The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 33

Two-Level Stochastic Search Variable Selection in GLMs with Missing Predictors

Robin Mitra, *Southampton Statistical Sciences Research
Institute*

David Dunson, *Duke University*

Recommended Citation:

Mitra, Robin and Dunson, David (2010) "Two-Level Stochastic Search Variable Selection in GLMs with Missing Predictors," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 33.

DOI: 10.2202/1557-4679.1173

Two-Level Stochastic Search Variable Selection in GLMs with Missing Predictors

Robin Mitra and David Dunson

Abstract

Stochastic search variable selection (SSVS) algorithms provide an appealing and widely used approach for searching for good subsets of predictors while simultaneously estimating posterior model probabilities and model-averaged predictive distributions. This article proposes a two-level generalization of SSVS to account for missing predictors while accommodating uncertainty in the relationships between these predictors. Bayesian approaches for allowing predictors that are missing at random require a model on the joint distribution of the predictors. We show that predictive performance can be improved by allowing uncertainty in the specification of predictor relationships in this model. The methods are illustrated through simulation studies and analysis of an epidemiologic data set.

KEYWORDS: missing at random, model averaging, multiple imputation, stochastic search, subset, variable selection

Author Notes: We would like to thank three anonymous reviewers for their helpful comments and suggestions.

1 Introduction

In regression, one issue that is routinely encountered is how to select a subset of the available predictors that are important in explaining the response. In many fields, this variable selection problem is faced in essentially every study that is conducted. For example, in epidemiologic studies of exposure-disease relationships, investigators typically collect information on multiple potential risk factors and confounding variables. Clearly, problems can be encountered if all these variables are included as predictors, so epidemiologists tend to discard covariates that do not have a significant impact on disease risk, unless these covariates are the primary exposures of interest or there is strong prior knowledge that they should be included.

Stepwise selection is the most widely-used automated algorithm for selecting variables to include in a regression model, with many variants possible depending on the starting model, the manner in which variables are added or deleted, and the criteria for deciding whether a predictor significantly improves goodness-of-fit. For example, forward selection sequentially adds predictors, keeping those that improve the AIC, BIC, or have p-values in a likelihood ratio, Wald or score test below some pre-specified threshold. For generalized linear models (GLMs), the order in which variables are added and the criteria used can have a substantial impact on the final subset of variables that are selected. In addition, basing inferences on the model selected from a stepwise procedure without accounting for uncertainty in the selection process can lead to highly misleading results. For example, there will be a greatly inflated type I error rate and the parameter estimates will be biased away from zero, particularly if there are many candidate predictors.

A number of strategies have been proposed to address such problems, with the focus in this article on Bayesian model averaging approaches allowing for missing predictors. There have been frequentist approaches for model assessment and selection with missing data; methods have recently been proposed by Verbeke, Mohlenbergs, and Beunckens (2008) and Claeskens and Consentino (2008). However, these approaches typically require selection of a best fitting model to proceed with further analysis and when there is a large candidate set of predictors to choose from this may not be a reasonable or even a feasible strategy to implement; we seek to avoid this limitation through model averaging. In the Bayesian paradigm, one can assign posterior probabilities to each of the models in a list of *a priori* plausible models. To incorporate uncertainty in model selection, one can then average over models in the list using posterior probability weights in performing inferences and predictions. In terms of prediction, Bayesian model averaging has been shown to have a lower risk, measured by a logarithmic scoring rule, compared with using any single model (Raftery, Madigan, and Hoeting, 1997). For a recent review of Bayesian model averaging, refer to Clyde and George (2004).

In variable selection problems, the list of models under consideration corresponds to the 2^p possible subsets of a set of p candidate predictors. Clearly, the number of models rapidly becomes enormous as p increases, so there is a need for efficient methods for searching for high posterior probability models, while also estimating posterior model probabilities and the posterior distributions for the coefficients in each model. A widely used strategy for addressing this problem is to embed all the models in a full model containing all the predictors, and then allow predictors to drop out by choosing a mixture prior for the coefficients with one component concentrated at zero (Mitchell and Beauchamp, 1988). One can then use a Gibbs sampling algorithm for simultaneous model search and posterior computation, with such an approach referred to as stochastic search variable selection (SSVS) (George and McCulloch, 1993, 1997).

When missing values are present in the covariates SSVS algorithms cannot be applied directly. One commonly used strategy is to discard subjects with any missing predictors (complete case analysis), but this can be a sizeable proportion of the subjects in variable selection contexts, as one would need to discard subjects with missing values in any of the candidate predictors. Further, when missing patterns are not missing completely at random (MCAR) this approach can lead to biased inferences. Bayesian models can easily accommodate missing predictors by placing a joint model on the distribution of the predictors and then imputing the missing values within an MCMC algorithm. In the variable selection setting, with the response and predictors following a multivariate normal distribution, such an approach was implemented by Yang, Bellin, and Boscardin (2006). This article addresses a much broader class of models involving mixed categorical and continuous variables, while also allowing model selection for the predictor component. Claeskens and Hjort (2008, ch. 10) reviews both frequentist and Bayesian model selection approaches in the presence of missing values. These approaches are not typically concerned with model averaging, and the approaches do not also consider the problem of model uncertainty in the model for the missing data.

Outside of the variable and model selection context, a standard approach for specifying the joint distribution of the predictors, while allowing these predictors to have different measurement scales, is to choose a sequence of GLMs (see, for example, Lipsitz and Ibrahim (1996), Ibrahim, Lipsitz, and Chen (1999)). However, following such an approach one faces uncertainty in how to specify the GLMs for X_1 , X_2 given X_1 , X_3 given X_1, X_2 , etc. This is essentially another level of variable selection, so it is natural to allow uncertainty in this component of the model as well. This article proposes a two-level SSVS approach to allow uncertainty in the exact form of the imputation models for the missing covariate data. By allowing more parsimonious modeling of the joint predictor distribution through model averaging, we anticipate an improvement in predictive performance.

Section 2 briefly reviews the Bayes approach to model uncertainty in variable selection, and describes how to accommodate missing predictors in this paradigm. Section 3 presents the priors and models used to implement a two-level SSVS algorithm. Section 4 provides theoretical support for the approach. Section 5 illustrates performance of the method through simulation studies. Section 6 presents an application to an epidemiologic study, and Section 7 concludes with a discussion.

2 Two Level Variable Selection

2.1 Framework for Bayesian Variable Selection

Suppose data for subject i ($i = 1, \dots, n$) consist of a response y_i and a vector of candidate predictors, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$. Let $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)'$ denote a vector of predictor inclusion indicators, with $\gamma_j = 1$ denoting that the j^{th} element of $\mathbf{x}_i$ should be included in the regression model for the response and $\gamma_j = 0$ otherwise. Then, we focus on the case in which the conditional likelihood of y_i given $(\mathbf{x}_i, \boldsymbol{\gamma})$ belongs to an exponential family with scale parameter τ and location parameter $\mu_i = E(y_i | \mathbf{x}_i, \boldsymbol{\gamma})$, with $g(\mu_i) = \mathbf{x}_{\boldsymbol{\gamma}i}' \boldsymbol{\beta}_{\boldsymbol{\gamma}}$, where $\mathbf{x}_{\boldsymbol{\gamma}i} = \{x_{ij}, j : \gamma_j = 1\}$ is the subset of predictors included in the model indexed by $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\beta_{\gamma 1}, \dots, \beta_{\gamma p_{\boldsymbol{\gamma}}})'$ denotes the coefficients for model $\boldsymbol{\gamma}$, $p_{\boldsymbol{\gamma}} = \sum_{j=1}^p \gamma_j$ is the number of predictors in model $\boldsymbol{\gamma}$, and $g(\cdot)$ is a known link function.

Hence we have defined a typical variable selection problem in the setting of a generalized linear model (GLM). There are 2^p possible indicator vectors $\boldsymbol{\gamma}$, with the model space corresponding to these different possibilities denoted by Γ . A Bayesian formalization of the variable selection problem requires a prior for $\boldsymbol{\gamma}$ with support on Γ , as well as a prior on the coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ for each $\boldsymbol{\gamma} \in \Gamma$. The posterior probability allocated to model $\boldsymbol{\gamma}$ is then defined via Bayes rule as:

$$p(\boldsymbol{\gamma} | \mathbf{y}, \mathbf{X}) = \frac{p(\boldsymbol{\gamma}) p(\mathbf{y} | \mathbf{X}, \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}^* \in \Gamma} p(\boldsymbol{\gamma}^*) p(\mathbf{y} | \mathbf{X}, \boldsymbol{\gamma}^*)}, \quad (1)$$

where $p(\mathbf{y} | \mathbf{X}, \boldsymbol{\gamma}) = \int \prod_{i=1}^n p(y_i | \mathbf{x}_{\boldsymbol{\gamma}i}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau) d p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$ is the marginal likelihood of the data under model $\boldsymbol{\gamma}$, $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$ is the prior on the coefficients and scale parameter in model $\boldsymbol{\gamma}$, and $p(\boldsymbol{\gamma})$ is the prior probability of model $\boldsymbol{\gamma}$.

For linear regression models and conjugate priors, the marginal likelihood under each model is available in closed form and the main practical issues that arise are (1) how to choose $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau)$, noting that model selection is sensitive to this choice (Liang, Paulo, Molina, Clyde, and Berger, 2008); and (2) how to efficiently search the model space given that the number of subsets increases rapidly with p . For non-normal GLMs, the Laplace approximation to the marginal likelihood

can be used (Raftery, 1996). Cottet, Kohn, and Nott (2008) propose a framework for Bayesian variable selection in a wider class of double exponential regression models; however, we do not consider such models in this article. In addition Cottet et al. (2008) do not explicitly propose a framework for Bayesian variable selection in the presence of missing covariates. In the example with missing data in Cottet et al. (2008), the approach taken was to only consider the complete cases; such an approach could introduce bias into the analysis if the missing data are not MCAR.

2.2 Bayes Variable Selection with Missing Predictors

Now consider the common setting in which only a subset of predictors are observed. In particular let $\mathbf{m}_i = (m_{i1}, \dots, m_{ip})'$ denote a vector of missingness indicators specific to subject i with $m_{ij} = 1$ denoting that the j^{th} predictor is missing. In this setting, the approach described in section 2.1 cannot be applied directly.

