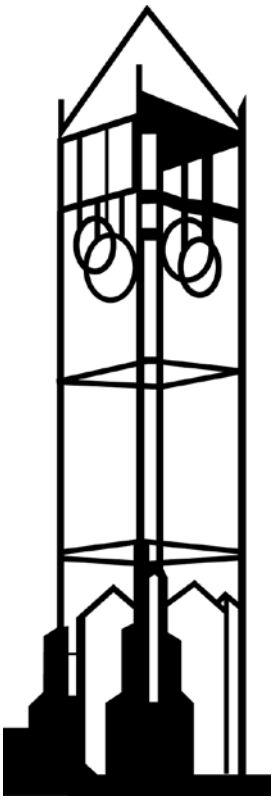
Model Uncertainty in Characterizing Recreation Demand

Babatunde Abidoye, Joseph A. Herriges



Working Paper No. 10035
October 2010

IOWA STATE UNIVERSITY
Department of Economics
Ames, Iowa, 50011-1070

Model Uncertainty in Characterizing Recreation Demand

Babatunde O. Abidoye
Economic Policy Analysis

Joseph A. Herriges[1]
Iowa State University

October 8, 2010

Abstract

A Bayesian variable selection procedure is used to control for uncertainty in the specification of a recreational demand model. In contrast to comparing models based on the likelihood values with unknown sampling properties (as in, e.g., Egan *et al*, 2009), we propose a model that draws on the Bayesian paradigm to integrate the variable selection process into the model and reflect the accompanying uncertainty about which is the "best" specification used for counterfactual predictions.

[1]Contact author information: 260 Heady Hall, Department of Economics, Iowa State University, Ames, IA 50011. email: *jaherrig@iastate.edu*. Phone: 515-294-4964

1 Introduction

Analysts and policymakers are often interested in understanding the impact that changing environmental conditions can have on the demand for recreational activities and quantifying the welfare implications of these changes. This information can be used to more efficiently direct scarce resources aimed at maintaining and restoring environmental quality. The modeling of recreation demand typically involves the specification of a functional relationship between individual demand and observable individual and site characteristics. Unfortunately, economic theory provides relatively little guidance regarding the form that this relationship should take and which variables ought to be included in the analysis. In many applications, limitations in the available data (e.g., describing the water quality conditions at a lake site) narrow the range of possibilities, but choices must still be made between, for example, level and logarithmic specifications for an environmental characteristic. The choices made by the researcher can have significant impact on the policy implications drawn from their analysis of recreational usage patterns.

While model selection criteria can be used to narrow the set of specifications, there is the risk that the analyst (even inadvertently) may engage in a "fishing" process among the available models, biasing the final outcome of the analysis. In a recent paper, Egan *et al.* (2009) attempt to ameliorate this problem by employing a split sample approach, using separate portions of the available data for model specification, estimation, and evaluation. They isolated one third of their sample in order to consider alternative models and functional forms, using a likelihood dominance criteria to pick their final model, which is in turn estimated using a separate sample. The final third of their sample was used for out-of-sample predictions. Though this approach can arguably reduce the impact of the specification search process on the final parameter estimates, it does not eliminate the problem. More importantly, the procedure inevitably requires the selection of a single model and does not account for the uncertainty in this process. Indeed, their selection of the final model is not based on a test among competing models (as the alternatives are non-nested), but on a log-likelihood based ranking.

In this paper, we consider an alternative approach that draws on the Bayesian paradigm to integrate the variable selection process into the model and to reflect the accompanying uncertainty about which is the "correct" specification into subsequent counterfactual predictions. Specifically, we describe a Bayesian posterior simulator that combines the literature on hierarchical modeling, Bayesian variable selection and data augmentation. Our underlying modeling framework is the class of repeated random utility models (See, e.g., Herriges

and Phaneuf, 2002) and follows closely the model proposed in Abidoye, Herriges and Tobias (2010). We then employ the stochastic search variable selection (SSVS) method described in George and McCulloch (1993) to determine the posterior probability that individual site characteristics influence the site selection decision. The model can be used to identify a preferred model specification. Alternatively, and we would argue preferably, the model can be used as part of the process of employing a Bayesian model averaging, integrating competing models into a single structure that can be used for welfare analysis and counterfactual predictions. The model is applied using data from the 2002 survey of Iowa Lakes Project, the same data underlying both the Egan *et al.* (2009) and the Abidoye, Herriges and Tobias (2010) analyses. We use our model to contrast our findings with those obtained from these earlier, highlighting the benefits of integrating model uncertainty into a unified framework. One advantage of this study is that we have large number of sites (130) and detailed information on both site attributes and lake water quality.

The outline of the chapter is as follows. Section 2 touches on the issue of model uncertainty in econometric analysis and also frames our approach in the context of other methods in the literature. Section 3 presents the model and how the parameters of interest are estimated. Section 4 describes a generated data experiment as a check for the performance of the sampler. Section 5 describes the data and application and section 6 provides posterior simulation and welfare analysis. The chapter concludes with a summary in section 7.

2 Related Literature

2.1 Model uncertainty

Researchers are often faced with the dilemma of which model specification or subset of explanatory variables will best fit their data. This problem is more pronounced in situations where economic theory does not dictate *a priori* the specific functional form or distributional assumption to be used. The inability to lay claim to a "best" model makes inference on the chosen model less certain and potentially inaccurate. This has led to widespread criticism of estimates presented for a "best" model (e.g., Leamer, 1983). For example, changing from linear to nonlinear specification or changing the functional form of some variables can lead to substantially different estimates. A number of studies, including Regal and Hook (1991) and Draper (1995), have shown the impact of ignoring uncertainty of the model on inference.

Various techniques has been proposed in the literature to account for this problem. The paper by Raftery (1995), among others, argues that the use of *p-values*, R^2 and other statistical tests based on them to search for the "best" model can lead to misleading inference and prediction. Poirier (1995) also discusses the problems with using hypothesis testing to select a specific model especially given that the procedure of pretesting introduces a level of uncertainty into the pretest estimator. Aside from the problem of choosing the significance level and balancing it with the power of the alternative hypothesis, most studies involve comparing more than two models. The sampling properties of the popular stepwise regression are usually unknown and making inference based on a model selected in this way is potentially misleading.

A solution to the model specification problem that has gained popularity among researchers in recent years is the use of Bayesian model selection and/or averaging. Bayesian model selection methods are used to select a model(s) with maximum posterior probabilities conditional on the data. Bayesian Model averaging (BMA), on the other hand, employs the rules of conditional probability to estimate a posterior probability for each considered model, with these probabilities used as weights in averaging results over all the models. The enormous number of possible explanatory variables and nonlinearity makes the use of model selection important for reducing the size of possible models before averaging among the most probable models. Variable selection methods can also be used to select a specific model (e.g., Raftery, 1995). There are a number of papers in the literature that have applied BMA in economics.[2] In the environmental and resource literature, some of the papers include Clyde (2000), Clyde, Guttorp and Sullivan (2000), Koop and Tole (2004), Layton and Lee (2006), Fernandez, Ley and Steel (2002), and Leon and Leon (2003). The message of all these papers is that model uncertainty can have a substantial impact on parameter estimates and should be accounted for explicitly.

For problems related to uncertainty regarding which predictors to include in a model, the stochastic search variable selection (SSVS) method proposed by George and McCulloch (1993) provides an insightful and easily implemented approach. The model works by capturing the entire range of possible model setups in a hierarchical Bayes mixture model. A series of latent binary variables $(\lambda_k, k = 0, \ldots, K)$ are used to indicate whether the data support inclusion of a given explanatory variable in the model. These latent variables are used to nest all of the possible models. The number of *visits* to a model including variable k through the course of an iterative sampling (Gibbs) process (i.e., the number of times $\lambda_k = 1$ versus $\lambda_k = 0$) determines how promising that variable is. SSVS makes use of both practical and statistical relevance of the model to select the "best" possible models. A major

[2]There are number of websites that are devoted to posting developments and research in this area. See http://www.research.att.com/~volinsky/bma.html for some of the papers and software.

practical advantage of the SSVS approach is that the researcher does not have to calculate the marginal likelihoods for each of the possible models.[3]

2.2 Model uncertainty in recreation demand

One primary reason for estimating recreation demand models is to quantify how site attributes (especially environmental attributes) influence the numbers of visits to the alternative sites. This is essential for policy analysis. Model estimates are used to justify important environmental policies such as pollution abatement programs. However, economic theory provides little or no guidance as to which characteristics should be in the model and subsequent welfare analysis. There have been relatively few studies to date addressing the issue of model uncertainty. Layton and Lee (2006) apply the procedure suggested by Buckland, Burnham and Augustin (1997) to control for model uncertainty in analyzing responses to a stated preference (SP) survey of saltwater angling in Alaska. They estimate weights for different model specifications and use those weights to calculate the expected willingness to pay. One problem with this procedure is that model uncertainty is incorporated *ex post* and does not account for uncertainty in the estimates of the parameters of the model.

As noted above, Egan *et al.* (2009) provide a split sample investigation into recreation demand model specification. Using data from the Iowa Lakes Project, including seven water quality measures, they consider thirty-two competing formulations of a repeated mixed logit model of Iowa lake usage. All of the models include the seven water quality measures, but differ in terms of whether these variables appeared in level or logarithmic form.[4] The "preferred" model was chosen based on the resulting log-likelihood values obtained using the first third of the sample and then re-estimated using the second third of the sample. While this does reduce the "fishing" problem, it still does not incorporate uncertainty in the final model estimated.

In this paper, we present a Random Utility Maximization (RUM) model that incorporates model uncertainty on the specification of site attributes in recreation demand. Specifically we apply the SSVS algorithm to identify the probability that a model is supported by the data.

[3]The marginal likelihood defined as $P(Y|m=j) = \int \mathbf{P}(Y|\theta_j)\mathbf{P}(\theta_j)d\theta_j$ are often difficult to estimate.

[4]To reduce the number of possible models, Total Nitrogen and Total Phosphorous are grouped together (i.e., always appearing in the same form), as are Inorganic and Organic Suspended Solids. The authors also investigate which single or pair of variables, when added to the model, yields the greatest increase in the log-likelihood function.

3 Model

As described in the previous sections, the model we present in this study incorporates model uncertainty in the site characteristics attributes in recreation demand. In addition, we want our model to be relatively flexible for posterior inference including welfare analysis. For the purpose of our model, we index individuals by $i = 1, 2, \ldots, N$, choice occasions by $t = 1, 2, \ldots, T$ and sites by $j = 1, 2, \ldots, J$.

3.1 Basic Structure

The model is similar to the repeated nested logit model (Morey, Rowe and Watson (1993)) and repeated mixed logit model (Herriges and Phaneuf (2002)). These models integrate individuals' choice among alternatives and the problem of allocating time between multiple recreation sites. The model of Morey, Rowe and Watson (1993) assumes that individuals face the decision to participate in recreation activities over fixed discrete occasions and at most one trip is taken at such an occasion. Furthermore, each decision is assumed conditionally independent across individuals and choice occasions. A summary of this framework and implications of the assumptions is presented in Herriges, Kling and Phaneuf (1999).

