APPLICATIONS OF BAYESIAN VARIABLE SELECTION IN TWO SOCIOLOGICAL DATA SETS

by

Suhai Liu

ISDS
Duke University

Date: _____

Approved:

_____
Dr. Peter Mueller, Supervisor

_____
Dr. Donald Berry

_____
Dr. Kenneth C. Land

_____
Dr. Bernadette Pelissier

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the ISDS
in the Graduate School of
Duke University

1998

Abstract

Variable Selection has been an important issue in sociological research. Many times, when a sociologist uses regression analysis for a fairly large number of regressors, he or she would face a practical problem – which variables should be included in the regression, and which variables should not. Traditional methods of classical statistics often fall short of the goal to include a meaningful subset of the regressors. Freedman's Paradox has exposed the limitations of using P-values to choose independent variables. Bayesian variable selection methods have revealed promising approaches to address the issues of variable selection. Those methods, including Occam's window and Monte Carlo variable selection, perform better than classical variable selection such as stepwise procedures using P-values. This thesis applies Bayesian variable selection approaches to two sociological data sets. One is macro-level national homicide data and the other is individual-level drug treatment data. I also emphasize the different focuses of two analyses: for the homicide data set, the focus is on identification of theoretically important independent variables, while for the drug treatment data, the focus is on prediction of completion of the treatment program. I also compare the results from different methods and comment on the consistency and discrepancy across their results.

Acknowledgements

It would have been impossible for me to accomplish this thesis without the help from many kind people. Here I express my sincere appreciation and gratefulness for all their support and assistance.

I would like to thank Dr. Peter Mueller for his persistent patience and encouragement in guiding me through the process. I would like to thank Dr. Ken Land for his generous support and invaluable insight for my research direction. I would like to thank Dr. Don Berry for his warm encouragement and helpful suggestions. I would like to thank Dr. Bernadette Pelissier for going extra miles to help me access the precious drug-treatment data set.

Besides, I would like to thank Dr. Adrian E. Raftery, Dr. Jennifer Hoeting and Dr. Chris Volinsky for their help in my understanding and applying Bayesian Model Averaging methods. I have also found great help in the comments on an earlier draft from Dr. Katja Ickstadt and Dr. Merlise Clyde. I also want to thank Dr. Steven Messner for his help in re-locating the national homicide data set and making it available for me.

Finally, I would also want to thank my friends in Chinese Christian Mission Church and Duke Chinese Christian Fellowship for their care, love and prayers.

Contents

List of Tables

Chapter 1

Introduction: Why and How should a Sociologist Select Independent Variables for Regression Analysis?

1.1 Why Variable Selection?

Variable selection is an unavoidable issue in regression analysis for a quantitative sociologist. Often, with a data set with a large number of potential regressors at hand, it is important for a sociologist to determine how to choose independent variables, either for theoretically-oriented research or for practically-oriented research. For a theoretically-oriented research, it is essential to include variables representing competing theoretical perspectives and identify the really important ones in the regression. Variables representing theories with more explanatory power would surpass in importance those standing for weaker theories. For a practically-oriented research, often prediction is more interesting. Using a 'promising' subset of covariates to predict outcomes is more economical and efficient in both measurement and computation than using all the possibly available covariates.

1.2 How to Select Variables?

Raftery (1995) noted that although, traditionally, P-values and significance tests have been used for statistical inference, some quantitative sociologists have attached less importance to them due to practical difficulties and counter-intuitive results associated with P-values. These difficulties are summarized by Raftery to the following two aspects:

1) Difficulties often arise with large samples where P-values tend to reject the null hypothesis even when the null model is theoretically valid and only has very small discrepancies with the data. Raftery also pointed out that in the early 1980s, some sociologists tackled this problem by basing model selection on theoretical considerations and informal assessment of discrepancies between model and data instead of by examining the results of P-value-based tests, which seemed to be counter-intuitive (Fienberg and Mason, 1979; Hout, 1983, 1984; Grusky and Hauser, 1984).

2) Difficulties also arise when many statistical models regarding choice of different variables are implicitly considered in the earlier stages of a data analysis. It happens when, having many control variables, a social scientist conducts variable selection to decide which ones to be included in the final model either by removing the insignificant ones from the full model or more formally by stepwise regression. On the one hand, P-values based on a model selected from among a large set of possibilities no longer have the same interpretation as when only two models were ever considered. On the other hand, one might have several different models which may all seem reasonable given the data but nevertheless lead to different conclusions about question of interest (Kass and Raftery, 1995; Raftery, 1993). In this situation, the standard approach of selecting a single model and basing inference on it underestimates uncertainty about quantities of interest because it ignores model uncertainty.

This thesis will mainly deal with the second kind of difficulties by introducing the

alternative variable selection approach – Bayesian variable selection, which takes into account model uncertainty. Before we go into that approach, I would like to revisit a classical example where a classical variable selection approach would fail to take into account model uncertainty and therefore provide misleading statistical inference when one has many candidate independent variables.

1.3 Freedman's Paradox Revisited: The Inadequacy of Classical Variable Selection

Freedman (1983) performs a simple simulation experiment on the reliability of classical variable selection. He created a matrix with 100 rows (data points) and 51 columns (variables). All the data in that matrix were independent observations drawn from the standard normal distribution. The 51st column was regarded as the response variable Y in a regression; the first 50 columns were taken as the covariates X_1,..,X_{50}. By construction, Y was independent of the X's, and R^2 should have been insignificant, by the standard F test. Similarly, the regression coefficients should have been insignificant, by the standard t test.

Raftery (1995) conducted this experiment and his results bear surprisingly high statistical significance. Looking at the full model, he found that 7 coefficients out of the 50 were significant at the .05 level. Stepwise regression shows a four-variable model with R^2= 0.18 and $P = 10^{-6}$, and all four coefficients significant at the .05 level.

I myself also replicated this experiment with a smaller data set of 100 cases, 30 independent variables, and one response variable. All the variables are simulated from standard normal distribution.

For the full model, I got X_{22} significant at .05 level, X_{13} and X_{20} significant at .10 level.

For the stepwise regression, I got a model with four covariates with Efroymson's forward selection method (F criterion =2). After I regress Y on the four X's, I obtained the results as shown in Table 1.1. There are two variables significant at .05 level with a model $R^2 = .12$ and P=.017. Those results from stepwise regression are misleading.

The inadequacy of the classical variable selection methods come from the fact that one does not take into account model uncertainty by just looking at one or two models. With such a large number of candidate covariates, the variable selection procedure often locates a model with satisfying statistical significance, even in the case of this pure-noise data.

This inadequacy can be overcome by taking into consideration a much larger number of models with different subset of covariates. In the next chapter, we can see that the Bayesian variable selection offers promising solutions to this problem.

Table 1.1: Stepwise regression results for simulated noise.

Variable	Coefficient	t	P
Intercept	-0.2200	-2.4233	0.0173
X_4	0.1955	1.8130	0.0730
X_{11}	-0.1821	-2.1133	0.0372
X_{13}	0.1261	1.4381	0.1537
X_{22}	-0.2135	-2.3842	0.0191

Chapter 2

The Promising Solution: Bayesian Variable Selection

This chapter introduces the available Bayesian variable selection methods. These methods take into account model uncertainty by considering a number of models instead of just looking at one or two of them. All of the methods introduced in this chapter will be applied in **Chapter 3** and **Chapter 4** to real sociological data sets.

2.1 Bayesian Model Averaging

Raftery et. al. (1997) provided an extensive summary of Bayesian Model Averaging methods. This section is based on their review.

There is a standard Bayesian solution to the problem of model uncertainty (Leamer, 1978). If $M = \{M_1, ..., M_k\}$ denotes the set of all models being considered and if Δ is the quantity of interest, then the posterior distribution of Δ given the data D is

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|M_k, D)Pr(M_k|D) \tag{2.1}$$

This is an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities. Raftery et al (1997) called this Bayesian

Model Averaging (BMA). In Equation (1) the posterior probability of model M_k is given by

$$Pr(M_k|D) = \frac{p(D|M_k)Pr(M_k)}{\sum_{l=1}^{K} p(D|M_l)Pr(M_l)}, \tag{2.2}$$

where

$$p(D|M_k) = \int p(D|\Theta_k, M_k)p(\Theta_k|M_k)d\Theta_k \tag{2.3}$$

is the marginal likelihood of model M_k, Θ_k is the vector of parameters of model M_k, $p(\Theta_k|M_k)$ is the prior density of Θ_k under model M_k, $p(D|\Theta_k, M_k)$ is the likelihood, and $Pr(M_k)$ is the prior probability that M_k is the true model. To implement BMA, we need to fulfill two tasks. First, we need to compute $p(D|M_k)$, the marginal likelihood of model M_k. Second, we need to address all the models in (2.1), the numbers of which can be enormous.

2.1.1 A Bayesian Framework

Raftery et al (1997) considered normal linear model of the form

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon = X\beta + \epsilon, \tag{2.4}$$

where X is an $n \times (p+1)$ matrix of the observed data on the covariates, Y is an n-vector of the data on the response variable. We assume that $\epsilon's$ have an independent normal distribution with mean zero and variance σ^2.