Using the formulation described in Little and Rubin (2002), we consider the full marginal likelihood under the model γ : $p(\mathbf{y}, \mathbf{M} | \gamma, \phi, \mathbf{X}_{obs})$, where $\mathbf{M} = (\mathbf{m}_1, \dots, \mathbf{m}_n)'$ is the $n \times p$ matrix of missingness indicators for all subjects, ϕ are parameters characterizing the likelihood of the missingness indicators, $\mathbf{X}_{obs} = \{x_{ij}, i = 1, \dots, n, j : m_{ij} = 0\}$ are the observed predictor values, and $\mathbf{X}_{mis} = \{x_{ij}, i = 1, \dots, n, j : m_{ij} = 1\}$ are the missing predictor values. We express this joint likelihood for $\mathbf{y}$ and $\mathbf{M}$ given ϕ and the observed predictors in a selection model form as follows:

$$p(\mathbf{y}, \mathbf{M} | \gamma, \phi, \mathbf{X}_{obs}) = \int p(\mathbf{y} | \mathbf{X}, \gamma) p(\mathbf{M} | \phi, \mathbf{y}, \mathbf{X}) p(\mathbf{X}_{mis} | \mathbf{X}_{obs}, \gamma) d\mathbf{X}_{mis}.$$

When predictors are MAR, $p(\mathbf{M} | \phi, \mathbf{y}, \mathbf{X}) = p(\mathbf{M} | \phi, \mathbf{y}, \mathbf{X}_{obs})$. In addition, when the parameters governing the observed data likelihood and the missing data mechanism are distinct, in that the prior distributions on these parameters are independent, the missing data mechanism is ignorable and we can base inferences on the observed data likelihood,

$$p(\mathbf{y} | \mathbf{X}_{obs}, \gamma) = \int p(\mathbf{y} | \mathbf{X}_\gamma, \beta_\gamma, \tau) p(\mathbf{X}_{\gamma mis} | \mathbf{X}_{\gamma obs}) d\mathbf{X}_{\gamma mis} d p(\beta_\gamma, \tau), \quad (2)$$

where $\mathbf{X}_{\gamma obs} = \{x_{ij}, i = 1, \dots, n, j : (1 - m_{ij})\gamma_j = 1\}$ and $\mathbf{X}_{\gamma mis} = \{x_{ij}, i = 1, \dots, n, j : m_{ij}\gamma_j = 1\}$.

We treat the missing covariate data $\mathbf{X}_{\gamma mis}$ in model γ as nuisance parameters to be integrated out of the likelihood. In this way we can estimate the posterior probability of model γ using equation (1), but with the marginal likelihood defined conditionally on the observed data. In order for (2) to be well defined, we require a probability model for the joint distribution of the predictors, so that one can obtain

the conditional likelihood of $\mathbf{X}_{mis}$ given $\mathbf{X}_{obs}$. We initially describe such a model without allowing for uncertainty in the choice.

In particular, following common practice in the literature on missing predictors having mixed measurement scales, we use the factorization:

$$p(\mathbf{X}) = p(\mathbf{x}_1) \prod_{j=2}^p p(\mathbf{x}_j | \mathbf{x}_1, \dots, \mathbf{x}_{j-1}), \quad (3)$$

where $p(\mathbf{x}_j | \mathbf{x}_1, \dots, \mathbf{x}_{j-1}) = \int \prod_{i=1}^n p(x_{ij} | x_{i1}, \dots, x_{i,j-1}, \boldsymbol{\theta}_j, \kappa_j) d p(\boldsymbol{\theta}_j, \kappa_j)$ is characterized as a distribution in the exponential family with dependence on previous predictors modeled via a GLM with $\boldsymbol{\theta}_j, \kappa_j$ the regression coefficients and dispersion parameter, respectively, in the j th GLM in the sequence, with $p(\boldsymbol{\theta}_j, \kappa_j)$ the prior distribution. We also model $\mathbf{x}_1$ to be in the exponential family conditional on location and scale parameters $\boldsymbol{\theta}_1$ and κ_1 . Then denote $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j, j = 1, \dots, p\}$ and $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_p)'$ to be the set of regression coefficients and dispersion parameters characterizing the joint distribution of the predictors. An alternative modeling strategy to deal with missing predictors known as regression switching was proposed by van Buuren, Boshuizen, and Knook (1999). However, this approach does not explicitly specify a joint distribution for the predictors and thus would not be compatible within a fully Bayes model selection procedure, we do not explore this approach here.

One could run an MCMC algorithm to generate samples from the conditional distribution of $\mathbf{X}_{mis}$ given $\mathbf{X}_{obs}$, $\mathbf{y}$ and $\boldsymbol{\gamma}$. These samples could be used to fill in the missing predictors at each sampling step of an SSVS analysis that accounts for uncertainty in the predictors to be included in the response model. However, this approach would not allow uncertainty in specification of the models characterizing (3).

2.3 Variable Selection for the Missing Data Model

When the number of predictors is large, questions arise in specification of each of the regression models, $p(\mathbf{x}_j | \mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \boldsymbol{\theta}_j, \kappa_j)$. We are faced with essentially the same issues that motivate variable selection in our ‘top level’ model relating the response to the predictors; in particular, there could be sparse relationships between variables and so it may not be necessary to include all $j - 1$ predictors in our model for $\mathbf{x}_j$.

A natural extension is to perform variable selection within each of the conditional regression models characterizing the joint distribution of the predictors. We do this by defining inclusion indicators $\boldsymbol{\gamma}_j^m = (\gamma_{j1}^m, \dots, \gamma_{j,j-1}^m)$, where $\gamma_{jk}^m = 1$ indicates

that x_{ik} should be included in the regression model for x_{ij} and $\gamma_{jk}^m = 0$ otherwise. Thus the joint distribution of the predictors in model γ^m is

$$p(\mathbf{X}|\boldsymbol{\theta}_{\gamma^m}, \boldsymbol{\kappa}) = p(\mathbf{x}_1|\boldsymbol{\theta}_1, \boldsymbol{\kappa}_1) \prod_{j=2}^p p(\mathbf{x}_j|\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \boldsymbol{\theta}_{j\gamma_j^m}, \boldsymbol{\kappa}_j) \quad (4)$$

where $\boldsymbol{\theta}_{j\gamma_j^m} = (\theta_{j\gamma_j^m 1}, \dots, \theta_{j\gamma_j^m p_{\gamma_j^m}})'$ are the coefficients in the regression model for $\mathbf{x}_j$ in model γ_j^m with $p_{\gamma_j^m} = \sum_{k=1}^{j-1} \gamma_{jk}^m$, $\gamma^m = (\gamma_1^{m'}, \dots, \gamma_p^{m'})' \in \Gamma^m$ indexes the model characterizing the joint distribution of the predictors, and $\boldsymbol{\theta}_{\gamma^m} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}'_{2\gamma_2^m}, \dots, \boldsymbol{\theta}'_{p\gamma_p^m})'$.

Thus, the distribution of $\mathbf{x}_j$ is conditional on a subset of the predictors $(\mathbf{x}_1, \dots, \mathbf{x}_{j-1})$ defined by the inclusion indicators in γ_j^m . Dropping a predictor $\mathbf{x}_k$, $1 \leq k \leq j-1$ from the regression model for $\mathbf{x}_j$ implies independence between $\mathbf{x}_j$ and $\mathbf{x}_k$ conditional on the other predictors in the model and so we are able to incorporate parsimonious relationships between predictors. In the special case when $(\mathbf{x}_1, \dots, \mathbf{x}_j)$ have a multivariate normal distribution this corresponds to putting zeroes in the $(j, k)^{th}$ and $(k, j)^{th}$ entries of its precision matrix.

Therefore, we are performing variable selection on two levels, (1) in the top level model relating the response to predictors, and (2) in the model characterizing the joint distribution of the predictors. In Section 4, we present a theoretical argument implying improved predictive performance for two level variable selection compared with a one level approach that bypasses level (2) and assumes a particular choice of model γ_0^m nested in Γ^m . Note that in the two-level case, there are $2^{\frac{p(p+1)}{2}}$ possible models in the joint model space, $\Gamma \otimes \Gamma^m$. Hence, even for modest p , the number is enormous. In the next section, we propose a two-level SSVS algorithm, SSVS², which extends the one-level algorithm, SSVS¹.

3 Stochastic Search Variable Selection

The SSVS² algorithm described in this section focuses on the case in which $p(\mathbf{x}_j|\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \boldsymbol{\theta}_{j\gamma_j^m}, \boldsymbol{\kappa}_j)$ is a normal linear regression model for continuous $\mathbf{x}_j$ and is a probit regression model for categorical $\mathbf{x}_j$. We also assume a normal or probit form for $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}_{\gamma}, \boldsymbol{\tau})$. These special cases are convenient in facilitating use of data augmentation, as proposed by Albert and Chib (1993), to obtain closed forms for conditional model probabilities and posterior distributions. However, the algorithm described can be modified to allow other GLMs through the use of a Laplace approximation to marginal likelihoods used in calculating conditional model probabilities, with adaptive rejection sampling used for updating the parameters from their full conditional posterior distributions given the model.

As for previously-proposed SSVS¹ algorithms, the goal of SSVS² is to simultaneously accomplish several goals through the use of an MCMC algorithm that alternates between updating the model indicators and the parameters within the current model. By sampling from full conditional posterior distributions sequentially, the samples converge in distribution to a stationary distribution that is the joint posterior distribution of the model indicators and the parameters within each model. For enormous model spaces, such as the ones encountered in the two-level variable selection problem or the one-level case for moderate to large numbers of candidate predictors, it is not realistic to expect accurate estimates of the exact posterior model probabilities and posterior distributions based on the number of samples it is feasible to collect. Nonetheless, it has been observed that marginal posterior densities of the coefficients for each predictor, marginal inclusion probabilities and predictive distributions tend to be well estimated by SSVS algorithms even in challenging cases (George and McCulloch, 1997).

In Section 3.1 we complete a Bayesian specification of the model with explicit models for each component of the likelihood and with prior distributions for the parameters and model indicators. In Section 3.2 we outline the steps involved in the SSVS² algorithm.

