Available online at www.sciencedirect.com

ScienceDirect

Journal of Econometrics 140 (2007) 781–801

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Modeling and calculating the effect of treatment at baseline from panel outcomes

Siddhartha Chib[a,*], Liana Jacobi[b]

[a] John M. Olin School of Business, Campus Box 1133, Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130, USA

[b] Department of Economics, The University of Melbourne, Victoria 3010, Australia

Available online 7 September 2006

Abstract

We propose and examine a panel data model for isolating the effect of a treatment, taken once at baseline, from outcomes observed over subsequent time periods. In the model, the treatment intake and outcomes are assumed to be correlated, due to unobserved or unmeasured confounders. Intake is partly determined by a set of instrumental variables and the confounding on unobservables is modeled in a flexible way, varying both by time and treatment state. Covariate effects are assumed to be subject-specific and potentially correlated with other covariates. Estimation and inference is by Bayesian methods that are implemented by tuned Markov chain Monte Carlo methods. Because our analysis is based on the framework developed by Chib [2004. Analysis of treatment response data without the joint distribution of counterfactuals. Journal of Econometrics, in press], the modeling and estimation does not involve either the unknowable joint distribution of the potential outcomes or the missing counterfactuals. The problem of model choice through marginal likelihoods and Bayes factors is also considered. The methods are illustrated in simulation experiments and in an application dealing with the effect of participation in high school athletics on future labor market earnings.

Keywords: Confounding; Heterogeneity; Instrumental variable; Marginal likelihood; Markov chain Monte Carlo; Metropolis–Hastings algorithm; Non-randomly assigned treatment; Panel data; Potential outcomes; Treatment effect

[*] Corresponding author. Tel.: +1 314 935 6359.
E-mail addresses: chib@wustl.edu (S. Chib), ljacobi@unimelb.edu.au (L. Jacobi).

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.07.014

1. Introduction

This paper develops a Bayesian framework to estimate the effect of a treatment taken at baseline from a panel of outcomes. Specifically, we consider the following situation. A group of individuals is observed at a base period, when some of the individuals receive a one-time *binary* treatment. They are then also observed over several subsequent time periods. The research question is whether the treatment intake at baseline enhances the performance of individuals in terms of some outcome measure in those succeeding time periods. An example of this situation is the determination of the effect of participation in high school athletics (the treatment intake) on later labor market outcomes. In this, and other similar cases, the central complication is that the outcome is likely to be affected not just by the intake but also by unmeasured or unobserved confounders.

One way of proceeding in such cases is to ignore the panel structure and use existing methods to model the treatment and the outcome as separate cross-section models, one for each time period. For example, one could model the treatment intake at baseline and the outcome in the first time period, ignoring the remaining outcomes. One could next model the treatment intake at baseline and the outcome in the second time period, ignoring the outcome in the first time period and the outcomes beyond the second time period. Naturally, one problem with this approach is that it ignores the joint dependence in the outcomes arising from the panel structure. There is thus the potential that the separate cross-section fitting will produce an incorrect sequence of treatment effect estimates.

In this paper we provide a different modeling remedy that is a variant of the Roy switching regression model (Lee, 1978), extended to the setting of panel outcomes and treatment at baseline. In this extension of the model, there are two potential outcome sequences, one for each level of the treatment. Intake at baseline is partly determined by a set of instrumental variables and the confounding on unobservables is modeled in a flexible way, varying both by time and potential outcome. Gaussianity is avoided by assuming that the joint distribution of the intake and each potential outcome sequence is multivariate-t. As is customary in Bayesian panel data modeling, covariate effects are assumed to be subject-specific and potentially correlated with other covariates.

The general modeling strategy is related to that of Chib and Hamilton (2002) except that they dealt with the situation of time varying treatments. Although our set-up is in some ways more restrictive, the single treatment intake at baseline leads to new estimation and inferential concerns. Yau and Little (2001) have also considered a related baseline treatment setting but the model in their paper is different, stemming from the work of Imbens and Rubin (1997) and Hirano et al. (2000). The two approaches are not compared because treatment intake in the latter set-up is only possible when the single binary instrument is one, an assumption that is not made in this paper.

The model proposed in this paper can be readily processed by Bayesian Markov chain Monte Carlo (MCMC) methods even though the likelihood function of the parameters is not in closed form. An important aspect of the modeling and estimation is that neither involves the unknowable joint distribution of the potential outcomes. As a result, the unidentified parameters of that joint distribution and the missing counterfactuals are not needed in the prior-posterior analysis, which simplifies the computations considerably. The fact that the analysis can proceed in this way is due to Chib (2004) where full details can be found. In addition, we discuss a predictive method for calculating the effect of the treatment on the outcomes, leading to various effect summaries, for example, the average and quantile treatment effects. Because we operate under the framework of Chib (2004), these predictive treatment effects are based only on the marginal distribution of the potential outcomes. This is in contrast to the predictive effects in Chib and Hamilton (2002) and Li et al. (2004) and other papers by these and many other authors, which need not be just the marginal, but also the unidentified joint distribution of the potential outcomes. Our final methodological contribution relates to the problem of model choice where we show how the approach of Chib (1995) can be used to find the model marginal likelihood and Bayes factors for competing model specifications.

We illustrate the efficacy of our modeling and inferential framework first in simulation experiments where we delineate what is achieved by the full modeling of the panel structure and what is lost by ignoring it. Next the ideas are illustrated in a problem concerned with the effect of participation in high school athletics on future labor market earnings. For data drawn from the NLSY for the period 1989–1992, the analysis uncovers a positive effect of participation on future earnings and evidence of negative confounding in the treatment group and positive confounding in the control group.

The rest of the paper is organized is follows. Section 2 introduces the modeling framework, together with the techniques for fitting the model, calculating the model marginal likelihood, and for finding treatment effects. Section 3 presents the simulation experiments while Section 4 is concerned with our real data example. Concluding remarks are given in Section 5.

2. Treatment problem with panel outcomes

2.1. Model

Suppose that a given subject i ($i \leqslant n$) in some base period $t_0 = 0$ has the possibility of taking a treatment once, indicated by the binary indicator variable x_i, where $x_i = 1$ indicates intake and $x_i = 0$ indicates non-intake of the treatment. Also suppose that for the observed treatment intake $x_i = j$, some outcome of interest $\mathbf{y}_{j,i} = (y_{j,i1}, \dots, y_{j,iT})$ is measured on each subject over T subsequent time periods, starting at the time point $t_1 = 1$. Under the assumption that the treatment intake at baseline is non-random, the goal is to find the effect of the treatment intake on the outcome of interest, allowing for the complication that the outcome is affected not just by the intake but also by unmeasured or unobserved confounders, even conditioned on covariates.

Our approach to this problem is model-based since non-parametric identification of the treatment effect is not possible in general with unobserved confounders. For the intake we assume the marginal model

$$x_i = I[\mathbf{v}_{i0}'\gamma + \mathbf{z}_{i0}'\boldsymbol{\delta} + u_i > 0], \tag{2.1}$$

where $I[\cdot]$ is the indicator function, $\mathbf{v}_{i0}$ is a l-dimensional vector of covariates in the baseline period (for example demographic variables) that also appear in the outcome model and $\mathbf{z}_{i0}$ is a m-dimensional vector of instruments (regressors that have a direct effect on the intake, have no direct affect on the outcomes, and are uncorrelated with the unobserved confounders), γ and $\boldsymbol{\delta}$ are conformable vectors of parameters, and u_i is an unobserved error whose distribution is specified below. To model the outcomes, assume that in treatment state $x_i = j$,

$$y_{j,it} = \mathbf{v}'_{1,it}\boldsymbol{\beta}_{j,1} + \mathbf{w}'_{it}\mathbf{c}_{j,i} + \varepsilon_{j,it}, \quad j = 0,1; \;\; t = 1,\ldots,T, \tag{2.2}$$

where $\mathbf{v}_{1,it} : k_1 \times 1$ and $\mathbf{w}_{it} : q \times 1$ are covariate vectors and $\mathbf{c}_{j,i} : q \times 1$ are subject and treatment specific random effects and $\varepsilon_{j,it}$ is the error. Assume also that the random effects $\mathbf{c}_{j,i}$ depend on an additional set of covariates $\mathbf{A}_i : q \times k_2$ through the model

$$\mathbf{c}_{j,i} = \mathbf{A}_i\boldsymbol{\beta}_{j,2} + \mathbf{b}_i, \tag{2.3}$$

$$\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}), \tag{2.4}$$

where the parameters $\boldsymbol{\beta}_{j,2}$ and $\mathbf{D}$ are unknown and $\mathcal{N}_q(\cdot,\cdot)$ denotes the multivariate normal distribution. It may be noted that the number of random effects is restricted for identification reasons; in general q must be less than T. Substituting $\mathbf{c}_{j,i}$ from (2.3) into (2.2), the outcome model can be expressed compactly for all T time periods as

$$\mathbf{y}_{j,i} = \mathbf{V}_i\boldsymbol{\beta}_j + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_{j,i}, \tag{2.5}$$

where

$$\mathbf{V}_i = \begin{pmatrix} \mathbf{v}'_{1,i1} & \mathbf{w}'_{i1}\mathbf{A}_i \\ \vdots & \vdots \\ \mathbf{v}'_{1,iT} & \mathbf{w}'_{iT}\mathbf{A}_i \end{pmatrix},$$

$\mathbf{W}_i = (\mathbf{w}_{i1}, \ldots, \mathbf{w}_{iT})'$, $\boldsymbol{\varepsilon}_{j,i} = (\varepsilon_{j,i1}, \ldots, \varepsilon_{j,iT})$ and $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j,1}, \boldsymbol{\beta}_{j,2}) : k \times 1$ $(k = k_1 + k_2)$.