With vague theory about the importance of the individual covariates, Raftery et al (1997) used the standard normal gamma conjugate priors:

$$\beta \sim N(\mu, \sigma^2 V)$$

and

$$\frac{\nu\lambda}{\sigma^2} \sim X^2_\nu.$$

Here ν, λ, the $(p+1)\times(p+1)$ matrix V, and the $(p+1)$ vector μ are hyperparameters to be chosen.

Therefore, we can obtain the marginal likelihood for Y under a model M_i, [1]

$$\begin{aligned} & p(Y|\mu_i, V_i, X_i, M_i) \\ &= \frac{\Gamma\left(\frac{\nu+n}{2}\right)(\nu\lambda)^{\frac{\nu}{2}}}{\pi^{\frac{n}{2}}\Gamma(\frac{\nu}{2})|I+X_iV_iX_i^t|^{1/2}} \\ &\quad \times [\lambda\nu + (Y-X_i\mu_i)^t(I+X_iV_iX_i^t)^{-1}(Y-X_i\mu_i)]^{-(\nu+n)/2}, \end{aligned} \tag{2.5}$$

where X_i is the design matrix and V_i is the covariance matrix for β corresponding to model M_i.

The Bayes factor for M_0 versus M_1, the ratio of equation (2.5) for $i=0$ and $i=1$, is then given by

$$B_{10} = \Big(\frac{|I+X_1V_1X_1^t|}{|I+X_0V_0X_0^t|}\Big)\Big[\frac{a_0}{a_1}\Big]^{-(\nu+n)/2}, \tag{2.6}$$

where $a_i = \lambda\nu + (Y-X_i\mu_i)^t \times (I+X_iV_iX_i^t)^{(-1)}(Y-X_i\mu_i)$, for i=0,1.

2.1.2 Selection of Prior Distributions

Raftery et. al. (1997) chose the following priors:

$$\mu = (\hat{\beta}_0, 0, 0, ..., 0),$$

where $\hat{\beta}_0$ is the ordinary least squares estimate of β_0, and

[1] $p(Y|\mu_i, V_i, X_i, M_i)$ is denoted as $p(D|M_k)$ in (2.3). It can be obtained by integrating out β and σ^{-2} from $f_N^{(n)}(y|X\beta, \sigma^{-2}I)f_{N\gamma}^{(p+1)}(\beta, \sigma^{-2}|\mu, V^{-1}, \nu, \lambda)$. For proof, see Hoeting (1994, p. 13).

$$\mathbf{V}(\beta) = \sigma^2 \begin{pmatrix} S_Y^2 & & & & & & & \\ & \phi^2 S_1^{-2} & & & & & & \\ & & \ddots & & & & & \\ & & & \phi^2 S_{i-1}^{-2} & & & & \\ & & & & \phi^2 (\frac{1}{n} X^{i^T} X^i)^{-1} & & & \\ & & & & & \phi^2 S_{i+1}^{-2} & & \\ & & & & & & \ddots & \\ & & & & & & & \phi^2 S_p^{-2} \end{pmatrix},$$

where S_Y^2 denotes the sample variance of Y, S_i^2 denotes the sample variance of X_i, and ϕ is a hyperparameter to be chosen. For quantitative predictor variables, the covariance matrix V is equal to σ^2 multiplied by a diagonal matrix with entries $(S_Y^2, \phi^2 S_1^{-2}, \phi^2 S_2^{-2}, ..., \phi^2 S_p^{-2})$; for a categorical predictor variable X_i with $(c+1)$ possible outcomes, the prior variance of $(\beta_{i1}, ..., \beta_{ic})$ is set to $\sigma^2 \phi^2 [\frac{1}{n} X^{i^T} X^i]^{-1}$, where X^i is the $n \times c$ design matrix for the dummy variables, where each dummy variable has been centered by subtracting its sample mean.

Assuming that all the variables have been standardized to have mean zero and sample variance 1, it is preferred that the prior density $p(\beta_1, ..., \beta_p)$ is reasonably flat over the unit hypercube $[-1, 1]^p$, $p(\sigma^2)$ is reasonably flat over $(\alpha, 1)$ for some small α, and $Pr(\sigma^2 \leq 1)$ is large. All these desired features warrant the priors to be vague.

Raftery et. al. (1997) stated that, for $\alpha = .05$, a set of priors of $\nu = 2.58, \lambda = .28$, and $\phi = 2.85$ are proper for linear regression models, resulting in $Pr(\sigma^2 \leq 1) = .81$. They are used in **Chapter 3**.

2.1.3 Occam's Window[2]

To deal with the large number of candidate models, Occam's window (OW) algorithm of Madigan and Rafterty(1994) is applied. Two basic principles underly this approach.[3]

First, if a model predicts the data far worse than the model that provides the best predictions, then it has effectively been discredited and should no longer be considered. Thus models not belonging to

$$A' = \left\{ M_k : \frac{max_l\{Pr(M_l|D)\}}{Pr(M_k|D)} \leq C \right\}$$

should be excluded from the equation above, where C is chosen by the data analyst and $max_l\{Pr(M_l|D)\}$ denotes the model with the highest posterior model probability.

Second, applying Occam's razor, models that receive less support from the data than any of their simpler submodels are excluded. That is, we should exclude models belonging to

$$B = \left\{ M_K : \exists M_l \in M, M_l \subset M_k, \frac{Pr(M_l|D)}{Pr(M_k|D)} > 1 \right\}$$

After these two principles are applied, only models $A = A'/B$ are considered. It has greatly reduced the number of the models in the sum of Equation (2.1). Madigan and Raftery (1994) provided a detailed description of their search algorithm and showed how averaging over the selected models provides better predictive performance than basing inference on a single model.

[2]For easy implementation, BIC approximation is used for Occam's window approach instead of the exact Bayesian setup in **2.1.1** and **2.1.2**. For more on BIC approximation, see Raftery (1995).

[3]The second principle of OW is not used in this thesis because it would lead to a very small number of models.

2.1.4 Markov Chain Monte Carlo Model Composition (MC^3)

To cope with the fact that the number of terms in (2.1) can be enormous, another BMA approach is considered. A Markov chain $\{M(t), t = 1, 2, ...\}$ is constructed with state space Ω and equilibrium distribution $Pr(M_i|D)$. For any given function $g(M_i)$, the average

$$\hat{G} = \frac{1}{N}\sum_{t=1}^{N} g(M(t)) \tag{2.7}$$

converges to E(g(M)) as $N \to \infty$. Setting $g(M) = p(\Delta|M, D)$, we can compute (2.1).

To construct the Markov chain, Raftery et. al. (1997) defined a neighborhood nbd(M) for each $M \in \Omega$ that consists of the model M itself and the set of models with either one variable more or one variable less than M. Define a transition matrix q by setting $q(M \to M')$ constant for all $M' \in$ nbd(M), and 0 for all the other $M's$. If the chain is currently in state M, then we proceed by drawing M' from $q(M \to M')$. According to the Metropolis algorithm with symmetric proposal, the acceptance probability for the candidate model M' is reduced to

$$min\left\{1, \frac{Pr(M'|D)}{Pr(M|D)}\right\}.$$

With the acceptance probability, the researcher decides to either accept M' and move to it, or stays in state M'. Such a procedure was described in Madigan and York (1995). In Raftery et al. (1997), the procedure which performs the above MCMC model composition is called MC^3.

2.2 Stochastic Search Variable Selection

George and McCulluoch (1993) proposed a hierarchical Bayesian model for variable selection. Each component β_i of the regression parameter vector β is modeled as

being generated either from a distribution with most of its mass concentrated about zero, or from a distribution with its mass spread out over plausible values (Figure 2.1). This is done by constructing a scale mixture of two normal distributions, which using a latent variable $\gamma_i = 0$ or 1, may be conveniently expressed as

$$p(\beta_i|\gamma) = (1-\gamma_i)N(0, {\tau_i}^2) + \gamma_i N(0, {c_i}^2{\tau_i}^2) \tag{2.8}$$

where τ and c are hyper priors for the dispersion,

and

$$P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = \omega_i \tag{2.9}$$

where ω_i is the prior probability of β_i being generated from a distribution with its mass spread out over plausible values.

The priors are set up as follows:

$$p(\beta|\gamma) = N_p(0, {D_\gamma}^2) \tag{2.10}$$

where D_γ is a diagonal matrix with i^{th} diagonal element equal to $(1-\gamma_i)\tau_i + \gamma_i c_i \tau i$,

$$p(\gamma) = \prod_{i=1}^{p+1} \omega_i^{\gamma_i}(1 - \omega_i)^{(1-\gamma_i)} \tag{2.11}$$

and

$$p(\sigma^{-2}) = Ga\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right) \tag{2.12}$$

where ν and λ are hyper priors.