Formally, we assume that an individual i at choice occasion t has to choose among J sites and also inactivity, or "staying at home." We represent the utility that an individual derives from making a particular choice on a given choice occasion as:

$$U_{ijt} = \begin{cases} \boldsymbol{Z}_i\gamma + \varepsilon_{ijt} & \text{if } j = 0 \text{ , i.e., stay at home} \\ \alpha_j + P_{ij}\beta + \varphi_i + \varepsilon_{ijt} & \text{for } j = 1, \ldots, J. \end{cases} \tag{1}$$

where α_j is the overall site-specific effect; β is the marginal utility of income; φ_i captures the individual specific effect and ε_{ijt} represents an idiosyncratic error that is assumed to be independent across the $J+1$ alternatives with variance normalized such that $\varepsilon_{ijt} \sim N(0, 1)$. We also assume that the demographic characteristics of an agent ($\boldsymbol{Z}_i$) have an effect on the likelihood of choosing the "stay at home" option, but not on the choice among recreation sites. Note that site attributes (e.g., water quality, facilities, etc.) do not appear directly in (1), but rather are subsumed in the α_j's (i.e., the alternative specific constants). As has been noted elsewhere (e.g., Murdock (2006) and Abidoye, Herriges and Tobias (2010)), including a full set of alternative specific constants controls for both observable and unobservable

site attributes and insulates the travel cost parameter from potential omitted variables bias stemming from unobserved site attributes. In Murdock (2006), the site attributes are linked to the alternative specific constants using a secondary regression. Similar to Abidoye, Herriges and Tobias (2010), we use a hierarchical structure, described in Section 3.2 below, to capture the impact of site attributes on the α_j's.

Given that it is the difference in utility that matters, we use the "stay at home" option as the base case and take the difference in utilities. Thus,

$$\tilde{U}_{ijt} = \alpha_j + P_{ij}\beta - \boldsymbol{Z}_i\gamma + \varphi_i + \tilde{\varepsilon}_{ijt} \tag{2}$$

where $\tilde{U}_{ijt} = U_{ijt} - U_{i0t}$; $\tilde{\varepsilon}_{ijt} = \varepsilon_{ijt} - \varepsilon_{i0t}$; for $j = 1, ..., J$. So that

$$\tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{i.t}} = \begin{bmatrix} \varepsilon_{i1t} - \varepsilon_{i0t} \\ \varepsilon_{i2t} - \varepsilon_{i0t} \\ \vdots \\ \varepsilon_{iJt} - \varepsilon_{i0t} \end{bmatrix} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}^*\right)$$

where

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ 1 & 1 & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{bmatrix}.$$

The observed choice y_{it} is linked to the latent variable vector $\tilde{\boldsymbol{U}}_{\boldsymbol{i\cdot t}}$ as follows:

$$y_{it}(\tilde{\boldsymbol{U}}_{\boldsymbol{i\cdot t}}) = \begin{cases} 0 \text{ if } \max\{\tilde{U}_{ijt}\}_{j=1}^{J} \leq 0 \\ k \text{ if } \max\{\tilde{U}_{ijt}\}_{j=1}^{J} = \tilde{U}_{ikt} > 0. \end{cases} \tag{3}$$

Stacking over the alternatives, we have:

$$\tilde{\boldsymbol{U}}_{\boldsymbol{i.t}} = \boldsymbol{\alpha} + \boldsymbol{P}_{\boldsymbol{i.}}\boldsymbol{\beta} - (\mathbf{1}_{\boldsymbol{J}} \otimes \boldsymbol{Z}_i)\gamma + \mathbf{1}_{\boldsymbol{J}}\varphi_i + \tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{i.t}}. \tag{4}$$

where $\mathbf{1}_{\boldsymbol{J}}$ is a $J \times 1$ vector of ones,

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_J \end{bmatrix}; \tilde{\boldsymbol{U}}_{\boldsymbol{i.t}} \begin{bmatrix} \tilde{U}_{i1t} \\ \tilde{U}_{i2t} \\ \vdots \\ \tilde{U}_{iJt} \end{bmatrix} \text{ and } \boldsymbol{P}_{\boldsymbol{i.}} = \begin{bmatrix} P_{i1} \\ P_{i2} \\ \vdots \\ P_{iJ} \end{bmatrix}.$$

We can then re-write the above equation concisely as

$$\tilde{\boldsymbol{U}}_{\boldsymbol{i.t}} = \boldsymbol{M}_{\boldsymbol{i.t}}\boldsymbol{\theta} + \mathbf{1}_{\boldsymbol{J}}\varphi_i + \tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{i.t}} \tag{5}$$

where

$$\boldsymbol{M}_{\boldsymbol{i.t}} = \begin{bmatrix} \boldsymbol{I}_{\boldsymbol{J}} & \boldsymbol{P}_{\boldsymbol{i.}} & \mathbf{1}_{\boldsymbol{J}} \otimes \boldsymbol{Z}_i \end{bmatrix}; \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha}' & \boldsymbol{\beta}' & \boldsymbol{\gamma}' \end{bmatrix}' .$$

Another way to write equation (5) is in terms of the error component. That is:

$$\tilde{\boldsymbol{U}}_{\boldsymbol{i.t}} = \boldsymbol{M}_{\boldsymbol{i.t}}\boldsymbol{\theta} + \boldsymbol{v}_{\boldsymbol{i.t}}$$

where

$$\begin{aligned} \boldsymbol{v}_{\boldsymbol{i.t}} &= \mathbf{1}_{\boldsymbol{J}}\varphi_i + \tilde{\boldsymbol{\varepsilon}}_{\boldsymbol{i.t}} \\ E(\boldsymbol{v}_{\boldsymbol{i.t}}\boldsymbol{v}_{\boldsymbol{i.t}}{}') &\equiv \boldsymbol{\Omega} = \sigma_{\varphi}^2 \mathbf{1}_{\boldsymbol{J}}\mathbf{1}'_{\boldsymbol{J}} + \boldsymbol{\Sigma}^{*}. \end{aligned}$$

3.2 Hierarchical Priors

As described earlier, the α_j's captures the overall site-specific effect. Given that these depend on the characteristics of the site, we specify an hierarchical prior on α_j with the assumption that its mean is the aggregate effect of the observed attributes, with the unobserved site characteristics determining deviations from the mean. Formally, the priors for the site-specific parameters is specified as:

$$\alpha_j \sim N(\boldsymbol{Q}_{\boldsymbol{j}}\boldsymbol{\alpha}_{\mathbf{0}}, \sigma_{\alpha}^2). \quad j = 1, 2, ...J \tag{6}$$

where $\boldsymbol{Q}_{\boldsymbol{j}}$ is a $1 \times (K+1)$ vector including a constant term and the K observed site characteristics that potentially influence demand for site j.

In investigating model uncertainty, we focus our attention on the parameters associated with the observed site attributes (i.e., the $\alpha_{0,k}$'s). We seek to calculate the probability that a given variable (or combination of variables) belong in the model using the SSVS approach. If a variable k is not supported by the data, we will expect that the true value of the parameter ($\alpha_{0,k}$) be zero. To capture this we introduce an additional level to the hierarchical structure described in (6). Following George and McCulloch (1993), we specify a prior for each regression coefficients ($\alpha_{0,k}$) as a mixture of two normal distributions with different variances and zero mean. That is conditional on a binary latent variable $\lambda_k = 0$ or 1, each k element of $\boldsymbol{\alpha}_{\mathbf{0}}$ can be defined as:

$$\alpha_{0,k}|\lambda_k \sim (1-\lambda_k)N(0,\tau_k^2) + \lambda_k N(0, c_k^2\tau_k^2) \tag{7}$$

and

$$P(\lambda_k = 1) = 1 - P(\lambda_k = 0) = p_k; \quad 0 \leq p_k \leq 1. \tag{8}$$

λ_k is a latent binary variable that indicates if the observed site characteristics is supported by the data or not. With the above representation, when $\lambda_k = 0$, $\alpha_{0,k} \sim N(0, \tau_k^2)$, whereas $\alpha_{0,k} \sim N(0, c_k^2\tau_k^2)$ when $\lambda_k = 1$. The variance term for the first normal distribution (τ_k^2) is assumed to be very small such that the distribution of the $\alpha_{0,k}$ is massed around zero, providing little evidence for its inclusion in the model. The second variance ($c_k^2\tau_k^2$), on the other hand, is large and signals evidence that the variable should be included in the model. p_k can be thought of as the prior probability that variable k should be included in the model. Thus, the prior on $\boldsymbol{\alpha_0}$ is represented as multivariate normal:

$$\boldsymbol{\alpha_0}|\boldsymbol{\lambda} \sim N_k(\mathbf{0}, \boldsymbol{D_\lambda V_\alpha D_\lambda}) \tag{9}$$

where $\boldsymbol{\lambda} = (\lambda_0, ..., \lambda_K)$, V_α is the prior correlation matrix and $\boldsymbol{D_\lambda} \equiv diag[L_0\tau_0, ..., L_K\tau_K]$, with $L_k = 1$ if $\lambda_k = 0$ and $L_k = c_k$ if $\lambda_k = 1$. $\boldsymbol{D_\lambda}$ is like a tuning parameter that ensures that the prior on $\alpha_{0,k}$ holds.

Finally, we set priors for the other parameters as

$$\begin{aligned} p(\boldsymbol{\lambda}) &= \prod_{k=1}^{K} p_k^{\lambda_k}(1-p_k)^{1-\lambda_k} && (10)\\ \sigma_\alpha^2 &\sim IG(a_\alpha, b_\alpha) && (11)\\ \sigma_\varphi^2 &\sim IG(a_\varphi, b_\varphi) && (12)\\ \boldsymbol{\gamma} &\sim N(\boldsymbol{\mu_\gamma}, \boldsymbol{V_\gamma}). && (13) \end{aligned}$$

The hyperparameters of the priors above are supplied by the researcher and are in general chosen to be relatively vague to allow dominance of the information from the data. The prior means (μ_β, $\boldsymbol{\mu_\gamma}$) in our empirical work and generated data experiments are set to zero vectors of appropriate dimensions with the respective prior variance for the parameters ($\boldsymbol{V_\alpha}$, V_β, and $\boldsymbol{V_\gamma}$) set to identity matrices of the appropriate dimensions. The hyperparameters of the variances are also chosen to have a reasonably non-informative prior for the variances.

3.3 Posterior Simulator

The posterior simulator uses the Gibbs sampler to generate draws from the posterior distribution for the parameters of our model.[5] In this subsection, we derive the necessary posterior conditionals and describe how to generate draws from these distributions. While the joint posterior distribution is complex, the conditional posterior distributions used in the Gibbs sampler take recognizable forms and are easy to draw from.