The marginal models are connected by assuming a joint distribution for the vector of errors $(\boldsymbol{\varepsilon}_{j,i}, u_i)$. Instead of assuming that this joint distribution is Gaussian, which can be rather strong in this context, we assume that the distribution is multivariate-t with fixed (but small) degrees of freedom v. In particular, we assume that

$$(\boldsymbol{\varepsilon}_{j,i}, u_i)|\lambda_i \sim \mathcal{N}_{T+1}(\mathbf{0}, \lambda_i^{-1}\boldsymbol{\Omega}_j), \tag{2.6}$$

where $\lambda_i \sim \mathcal{G}\left(\frac{v}{2}, \frac{v}{2}\right)$, so that marginally of λ_i, the joint distribution of $(\boldsymbol{\varepsilon}_{j,i}, u_i)$ is $\mathcal{T}_{T+1}(\mathbf{0}, \boldsymbol{\Omega}_j, v)$. The dispersion matrix $\boldsymbol{\Omega}_j$ is of the form

$$\boldsymbol{\Omega}_j = \begin{pmatrix} \sigma^2_{j,1} & 0 & \cdots & \omega_1 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \cdots & \sigma^2_{j,T} & \omega_T \\ \omega_1 & \cdots & \omega_T & 1 \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} \boldsymbol{\Sigma}_j & \boldsymbol{\omega}_j \\ \boldsymbol{\omega}'_j & 1 \end{pmatrix}, \tag{2.7}$$

where $\boldsymbol{\omega}_j = (\omega_{j,1}, \omega_{j,2}, \ldots, \omega_{j,T})' : T \times 1$ are the parameters that capture the confounding on unobservables in each treatment state j. The $\boldsymbol{\omega}_j$'s are important objects of interest. The covariance matrix $\boldsymbol{\Sigma}_j : T \times T$ of the outcomes is diagonal, with elements $\boldsymbol{\sigma}^2_j = (\sigma^2_{j,1}, \ldots, \sigma^2_{j,T})$, because under the maintained assumption that the matrix $\mathbf{D}$ is full, the off-diagonal elements of $\boldsymbol{\Sigma}_j$ are unidentified and redundant. In some cases a simpler form of $\boldsymbol{\Omega}_j$ may be adequate. For example, the matrix $\boldsymbol{\Sigma}_j$ could be parameterized as $\sigma^2\mathbf{I}_T$ in terms of a scalar variance parameter σ^2 and the identity matrix $\mathbf{I}_T$, and the covariances $\omega_{j,t}$ as $\omega_{j,t} = \omega_j$, constant across time. These different modeling assumptions about $\boldsymbol{\Omega}_j$, which lead to what we call the unrestricted and restricted models, can be compared through the computation of marginal likelihoods and Bayes factors as we discuss below.

We stress that in the above we make no assumptions about the unidentified joint distribution of $\mathbf{y}_{0,i}$ and $\mathbf{y}_{1,i}$; the distribution is unidentified because the outcomes $\mathbf{y}_{0,i}$ and $\mathbf{y}_{1,i}$ cannot be observed simultaneously. If alternatively we had insisted on assuming that the joint distribution is of a certain form (say multivariate-t) then this would have introduced $T(T+1)/2$ unidentified covariances into the problem. Another drawback would have been that one would also have had to involve the T dimensional unobserved counterfactuals for each subject. These twin problems are finessed in our modeling. That the modeling and subsequent estimation of such models can be conducted in this way without the joint distribution of the potential outcomes is due to Chib (2004).

2.2. A reparameterization

In anticipation of the estimation procedure we develop in the sequel, we note that the parameterization of $\mathbf{\Omega}_j$ by means of $\boldsymbol{\sigma}_j^2$ and $\boldsymbol{\omega}_j$ is not convenient because the required positive-definiteness constraint dictates that these parameters must be restricted to the region $\{\boldsymbol{\sigma}_j^2, \boldsymbol{\omega}_j : |\mathbf{\Omega}_j| > 0\}$. This complicates both the formulation of a prior distribution and the development of an efficient estimation procedure. It is more natural to work with the parameters that appear in the cholesky decomposition of $\mathbf{\Omega}_j$. From a simple calculation we see that $\mathbf{\Omega}_j = \mathbf{L}_j\mathbf{L}_j'$ where $\mathbf{L}_j$ is the lower-triangular matrix

$$\mathbf{L}_j = \begin{pmatrix} \sigma_{j,1} & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \cdots & \sigma_{j,T} & 0 \\ \rho_{j,1} & \cdots & \rho_{j,T} & (1-\sum \rho_{j,t}^2)^{1/2} \end{pmatrix} = \begin{pmatrix} e^{\ln \sigma_{j,1}} & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \cdots & e^{\ln \sigma_{j,T}} & 0 \\ \rho_{j,1} & \cdots & \rho_{j,T} & (1-\sum \rho_{j,t}^2)^{1/2} \end{pmatrix}$$

and $\rho_{j,t} = \omega_{j,t}/\sigma_{j,t}$ is the correlation between $\varepsilon_{j,it}$ and u_i. We can now define the parameters of $\mathbf{\Omega}_j$ with

$$\zeta_j = (\ln \sigma_{j,1}, \ln \sigma_{j,2}, \ldots, \ln \sigma_{j,T}, \rho_{j,1}, \ldots, \rho_{j,T}) \overset{\text{def}}{=} (\ln \boldsymbol{\sigma}_j, \boldsymbol{\rho}_j),$$

the log standard-deviations and the correlations. The positive-definiteness of $\mathbf{L}_j$, and hence of $\mathbf{\Omega}_j$, is achieved by restricting $\boldsymbol{\rho}_j$ to the interior of the T-dimensional unit hyper-sphere, namely to the region $S_j = \{\boldsymbol{\rho}_j : \sum \rho_{j,t}^2 < 1\}$. This is a simple constraint to impose. Of course, the elements of $\ln \boldsymbol{\sigma}_j$ are unrestricted.

2.3. Prior

Our approach to inference is Bayesian so we complete the model specification by defining the prior distribution. Let

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \zeta_0, \zeta_1, \mathbf{D})$$

denote the model parameters where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \gamma, \boldsymbol{\delta})$ is of dimension $p = 2k + l + m$, $\zeta_0 = (\ln \boldsymbol{\sigma}_0, \boldsymbol{\rho}_0)$ and $\zeta_1 = (\ln \boldsymbol{\sigma}_1, \boldsymbol{\rho}_1)$. Following common practice, we specify a multivariate normal $\mathcal{N}_p(\boldsymbol{\beta}|\boldsymbol{b}_0, \mathbf{B}_0)$ prior for $\boldsymbol{\beta}$ and a Wishart $\mathcal{W}_q(\mathbf{D}^{-1}|v_0, \mathbf{R}_0)$ prior for $\mathbf{D}^{-1}$. Further, we assume that $\ln \boldsymbol{\sigma}_j$ has a T-variate normal prior distribution $\mathcal{N}_T(\ln \boldsymbol{\sigma}_j|\mathbf{c}_{j0}, \mathbf{C}_{j0})$ and independently that $\boldsymbol{\rho}_j$ is truncated T-variate normal which we denote as $\mathcal{TN}_T(\boldsymbol{\rho}_j|\mathbf{m}_{j0}, \mathbf{M}_{j0})I[S_j]$. The normalizing constant of the latter distribution, which is not needed in the estimation, can be found by simulation, by drawing a large number of variates $\boldsymbol{\rho}_j$ from $\mathcal{N}_T(\boldsymbol{\rho}_j|\mathbf{m}_{j0}, \mathbf{M}_{j0})$ and counting the proportion that lie in S_j. Assuming that $\boldsymbol{\zeta}_0$ and $\boldsymbol{\zeta}_1$ are a priori independent, our prior distribution of $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}) = \mathcal{N}_p(\boldsymbol{\beta}|\boldsymbol{b}_0, \mathbf{B}_0)\mathcal{W}_q(\mathbf{D}^{-1}|v_0, \mathbf{R}_0)\prod_{j=0}^{1} \mathcal{N}_T(\ln \boldsymbol{\sigma}_j|\mathbf{c}_{j0}, \mathbf{C}_{j0})\mathcal{TN}_T(\boldsymbol{\rho}_j|\mathbf{m}_{j0}, \mathbf{M}_{j0})I[S_j]. \tag{2.8}$$