3.1 Model and Prior Specification

We first model the top level which relates the p predictors x_i to the response y_i for each individual i under model γ . As we are considering y_i to be either continuous or categorical define $y_i = g_y(y_i^*, \xi_y)$, where

$$p(y_i^* | x_{i1}, \dots, x_{ip}, \beta, \gamma, \tau) = N(y_i^*; \beta_0 + x'_{\gamma i} \beta_\gamma, \tau), \quad (5)$$

$x_{\gamma i}$, β_γ as defined in section 2 and τ is the residual variance, with $\phi = \tau^{-1}$. When y_i is continuous g_y is the identity so that $y_i = g_y(y_i^*) = y_i^*$. With an ordered categorical response with $y_i \in \{1, \dots, c_y\}$, we set $\tau = 1$ and define $\xi_y = (\xi_{y0}, \xi_{y1}, \dots, \xi_{yc_y})'$ to represent the threshold parameters in a generalized probit model with $\xi_{y0} = -\infty$, $\xi_{yc_y} = \infty$, $\xi_{y1} = 0$, and $y_i = g_y(y_i^*, \xi_y) = \sum_{k=1}^{c_y} kI(\xi_{y,k-1} < y_i^* \leq \xi_{yk})$. In the special case that $c_y = 2$ $y_i = I(y_i^* > 0)$.

We embed all the models within one full model that includes main effects for all the predictors. To simultaneously specify a prior over the model space and for the coefficients within each model, we let

$$\beta_j \sim (1 - \pi_j)\delta_0 + \pi_j N(0; \phi_{\beta_j}^{-1}), \quad \phi_{\beta_j} \sim Ga(1/2, 1/2), \quad (6)$$

where δ_0 is a unit probability mass at zero, $\gamma_j = 1(\beta_j \neq 0)$, and π_j is the prior probability of including the j th predictor, with $\pi_j = 0.5$ if inclusion and exclusion

are equally likely. Predictors having zero coefficients are effectively excluded from the model, while a heavy-tailed Cauchy prior is induced through a scale mixture of Gaussians for the coefficients for the included predictors. The marginal Cauchy prior for β_j arises by integrating out the hyper-parameter ϕ_{β_j} with respect to its prior distribution. We place a Jeffreys prior on ϕ for a continuous response and a uniform improper prior for $\xi_y^* = (\xi_{y2}, \dots, \xi_{y,c_y-1})'$ on the restricted space $\Omega = \{\xi_y^* : 0 < \xi_{y2} < \xi_{y3} < \dots < \xi_{y,c_y-1}\} \subset R^{c_y-2}$ for an ordered categorical response.

Focusing now on the predictor component model and using the specification of Section 2.3, we let $x_{ij} = g_{x_j}(x_{ij}^*, \xi_{x_j})$ where,

$$p(x_{ij}^* | x_{i1}, \dots, x_{i,j-1}, \theta_j, \kappa_j, \gamma_j^m) = N(x_{ij}^*; \theta_{j0} + \mathbf{x}_{j\gamma_j^m}^* \theta_{j\gamma_j^m}, \kappa_j), \quad (7)$$

where $\mathbf{x}_{j\gamma_j^m}^* = (x_{j\gamma_j^m i1}^*, \dots, x_{j\gamma_j^m ip\gamma_j^m}^*)'$ are the predictors in model γ_j^m , $\theta_{j\gamma_j^m}$ are the coefficients for these predictors, and $\psi_j = \kappa_j^{-1}$. For continuous x_{ij} , g_{x_j} is the identity so $x_{ij} = g_{x_j}(x_{ij}^*) = x_{ij}^*$ and for an ordered categorical predictor $x_{ij} \in \{1, \dots, c_{x_j}\}$ set $\psi_j = 1$ and use threshold parameters $\xi_{x_j} = (\xi_{x_j0}, \xi_{x_j1}, \dots, \xi_{x_jc_{x_j}})'$ to model $x_{ij} = g_{x_j}(x_{ij}^*, \xi_{x_j}) = \sum_{k=1}^{c_{x_j}} kI(\xi_{x_j,k-1} < x_{ij}^* \leq \xi_{x_jk})$ where, $\xi_{x_j0} = -\infty$, $\xi_{x_jc_{x_j}} = \infty$ and $\xi_{x_j1} = 0$. For binary predictors, we let $x_{ij} = I(x_{ij}^* > 0)$.

Define $\kappa = \{\kappa_j, j : x_{ij} = x_{ij}^*\}$ to be the set of scale parameters in the joint distribution of the predictors and $\mathbf{X}^* = \{x_{ij}^*, i = 1, \dots, n, j : x_{ij} \neq x_{ij}^*\}$ to be the set of latent variables corresponding to categorical predictors in our data set. To complete a prior specification using a similar specification to (6), we let

$$\theta_{jk} \sim (1 - \pi_{jk})\delta_0 + \pi_{jk}N(0, \phi_{\theta_{jk}}^{-1}), \quad \phi_{\theta_{jk}} \sim Ga(1/2, 1/2), \quad (8)$$

for $j = 1, \dots, p, k = 0, \dots, j-1$, where $\pi_{jk} = 0.5$ as a default, $\gamma_{jk}^m = 1(\theta_{jk} \neq 0)$, and $\phi_{\theta_{\gamma_j^m}} = \{\phi_{\theta_{jk}}, (j, k) : \gamma_{jk} = 1\}$. In the SSVS¹ approach we do not perform SSVS on the missing data model, instead we put Jeffreys priors on all regression coefficients and intercepts so that $p(\theta_{jk}) \propto 1$ $j = 1, \dots, p, k = 0, \dots, j-1$, this implicitly assumes that $\gamma_{jk}^m = 1$ for all j, k . In both approaches SSVS² and SSVS¹ we can again place Jeffreys priors for any residual variances in the regression models and improper uniform priors on the restricted support of the threshold parameters for each categorical predictor.

3.2 Posterior Computation

We now outline the basic steps of the SSVS² algorithm, focusing for simplicity on the case in which the response is binary and the predictors are binary or continuous. SSVS² proceeds by sampling from the joint posterior of the model space

(γ, γ^m) , parameters within each model $(\beta_\gamma, \phi_{\beta_\gamma}, \theta_{\gamma^m}, \phi_{\theta_{\gamma^m}}, \kappa)$, and the latent variables $(\mathbf{y}^*, \mathbf{X}^*, \mathbf{X}_{mis})$ conditional on the observed data $(\mathbf{y}, \mathbf{X}_{obs})$.

Under the likelihood and prior specification of Section 3.1, full conditional posterior distributions of each unknown have a simple form allowing Gibbs sampling. These full conditionals are provided in Appendix 1, and we focus here on updating of β and θ . The full conditional posterior of β_j can be expressed as

$$(1 - \hat{\pi}_j)\delta_0 + \hat{\pi}_j N(E_j, V_j), \quad (9)$$

where $\hat{\pi}_j$ is the conditional posterior probability of $\gamma_j = 1$, which is

$$\hat{\pi}_j = 1 - \frac{1 - \pi_j}{1 - \pi_j + \pi_j \frac{\sqrt{\phi_{\beta_j} \phi(0)}}{V_j^{-\frac{1}{2}} \phi(V_j^{-\frac{1}{2}} E_j)}},$$

and the conditional expectation and variance of β_j given $\gamma_j = 1$ are

$$E_j = V_j \sum_{i=1}^n x_{ij} \tilde{y}_{ij}^*, \quad V_j = \left(\phi_{\beta_j} + \sum_{i=1}^n x_{ij}^2 \right)^{-1},$$

with $\tilde{y}_{ij}^* = y_i^* - \beta_0 - \sum_{h \neq j} x_{ih} \beta_h$ and $\phi(\cdot)$ the standard normal density.

Note that in updating β_j , we automatically update $\gamma_j = 1 (\beta_j \neq 0)$. Upon convergence, samples of γ are drawn from the marginal posterior distribution $p(\gamma | \mathbf{y}, \mathbf{X}_{obs})$. A model's posterior probability can then be estimated by the proportion of samples in that model. In addition, marginal inclusion probabilities, $\Pr(\gamma_j = 1 | \mathbf{y}, \mathbf{X}_{obs})$, provide a convenient weight of evidence that the j th predictor should be included.

The full conditional posterior of θ_{jk} is

$$(1 - \hat{\pi}_{jk})\delta_0 + \hat{\pi}_{jk} N(E_{jk}, V_{jk}), \quad (10)$$

where the conditional posterior probability of $\gamma_{jk}^m = 1$ is

$$\hat{\pi}_{jk} = 1 - \frac{1 - \pi_{jk}}{1 - \pi_{jk} + \pi_{jk} \frac{\sqrt{\phi_{\theta_{jk}} \phi(0)}}{V_{jk}^{-\frac{1}{2}} \phi(V_{jk}^{-\frac{1}{2}} E_{jk})}},$$

and the conditional posterior mean and variance given inclusion is

$$E_{jk} = V_{jk} \psi_j \sum_{i=1}^n x_{ik}^* \tilde{x}_{ijk}, \quad V_{jk} = \left(\phi_{\theta_{jk}} + \psi_j \sum_{i=1}^n x_{ik}^{*2} \right)^{-1},$$

with $\tilde{x}_{ijk} = x_{ij}^* - \theta_{j0} - \sum_{h=1, h \neq k}^{j-1} x_{ih}^* \theta_{jh}$. All other parameters can be sampled from their full conditionals as standard in regression models. For details of all the full conditionals to implement the Gibbs sampler please refer to Appendix 1.

The missing predictors are also imputed from their full conditional distributions, which are available in closed form, and so we embed the imputation of missing covariates within our stochastic search of the model space, allowing simultaneous treatment of the missing data and variable selection problems. We evaluate both SSVS¹ and SSVS² by considering posterior model inferences as well as out of sample predictive performance in a simulation study. We compare our results to model averaging performed on the original completely observed data (prior to introducing covariate missingness). More details on this are presented in Section 5. In the next section we provide a theoretical argument supporting the use of the SSVS² approach.

4 Improved Predictive Performance

Raftery et al. (1997) showed that Bayesian model averaging has better predictive performance than using any single model alone. We present a similar argument here extending this to the case of model averaging over the missing data models as opposed to using a single model for imputations.

