Studies in Nonlinear Dynamics & Econometrics

Volume 10, *Issue* 4 2006 *Article* 7

Bayesian Analysis of Structural Effects in an Ordered Equation System

Mingliang Li* Justin L. Tobias†

*State University of New York at Buffalo, mli3@buffalo.edu
†Iowa State University, tobiasj@iastate.edu

Bayesian Analysis of Structural Effects in an Ordered Equation System*

Mingliang Li and Justin L. Tobias

Abstract

We describe a new simulation-based algorithm for Bayesian estimation of structural effects in models where the outcome of interest and an endogenous treatment variable are ordered. Our algorithm makes use of a reparameterization, suggested by Nandram and Chen (1996) in the context of a single equation ordered-probit model, which significantly improves the mixing of the standard Gibbs sampler. We illustrate the improvements afforded by this new algorithm (relative to the standard Gibbs sampler) in a generated data experiment and also make use of our methods in an empirical application. Specifically, we take data from the National Longitudinal Survey of Youth (NLSY) and investigate the impact of maternal alcohol consumption on early infant health. Our results show clear evidence that the health outcomes of infants whose mothers drink while pregnant are worse than the outcomes of infants whose mothers never consumed alcohol while pregnant. In addition, the estimated parameters clearly suggest the need to control for the endogeneity of maternal alcohol consumption.

*We would like to thank a referee, the associate editor, and the editor for helpful comments and suggestions. All errors are our own. Mingliang Li: Department of Economics, State University of New York at Buffalo, 433 Fronczak Hall, Buffalo, NY 14260, email: mli3@buffalo.edu; Justin L. Tobias: Department of Economics, Iowa State University, 260 Heady Hall, Ames, IA 50011, email: tobiasj@iastate.edu.

1 Introduction

The isolation of "structural," "causal" or "treatment" effects is a topic of central importance to the social, biological and statistical sciences. In recent years, we have witnessed an explosion of applied work that seeks to identify and estimate causal impacts of various endogenous variables on outcomes of interest. For example, in the econometrics literature, recent studies have attempted to identify the causal effect of education on wages [e.g., Ashenfelter and Krueger (1994), Card (1999)], the effect of family size on female labor supply [e.g., Angrist and Evans (1998)], the effect of maternal inputs on subsequent outcomes of children [e.g., Li and Poirier (2001,2003a,2003b,2003c)], and the effect of health insurance on medical expenditure and number of physician visits [e.g., Munkin and Trivedi (2003)].

The applied studies in this literature seek to surmount a common problem and share a similar statistical structure. In each case, based on our understanding of the institution under study, we are concerned that factors not directly observable by the researcher are potentially correlated with both the endogenous variable and the outcome of interest. For example, when trying to estimate the causal effect of educational attainment on wages, it is important to recognize that the unobserved "ability" of the agent may influence both the quantity of education received as well as labor market earnings.

To overcome this potential problem of confounding on unobservables, many researchers - particularly those in economics - have made use of instrumental variable (IV) estimators. This approach requires that the researcher can find, and provide a convincing justification for, some characteristic that is correlated with the endogenous treatment variable, but is conditionally uncorrelated with the outcome of interest. While empirical researchers have become quite adept in finding these excluded covariates, and in many cases offer compelling stories as to the validity of the instruments employed, in another sense, IV studies are somewhat limiting. First, IV estimation methods typically focus only on the causal effect of interest and do not seek to recover other parameters of the full statistical model. For example, IV studies typically do not quantify the degree of confounding on unobservables, which is seemingly a parameter of significant interest. Second, it is often difficult to move beyond estimation to conduct policy experiments or out of sample predictive exercises unless the full statistical model is supplied and estimates of parameters associated with that model are obtained. Finally, there is growing awareness of problems associated with weak instruments [e.g., Staiger and Stock (1997)] and an increasing emphasis on the proper interpretation of the IV estimator when returns are heterogeneous or the model is nonlinear [e.g., Imbens and Angrist (1994)].

In this paper, we aim to make an additional contribution to this literature and investigate a particular treatment effects or causal[1] effect model where both the outcome and

[1] Our repeated use of the word "causal" may be something of an abuse of language. In the context of this study, a causal effect model is an empirical specification that seeks to consistently estimate parameters of the structural equation system (most notably the slope coefficient on the endogenous variable) in the presence of unobserved confounding. This use of language seems consistent with its current use in the applied literature, where causal effects are reported to be obtained when a convincing instrument or natural experiment has been used to surmount the endogeneity problem.

endogenous variable of interest are generated by nonlinear specifications. Specifically, we investigate the particular case of a two equation triangular simultaneous equations model where both the outcome and a potentially endogenous variable appearing in that outcome equation are ordered in nature.

Our estimation algorithm, which we feel is useful for other studies sharing a similar structure, mitigates autocorrelation in our posterior simulations by making use of a reparameterization building off the suggestion of Nandram and Chen (1996). Among other benefits, this reparameterization effectively eliminates one unknown cutpoint from each ordered equation so that, for example, if both the outcome and the endogenous variable take on only three values, the model will contain no unknown cutpoints. In addition, the reparameterization eliminates restrictions initially imposed on the structural covariance matrix (which can complicate a posterior simulator), and thus posterior simulation for the (reparameterized) covariance matrix can proceed using standard conjugate analysis.

We apply our techniques in practice to investigate the impact of maternal alcohol consumption during pregnancy on infant health during the first year of life. Our study makes use of rich data provided in the National Longitudinal Survey of Youth (NLSY). Both alcohol consumption and early infant health are recorded as ordered variables in the NLSY data we employ, and thus our application fits directly into the framework of our maintained model. Unlike the majority of biomedical research on this topic, we recognize that maternal alcohol consumption may be endogenous, and thus allow for potential confounding on unobservables, even conditioned on a variety of controls.[2] For our instrument, which is assumed to have a structural effect on maternal alcohol consumption but no effect on infant health given our included controls, we exploit data on whether or not the mother has a biological parent who has a drinking problem or is an alcoholic.[3] Our argument is that individuals with at least one parent with a drinking problem would be more likely to drink themselves, while *grand*parental alcohol consumption patterns will have no structural effect on infant health conditioned on the health status and alcohol consumption patterns of the mother. We find that alcohol consumption has a large undesirable impact on number of doctor visits during the first year of life, and also find significant evidence regarding the endogeneity of maternal alcohol consumption (i.e., unobserved confounding).

The outline of the paper is as follows. The following section describes the empirical specification and our suggested reparameterization. A generated data experiment illustrating the performance of our posterior simulator is presented in section 3 and a description of the data used in our empirical investigation is provided in Section 4. Section 5 contains the empirical results of our application, and the paper concludes with a summary in section 6. Technical details regarding our posterior simulator are completely provided in the appendix.

[2]For research in the biomedical literature, see, for example, Jacobson et al (1993, 1994) and Goldschmidt et al (1996). Notable exceptions include the careful and general structural equations analyses of Li and Poirier (2001, 2003a,2003b,2003c) who examine the impact of a variety of endogenous maternal inputs (including alcohol use) on early and subsequent child outcomes. Unlike our work, however, Li and Poirier focus on a different set of birth outcomes (such as birth weight, birth length, gestational age and childhood test scores) and model variables that are either continuous or binary.

[3]This determination is subjective, as the mothers in the sample are asked whether or not either of her parents had or has a "drinking problem."

2 The Model

In general terms, the model we consider is a two-equation system containing two endogenous variables, denoted y and r. Both of these variables are discrete and ordered with $y_i \in \{1, 2, \cdots, Y\}$ and $r_i \in \{1, 2, \cdots, R\}$ $\forall i$. To remain consistent with a typical treatment-response formulation of the model, we take up the case of a two equation triangular system where y has a structural dependence on r, and r is generated from a reduced form specification. To formally account for the discrete, ordered nature of each response we begin with a latent variable representation of the model:

$$z_{yi} = x_{yi}\beta_y + d_{ri}\theta + \epsilon_{yi} \tag{1}$$
$$z_{ri} = x_{ri}\beta_r + \epsilon_{ri}, \tag{2}$$

where d_{ri} is the $1 \times R$ dummy variable vector for r_i which contains a one in the r_i^{th} column and zeros elsewhere. We interpret the parameter vector θ as quantifying the *treatment effect* of levels of r on y. In our triangular system, like other "causal" effect models with continuous outcome and endogenous variables, we model the endogeneity of r_i by permitting correlation between ϵ_y and ϵ_r. Throughout this paper, we therefore assume[4]

$$\begin{bmatrix} \epsilon_{yi} \\ \epsilon_{ri} \end{bmatrix} \stackrel{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{yr} \\ \sigma_{yr} & 1 \end{pmatrix}\right], \tag{3}$$

where the error variances in each of the latent variable equations have been normalized to unity for identification purposes.

