Beta	C_p	AICc	BIC	Press	P	Risk	RIC	BF	F_h	(equal prior)	$D_0 = 1$	$D_0 = 16$
0.01	1	5	2	5	5	5	1	5	5	10	10	5	2
		10	1	10	2	10	2	10	2	26	5	2	5
		2	5	26	13	26	5	26	10	5	21	21	1
	2	5	3	5	5	5	3	5	5	10	5	5	5
		10	5	10	26	10	4	10	10	26	10	21	3
		26	4	26	10	26	7	26	26	5	26	3	6
	3	5	5	5	5	5	5	5	5	26	26	5	5
		26	2	26	26	26	10	10	10	10	5	21	2
		10	26	10	10	10	2	26	26	5	21	2	18
	4	5	5	5	5	5	7	5	5	26	26	5	5
		26	6	26	26	26	6	10	26	10	21	21	2
		10	26	10	10	10	5	26	10	5	5	26	3
10	1	5	5	5	5	5	5	5	5	10	10	5	5
		10	2	10	26	10	10	26	19	26	21	10	2
		26	10	26	10	26	26	10	26	5	5	21	10
	2	5	5	5	5	5	5	5	5	10	10	5	5
		10	1	10	26	10	26	26	10	26	26	21	1
		26	2	26	10	26	10	10	26	5	21	10	2
	3	5	2	5	5	5	5	5	5	26	5	5	5
		26	5	26	10	26	26	26	10	10	26	21	2
		10	13	10	2	10	2	10	26	5	10	2	13
	4	5	5	5	5	5	5	5	5	10	10	5	5
		10	10	10	26	10	10	10	10	26	5	21	3
		26	26	26	10	26	26	26	26	5	21	26	21

The Zellner Approach

The Zellner approach preferred strongly the best subset number 5 for all the different cases of beta and σ^2. The BF rates for the subset 5 are different from those for the alternative subsets 10 and 26. This difference is also reflected in the interpretations of Jeffreys, Kass and Rafftery. The results of the Zellner approach on standardized data are given in Table 4.3. When the rates F_h are considered, it is observed that the alternative subsets 10 and 26 are preferred. Since these subsets have a larger dimension than the best subset, they give better results for the purpose of this study. However, generally when the Zellner approach is used for normally distributed variables, β and σ^2 have no affect.

Table 4.3. Results of the Zellner Approach

	$\sigma^2 = 0.01$							$\sigma^2 = 10.0$						
	for max BF				**for min F_h**			**for max BF**				**for min F_h**		
Beta	**Subset No.**	**BF**	**J**	**KR**	**Subset No.**	**F_h**	**T**	**Subset No.**	**BF**	**J**	**KR**	**Subset No.**	**F_h**	**T**
1	5	13.68	S	S	10	0.09	A	5	12.16	S	H	10	0.02	A
	2	7.17	H	H	26	0.31	A	10	4.46	H	H	26	0.57	A
	10	4.50	H	H	5	0.39	A	26	3.61	H	H	5	0.65	A
2	5	15.29	S	S	10	0.06	A	5	13.37	S	S	10	0.11	A
	10	4.55	H	H	26	0.11	A	10	4.48	H	H	26	0.20	A
	26	4.46	H	H	5	0.14	A	26	4.29	H	H	5	0.44	A
3	5	14.52	S	S	26	0.04	A	5	15.50	S	S	26	0.00	A
	10	4.59	H	H	10	0.24	A	10	4.69	H	H	10	0.10	A
	26	4.21	H	H	5	0.25	A	26	4.47	H	H	5	0.11	A
4	5	0.38	H	H	26	0.04	A	5	13.74	S	S	10	0.02	A
	26	4.60	H	H	10	1.22	A	10	4.65	H	H	26	0.37	A
	10	2.27	W	H	5	1.25	A	26	3.96	H	H	5	0.38	A

J: Jeffreys, **KR**: Kass ve Rafftery, **T**: according to the F table
A: Acceptance, **S**: Strong, **H**: High, **W**: Weak, **R**: Rejection