Let Δ be the quantity we are interested in predicting (e.g. the disease outcome of a patient). Denote $(\mathbf{y}, \mathbf{X}_{obs})$ to represent the observed data and $\mathbf{X}_{mis}$ the missing data. As before let γ index the set Γ of the 2^p possible models for the response $\mathbf{y}$ and let γ^m index the set Γ^m of the $2^{\frac{p(p-1)}{2}}$ models for the missing data $\mathbf{X}_{mis}$, with γ_0^m representing a particular choice of model nested in Γ^m . Then in our SSVS² approach we define the predictive distribution of Δ as f , where

$$f = \sum_{\gamma \in \Gamma} \sum_{\gamma^m \in \Gamma^m} \left[\int p(\Delta | \gamma, \gamma^m, \mathbf{y}, \mathbf{X}_{obs}, \mathbf{X}_{mis}) p(\mathbf{X}_{mis} | \gamma, \gamma^m, \mathbf{y}, \mathbf{X}_{obs}) d\mathbf{X}_{mis} \right] \times p(\gamma, \gamma^m | \mathbf{y}, \mathbf{X}_{obs})$$

and in a one level approach the predictive distribution of Δ , defined as g is,

$$g = \sum_{\gamma \in \Gamma} \left[\int p(\Delta | \gamma, \gamma_0^m, \mathbf{y}, \mathbf{X}_{obs}, \mathbf{X}_{mis}) p(\mathbf{X}_{mis} | \gamma, \gamma_0^m, \mathbf{y}, \mathbf{X}_{obs}) d\mathbf{X}_{mis} \right] \times p(\gamma | \gamma_0^m, \mathbf{y}, \mathbf{X}_{obs})$$

The following theorem then holds:

Theorem 4.1. $-E_f[\log(f(x))] \leq -E_f[\log(g(x))]$

Hence under a logarithmic scoring rule we see that the estimate that model averages over both the set of models for the response as well as the missing data has a lower risk than the one which does not model average over the missing data.

Proof. Consider the Kullback-Leibler divergence for f and g :

$$\begin{aligned}
 K(f : g) &= \int_{-\infty}^{\infty} f(x) \log \left(\frac{f(x)}{g(x)} \right) dx \geq 0 \quad (\text{by non-negativity of K-L divergence}). \\
 \Rightarrow \int_{-\infty}^{\infty} \{f(x) \log(f(x)) - f(x) \log(g(x))\} dx &\geq 0 \\
 \Rightarrow \int_{-\infty}^{\infty} f(x) \log(f(x)) dx &\geq \int_{-\infty}^{\infty} f(x) \log(g(x)) dx \\
 \Rightarrow E_f[\log(f(x))] &\geq E_f[\log(g(x))] \\
 \Rightarrow -E_f[\log(f(x))] &\leq -E_f[\log(g(x))]
 \end{aligned}$$

□

For a proof of non-negativity of the Kullback-Leibler divergence please see Claeskens and Hjort (2008, ch. 2). We can use the above theorem to justify our SSVS² approach having better predictive performance for a future units' observation y^{new} conditional on observing some of its covariates $\mathbf{x}^{new}$ with $\mathbf{m}^{new}$ indicating which elements in $\mathbf{x}^{new}$ are missing. Defining $\mathbf{x}_{obs}^{new} = \{x_j^{new}, j : m_j^{new} = 0\}$ we express the predictive density of the future observation y^{new} in the SSVS² approach as:

$$p(y^{new} | \mathbf{x}_{obs}^{new}) = \sum_{\gamma \in \Gamma} \sum_{\gamma^m \in \Gamma^m} p(y^{new} | \gamma, \gamma^m, \mathbf{x}_{obs}^{new}) p(\gamma, \gamma^m | \mathbf{y}, \mathbf{X}_{obs})$$

which we know, by using the result of Theorem 4.1, has better predictive properties than restricting imputation models to be based on using a single model γ_0^m , where

$$p(y^{new} | \mathbf{x}_{obs}^{new}, \gamma_0^m) = \sum_{\gamma \in \Gamma} p(y^{new} | \gamma, \gamma_0^m, \mathbf{x}_{obs}^{new}) p(\gamma | \gamma_0^m, \mathbf{y}, \mathbf{X}_{obs})$$

It is interesting to note that in the applied statistician community, the consensus seems to be to specify as complex an imputation model as is feasible, and that model selection is not a concern; however, this is the case where only one model for imputations is being considered. We see that the SSVS² approach achieves gains in predictive performance by incorporating model uncertainty into the imputation model. The SSVS² approach considers, among other models, the most complex model, appropriately weighting each model by its posterior probability. The SSVS² approach thus does not aim to eliminate relationships present in the data; rather it takes advantage of the fact that there may be sparse relationships among the variables. We would like to thank an anonymous reviewer for highlighting this issue.

Note that for Theorem 4.1 to apply we must have model γ_0^m nested within the set of models being considered in the SSVS² approach, Γ^m . This is not the case with the SSVS¹ approach described in Section 3.1 as we are using Jeffreys

priors, not Cauchy priors, for the regression coefficients in the predictor component of the model. The SSVS¹ approach does correspond to a commonly used strategy to impute missing covariate values, and we would like to see whether benefits of improved predictive performance still exist in using the SSVS² approach, as the SSVS¹ approach does not incorporate model uncertainty into its imputation model. In the next section we investigate possible gains in predictive performance as well as additional benefits in obtaining posterior inferences, such as marginal inclusion probabilities of the predictors, when using the SSVS² approach in a simulation study.

5 Simulation Studies

We simulate 1000 units with 17 predictors ($x_1, \dots, x_{17}$) and a binary response y , of the 17 predictors x_3, x_8 , and x_{10} are simulated as binary predictors while the rest are continuous. We simulate the predictors from a Directed Acyclic Graph (DAG) to have sparse relationships in the following way,

$$\begin{aligned} p(x_1, \dots, x_{17}) = & p(x_1)p(x_2)p((x_3|x_1)p(x_4|x_1, x_3)p(x_5|x_1)p(x_6|x_1, x_5) \\ & p(x_7|x_3, x_6)p(x_8|x_1, x_5)p(x_9|x_3, x_6)p(x_{10}|x_4, x_7) \\ & p(x_{11}|x_7)p(x_{12}|x_3, x_4)p(x_{13}|x_6, x_9)p(x_{14}|x_1, x_4) \quad (11) \\ & p(x_{15}|x_1, x_{12})p(x_{16})p(x_{17}). \end{aligned}$$

Continuous predictors are simulated using normal linear models while the binary predictors are simulated from probit models in the same way that the response is simulated as described below.

We simulate y using a probit model. This is done by first simulating a latent variable for each unit i , y_i^* for $i = 1, \dots, 1000$ in the following way,

$$y_i^* = N(\beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \beta_5 x_{i5}, \sigma^2) \quad (12)$$

and then simulating the response by,

$$y_i = I(y_i^* > c) \quad (13)$$

where $I(\cdot)$ is the indicator function and c is a suitable cut point that results in approximately half the units' responses assigned to either 0 or 1. With this simulation design only 3 out of the 17 predictors are used to generate the response variable, and we denote these (and the intercept) to be our true predictors, the rest we denote as null predictors. Any relationship between the null predictors and the response is spurious and is due to correlations with the true set of predictors. Half of the

observations are assigned to be in our training data set and we use the other half as an out of sample test data set.

To evaluate our two approaches (SSVS² and SSVS¹) we introduce missing values in the predictors using relationships similar to those in (11). For example missing probabilities for x_4 are simulated using a logistic regression model on x_1 and x_3 . Each predictor is set to have approximately 40% of its values missing. When modeling the joint distribution of the predictors to impute the missing values we use the same ordering that was used to simulate the predictors. The SSVS¹ approach thus models $p(x_1, \dots, x_{17})$ by,

$$p(x_1, \dots, x_{17}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_{17}|x_1, \dots, x_{16}). \quad (14)$$

The SSVS² approach uses the same ordering of predictors to factorize the joint distribution with each regression conditional on a subset of the previous predictors in the list.

We can now perform variable selection via the Gibbs sampler outlined in section 3 using both approaches and compare posterior model inferences. In addition, we can consider posterior inferences in the situation where there is no missing data (SSVS^{obs}). In all cases the Markov Chains are run for 250,000 iterations with a burn-in of 50,000 iterations. We apply both approaches on increasing fractions of our training data set, each larger training data set contains all the units from the smaller training data sets. Figure 1 presents the mean inclusion and exclusion probabilities of the true and null predictor sets respectively for training data sizes of 50, 100, 150, 200, 300, 400 and 500.

The closer the line is to 1 in both plots the better the method is performing. As expected the case of fully observed covariate information does the best with performance increasing with training data size. In the first plot the SSVS² approach also exhibits a similar monotone pattern with gains in estimation of the true predictors' inclusion probabilities over SSVS¹ evident. In the second plot there is not much difference between the two approaches, with small gains in estimation of the null predictors' exclusion probabilities as the training data size increases.

To add measures of uncertainty to each of the point probability estimates in Figure 1, we estimate the Monte Carlo standard error using the consistent batch means (CBM) estimate suggested by Flegal, Haran, and Jones (2008). In Figure 2, we present the half-widths from 95% confidence intervals for each of the point probability estimates; the intervals were obtained using the CBM approach, with batch widths selected to be the square root of the Monte Carlo sample size, as recommended by Flegal et al. (2008). From Figure 2, we see that all except three of the half-width intervals are less than 0.005; the point probability estimates with half-width intervals less than 0.005 are taken to be accurate to the second decimal place. We obtained similar results for the probability estimates with Markov Chains

run for only 50000 iterations, and have noted that estimates from our proposed SSVS² approach are robust to different starting points.

Figure 1: Mean Inclusion/Exclusion Probabilities for the True/Null Predictors respectively for the three cases across different training data sizes.

Figure 2: Half-widths from 95% confidence intervals of the mean marginal Inclusion/Exclusion Probabilities for the True/Null Predictor sets respectively, for the three cases across different training data sizes.

Both the $SSVS^1$ and $SSVS^2$ approach use the same ordering of predictors to factorize the joint distribution as was used to simulate the data. We consider four random permutations of the ordering of predictors when factorizing the joint distribution in both approaches. Plots of the inclusion and exclusion probabilities of the true and null predictor sets similar to Figure 1 are presented in Appendix 2 for each of the reorderings. In each of the plots we see the relationship is similar to that of Figure 1 with the $SSVS^2$ approach showing gains over the $SSVS^1$ approach in estimating marginal inclusion probabilities for the true predictor set for smaller training data sizes. Lipsitz and Ibrahim (1996), Ibrahim et al. (1999) also explore the effect of different orderings of predictors when imputing missing values using this type of decomposition and find that inferences are robust.

In addition we can use the out of sample test data set to evaluate predictive performance of the methods. We impute missing covariate values in the test data set from their full conditional distributions within each iteration of the MCMC (see Appendix 1 for details) and can thus generate predictions for the response which can be compared with the actual values. As we have a binary response this can be conveniently summarized by the percentage of units correctly classified. Figure 3 presents the correct classification rates for the two approaches plus the situation when there is no missing data when applied to the training data sets described above.