We relate the observed ordered y and r variables to the latent z_y in (1) and z_r in (2) through the restrictions:

$$y_i = j \text{ iff } \gamma_j < z_{yi} \leq \gamma_{j+1}, \quad j = 1, 2, \cdots, Y \tag{4}$$

and

$$r_i = k \text{ iff } \tilde{\gamma}_k < z_{ri} \leq \tilde{\gamma}_{k+1}, \quad k = 1, 2, \cdots, R. \tag{5}$$

For identification purposes, we also impose standard restrictions on certain values of the *cutpoints* γ_j and $\tilde{\gamma}_k$, namely: $\gamma_1 = \tilde{\gamma}_1 = -\infty, \gamma_2 = \tilde{\gamma}_2 = 0$ and $\gamma_{Y+1} = \tilde{\gamma}_{R+1} = \infty$.

2.1 Estimation

While one could simply use a standard Gibbs sampler to fit the above model (upon specifying a prior for all of the model parameters), there is a potential concern associated with adopting such an approach. As described in Cowles (1996) in the context of a simplified one-equation ordered probit model, use of the standard Gibbs sampler in applications of moderate size can suffer from slow mixing due to high correlation between the simulated cutpoints and latent data. This slow mixing problem is likely to be even more severe in our two-equation ordered

[4]To relax the Normality assumption without significantly complicating the posterior simulator, one could add mixing variables to the disturbance variance or pursue a mixture of Normals approach. Since these techniques are well-known, we do not discuss them here.

outcome model with an endogeneity problem. We will provide rather striking evidence of this slow mixing problem in the generated data experiment of the following section.

To mitigate the high degree of autocorrelation in our posterior simulations, we choose to work with a reparameterization of the model, building off the suggestion of Nandram and Chen (1996), which both improves the performance of the posterior simulator and offers some computational simplifications. In this regard it will prove useful to first separate the largest two unknown cutpoints γ_Y and γ_R from the remaining vector of cutpoints. We accomplish this by defining the following $(Y-3)\times 1$ and $(R-3)\times 1$ vectors (respectively),

$$\gamma = [\gamma_3\ \gamma_4 \cdots \gamma_{(Y-1)}]' \quad \text{and} \quad \tilde{\gamma} = [\tilde{\gamma}_3\ \tilde{\gamma}_4 \cdots\ \tilde{\gamma}_{(R-1)}]',$$

and let $\delta = [\beta_y'\ \beta_r'\ \gamma'\ \tilde{\gamma}'\ \gamma_Y^2\ \tilde{\gamma}_R^2\ \sigma_{yr}]'$ denote the vector of parameters in our model.[5]

Before discussing our strategy for reparameterization, it is instructive to first derive the joint posterior for the "structural" parameters of this model. With an eye toward our eventual reparameterization, we employ priors for these parameters of the following forms:

$$\begin{aligned}
\beta_y|\gamma_Y &\sim N(0, \gamma_Y^2 V_y) && (6)\\
\beta_r|\tilde{\gamma}_R &\sim N(0, \tilde{\gamma}_R^2 V_r) && (7)\\
p(\gamma_3, \gamma_4, \cdots, \gamma_{Y-1}|\gamma_Y) &\propto \gamma_Y^{-Y+3} I\left[0 < \gamma_3 < \gamma_4 < \cdots < \gamma_{Y-1} < \gamma_Y\right] && (8)\\
p(\tilde{\gamma}_3, \tilde{\gamma}_4, \cdots, \tilde{\gamma}_{R-1}|\tilde{\gamma}_R) &\propto \tilde{\gamma}_R^{-R+3} I\left[0 < \tilde{\gamma}_3 < \tilde{\gamma}_4 < \cdots < \tilde{\gamma}_{R-1} < \tilde{\gamma}_R\right] && (9)\\
\gamma_Y^2|\sigma_{yr} &\sim G[a, 2h_1(1-\sigma_{yr}^2)] && (10)\\
\tilde{\gamma}_R^2|\sigma_{yr} &\sim G[a, 2h_2(1-\sigma_{yr}^2)] && (11)\\
p(\sigma_{yr}) &\propto (1-\sigma_{yr}^2)^{a-(3/2)}, \quad |\sigma_{yr}| < 1, && (12)
\end{aligned}$$

with G denoting a Gamma distribution[6] and $I(\cdot)$ denoting the standard indicator function.

Perhaps with the exception of (12), these prior specifications take somewhat nonstandard forms. In particular, the priors in (6) and (7) include the largest cutpoints γ_Y and γ_R in the variance function for β_y and β_r, respectively, and in (10) and (11), specify correlation between the largest cutpoints and the covariance term σ_{yr}. The priors explicitly impose a sequential ordering truncation on the cutpoints γ_j and $\tilde{\gamma}_j$,[7] and impose that the covariance

[5]Because both the covariates x_{yi} and the dummy variable vector d_{ri} enter equation (1) linearly, we simplify our notation henceforth by including the dummy variable vector d_{ri} in the covariates x_{yi}, and by including the parameter vector θ in the coefficients β_y.

[6]See, e.g., Poirier (1995, page 98). With this parameterization $x \sim G(a,b)$ implies $p(x) \propto x^{a-1}\exp(-x/b)$.

[7]To gain some further insights into the priors we specify on the cutpoints, say, for example, γ_3, γ_4, $\cdots$, γ_{Y-1}, define u_3, u_4, $\cdots$, u_{Y-1} as $Y-3$ random variables that are, conditional on γ_Y, independently and identically drawn from the uniform distribution $U(0,\gamma_Y)$, where $U(x_1,x_2)$ denotes the uniform distribution over the interval $(x_1,\ x_2)$. Order the random variables u_3, u_4, $\cdots$, u_{Y-1} and define γ_3 as the smallest among them, γ_4 the second smallest among them, γ_{Y-1} the largest among them, and so on. It turns out that $p(\gamma_3,\gamma_4,\cdots,\gamma_{Y-1}|\gamma_Y) \propto \gamma_Y^{-Y+3}$ as specified in the above priors. To see this, note that $p(\gamma_3 = x_3, \gamma_4 = x_4, \cdots, \gamma_{Y-1} = x_{Y-1}|\gamma_Y) = p(\text{one of the } u_j\text{'s} = x_3, \text{one of the } u_j\text{'s} = x_4, \cdots, \text{one of the } u_j\text{'s} = x_{Y-1}|\gamma_Y) = (Y-3)!p(u_3 = x_3, u_4 = x_4, \cdots, u_{Y-1} = x_{Y-1}|\gamma_Y) = (Y-3)!p(u_3 = x_3|\gamma_Y)p(u_4 = x_4|\gamma_Y)\cdots p(u_{Y-1} = x_{Y-1}|\gamma_Y) = (Y-3)!\gamma_Y^{-Y+3}$. The second equality in the above derivation holds because there are in total $(Y-3)!$ different possible outcomes of the ordering of the random variables u_3, u_4, $\cdots$, u_{Y-1}.

matrix in (3) is positive definite since $|\sigma_{yr}| < 1$. As we will show below, use of these particular priors for the "structural" parameters proves to be computationally advantageous, as they will imply the use of "standard" conjugate priors for parameters in our reparameterized model. We will revisit this point and take up a more detailed discussion of the priors in (6) - (12) following our discussion of this reparameterization.

The priors described above, combined with the *complete data likelihood* implied by (1)-(3), give the *augmented posterior* distribution of the model parameters and latent data. This augmented posterior can be shown to be of the form:

$$\begin{aligned} p(\delta, z_y, z_r|y, r) &\propto \{\prod_{i=1}^{n} \phi_2 \left[\begin{pmatrix} z_{yi} \\ z_{ri} \end{pmatrix}; \begin{pmatrix} x_{yi}\beta_y \\ x_{ri}\beta_r \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{yr} \\ \sigma_{yr} & 1 \end{pmatrix} \right] \\ &\times \quad I(\gamma_{y_i} < z_{yi} \leq \gamma_{y_i+1}) I(\tilde{\gamma}_{r_i} < z_{ri} \leq \tilde{\gamma}_{r_i+1})\} p(\delta), \end{aligned}$$

with $\phi_k(x; \mu, V)$ denoting a k-dimensional normal density for x with mean μ and variance V. The prior density $p(\delta)$ is given in equations (6) - (12).

2.2 A reparameterization

Let us now take the augmented joint posterior just discussed, and consider making a change of variables. Specifically, let

$$\sigma_y = 1/[\gamma_Y^2], \quad \beta_y^* = \sqrt{\sigma_y}\beta_y, \quad z_y^* = \sqrt{\sigma_y} z_y, \quad \gamma^* = \sqrt{\sigma_y}\gamma$$

and

$$\sigma_r = 1/[\tilde{\gamma}_R^2], \quad \beta_r^* = \sqrt{\sigma_r}\beta_r, \quad z_r^* = \sqrt{\sigma_r} z_r, \quad \tilde{\gamma}^* = \sqrt{\sigma_r}\tilde{\gamma}, \ \tilde{\sigma}_{yr} = \sqrt{\sigma_y}\sqrt{\sigma_r}\sigma_{yr}.$$

With some work, one can derive that the Jacobian of the transformation from $[\delta \ z_y \ z_r]$ to $[\delta^* \ z_y^* \ z_r^*]$ (with $\delta^* \equiv [\beta_y^{*\prime} \ \beta_r^{*\prime} \ \gamma^{*\prime} \ \tilde{\gamma}^{*\prime} \ \sigma_y \ \sigma_r \ \tilde{\sigma}_{yr}]'$) is $\sigma_y^{-[k_y+(Y-3)+n+5]/2} \sigma_r^{-[k_r+(R-3)+n+5]/2}$, with k_y and k_r denoting the number of elements in β_y and β_r, respectively.