2.4. Likelihood function

Assume that the outcomes across individuals are distributed independently and let N_j denote the set of subjects in treatment state j, i.e., $N_j = \{i : x_i = j\}$. Also let $\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1)$ where $\mathbf{y}_0 = (\mathbf{y}_{0,i} : i \in N_0)$ and $\mathbf{y}_1 = (\mathbf{y}_{1,i} : i \in N_1)$ represent the observations on the controls and the treated, respectively; with a similar convention for the intake $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1)$. Let $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{D})$ denote the likelihood function. Then, the likelihood function is composed of two distinct terms, one from the $x_i = 0$ observations and the other from the $x_i = 1$ observations. Specifically, it has the form

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{D}) = \prod_{i \in N_0} p_0(\mathbf{y}_{0,i}, x_i = 0|\boldsymbol{\beta}, \zeta_0, \mathbf{D}) \prod_{i \in N_1} p_1(\mathbf{y}_{1,i}, x_i = 1|\boldsymbol{\beta}, \zeta_1, \mathbf{D}), \tag{2.9}$$

where $p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}, \zeta_j, \mathbf{D})$ is the contribution of the ith observation in treatment state j. To derive this contribution we introduce the latent treatment variable x_i^* such that $x_i = I\{x_i^* > 0\}$ and let $\mathbf{y}_{j,i}^* = (\mathbf{y}_{j,i}, x_i^*)$. Then, from our assumptions in Section 2.1 it follows that the generating model for $\mathbf{y}_{j,i}^*$ is

$$\underbrace{\begin{pmatrix} \mathbf{y}_{j,i} \\ x_i^* \end{pmatrix}}_{\mathbf{y}_{j,i}^*} = \underbrace{\begin{pmatrix} \mathbf{V}_i(1 - x_i) & \mathbf{V}_i x_i & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & \mathbf{v}_{i0}' & \mathbf{z}_{i0}' \end{pmatrix}}_{\mathbf{V}_{j,i}} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \gamma \\ \boldsymbol{\delta} \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \mathbf{W}_i \\ 0 \end{pmatrix}}_{\mathbf{W}_{j,i}} \mathbf{b}_i + \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_{j,i} \\ u_i \end{pmatrix}}_{\mathbf{v}_{j,i}}, \tag{2.10}$$

where $\mathbf{v}_{j,i}|\lambda_i \sim \mathcal{N}_T(\mathbf{0}, \lambda_i^{-1}\boldsymbol{\Omega}_j)$. In other words, the joint density of $\mathbf{y}_{j,i}^*$ given $(\boldsymbol{\beta}, \zeta_j, \mathbf{b}_i, \lambda_i)$ is

$$p_j(\mathbf{y}_{j,i}, x_i^*|\boldsymbol{\beta}, \zeta_j, \mathbf{b}_i, \lambda_i) = \mathcal{N}_{T+1}(\mathbf{y}_{j,i}, x_i^*|\mathbf{V}_{j,i}\boldsymbol{\beta} + \mathbf{W}_{j,i}\mathbf{b}_i, \lambda_i^{-1}\boldsymbol{\Omega}_j) \tag{2.11}$$

and the contribution of interest is

$$p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}, \zeta_j, \mathbf{D}) = \int_0^\infty \int_{A_j} \int_{\Re^q} p_j(\mathbf{y}_{j,i}, x_i^*|\boldsymbol{\beta}, \zeta_j, \mathbf{b}_i, \lambda_i)\mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}) \times \mathcal{G}\left(\lambda_i \middle| \frac{v}{2}, \frac{v}{2}\right) \mathrm{d}\mathbf{b}_i \, \mathrm{d}x_i^* \, \mathrm{d}\lambda_i,$$

where A_j is the set $(-\infty, 0)$ if $j = 0$ or $(0, \infty)$ if $j = 1$. If we integrate out λ_i first the integrand for the remaining two integrals is

$$p_j(\mathbf{y}_{j,i}, x_i^*|\boldsymbol{\beta}, \zeta_j, \mathbf{b}_i) = \mathcal{T}_{T+1}(\mathbf{y}_{j,i}, x_i^*|\mathbf{V}_{j,i}\boldsymbol{\beta} + \mathbf{W}_{j,i}\mathbf{b}_i, \boldsymbol{\Omega}_j, v) \tag{2.12}$$

from where x_i^* can be integrated out analytically to give

$$p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}, \zeta_j, \mathbf{D}, \mathbf{b}_i) = \mathscr{T}_v(\mathbf{y}_{j,i}|\mathbf{V}_i\boldsymbol{\beta}_j + \mathbf{W}_i\mathbf{b}_i, \boldsymbol{\Sigma}_j)\mathbf{T}((2j-1)\hat{m}_{j,i}(\hat{h}^2_{j,i}\hat{\eta}^2_{j,i})^{-1/2}, v + T), \tag{2.13}$$

where $\mathbf{T}$ is the cdf of the standard student-t density with $v + T$ degrees of freedom, $\hat{m}_{j,i} = \mathbf{v}'_{i0}\gamma + z'_{i0}\boldsymbol{\delta} + \boldsymbol{\omega}'_j\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_{j,i} - \mathbf{V}_i\boldsymbol{\beta}_j - \mathbf{W}_i\mathbf{b}_i)$,
$\hat{h}^2_{j,i} = v/v + T[1 + (\mathbf{y}_{j,i} - \mathbf{V}_i\boldsymbol{\beta}_j - \mathbf{W}_i\mathbf{b}_i)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_{j,i} - \mathbf{V}_i\boldsymbol{\beta}_j - \mathbf{W}_i\mathbf{b}_i)/v)]$, and $\hat{\eta}^2_{j,i} = 1 - \boldsymbol{\omega}'_j\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\omega}_j$. This leaves a final integral over $\mathbf{b}_i$ that must be computed numerically.

We can avoid the latter q-dimensional numerical integration by first integrating out $\mathbf{b}_i$ which gives

$$\begin{aligned} p_j(\mathbf{y}_{j,i}, x_i^*|\boldsymbol{\beta}, \zeta_j, \mathbf{D}, \lambda_i) &= \int_{\Re^q} p_j(\mathbf{y}_{j,i}, x_i^*|\boldsymbol{\beta}, \zeta_j, \mathbf{b}_i, \lambda_i)\mathscr{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D})\,\mathrm{d}\mathbf{b}_i \\ &= \mathscr{N}_{T+1}(\mathbf{y}_{j,i}, x_i^*|\mathbf{V}_{j,i}\boldsymbol{\beta}, \boldsymbol{\Lambda}_{j,i}), \end{aligned} \tag{2.14}$$

where

$$\boldsymbol{\Lambda}_{j,i} = \begin{pmatrix} \overbrace{\lambda_i^{-1}\boldsymbol{\Sigma}_j + \mathbf{W}_i\mathbf{D}\mathbf{W}'_i}^{\boldsymbol{\Lambda}^\dagger_{j,i}} & \lambda_i^{-1}\boldsymbol{\omega}_j \\ \lambda_i^{-1}\boldsymbol{\omega}'_j & \lambda_i^{-1} \end{pmatrix}. \tag{2.15}$$