Stochastic Search Variable Selection (SSVS) uses the Gibbs sampler to generate a sequence

$$\gamma^{(1)}, \gamma^{(2)}, \ldots$$

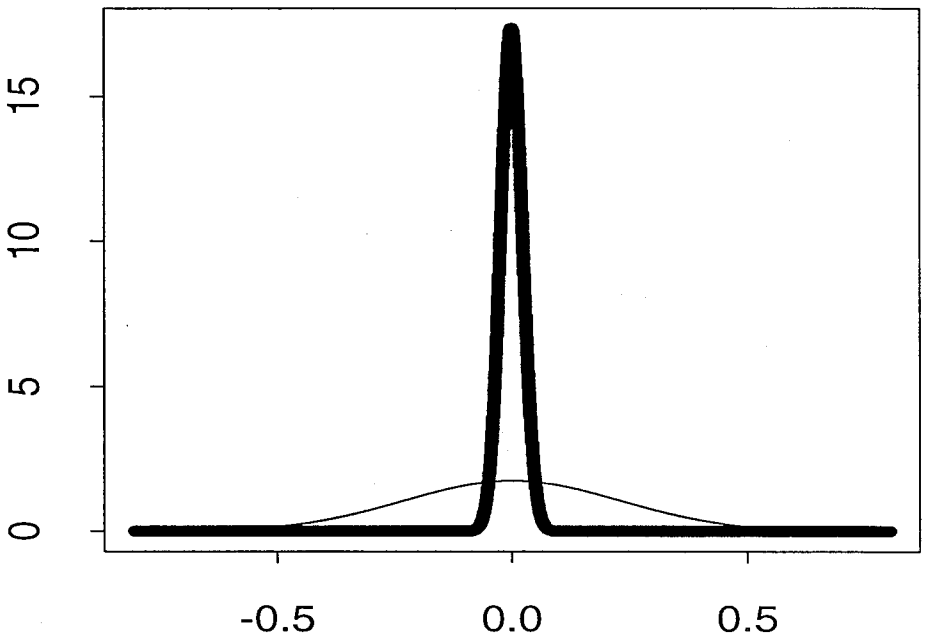

Figure 2.1: Concentrated density $N(0, \tau_i^2)$ and diffuse density $N(0, c_i^2\tau_i^2)$.

which converges in distribution to $\gamma \sim p(\gamma|Y)$.

The full conditional distributions are as follows

$$p(\beta|\sigma^2, \gamma, Y) = N_p((X^TX + \sigma^2 D_\gamma^{-2})^{-1}X^TY, \sigma^2(X^TX + \sigma^2 D_\gamma^{-2})^{-1}) \tag{2.13}$$

$$p(\sigma^{-2}|\beta, Y) = Ga\left(\frac{n+\nu}{2}, \frac{|Y - X\beta|^2 + \nu\lambda}{2}\right) \tag{2.14}$$

and

$$P(\gamma_i = 1|\beta_i) = \frac{a}{a+b} \tag{2.15}$$

where

$$a = p(\beta_i|\gamma_i = 1)\omega_i \tag{2.16}$$

$$b = p(\beta_i|\gamma_i = 0)(1 - \omega_i) \tag{2.17}$$

To set the prior input for SSVS, $\delta_i = .05$ is used as the threshold for practical significance. This determination was based on the sense that when $|\beta| \leq .05$, the 'effect' of X_i on Y would be negligible so that it would be preferable to exclude X_i from the subset. George and McCulluoch (1996) then chose $c_i = 10$, $\tau_i = .023$ to achieve this δ. Further, $\omega_i = .5$, $\nu = 5$, $\lambda = .004$ are set for other priors.

2.3 Freedman's Paradox Solved

Bayesian variable selection methods introduced in this chapter provide a promising solution to the practical difficulty in classical methods such as stepwise regression. Hoeting (1995) gives a detailed explanation of how the BMA methods can solve Freedman's Paradox. Here I simply use OW method to reanalyze the simulated data which we dealt with in **1.3** and highlight how the difficulty is solved.

Table 2.1 shows that OW returns the null model as the top model with $P(M|D) = 10.1\%$. It means that if we look for the model with the highest probability of being the true model given the data, it would be the null model. In other words, none of the independent variables would help in explaining the variability in the response variable. This result **agrees** with the fact that all the variables in this data set are noise.

In **Chapter 3** and **Chapter 4**, I will use the Bayesian variable selection methods to analyze two sociological data sets.

Table 2.1: Bayesian variable selection results for simulated noise: top 10 models

Models	$Pr(M_i\|D)\%$
NULL	10.9
X_{22}	9.6
X_{11}, X_{22}	6.2
X_{11}	5.9
X_4	5.8
X_4, X_{22}	4.3
X_{20}	4.2
X_{28}	4.1
X_{23}	3.9
X_4, X_{11}, X_{22}	2.9

Chapter 3

Bayesian Variable Selection in Normal Linear Models: A Macro-Level National Homicide Analysis

Using linear regression, Messner (1989) examined the effects of economic discrimination on national rates of homicide. He found that the effects of discrimination exceeded those of income inequality, based on the full model and the model chosen by stepwise regression. However, as it is known, such a classical model selection approach tends to underestimate the uncertainty of quantities of interest since it chooses only one model and ignores all the others. Using OW, MC^3 and SSVS procedures, I will reanalyze Messner's data and generate posterior information about the variables of interest. I will also compare OW, MC^3 and SSVS results with those from Messner's classical approach. Results show that Bayesian and classical approaches bear similar results when there is less multicollinearity between the covariates. However, when there is serious multicollinearity, OW, MC^3 and SSVS take into account the model uncertainty and provide more meaningful inferences on regressor effect.

3.1 Theoretical Question

Messner (1989) examined the effects of economic discrimination against social groups on national rates of homicide. Drawing on Peter Blau's macrostructural theory (Blau and Schwartz, 1984), Messner hypothesized that nations with intense and pervasive discrimination would exhibit comparatively high levels of homicide, and that the effects of discrimination would exceed those of income inequality. Classical regression analyses using cross-national homicide data supported both hypotheses. Indicators of economic discrimination against social groups were found significantly and positively related to homicide rates despite fairly extensive control for the other theoretically relevant national characteristics. Messner also found that the effect for discrimination was consistently stronger than that for income concentration. Messner concluded that the structuring of economic inequality on the basis of ascribed characteristics was a particularly important source of lethal violence in contemporary societies.

Messner's results were based on a method commonly used in sociological research. First, a full model is visited with all the covariates included, and significant variables are identified. Then, a stepwise regression is used to choose a model with only significant covariates by adding/deleting variables to/from subsets of covariates. Inferences are drawn based on the consistency of the two models. However, it is known that, with only one or two models selected, model uncertainty is ignored and the uncertainty about quantities of interest is underestimated. This problem is usually prominent when the sample size is small and multicollinearity is present between the covariates, as it is in Messner's case.

To tackle such a problem, I will use BMA and SSVS methods to reanalyze Messner's data and revisit his research question. As we have seen in **Chapter 2**, BMA averages the quantity of interest ($\beta_j \neq 0$) over all the models weighted by their posterior model probability, and SSVS traverses the model space by simulating Bernoulli

outcomes for the impact of each independent variable. Therefore, information from a large number of models are taken into account. Bayesian methods provide posterior probability of $\beta_j \neq 0$. By doing so, I try to answer the same research question – Is Economic Discrimination more important than Income Inequality in their effects on Homicide Rates?

3.2 Data

Gathering data from various sources, Messner (1989) obtained a sample of 52 contemporary nation-states. The years of data vary depending on variables, and some variables are averaged values over several years while some are from one particular year. But the time spans from late 60's to the early 80's. Table 3.1 displays the means and standard deviation of all the variables. Natural log transformation indicated by (ln) is performed when it is appropriate.

Table 3.1: Description of the Variables

Variable	Mean	Std. Dev.
Average INTERPOL Homicide Rates(ln)	1.35	.89
Economic Discrimination Dummy	.46	.50
Income Inequality	.46	.09
Infant Mortality Rate	47.18	40.62
Population less than 15	35.42	9.75
GNP/capita(ln)	7.82	1.26
Annual Rate of Population Increase	1.87	1.16
Percent Urban Population	55.48	24.93
Ethno-linguistic Heterogeneity Index	.33	.29
Population(ln)	16.50	1.27
Population Density(ln)	4.16	1.53
Political Democracy Index	71.72	28.81
Percent Male 15-29	13.09	1.63
Development Index	0	.98
Life Expectancy	65.37	8.65

The dependent variable is the average rate of homicides for 1977-82 known to the police per 100,000 inhabitants. The source for this measure is International Crime Statistics, published by the International Criminal Police Organization (INTERPOL).

The primary independent variable – economic discrimination – is coded as a dummy variable, with nations with economic discrimination assigned '1', and all the others '0'. It is defined as 'the deliberate, invidious exclusion of social groups from some desired economic goods or conditions (values) because of the groups' ascribed characteristics' (Taylor and Jodice 1983, 52).

Representing a competing theory, the second independent variable is income inequality, which is operationalized by the Gini coefficient of household or individual income concentration.