Let

$$\boldsymbol{\Xi} = \begin{bmatrix} \boldsymbol{\theta} & \boldsymbol{\alpha_0} & \boldsymbol{\lambda} & \sigma_\alpha^2 & \boldsymbol{\varphi}_\cdot & \sigma_\varphi^2 \end{bmatrix}$$

denote all the parameters of the model with $\boldsymbol{\varphi}_\cdot$ denoting φ_i stacked over individuals. The joint posterior distribution of $\boldsymbol{\Xi}$ and the latent utility data $\tilde{\boldsymbol{U}}$ gives us the posterior density for the parameters in our model. We use blocking steps (e.g., Chib and Carlin, 1999) to obtain draws from the joint posterior conditional of the individual random effects and the site specific effects to improve the mixing of the sampler.

Using Bayes theorem, we can write the posterior density as:

$$\begin{aligned} p(\Xi, \tilde{\boldsymbol{U}}|\boldsymbol{y}) \propto & \prod_{t=1}^{T}\prod_{i=1}^{N} \phi(\tilde{\boldsymbol{U}}_{\boldsymbol{i.t}}, \boldsymbol{M_{i.t}}\boldsymbol{\theta}, \boldsymbol{\Omega}) \\ & \times \left\langle I(y_{i.t} = j) I(\tilde{U}_{ijt} > max[\tilde{U}_{i,-j,t}, 0]) + I(y_{i.t} \neq j) I(\tilde{U}_{ijt} < max[\tilde{U}_{i,-j,t}, 0]) \right\rangle \\ & \times \left[\prod_{j=1}^{J} p(\alpha_j|\boldsymbol{\alpha}_0, \boldsymbol{\lambda}, \sigma_\alpha^2)\right] \left[\prod_{i=1}^{N} p(\varphi_i|\sigma_\varphi^2)\right] p(\boldsymbol{\alpha_0}|\boldsymbol{\lambda}) p(\beta) p(\boldsymbol{\gamma}) p(\boldsymbol{\alpha}_0) p(\sigma_\alpha^2) p(\sigma_\varphi^2) p(\boldsymbol{\lambda}). \end{aligned} \tag{14}$$

We outline each posterior conditional distribution below.

Step 1: Draw the hierarchical parameter conditional on the latent utility and the hierarchical prior $(\boldsymbol{\theta}|\boldsymbol{\Xi}_{-\boldsymbol{\theta}}, \tilde{\boldsymbol{U}}, \boldsymbol{y})$ using the results of Lindley and Smith (1972) with blocking step.[6] The posterior conditional for $\boldsymbol{\theta}$ is given as:

$$\boldsymbol{\theta}|\boldsymbol{\Xi}_{-\boldsymbol{\theta}}, \tilde{\boldsymbol{U}}, \boldsymbol{y} \sim N(\mathbf{D}_\theta \mathbf{d}_\theta, \mathbf{D}_\theta). \tag{15}$$

[5] The simulator itself was programmed in MATLAB and the associated code is available from the authors upon request.

[6] The notation $\boldsymbol{\Xi}_{-\boldsymbol{a}}$ is used to denote the vector $\boldsymbol{\Xi}$ excluding the parameters in $\boldsymbol{a}$.

where

$$\mathbf{D}_\theta \equiv \left[T \sum_{i=1}^{N} \mathbf{M}_{\mathbf{it}}' \mathbf{\Omega}^{-\mathbf{1}} \mathbf{M}_{\mathbf{it}} + \mathbf{\Sigma}_\theta^{-\mathbf{1}} \right]^{-1}$$
$$\mathbf{d}_\theta \equiv \sum_t \sum_i \mathbf{M}_{\mathbf{it}}' \mathbf{\Omega}^{-\mathbf{1}} \mathbf{w}_{\mathbf{it}} + \mathbf{\Sigma}_\theta^{-\mathbf{1}} \mu_\theta$$

and

$$\mathbf{\Sigma}_\theta = \begin{bmatrix} \sigma_\alpha^2 I_J & 0 & \mathbf{0} \\ \mathbf{0} & V_\beta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_\gamma \end{bmatrix}, \ \mu_\theta = \begin{bmatrix} \mathbf{Q}\boldsymbol{\alpha_0} \\ \mu_\beta \\ \boldsymbol{\mu_\gamma} \end{bmatrix}.$$

Step 2: $\boldsymbol{\alpha}_0 | \mathbf{\Xi}_{-\boldsymbol{\alpha_0}}, \tilde{\boldsymbol{U}}, \boldsymbol{y}$

Once we condition on the $\boldsymbol{\alpha}$, the posterior conditional for $\boldsymbol{\alpha}_0$ is similar to that of a linear regression parameter. However, the introduction of the latent variable λ into the model specification helps account for model uncertainty such that less weight is put on specifications not supported by the data.

$$\boldsymbol{\alpha}_0 | \mathbf{\Xi}_{-\boldsymbol{\alpha_0}}, \tilde{\boldsymbol{U}}, \boldsymbol{y} \sim N(\mathbf{D}_{\alpha_\mathbf{0}} \mathbf{d}_{\alpha_\mathbf{0}}, \mathbf{D}_{\alpha_\mathbf{0}}) \tag{16}$$

where

$$\mathbf{D}_{\alpha_\mathbf{0}} = (\mathbf{Q}'\mathbf{Q}/\sigma_\alpha^2 + (\boldsymbol{D_\lambda V_\alpha D_\lambda})^{-1})^{-1} \text{ and } \mathbf{d}_{\alpha_\mathbf{0}} = \mathbf{Q}'\boldsymbol{\alpha}/\sigma_\alpha^2 + (\boldsymbol{D_\lambda V_\alpha D_\lambda})^{-1} \boldsymbol{\mu_\alpha}.$$

Step 3: $\sigma_\alpha^2 | \mathbf{\Xi}_{-\sigma_\alpha^2}, \tilde{\boldsymbol{U}}, \boldsymbol{y}$

$$\sigma_\alpha^2 | \mathbf{\Xi}_{-\sigma_\alpha^2}, \tilde{\boldsymbol{U}}, \boldsymbol{y} \sim IG \left[\frac{J}{2} + a_\alpha, \left(b_\alpha^{-1} + .5 \sum_{j=1}^{J} (\alpha_j - \mathbf{Q_j}\boldsymbol{\alpha_0})^2 \right)^{-1} \right]. \tag{17}$$

Step 4: Draw the λ_k

As described earlier, the marginal posterior distribution $p(\boldsymbol{\lambda}|\alpha)$ carries information on the relevance of each model and variable specification. However, since the only link between λ and the alternative specific constants (α) is through the mean parameters α_0, the distribution of λ_k simplifies to a Bernoulli distribution with probability

$$P(\lambda_k | \alpha_0, \boldsymbol{\lambda_{-k}}) = \frac{p(\boldsymbol{\alpha_0} | \boldsymbol{\lambda_{-k}}, \lambda_k = 1) p_k}{p(\boldsymbol{\alpha_0} | \boldsymbol{\lambda_{-k}}, \lambda_k = 0)(1 - p_k)} \tag{18}$$

where $\boldsymbol{\lambda_{-k}}$ represents all $\boldsymbol{\lambda}$ except λ_k.

Step 5: $\boldsymbol{\varphi}_{\cdot}|\boldsymbol{\Xi}_{-\varphi_i}, \tilde{\boldsymbol{U}}, \boldsymbol{y}$

$$\boldsymbol{\varphi}_{\cdot}|\boldsymbol{\Xi}_{-\varphi_i}, \tilde{\boldsymbol{U}}, \boldsymbol{y} \sim N(D_\varphi d_\varphi, D_\varphi) \tag{19}$$

where

$$D_\varphi^{-1} = JT + \frac{1}{\sigma_\varphi};\ \text{and}\ d_\varphi = \sum_{t=1}^{T}(\mathbf{U}_{\mathbf{i.t}}^{\varphi} - \mathbf{M}_{\mathbf{i.t}}^{\varphi}\theta^{\varphi})$$

and $\mathbf{U}_{\mathbf{i.t}}^{\varphi}, \mathbf{M}_{\mathbf{i.t}}^{\varphi}$, and θ^{φ} are stacked over the sites j ($j = 1...J$) and choice occasion for each individual without the stay at home equation. That is

$$\mathbf{M}_{\mathbf{i.t}}^{\varphi} = \begin{bmatrix}\mathbf{I_J} & \mathbf{P_i}\end{bmatrix}; \boldsymbol{\theta^{\varphi}} = \begin{bmatrix}\boldsymbol{\alpha}_{\cdot}' & \beta\end{bmatrix}'.$$

Step 6: $\sigma_\varphi^2|\boldsymbol{\Xi}_{-\sigma_\varphi^2}, \tilde{\boldsymbol{U}}_{\boldsymbol{i.t}}$

$$\sigma_\varphi^2|\boldsymbol{\Xi}_{-\sigma_\varphi^2}, \tilde{\boldsymbol{U}}_{\boldsymbol{i.t}} \sim IG\left[\frac{N}{2} + \alpha_\varphi, \left(b_\varphi^{-1} + .5\sum_{i=1}^{N}\varphi_i^2\right)^{-1}\right]. \tag{20}$$

Step 7: Draw the $\tilde{\boldsymbol{U}}_{\boldsymbol{i\cdot t}}|\boldsymbol{\Xi}, \boldsymbol{y}$

Given the structure of our model and to ease computation, we draw the latent utilities that individual i derives from visiting site j using utility levels instead of differences. That is, we sample the U_{ijt} and then take the differences to get the $\tilde{U}_{ijt}$. At the structural level of the U_{ijt} in equation (1), there is no correlation among the alternatives conditional on α_j, β, γ, and $\boldsymbol{\varphi}_{\cdot}$.

Each of the U_{ijt}'s are conditionally normal with mean μ and variance of 1 with truncation point that depends on the choice of the individual. That is, if an alternative is chosen, it must be the alternative that gives the maximum utility - this gives the upper truncation point for all the other alternatives.

We therefore follow the following steps to draw the $\tilde{U}_{ijt}$'s at a given draw r :

Assuming that individual i chooses alternative k at choice occasion t,

1: Draw U_{ijt}^r for all $j \neq k$ from a truncated normal distribution with mean and variance from equation (1) and upper truncation point $U_{ikt} = U_{ikt}^{r-1}$.

2: Draw U_{ikt}^r from a truncated normal distribution with its mean and variance with lower truncation point at the $\max(U_{ijt}^r)$ for all $j \neq k$.

3: Calculate $\tilde{U}_{ijt}$ by taking the difference between utilities from all sites and the stay at home option: $\tilde{U}_{ijt}^r = U_{ijt}^r - U_{ibt}^r$.