We then integrate over x_i^* to get

$$\begin{aligned} p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}, \zeta_j, \mathbf{D}, \lambda_i) &= \int_{A_j} p_j(\mathbf{y}_{j,i}, x_i^*|\boldsymbol{\beta}, \zeta_j, \mathbf{D}, \lambda_i)\,\mathrm{d}x_i^* \\ &= p_j(\mathbf{y}_{j,i}|\boldsymbol{\beta}_j, \boldsymbol{\sigma}_j^2, \mathbf{D}, \lambda_i)\int_{A_j} p_j(x_i^*|\mathbf{y}_{j,i}, \boldsymbol{\beta}, \zeta_j, \mathbf{D}, \lambda_i)\,\mathrm{d}x_i^* \\ &= \mathscr{N}(\mathbf{y}_{j,i}|\mathbf{V}_i\boldsymbol{\beta}_j, \boldsymbol{\Lambda}^\dagger_{j,i})\Phi((2j-1)m_{j,i}(\eta^2_{j,i})^{-1/2}), \end{aligned} \tag{2.16}$$

where Φ is the cdf of the standard normal density, $m_{j,i} = \mathbf{v}'_{i0}\gamma + z'_{i0}\boldsymbol{\delta} + \lambda_i^{-1}\boldsymbol{\omega}'_j\boldsymbol{\Lambda}^{\dagger-1}_{j,i}$ $(\mathbf{y}_{j,i} - \mathbf{V}_i\boldsymbol{\beta}_j)$, and $\eta^2_{j,i} = \lambda_i^{-1}(1 - \lambda_i^{-1}\boldsymbol{\omega}'_j\boldsymbol{\Lambda}^{\dagger-1}_{j,i}\boldsymbol{\omega}_j)$. The final integral, that over λ_i,

$$p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}, \zeta_j, \mathbf{D}) = \int p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}, \zeta_j, \mathbf{D}, \lambda_i)\mathscr{G}\left(\lambda_i\middle|\frac{v}{2}, \frac{v}{2}\right)\mathrm{d}\lambda_i, \quad i \in N_j \tag{2.17}$$

is now one-dimensional.

It follows that the integral in (2.17) is not in closed form. In Section 4 below, where we consider the problem of model choice, we compute it numerically by the method of importance sampling at one particular value of $\boldsymbol{\theta}$. The importance sampling option, however, is not convenient while estimating $\boldsymbol{\theta}$ because then it must be applied repeatedly for each subject and every new value of $\boldsymbol{\theta}$ in the search process. A less intensive approach based on Markov chain Monte Carlo methods is possible which we now describe.

2.5. Posterior distribution and MCMC based fitting

We follow Albert and Chib (1993) and focus on the posterior distribution of $\boldsymbol{\psi} = (\boldsymbol{\theta}, \mathbf{z})$, where $\mathbf{z} = (\mathbf{b}, \mathbf{x}^*, \boldsymbol{\lambda})$ and $\mathbf{b} = \{\mathbf{b}_i\}, \mathbf{x}^* = \{x_i^*\}, \boldsymbol{\lambda} = \{\lambda_i\}$. From Bayes theorem the joint distribution of interest, $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, \mathbf{x})$, is proportional to $p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \mathbf{z})$ where

$$p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \mathbf{z}) = \pi(\boldsymbol{\theta}) \prod_{i \in N_0} \mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}) \mathcal{G}\left(\lambda_i \middle| \frac{v}{2}, \frac{v}{2}\right) p_0(\mathbf{y}_{0,i}, x_i^* | \boldsymbol{\beta}, \zeta_0, \mathbf{b}_i, \lambda_i) I(x_i^* < 0) \times \prod_{i \in N_1} \mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}) \mathcal{G}\left(\lambda_i \middle| \frac{v}{2}, \frac{v}{2}\right) p_1(\mathbf{y}_{1,i}, x_i^* | \boldsymbol{\beta}, \zeta_1, \mathbf{b}_i, \lambda_i) I(x_i^* > 0) \tag{2.18}$$

and $p_j(\mathbf{y}_{j,i}, x_i^* | \boldsymbol{\beta}, \boldsymbol{\eta}_j, \mathbf{b}_i, \lambda_i)$ from (2.11) is $\mathcal{N}_{T+1}(\mathbf{y}_{j,i}, x_i^* | \mathbf{V}_{j,i}\boldsymbol{\beta} + \mathbf{W}_{j,i}\mathbf{b}_i, \lambda_i^{-1}\boldsymbol{\Omega}_j)$. This joint distribution is of a type that can be efficiently processed by MCMC methods. Recall that the goal of a MCMC simulation is to obtain draws from the posterior distribution by simulating a suitably constructed Markov chain whose invariant distribution is the posterior distribution of interest (for an extensive discussion of MCMC methods in Bayesian inference see Chib, 2001).

The specific form of the posterior distribution in (2.18) leads to interesting design options, especially in connection with the sampling of $\boldsymbol{\zeta}_j = (\ln \boldsymbol{\sigma}_j, \boldsymbol{\rho}_j)$. For example, the $\boldsymbol{\zeta}_j$'s can be generated marginalized over $(\mathbf{x}^*, \mathbf{b})$ from the conditional density

$$\pi(\boldsymbol{\zeta}_j | \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_{-\zeta_j}, \mathbf{z}_{-(\mathbf{x}^*, \mathbf{b})}) \propto \pi(\boldsymbol{\zeta}_j) \times \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j, \mathbf{D}, \boldsymbol{\lambda}_i).$$

Sampling of this density requires a Metropolis–Hastings (M–H) step with probability of move $\alpha_j = \alpha(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_j' | \mathbf{y}_j, \mathbf{x}_j, \boldsymbol{\theta}_{-\zeta_j}, \mathbf{z}_{-(\mathbf{x}^*, \mathbf{b})})$ given by

$$\alpha_j = \min\left\{1, \frac{\pi(\boldsymbol{\zeta}_j') I_{S_j} \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j', \mathbf{D}, \lambda_i)}{\pi(\boldsymbol{\zeta}_j) I_{S_j} \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j, \mathbf{D}, \lambda_i)}\right\}, \tag{2.19}$$

where $p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j', \mathbf{D}, \lambda_i)$ is given in (2.16), and $q(\boldsymbol{\zeta}_j | \mathbf{y}, \mathbf{x}, \mathbf{D}, \boldsymbol{\beta}, \lambda)$ is the proposal density. Following Chib and Greenberg, 1994, 1995 we can let the proposal density be $\mathcal{T}_{2T}(\boldsymbol{\zeta}_j | \boldsymbol{\mu}_j, \mathbf{V}_j, v_0)$ (for some choice of v_0) such that $\boldsymbol{\mu}_j$ and $\mathbf{V}_j$ are, respectively, the mode and negative inverse Hessian of $\ln \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j, \mathbf{D}, \lambda_i)$. Then, in the remaining steps of the MCMC algorithm, one samples $(\mathbf{x}^*, \boldsymbol{\beta})$ marginalized over $\mathbf{b}$, followed by $\mathbf{b}$, λ_i and $\mathbf{D}^{-1}$ from the full conditional distributions that are derived from (2.18).

Alternatively, one can sample $\boldsymbol{\zeta}_j$ marginalized over $(\mathbf{x}^*, \boldsymbol{\lambda})$, again with a M–H step, from the conditional density

$$\pi(\boldsymbol{\zeta}_j | \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_{-\zeta_j}, \mathbf{z}_{-(\mathbf{x}^*, \lambda)}) \propto \pi(\boldsymbol{\zeta}_j) \times \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j, \mathbf{D}, \mathbf{b}_i),$$

where now the probability of move $\hat{\alpha}_j = \alpha_j(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_j' | \mathbf{y}_j, \mathbf{x}_j, \boldsymbol{\beta}, \mathbf{D}, \mathbf{z}_{-(\mathbf{x}^*, \lambda)})$ is given by

$$\hat{\alpha}_j = \min\left\{1, \frac{\pi(\boldsymbol{\zeta}_j') I_{S_j} \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j', \mathbf{D}, \mathbf{b}_i)}{\pi(\boldsymbol{\zeta}_j) I_{S_j} \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j, \mathbf{D}, \mathbf{b}_i)} \frac{q(\boldsymbol{\zeta}_j | \mathbf{y}, \mathbf{x}, \mathbf{D}, \boldsymbol{\beta}, \mathbf{b})}{q(\boldsymbol{\zeta}_j' | \mathbf{y}, \mathbf{x}, \mathbf{D}, \boldsymbol{\beta}, \mathbf{b})}\right\}, \tag{2.20}$$

and $p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j', \mathbf{D}, \mathbf{b}_i)$ is given in (2.13). As before, the proposal density $q(\boldsymbol{\zeta}_j | \mathbf{y}, \mathbf{x}, \mathbf{D}, \boldsymbol{\beta}, \mathbf{b})$ can be chosen to be $\mathcal{T}_{2T}(\boldsymbol{\zeta}_j | \boldsymbol{\mu}_j, \mathbf{V}_j, v_0)$ where $(\boldsymbol{\mu}_j, \mathbf{V}_j)$ are found as the mode and negative inverse Hessian of $\ln \prod_{i \in N_j} p_j(\mathbf{y}_{j,i}, x_i = j | \boldsymbol{\beta}, \boldsymbol{\zeta}_j, \mathbf{D}, \mathbf{b}_i)$.