Occam's Window Approach

The OW approach primarily determines the subsets to be studied, rather than determining the best subset. As a result, the researcher chooses the suitable subset for his aims. On artificial data, while the posterior odds are calculated for the values $O_L = 0.5$ and $OR_{=}1$, the prior probability for all subsets is considered to be equal at the first stage. The OW approach evaluates the full set as a good model. However, the OW approach concentrates more on larger dimensional subsets. The results for the OW approach are given in Table 4.4. In this study, the best subset number 5 is generally selected. However the posterior probabilities of subsets in the classification are close to one another. It may be observed that the OW approach is not much affected by a change in the β and σ^2 structures. At the second stage, Case I and II are defined to observe the effects of prior probabilities on the results. In Case I, the prior probability of subset 2 and 5 is 0.10. In Case II, a 0.10 prior probability is given to subset 26. The prior probabilities of the other subsets are 0.90/29 for Case I and 0.90/30 for Case II. It may be observed that the subset prior probabilities have a considerable effect on the OW results. The effect in Case I is more than in Case II. In Case I, the prior probability determined for subset 2 is not sufficient for it to be taken into the classification, but subset 5 has a high posterior probability.

The Gibbs Sampling Approach

The Gibbs sampling approach used by Kuo and Mallick [10] is applied using the BARVAS program, and the results given in Table 4.5. Gibbs sampling is applied for all cases using simulated data over 10.000 iterations. The sampling time lasted approximately two minutes for each of the σ^2, β and D_0 levels. By considering the

prior parameter estimation, which is $\beta_0 = 0$, the noninformative prior information has been utilized in the Kuo and Mallick approach to Gibbs sampling.

Table 4.4. The Results for Occam's Window Approach

	$\sigma^2 = 0.01$					
	Equal prior		**Case 1**		**Case 2**	
Beta	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**
	10	0.25	5	0.50	26	0.47
1	5	0.23	10	0.15	10	0.17
	21	0.21	21	0.13	5	0.15
	5	0.27	5	0.58	26	0.52
2	10	0.25	10	0.15	5	0.17
	26	0.24	26	0.14	10	0.16
	26	0.27	5	0.56	26	0.56
3	5	0.26	26	0.16	5	0.16
	21	0.23	21	0.14	21	0.14
	26	0.36	5	0.44	26	0.65
4	21	0.30	26	0.24	21	0.17
	5	0.18	21	0.21	5	0.09
	$\sigma^2 = 10.0$					
	Equal prior		**Case 1**		**Case 2**	
Beta	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**
	10	0.31	5	0.51	26	0.46
1	21	0.26	10	0.20	10	0.21
	5	0.22	21	0.17	21	0.17
	10	0.27	5	0.53	26	0.52
2	26	0.25	10	0.17	10	0.16
	21	0.24	26	0.15	21	0.15
	5	0.26	5	0.56	26	0.53
3	26	0.25	26	0.15	5	0.17
	10	0.23	10	0.14	10	0.15
	10	0.28	5	0.55	26	0.48
4	5	0.25	10	0.17	10	0.19
	21	0.24	21	0.14	5	0.17

Case 1: 0.10 prior probability is given to subsets 2 and 5;

Case 2: 0.10 prior probability is given to subset 26.

Pos. P: Subset posterior probability.

The effect on the results have been observed by using prior variance values of $D_0 = 1$ and 16. The number of zero subsets (Zero) reached, and the subset numbers (Processed) which were processed were given by the BARVAS program for Gibbs sampling. It may be observed that the number of zero subsets is affected

by the β structure. A lot of computation time was needed for the zero subsets for $\sigma^2 = 0.01$ and $\sigma^2 = 1.0$ with Beta 2, and for $\sigma^2 = 10$ with Beta 4. It may be observed that D_0, which is regarded as representing the prior information level, has no effect on the number of processed subsets, but is effective on the appearance of zero subsets. In this study, it may be observed that β and σ^2 are not very influential in obtaining subsets. However for $\sigma^2 = 10$, the posterior model probabilities for the correct subset are greater than the others. During this study, the high prior probability rates and the different prior probabilities in the model are given for redundant variables in all β structures in order to observe the effects of the independent variables on the process. As a result of this application, it has been observed that the prior probabilities reversibly determined for variables in the model have no effect on the results.