Adding this Jacobian term to our previous expression of the joint posterior, and completing our change of variables, we obtain

$$\begin{aligned} p(\delta^*, z_y^*, z_r^*|y, r) &\propto \{\prod_{i=1}^{n} \phi_2 \left[\begin{pmatrix} z_{yi}^* \\ z_{ri}^* \end{pmatrix}; \begin{pmatrix} x_{yi}\beta_y^* \\ x_{ri}\beta_r^* \end{pmatrix}, \begin{pmatrix} \sigma_y & \tilde{\sigma}_{yr} \\ \tilde{\sigma}_{yr} & \sigma_r \end{pmatrix} \right] \\ &\times \quad I(\gamma_{y_i}^* < z_{yi}^* \leq \gamma_{y_i+1}^*) I(\tilde{\gamma}_{r_i}^* < z_{ri}^* \leq \tilde{\gamma}_{r_i+1}^*)\} p(\delta^*), \end{aligned}$$

where the priors for the transformed parameters take on the convenient forms:[8]

$$\beta_y^* \sim N(0, V_y) \tag{13}$$

$$\beta_r^* \sim N(0, V_r) \tag{14}$$

$$p(\gamma_3^*, \gamma_4^*, \cdots, \gamma_{Y-1}^*) \propto I\left[0 < \gamma_3^* < \gamma_4^* < \cdots < \gamma_{Y-1}^* < 1\right] \tag{15}$$

$$p(\tilde{\gamma}_3^*, \tilde{\gamma}_4^*, \cdots, \tilde{\gamma}_{R-1}^*) \propto I\left[0 < \tilde{\gamma}_3^* < \tilde{\gamma}_4^* < \cdots < \tilde{\gamma}_{R-1}^* < 1\right] \tag{16}$$

$$\Sigma^* \sim IW(2a, H) \tag{17}$$

[8]Note that the transformed cutpoints γ_j^* and $\tilde{\gamma}_k^*$ must lie between 0 and 1, and thus the priors in (15) and (16) are quite natural choices.

where IW denotes an inverted Wishart distribution,

$$\Sigma^* \equiv \begin{bmatrix} \sigma_y & \tilde{\sigma}_{yr} \\ \tilde{\sigma}_{yr} & \sigma_r \end{bmatrix}, \qquad H \equiv \begin{bmatrix} h_1 & 0 \\ 0 & h_2 \end{bmatrix},$$

and a, h_1 and h_2 are the hyperparameters employed in prior specifications (10)-(12).

We argue that there are several advantages of working with this reparameterization. First, as discussed in Nandram and Chen (1996) in the context of a single equation ordered probit, and clearly demonstrated in the following section for this system of ordered equations, the rescaling helps to mitigate correlation between the simulated cutpoints and latent data and thus improves the performance of our posterior simulator. Second, the reparameterization effectively eliminates one cutpoint from each equation in the model. For example, if there are three ordered choices for both y and r (i.e., $Y = 3$ and $R = 3$), then there are no unknown cutpoints in this specification; sampling the cutpoints follows from standard sampling of elements of the covariance matrix and no additional Metropolis-within-Gibbs steps are required. Finally, our reparameterization eliminates the diagonal restrictions on the covariance matrix in (3) and produces an unrestricted covariance matrix for the transformed latent data. This simplifies posterior simulation of the covariance parameter σ_{yr}, which cannot be drawn using standard conjugate (i.e., inverse Wishart) sampling given the restrictions on the diagonal elements in (3).[9]

Given the computational benefits afforded by the reparameterization, our recommended empirical approach is to use the reparameterized model as the working model, yet to proceed with caution, as the priors employed for the transformed parameters could potentially have unexpected implications for priors regarding the structural parameters. To this end, the expressions in (6)-(12) are particularly useful, since they provide the implied priors on the structural coefficients of interest. Our view is that with suitably chosen hyperparameters, the implied priors on the structural coefficients are still sensible and suitably diffuse, and any costs associated with the choice of prior are more than outweighed by the computational benefits offered by the reparameterization.

To investigate features of the priors in (6)-(7) in more detail, one can derive the following marginal moments for the largest cutpoints γ_Y^2 and γ_R^2:[10] $\mathrm{E}(\gamma_Y^2) = h_1(2a-1)$, $\mathrm{E}(\gamma_R^2) = h_2(2a-1)$, $\mathrm{Var}(\gamma_Y^2) = 2h_1^2(2a-1) = 2h_1 E(\gamma_Y^2)$ and $\mathrm{Var}(\gamma_R^2) = 2h_2^2(2a-1) = 2h_2 E(\gamma_Y^2)$. In terms of the marginal prior for σ_{yr} in (12), it is a reasonably "default" choice with a prior mean of zero and a variance equal to $1/[2a]$. The prior is symmetric about zero and has a mode equal to zero for $a \geq 3/2$.[11] The marginal priors for β_r and β_y can be made suitably "flat" by simply choosing V_r and V_y to be diagonal with large elements on the diagonal. In

[9] This feature has considerable appeal in an elaborated model containing more than one ordered endogenous variable. In such a model, the same reparameterization can be used to effectively "restore" the conditional conjugacy.

[10] This follows by noting from (12) that $\psi \equiv (1-\sigma_{yr}^2)$ has a $\mathrm{Beta}(a-[1/2],[1/2])$ density. Using the mean and variance of this random variable, one can then derive the unconditional prior mean and variance of γ_Y^2 and γ_R^2.

[11] It seems quite natural to us to center the model over a specification where endogeneity is not a problem (i.e. $\sigma_{yr} = 0$), yet to remain reasonably vague with this prior belief and permit the data to reveal that endogeneity is a concern.

our empirical work, we settle on $a = 2$, $h_1 = h_2 = 1$, $V_y = 1000I_{k_y}$, and $V_r = 1000I_{k_r}$. This implies a reasonably diffuse prior on the covariance parameter σ_{yr}, with a mean of zero and a variance of 1/4. With these hyperparameter values, we also obtain $\mathrm{E}(\gamma_Y^2) = \mathrm{E}(\tilde{\gamma}_R^2) = 3$, and $\mathrm{Var}(\gamma_Y^2) = \mathrm{Var}(\tilde{\gamma}_R^2) = 6$.

3 A Generated Data Experiment

We illustrate the potential benefits of working with our reparameterization through a generated data experiment. Specifically, we generate 5,000 observations from the following three-alternative [i.e., $(Y = R = 3)$] ordered equation system:

$$\begin{aligned} z_{yi} &= \beta_{y0} + x_{1i}\beta_{y1} + I(0 < z_{ri} \leq \tilde{\gamma})\theta_1 + I(\tilde{\gamma} < z_{ri})\theta_2 + \epsilon_{yi} \\ z_{ri} &= \beta_{r0} + x_{1i}\beta_{r1} + x_{2i}\beta_{r2} + x_{3i}\beta_{r3} + \epsilon_{ri}, \end{aligned}$$

where x_{1i}, x_{2i} and x_{3i} are drawn independently from a $N(0, 1)$ distribution, and $[\epsilon_{yi}\ \epsilon_{ri}]'$ are drawn jointly from a bivariate normal distribution with

$$\begin{bmatrix} \epsilon_{yi} \\ \epsilon_{ri} \end{bmatrix} \stackrel{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -.5 \\ -.5 & 1 \end{pmatrix}\right].$$

For the regression parameters, we set $\beta_{y0} = 0.5$, $\beta_{y1} = -0.4$, $\beta_{r0} = 0.3$, $\beta_{r1} = -0.6$, $\beta_{r2} = 0.2$ and $\beta_{r3} = -0.5$. For the "causal" effects (i.e., the impacts of the endogenous variable r on y), we set $\theta_1 = 1$ and $\theta_2 = 2$. The latent variables z_y and z_r are related to the observed ordered variables y and r through the following restrictions:

$$y_i = \begin{cases} 1 & \text{if } z_{yi} \leq 0 \\ 2 & \text{if } 0 < z_{yi} \leq \gamma \\ 3 & \text{if } \gamma < z_{yi} \end{cases} \quad \text{and} \quad r_i = \begin{cases} 1 & \text{if } z_{ri} \leq 0 \\ 2 & \text{if } 0 < z_{ri} \leq \tilde{\gamma} \\ 3 & \text{if } \tilde{\gamma} < z_{ri} \end{cases}.$$

Finally, we choose the cutpoint values as follows: $\gamma = 3$ and $\tilde{\gamma} = 2$.[12]

To evaluate the potential merits of using such a reparameterization, we fit the model using both our reparameterized Gibbs sampler and the standard Gibbs sampler. In each case, we use the priors described in (13)-(17) with hyperparameter values chosen as described in the previous section. Though not explicitly derived here, the standard Gibbs sampler involves sequentially simulating the cutpoints from their complete conditional distributions, which are uniform with bounds depending on the values of the neighboring cutpoints and latent data. For each parameterization, we ran the corresponding sampler for 1,000 iterations and discarded the first 200 draws as the burn-in period.