We favor the latter approach because the likelihood contribution in (2.13) can be vectorized over subjects whereas the contribution in (2.16) on the other hand cannot, because it involves a subject-specific covariance matrix $\boldsymbol{\Lambda}_{j,i}$. As a result, the second scheme is faster. However, it also tends to produce output that is more serially correlated due to the fact of sampling $\boldsymbol{\zeta}_j$ conditioned on the random effects (Chib and Carlin, 1999).

Nonetheless, the loss of efficiency tends not to be severe. In detail, our suggested sampling approach may be described as follows:

MCMC Algorithm.

```
1. Initialize $\ln \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\beta}, \mathbf{D}, \boldsymbol{\lambda}$
2. Sample $(\ln \boldsymbol{\sigma}, \boldsymbol{\rho}, \mathbf{x}^*, \boldsymbol{\lambda})$ by sampling
   (a) $(\ln \boldsymbol{\sigma}_j, \boldsymbol{\rho}_j)|\mathbf{y}_j, \mathbf{x}_j, \boldsymbol{\theta}_{-\zeta_j}, \mathbf{z}_{-(\mathbf{x}^*,\boldsymbol{\lambda})}$ by a M–H step with probability of move in (2.20)
   (b) $(\mathbf{x}^*, \boldsymbol{\lambda})|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \mathbf{z}_{-(\mathbf{x}^*,\boldsymbol{\lambda})}$ by sampling
       (i) $x_i^*|\mathbf{y}_i, x_i, \boldsymbol{\theta}, \mathbf{z}_{-\mathbf{x}^*}$, a $\mathcal{N}(\hat{m}_{j,i}, \hat{\eta}_{j,i}^2)$ distribution truncated to the interval $(-\infty, 0)$ if
           $x_i = 0$, and $(0, \infty)$ if $x_i = 1$
       (ii) $\lambda_i|\mathbf{y}_i, x_i, \boldsymbol{\theta}, \mathbf{z}_{-\boldsymbol{\lambda}} \sim \mathcal{G}\left(\frac{v+T+1}{2}, \frac{v+\boldsymbol{\varepsilon}_{j,i}'\boldsymbol{\Omega}_j^{-1}\boldsymbol{\varepsilon}_{j,i}}{2}\right)$, where $\boldsymbol{\varepsilon}_{j,i} = (\mathbf{y}_{j,i}^* - \mathbf{V}_{j,i}\boldsymbol{\beta} - \mathbf{W}_{j,i}\mathbf{b}_i)$
3. Sample $(\boldsymbol{\beta}, \mathbf{b})$ by sampling
   (a) $\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_{-\boldsymbol{\beta}}, \mathbf{z}_{-\mathbf{b}} \sim \mathcal{N}_p(\mathbf{B}\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{B}\sum_{i=1}^n \mathbf{V}_{j,i}'\boldsymbol{\Lambda}_{j,i}^{-1}\mathbf{y}_{j,i}^*, \mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{V}_{j,i}'\boldsymbol{\Lambda}_{j,i}^{-1}\mathbf{V}_{j,i}))$.
   (b) $\mathbf{b}_i|\mathbf{y}_i, x_i, \boldsymbol{\theta}, \mathbf{z}_{-\mathbf{b}} \sim \mathcal{N}_q(\mathbf{B}_i[\lambda_i\mathbf{W}_{j,i}'\boldsymbol{\Omega}_j^{-1}(\mathbf{y}_{j,i}^* - \mathbf{V}_{j,i}\boldsymbol{\beta})], \mathbf{B}_i = (\mathbf{D}^{-1} + \lambda_i\mathbf{W}_{j,i}'\boldsymbol{\Omega}_j^{-1}\mathbf{W}_{j,i}))$.
4. Sample $\mathbf{D}^{-1}|\mathbf{z}_{-(\mathbf{x}^*,\boldsymbol{\lambda})} \sim \mathcal{W}_q(v_0 + n, [\mathbf{R}_0^{-1} + \sum_{i=1}^n \mathbf{b}_i\mathbf{b}_i']^{-1})$.
5. Goto 2.
```

It may be noted that the model in which $\boldsymbol{\Omega}_j$ is restricted (the model where $\boldsymbol{\Sigma}_j$ is parameterized as $\sigma^2\mathbf{I}_T$ and the covariances $\omega_{j,t}$ are constant across time) is fit in the same way as above. Of course, because $\boldsymbol{\zeta}_j$ is now two-dimensional, the M–H step in 2a becomes even simpler to implement.

2.6. Treatment effects

In the potential outcomes framework, the individual-level treatment effect is defined as the difference between the potential outcomes $(y_{1,i} - y_{0,i})$, but this quantity is not observable. It is known that under certain assumptions, primarily random assignment of treatment, the average treatment effect (ATE), which is the difference between the average value of the outcome if everyone were (hypothetically) given the treatment, and the average value of the outcome if everyone were (hypothetically) not given the treatment, is identified non-parametrically. In the case of non-random treatment intake and confounding on unobservables, even the ATE is, in general, not identified. However, under our parametric assumptions the ATE can be found. Instead of focusing on the ATE we prefer the predictive ATE, which we define as

$$\mathrm{E}(\mathbf{y}_{1,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{V}, \mathbf{z}) - \mathrm{E}(\mathbf{y}_{0,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{V}, \mathbf{z}),$$

where $(n + 1)$ refers to a new subject drawn from the population and $(\mathbf{y}, \mathbf{x}, \mathbf{V}, \mathbf{z})$ denotes the sample data. The quantities $\mathrm{E}(\mathbf{y}_{j,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{V}, \mathbf{z})$, $j = 0, 1$, which are conditional on the observed data, can be readily calculated as a by-product of our MCMC simulation. In fact, we can calculate the entire predictive distribution of the potential outcomes, and use these distributions to find the difference in expectations and the difference in, for example, the predictive quantiles. These various prediction-based treatment effects can be quite informative.

By definition, the Bayesian predictive distribution of each potential outcome sequence is given by

$$\begin{aligned} p(\mathbf{y}_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z}) = & \int p_j(\mathbf{y}_{j,n+1}|\mathbf{V}_{n+1},\boldsymbol{\beta},\boldsymbol{\sigma}_j^2,\mathbf{D},\mathbf{b}_{n+1},\lambda_{n+1}) \\ & \times \pi(\boldsymbol{\beta},\boldsymbol{\sigma}_j^2,\mathbf{D},\mathbf{b}_{n+1},\lambda_{n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z}) \\ & \times p(\mathbf{V}_{n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z})\,\mathrm{d}\mathbf{V}_{n+1}\,\mathrm{d}\boldsymbol{\beta}_j\,\mathrm{d}\boldsymbol{\sigma}_j^2\,\mathrm{d}\mathbf{D}\,\mathrm{d}\mathbf{b}_{n+1}\,\mathrm{d}\lambda_{n+1} \end{aligned}$$

without involvement of the treatment intake model. As in any Bayesian predictive calculation, the different unknowns of the model are marginalized out of the sampling density of $\mathbf{y}_{j,n+1}$ given the sample data. In particular, $(\boldsymbol{\beta},\boldsymbol{\sigma}_j^2,\mathbf{D},\mathbf{b}_{n+1},\lambda_{n+1})$ are marginalized with respect to their posterior distributions. We also marginalize out $\mathbf{V}_{n+1}$ but with respect to its empirical distribution. Note that because subject $(n+1)$ is randomly drawn from the population, $p_j(\mathbf{y}_{j,n+1}|\mathbf{V}_{n+1},\boldsymbol{\beta},\boldsymbol{\sigma}_j^2,\mathbf{D},\mathbf{b}_{n+1},\lambda_{n+1})$ does not depend on the sample data.