Figure 3: Out of sample predictive performance for the two different methods compared to the case with fully observed covariates.

We see that of course the situation when the covariates are all fully observed has the best classification rates. The $SSVS^2$ method outperforms the $SSVS^1$ approach across all training data sizes. In $SSVS^{obs}$ and $SSVS^2$ the correct classification rate tends to increase with training data size, while the increasing trend is not so clear with $SSVS^1$.

We also explore whether the correct classification rates are sensitive to reorderings; we calculate the correct classification rates from both approaches under the same four reorderings of the predictors. The results are presented in Appendix 2 using plots similar to Figure 3. We see that in all the reorderings examined the $SSVS^2$ approach outperforms the $SSVS^1$ approach across all training data sizes.

6 Reproductive Epidemiology Application

We now apply our methods to data from the Longnecker, Klebanoff, Zhou, and Brock (2001) sub-study of the US Collaborative Perinatal Project (CPP). We are interested in predicting high risk pregnancies when there are missing predictors, for related work refer to Eastaugh, Smye, Snowden, Walker, Dear, and Farrin (1997). We took our response to be whether a preterm birth was observed or not (0 indicating that a preterm birth occurred and 1 indicating a normal term birth), and chose thirteen fully observed variables (binary and continuous) as candidate predictors. We include mother's age, height, pre-pregnancy BMI, pregnancy weight gain, smoking status, race (an indicator variable of African American ethnicity), serum total cholesterol, triglycerides, sum of PCBs, and p,p'-DDE (lipid adjusted) which is a persistent metabolite of DDT measured in serum for each woman. We also include the child's gender, the socio-economic index and whether the prenatal care was adequate. Of these predictors, smoking status, race, child's gender and prenatal care are binary and the other nine predictors are continuous. We focus on the sub-sample of women in the study with advanced maternal age (35 or older) which results in a sample size of 182. In this sub-sample approximately 20% of women experienced preterm birth.

We then introduced approximately 40% missing data in each predictor. In particular we generate missing values in the predictors race, pre-pregnancy BMI and socio-economic index using an underlying latent lifestyle factor that we assume is related to these three predictors and the response preterm birth. For all other predictors we generate missing values using an MCAR mechanism. To impute the missing values we model the joint distribution of the predictors as in section 5,

$$p(\mathbf{x}_1, \dots, \mathbf{x}_{13}) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1) \dots p(\mathbf{x}_{13}|\mathbf{x}_1, \dots, \mathbf{x}_{12}) \quad (15)$$

where we take adequate prenatal care as x_1 , race as x_2 , child's gender as x_3 , smoking status as x_4 , pregnancy weight gain as x_5 , serum total cholesterol as x_6 , triglycerides as x_7 , sum of PCBs as x_8 , socio-economic index as x_9 , mother's height as x_{10} , pre-pregnancy BMI as x_{11} , mother's age as x_{12} , and p,p'-DDE (lipid adjusted) as x_{13} . We thus chose to factorize the predictors by first modeling the binary predictors and then the continuous predictors. Another possibility could be to factorize the predictors so that posterior inferences of the regression coefficients from the predictor component of the model would correspond to inferences of scientific interest.

Table 1 presents posterior summaries of the regression coefficients for the response component of the model when using the SSVS² approach. The table presents the marginal inclusion probability, the posterior mean, the posterior standard deviation and a 95% credible interval for each coefficient. We see that only the coefficients for race and triglycerides have marginal inclusion probabilities greater than 0.6. The posterior means for the coefficients suggest that having adequate prenatal care may decrease the risk of preterm birth while African American women and subjects with higher levels of triglycerides tend to have an increased risk of preterm birth, though the intervals are wide so the results are just suggestive.

In Appendix 3 we also present posterior summaries of the regression coefficients in the SSVS² approach for the predictor component of the model. From these we find that there are typically sparse relationships between the predictors with most coefficients having marginal inclusion probabilities less than 0.4 and this further supports the use of the SSVS² approach. The variable race appears to be an important predictor in several of the conditional regressions with marginal inclusion probabilities greater than 0.6 in the regression for triglycerides, sum of PCBs, and socio-economic index. The variable sum of PCBs also appears to be an important predictor in the regression for p,p'-DDE with a large marginal inclusion probability. These findings are consistent with our prior expectations and the literature. In particular, we would expect African Americans to tend to have a higher level of triglycerides and lower socio-economic status, while potentially having higher levels of exposures due to dietary factors. It is also reasonable to expect the sum of PCBs to be correlated with p,p'-DDE.

For comparison, Table 2 presents posterior summaries for the regression coefficients obtained in analyzing the complete data without missing values introduced. In addition to this SSVS^{obs} analysis, we also include frequentist p-values for each coefficient obtained from a probit model fit to the complete data. In general posterior means are fairly similar between SSVS² and SSVS^{obs}, with the main differences in the posterior means of the coefficients for the predictors measuring adequate prenatal care and race. Nevertheless, the 95% credible intervals for these coefficients from SSVS² overlap greatly with the corresponding 95% credible in-

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	1.033	3.833	(-8.422, 7.826)
Adequate prenatal care	0.434	0.268	0.452	(-0.120, 1.459)
Race	0.729	-0.520	0.434	(-1.369, 0.000)
Gender	0.253	-0.066	0.195	(-0.659, 0.140)
Smoking status	0.298	0.101	0.244	(-0.126, 0.823)
Pregnancy weight gain	0.143	-0.000	0.000	(-0.001, 0.000)
Cholesterol	0.100	0.000	0.001	(-0.001, 0.001)
Triglycerides	0.657	-0.003	0.003	(-0.008, 0.000)
Sum of PCBs	0.163	0.000	0.001	(-0.000, 0.002)
Socio-economic index	0.236	0.022	0.054	(0.000, 0.184)
Height	0.289	0.026	0.051	(0.000, 0.168)
Pre-pregnancy BMI	0.100	-0.000	0.008	(-0.021, 0.013)
Age	0.268	-0.024	0.050	(-0.166, 0.000)
p,p'-DDE	0.123	-0.000	0.000	(-0.000, 0.000)

Table 1: Posterior summaries of the coefficients in the regression for preterm from SSVS².

tervals from SSVS^{obs}. In general, the 95% credible intervals from SSVS² tend to be wider than the intervals from SSVS^{obs} as expected.

We also compare the results obtained from the SSVS² approach and SSVS^{obs} to those obtained from the SSVS¹ approach. Table 3 presents posterior summaries of the regression coefficients from the SSVS¹ approach. We see again that there are similarities in the posterior means of many of the regression coefficients to the posterior means from SSVS^{obs}. The main differences in posterior means between SSVS^{obs} and SSVS¹ are the regression coefficients for the intercept and socio-economic index. The marginal inclusion probability of the coefficient for socio-economic index is also quite different in SSVS¹ to the one obtained in SSVS^{obs}.

From the many numbers in these three tables, it is hard to compare overall differences. Hence, in Figure 4 we present a comparison plot of the absolute differences in the posterior means between SSVS^{obs} and SSVS¹ against absolute differences in the posterior means between SSVS^{obs} and SSVS². Points above the line $y = x$ (included on the plot) indicate better performance in SSVS² over SSVS¹ and vice versa. We see that there are several points quite far above the line, and so there is some evidence to suggest that the SSVS² estimates are closer to the full data results than the SSVS¹ estimates.

Predictor	incl. probability	mean	std dev	95% CI	p-value
Intercept	1.000	0.721	1.705	(-4.261, 4.257)	0.872
Adequate prenatal care	0.793	0.605	0.440	(0.000, 1.459)	0.085
Race	0.298	-0.089	0.189	(-0.629, 0.022)	0.134
Gender	0.148	-0.007	0.084	(-0.249, 0.160)	0.894
Smoking status	0.157	0.012	0.092	(-0.148, 0.293)	0.968
Pregnancy weight gain	0.081	0.000	0.000	(-0.000, 0.000)	0.753
Cholesterol	0.085	-0.000	0.000	(-0.001, 0.000)	0.919
Triglycerides	0.203	-0.000	0.001	(-0.004, 0.000)	0.088
Sum of PCBs	0.084	-0.000	0.000	(-0.001, 0.000)	0.772
Socio-economic index	0.212	0.016	0.040	(0.000, 0.141)	0.504
Height	0.141	0.007	0.022	(0.000, 0.083)	0.372
Pre-pregnancy BMI	0.080	0.000	0.005	(-0.005, 0.013)	0.553
Age	0.138	-0.007	0.023	(-0.088, 0.000)	0.304
p,p'-DDE	0.112	-0.000	0.000	(-0.000, 0.000)	0.564

Table 2: Posterior summaries of the coefficients in the regression for preterm from SSVS^{obs}. Frequentist p-values from a probit model fit on the fully observed data are also included.

In addition we compare the quality of the imputed values from both approaches. In Figure 5 we plot the mean of the squared differences between the imputed and true values for each predictor, averaged over the imputations from the SSVS¹ approach, against the corresponding statistics obtained from the SSVS² approach. Points above the $y = x$ line (included on the plot) indicate favor for the SSVS² approach; we see that there are more points above the line, and points below are still fairly close to the $y = x$ line. Hence, we see that the SSVS² approach tends to impute missing values closer to their true values than the SSVS¹ approach.