Results of this experimental exercise are summarized in Figure 1 and Table 1. In Figure 1, we plot the lagged autocorrelations up to order 20 for several selected parameters: β_{r0}, γ and σ_{yr}. As can be clearly seen from the figure, the posterior simulations from our recommended algorithm mix quite well and the lagged autocorrelations drop away reasonably

[12] Under this design, 15.1%, 79.1%, and 5.8% of the y's fall into the categories of $y = 1$, $y = 2$, and $y = 3$, respectively. Additionally, 41.4%, 49.3%, and 9.3% of the r's have the values of $r = 1$, $r = 2$, and $r = 3$, respectively.

Table 1: True values and posterior estimates of the parameters, using reparameterized Gibbs sampler and standard Gibbs sampler

	True	Reparameterized Gibbs Sampler			Standard Gibbs Sampler		
	Value	E($\beta\|D$)	Std($\beta\|D$)	P($\beta > 0\|D$)	E($\beta\|D$)	Std($\beta\|D$)	P($\beta > 0\|D$)
β_{r0}	0.3	0.284	0.0203	1	0.148	0.0278	1
β_{r1}	-0.6	-0.621	0.0195	0	-0.433	0.03	0
β_{r2}	0.2	0.192	0.0165	1	0.0736	0.0108	1
β_{r3}	-0.5	-0.505	0.018	0	-0.192	0.0171	0
β_{y0}	0.5	0.427	0.0633	1	-0.353	0.0195	0
β_{y1}	-0.4	-0.349	0.0319	0	0.0829	0.015	1
θ_1	1	1.1	0.0676	1	1.29	0.0597	1
θ_2	2	2.1	0.129	1	2.46	0.117	1
$\tilde{\gamma}$	2	1.99	0.0368	1	1.3	0.104	1
γ	3	2.96	0.0672	1	1.27	0.0761	1
σ_{yr}	-0.5	-0.538	0.0424	0	-0.972	0.00177	0

quickly.[13] In sharp contrast, the lagged autocorrelations obtained when using the standard Gibbs sampler exhibit much slower rates of decay, thus requiring substantially more simulations in order to obtain an equivalent level of numerical precision. In fact, for the cutpoint parameter γ, the lagged correlations were approximately 1 even up to order 50. Specifically, we found the lagged autocorrelations to be 0.9995, 0.9987, 0.9978, 0.9971 and 0.9963 at orders 10, 20, 30, 40 and 50, respectively.

Table 1 shows how this slow mixing problem can potentially lead to misleading inference regarding the regression and variance parameters. Clearly, given the number of iterations we ran, the posterior means of the parameters obtained using the standard Gibbs sampler do not match the actual values used to generate the data, while estimates obtained using the reparameterized sampler match the actual values quite closely. In our view, this example clearly illustrates the limitations associated with use of the standard Gibbs sampler in our ordered equation system, and motivates the potential benefits afforded by working with our suggested reparameterization.

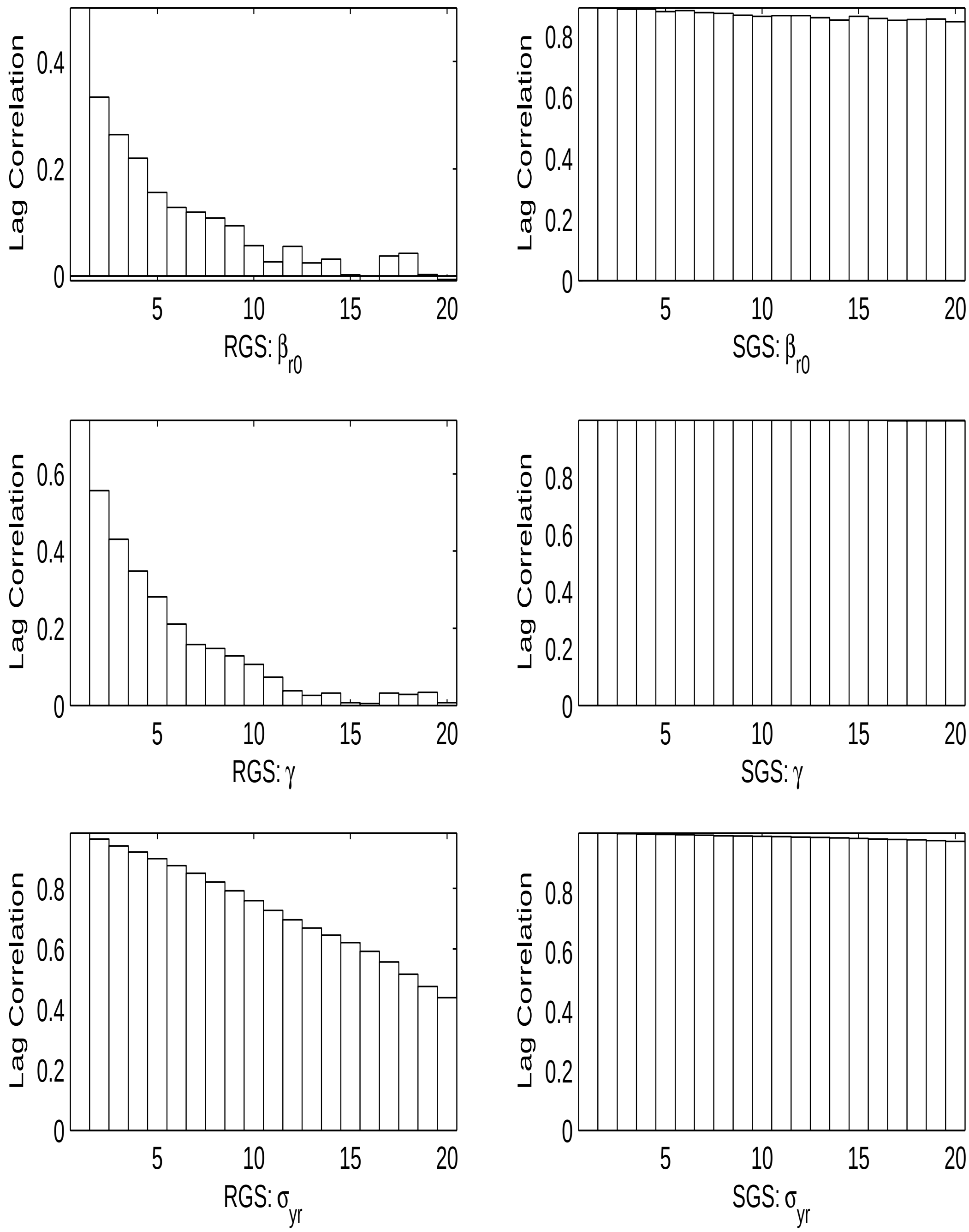
4 The Data

In the following section we provide an application of the described methodology using data from the National Longitudinal Survey of Youth (NLSY79). This data set is a widely-used panel survey of young men and women ranging in age from 14-22 in the base year (1979), and contains a wealth of information on the labor market experiences, family background characteristics, health outcomes and other demographic information of the sampled individuals. In this study, we focus primarily on variables related to fertility and associated health outcomes, and describe the variables of primary interest below.

During the 1982 and 1983 interview waves, the NLSY significantly expanded its set

[13]The lagged correlations for σ_{yr} were found to be approximately 0 for lag orders exceeding 30.

Figure 1: Lagged Autocorrelations of simulated posterior draws for β_{r0}, γ and σ_{yr}, using reparameterized Gibbs sampler (RGS) and standard Gibbs sampler (SGS)



of questions related to fertility. These expanded questions provide unusually rich information on maternal inputs during pregnancy and a variety of birth outcomes including birth weight, birth length, and information regarding early infant health. Most importantly for the purposes of our application, questionnaires in 1983 asked the female respondents to report (when appropriate) the frequency of alcohol consumption during her most recent pregnancy. In our data set, the alcohol consumption measure is recorded as a categorical (and naturally ordered) response: the mothers choose among 5 different categories which range from no alcohol consumption to consuming alcohol at least one or two days a week.[14] Additionally, in this wave of interviews, the mothers in the sample are asked to provide information on the early health outcomes of the youngest child. For the purposes of this paper, we focus on the number of months during the first year of life that the child was taken to the doctor for reasons related to illness or injury, and use this variable as our measure of infant health. By "number of months," we mean an aggregation of a set of 12 individual indicators that denote if the child was taken to see the doctor during her first month of life, during her second month of life, etc.[15] In terms of the model and notation described in (1) and (2), we make use of this data to investigate the "causal" effect of maternal alcohol consumption r on infant health y.