Table 4.5 The Results for the Gibbs Sampling Approach

	$\sigma^2 = 0.01$				$\sigma^2 = 10$			
	$D_0 = 1$		$D_0 = 16$		$D_0 = 1$		$D_0 = 16$	
Beta	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**	**Subset No.**	**Pos. P**
1	5	0.37	2	0.63	5	0.62	5	0.70
	2	0.31	5	0.27	10	0.12	2	0.21
	21	0.24	1	0.05	21	0.10	10	0.05
	Zero = 4, Processed = 16		Zero = 0, Processed = 14		Zero = 48, Processed = 14		Zero = 9, Processed = 10	
2	5	0.72	5	0.66	5	0.74	5	0.70
	21	0.17	3	0.22	21	0.15	1	0.20
	3	0.05	6	0.03	10	0.05	2	0.05
	Zero = 46, Processed = 25		Zero = 389, Processed = 19		Zero = 57, Processed = 19		Zero = 125, Processed = 19	
3	5	0.70	5	0.75	5	0.71	5	0.68
	21	0.10	2	0.19	21	0.14	2	0.26
	2	0.05	18	0.02	2	0.09	13	0.03
	Zero = 0, Processed = 13		Zero = 5, Processed = 13		Zero = 0, Processed = 10		Zero = 16, Processed = 9	
4	5	0.64	5	0.59	5	0.80	5	0.88
	21	0.13	2	0.15	21	0.16	3	0.06
	26	0.11	3	0.12	26	0.03	21	0.03
	Zero = 0, Processed = 16		Zero = 26, Processed = 18		Zero = 48, Processed = 14		Zero = 448, Processed = 19	

5. Application

An application using air pollution data [3] is also used for comparing the criteria. The dependent variable is the quantity of SO_2. There are 8 independent variables and 255 subsets for these data. In our study, the classical criterion such as Cp, BIC, PRESSp, P (posterior model probability) and RIC have selected the same

subsets which given with model numbers 162, 163 in Table 5.1. AICc has selected the models 156, 163. The results of Bayesian criterion are also shown in Table 5.1. The models selected by Bayesian criterion are fairly like to be the results of classical selection. While Occam's Window method tends to select the more large subsets, Gibbs sampling used by Kuo and Mallick tends to select the subsets with less variables. The advantage of Gibbs sampling is to have processed only 54 instead of 255 subsets for D0=1 or 42 subsets for D0=4. If the models with less variables is preferred according the parsimony rule or the aim of prediction, the models 156 or 163 can be used to make the estimation of parameters of these models. Using least square or Bayesian methods can make the estimation of parameters of these models.

Table 5.1 The Variables and Results for Bayesian Selection.

No.	Variables	Zellner		Occams Window		Gibbs	
		Max. BF	Min. F_h	Pos. P[1]	Pos. P[2]	Pos. P ($D_0 = 1$)	Pos. P ($D_0 = 4$)
99	x_2, x_5, x_7						0.22
156	x_2, x_5, x_7, x_8	200.0	0.07			0.20	0.33
162	$x_1, x_2, x_5, x_6, x_7, x_8$	162.2	0.07	0.179	0.160	0.22	
163	x_2, x_5, x_6, x_7, x_8	255.1				0.31	0.23
165	$x_1, x_2, x_3, x_5, x_6, x_7, x_8$		0.02	0.179	0.160		
170	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$			0.176	0.157		

1) Posterior probability for equal prior, **2)** 0.10 prior probability for model 156.

6. Conclusion

As a result of this study, it can be claimed that Bayesian approaches are affected by prior changes. Generally, the β and σ^2 structures do not have any effect on the results. However Bayesian approaches define the correct subset most of the time.

In Gibbs sampling, the prior probabilities defined for variable selection are not effective. However, it can be said that prior information as used in parameter estimation directly affects the results. It is also observed that the prior information level is influential in obtaining zero subsets. In Gibbs sampling it is also observed that few subsets having high posterior probability, are dealt with.

It has been observed that the risk criterion proposed by Yardımcı [17] prefers smaller dimensional subsets than the other criteria, and that these subsets have average posterior probabilities. In other words, analyzing and interpreting together the posterior probabilities of subsets with low risk will help the researcher to determine the good subsets. In the Zellner and Occam's window approaches, subset prior probability is effective on the results. It has also been observed that Occam's window prefers larger dimensional subsets than both the Zellner approach and Gibbs sampling.

The Bayesian approaches are more effective, though not very much so, than all possible subset selection. They have the advantage that only the evaluation of subsets with a high posterior probability is needed.