4 Generated Data experiment

In this section we illustrate the performance of the algorithm described above in accounting for model uncertainty in a RUM model of recreation demand.[7] Specifically, we generated a pseudo-data set consisting of $N = 3000$ individuals who are assumed to choose among $J = 10$ sites and the "stay at home" option on each of $T = 52$ choice occasions. The vector of individual characteristics in equation (1) (i.e., $\boldsymbol{Z_i}$) consisted of a uniform random variable that signifies the age of the individual and a gender dummy variable generated from a Bernoulli distribution with equal probability of success and failure. The alternative specific constant for site j (α_j) was drawn from a normal distribution with mean $\boldsymbol{Q_j \alpha_0}$ and variance $\sigma_\alpha^2 = 0.25$ where $\boldsymbol{Q_j}$ included an intercept term and a uniformly generated random variable $\boldsymbol{Q_{j,1}}$, which can be thought of as water pollution. Travel costs for each individual/site combination were generated as a linear combination of a standard normal variable and the alternative specific constants (i.e., the α_j's), thus inducing correlation between travel costs and unobserved site characteristics. The remaining parameters of the model in equation (1) were fixed (with their values reported in Table 1). These were then used to generate the latent utility values U_{ijt} for $j = 0, .., J$ which were in turn mapped into the observed choice of the individuals.

Two experiments were conducted using the pseudo-data set. In the first experiment, we considered the inclusion of an (erroneous) additional predictor ($\boldsymbol{Q_{j,2}}$) when modeling recreation demand, where $\boldsymbol{Q_{j,2}}$ is generated from a standard normal distribution that is independent of $\boldsymbol{Q_{j,1}}$. In a second experiment, the added predictor is generated such that it is equal to $\boldsymbol{Q_{j,1}}$ plus a randomly generated uniformly distributed variable. That is $\boldsymbol{Q_{j,2}} = \boldsymbol{Q_{j,1}} + U(0, 1)$. This is to test the performance of our model when high correlation exists between the two

[7] We also use the generated data as a guide to know how many draws will be needed for our application to achieve the same level of precision under independent and identical distribution (*iid*) sampling.

observed site characteristics, as can be the case in practice with some water quality measures and other site attributes.

The Gibbs sampler described in section 3.3 was implemented using 50000 iterations, with 5,000 iterations discarded as burn-in. The results are presented in Table 1. The table reports both the posterior mean for each parameter and its posterior probability of being positive [denoted $P(\cdot > 0|y)$]. Starting with experiment #1, in which an additional site attribute $\boldsymbol{Q_{j,2}}$ (uncorrelated with $\boldsymbol{Q_{j,1}}$) is erroneously included in the analysis, we see that posterior means of the parameters are all close to their true values. Moreover, all of the posterior means lie well within two standard deviations of their true values. The posterior distribution for the parameter associated with the site characteristic $\boldsymbol{Q_{j,1}}$ (i.e., $\alpha_{0,1}$) is largely bounded away from zero, with approximately ninety percent of the posterior distribution being negative. In contrast, the the posterior distribution for the parameter associated with the erroneously included site characteristic $\boldsymbol{Q_{j,2}}$ (i.e., $\alpha_{0,2}$) is more evenly distributed between positive and negative values. Figure 1 and 2 provides graphical depictions of the posterior distributions for these two parameters. The distribution for $\alpha_{0,1}$, as expected, looks like a mixture of two normal distribution with majority of the mass in the distribution being negative. However, the distribution for $\alpha_{0,2}$ is largely massed around zero.[8] The draws for the latent binary variables (i.e., the λ_k's) confirm these basic findings. For the intercept term, $\lambda_0 = 1$ appeared in almost all iterations (49,976 out of the 50,000 times), while $\lambda_1 = 1$ 36,038 times for the correctly included $\boldsymbol{Q_{j,1}}$. In contrast, for the erroneously added predictor $\boldsymbol{Q_{j,2}}$, we find that $\lambda_2 = 1$ only 9400 times, signaling that the variable is not a promising part of the model.

Turning to the second experiment, in which the added variable $\boldsymbol{Q_{j,2}}$ can act as a proxy for $\boldsymbol{Q_{.,1}}$ given the high level of correlation (0.92) between the two variables, the posterior means are again generally close to their true values, with the important exception of the $\alpha_{0,k}$'s. Not surprisingly, the posterior distribution has a difficult time clearly isolating the contribution of $\boldsymbol{Q_{.,1}}$ to the appeal of a given site. Indeed, while we find that one of the two variables is almost always visited, $\lambda_1 = 1$ in only 47% of the iterations, while $\lambda_2 = 1$ 65% of the time. The simulated posterior distributions for $\alpha_{0,1}$ and $\alpha_{0,2}$ are presented in Figures 3 and 4, respectively.

The benefit of controlling for model uncertainty can be seen if we naively estimate the model assuming that both site attributes $\boldsymbol{Q_{j,1}}$ and $\boldsymbol{Q_{j,2}}$ should be included in the analysis. The result, reported in the last two columns of Table 1, suggest that even the posterior mean for $\alpha_{0,2}$ is massed away from zero (Figure 5). However, policies directed to improve $\boldsymbol{Q_{j,2}}$ would be a waste of resources.

[8] Since the variable was included in the model 9,400 times, the distribution is not fully centered on zero.

5 Application

The methods described above is applied to data from the Iowa lakes Valuation project at Iowa State University. This is the same data described in Egan *et al.* (2009). The Iowa Lakes Project is a four year panel data study, sponsored by the Iowa Department of Natural Resources and the US EPA, eliciting the visitation patterns of Iowan residents to the primary recreational lakes in the state. The data set is appropriate for our study for a number of reasons. The Iowa Lakes Project not only covers all the major lakes in the state but also provides information on a wide variety of site characteristics. The observed site characteristics ($\boldsymbol{Q}$) include both site attributes, such as lake acreage and indicators for paved boat ramps and handicap accessibility, and an unusually large number of water quality attributes, such as Secchi Transparency (a measure of the depth of water clarity), Nitrogen, and Chlorophyll.[9] In addition, the exact same data was used by Egan *et al.* (2009) to investigate model specification and by Abidoye, Herriges and Tobias (2010) to investigate the importance of controlling for unobserved site attributes in models of recreation demand. Whereas Egan *et al.* (2009) found water quality to significantly impact recreation demand, Abidoye, Herriges and Tobias (2010) concluded that this result is no longer clear once the analysis include a full set of alternative specific constants in the model to control for unobserved site attributes.

Although data for the project was collected over a four year period (2002-2005), we focus on the 2002 survey. The initial survey was sent by mail to 8,000 randomly selected Iowa residents. The response rate among deliverable surveys was 62%, yielding a total of 4,423 returned surveys. We exclude from our analysis those individuals who (a) were not Iowa residents (42), (b) failed to complete the section of the survey asking for lake visitation patterns (360), or (c) reported taking more than fifty-two day trips per-year (223). The latter sample exclusion follows the procedure used in Egan *et al.* (2009), wherein the authors note that individuals taking such frequent trips are usually local residents who are counting casual visits to or the passing by of their local lake. Instead, our analysis, like theirs, is concerned with day-trips taken to lake sites solely for the purpose of recreation.[10] The cut-off of fifty-two trips per year allows for a day-trip each week.

Table 2 provides summary statistics for our sample, both in terms of household demographics

[9]The water quality attributes were measured by Iowa State University's Limnology Laboratory three times a year at each lake. The values used in our analysis are simple averages of these measures, following the approach used in Egan *et al.* (2009).

[10]Egan *et al.* (2009) also found that their qualitative results were not sensitive to the specific cut-off of fifty-two trips per year.

and individual site characteristics. As the table indicates, the survey respondents in our data set are, on average, older males with some college or trade/vocational school. The average household size is 2.61. Travel cost (P_{ij}) is calculated using 25 cents per mile for the round-trip travel distance [computed using *PCMiler (Streets Version 17)*] plus one-third the respondent's wage rate multiplied by the travel time.[11] Overall, round-trip travel costs average just under $140, ranging from less than $1 to $1366.

One of the appealing features of the Iowa Lakes Project is that, not only is there a wealth of information available regarding the site attributes and lake water quality, but there is also considerable variation across the lakes in terms of these characteristics. The lakes in the Iowa Lakes Project are, on average, 667 acres in size, ranging from 10 acres to approximately 19,000 acres. The other site attributes are represented with dummy variables that indicate the availability of amenities of interest. The majority of the lakes in our sample have a paved boat ramp (85%) and wake restrictions (i.e., $Wake = 1$) (65%), while less than forty percent of the lakes have handicap facilities or are part of a local state park. There is also a wide range of water quality in Iowa lakes. For example, Secchi Transparency (which measures the depth into the lake that one can see) averages just over one meter, but varies from less than 0.1 meters (approximately 3.5 inches) to 5.67 meters (well over 18 feet). Similar ranges are found for the other water quality measures, including Total Nitrogen, Total Phosphorus, and Cyanobacteria. Moreover, these water quality measures are not highly correlated, as the source and nature of the water quality problems in individual lakes varies considerably across the state.

For the purpose of this application, the observed site characteristics ($\boldsymbol{Q}$) include the levels and natural *log* form of both site and water quality attributes. In contrast to Egan *et al.* (2009), who estimated a series of alternative specifications in a split sample analysis (ultimately choosing a single specification), we estimate a single model allowing the data to dictate the model with high posterior density that incorporates model uncertainty. In contrast to Abidoye, Herriges and Tobias (2010), in which the authors rely on a single specification (the "preferred" model identified in Egan *et al.*, 2009), we consider a wider range of possible functional forms for the set of site attributes impacting site selection.

[11] The "average wage rate" is calculated for all respondents as their household's income divided by 2,000. This allows for a 40 hour work week with two weeks of vacation.

5.1 Empirical Results

Using the model and posterior simulator detailed in the previous sections, we fit the site choice model using the Iowa Lakes data. Specifically, the Gibbs algorithm was first run for just over 20,000 iterations. The last iteration from this process was then used to initiate four different chains, run simultaneously on four different machines with different seeds. Discarding the first 20,000 iterations as burn-in, the four runs produced a total of 300,100 post-convergence draws to calculate posterior means, standard deviations and to make posterior inference.

5.2 Estimation Results

We are primarily interested in applying the algorithm described above to data from Iowa Lakes Project to illustrate its use in controlling for uncertainty in model specification. We addressed this question by considering a general model that includes all of the water quality and site characteristics in both their linear and logarithmic forms. We report parameter posterior means and posterior probabilities of being positive [denoted $P(. > 0|y)$] for key parameters of the model in Tables 3 through 5.

In general, many of the basic results are similar to those obtained in Egan *et al.* (2009) and Abidoye, Herriges and Tobias (2010). Starting with Table 3, we find that the marginal utility of income (i.e., negative of the coefficient on travel cost, $-\beta$) has a posterior mean of 0.0134 and a posterior distribution that is clearly massed away from zero. Turning to socio-demographic characteristics, older individuals, females, and the less educated are found to be more likely to stay at home, whereas households with more adults and more children are more likely to take trips. In Table 5, the alternative specific constants for each site (i.e., the α_j's) are all negative with over 99.9% of the posterior mass for each parameter lying below zero. This is consistent with the fact that households typically took relatively few trips during the course of a season.