In the absence of an instrumental variable - some characteristic that is correlated with maternal alcohol consumption but has no structural effect on infant health conditioned on the employed controls - the parameters of the model in (1) and (2) are still identifiable based on maintained distributional assumptions and the nonlinearity of the model. In empirical practice, of course, we prefer to make use of an instrumental variable (when available) so that identification does not rely solely on functional form assumptions. To this end, we are able to gather information on whether or not the respondent's (i.e., the mother's) biological mother or father "has been an alcoholic or problem drinker at any time in his or her life," and argue that this variable can serve as an adequate instrument. Aside from environmental factors that would seem to generate correlation between the alcohol consumption of parents and their children, there is a growing awareness in the medical literature that alcoholism, like many other diseases, has a genetic component (*e.g.*, NIAAA 2003). Seemingly, then, a strong case can be made that this instrument should be correlated with maternal alcohol

[14]The raw NLSY data actually contain 8 different categories. Due to the small number of mothers found in the highest consumption categories (and in an effort to reduce measurement error in the responses), we chose to group the respondents in the "one or two days a week", "three or four days a week", "Nearly Every Day", and "Every Day" consumption categories into a single group. Our highest consumption group which combines these four categories is labeled "at least one or two days a week."

[15]Because of small cell sizes, we create and analyze the following six categories: "no months", "one month", "two months", "three months", "four to six months", and "seven to twelve months." Admittedly, this outcome variable may not be an ideal indicator of infant health for two reasons. First, it may be the case that the number of months the child was taken to the doctor is a poor proxy for the overall number of visits during the first year of life. (That is, one child may have gone to the doctor 10 times in a particular month but still be recorded as only visiting the doctor 1 month of her first year of life.) We believe, however, that this problem is likely to be minor and our indicator should be strongly correlated with the overall number of physician visits. Second, for a child to visit the doctor for reasons related to illness, she must both be sick and the mother must actually make the effort to take the child to the doctor. That is, a child may be ill frequently, but a particular mother may simply choose not to take that child to visit the doctor. As described in the following section, we attempt to mitigate this last concern by controlling for maternal "indifference" through correlation between the errors in equations (1) and (2).

consumption. We also argue that grandparental alcohol consumption should have no structural effect on infant health conditioned on maternal alcohol consumption and other included maternal characteristics such as education, "ability," family income, age and, importantly, a proxy for maternal health.[16]

In addition to the key variables listed above, we also include age of the mother at the time of childbirth, highest grade completed by the mother, an "ability" (test score)[17] measure, family income and an indicator for whether or not there are other (older) children present in the household in equations (1) and (2). These controls are added to account for characteristics of the mother that we believe may be associated with maternal alcohol consumption and infant health. After restricting our attention to models with complete data[18] for the requisite variables, we are left with a final sample of 1,122 observations. Descriptive statistics for variables used in our final sample are provided in Table 2.

4.1 The endogeneity of maternal alcohol consumption and a conceptual framework

In equation (1) z_{yi} is a latent variable which is assumed to generate our observed infant health outcome, number of months during the first year of life in which the child visited the doctor for reasons related to illness or injury. As discussed in the previous section, this latent variable can be decomposed into two parts: one part which proxies the true health status of the child and another portion which picks up the mother's willingness to take the child to the doctor for a given health level. To this end, we think that the error term ϵ_{yi} in (1) contains a mother-specific component (which is potentially correlated with observed maternal alcohol consumption), which quantifies the level of (for lack of a better term) "indifference" of the mother. That is, for a given health status of the child, more indifferent mothers are probably less likely to make the effort to take their children to the doctor, resulting in fewer observed doctor visits on average. This "indifference," of course, is also likely to have a positive effect on the propensity to consume alcohol during pregnancy, thus contributing to a negative correlation between the errors in (1) and (2). On the other hand, more indifferent mothers are probably less likely to perform acts that are typically associated with positive health outcomes for the child (e.g., providing a well-balanced diet), thus increasing the latent index z_{yi} and increasing the number of observed doctor visits. The

[16] Of course, one might argue that grandparental alcohol consumption influences maternal health independently of her decision to drink, and this unobserved maternal health may also affect infant health. As a result, the instrument would not be valid. To this end, we construct a proxy for maternal health as a dummy variable indicating if there are any health problems that prevents the mother from working at a job for pay, or limits the kind or the amount of work she can do on a job for pay prior to 1983. Though this may not be an ideal health status proxy, our belief is that it may pick up significant maternal health issues which, potentially, could arise due to the alcohol consumption patterns of her parents. We were also able to obtain subjective classifications of maternal health status for *later* years of the NLSY survey (since these are not available in 1983 or earlier), and then included these subjective health classifications in the doctor visit equation. Results were found to be robust to the inclusion of this maternal health status variable.

[17] This ability measure is the AFQT score provided in the NLSY. This score is included after first being standardized by age.

[18] In cases where the age of the youngest child at the time of the interview was less then a year, we deleted these observations, as they could not provide information on doctor visits for the full 12 month period.

latter part of this story is suggestive of a possible positive correlation between the errors in (1) and (2). *A priori*, we are not sure which of these two effects should dominate, but both stories suggest the potential for correlation among the outcome errors in our system and thus motivate the need to control for the endogeneity of alcohol consumption.

Relatedly, in some cases the coefficients on the observables in our health outcome equation (1) may be difficult to interpret. To show why this might be the case, let us introduce an illustrative conceptual model which formally incorporates the effects described in the previous paragraph:

$$z_{yi} = \alpha_1 H_i + \alpha_2 P_i + \tilde{\epsilon}_{yi} \tag{18}$$

$$H_i = x_{1i}\beta_1 + d_{ri}\theta_1 + \nu_i \tag{19}$$

$$P_i = x_{2i}\theta_2 + \omega_i \tag{20}$$

$$z_{ri} = x_{ri}\beta + I_i + \tilde{\epsilon}_{ri} \tag{21}$$

Equation (18) writes the latent variable generating observed doctor visits, z_{yi}, as a combination of two factors: the "true" (but, unfortunately, unobserved) child health status, denoted H_i, and an effect arising from a second variable P_i, which denotes the mother's unobserved propensity to take the child to the doctor (for a given health level). Both of these variables combine to form an overall index, which is assumed to generate the observed number of doctor visits. Quite naturally, we expect $\alpha_1 < 0$ and $\alpha_2 > 0$ so that children in poorer health are likely to have more doctor visits, and mothers with a predisposition to have their child examined by a physician at the onset of an illness are more likely to take their children to the doctor.

In equation (19), infant health level H is written as a function of observables x_1 and d_r, the latter denoting the observed amount of maternal alcohol consumption. Similarly, in (20) P_i is written as a linear function of observables x_2 [which could potentially contain the covariates in (19)]. In the final equation, the latent variable generating alcohol consumption z_r is written as a function of observables x_r, and we have decomposed the error term into unobserved maternal "indifference" I and a random error $\tilde{\epsilon}_r$. As discussed previously, we expect that unobserved "indifference" I is negatively correlated with both ω and ν.

Of course, upon substituting the equations for H and P into equation (18), the four equation system in (18)-(21) reduces to a two equation ordered system as in (1) and (2). The resulting sign of the correlation between the final composite errors of the two equations is unclear, though we have strong reason to suspect that a non-zero correlation may exist. As an additional concern, if x_1 and x_2 share common elements, then the coefficients appearing on those variables common to (19) and (20) must be interpreted as picking up the combined contribution of those covariates on both H and P. For example, it seems reasonable to expect that having other children in the household may lower infant health H (since the infant will typically be exposed to more illnesses), but at the same time, the experience of having children before may make the mother more confident in handling health issues without needing to see a doctor (thus lowering P). So, this variable would presumably be present as a covariate in both equations H and P (with offsetting effects on our observed outcome variable), and upon estimating the model in (1) and (2), we are not able to identify its individual contribution to each of these equations.

We are less concerned, however, about drawing correct conclusions regarding the "significance" of our alcohol consumption variable, the primary covariate of interest. In particular, if one were to argue that maternal alcohol consumption should also be included as an explanatory variable in the equation generating P, then presumably it would be negatively correlated with P - mothers who are more likely to drink probably have a lower propensity to take their children to the doctor, holding other factors constant. Therefore, if we find that the coefficient estimates on maternal alcohol consumption in (1) are positive, then it must be the case that maternal alcohol consumption has a negative effect on infant health; there is no other way to obtain such a positive coefficient under the seemingly reasonable assumptions that $\alpha_1 < 0$, $\alpha_2 > 0$ and alcohol consumption correlates negatively with P.[19] As we describe in the following section, we find strong evidence of positive coefficients associated with maternal alcohol consumption, and thus draw the conclusion that maternal alcohol consumption has a negative and significant impact on early infant health.

5 Empirical Results

Before providing coefficient estimates from our preferred simultaneous equations model, it is, perhaps, useful to present estimates obtained when considering equations (1) and (2) separately. Such an estimation procedure would be appropriate if $\sigma_{yr} = 0$, thus eliminating the need to control for the endogeneity of maternal alcohol consumption. This assumption is often implicitly made in biomedical research on this topic, as single-equation methods are routinely used, and mean independence assumptions are assumed to hold given the inclusion of sufficient controls.