We see from the results in Tables 1 to 3 that there are no predictors that are highly “significant” based on commonly used cutoffs. However, this is not a feature of our prior distributions but is instead a common issue in studies in which there are moderate to large numbers of candidate predictors and the sample size is not that big. In our case, we have $p = 14$ candidate predictors including the intercept and only $n = 182$ women, which is a smallish sample size for a binary outcome. In reproductive epidemiology studies, the signal-to-noise ratio is typically small and it is very challenging to accurately predict preterm birth as there are unmeasurable factors (based on current knowledge and technology) that have a large influence. From Table 2, we see a frequentist analysis of the fully observed data produced p-values greater than 0.05 for all predictors. That said, from the SSVS² analysis there was evidence that advanced maternal age African Americans had increased

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	-0.546	3.689	(-9.378, 7.343)
Adequate prenatal care	0.902	1.096	0.626	(0.000, 2.334)
Race	0.281	0.003	0.246	(-0.672, 0.594)
Gender	0.206	-0.039	0.139	(-0.481, 0.106)
Smoking status	0.200	0.006	0.136	(-0.314, 0.375)
Pregnancy weight gain	0.096	-0.000	0.000	(-0.000, 0.000)
Cholesterol	0.134	0.000	0.001	(-0.000, 0.003)
Triglycerides	0.272	-0.001	0.002	(-0.005, 0.000)
Sum of PCBs	0.097	0.000	0.000	(-0.000, 0.001)
Socio-economic index	0.778	0.123	0.085	(0.000, 0.277)
Height	0.275	0.022	0.044	(0.000, 0.149)
Pre-pregnancy BMI	0.209	0.007	0.018	(0.000, 0.062)
Age	0.206	-0.017	0.047	(-0.169, 0.003)
p,p'-DDE	0.118	0.000	0.000	(-0.000, 0.000)

Table 3: Posterior summaries of the coefficients in the regression for preterm from SSVS¹.

risk of preterm birth. In particular, the estimated marginal inclusion probability was 0.729 for African American ethnicity, with the estimated coefficient being -0.520. Although it is well known that African Americans have an increased risk of preterm birth, it is interesting that our results suggest that this increased risk also holds among women in their late 30s. Many of the African American preterm births that occur in the US are for younger women. There is substantial interest among reproductive epidemiologists in looking for factors explaining the difference between African Americans and Caucasians in preterm birth risk, as adverse exposures and lifestyle factors arising from low socio-economic status does not seem to be an explanation. Hispanics do not have an increased risk relative to Caucasians based on previous studies. The suggestion of a higher rate of preterm birth even in advanced maternal age mothers is a scientifically interesting finding. Interestingly, the marginal inclusion probability for African American ethnicity was only 0.281 in the SSVS¹ analysis, with a coefficient estimate of 0.003 being very close to zero.

To verify that our prior distributions are not overly parsimonious, we perform an analysis of the incomplete data without performing any variable selection. We do this by placing Jeffrey's priors on the regression coefficients in both the response and predictor components of the model. Table 4 presents posterior summaries of the regression coefficients from the response component of the model; as we are not performing SSVS, marginal inclusion probabilities are not presented in

Figure 4: Absolute difference in posterior means of regression coefficients from $SSVS^1$ against $SSVS^2$ as compared to $SSVS^{obs}$, the line $y = x$ is included.

the table. We see that the posterior means from this approach are fairly similar to the posterior means obtained when using the $SSVS^2$ approach. The main differences are in the posterior means of the regression coefficients for the intercept and the predictor measuring adequate prenatal care. Hence, our Bayes variable selection approach does not appear to be overly favoring parsimony at the expense of discarding important predictors.

We also perform a frequentist analysis of the incomplete data, where we multiply impute missing values using the MICE package in R. The imputations are based on the switching regression approach of van Buuren et al. (1999) and estimates of the regression coefficients are derived from the multiple imputation combining rules of Rubin (1987). Table 5 summarizes the results from the analysis based on 10 multiply imputed data sets generated using MICE. As with the results from Table 4, we see that there are few significant predictors. All the p-values are greater than 0.05, and all the 95% confidence intervals include zero. The coefficient for Race, an indicator variable for African American ethnicity, is substantially closer to zero than the estimate from the $SSVS^2$ analysis. We might thus make different conclusions about the effect of this predictor if we used this approach to deal with the missing values.

Figure 5: Mean of the squared differences between the imputed and true predictor values, for each predictor, averaged over the imputations from the SSVS¹ approach, against the corresponding statistics from the SSVS² approach. The line $y = x$ is included.

7 Conclusion

In this paper we presented an efficient way to accommodate the problem of missing continuous and binary covariates when model averaging in Generalized Linear Models. We illustrated the benefits of additionally model averaging over the imputation models in posterior inferences and out of sample predictive performance through a simulation study. Finally, we applied our method to a reproductive epidemiology study to evaluate the benefits in using our two level approach. We found that SSVS² obtained estimates closer to the estimates from SSVS^{obs} than those from SSVS¹.

It would be interesting to extend our models to incorporate a wider range of Generalized Linear Models such as count response data, perhaps using approximations to the Marginal Likelihood developed by Raftery (1996) or Cai and Dunson (2006). We could also in principle extend our method to mixed effects data where variable selection could be performed on both the fixed effects regression coefficients as well as the variances of the random effects (Kinney and Dunson, 2007). A prior specification that takes into account the scale of the predictors such as Zellner's g prior might also be preferable to the ridge type priors used in this paper.

Predictor	mean	std dev	95% CI
Intercept	0.561	3.319	(-5.977, 7.029)
Adequate prenatal care	0.717	0.390	(-0.004, 1.527)
Race	-0.452	0.300	(-1.044, 0.136)
Gender	-0.036	0.238	(-0.502, 0.429)
Smoking status	-0.009	0.261	(-0.518, 0.508)
Pregnancy weight gain	0.000	0.000	(-0.001, 0.001)
Cholesterol	-0.000	0.002	(-0.003, 0.003)
Triglycerides	-0.003	0.002	(-0.006, 0.000)
Sum of PCBs	0.000	0.001	(-0.001, 0.002)
Socio-economic index	0.045	0.064	(-0.080, 0.169)
Height	0.041	0.044	(-0.046, 0.128)
Pre-pregnancy BMI	0.013	0.020	(-0.026, 0.053)
Age	-0.050	0.047	(-0.142, 0.043)
p,p'-DDE	-0.000	0.000	(-0.000, 0.000)

Table 4: Posterior summaries of the coefficients in the regression for preterm without performing any SSVS.

8 Appendix 1 - Full Conditionals

We present here the joint posterior distribution and the resulting full conditionals required for the Gibbs sampler in the SSVS² approach, focusing for simplicity on the case in which the response is binary and the predictors are binary or continuous. SSVS proceeds by sampling from the joint posterior of the model space (γ, γ^m) , parameters within each model $(\beta_\gamma, \phi_{\beta_\gamma}, \theta_{\gamma^m}, \phi_{\theta_{\gamma^m}}, \kappa)$, and the latent/unobserved variables $(\mathbf{y}^*, \mathbf{X}^*, \mathbf{X}_{mis})$ conditional on the observed data $(\mathbf{y}, \mathbf{X}_{obs})$. The joint posterior is expressed below,

$$\begin{aligned}
& \pi(\gamma, \gamma^m, \mathbf{y}^*, \beta_\gamma, \phi_{\beta_\gamma}, \theta_{\gamma^m}, \phi_{\theta_{\gamma^m}}, \kappa, \mathbf{X}^*, \mathbf{X}_{mis} | \mathbf{y}, \mathbf{X}_{obs}) \\
& \propto \left\{ \prod_i^n p(y_i | y_i^*) p(y_i^* | \beta_\gamma, \mathbf{x}_i) p(\mathbf{x}_i, \mathbf{x}_i^* | \theta_{\gamma^m}, \kappa) \right\} p(\beta_\gamma, \phi_{\beta_\gamma} | \gamma) p(\gamma) \\
& \quad \times \pi(\theta_{\gamma^m}, \phi_{\theta_{\gamma^m}} | \gamma^m) p(\gamma^m) p(\kappa) \\
& \propto \left\{ \prod_i^n \left(y_i I(y_i^* \geq 0) + (1 - y_i) I(y_i^* < 0) \right) N(y_i^*; \beta_0 + \mathbf{x}'_{\gamma i} \beta_\gamma, 1) \right. \\
& \quad \times \left. \left[\prod_{j=1}^p p(x_{ij} | x_{ij}^*) N(x_{ij}^*; \theta_{j0} + \mathbf{x}'_{j\gamma^m i} \theta_{j\gamma^m}, \kappa_j) \right] \right\} \left\{ \prod_{j=0}^p p(\beta_j | \phi_{\beta_j}, \gamma_j) p(\gamma_j) \right. \\
& \quad \times \left. p(\phi_{\beta_j} | \gamma_j) \right\} \left\{ \prod_{j=1}^p \prod_{k=0}^{j-1} \left[p(\theta_{jk} | \phi_{\theta_{jk}}, \gamma_{jk}^m) p(\phi_{\theta_{jk}} | \gamma_{jk}^m) p(\gamma_{jk}^m) \right] \right\} p(\kappa)
\end{aligned}$$

Predictor	estimate	std. error	p-value	95% CI
Intercept	-1.779	7.141	0.806	(-16.877, 13.318)
Adequate prenatal care	0.742	0.567	0.201	(-0.415, 1.898)
Race	-1.257	0.778	0.132	(-2.947, 0.434)
Gender	-0.450	0.673	0.517	(-1.924, 1.024)
Smoking status	0.458	0.576	0.440	(-0.773, 1.688)
Pregnancy weight gain	-0.001	0.001	0.095	(-0.003, 0.000)
Cholesterol	0.000	0.003	0.995	(-0.007, 0.007)
Triglycerides	-0.008	0.004	0.067	(-0.016, 0.001)
Sum of PCBs	0.003	0.002	0.101	(-0.001, 0.007)
Socio-economic index	-0.065	0.122	0.603	(-0.324, 0.194)
Height	0.136	0.093	0.158	(-0.058, 0.330)
Pre-pregnancy BMI	0.048	0.039	0.227	(-0.032, 0.128)
Age	-0.115	0.092	0.224	(-0.307, 0.077)
p,p'-DDE	-0.000	0.000	0.354	(-0.001, 0.000)

Table 5: Estimates of the regression coefficients when multiply imputing the missing data using the MICE package in R

where, $p(x_{ij}|x_{ij}^*) = x_{ij}I(x_{ij}^* \geq 0) + (1 - x_{ij})I(x_{ij}^* < 0)$, $\kappa_j = 1$ for binary x_{ij} and $p(x_{ij}|x_{ij}^*) = \delta_{x_{ij}^*}(x_{ij})$ for continuous x_{ij}