The other covariates are control variables. Among them, five are highly correlated: population less than 15 years of age; the annual rate of population increase; life expectancy; GNP/capita (ln); and the infant mortality rate. To avoid multicollinearity, Messner constructed a latent variable 'development index' based on a principal factor analysis. As an alternative approach, he also used 'population less than 15' as the representative variable.

3.3 Analyses

I will use OW, MC^3 and SSVS to reanalyze Messner's data and compare the results with those from the classical approach.[1] Imitating Messner's steps in his study, I first address variable selection for 9 covariates, with 'Development Index' included and the five original correlated covariates excluded. Then I address variable selection for the same subset of covariates except that the representative variable 'Population less

[1]Due to the expensive computational demand of SSVS, it is only used in **3.3.3** of this chapter

than 15' is used instead of 'Development Index'. Finally, I carry out OW, MC^3 and SSVS when all the original covariates are allowed to enter the regression despite of the multicollinearity between some of them. All the variables have been standardized for MC^3 and SSVS.

3.3.1 Models with 9 covariates containing 'Development Index'

Table 3.2 reports the top 10 models selected by OW. These models are built on different subsets from the nine covariates (labeled in Table 3.4). The top one has only the Economic Discrimination variable included, with a posterior model probability of about 22.3%. The second model comprises $X1$ and $X9$ with a probability of 12.5% that it is the true model. The other models are all below 8.0%.

Table 3.3 presents the top 10 models with the highest posterior model probability selected by MC^3. The top one has only the Economic Discrimination variable included, with a posterior model probability of about 28.5%. The second model is surprisingly the null model, with about a 22.2% posterior model probability. The other models are all below 6.0%.

Table 3.4 compares the results from Messner's classical approach of variable selection and the Bayesian inference on $Pr(\beta_i \neq 0|D)$.[2] In the classical full model, which has all the nine covariates included, only 'Economic Discrimination' is significant at .05 level. The stepwise variable selection procedure also chooses only 'Economic Discrimination,' which is significant at .05 level[3]. The Bayesian approach presents $Pr(\beta_i \neq 0|D)$ for all the candidate variables. 'Economic Discrimination' is assigned

[2] $Pr(\beta_i \neq 0|D)$ is computed by summing the posterior probabilities across models for each predictor. If a predictor is present in a model, it will bear the posterior probability for that model. If not, it will have a zero probability in that particular model.

[3] The Efroymson stepwise procedure is used with an F value of 2 as criteria for adding variables to the subsets of explanatory variables. It is consistent for the other parts of **Chapter 3**.

Table 3.2: Nine-covariate Models with 'Development Index': Models from Occam's window with top 10 posterior model probabilities

Models									$Pr(M_i\|D)\%$
1									22.3
1								9	12.5
1		3							7.8
									6.8
1	2								5.8
								9	4.2
1			4						4.0
1						7			3.8
1							8		3.3
1					6				3.2

Table 3.3: Nine-covariate Models with 'Development Index': Models from MC^3 with top 10 posterior model probabilities

Models									$Pr(M_i\|D)\%$
1									28.5
									22.2
1								9	5.9
								9	5.1
1		3							3.6
	2								3.1
		3							3.0
1	2								2.7
1			4						1.8
1						7			1.7

Table 3.4: Nine-covariate Models with 'Development Index': Comparing classical and Bayesian Inference

Variables	Full Model	Stepwise	Occam's Window	MC^3
X_1 Eco. Discrimination Dummy	**	**	82.5	54.8
X_2 Income Inequality	N		11.8	9.3
X_3 Percent Urban Pop.	N		16.1	10.6
X_4 Ethno-linguistic Hetero. Index	N		7.1	6.1
X_5 Population (ln)	N		6.0	4.8
X_6 Population Density (ln)	N		4.9	5.0
X_7 Political Democracy Index	N		5.6	5.2
X_8 Percent Male 15-29	N		6.5	5.0
X_9 Development Index	N		29.5	16.6

N : not significant;
$*p \leq .10, **p \leq .05$, two-tailed
For Occam's window (only first principle is used) and MC^3, $Pr(\beta_i \neq 0|D)$ is computed by summing the posterior model probabilities over all the models visited.

the highest probabilities of not equal to zero, 82.5% and 54.8% by OW and MC^3 respectively. The second important variable is 'Development Index,' which bears 29.5% (OW) and 16.6%(MC^3). Therefore, we observe that both the classical and Bayesian approaches provide similar information. However, taking into account all the potential models, $Pr(\beta \neq 0|D)$ is a more reasonable way of identifying important variables than the P values for covariates based on a single model.

3.3.2 Models with 9 covariates containing 'Population less than 15'

Messner (1989) noted that 'Population less than 15' had the highest loading when he conducted the principle factor analysis over the five highly correlated variables. Therefore, he decided to try replacing 'Development Index' with 'Population less than 15' and double check the model. Here I replicate his research with OW and MC^3. Results of top 10 models selected are reported in Table 3.5 and 3.6, which are

Table 3.5: Nine-covariate Models with 'Population less than 15': Models from Occam's window with top 10 posterior model probabilities

Models									$Pr(M_i\|D)\%$
1								9	16.9
1									16.9
								9	8.5
1			3						5.9
									5.1
1	2								4.4
1							8	9	3.4
1				4					3.0
1						7			2.8
1						7		9	2.6

similar.

Again, in Table 3.7, the classical approach and Bayesian approach provide consistent information. The classical full model still only finds 'Economic Discrimination' significant at .05 level. The stepwise selection procedure chooses both 'Economic Discrimination' and 'Population less than 15,' significant at .05 and .10 level respectively. The Bayesian methods also identify 'Economic Discrimination' and 'Population less than 15' as covariates with the top two $Pr(\beta \neq 0|D)$. However, for either OW or MC^3, $Pr(\beta \neq 0|D)$ for 'Population less than 15' in Table 3.7 is about 20% larger than that for 'Development Index' in Table 3.4. It indicates that the former is a variable with a larger effect.

3.3.3 Models with all the 13 covariates

It is not surprising that in **2.3.1** and **2.3.2**, the classical approach and the Bayesian approach tend to identify similar important variables since the covariates are less correlated with each other and there is less uncertainty in the model. I will now ex-

Table 3.6: Nine-covariate Models with 'Population less than 15': Models from MC^3 with top 10 posterior model probabilities

Models									$Pr(M_i\|D)\%$
1									23.5
									18.3
							9		11.6
1							9		9.0
1		3							3.0
	2								2.5
		3							2.5
1	2								2.2
1			4						1.5
1					7				1.4

Table 3.7: Nine-covariate Models with 'Population less than 15': Comparing classical and Bayesian Inference

Variables	Full Model	Stepwise	Occam's Window	MC^3
X_1 Eco. Discrimination Dummy	**	**	78.0	51.4
X_2 Income Inequality	N		9.9	8.6
X_3 Percent Urban Pop.	N		13.3	9.6
X_4 Ethno-linguistic Hetero. Index	N		6.4	5.8
X_5 Population (ln)	N		5.6	4.7
X_6 Population Density (ln)	N		4.8	4.9
X_7 Political Democracy Index	N		7.6	6.1
X_8 Percent Male 15-29	N		8.0	5.3
X_9 Population less than 15	N	*	46.8	31.2

N : not significant;
$*p \leq .10, **p \leq .05$, two-tailed
For Occam's window (only first principle is used) and MC^3, $Pr(\beta_i \neq 0|D)$ is computed by summing the posterior model probabilities over all the models visited.

Table 3.8: Models with 13 Covariates: Models from Occam's Window with top 10 posterior model probabilities

Models													$Pr(M_i\|D)\%$
1			4										13.0
1													12.9
			4										6.5
1				5									6.0
1						6							4.6
1								7					4.5
1											13		4.2
													3.9
1			3										3.5
1	2												3.4

periment with models with all the 13 original covariates and see how the Bayesian and the classical approach will differ when there is much multicollinearity and uncertainty about the model.

Table 3.8 and 3.9 report similar findings of OW and MC^3 as to the top 10 selected models. Table 3.10 reports the findings from SSVS, which are similar to those of OW and MC^3 except that SSVS tends to give X_4 more importance. Table 3.11 reveals larger differences between the classical and Bayesian approaches in the results of variable identification. For classical approach, the full model finds no significant covariates at all. The stepwise method identifies X_1 and X_4. However, taking into account all the potential models, the Bayesian approach does not only still identify the 'Economic Discrimination' and 'Population less than 15' as the top two variables with largest effect, but also assigns various posterior non-zero probabilities to all the other covariates considered despite the multicollinearity.