Results of these equation-by-equation ordered probit analyses are presented in Table 3.[20] What is most important to note from this first pass at the data are the coefficients

[19] Upon substitution, the "final" coefficient on the alcohol consumption variable is of the form $\alpha_1\theta_1+\alpha_2\theta_{2r}$, where θ_{2r} denotes the coefficient on the alcohol consumption variable in (20). If $\alpha_1 < 0$, $\alpha_2 > 0$, $\theta_{2r} < 0$ and the entire sum is found to be positive, then it must be the case that $\theta_1 < 0$ so that maternal alcohol consumption has a negative impact on infant health.

[20] The last column of Table 3 calculates the marginal effect associated with each covariate. For example, the quantity $\Delta\mathrm{P}(r > 1|\Delta x_{r,k_r}, Data)$ measures the marginal effect of the k_rth control variable x_{r,k_r} in the drinking frequency equation on the probability of having any alcohol consumption during pregnancy, $\mathrm{P}(r > 1)$. Formally, $\Delta\mathrm{P}(r > 1|\Delta x_{r,k_r}, Data) = \mathrm{P}(r > 1|x_{r,-k_r} = x^*_{r,-k_r}, x_{r,k_r} = x^*_{r,k_r} + 1, Data) - \mathrm{P}(r > 1|x_{r,-k_r} = x^*_{r,-k_r}, x_{r,k_r} = x^*_{r,k_r}, Data)$, where $x_{r,-k_r}$ denotes the covariates in the drinking frequency equation other than the k_rth variable x_{r,k_r}, and $x^*_{r,-k_r}$ and x^*_{r,k_r} are original levels of the control variables. For the noncategorical variables in the drinking frequency equation, we set them originally at their sample averages as listed in the descriptive statistics shown in Table 2. For categorical variables, they are originally set to their default categories as defined in Table 2.

Likewise, in the last column of Table 3, the variable $\Delta\mathrm{P}(y > 1|\Delta x_{y,k_y}, Data)$ captures the marginal effect of the k_yth control variable x_{y,k_y} in the doctor visit equation on the probability of having the youngest child treated by doctor for illness during any months of the first year, $\mathrm{P}(y > 1)$. Consistent with our practice in the drinking frequency equation, to calculate the marginal effects in the doctor visit equation, we set the noncategorical covariates originally at their sample averages and the categorical variables in their omitted categories. For example, for the endogenous dummy variables characterizing the drinking frequency of the mother, the default group refers to mothers who do not drink during their pregnancies.

on the maternal alcohol consumption variables.[21] As we can see from the table, children of mothers consuming moderate to large amounts of alcohol while pregnant (i.e., drinking at least one or two days a week while pregnant) tend to experience more doctor visits during the first year of life than children whose mothers never consumed alcohol. The evidence in this regard, however, is not overwhelming, as we find a small posterior probability that the coefficient for the highest consumption category is positive. In addition, there is little evidence that children whose mothers consumed small to moderate quantities of alcohol while pregnant are associated with any increase in the number of doctor visits.

Of course, one might question these results and suspect that selection bias remains an important concern, as unobservable confounding may still exist even with the given set of controls. If the correlation among the unobservables in equations (1) and (2) is negative, for example, (as could be the case if more indifferent mothers have a higher propensity to drink and are also less likely to take their children to the doctor when sick), then we might suspect that the single-equation ordered probit estimates actually understate the true impact of alcohol consumption on infant health. To this end, we now take up the case of our more general model which allows for the endogeneity of maternal alcohol consumption.

5.1 Posterior results for the two equation system

We fit our system of ordered outcomes using the posterior simulator described in the appendix. This posterior simulator makes use of Gibbs steps to simulate the majority of parameters in the model, but uses Metropolis-within-Gibbs steps based on Dirichlet proposal densities to simulate the transformed cutpoints γ^* and $\tilde{\gamma}^*$. The posterior simulator is run for 20,000 iterations and the first 4,000 are discarded as the burn-in. Coefficient posterior means, standard deviations, probabilities of being positive, and point estimates of marginal effects are provided in Table 4.[22]

The top panel of Table 4 presents posterior results for the parameters of equation (2) describing the quantity of alcohol consumption during pregnancy. We first see that the instrument, an indicator denoting if the mother had a biological parent who was a "problem drinker or alcoholic," is strongly correlated with maternal alcohol consumption. Mothers with at least one parent who was a problem drinker are approximately 9 percent more likely to consume at least some amount of alcohol during pregnancy than those mothers whose parents were not problem drinkers. Somewhat surprisingly, the coefficients associated with the other variables are often insignificant and typically possess unexpected signs. The coefficients associated with education and test scores, for example, are positive, and for the case of test scores, have a very low posterior probability of being negative. As can be seen from the magnitude of the marginal effect estimates, however, we should not make too much of these positive coefficients, since these variables seemingly play minor roles in explaining maternal alcohol consumption decisions.

[21]Mothers who report to have "never" consumed alcohol while pregnant are the excluded group, so the coefficients on the remaining dummies should be interpreted as relative to that group.

[22]For the sake of brevity, we do not present posterior information regarding the cutpoints from each equation, though these details are available upon request.

Posterior results for our health outcome equations are presented in the bottom panel of Table 4. Most importantly, we see positive coefficients associated with our maternal alcohol consumption variables, and these coefficients are generally increasing with the level of alcohol consumption. Specifically, mothers who reported to drink at least one or two days a week during their pregnancies are estimated to be 30 percent more likely to take their children to the doctor at least once during the child's first year of life than mothers who report "never" consuming alcohol while pregnant. Those children of mothers consuming relatively "small" but positive amounts of alcohol while pregnant (say, at most 3 or 4 days a month) are approximately 20 percent more likely to take their child to a doctor for reasons related to illness than children of mothers who completely avoid alcohol while pregnant. Older and more educated mothers tend to have children associated with more doctor visits, while more experienced mothers (i.e., those who have had children before) tend to make fewer physician visits. Again, similar to our discussion in the previous sections, these variables are likely to proxy a mother's unobserved propensity to take the child to the physician rather than reflecting a structural effect related to infant health. Finally, we also note that our proxy for maternal health status (denoted "Health Problem" in Tables 3 and 4) plays some role in our equation describing the number of observed doctor visits. In particular, children of mothers who report having a health problem limiting their ability to work are associated with a 3.6 percent larger probability of doctor visits, potentially suggesting that maternal health (or lack thereof) is, perhaps to a small degree, transmitted to the child.

We conclude by noting that the magnitude of the impacts of maternal alcohol consumption from our joint estimation procedure are larger than those suggested by our previous single equation analyses. We expected to observe such an increase if selection bias was indeed an important concern, and in particular, if there was a negative correlation between the unobservables of equations (1) and (2). The last row of Table 4 provides rather strong evidence that a non-zero correlation exists. The posterior mean of the correlation between the errors of our two equations is -.4, and the marginal posterior density places most of its mass over negative values (i.e., the posterior probability that the coefficient is negative is .973). This result clearly suggests the need to allow for the potential of unobservable confounding in our application, even with a reasonably rich set of employed controls.

6 Conclusion

We have described a new simulation-based Bayesian algorithm for fitting "treatment" effect models when both the outcome of interest and the endogenous treatment variable are ordered. A generated data experiment shows how this new posterior simulator (based on a rescaling transformation) may lead to improved mixing of the simulated parameters relative to use of the standard Gibbs sampler. This rescaling transformation was also shown to simplify some of the posterior computations, and in the specific case where there are 3 possible alternatives for either outcome, the need for traditional methods to simulate cutpoints associated with that variable is eliminated. It is our hope that the algorithm provided will be useful to other empirical researchers seeking to estimate treatment effect models with a similar structure.

Using data from the National Longitudinal Survey of Youth (NLSY) we also applied our methods in practice and investigated the effect of maternal alcohol consumption on early infant health. Our results revealed clear evidence that the health outcomes of infants whose mothers consumed alcohol while pregnant were worse than the outcomes of infants whose mothers never consumed alcohol while pregnant. In addition, the estimated parameters of our model clearly suggested the need to allow for the endogeneity of maternal alcohol consumption when seeking to identify its effect on early infant health.

Appendix: The Gibbs Algorithm

We employ the Gibbs sampler to fit the model described by (1) and (2). As mentioned in section 2, to improve the performance of the standard sampler, we follow Nandram and Chen (1996) and introduce a rescaling transformation in each of the ordered outcome equations. Specifically, we introduce the reparameterizations:

$$\sigma_y = 1/[\gamma_Y^2], \quad \beta_y^* = \sqrt{\sigma_y}\beta_y, \quad z_y^* = \sqrt{\sigma_y}z_y, \quad \gamma^* = \sqrt{\sigma_y}\gamma$$

and

$$\sigma_r = 1/[\tilde{\gamma}_R^2], \quad \beta_r^* = \sqrt{\sigma_r}\beta_r, \quad z_r^* = \sqrt{\sigma_r}z_r, \quad \tilde{\gamma}^* = \sqrt{\sigma_r}\tilde{\gamma}, \;\; \tilde{\sigma}_{yr} = \sqrt{\sigma}_y\sqrt{\sigma}_r\sigma_{yr}.$$

Multiplying the latent variable equation in (1) on both sides by $\sqrt{\sigma_y}$, multiplying (2) on both sides by $\sqrt{\sigma_r}$, and using the parameterization above, we obtain an equivalent model of the form

$$z_{yi}^* = x_{yi}\beta_y^* + u_{yi} \tag{22}$$
$$z_{ri}^* = x_{ri}\beta_r^* + u_{ri} \tag{23}$$

with $y_i = j$ if $\gamma_j^* < z_{yi}^* \leq \gamma_{j+1}^*$ and $r_i = k$ if $\tilde{\gamma}_k^* < z_{ri}^* \leq \tilde{\gamma}_{k+1}^*$. The Normality assumption in (3) implies

$$\begin{bmatrix} u_{yi} \\ u_{ri} \end{bmatrix} \overset{iid}{\sim} N(0_2, \Sigma) \quad \text{where} \quad \Sigma = \begin{bmatrix} \sigma_y & \tilde{\sigma}_{yr} \\ \tilde{\sigma}_{yr} & \sigma_r \end{bmatrix} \tag{24}$$

and $\tilde{\sigma}_{yr} \equiv \sigma_{yr}\sqrt{\sigma_y}\sqrt{\sigma_r}$. In the algorithm below, we employ *blocking steps* where the transformed cutpoints and transformed latent data are drawn together in a single block to improve the overall performance of our sampler.