The distinguishing feature of our model, relative to the earlier studies, lies in the hierarchical parameters reported in Table 4. For these parameters, we report not only the posterior means and $P(\cdot > 0|y)$, but also the frequency with which the Gibbs sampler yields $\lambda_k = 1$. The latter proportions indicate the extent to which the data suggests that an individual variable should be included as determining factor in recreation demand. As such, we use it to rank the various site and water quality characteristics in Table 4.

Several results emerge from examining the hierarchical parameter results. First, as both Egan *et al.* (2009) and Abidoye, Herriges and Tobias (2010) suggest, site characteristics are important determinants of where households choose for recreation. Lake size (in logarithmic form), the presence of wake restrictions, the inclusion of a lake in a state park and the availability of handicap facilities all have the expected positive signs, have a posterior distribution massed away from zero, and (excluding the intercept) account for four of the top six variables in terms of the proportion of draws with $\lambda_k = 1$. Notice too that the linear form for lake size (i.e., Acres) is clearly dominated by the logarithmic form, as was found in Egan *et al.* (2009).

Turning to the water quality attributes, we find that the coefficient on Total Phosphorous (in its logarithmic form) has a negative posterior mean of -0.16 and a posterior distribution that is clearly massed away from zero (with $P(\cdot > 0|y) = 0.059$), suggesting that high phosphorous levels negatively influence the appeal of a site. Moreover, ln(Total Phosphorous) is the highest ranking variable in terms of the proportion of draws (over thirty-seven percent) with $\lambda_k = 1$. The importance of Phosphorous in influencing recreation demand is not surprising as it is often a determining factor in algae growth, a clearly visible indicator of water quality. The linear form for Total Phosphorous, in contrast, appears in just over nine percent of the models and has a posterior distribution that is massed fairly evenly on either side of zero. Inorganic suspended solids (ISS) is the second most highly ranked water quality variable. In its logarithmic form, the associated coefficient has a positive posterior mean and is clearly massed away from zero (with $P(\cdot > 0|y) = 0.998$). At the same time, the coefficient associated with the linear form for ISS has a posterior distribution that is largely negative (with $P(\cdot > 0|y) = 0.043$). The marginal impact of a change in ISS will be a combination of these two coefficients, with logarithmic term diminishing in relative importance as ISS increases. Overall, when ISS is low (e.g., at the minimum of ISS in the sample), the marginal impact of ISS on site utility is positive (with a posterior probability greater than 0.90). However, when ISS is large (e.g., at the maximum of ISS in the sample), the marginal impact of ISS on site utility is negative (with a posterior probability greater than 0.90). This suggests that the impact of ISS is not captured effectively by either functional representation alone (i.e., linear or logarithmic), but is captured more effectively by the combination. The only other water quality variable that has a clear impact on site utility is Chlorophyll. In its logarithmic form, the associated coefficient is generally positive (with $P(\cdot > 0|y) = 0.918$), suggesting that an increase in Chlorophyll improves the appeal of a site. This result is consistent with earlier findings (e.g., Egan *et al.* (2009)). Interestingly, Secchi Transparency (which indicates the depth to which one can clearly see into a body of water) is not a significant factor, with a posterior distribution for both the linear and logarithmic terms massed evenly on either side of zero and with around twelve percent of

the draws having $\lambda_k = 1$ for either variable. This result is in sharp contrast to Egan *et al.* (2009), who suggest that Secchi is the best single water quality measure.

The results from the variable selection portion of our hierarchical model can be used in several ways. If the goal is to pick a single "best" model or to narrow the range of specifications for a more extensive Bayesian Model Averaging exercise, one can use the rankings to select a subset of the variables by choosing a cutoff for the frequency of $\lambda_k = 1$. For example, choosing variables that are visited (i.e., have $\lambda_k = 1$) at least 15% of the time leaves us with: $ln(TP)$, $ln(Acres)$, wake restrictions, $ln(ISS)$, state park classification, the availability of handicap facilities, $ln(TN)$ and $ln(Chlorophyll)$.[12] George and McCulloch (1997) employ a different strategy, considering further only those models whose relative probability was within 0.00674 (= -5 on a log posterior scale) of the best model. Ignoring the uncertainty regarding the site attributes, there are a total of 65,536 models, varying in terms of the inclusion or exclusion of each of the 16 water quality variables in Table 4. Using the criteria of George and McCulloch, this would reduce the number of models down to 228, with the top 40 models listed in Table 6. Notice that all of these specifications are relatively simple, typically including only a couple of the water quality variables. Also, the proportion of times any one model is visited is relatively small, though the top 40 models combine account for nearly half (47%) of the posterior draws.

Finally, and we would argue preferably, one can use the model as is, providing a basis for integrating the impact of all of the variables (in both their linear and logarithmic forms) into a policy evaluation, averaging over the range of possible model specifications (i.e., the various combinations of λ_k's). While any one model (or specific variable) may have a low posterior probability, the joint effect of the group of models or variables may still be significant. One indication of this in our application is that, while few of the water quality variables have a clear impact on site selection, it is clear that as a group they matter. Fewer than 10% of the models visited during the posterior simulation exclude all of the water quality variables. Note that this is in contrast to the conclusions reached in Abidoye, Herriges and Tobias(2010). Employing a single model specification (i.e., with Secchi entering the model linearly and all other water quality variables entering in logarithmic form), the authors use Bayes factors to conclude that water quality attributes are not an important determinant of recreation demand. However, this may reflect the selected model.[13] In discrete choice models with a full set of alternative specific constants, the impact of site attributes (including water quality) on site selection is reflected entirely in the alternative specific constants. This

[12] One might also include ISS given its strongly negative posterior distribution.

[13] Indeed, the specific model used by Abidoye, Herriges and Tobias is never visited in our posterior simulator. It was chosen, however, to be consistent with the earlier analysis of Egan *et al.* (2009).

effectively reduces that sample size used in measuring the role of site characteristics to the number of sites (i.e., J). In such settings, it seems prudent to allow flexibility in terms of model specification, rather than relying upon a single functional form.

6 Posterior Calculation and Welfare

Recreational demand models are used primarily to predict how exogenous changes in the attributes of the sites will affect the welfare of the household. These posterior calculations are intuitive and relatively easy to implement in the Bayesian framework. The approach we propose is in the spirit of implementing a Bayesian Model Averaging for posterior inference purposes. This approach of averaging over all the variables of the model rather than selecting a subset of the model for welfare analysis takes into consideration the uncertainty related to each of the variables.

As in Abidoye *et al.* (2010), let Υ_{it}^{s} denote the maximum utility achieved by agent i on choice occasion t under scenario s ($s = 0, 1$). That is,

$$\boldsymbol{\Upsilon_{it}^{s}(\Xi_{-\alpha.}, Q^{s}) = \max_{j}(U_{ijt}^{s}|\Xi_{-\alpha.}, Q^{s}) \quad s = 0, 1} \tag{21}$$

where $\boldsymbol{\alpha}_{\cdot} = (\alpha_1, \ldots, \alpha_J)$ denotes the vector of alternative specific constants. Changes in the site characteristics impact individual consumers by altering the overall appeal of the sites, as reflected in the α_j's. Thus, we no longer have a single set of alternative specific constants, but a set for each scenario (denoted $\boldsymbol{\alpha}_{\cdot}^{s}$). We use the hierarchical structure in equation (6) to simulate the changes to these constants resulting from a change in the site attributes. However, given the structure of the α_0 parameter, we average over the parameter instead of choosing a subset. Thus, the first step of drawing the alternative specific constant will proceed as follows:

Step 1: Draw $\boldsymbol{\alpha}_{(r)}^{s}, \quad s = 0, 1$ using (6).

That is, draw $\boldsymbol{\alpha}_{(r)}^{s}$ from a normal distribution with mean $\boldsymbol{Q^{s}[\alpha_{0(r)} * p(\lambda(r)|Y)]}$ and variance $\sigma_{\alpha(r)}^{2}$. where (r) indexes each iteration of the *posterior* simulator of the stated parameter. What this does is that at each iteration, only variables visited are used to simulate the alternative specific constant which will be used to average the CV estimate. This way the frequency that a variable is included in the model is used to weight the variable and follows the procedure proposed by Chipman, George and McCulloch (2001).

Once we draw the alternative specific constants, the other steps in the algorithm are straightforward. The utility levels are drawn using the simulated parameters and used to calculate a simulation based estimate of the compensating variation defined as:

$$\widehat{CV} = \frac{1}{R}\sum_{r=1}^{R} \frac{T}{-\beta}\left[\left(\max_j U_{ijt}^{1(r)}\right) - \left(\max_j U_{ijt}^{0(r)}\right)\right]. \tag{22}$$

The above algorithm is applied to the Iowa Lakes data. The scenario considered is one in which the water quality attributes of nine key zonal lakes (spread throughout the state) are upgraded to the quality of West Lake Okoboji (the cleanest lake in the state). [14] This same scenario was evaluated by both Egan *et al.* (2009) and similar to the scenario consider by Abidoye, Herriges and Tobias (2010). The result using the model estimates from section 5 is a posterior mean compensated variation of \$30.08, with $P(\widehat{CV} > 0)$ of approximately 92%. This result is comparable to Egan *et al.* (2009), who obtained CV estimates for the same water quality improvement scenario ranging from \$8 and \$40, depending on the model used, with their "best" model yielding a CV estimate of about \$28.92. The key difference here is that Egan *et al.* (2009) obtain very tight confidence bands around their welfare estimates, whereas our model suggests that there remains considerable uncertainty regarding the exact welfare gains from the water quality improvement. This stems largely from the fact that our model, like that of Abidoye, Herriges and Tobias (2010), incorporates a full set of alternative specific constants to control for unobservable site attributes. However, in contrast to Abidoye, Herriges and Tobias, in which the posterior mean compensated variation is actually negative ($\widehat{CV} = -1.61$, with $P(\widehat{CV} > 0) = 0.374$), our analysis provides for greater flexibility in terms of *how* water quality impacts recreation demand and yields a greater probability that the water quality changes represent a welfare improving policy. Similar results are obtained if the water quality improvements are considered on a lake-by-lake basis. The mean compensating variation ranges from \$0.53 (with $P(\widehat{CV} > 0) = 0.62$) for Briggs Wood lake to \$6.42 (with $P(\widehat{CV} > 0) = 0.93$) for Lake McBride.

7 Summary

In modeling the demand for recreation, analysts typically have relatively little *a priori* basis for specifying which site attributes should be included in their analysis and the functional

[14] Water quality attributes that are already better than those of West Lake Okobiji are left unchanged.

form representation to use. This paper presents a Bayesian variable selection model that can be used to either narrow the range of models to be considered further or as means of integrating a wide range of possible model specifications in what is akin to Bayesian model averaging. The Gibbs sampler, combined with data augmentation, makes characterizing the posterior distribution of the models parameters straightforward.