Table 3.9: Models with 13 Covariates: Models from MC^3 with top 10 posterior model probabilities

Models													$Pr(M_i\|D)\%$
1													14.5
													11.3
			4										7.2
1			4										5.5
1				5					11	12	13		2.5
				5									2.5
1					6								1.9
1						7							1.9
					6								1.8
1											13		1.7

Table 3.10: Models with 13 Covariates: Models from SSVS with top 10 posterior model probabilities

Models													$Pr(M_i\|D)\%$
1				4									17.1
1				4					12				6.7
1				4			7						6.0
1				4	5								3.7
1			3	4									2.5
1				4					11				2.4
1	2			4									2.3
1				4							13		2.0
1				4			7		12				2.0
1				4		6							1.9

Table 3.11: Models with 13 covariates: Comparing classical and Bayesian Inference

Variables	Full Model	Stepwise	SSVS	MC^3	Occam's window
X_1 Eco. Discrimination Dummy	N	**	100.0	49.4	77.6
X_2 Income Inequality	N		12.0	7.6	4.8
X_3 Infant Mortality Rate	N		15.4	10.2	8.8
X_4 Population less than 15	N	*	99.8	32.7	42.7
X_5 GNP/capita(ln)	N		18.0	12.3	10.1
X_6 Annual Rate of Pop.Increase	N		10.7	12.7	9.4
X_7 Percent Urban Pop.	N		26.3	8.6	7.9
X_8 Ethno-linguistic Hetero. Index	N		9.8	5.4	4.9
X_9 Population (ln)	N		9.6	4.7	1.8
X_{10} Population Density (ln)	N		9.6	4.7	1.8
X_{11} Political Democracy Index	N		12.4	5.9	4.2
X_{12} Percent Male 15-29	N		28.6	5.1	2.6
X_{13} Life Expectancy	N		10.9	10.0	7.3

N : not significant;
$*p \leq .10, **p \leq .05$, two-tailed
For SSVS, MC^3, and Occam's window, $Pr(\beta_i \neq 0|D)$ is computed by summing the posterior model probabilities over all the models visited.

3.4 Discussion

Based on the foregoing analyses, I summarize my observations on the following aspects:

1) Bayesian methods support Messner's conclusion that 'Economic Discrimination Dummy' variable is more important than 'Income Inequality'. With the highest $Pr(\beta \neq 0|D)$, 'Economic Discrimination' stays at the top of all the X's for different methods and is always above 49%, while that for 'Income Inequality never exceeds 13%;

2) Bayesian methods support Messner's choice of 'Population less than 15' as the representative variable for the 5 highly correlated covariates. It is a regressor with larger $Pr(\beta \neq 0|D)$, and hence a larger effect than that of 'Development Index';

3) For this research, most of the time, Bayesian methods provide findings consistent with those of classical approach in various situations. However, there are two advantages for using Bayesian methods. First, when there is much multicollinearity (Table 3.11), the full model finds nothing due to inflated standard errors, while Bayesian methods still identify the important variables. Second, although stepwise regression seems to perform as well as Bayesian methods, it does not take into account model uncertainty and ignores covariates excluded, while Bayesian methods provide more complete and meaningful posterior information about every candidate covariates by summarizing all the models considered.

4) Although most of the time, OW, MC^3 and SSVS provide similar result, there are still some differences on $Pr(M_i|D)$ for models and $Pr(\beta \neq 0|D)$ for covariates. Those differences can be attributed to different priors.

3.5 Implementation Notes

3.5.1 Computer Programs

For OW, I used BICREG software developed by Adrian E. Raftery and Chris Volinsky. It is available at http://lib.stat.cmu.edu/S/bicreg

For MC^3, I used BMA software developed by Jennifer Hoeting. It is available at http://lib.stat.cmu.edu/S/bma. I conducted 30,000 iterations.

For SSVS, I wrote the Gibbs sampler code in S-PLUS. I started from the full model and conducted 10,000 iterations.

3.5.2 Convergence Check for SSVS

To check the convergence of SSVS, I input the Markov Chain results into CODA, a software for convergence diagnostics. Outputs demonstrate that convergence is achieved. Figure 3.1 and 3.2 are only two of them.