With the models for the data and the prior distributions for the parameters discussed in section 3, the full conditionals are available in closed form. First consider the full conditional distributions for the parameters in the predictor component of the model. Sample θ_{jk} from,

$$(1 - \hat{\pi}_{jk})\delta_0 + \hat{\pi}_{jk}N(E_{jk}, V_{jk}), \quad (16)$$

where the conditional posterior probability of $\gamma_{jk}^n = 1$ is

$$\hat{\pi}_{jk} = 1 - \frac{1 - p_{jk}}{1 - p_{jk} + p_{jk} \frac{\sqrt{\phi_{\theta_{jk}} \phi(0)}}{V_{jk}^{-\frac{1}{2}} \phi(V_{jk}^{-\frac{1}{2}} E_{jk})}},$$

and the conditional posterior mean and variance given inclusion is

$$E_{jk} = V_{jk} \psi_j \sum_{i=1}^n x_{ik}^* \tilde{x}_{ijk}^*, \quad V_{jk} = \left(\phi_{\theta_{jk}} + \psi_j \sum_{i=1}^n x_{ik}^{*2} \right)^{-1},$$

with $\tilde{x}_{ijk}^* = x_{ij}^* - \theta_{j0} - \sum_{h=1, h \neq k}^{j-1} x_{ih}^* \theta_{jh}$ and $\phi(\cdot)$ the standard normal density. Also update $\phi_{\theta_{jk}}$ for predictors included in the model from,

$$Ga\left(1, \frac{\theta_{jk}^2 + 1}{2}\right). \quad (17)$$

Next sample $\psi_k = \kappa_k^{-1}$ for continuous x_k from,

$$Ga\left(\frac{n}{2}, \sum_{i=1}^n x_{ik}^* - \left(\theta_{j0} - \sum_{j=1}^k x_{ij}^* \theta_{jk}\right)^2\right) \quad (18)$$

while $\psi_k = \kappa_k^{-1} = 1$ for binary x_k . Now for the i^{th} missing continuous covariate value x_{ij} , we impute from a normal distribution,

$$N\left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) \quad (19)$$

where,

$$\begin{aligned} \tilde{\psi}_j &= \beta_j^2 + \psi_j + \sum_{k=j+1}^p \psi_k \theta_{kj}^2, \\ \tilde{\mu}_{ij} &= \tilde{y}_{ij}^* \beta_j + \psi_j \mu_{ij} + \sum_{k=j+1}^p \psi_k \theta_{kj} \tilde{x}_{ikj}^* \end{aligned}$$

and,

$$\begin{aligned} \tilde{y}_{ij}^* &= y_i^* - \beta_0 - \sum_{h \neq j} x_{ih} \beta_h, \\ \mu_{ij} &= \theta_{j0} + \sum_{k=1}^{j-1} x_{ik}^* \theta_{jk}, \\ \tilde{x}_{ikj}^* &= x_{ik}^* - \theta_{k0} - \sum_{h=1, h \neq j}^{k-1} x_{ih}^* \theta_{kh}. \end{aligned}$$

While when x_{ij} is binary and missing, we first impute its underlying latent variable x_{ij}^* from the full conditional,

$$\tilde{\pi}_{ij} N_+\left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) + (1 - \tilde{\pi}_{ij}) N_-\left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1}\right) \quad (20)$$

where,

$$\begin{aligned}\tilde{\psi}_j^{-1} &= \psi_j + \sum_{k=j+1}^p \psi_j \theta_{kj}^2, \\ \tilde{\mu}_{ij} &= \psi_j \mu_{ij} + \sum_{k=j+1}^p \psi_k \theta_{kj} \tilde{x}_{ik_j}^*, \\ \tilde{\pi}_{ij} &= \frac{(1 - \Phi\left(\frac{\tilde{\mu}_{ij}}{\sqrt{\tilde{\psi}_j}}\right)) \phi\left(\tilde{y}_{ij}^* - \beta_j\right)}{(1 - \Phi\left(\frac{\tilde{\mu}_{ij}}{\sqrt{\tilde{\psi}_j}}\right)) \phi\left(\tilde{y}_{ij}^* - \beta_j\right) + \Phi\left(\frac{\tilde{\mu}_{ij}}{\sqrt{\tilde{\psi}_j}}\right) \phi\left(\tilde{y}_{ij}^*\right)}\end{aligned}$$

and,

$$\begin{aligned}\mu_{ij} &= \theta_{j0} + \sum_{k=1}^{j-1} x_{ik}^* \theta_{jk}, \\ \tilde{x}_{ik_j}^* &= x_{ik}^* - \theta_{k0} - \sum_{h=1, h \neq j}^{k-1} x_{ih}^* \theta_{kh}, \\ \tilde{y}_{ij}^* &= y_i^* - \beta_0 - \sum_{h \neq j} x_{ih} \beta_h,\end{aligned}$$

and then impute $x_{ij} = I(x_{ij}^* > 0)$. We also update latent x_{ij}^* for observed binary x_{ij} from the following distribution:

$$x_{ij} N_+ \left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1} \right) + (1 - x_{ij}) N_- \left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1} \right) \quad (21)$$

where, $\tilde{\psi}$ and $\tilde{\mu}_{ij}$ are as in (20). Note that for individual i predictors other than x_{ij} may be missing, in the imputations we condition on the most recently imputed values of other missing predictors. Now conditional on the observed and imputed predictors we can sample from the full conditionals in the top level models for the response. We sample β_j from its full conditional posterior,

$$(1 - \hat{\pi}_j) \delta_0 + \hat{\pi}_j N(E_j, V_j), \quad (22)$$

where $\hat{\pi}_j$ is the conditional posterior probability of $\gamma_j = 1$, which is

$$\hat{\pi}_j = 1 - \frac{1 - \pi_j}{1 - \pi_j + \pi_j \frac{\sqrt{\phi_{\beta_j} \phi(0)}}{V_j^{-\frac{1}{2}} \phi(V_j^{-\frac{1}{2}} E_j)}},$$

and the conditional expectation and mean of β_j given $\gamma_j = 1$ are

$$E_j = V_j \sum_{i=1}^n x_{ij} \tilde{y}_{ij}^*, \quad V_j = \left(\phi_{\beta_j} + \sum_{i=1}^n x_{ij}^2 \right)^{-1},$$

with $\tilde{y}_{ij}^* = y_i^* - \beta_0 - \sum_{h \neq j} x_{ih} \beta_h$ and $\phi(\cdot)$ the standard normal density. We sample ϕ_{β_j} for predictors included in the model from its full conditional,

$$Ga \left(1, \frac{\beta_j^2 + 1}{2} \right). \quad (23)$$

Finally sample y_i^* from its full conditional,

$$y_i N_+(\mathbf{x}_i' \boldsymbol{\beta}, 1) + (1 - y_i) N_-(\mathbf{x}_i' \boldsymbol{\beta}, 1) \quad (24)$$

In this way within one Gibbs sampler we repeatedly impute values for the missing covariates from their full conditional distributions and conditional on the completed data set perform variable selection on the model relating the response to the predictors.

When imputing missing values in the out of sample test data we do not observe the response y and so we must impute from modified full conditionals. For x_{ij} missing and continuous impute from,

$$N \left(\tilde{\psi}_j^{-1} \tilde{\mu}_{ij}, \tilde{\psi}_j^{-1} \right) \quad (25)$$

where, $\tilde{\psi}_j$ and $\tilde{\mu}_{ij}$ are as in (20). For x_{ij} missing and binary impute its underlying latent variable x_{ij}^* from (25) and impute $x_{ij} = I(x_{ij}^* > 0)$. For x_{ij} observed and binary update x_{ij}^* from the same distribution as (21).

9 Appendix 2 - Sensitivity to reordering

In this section we examine the sensitivity of the results obtained in Section 5 to four different reorderings of the predictors used to factorize the joint density of the predictors. This is done by obtaining a random permutation of $(1, 2, \dots, 17)$, $(a_1, a_2, \dots, a_{17})$ and decomposing the joint distribution by,

$$p(x_1, x_2, \dots, x_{17}) = \prod_{i=2}^{17} p(x_{a_i} | x_{a_1}, \dots, x_{a_{i-1}})$$

The four different reorderings are as follows:

1. $(a_1, a_2, \dots, a_{17}) = (15, 8, 9, 14, 11, 6, 4, 17, 13, 3, 10, 7, 2, 12, 5, 16, 1)$
2. $(a_1, a_2, \dots, a_{17}) = (17, 16, 3, 11, 13, 7, 15, 8, 5, 4, 1, 9, 12, 14, 2, 10, 6)$
3. $(a_1, a_2, \dots, a_{17}) = (7, 4, 10, 12, 1, 11, 15, 8, 14, 6, 9, 13, 17, 16, 5, 2, 3)$
4. $(a_1, a_2, \dots, a_{17}) = (11, 7, 16, 8, 2, 6, 17, 14, 10, 1, 9, 5, 15, 13, 4, 3, 12)$

For each reordering we present the mean inclusion and exclusion probabilities of the true and null predictor sets respectively. We see similar results to those presented in Figure 1. We also present the out of the sample correct classification rates, and observe similar results to those obtained in Figure 3.

9.1 Reordering 1

Figure 6: Mean Inclusion/Exclusion Probabilities for the True/Null Predictors respectively for the three cases across different training data sizes

Figure 7: Out of sample predictive performance for the two different methods compared to the case with fully observed covariates

9.2 Reordering 2

Figure 8: Mean Inclusion/Exclusion Probabilities for the True/Null Predictors respectively for the three cases across different training data sizes

Figure 9: Out of sample predictive performance for the two different methods compared to the case with fully observed covariates

9.3 Reordering 3

Figure 10: Mean Inclusion/Exclusion Probabilities for the True/Null Predictors respectively for the three cases across different training data sizes

Figure 11: Out of sample predictive performance for the two different methods compared to the case with fully observed covariates

9.4 Reordering 4

Figure 12: Mean Inclusion/Exclusion Probabilities for the True/Null Predictors respectively for the three cases across different training data sizes

Figure 13: Out of sample predictive performance for the two different methods compared to the case with fully observed covariates

10 Appendix 3 - Posterior summaries from the SSVS² analysis of the reproductive epidemiology study

In this section we present posterior summaries of the coefficients in each of the conditional regressions for the predictors when applying the SSVS² approach. Specifically for each coefficient we present its marginal inclusion probability, posterior mean, standard deviation and 95% credible interval.