As discussed in Chib (2004), we obtain these predictive distributions by the method of composition, appending the following steps at the end of each cycle of the MCMC algorithm:

- sample $\mathbf{V}_{n+1}^{(g)}$ by assigning probability of $1/n$ to each row of $\mathbf{V}$;
- sample $\mathbf{b}_{n+1}^{(g)}$ from $\mathcal{N}(0,\mathbf{D}^{(g)})$;
- sample $\lambda_{n+1}^{(g)}$ from $\mathcal{G}\left(\frac{\nu}{2},\frac{\nu}{2}\right)$;
- sample $\mathbf{y}_{j,n+1}^{(g)}$ from $\mathcal{N}_T(\mathbf{V}_{n+1}\boldsymbol{\beta}_j^{(g)}+\mathbf{W}_{n+1}\mathbf{b}_{n+1}^{(g)},\boldsymbol{\Sigma}_j^{(g)}/\lambda_{n+1}^{(g)})$ for $j=0,1$.

The output from these steps, $\{\mathbf{y}_{j,n+1}^{(1)},\ldots,\mathbf{y}_{j,n+1}^{(M)}\}$, are draws from $p(\mathbf{y}_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z})$. Given these draws, the expected values of $\mathbf{y}_{j,n+1}$, can be computed in the usual manner as a sample average of the draws. The predictive quantiles and other summaries can also be obtained from these draws.

It is possible to also calculate $\mathrm{E}(\mathbf{y}_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z})$ by the law of the iterated expectation, averaging $\mathrm{E}(\mathbf{y}_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z},\mathbf{V}_{n+1},\boldsymbol{\theta})=\mathbf{V}_{n+1}\boldsymbol{\beta}_j$ over the distribution of $\{\mathbf{V}_{n+1},\boldsymbol{\beta}_j\}$ given $(\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z})$. This approach can be used when the response variable is in logarithms, as in the application presented in Section 5 below. In this case, the conditional expectation of the response on the original scale, marginalized over $\mathbf{b}_i$, is given by

$$\exp\{\mathbf{V}_{n+1}\boldsymbol{\beta}_j+0.5\,\mathrm{diag}(\lambda_{n+1}^{-1}\boldsymbol{\Omega}_j^{\dagger}+\mathbf{W}_{n+1}\mathbf{D}\mathbf{W}_{n+1}')\},$$

which can be averaged over the draws of $\{\mathbf{V}_{n+1},\boldsymbol{\beta}_j,\lambda_{n+1}\}$ given $(\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z})$. In practice, this conditional expectation is evaluated at the end of each MCMC iteration, given the current draws of the parameters, and $\mathbf{V}_{n+1}$ and λ_{n+1} are drawn as above. By the Rao–Blackwell theorem the simulation variance of this estimate of $\mathrm{E}(\mathbf{y}_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{V},\mathbf{z})$ can be expected to be lower than that of sample average of the draws.

3. Simulation studies

3.1. Generating processes

We illustrate the performance of our inferential techniques with artificial data sets that are generated from a model with two covariates, two random effects, a student-t joint distribution with $v = 10$ degrees of freedom, $T = 4$ panel periods and sample sizes of $n = 500$, 1000, and 1500 subjects. One of the covariates is continuous and the other is binary. We also assume that the single instrument is continuous. In detail, the generating model is given by

$$x_i = I[-1 + 0.6\,v_{1,i1} + 1\,v_{2,i1} + 1.5\,z_i + u_i > 0],$$
$$y_{j,it} = (2 + j) + (3 + j)\,v_{1,it} + 1.5\,v_{2,it} + b_{1,i} + b_{2,i}\,v_{1,it} + \varepsilon_{j,it}, \quad t \leqslant 4;\; i \leqslant n,$$

where $v_{1,it}$ (the continuous covariate whose effect is heterogenous) is $\mathcal{N}(0.5 + 0.5t, 0.8)$, $v_{2,it}$ is a binary covariate with mean 0.5 and the instrument z_i is drawn as $\mathcal{N}(0, 1)$. The random effects $b_{1,i}$ and $b_{2,i}$ are drawn independently from zero mean normal distributions with variances of 1.5 and 1, respectively. The matrices $\mathbf{\Omega}_j$ are specified in restricted form with $\boldsymbol{\sigma}_j^2 = (1.5, 1.5, 1.5, 1.5)$ and $\boldsymbol{\omega}_j = (0.49, 0.49, 0.49, 0.49)$, which implies that the pairwise correlation coefficients between the intake and outcome errors are $\boldsymbol{\rho}_j = (0.4, 0.4, 0.4, 0.4)$. Under this design, approximately half of the subjects participate in the treatment. Finally, based on the data that is generated in this simulation, we are able to calculate the kernel-smoothed distribution of each potential outcome, and the true average treatment effects, across the four time periods.

3.2. Estimation results

For these three data sets, we fit models with $\mathbf{\Omega}_j$ both unrestricted and restricted (the form under the true generating process). From the former model we are able to see if parameter estimates of $\boldsymbol{\sigma}_j^2$ and $\boldsymbol{\omega}_j$ concentrate on the single true values. For our prior, we let $\boldsymbol{b}_0 = 0\mathbf{i}_{10}$, $\mathbf{B}_0 = 2\mathbf{I}_{10}$, $\mathbf{m}_{j0} = 0\mathbf{i}_T$ and $\mathbf{M}_{j0} = 2\mathbf{I}_T$. Moreover, we let the prior mean and variance of $\ln \sigma_{j,t}$ be 1 and let $\mathrm{E}(\mathbf{D}) = \mathrm{diag}(1, 1)$ with $v_0 = 6$ degrees of freedom. Our results are based on 20,000 MCMC iterations (beyond a burn-in of a 1000 cycles).

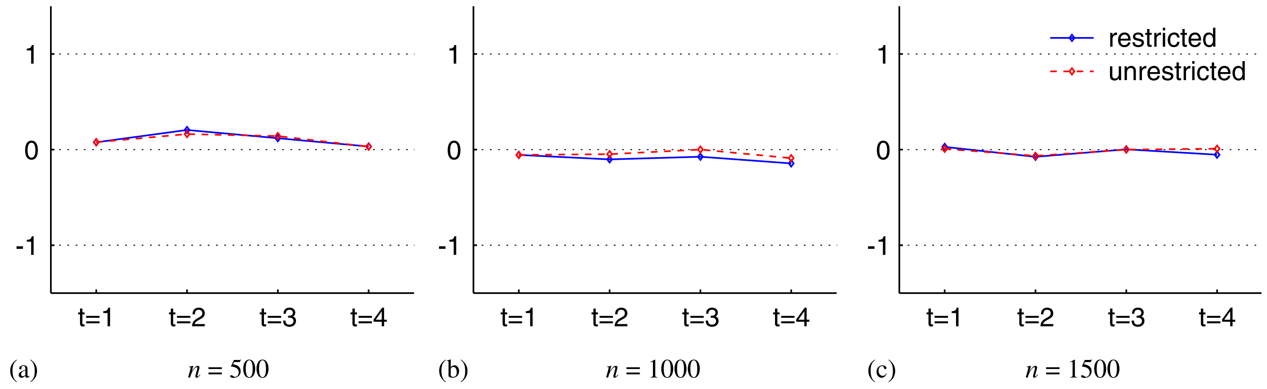
In Table 1 we give the posterior means and standard deviations of $\boldsymbol{\beta}_j$ and the correlations $\boldsymbol{\rho}_j$ from the fitting of each model. The true values of the parameters are given in the first column. It should be noted that in the case of the restricted model, $\boldsymbol{\rho}_j$ is a scalar and therefore six rows in columns 3–5 are empty. It can be seen from this table that the true parameters are well estimated even from the (incorrect) unrestricted model.

The performance of the fitting algorithm can also be examined in terms of the inferences about the treatment effects. We first consider the estimation of the ATEs, comparing the estimates we get from the model fitting with the true values (calculated when the data was generated). In Fig. 1 we plot the difference between the predictive ATE and the true ATE across the four time periods; the solid line gives the results from the restricted model whereas the dashed line gives those from the unrestricted model. We see that the estimates from the two models are essentially identical and close to the true value, even for the smallest sample size.

The predictive ATE is calculated from the predictive distribution of the potential outcomes. It is worthwhile, therefore, to examine these predictive distributions for each t and each j and to compare them with the true distribution of the potential outcomes. This comparison is reported in Fig. 2. For simplicity we only report the results from the fitting of the restricted model for the sample size of $n = 1000$. It can be seen from these plots that the estimated predictive densities are quite close to the true density of the potential outcomes.