Gibbs Algorithm

1. Sample the coefficients $\beta^* = [\beta_y^{*\prime} \;\; \beta_r^{*\prime}]'$:

$$\beta^*|\Xi_{-\beta^*}, Data \quad \sim \quad N(D_\beta d_\beta, D_\beta),$$

where $\Xi_{-\theta}$ denotes all the parameters other than the parameter θ, $D_\beta = [X'(\Sigma^{*-1} \otimes I_n)X + V_\beta^{-1}]^{-1}$, $d_\beta = X'(\Sigma^{*-1} \otimes I_n)z^* + V_\beta^{-1}\beta_0$, $X = \begin{pmatrix} X_y & 0_{n\times k_r} \\ 0_{n\times k_y} & X_r \end{pmatrix}$, $V_\beta = \begin{pmatrix} V_y & 0_{k_y\times k_r} \\ 0_{k_r\times k_y} & V_r \end{pmatrix}$, $\beta_0 = \begin{pmatrix} 0_{k_y\times 1} \\ 0_{k_r\times 1} \end{pmatrix}$, and $z^* = \begin{pmatrix} z_y^* \\ z_r^* \end{pmatrix}$.

2. Sample the truncation points in the doctor visit equation, $\{\gamma_j^*\}_{j=3}^{Y-1}$, from its conditional posterior marginalized over z_y^*:

$$p(\{\gamma_j^*\}_{j=3}^{Y-1}|\Xi_{-\{\gamma_j^*\}_{j=3}^{Y-1}, z_y^*}, Data) \quad \propto \quad \prod_{i=1}^{n}[\Phi([\gamma_{y_i+1}^* - \mu_{y|r}]/\sqrt{\sigma_{y|r}}) - \Phi([\gamma_{y_i}^* - \mu_{y|r}]/\sqrt{\sigma_{y|r}})],$$

where Φ denotes the cumulative distribution function of the Normal density, $\mu_{y|r} \equiv x_{yi}\beta_y^* + \tilde{\sigma}_{yr}\sigma_r^{-1}(z_{ri}^* - x_{ri}\beta_r^*)$, and $\sigma_{y|r} \equiv \sigma_y - \tilde{\sigma}_{yr}^2\sigma_r^{-1}$. Following the reasoning of Nandram and Chen (1996), we use a Dirichlet proposal density to sample the *differences* between cutpoint values, $q_j \equiv \gamma_{j+1}^* - \gamma_j^*,\ j = 3, \cdots Y-1$, and then solve back for $\{\gamma_j^*\}$. Specifically, we sample a candidate value, say $\{q_j^{can}\}_{j=3}^{Y-1} \sim Dirichlet(\{\alpha_j n_j + 1\}_{l=3}^{Y-1})$, where "can" denotes the candidate draw, $\{\alpha_j\}_{j=3}^{Y-1} = 0.1$ are tuning parameters, and $n_j \equiv \sum_{i=1}^n I(y_i = j),\ j = 3, \cdots Y-1$ are the numbers of individuals falling into each category of the outcome variable. The probability of accepting the candidate draw is $\min(R, 1)$, where

$$R = [\prod_{i=1}^n \frac{\Phi([\gamma_{y_i+1}^{can} - \mu_{y|r}]/\sqrt{\sigma_{y|r}}) - \Phi([\gamma_{y_i}^{can} - \mu_{y|r}]/\sqrt{\sigma_{y|r}})}{\Phi([\gamma_{y_i+1,l-1}^* - \mu_{y|r}]/\sqrt{\sigma_{y|r}}) - \Phi([\gamma_{y_i,l-1}^* - \mu_{y|r}]/\sqrt{\sigma_{y|r}})}][\prod_{j=3}^{Y-1}(\frac{q_{j,l-1}}{q_j^{can}})^{\alpha_j n_j}],$$

and "$_{l-1}$" denotes the current value of the algorithm.

3. Sample the latent outcome in the doctor visit equation, $z_{yi}^*,\ i = 1, 2, \cdots n$, from the complete conditional:

$$z_{yi}^*|\Xi_{-z_{yi}^*}, Data \quad \overset{ind}{\sim} \quad TN_{(\gamma_{y_i}^*, \gamma_{y_i+1}^*]}(\mu_{y|r}, \sigma_{y|r}),$$

where $TN_{(a,b]}(\mu, \sigma)$ denotes a Normal distribution with mean μ and variance σ truncated to the interval between a and b.

4. Sample the cutpoints relevant to the drinking frequency equation from the posterior conditional for $\{\tilde{\gamma}_k^*\}_{k=3}^{R-1}$ marginalized over z_r^*:

$$p(\{\tilde{\gamma}_k^*\}_{k=3}^{R-1}|\Xi_{-\{\tilde{\gamma}_k^*\}_{k=3}^{R-1}, z_r^*}, Data) \quad \propto \quad \prod_{i=1}^n [\Phi([\tilde{\gamma}_{r_i+1}^* - \mu_{r|y}]/\sqrt{\sigma_{r|y}}) - \Phi([\tilde{\gamma}_{r_i}^* - \mu_{r|y}]/\sqrt{\sigma_{r|y}})],$$

where $\mu_{r|y} = x_{ri}\beta_r^* + \tilde{\sigma}_{yr}\sigma_y^{-1}(z_{yi}^* - x_{yi}\beta_y^*)$, and $\sigma_{r|y} = \sigma_r - \tilde{\sigma}_{yr}^2\sigma_y^{-1}$. As in step (2) we use a Dirichlet proposal density to sample the differences between the cutpoints $\tilde{q}_k \equiv \tilde{\gamma}_{k+1}^* - \tilde{\gamma}_k^*,\ k = 3, \cdots R-1$, and specifically sample these differences from a $Dirichlet(\{\tilde{\alpha}_k\tilde{n}_k + 1\}_{k=3}^{R-1})$ proposal density, where $\{\tilde{\alpha}_k\}_{k=3}^{R-1} = 0.05$ are the tuning parameters and $\tilde{n}_k \equiv \sum_{i=1}^n I(r_i = k),\ k = 3, \cdots R-1$ are the numbers of individuals falling into each category of the outcome variable. The probability of accepting the candidate draw is $\min(R, 1)$, where

$$R = [\prod_{i=1}^n \frac{\Phi([\tilde{\gamma}_{r_i+1}^{can} - \mu_{r|y}]/\sqrt{\sigma_{r|y}}) - \Phi([\tilde{\gamma}_{r_i}^{can} - \mu_{r|y}]/\sqrt{\sigma_{r|y}})}{\Phi([\tilde{\gamma}_{r_i+1,l-1}^* - \mu_{r|y}]/\sqrt{\sigma_{r|y}}) - \Phi([\tilde{\gamma}_{r_i,l-1}^* - \mu_{r|y}]/\sqrt{\sigma_{r|y}})}][\prod_{k=3}^{R-1}(\frac{\tilde{q}_{k,l-1}}{\tilde{q}_k^{can}})^{\tilde{\alpha}_k\tilde{n}_k}].$$

5. Sample the latent outcomes in the drinking frequency equation:

$$z_{ri}^*|\Xi_{-z_{ri}^*}, Data \quad \overset{ind}{\sim} \quad TN_{(\tilde{\gamma}_{r_i}^*, \tilde{\gamma}_{r_i+1}^*]}(\mu_{r|y}, \sigma_{r|y}),$$

for $i = 1, 2, 3, \cdots, n$.

6. Sample the covariance matrix Σ^*:

$$\begin{aligned} &\Sigma^*|\Xi_{-\Sigma^*}, Data \\ \sim \quad & IW\left(2a + n, \{H^{-1} + \sum_{i=1}^n [(z_{yi}^* - x_{yi}\beta_y^*)\ (z_{ri}^* - x_{ri}\beta_r^*)]'[(z_{yi}^* - x_{yi}\beta_y^*)\ (z_{ri}^* - x_{ri}\beta_r^*)]\}^{-1}\right). \end{aligned} \tag{25}$$

After each iteration we rescale all the parameters by dividing $\sqrt{\sigma_y}$ into β_y^* and $\{\gamma_j^*\}_{j=3}^{Y}$, and by dividing $\sqrt{\sigma_r}$ into β_r^* and $\{\tilde{\gamma}_k^*\}_{k=3}^{R}$. We provide the posterior summary statistics for these coefficients in the tables.