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	-0.884	0.146	(-1.173, -0.604)

Table 6: Posterior summaries of the coefficients in the regression for adequate pre-natal care

Predictor	incl. probability	mean	std dev	95% CI
Intercept	0.117	-0.017	0.082	(-0.273, 0.003)
Adequate prenatal care	0.118	-0.018	0.083	(-0.279, 0.000)

Table 7: Posterior summaries of the coefficients in the regression for race

Predictor	incl. probability	mean	std dev	95% CI
Intercept	0.552	-0.167	0.191	(-0.587, 0.000)
Adequate prenatal care	0.199	0.000	0.113	(-0.322, 0.271)
Race	0.232	0.051	0.131	(-0.010, 0.461)

Table 8: Posterior summaries of the coefficients in the regression for child's gender

Predictor	incl. probability	mean	std dev	95% CI
Intercept	0.869	-0.334	0.187	(-0.688, 0.000)
Adequate prenatal care	0.217	0.007	0.123	(-0.302, 0.339)
Race	0.170	0.024	0.096	(-0.054, 0.338)
Gender	0.146	0.006	0.084	(-0.165, 0.244)

Table 9: Posterior summaries of the coefficients in the regression for smoking status

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	332.254	10.556	(317.786, 349.634)
Adequate prenatal care	0.087	1.267	11.116	(-4.188, 29.676)
Race	0.246	13.881	31.099	(0.000, 106.207)
Gender	0.153	-5.360	19.623	(-72.345, 2.266)
Smoking status	0.121	-2.968	17.031	(-59.152, 11.243)

Table 10: Posterior summaries of the coefficients in the regression for pregnancy weight gain

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	236.513	5.575	(218.612, 244.963)
Adequate prenatal care	0.077	-0.181	2.582	(-5.847, 1.697)
Race	0.308	-5.567	10.203	(-32.601, 0.000)
Gender	0.183	-2.420	7.533	(-27.419, 1.001)
Smoking status	0.088	-0.277	3.030	(-8.684, 2.460)
Pregnancy weight gain	0.109	0.003	0.015	(-0.006, 0.053)

Table 11: Posterior summaries of the coefficients in the regression for Cholesterol

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	231.997	16.556	(177.367, 249.500)
Adequate prenatal care	0.082	-0.385	2.729	(-8.481, 0.000)
Race	0.605	-14.398	14.149	(-41.728, 0.000)
Gender	0.163	1.848	6.224	(-0.448, 23.030)
Smoking status	0.125	1.060	4.488	(0.000, 16.864)
Pregnancy weight gain	0.086	-0.001	0.010	(-0.027, 0.010)
Cholesterol	0.160	0.021	0.068	(0.000, 0.251)

Table 12: Posterior summaries of the coefficients in the regression for triglycerides

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	328.530	41.001	(211.320, 406.890)
Adequate prenatal care	0.096	-1.104	6.617	(-22.345, 0.000)
Race	0.909	61.407	28.694	(0.000, 109.367)
Gender	0.327	12.996	23.167	(0.000, 73.640)
Smoking status	0.113	1.445	8.607	(-4.516, 30.090)
Pregnancy weight gain	0.150	-0.011	0.040	(-0.147, 0.000)
Cholesterol	0.108	-0.000	0.095	(-0.213, 0.207)
Triglycerides	0.130	0.026	0.130	(-0.072, 0.466)

Table 13: Posterior summaries of the coefficients in the regression for sum of PCBs

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	5.163	1.120	(3.040, 7.628)
Adequate prenatal care	0.082	0.000	0.082	(-0.100, 0.152)
Race	0.980	-1.104	0.377	(-1.857, -0.296)
Gender	0.222	-0.099	0.242	(-0.843, 0.000)
Smoking status	0.161	0.052	0.167	(0.000, 0.619)
Pregnancy weight gain	0.117	-0.000	0.000	(-0.002, 0.000)
Cholesterol	0.165	0.001	0.002	(0.000, 0.007)
Triglycerides	0.402	-0.003	0.004	(-0.013, 0.000)
Sum of PCBs	0.164	0.000	0.001	(-0.001, 0.005)

Table 14: Posterior summaries of the coefficients in the regression for socio-economic index

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	63.019	1.147	(60.330, 65.342)
Adequate prenatal care	0.077	0.000	0.091	(-0.132, 0.134)
Race	0.350	0.272	0.468	(0.000, 1.513)
Gender	0.259	-0.160	0.353	(-1.203, 0.000)
Smoking status	0.175	0.075	0.220	(0.000, 0.802)
Pregnancy weight gain	0.120	0.000	0.001	(0.000, 0.002)
Cholesterol	0.147	0.001	0.002	(0.000, 0.009)
Triglycerides	0.166	-0.001	0.002	(-0.009, 0.000)
Sum of PCBs	0.136	0.000	0.001	(-0.001, 0.004)
Socio-economic index	0.210	0.054	0.139	(0.000, 0.496)

Table 15: Posterior summaries of the coefficients in the regression for mother's height

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	27.486	4.630	(21.944, 40.349)
Adequate prenatal care	0.135	-0.122	0.459	(-1.762, 0.000)
Race	0.272	0.331	0.708	(0.000, 2.349)
Gender	0.115	-0.017	0.293	(-0.823, 0.515)
Smoking status	0.238	-0.245	0.546	(-1.880, 0.000)
Pregnancy weight gain	0.342	0.001	0.002	(0.000, 0.007)
Cholesterol	0.184	-0.002	0.005	(-0.017, 0.000)
Triglycerides	0.094	0.000	0.003	(-0.001, 0.009)
Sum of PCBs	0.104	-0.000	0.002	(-0.007, 0.000)
Socio-economic index	0.492	-0.288	0.345	(-1.001, 0.000)
Height	0.084	-0.009	0.066	(-0.215, 0.002)

Table 16: Posterior summaries of the coefficients in the regression for mother's pre-pregnancy BMI

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	38.957	4.558	(34.033, 53.397)
Adequate prenatal care	0.111	-0.027	0.152	(-0.522, 0.054)
Race	0.146	0.043	0.176	(-0.049, 0.639)
Gender	0.163	0.057	0.199	(-0.031, 0.737)
Smoking status	0.423	-0.297	0.428	(-1.316, 0.000)
Pregnancy weight gain	0.104	-0.000	0.000	(-0.001, 0.001)
Cholesterol	0.099	0.000	0.001	(-0.002, 0.003)
Triglycerides	0.230	0.001	0.003	(0.000, 0.010)
Sum of PCBs	0.098	0.000	0.001	(-0.000, 0.002)
Socio-economic index	0.224	-0.042	0.099	(-0.344, 0.000)
Height	0.180	-0.024	0.068	(-0.249, 0.000)
Pre-pregnancy BMI	0.181	0.012	0.034	(0.000, 0.123)

Table 17: Posterior summaries of the coefficients in the regression for mother's age

Predictor	incl. probability	mean	std dev	95% CI
Intercept	1.000	249.131	5468.912	(-12930.480, 7217.436)
Adequate prenatal care	0.089	14.418	88.435	(0.000, 275.790)
Race	0.584	507.475	528.756	(0.000, 1504.418)
Gender	0.134	30.835	140.263	(-50.568, 510.585)
Smoking status	0.159	45.594	143.563	(0.000, 530.637)
Pregnancy weight gain	0.132	-0.126	0.454	(-1.702, 0.000)
Cholesterol	0.512	-3.414	3.954	(-11.520, 0.000)
Triglycerides	0.084	-0.042	0.904	(-2.173, 1.065)
Sum of PCBs	0.807	4.605	2.791	(0.000, 8.686)
Socio-economic index	0.425	-98.569	134.893	(-391.975, 0.000)
Height	0.346	50.865	83.841	(0.000, 258.394)
Pre-pregnancy BMI	0.091	-1.685	14.019	(-44.336, 6.954)
Age	0.084	-4.540	28.570	(-98.243, 0.000)

Table 18: Posterior summaries of the coefficients in the regression for p,p'-DDE

References

- Albert, J. and S. Chib (1993): “Bayesian analysis of binary and polychotomous data response data,” *Journal of the American Statistical Association*, 88, 669–679.
- Cai, B. and D. Dunson (2006): “Bayesian covariance selection in generalized linear mixed models,” *Biometrics*, 62, 446–457.
- Claeskens, G. and F. Consentino (2008): “Variable Selection with Incomplete Covariate Data,” *Biometrics*, 64, 1062–1069.
- Claeskens, G. and N. L. Hjort (2008): *Model Selection and Model Averaging*, Cambridge University Press.
- Clyde, M. and E. George (2004): “Model uncertainty,” *Statistical Science*, 19, 81–94.
- Cottet, R., R. J. Kohn, and D. J. Nott (2008): “Variable Selection and Model Averaging in Semiparametric Overdispersed Generalized Linear Models,” *Journal of the American Statistical Association*, 103, 661–671.
- Eastaugh, J., S. Smye, S. Snowden, J. Walker, P. Dear, and A. Farrin (1997): “Comparison of neural networks and statistical models to predict gestational age at birth,” *Neural Computing & Applications*, 6, 158–164.
- Flegal, J. M., M. Haran, and G. L. Jones (2008): “Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?” *Statistical Science*, 23, 250–260.
- George, E. and R. McCulloch (1993): “Variable selection via gibbs sampling,” *Journal of the American Statistical Association*, 88, 881–889.
- George, E. and R. McCulloch (1997): “Approaches for bayesian variable selection,” *Statistica Sinica*, 7, 339–373.
- Ibrahim, J., S. Lipsitz, and M.-H. Chen (1999): “Missing covariates in generalized linear models when the missing data mechanism is nonignorable,” *Journal of the Royal Statistical Society, Ser. B*, 61, 173–190.
- Kinney, S. K. and D. Dunson (2007): “Fixed and random effects selection in linear and logistic models,” *Biometrics*, 63, 690–698.
- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. Berger (2008): “Mixtures of g Priors for Bayesian Variable Selection,” *Journal of the American Statistical Association*, 103, 410–423.
- Lipsitz, S. and J. Ibrahim (1996): “A conditional model for incomplete covariates in parametric regression models,” *Biometrika*, 83, 916–922.
- Little, R. J. A. and D. B. Rubin (2002): *Statistical Analysis with Missing Data: Second Edition*, New York: John Wiley & Sons.
- Longnecker, M. P., M. A. Klebanoff, H. Zhou, and J. W. Brock (2001): “Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at birth,” *Lancet*, 358, 110–114.

- Mitchell, T. and J. Beauchamp (1988): "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.
- Raftery, A. (1996): "Approximate bayes factors and accounting for model uncertainty in generalised linear models," *Biometrika*, 83, 251–266.
- Raftery, A., D. Madigan, and J. Hoeting (1997): "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92, 179–191.
- Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*, Wiley.
- van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999): "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine*, 18, 681–694.
- Verbeke, G., G. Mohlenbergs, and C. Beunckens (2008): "Formal and Informal Model Selection with Incomplete Data," *Statistical Science*, 23, 201–218.
- Yang, X., T. Bellin, and W. Boscardin (2006): "Imputation and variable selection in linear regression models with missing covariates," *Biometrics*, 61, 498–506.