In our application, evaluating demand for recreational lake usage in Iowa, we find clear evidence that site attributes, such as lakes size, handicap facilities and wake restrictions, do impact lake usage. There is also evidence that water quality matters in household recreation choices. Total Phosphorus, inorganic suspended solids and chlorophyll levels all matter, but the influence of most other water quality measures, including Secchi Transparency, are only imprecisely measured. Yet, contrary to Abidoye, Herriges and Tobias (2010), in which only a single functional form is considered, we find clear evidence that water quality matters, with posterior probability of less that 10% associated with a model without any water quality variables. This suggests that the flexibility that the Bayesian variable selection model affords in capturing the linkage between recreation demand and site characteristics can be important. This is particularly true in RUM models of recreation demand with a full set of alternative specific constants, since the effective degrees of freedom available in measuring the impact of site characteristics has been reduced to the number of sites.

References

[1] Abidoye, B., J. A. Herriges and J. Tobias. (2010). Controlling for Observed and Unobserved Site Characteristics in RUM Models of Recreation Demand, Iowa State University Working Paper #10011, May.

[2] Buckland, S. T., K.P. Burnham, and N.H. Augustin (1997). Model Selection: An Integral Part of Inference. *Biometrics* **53**: 603-618.

[3] Chipman, H., E., George, and R.E. McCulloch (2001). The Practical Implementation of Bayesian Model Selection, Hayward, CA: IMS, pp. 67-116.

[4] Clyde, M. (2000). Model Uncertainty and Health Effect Studies for Particulate Matter, *Environmetrics* **11**: 745-763.

[5] Clyde, M., Guttorp, P., Sullivan, E. (2000). Effects of Ambient Fine and Coarse Particles on Mortality in Phoenix, Arizona ISDS Discussion Paper 00-05.

[6] Draper, D. (1995), Assessment and Propagation of Model Uncertainty (with discussion), *Journal of the Royal Statistical Society, Ser. B* **57**: 45-97.

[7] Egan, K.J., J.A. Herriges, C.L. Kling and J.A Downing (2009). Valuing Water Quality as a Function of Water Quality Measures, *American Journal of Agricultural Economics.*, **91**(1): 106-123.

[8] Fernandez, C., E. Ley, and M.F. Steel (2002) Bayesian Modeling of Catch in a North-West Atlantic Fishery, *Applied Statistics* **51**: 257-280.

[9] George, E.I., and R. E. McCulloch (1993), Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88**(423): 881-89.

[10] George, E.I., and R. E. McCulloch (1997), Approaches for Bayesian Variable Selection. *Statistica Sinica* **7**: 339-373.

[11] Herriges, J.A, C.L. Kling and D. Phaneuf (1999). Corner Solution Models of Recreation Demand: A Comparison of Competing Frameworks in: *Valuing Recreation and the Environment: Revealed Preference Methods in Theory and Practice* (ed. Herriges, J.A., and C.L. Kling), Cheltenham, UK: Edward Elgar, pp. 141-161. 163-197.

[12] Herriges, J., and D. Phaneuf (2002). Inducing Patterns Correlation and Substitution in Repeated Logit Model of Recreation Demand, *American Journal of Agricultural Economics*, **84**: 1076-1090.

[13] Koop, G and L. Tole (2004). Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air? *Journal of Environmental Economics and Management* **47**: pp. 3054.

[14] Layton, D. F. and S. T. Lee (2006). Embracing Model Uncertainty: Strategies for response Pooling and Model Averaging, *Environmental and Resource Economics* **34**: 51-85.

[15] Leamer, E. E. (1983), Lets Take the Con out of Econometrics. *The American Economic Review* **73**(1): 31-43.

[16] Leon, R., and C. J. Leon (2003) Single or Double Bounded Contingent Valuation? A Bayesian Test, *Scottish Journal of Political Economy* **50**: 174-188.

[17] Morey, E.R., R.D. Rowe, and M. Watson (1993). A Repeated Nested-Logit Model of Atlantic Salmon Fishing, *American Journal of Agricultural Economics*, **75**: 578-592.

[18] Murdock, Jennifer (2006). Handling unobserved site characterisitics in random utility models of recreation demand, *Journal of Environmental Economics and Management*, 51, 1-25.

[19] Poirier, Dale J. (1995), *Intermediate Statistics and Econometrics, A Comparative Approach*, Cambridge, MA: The MIT Press.

[20] Raftery, Adrian E. (1995). Bayesian Model Selection in Social Research [with discussion]. in: Marsden, P. V. (ed.), *Sociological Methodology*. Cambridge, MA: Blackwell, pp. 111-195 .

[21] Regal, R. and Hook, E. B. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* **10**: 717-721.

8 Tables and Figures

Table 1: Posterior Results for Generated Data Experiment

		Model with SSVS					Ignoring Model Uncertainty	
		Exper. #1: No correlation			Exper. #2: High corr.			
Parameter	True	Mean	$P(\cdot > 0\|y)$	$\sqrt{\text{Ineff. Factor}}$	Mean	$P(\cdot > 0\|y)$	Mean	$P(\cdot > 0\|y)$
$\alpha_{0,0}$	-1.91	-1.80	0.00	2.05	-1.58	0.00	-1.60	0.00
$\alpha_{0,1}$	-0.53	-0.57	0.10	2.26	-0.15	0.40	-0.90	0.02
$\alpha_{0,2}$	0	-0.07	0.22	1.26	-0.45	0.13	-0.16	0.10
β	-3.71	-3.72	0.00	20.40	-3.73	0.00	-3.73	0.00
γ_{01}	0.40	0.40	1.00	7.93	0.46	1.00	0.44	1.00
γ_{02}	0.50	0.56	1.00	6.52	0.56	1.00	0.57	1.00
σ^2_φ	0.50	0.49	1.00	20.81	0.50	1.00	0.48	1.00
σ^2_α	0.1	0.19	1.00	1.19	0.20	1.00	0.19	1.00
		Alternative specific constants						
α_1	-2.49	-2.48	0.00	15.32	-2.48	0.00	-2.46	0.00
α_2	-2.42	-2.40	0.00	12.97	-2.41	0.00	-2.39	0.00
α_3	-2.22	-2.20	0.00	13.63	-2.21	0.00	-2.19	0.00
α_4	-1.94	-1.91	0.00	13.50	-1.91	0.00	-1.89	0.00
α_5	-1.99	-1.99	0.00	15.26	-1.98	0.00	-1.97	0.00
α_6	-1.60	-1.59	0.00	11.49	-1.59	0.00	-1.57	0.00
α_7	-2.26	-2.24	0.00	14.33	-2.25	0.00	-2.23	0.00
α_8	-2.76	-2.73	0.00	17.94	-2.74	0.00	-2.72	0.00
α_9	-2.54	-2.50	0.00	14.81	-2.51	0.00	-2.49	0.00
α_{10}	-1.31	-1.29	0.00	11.53	-1.29	0.00	-1.27	0.00

Table 2: Summary Statistics

Variable	Model Variable	Mean	Std. Dev.	Min	Max
Total Day Trips (2002)[15]	T_i	6.33	9.97	0	50
Travel Cost ($100's)	P_{ij}	1.37	.83	0.0044	13.66
Age	$D_{i(1)}$	54.38	15.93	15	82
Male	$D_{i(2)}$	0.69	0.46	0	1
School	$D_{i(3)}$	0.67	0.47	0	1
Household Size	$D_{i(4)}$	2.61	1.30	0	12
Lake Attributes					
Acres	$Q_{j(1)}$	667.20	2112.83	10	19000
Ramps	$Q_{j(2)}$	0.85	0.36	0	1
Wake	$Q_{j(3)}$	0.65	0.48	0	1
Handicap	$Q_{j(4)}$	0.38	0.49	0	1
State Park	$Q_{j(5)}$	0.39	0.49	0	1
Water Quality					
Secchi Transparency (m)	$Q_{j(6)}$	1.17	0.92	0.09	5.67
Total Nitrogen (mg/l)	$Q_{j(7)}$	2.19	2.53	0.55	13.37
Total Phosphorus (μg/l)	$Q_{j(8)}$	105.45	80.33	17.10	452.55
Volatile SS (mg/l)	$Q_{j(9)}$	9.30	7.98	0.25	49.87
Inorganic SS (mg/l)	$Q_{j(10)}$	10.12	17.79	0.57	177.60
Cyanobacteria (mg/l)	$Q_{j(11)}$	298.08	831.51	0.02	7178.13
Chlorophyll (μg/l)	$Q_{j(12)}$	40.64	38.01	2.45	182.92

Table 3: Posterior Means of Travel Cost, Demog. Variables and Variance Parameters

Parameter	Mean	$P(\cdot > 0\|y)$
Travel cost	-0.0138	0.0000
Demographic Variables		
Age	0.0164	1.0000
Male	-0.2425	0.0000
School	-0.1859	0.0010
Household Size	-0.0430	0.0049
Variance parameters		
σ^2_φ	2.04	1.0000
σ^2_α	0.06	1.0000

Table 4: Posterior Means of hierarchical Parameters (Site Characteristics)

Site Characteristics	Posterior Mean	$P(\cdot > 0\|y)$	Proportion $[P(\lambda\|Y)]$
α_0	-3.990	0	1
ln(Total Phosphorus)	-0.1623	0.0593	0.3715
ln(Acres)	0.1736	1.0000	0.3181
Wake	0.1523	0.9986	0.2871
ln(ISS)	0.1361	0.9979	0.2389
State Park	0.1148	0.9888	0.1962
Handicap	0.1091	0.9872	0.1828
ln(Total Nitrogen)	0.0252	0.5849	0.1763
ln(Chlorophyll)	0.0827	0.9179	0.1564
ln(VSS)	-0.0263	0.3669	0.1322
Ramp	0.0513	0.7906	0.1303
ln(Secchi)	0.0020	0.5109	0.1267
Total Nitrogen	-0.0237	0.3634	0.1203
N03	0.0307	0.6931	0.1165
Secchi	0.0342	0.7271	0.1146
ln(N03)	-0.0164	0.3551	0.1015
ln(Cyanobacteria)	-0.0125	0.2433	0.0925
VSS	-0.0071	0.1556	0.0916
Chlorophyll	0.0008	0.7020	0.0910
Cyanobacteria	-1.26E-06	0.3669	0.0910
Total Phosphorus	0.0005	0.6384	0.0908
ISS	-0.0038	0.0483	0.0904
Acres	-2.17E-06	0.4435	0.0901