7. Appendix

List of subsets for $k = 5$

Subset No	Variables	Subset No	Variables
1	x_1	17	$x_1,\quad x_4,\quad x_5$
2	$x_1,\quad x_2$	18	$x_1,\quad x_2,\quad x_4,\quad x_5$
3	x_2	19	$x_2,\quad x_4,\quad x_5$
4	$x_2,\quad x_3$	20	$x_2,\quad x_3,\quad x_4,\quad x_5$
5	$x_1,\quad x_2,\quad x_3$	21	$x_1,\quad x_2,\quad x_3,\quad x_4,\quad x_5$
6	$x_1,\quad x_3$	22	$x_1,\quad x_3,\quad x_4,\quad x_5$
7	x_3	23	$x_3,\quad x_4,\quad x_5$
8	$x_3,\quad x_4$	24	$x_3,\quad x_5$
9	$x_1,\quad x_3,\quad x_4$	25	$x_1,\quad x_3,\quad x_5$
10	$x_1,\quad x_2,\quad x_3,\quad x_4$	26	$x_1,\quad x_2,\quad x_3,\quad x_5$
11	$x_2,\quad x_3,\quad x_4$	27	$x_2,\quad x_3,\quad x_5$
12	$x_2,\quad x_4$	28	$x_2,\quad x_5$
13	$x_1,\quad x_2,\quad x_4$	29	$x_1,\quad x_2,\quad x_5$
14	$x_1,\quad x_4$	30	$x_1,\quad x_5$
15	x_4	31	x_5
16	$x_4,\quad x_5$		

References

[1] Carlin, B.P. and Chib, S. Bayesian model choice via markov chain monte carlo methods, J. R. Statist. Soc. B 57, (3), 473–484, 1995.

[2] Cavanaugh, J. E. Unifying the derivations for the Akaike and corrected Akaike information criteria, Statistics and Probability Letters 33, 201–208, 1997.

[3] Candan, M. The Robust Estimation in Linear Regression Analysis, Unpublished MSc. Thesis (in Turkish), Hacettepe University, Ankara, 2000.

[4] Draper, N. and Smith, H. Applied Regression Analysis, John Wiley, 709 p, 1981.

[5] Foster, D. P. and George, E. I. The risk inflation criterion for multiple regression, The Annals of Statistics 22 (4), 1947–1975, 1995.

[6] Gamerman, D. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman and Hall, London, 245p, 1997.

[7] George, E. I. and McCulloch, R. E. Approaches for Bayesian variable selection, Statistica Sinica 7, 339–373, 1997.

[8] Hoeting, J. A. Accounting for model uncertainty in linear regression, PhD. Thesis, University of Washington, 1994.

[9] Kass, R. E. and Raftery, A. E. Bayes factors, JASA 90, 773–795, 1995.

[10] Kuo, L. and Mallick, B. Variable selection for regression models, Sankhya B 60, 65–81, 1998.

[11] Lee, P. M. Bayesian Statistics: An Introduction, Oxford, 264p, 1989.

[12] Madigan, D. and Raftery, A. E. Model selection and accounting for model uncertainty in graphical models using Occam's window, JASA 89, 1535–1546, 1994.

[13] Moulton, B. I. A Bayesian approach to regression selection and estimation, with application to price index for radio services, Journal of Econometrics 49, 169–193, 1991.

[14] Raftery, A., Madigan, D. and Hoeting, J. Model selection and accounting for model uncertainty in linear regression models, University of Washington, Tech. Report. No. 262., 24p, 1993.

[15] Toutenburg, H. Prior information in linear models, John Wiley Inc., 215p, 1982.

[16] Wasserman, L. Bayesian model selection, Symposium on Methods for Model Selection, Indiana Uni., Bloomington, 1997.

[17] Yardımcı, A. Comparison of Bayesian Approaches to Variable Selection in Linear Regression, Unpublished PhD. Thesis (in Turkish), Hacettepe University, Ankara, 2000.

[18] Yardımcı, A. and Erar, A. Bayesian variable selection in linear regression and BARVAS Software Package (in Turkish), Second Statistical Congress, Antalya, 386–390, 2001.

[19] Zellner, A. and Siow, A. Posterior odds ratio for selected regression hypothesis, Bayesian Statistics, University Press, Valencia, 585–603, 1980.