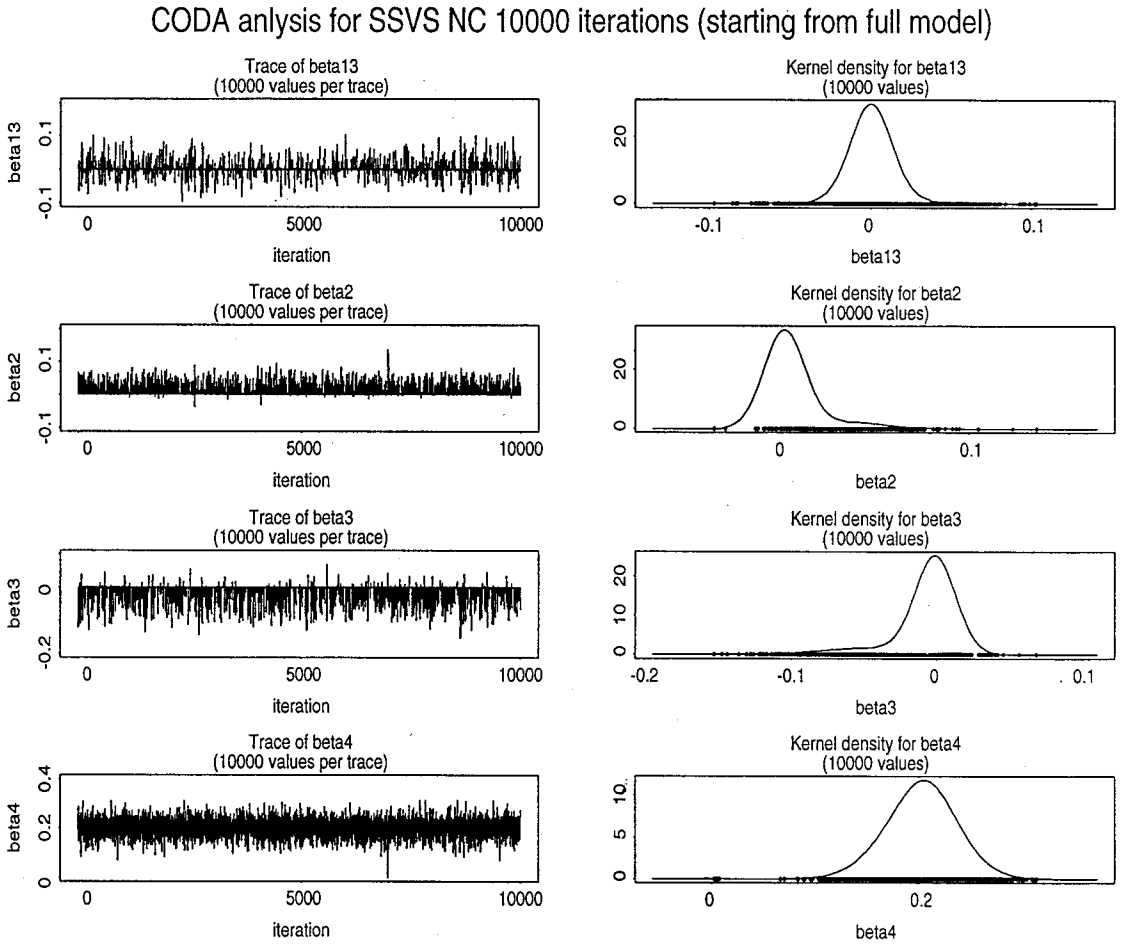

Figure 3.1: Trace plots from SSVS for $\beta_2, \beta_3, \beta_4$ and β_{13}

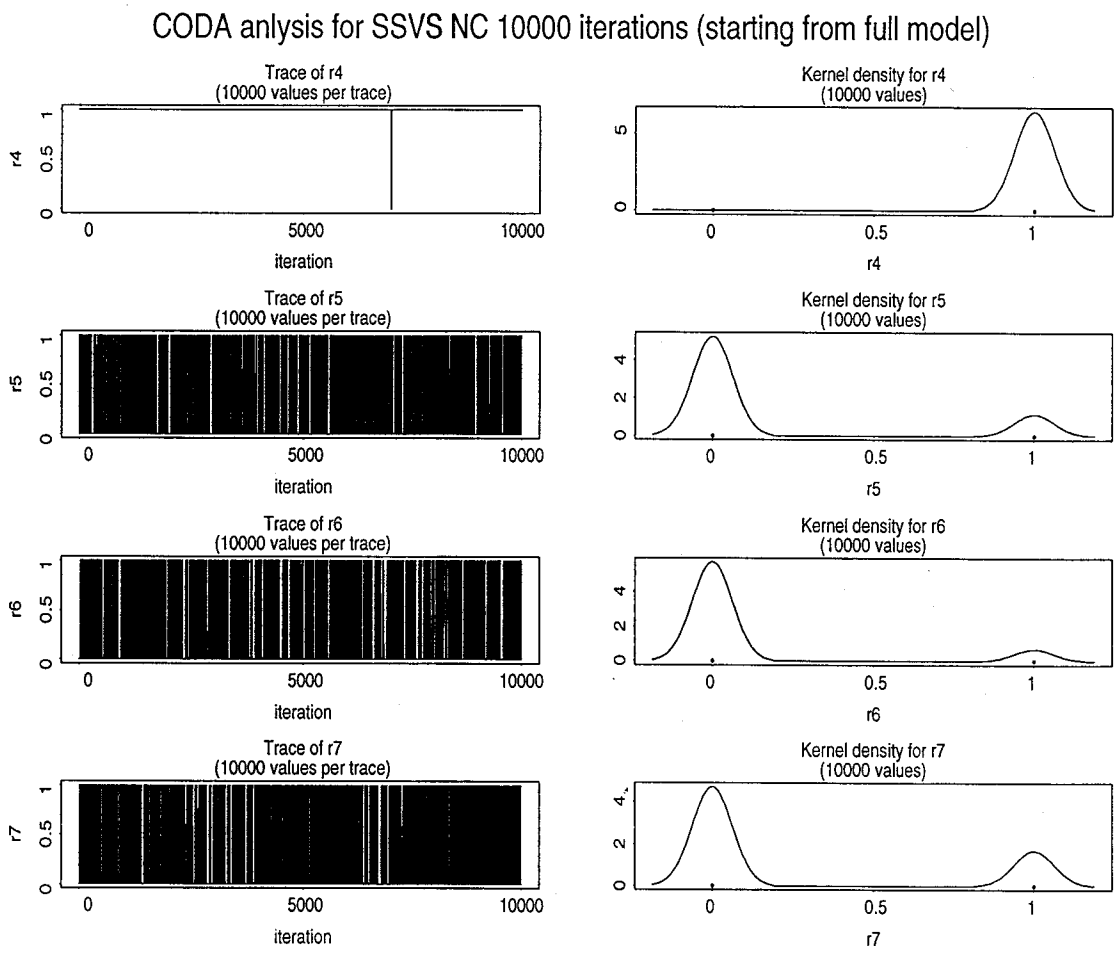

Figure 3.2: Trace plots from SSVS for $\gamma_4, \gamma_5, \gamma_6$ and γ_7

Chapter 4

Bayesian Variable Selection in Binary Response Models: A Micro-Level Drug Treatment Analysis

4.1 Question and Data

Since the early 1990s, the Federal Bureau of Prison has started a large-scaled treatment of drug addictions for inmates in federal correctional institutions across the country. Data on treatment outcome and socio-demographic variables have been collected. With millions of dollars spent on the treatment program, program evaluation appears to be more and more important. Interest has been particularly focused on assessment of treatment completion and dropouts. In this section, I intend to find a subset of variables by OW and SSVS to **predict** whether a subject is likely to complete or drop out the treatment program.

The original data set has 1,414 cases and 63 variables. After deleting missing cases and recode variables, I got a data set with 821 cases and 18 variables. The variables are as follows:

RSNEND: 1 - complete program; 0 - drop out voluntarily or with discipline discharge

ACTION: Score (1-5) of psychological readiness to change addiction behavior by tak-

ing action

AGESUB: Age of subject

CONTEM: Score (1-5) of psychological readiness to change addiction behavior by contemplating and planning action

DEPLOGTA: Estimated log odds of dependency on alcohol

DEPLOGTD: Estimated log odds of dependency on drugs

DRUG_FRQ: Frequency of drug use (scale 1-5)

DSMASP: Anti-Social Personality

DSMDEP: Depression (1 - yes; 0 - no)

EVFUTIME: Ever working full-time(1 - yes; 0 - no)

FEMALE: Being female (1 - yes; 0 - no)

GRADEA: Level of education

HSPANIC1: Hispanic ethnicity (1 - yes; 0 - no)

MAINTE: Score (1-5) of psychological readiness to change addiction behavior by maintaining progress achieved

NONWITE: Being nonwhite(1 - yes; 0 - no)

PASTETOH: Received past inpatient or outpatient alcohol treatment (1 - yes; 0 - no)

PRECON: Score (1-5) of psychological readiness to change addiction behavior in the most primitive stage

PRIOR1: Prior commitments (1 - yes; 0 - no)

VIOL1: History of violence(1 - yes; 0 - no)

4.2 SSVS in Probit Model

With binary response data, a SSVS approach for probit models is proposed by George, McCulloch and Tsay (1996). They pointed out that in SSVS for generalized linear models, the posteriors can be obtained from

$$\beta|\gamma, \phi, y; \quad \phi|\beta, y; \quad \gamma_j|\gamma_{-j}, \beta$$

where ϕ is some dispersion parameter.

In the probit model, $Y_i \in \{0, 1\}$, and $\Pr(Y_i = 1|x_i, \beta) = \Phi(x_i\beta)$, where Φ is the standard normal cumulative distribution function. With $\phi = 1$, only $\beta|\gamma, y$ and $\gamma_j|\gamma_{-j}, \beta$ need to be obtained by Gibbs sampler. The priors $p(\beta|\gamma)$ and $p(\gamma)$ are set up exactly the same as in (2.10) and (2.11). The full conditional distribution of γ is specified as in (2.15). The remaining question is: how to obtain the full conditional distribution of β?

4.2.1 Metroplis Algorithm

George, McCulloch and Tsay (1996) proposed the Metropolis algorithm to obtain the posterior of β. The full conditional distribution of β can be written as:

$$p(\beta|\gamma, Y) \propto p(\beta|\gamma)p(Y|\beta, \gamma) = p(\beta|\gamma) \prod_{i=1}^{n} \{[\Phi(x_i\beta)]^{Y_i}[1 - \Phi(x_i\beta)]^{(1-Y_i)}\} \qquad (4.1)$$

Since the calculation is analytically intractable, the Metroplis algorithm with a candidate proposal distribution $q(.|\beta)$ simulates $\beta^{(1)}, ..., \beta^{(j)}$ as follows:

1) Sample β^* from $q(.|\beta^{(j)})$.

2) Set $\beta^{(j+1)} = \beta^*$ with acceptance probability

$$\alpha = min\left\{1, \frac{p(\beta^*|\gamma, Y)}{p(\beta^{(j)}|\gamma, Y)}\right\} \tag{4.2}$$

Otherwise, set $\beta^{(j+1)} = \beta^{(j)}$

If the candidate distribution $q(.|\beta^{(j)})$ is chosen such that the Markov chain is both irreducible and aperiodic, then the convergence to $p(\beta|\gamma, Y)$ is guaranteed.

4.2.2 A Latent Variable Approach

Here I propose a simpler method by simulating a latent variable Z and transform the issue of variable selection back to SSVS in normal linear model[1].

First, I sample Z_i from the following distribution

$$p(Z_i|\beta, Y_i) = N(X_i\beta, 1), \qquad with \quad \begin{cases} Z_i > 0 \text{ if } Y_i = 1 \\ Z_i \leq 0 \text{ if } Y_i = 0 \end{cases}$$

Then the full conditional distribution of β is specified as in (2.13) with Z replacing Y

$$p(\beta|\sigma^2, \gamma, Z) = N_p((X^TX + \sigma^2 D_\gamma^{-2})^{-1}X^TZ, \sigma^2(X^TX + \sigma^2 D_\gamma^{-2})^{-1}) \tag{4.3}$$

To solicit reasonable priors on τ, I chose $\delta_i = .1/\delta x_j$ and $c = 10$, where $\delta x_j = x_{j,.75} - x_{j,.25}$ (the interquartile range) as suggested in George, McCulloch and Tsay (1996). Since SSVS is computationally demanding in time, I randomly sampled 250 cases from the 821 cases for the Gibbs sampler.

Table 4.1: Models from Occam's window with top 10 posterior model probabilities

Models						$Pr(M_i\|D)\%$
2			10			26.4
2			10		18	18.4
2			10		17	12.6
2	7		10			11.6
2		8	10			5.7
2	7		10		18	4.2
2		8	10		18	3.6
2		8	10	17		3.0
2	7		10	17		2.4
2			10	13		2.3

4.3 Variable Selection and Model Prediction

Table 4.1 reports the top ten models selected by OW.[2] The top model comprise X_2 and X_{10}, with a probability of 26.4% for being the true model. All the other models comprise X_2 and X_{10}.

Table 4.2 reports the findings on variable identification by both classical and Bayesian methods. The result of full model is not reliable since multicollinearity has affected some standard errors and thus P-values. Stepwise probit regression finds three variables X_2, X_{10}, and X_{18}. SSVS and OW share five out of the top six X's regarding $Pr(\beta_i \neq 0|D)$. SSVS has $X_2, X_7, X_8, X_{10}, X_{14}$ and X_{17}, while OW has $X_2, X_7, X_8, X_{10}, X_{17}$ and X_{18}.

I used the subsets of six variables selected by SSVS and by OW, respectively, to

[1]Thank Dr. Peter Mueller for his suggestion on absorbing the data augmentation idea in Albert and Chib (1990)

[2]Due to the large number of possible models (2^{18}), my SSVS of 2,2000 iterations did not give me expected result for top 10 models selected. The models obtained have quite small probabilities. It is possible that convergence has not been fully achieved. It is subject to more iterations and further analysis. However, SSVS and OW bear similar results in Table 4.2 about $Pr(\beta \neq 0|D)$.