Table 5: Posterior Means of Alternative Specific Constants

Lake	Mean	$P(\cdot > 0\|y)$	Lake	Mean	$P(\cdot > 0\|y)$	Lake	Mean	$P(\cdot > 0\|y)$
Arbor	-3.4663	0.0000	Hooper	-3.8797	0.0000	North Twin	-2.8972	0.0000
Arrowhead	-3.1154	0.0000	Indian	-3.3165	0.0000	Oldham	-3.9083	0.0000
Arrowhead	-3.6949	0.0000	Ingham	-2.9736	0.0000	Otter Creek	-3.5688	0.0000
Ave. of the Saints	-3.7665	0.0000	Kent Park	-3.2356	0.0000	Ottumwa Lagoon (proper)	-2.9801	0.0000
Badger Creek	-3.2967	0.0000	Lacey-Keosauqua	-2.9466	0.0000	Pierce Creek	-3.5287	0.0000
Badger	-2.9396	0.0000	Ahquabi	-3.0697	0.0000	Pleasant Creek	-2.9714	0.0000
Beaver	-3.7162	0.0000	Anita	-3.0201	0.0000	Pollmiller	-3.2584	0.0000
Beed's	-2.9717	0.0000	Cornelia	-3.0652	0.0000	Prairie Rose	-3.0607	0.0000
Big Creek	-2.5480	0.0000	Darling	-3.0141	0.0000	Rathbun	-2.3455	0.0000
Spirit Lake	-2.2197	0.0000	Geode	-2.8668	0.0000	Red Haw	-3.1610	0.0000
Black Hawk	-2.7683	0.0000	Hendricks	-3.3206	0.0000	Red Rock	-2.3615	0.0000
Blue	-3.0173	0.0000	Icaria	-2.7834	0.0000	Robert's Creek	-3.4885	0.0000
Bob White	-3.6031	0.0000	Iowa	-3.3746	0.0000	Rock Creek	-3.1649	0.0000
Brigg's Woods	-3.2081	0.0000	Keomah	-3.3388	0.0000	Rogers	-3.7456	0.0000
Brown's	-2.9291	0.0000	Manawa	-2.4587	0.0000	Saylorville	-2.3646	0.0000
Brushy Creek	-2.7864	0.0000	Macbride	-2.6905	0.0000	Silver	-2.8854	0.0000
Carter	-3.1461	0.0000	Miami	-3.3164	0.0000	Silver	-3.7448	0.0000
Casey	-3.4547	0.0000	Minnewashata	-2.9932	0.0000	Silver	-3.5842	0.0000
Center	-3.1749	0.0000	Lake of The Hills	-3.2113	0.0000	Silver	-3.0613	0.0000
Central	-3.5290	0.0000	Three Fires	-2.9127	0.0000	Slip Bluff	-3.9377	0.0000
Clear	-2.1675	0.0000	Orient	-3.6675	0.0000	South Prairie	-3.7247	0.0000
Cold Springs	-3.2940	0.0000	Pahoja	-2.9762	0.0000	Spring	-3.4325	0.0000
Coralville	-2.5010	0.0000	Smith	-3.2590	0.0000	Springbrook	-3.1328	0.0000
Crawford Creek	-3.5048	0.0000	Sugema	-2.9286	0.0000	Storm Lake	-2.4095	0.0000
Crystal	-3.0117	0.0000	Wapello	-2.9019	0.0000	Swan	-2.8845	0.0000
Dale Maffit	-3.5259	0.0000	Little River	-3.2240	0.0000	Thayer	-3.8172	0.0000
DeSoto Bend	-2.8251	0.0000	Little Sioux Park	-3.1550	0.0000	Three Mile	-2.8776	0.0000
Diamond	-3.3826	0.0000	Little Spirit	-2.6425	0.0000	Trumbull	-3.3258	0.0000
Dog Creek	-3.3488	0.0000	Little Wall	-3.4092	0.0000	Tuttle	-3.4103	0.0000
Don Williams	-3.0025	0.0000	Littlefield	-3.3476	0.0000	Twelve Mile	-2.9510	0.0000
East Osceola	-3.2562	0.0000	Lost Island	-2.7296	0.0000	Union Grove	-3.4362	0.0000
East Okoboji	-2.0441	0.0000	Lower Gar	-2.8834	0.0000	Upper Gar	-2.9968	0.0000
Easter	-3.1976	0.0000	Lower Pine	-3.2059	0.0000	Upper Pine	-3.0827	0.0000
Eldred Sherwood	-3.6085	0.0000	Manteno Pond	-3.8601	0.0000	Viking	-2.9418	0.0000
Five Island	-2.9795	0.0000	Mariposa	-3.6137	0.0000	Volga	-2.9555	0.0000
Fogle	-3.5821	0.0000	Meadow	-4.0143	0.0000	West Okoboji	-1.8592	0.0000
George Wyth	-2.9115	0.0000	Meyers	-3.6530	0.0000	West Osceola	-3.1686	0.0000
Green Belt	-3.9644	0.0000	Mill Creek	-3.3010	0.0000	White Oak	-4.0582	0.0000
Green Castle	-3.7493	0.0000	Mitchell Impoundment	-3.9798	0.0000	Williamson Pond	-4.1626	0.0000
Green Valley	-3.0887	0.0000	Moorehead	-3.3896	0.0000	Willow	-3.4959	0.0000
Greenfield Lake	-3.6353	0.0000	Mormon Trail	-3.6413	0.0000	Wilson	-3.6920	0.0000
Hannen	-3.4521	0.0000	Nelson Park	-3.7240	0.0000	Windmill	-3.4613	0.0000
Hawthorn	-3.3248	0.0000	Nine Eagles	-3.3471	0.0000	Yellow Smoke	-3.0525	0.0000
Hickory Grove	-3.3508	0.0000						

Table 6: Posterior Model Frequencies

	Included Water Quality Variables							
Percent	TP	TN	NO3	Cyan.	Chlor.	Secchi	VSS	ISS
9.20%								
5.02%	Log							
2.67%								Log
1.90%		Log						
1.70%	Log							Log
1.50%					Log			
1.42%							Log	
1.32%						Log		
1.20%						Linear		
1.17%		Linear						
1.15%			Linear					
1.08%	Log	Log						
1.07%	Log				Log			
1.01%			Log					
0.93%					Linear			
0.92%								Linear
0.92%	Linear							
0.91%				Log				
0.90%				Linear				
0.90%							Linear	
0.73%	Log						Log	
0.73%	Log					Log		
0.66%	Log	Linear						
0.65%	Log		Linear					
0.59%	Log					Linear		
0.56%	Log		Log					
0.54%		Log						Log
0.51%	Log			Log				
0.50%	Log							Linear
0.49%	Log				Linear			
0.49%	Log,Linear							
0.49%	Log						Linear	
0.49%	Log			Linear				
0.43%					Log			Log
0.43%							Log	Log
0.38%						Log		Log
0.37%	Log				Log			Log
0.36%						Linear		Log
0.36%			Linear					Log
0.36%	Log	Log						Log

Figure 1: Posterior Distribution for the true regressor ($\alpha_{0,1}$) - Experiment #1: No correlation

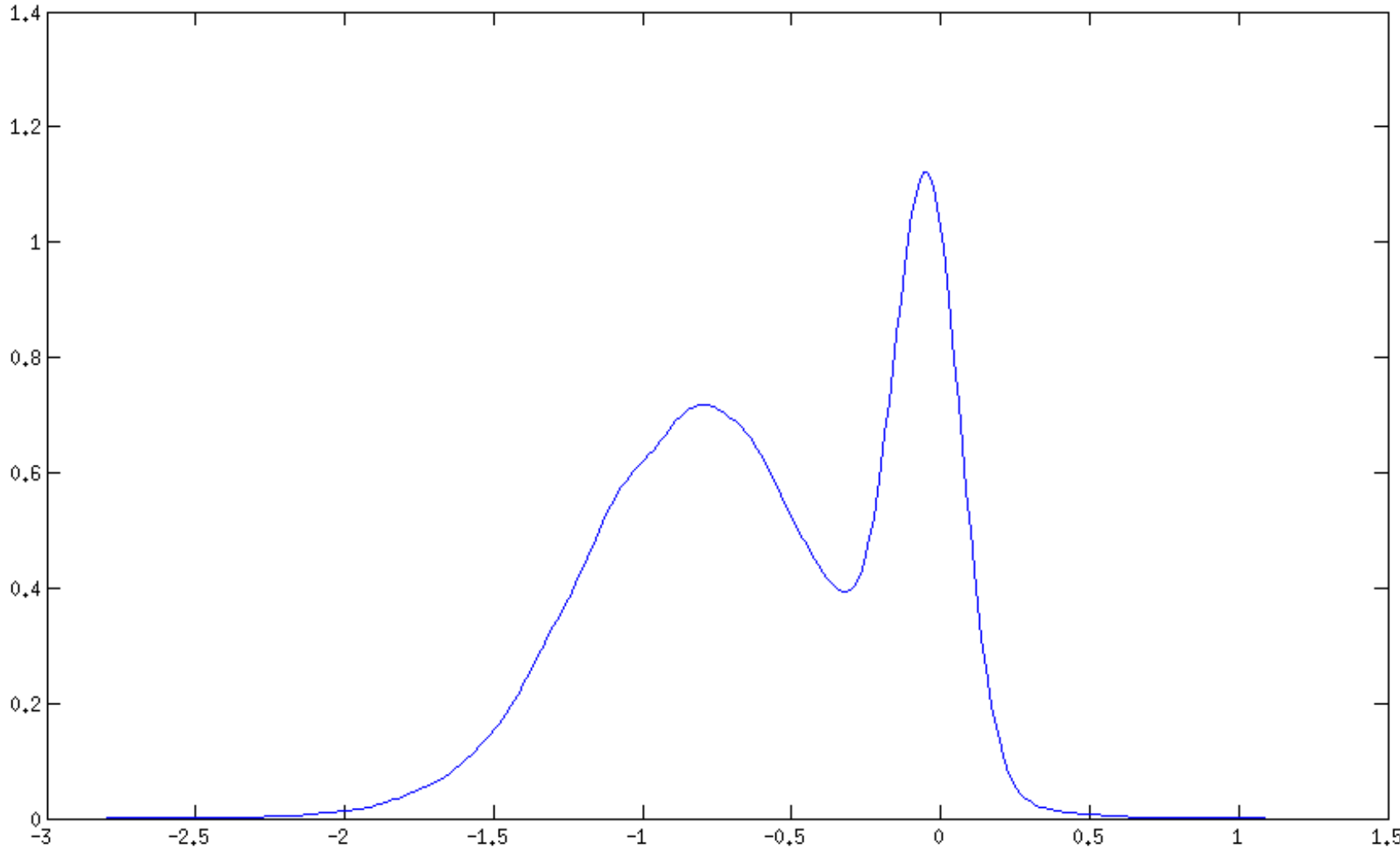


Figure 2: Posterior Distribution for the added regressor ($\alpha_{0,2}$) - Experiment #1: No correlation