References

Ashenfelter, O., and A. Krueger (1994): "Estimates of the economic return to schooling from a new sample of twins," *American Economic Review*, 84(5), 1157-1173.

Angrist, J., and W. Evans (1998): "Children and their parents' labor supply: Evidence from exogenous variation in family size," *American Economic Review*, 88, 450-477.

Card, D. (1999): "The causal effect of education on earnings," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics Vol 3A*, 1801-1863.

Cowles, M.K. (1996): "Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models," *Statistics and Computing*, 6, 101-111.

Goldschmidt, L., G. A. Richardson, D. S. Stoffer, D. Geva, and N.L. Day (1996): "Prenatal alcohol exposure and academic achievement at age six: a nonlinear fit," *Alcoholism: Clinical and Experimental Research*, 20(4), 763-770.

Imbens, G. W., and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467-475.

Jacobson, S. W., J. L. Jacobson, R. J. Sokol, S. S. Martier, and J. W. Ager (1993): "Prenatal alcohol exposure and infant information processing ability," *Child Development*, 64(6), 1706-1721.

Jacobson, S. W., J. L. Jacobson, and R. J. Sokol (1994): "Effects of fetal alcohol exposure on infant reaction time," *Alcoholism: Clinical and Experimental Research*, 18(5), 1125-1132.

Li, K., and D. J. Poirier (2001): "Using the national longitudinal survey of youth in the US to study the birth process: A Bayesian approach," *Research in Official Statistics*, 4, 127-150.

Li, K., and D. J. Poirier (2003a): "An econometric model of birth inputs and outputs for native Americans," *Journal of Econometrics*, 113, 337-361.

Li, K., and D. J. Poirier (2003b): "The roles of birth inputs and outputs in predicting health, behaviour and test scores in early childhood," *Statistics in Medicine*, 22, 3489-3514.

Li, K., and D. J. Poirier (2003c): "Bayesian analysis of an econometric model of birth inputs and outputs," *Journal of Population Economics*, 16, 597-625.

Munkin, M. K., and P. K. Trivedi (2003): "Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare," *Journal of Econometrics*, 114(2), 197-220.

Nandram, B., and M.-H. Chen (1996): "Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence," *Journal of Statistical Computation and Simulation*, 54, 129-144.

National Institute on Alcohol Abuse and Alcoholism (NIAAA) (2003): "The genetics of alcoholism," *Alcohol Alert*, No. 60, Rockville, MD.

Poirier, D. J. (1995): *Intermediate Statistics and Econometrics*, MIT Press: Cambridge, MA.

Staiger, D., and J. H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557-586.

Table 2: Descriptive statistics: drinking frequency equation and doctor visit equation

	Drinking Frequency Equation	
	Sample Mean	Standard Error
Drinking frequency		
Never	0.656	0.475
Less than once a month	0.168	0.373
About once a month	0.0722	0.259
3 or 4 days a month	0.0517	0.221
At least 1 or 2 days a week	0.0526	0.223
Race[a]		
Hispanic	0.165	0.371
Black	0.335	0.472
Cognitive test score	0	1
Education	11.4	1.86
Age	20.1	2.21
Having children before	0.386	0.487
Family income ($10,000)	1.57	1.2
Biological parent alcoholic	0.299	0.458

	Additional Variables in Doctor Visit Equation	
	Sample Mean	Standard Error
Number of months treated for illness		
No months	0.387	0.487
One month	0.307	0.461
Two months	0.127	0.333
Three months	0.0722	0.259
Four to six months	0.0561	0.23
Seven to twelve months	0.0499	0.218
Youngest child female	0.477	0.499
Health problem	0.369	0.483

[a]For race, the excluded group are individuals other than Hispanics and blacks.

Table 3: Single equation estimates: drinking frequency equation and doctor visit equation

	Drinking Frequency Equation			
	$E(\beta\|D)$	$Std(\beta\|D)$	$P(\beta > 0\|D)$	$\Delta P(r > 1\| \Delta x_{r,k_r}, D)$[a]
Constant	-1.11	0.396	0.00338	
Race				
Hispanic	-0.211	0.11	0.0257	-0.0715
Black	-0.0447	0.0963	0.32	-0.0152
Cognitive test score	0.0979	0.0457	0.984	0.0366
Education	0.0169	0.0257	0.745	0.00625
Age	0.0241	0.0193	0.894	0.00884
Having children before	0.0767	0.0847	0.817	0.0288
Family income ($10,000)	-0.0207	0.032	0.259	-0.00755
Biological parent alcoholic	0.21	0.0774	0.996	0.0796
	Doctor Visit Equation			
	$E(\beta\|D)$	$Std(\beta\|D)$	$P(\beta > 0\|D)$	$\Delta P(y > 1\| \Delta x_{y,k_y}, D)$[b]
Constant	-1	0.353	0.00219	
Race				
Hispanic	0.00922	0.0955	0.537	0.00321
Black	-0.317	0.0871	0.000125	-0.125
Cognitive test score	-0.0353	0.0422	0.2	-0.0136
Education	0.0405	0.0228	0.963	0.0154
Age	0.0481	0.017	0.998	0.0183
Having children before	-0.0825	0.0756	0.136	-0.0322
Family income ($10,000)	-0.0164	0.029	0.285	-0.0063
Youngest child female	-0.0745	0.0652	0.127	-0.0289
Health problem	0.101	0.0671	0.935	0.0376
Drinking frequency				
Less than once a month	0.157	0.0903	0.958	0.0575
About once a month	-0.0568	0.129	0.332	-0.023
3 or 4 days a month	-0.135	0.151	0.187	-0.0533
At least 1 or 2 days a week	0.134	0.146	0.822	0.0486

[a] The quantity $\Delta P(r > 1|\Delta x_{r,k_r}, Data)$ measures the marginal effect of the k_rth control variable x_{r,k_r} in the drinking frequency equation on the probability of having any alcohol consumption during pregnancy, $P(r > 1)$.

[b] The variable $\Delta P(y > 1|\Delta x_{y,k_y}, Data)$ captures the marginal effect of the k_yth control variable x_{y,k_y} in the doctor visit equation on the probability of having the youngest child treated by doctor for illness during any months of the first year, $P(y > 1)$.

Table 4: Simultaneous equation estimates: drinking frequency equation and doctor visit equation

	Drinking Frequency Equation			
	$E(\beta\|D)$	$Std(\beta\|D)$	$P(\beta > 0\|D)$	$\Delta P(r > 1\| \Delta x_{r,k_r}, D)^a$
Constant	-1.14	0.393	0.00119	
Race				
Hispanic	-0.203	0.109	0.0331	-0.0695
Black	-0.0115	0.0972	0.455	-0.00358
Cognitive test score	0.105	0.0449	0.99	0.0395
Education	0.0121	0.0254	0.685	0.00446
Age	0.0275	0.019	0.928	0.0101
Having children before	0.0725	0.0834	0.812	0.0275
Family income ($10,000)	-0.0185	0.0314	0.279	-0.00657
Biological parent alcoholic	0.243	0.0733	0.999	0.0926
	Doctor Visit Equation			
	$E(\beta\|D)$	$Std(\beta\|D)$	$P(\beta > 0\|D)$	$\Delta P(y > 1\| \Delta x_{y,k_y}, D)^b$
Constant	-1.04	0.347	0.00144	
Race				
Hispanic	0.0623	0.0974	0.736	0.0233
Black	-0.281	0.0878	0.0005	-0.111
Cognitive test score	-0.0623	0.0429	0.0737	-0.0241
Education	0.0357	0.0224	0.944	0.0136
Age	0.0391	0.0174	0.988	0.0149
Having children before	-0.0975	0.0756	0.0978	-0.0382
Family income ($10,000)	-0.00886	0.0292	0.379	-0.00344
Youngest child female	-0.0658	0.0608	0.141	-0.0256
Health problem	0.0942	0.0633	0.934	0.0355
Drinking frequency				
Less than once a month	0.63	0.214	0.99	0.205
About once a month	0.619	0.307	0.959	0.197
3 or 4 days a month	0.674	0.369	0.945	0.207
At least 1 or 2 days a week	1.16	0.448	0.981	0.299
Correlation: $\sigma_{yr} = \rho_{yr}$	-0.399	0.166	0.0268	

[a]The quantity $\Delta P(r > 1|\Delta x_{r,k_r}, Data)$ measures the marginal effect of the k_rth control variable x_{r,k_r} in the drinking frequency equation on the probability of having any alcohol consumption during pregnancy, $P(r > 1)$.

[b]The variable $\Delta P(y > 1|\Delta x_{y,k_y}, Data)$ captures the marginal effect of the k_yth control variable x_{y,k_y} in the doctor visit equation on the probability of having the youngest child treated by doctor for illness during any months of the first year, $P(y > 1)$.