Table 4.2: Comparing classical and Bayesian Inference on the covariates

Variables	Full Model	Stepwise	SSVS	Occam's window
X_1 ACTION	N		14.6	0.0
X_2 AGESUB	**	**	49.1	100.0
X_3 CONTEM	N		14.3	0.0
X_4 DEPLOGTA	N		16.4	0.0
X_5 DEPLOGTD	N		10.9	1.6
X_6 DRUG_FRQ	N		11.5	0.0
X_7 DSMASP	N		51.0	19.7
X_8 DSMDEP	*		78.1	13.8
X_9 EVFUTIME	N		34.4	0.0
X_{10} FEMALE	**	**	85.1	100.0
X_{11} GRADEA	N		11.5	1.4
X_{12} HSPANIC1	N		20.6	0.0
X_{13} MAINTE	N		19.1	3.7
X_{14} NONWHITE	N		54.5	2.2
X_{15} PASTETOH	N		31.0	0.0
X_{16} PRECON	N		14.1	0.0
X_{17} PRIOR1	N		65.8	19.6
X_{18} VIOL1	N	**	27.2	29.3

N : not significant;
$*p \leq .10, **p \leq .05$, two-tailed
For SSVS and Occam's window, $Pr(\beta_i \neq 0|D)$ is computed by summing the posterior model probabilities over all the models visited.

perform model validation.[3] The procedures are as follows:

1) Randomly split the 821 cases into half, each with same proportions of completion and dropout as the parent data set;

2) Use first half data set, regress Y on a subset of X's with a probit link function, obtain coefficient estimates;

3) Use those coefficient estimates and the same subset of X's in second half data set to compute $\hat{P}$, the predicted probability of completion;

4) Compute the mean values for $\hat{P}$ for those who actually completed and dropped out in the second data set, respectively, and further calculate d, the difference between the two means;

5) Repeat step (1) - (4) for 1,000 times, and obtain the univariate summary for the d's.

Figure 4.1 reports the histograms for the two sets of margins. By both visual inspection and a two-sample t-test, no significant departure from the two sets of margins was found.

4.4 Discussion

I summarize my observations in this chapter as follows:

1) Several variables are identified by Bayesian methods to be predictive about treatment completion: age of subject (X_2), anti-social personality (X_7), depression (X_8), female (X_{10}), prior commitments (X_{17}), and history of violence (X_{18}) or being nonwhite (X_{14}).

2) The model chosen by stepwise probit regression (X_2, X_{10}, and X_{18}) is merely the second model in the top 10 models chosen by OW, with a 18.4% probability of being the true model. OW shows that the top model consists of X_2 and X_{10} with

[3]This model-validation approach is suggested by Dr. Don Berry.

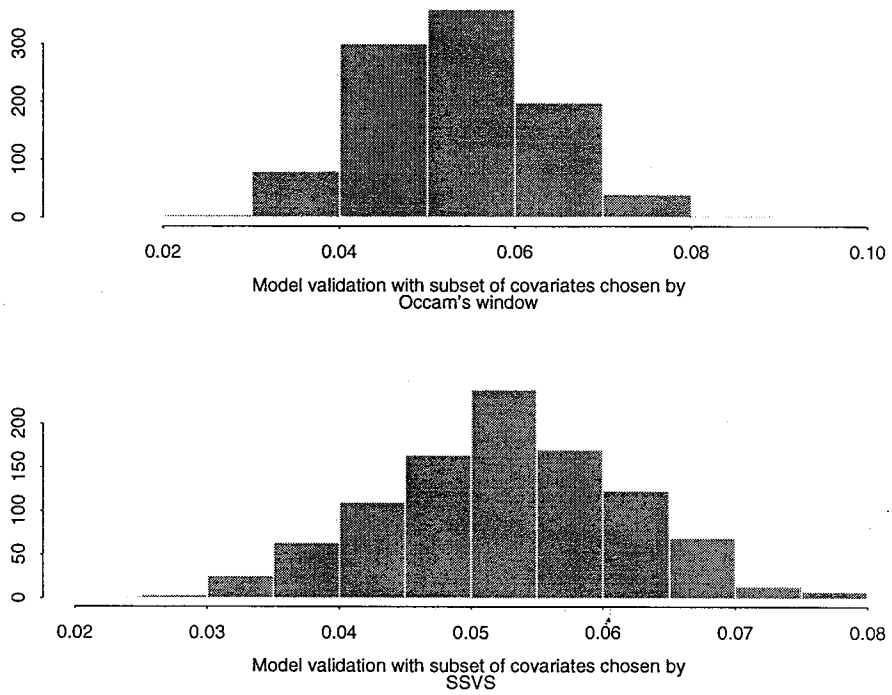

Figure 4.1: Histograms for 1,000 margins for model validation

a $Pr(M_i|D)$ of 26.4%. Therefore, We are far less willing to say that the model of X_2, X_{10}, and X_{18} is the true model.

3) SSVS and OW provide similar results as to $Pr(\beta_i \neq 0|D)$ except for X_{14}. The differences can be attributed to different setup of priors.

4.5 Implementation Notes

4.5.1 Computer Programs

For OW, I used BIC.GLM software developed by Chris Volinsky. It is available at http://www.stat.washington.edu/volinsky/software/bic.glm

For SSVS, I wrote the Gibbs sampler code in MATLAB. I started from the full model, and conducted 22,000 iterations, with 20,000 burn-ins and 2,000 for analysis.

4.5.2 Convergence Check for SSVS

To check the convergence of SSVS, I input the Markov Chain results into CODA. It is unclear from the outputs whether the convergence is achieved or not since the density of some of the chains is bimodal. Figure 4.1 and 4.2 are only two of them.

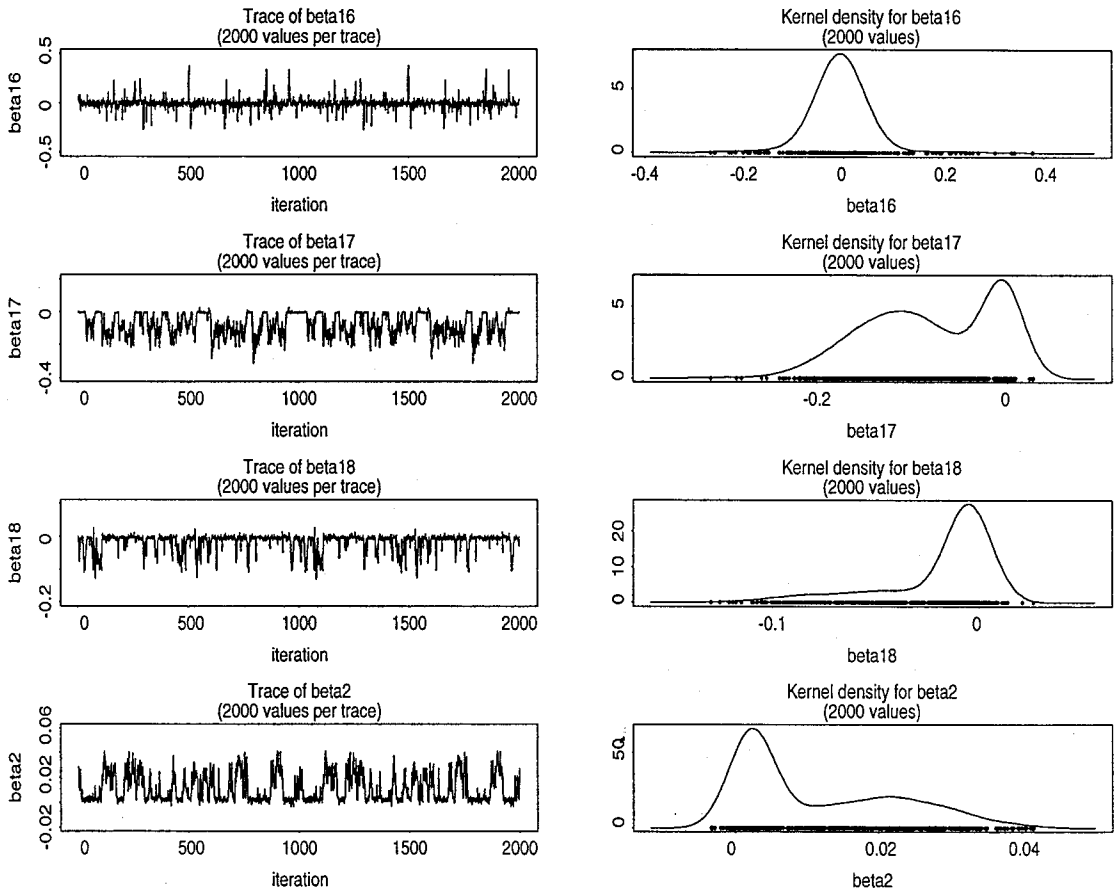

Figure 4.2: Trace plots from SSVS for $\beta_{16}, \beta_{17}, \beta_{18}$ and β_2