Table 1
Posterior means and standard deviations (in parentheses) for the coefficient vectors $\boldsymbol{\beta}_j$ and the correlation vectors $\boldsymbol{\rho}_j$ from the fitting of the panel model to the simulated data

	True	Restricted			Unrestricted		
		$n = 500$	$n = 1000$	$n = 1500$	$n = 500$	$n = 1000$	$n = 1500$
$\boldsymbol{\beta}_0$	2.00	2.13 (0.13)	2.12 (0.10)	1.97 (0.08)	2.12 (0.13)	2.11 (0.10)	1.97 (0.08)
	3.00	3.00 (0.08)	2.98 (0.05)	3.02 (0.05)	3.00 (0.08)	2.97 (0.06)	3.02 (0.05)
	1.50	1.09 (0.18)	1.35 (0.14)	1.51 (0.12)	1.10 (0.18)	1.34 (0.14)	1.51 (0.12)
$\boldsymbol{\beta}_1$	3.0	2.99 (0.21)	2.87 (0.13)	3.05 (0.10)	2.98 (0.21)	2.85 (0.14)	3.04 (0.11)
	4.0	3.97 (0.09)	3.99 (0.06)	4.02 (0.05)	3.98 (0.10)	4.02 (0.06)	4.04 (0.05)
	2.0	2.08 (0.21)	2.12 (0.15)	1.81 (0.12)	2.10 (0.21)	2.13 (0.15)	1.81 (0.12)
$\boldsymbol{\rho}_0$	0.40	0.31 (0.12)	0.40 (0.08)	0.38 (0.07)	0.27 (0.15)	0.30 (0.10)	0.39 (0.08)
	0.40	—	—	—	0.31 (0.14)	0.42 (0.10)	0.41 (0.09)
	0.40	—	—	—	0.35 (0.16)	0.46 (0.10)	0.29 (0.10)
	0.40	—	—	—	0.28 (0.16)	0.31 (0.11)	0.42 (0.10)
$\boldsymbol{\rho}_1$	0.40	0.28 (0.14)	0.40 (0.07)	0.43 (0.05)	0.28 (0.16)	0.46 (0.10)	0.41 (0.07)
	0.40	—	—	—	0.22 (0.16)	0.34 (0.10)	0.48 (0.07)
	0.40	—	—	—	0.32 (0.17)	0.45 (0.10)	0.44 (0.08)
	0.40	—	—	—	0.21 (0.18)	0.24 (0.12)	0.35 (0.09)



(a) $n = 500$ (b) $n = 1000$ (c) $n = 1500$

Fig. 1. Difference between estimated and true average treatment effect from the fitting of the two panel models to the various simulated data; the solid line gives the results for the restricted version and the dashed line for the unrestricted version of the model.

3.3. Comparison with cross-section fitting

We conclude our simulation study by considering the results that emerge if the panel data generated above were processed by a sequence of cross-section models, one for each time period. As described in the introduction, this fitting would miss the dependence across outcomes (arising from the heterogenous effects) and could potentially produce an incorrect sequence of treatment effect estimates. We do not go into details about the specifics of the cross-section fitting except to say that the algorithm we use comes from Chib (2004). In addition, the prior distribution we specify is similar to the one given above. We find (as shown in Table 2) that the cross-section fitting produces roughly comparable

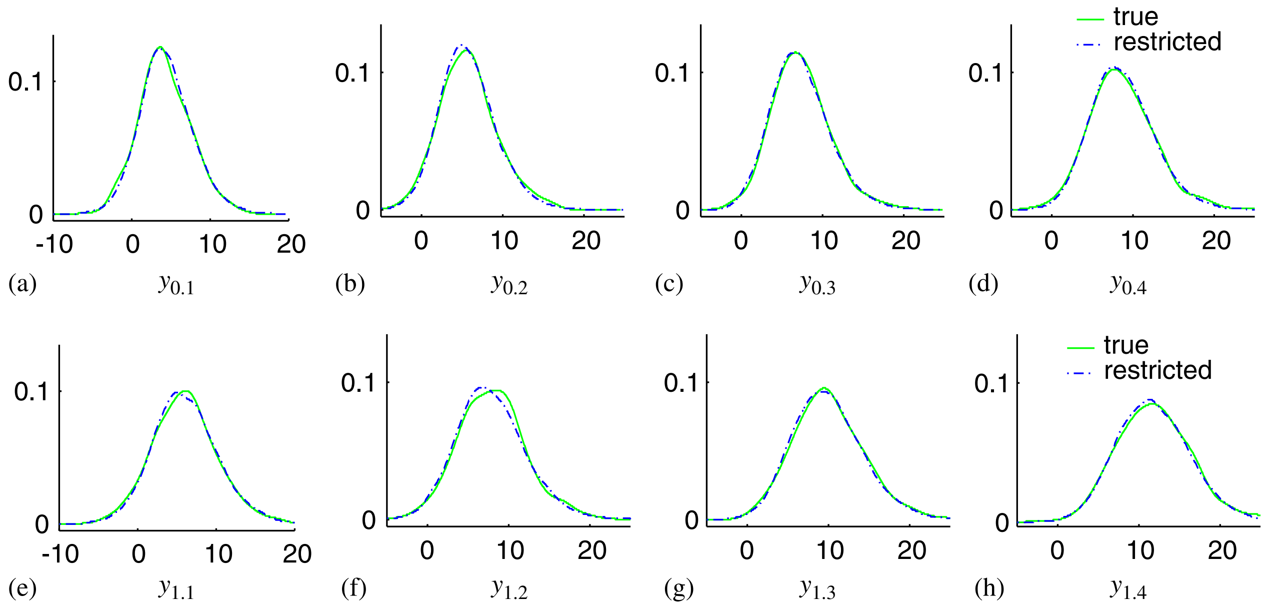
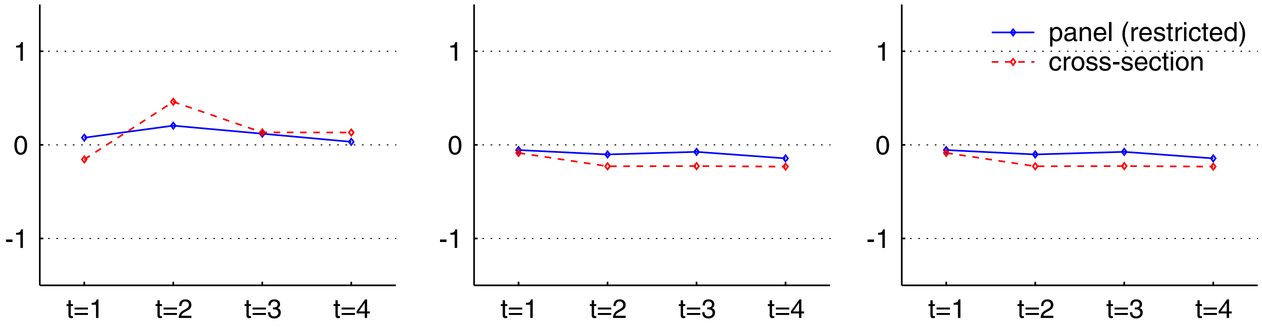


Fig. 2. Estimates of the predictive distributions of the potential outcomes for the four time periods from the fitting of the restricted version of the panel model to the data with $n = 1000$, and the corresponding empirical distributions.

Table 2
Posterior means and standard deviations (in parentheses) for the coefficient vectors $\boldsymbol{\beta}_j$ and the correlation vectors $\boldsymbol{\rho}_j$ from the fitting of a sequence of cross-section models to the simulated data with $n = 1000$

	True	$t = 1$	$t = 2$	$t = 3$	$t = 4$
$\boldsymbol{\beta}_0$	2.00	2.10 (0.13)	2.12 (0.18)	2.05 (0.25)	1.84 (0.31)
	3.00	3.11 (0.11)	2.99 (0.11)	3.05 (0.13)	3.09 (0.13)
	1.50	1.27 (0.18)	1.62 (0.20)	1.31 (0.22)	1.71 (0.24)
$\boldsymbol{\beta}_1$	3.0	2.89 (0.23)	2.89 (0.28)	2.80 (0.33)	2.55 (0.40)
	4.0	3.85 (0.12)	3.97 (0.12)	3.89 (0.14)	4.13 (0.15)
	2.0	2.17 (0.20)	2.07 (0.22)	2.20 (0.25)	2.33 (0.28)
$\boldsymbol{\rho}_0$	0.40	0.23 (0.10)	—	—	—
	0.40	—	0.35 (0.10)	—	—
	0.40	—	—	0.23 (0.10)	—
	0.40	—	—	—	0.25 (0.10)
$\boldsymbol{\rho}_1$	0.40	0.31 (0.11)	—	—	—
	0.40	—	0.18 (0.13)	—	—
	0.40	—	—	0.35 (0.12)	—
	0.40	—	—	—	0.08 (0.12)

estimates of the model parameters, although the confounding parameters are not as well estimated. This leads to less precise estimates of the ATE (as shown in the upper panel in Fig. 3). This is shown in Fig. 3 where we graph the difference between the estimated and the true ATE from the panel and cross-section models. We see that the panel model produces estimates that are closer to the true value for all sample sizes.