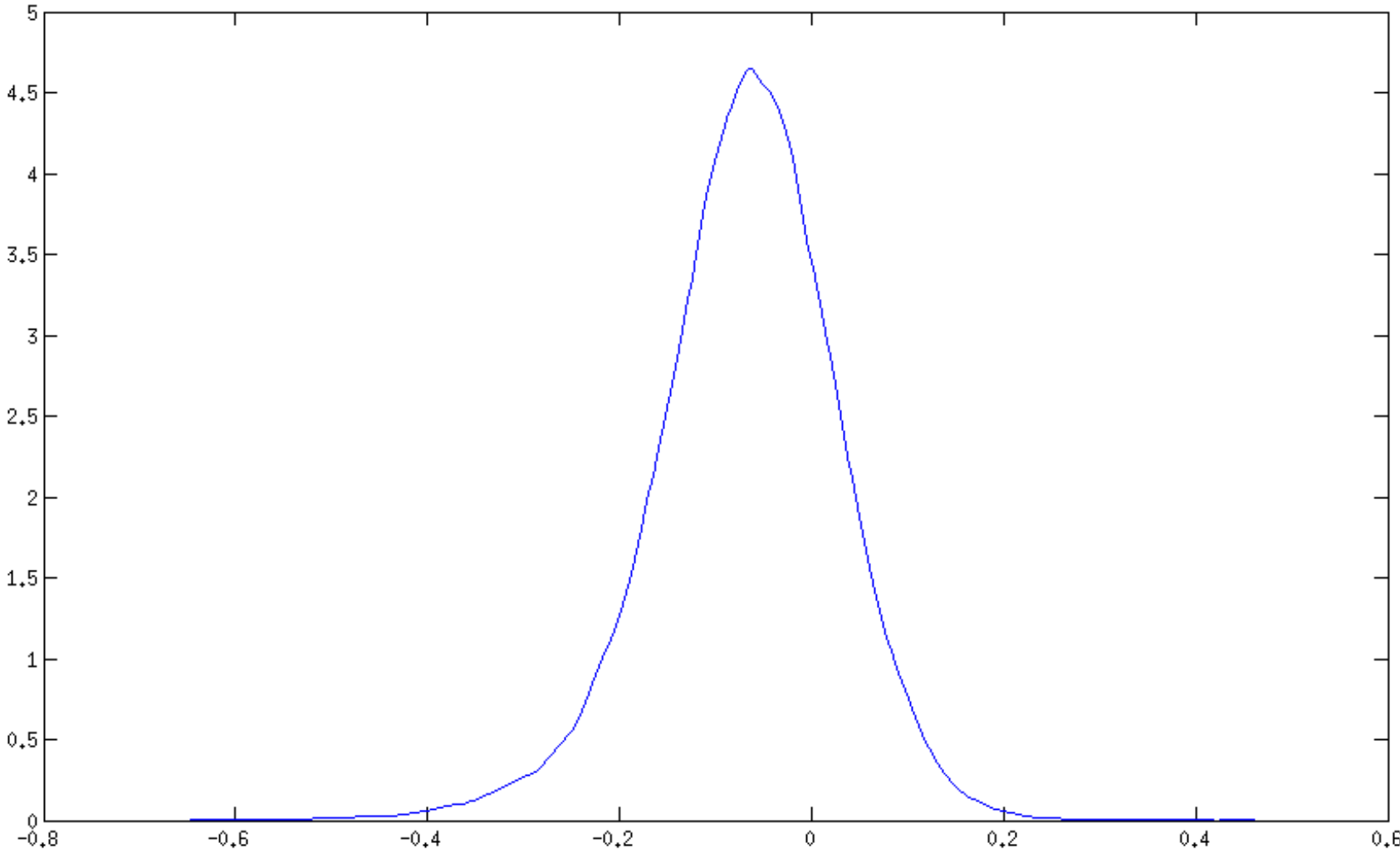


Figure 3: Posterior Distribution for the true regressor ($\alpha_{0,1}$) - Experiment #2: High correlation

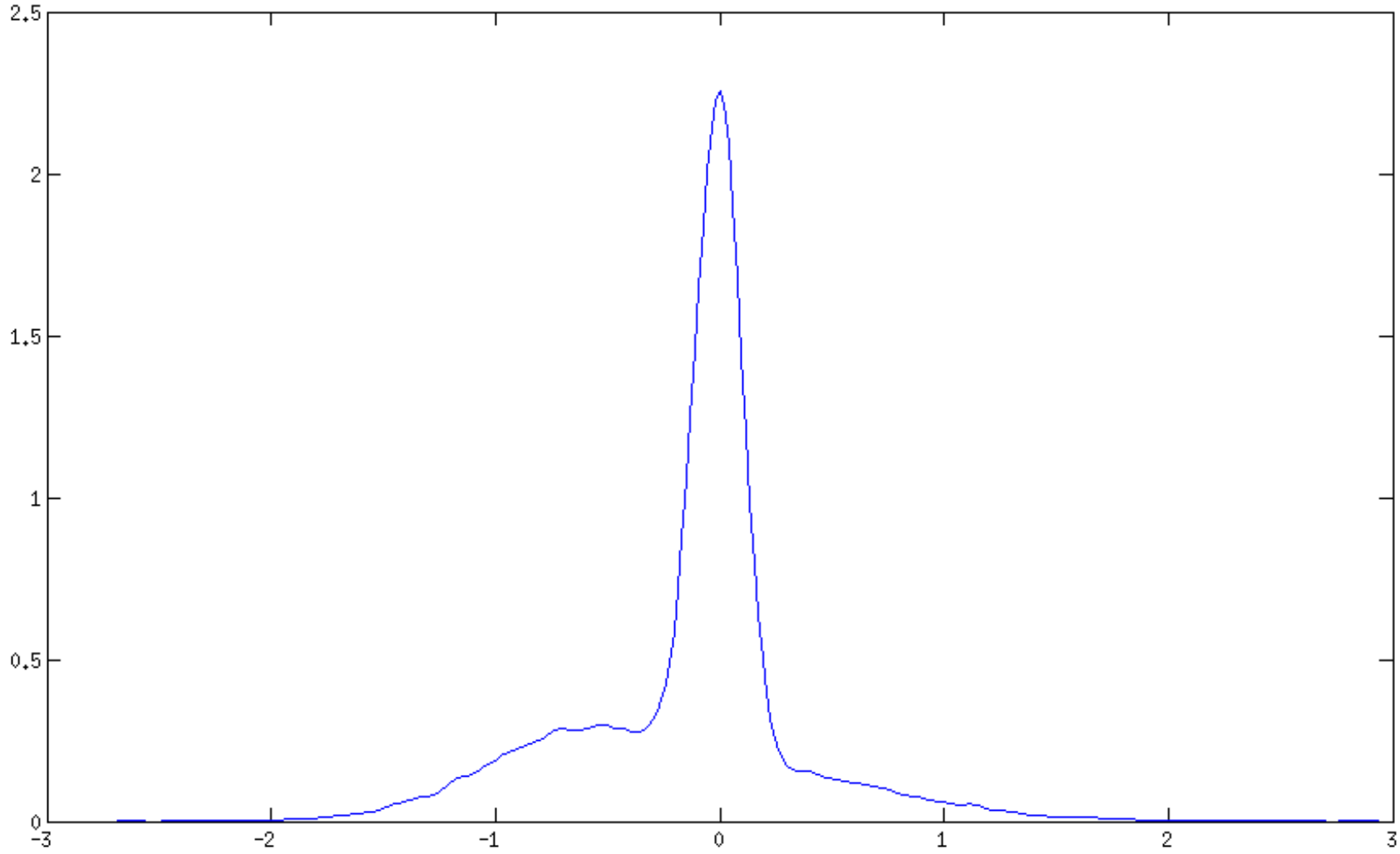


Figure 4: Posterior Distribution for the added regressor ($\alpha_{0,2}$) - Experiment #2: High correlation

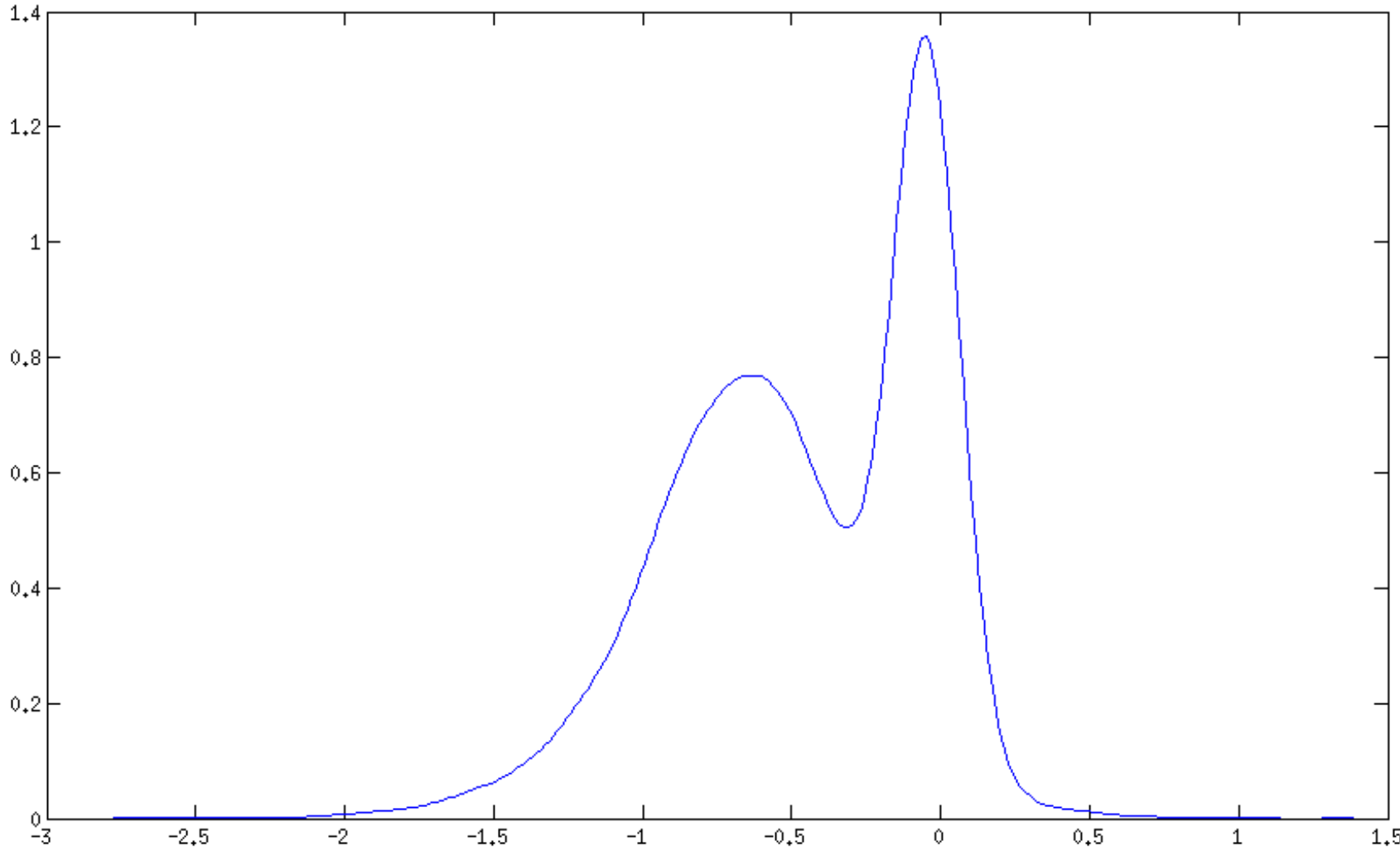


Figure 5: Posterior Distribution for the added regressor ($\alpha_{0,2}$) - Ignoring Model Uncertainty