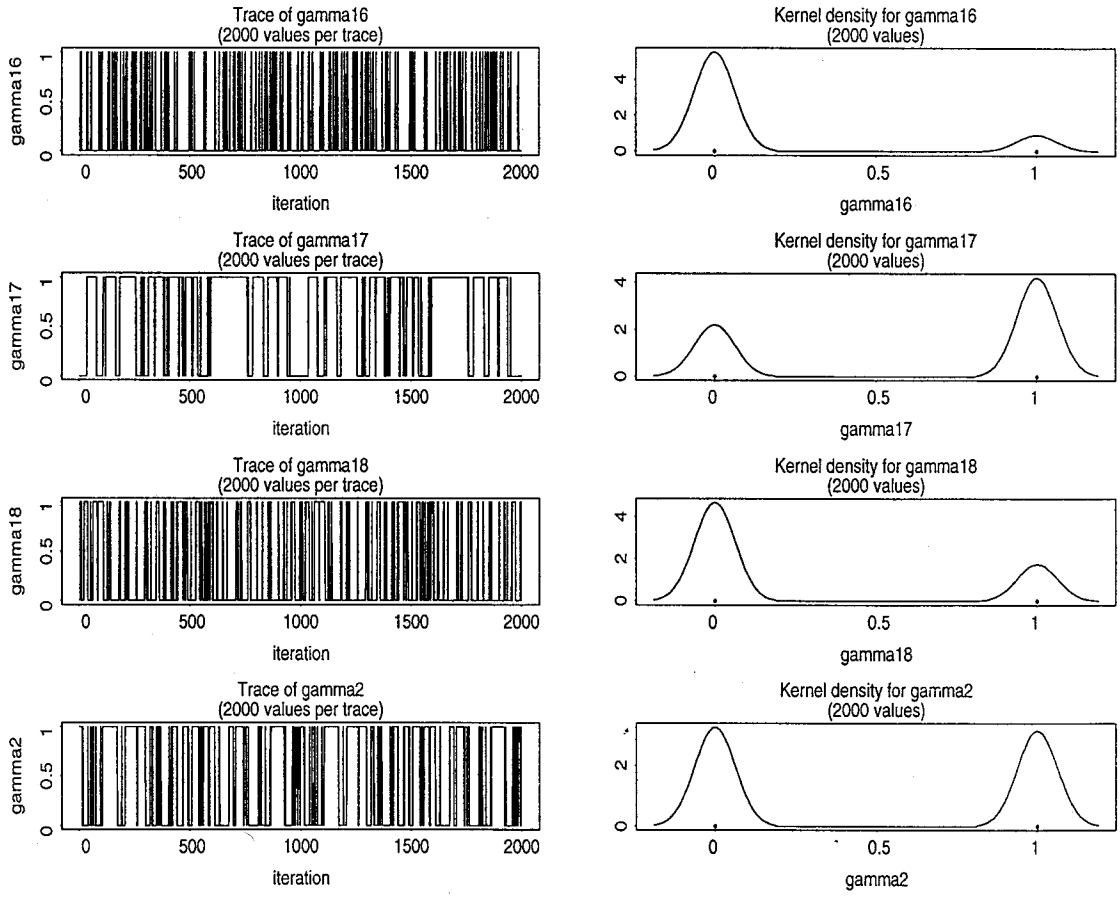

Figure 4.3: Trace plots from SSVS for $\gamma_{16}, \gamma_{17}, \gamma_{18}$ and γ_2

Chapter 5

Discussions and Conclusions

The applications of Bayesian variable selection methods in this thesis suggest that

1) Bayesian methods are more reasonable and reliable than classical methods by taking into account model uncertainty. Assuming that there is only one true model in the world, the classical approach only considers one single model, and loses information on the effect of the other covariates excluded. On the contrary, Bayesian methods assume that every possible model has a probability to be true (i.e. $0 \leq Pr(M_i|D) \leq 1$) and consider a much larger number of models to take into account model uncertainty.

2) $Pr(\beta_i \neq 0|D)$ is a much more meaningful way than P-values to identify important variables either for theoretical or practical purpose. As our quantity of interest throughout this thesis, $Pr(\beta_i \neq 0|D)$ is an average with the posterior model probabilities for all the models considered. Although most of the comparison between stepwise regression and Bayesian methods bear similar results, we can see that there are cases when stepwise regression ignores variables with fair importance. By either including or excluding a variable, there is no room in stepwise regression for a variable to have a probability of being in the model somewhere between 0 and 1. This inadequacy, in certain situations, can lead to incomplete or inaccurate inferences on

variable effect.

3) Although all the Bayesian analyses in this thesis used generic priors, Bayesian variable selection methods open a door for sociologists to incorporate prior knowledge about the importance of variables. For example, in SSVS, we can alter the prior values for certain γ_i's according to our prior knowledge about $Pr(\beta_j \neq 0)$ instead of setting every γ_i to .5.

4) The latent variable approach I proposed in **4.2.2** for SSVS in probit regression model takes advantage of the simplicity inherited from SSVS in normal linear model. It avoids the Metropolis algorithm, which highly depends upon the selection of a proper proposal distribution and is computationally costly.

To make Bayesian variable selection methods more available for social scientists, certain future work must be done:

1) More comparison between classical and Bayesian variable selection methods has to be done. Therefore, certain statistical situations should be specified where Bayesian methods would perform better than classical methods.

2) Simulated data should be used to evaluate the validity across OW, MC^3 and SSVS so that we can know which direction each method tends to go. In that way, we can probably know why there are differences in the outcomes for top models selected and for $Pr(\beta_i \neq 0|D)$.

3) Although the latent-variable SSVS method for probit model is simpler in implementation than that of George, McCulloch and Tsay (1996), it is unclear which one is more efficient in terms of convergence of the simulations. More work needs to be done to compare and evaluate them.

Bibliography

[1] James Albert and Siddhartha Chib. Bayesian Regression Analysis of Binary Data. October, 1990.

[2] Peter M. Blau and Joseph E. Schwarz. *Crosscutting Social Circles.* Academic, Orland, FL, 1984.

[3] Steven E. Fienberg and William M.M. Mason. Identification and Estimation of Age-Period-Cohort Effects in the Analysis of Discrete Archival Data. *Sociological Methodology 1979* 1-67, 1979

[4] David A. Freedman A Note on Screening Regression Equations. *The American Statistician,*37, No. 2, 152-155, 1983

[5] Edward I. George and Robert E. McCulluch. Variable Selection via Gibbs Sampling. *Journal of American Statistical Association,*85,398-409, 1993

[6] Edward I. George and Robert E. McCulluch. Stochastic Search Variable Selection. *Markov Chain Monte Carlo in Practice.* (ch 12). by Gilks, Richardson, and Spiegelhalter, 1996

[7] Edward I. George, Robert E. McCulluch, and Ruey S. Tsay *Two Approaches to Bayesian Model Selection with Applications* Bayesian Analysis in Statistics and Econometrics, Edited by Donald A. Berry, Kathryn M. Chaloner, and John K. Geweke. John Wiley & Sons, 1996

[8] David B. Grusky and Robert M. Hauser Comparative Social Mobility Revisited: Models of Convergence and divergence in 16 Countries. *American Sociological Review,* (49):19-38, 1984

[9] Jennifer A Hoeting. *Accounting for Model Uncertainty in Linear Regression.* PhD thesis, University of Washington, 1995.

[10] Michael Hout. *Mobility Tables.* (3rd ed.), Oxford University Press, 1983.

[11] Michael Hout. Status, Autonomy and Training in Occupational Mobility. *American Journal of Sociology,*(93): 1358-1400, 1984

[12] Robert E. Kass and Adrian E. Raftery Bayes Factors. 1995, *Journal of the American Statistical Association.* to appear

[13] E E. Leamer. *Specification Searches.* New York: Wiley, 1978

[14] David Madigan and A. E. Raftery. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of American Statistical Association*, (89), 1535-1546, 1994

[15] David Madigan and Jeremy York. Bayesian graphical models for discrete data. *International Statistical Review*, (63):215–232, 1995.

[16] Steven Messner. Economic discrimination and societal homicide rates: Further evidence on the cost of inequality. *American Sociological Review*, (54):597–611, 1989.

[17] Adrian E. Raftery. Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. *Technical Report 255*, Department of Statistics, University of Washington, 1993.

[18] Adrian E. Raftery. Bayesian Model Selection in Social Research (with Discussion by Andrew Gelman and Donald B. Rubin, and Robet M. Hauser, and a Rejoinder). *Sociological Methodology 1995* edited by Peter V. Marsden Cambridge, Mass.: Blackwells, 1995

[19] Adrian E. Raftery, David Madigan, and Jennifer Hoeting Bayesian Model Averaging for Linear Regression Models *Journal of American Statistical Association.* March 1997, Vol. 92, No. 437, Theory and Methods.

[20] Charles Taylor and David A. Jodice. *World Handbook of Political and Social Indicator.* Yale University Press, New Haven, 3rd edition, 1983.