(a) $n = 500$ (b) $n = 1000$ (c) $n = 1500$

Fig. 3. Difference between estimated and true average treatment effect from the fitting of the restricted panel model and a sequence of comparable cross-section models; the solid line gives the results for the panel model and the dashed line for the cross-section model.

4. Model comparison

We now discuss an approach for comparing different specifications of our panel data model, for example models specified with a different set of covariates or with a restricted versus an unrestricted covariance matrix, and also for comparing our panel model with alternative models, for example models specified without confounding. We do this comparison through the marginal likelihood of the contending models or equivalently by the pairwise Bayes factors, obtained as ratios of marginal likelihoods. As has been demonstrated in several papers, the method of Chib (1995) makes it possible to find the marginal likelihood with a modest amount of effort.

By definition, the marginal likelihood is given by

$$m(\mathbf{y},\mathbf{x}) = \int p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}.$$

Although direct evaluation of this quantity is not easy, it can be obtained indirectly. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \zeta_0^*, \zeta_1^*, \mathbf{D}^*)$ be a given point in the parameter space, say the posterior mean of $\boldsymbol{\theta}$ estimated from the MCMC output. Following Chib (1995) we then have the basic marginal likelihood identity

$$\ln m(\mathbf{y},\mathbf{x}) = \ln p(\mathbf{y},\mathbf{x}|\boldsymbol{\beta}^*,\zeta^*,\mathbf{D}^*) + \ln \pi(\boldsymbol{\beta}^*,\zeta_0^*,\zeta_1^*,\mathbf{D}^{*-1}) - \ln \pi(\boldsymbol{\beta}^*,\boldsymbol{\eta}_0^*,\zeta_1^*,\mathbf{D}^*|\mathbf{y},\mathbf{x}), \quad (4.1)$$

where $p(\mathbf{y},\mathbf{x}|\boldsymbol{\beta}^*,\zeta^*,\mathbf{D}^*)$ is the likelihood in (2.9), and $\pi(\boldsymbol{\beta}^*,\zeta^*,\mathbf{D}^*)$ and $\pi(\boldsymbol{\beta}^*,\zeta^*,\mathbf{D}^*|\mathbf{y},\mathbf{x})$ are the prior and posterior densities, each evaluated at $\boldsymbol{\theta}^*$. An estimate of the marginal likelihood is found by separately estimating $p(\mathbf{y},\mathbf{x}|\boldsymbol{\beta}^*,\zeta^*,\mathbf{D}^*)$ and $\pi(\boldsymbol{\beta}^*,\zeta^*,\mathbf{D}^{*-1}|\mathbf{y},\mathbf{x})$.

4.1. *Estimation of* $p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta}^*)$

In the discussion surrounding (2.9) it was pointed out that the likelihood is not available in closed form. This is not a complication, however, because the likelihood ordinate must be computed at a single point $\boldsymbol{\theta}^*$. In addition, $p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta},\zeta_j,\mathbf{D},\lambda_i)$, the likelihood contribution of the ith subject conditioned on λ_i, is in closed form. From here λ_i can be marginalized out in several different ways. For example, one convenient approach is by the method of importance sampling. To efficiently apply this method, we alter the range of integration by a change of variable from λ_i to $\lambda_i^* = \ln(\lambda_i)$, so that $\lambda_i = \exp(\lambda_i^*)$ with Jacobian $\exp(\lambda_i^*)$. Now let $h(\lambda_i^*)$ be an importance function, which we specify as a matched student-t density $\mathscr{T}(\lambda_i^*|a_i, b_i, \xi)$, where the parameters a_i and b_i are the (approximate) mode of the function $\ln\{p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}^*, \zeta_j^*, \mathbf{D}^*, \exp(\lambda_i^*))\mathscr{G}\big(\exp(\lambda_i^*)|\frac{\nu}{2}, \frac{\nu}{2}\big)\ \exp(\lambda_i^*)\}$ and the inverse of the observed information of this function evaluated at the mode, respectively, and the degrees of freedom ξ is set arbitrarily at (say) 5. Then, our importance sampling estimate of the likelihood contribution $p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}^*, \zeta_j^*, \mathbf{D}^*)$ is given by

$$G^{-1}\sum_{g=1}^{G}\frac{p_j(\mathbf{y}_{j,i}, x_i = j|\boldsymbol{\beta}^*, \zeta_j^*, \mathbf{D}^*, \exp(\lambda_i^{*(g)}))\ \mathscr{G}\Big(\exp(\lambda_i^{*(g)})|\frac{\nu}{2}, \frac{\nu}{2}\Big)\ \exp(\lambda_i^{*(g)})}{\mathscr{T}(\lambda_i^{*(g)}|a_i, b_i, \xi)},$$

where $\lambda_i^{*(g)}$ is the gth draw from the importance sampling function. The sum of the log of these estimates is ours estimate of the log likelihood ordinate.

4.2. Estimation of $\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x})$

We express the posterior ordinate as

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x}) = \pi(\mathbf{D}^{-1^*}|\mathbf{y}, \mathbf{x})\pi(\zeta_0^*, \zeta_1^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*)\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*, \zeta^*)$$

in which the first term can be estimated from the output of the full MCMC run by averaging the full conditional density of $\mathbf{D}^{-1}$ over the simulated draws. Specifically, letting g index the MCMC iteration, we calculate

$$\hat{\pi}(\mathbf{D}^{-1^*}|\mathbf{y}, \mathbf{x}) = G^{-1}\sum_{g=1}^{G}\mathscr{W}_q\left(\mathbf{D}^{-1^*}|\nu_0 + n, \left[\mathbf{R}_0^{-1} + \sum_{i=1}^{n}\mathbf{b}_i^{(g)}\mathbf{b}_i^{(g)'}\right]^{-1}\right), \tag{4.2}$$

where $\{\mathbf{b}_i^{(g)}\}$ are the draws from the full MCMC run. This is a simulation-consistent estimate of $\pi(\mathbf{D}^{-1^*}|\mathbf{y}, \mathbf{x})$.

Next, to estimate $\pi(\zeta_0^*, \zeta_1^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*)$, we employ the approach of Chib and Jeliazkov (2001). Even though $\boldsymbol{\zeta}$ is sampled in a sequence of M–H steps, the fact that $\boldsymbol{\zeta}_0$ is independent of $\boldsymbol{\zeta}_1$, given $(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}_{-\zeta}, \mathbf{z}_{-(\mathbf{x}^*,\lambda)})$, yields the result that

$$\pi(\zeta_0^*, \zeta_1^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*) = \prod_{j=0}^{1}\frac{\mathbb{E}_1[\alpha_j(\zeta_j, \zeta_j^*|\mathbf{y}_j, \mathbf{x}_j, \boldsymbol{\beta}, \mathbf{D}^*, \mathbf{z}_{-(\mathbf{x}^*,\lambda)})q(\zeta_j^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*, \boldsymbol{\beta}, \mathbf{b})]}{\mathbb{E}_2[\alpha_j(\zeta_j^*, \zeta_j|\mathbf{y}_j, \mathbf{x}_j, \boldsymbol{\beta}, \mathbf{D}^*, \mathbf{z}_{-(\mathbf{x}^*,\lambda)})]}, \tag{4.3}$$

where the expectation $\mathbb{E}_1$ in the numerator is with respect to $\pi(\boldsymbol{\beta}, \zeta_j, \mathbf{b}|\mathbf{y}, \mathbf{x}, \mathbf{D}^*)$ and the expectation $\mathbb{E}_2$ in the denominator is with respect to $\pi(\boldsymbol{\beta}, \mathbf{b}|\mathbf{y}, \mathbf{x}, \zeta_j^*, \mathbf{D}^*)q(\zeta_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \mathbf{D}^*, \mathbf{b})$. Each expectation can be estimated from the output of suitable reduced runs (Chib, 1995). To estimate the numerator, we fix $\mathbf{D}$ at $\mathbf{D}^*$ and continue the MCMC iterations with the quantities $\boldsymbol{\theta}_{-\mathbf{D}}$ and $\mathbf{z}$, and then average $\alpha(\zeta_j, \zeta_j^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \mathbf{D}^*, \mathbf{z}_{-(\mathbf{x}^*,\lambda)})q(\zeta_j^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*, \boldsymbol{\beta}, \mathbf{b})$ over the resulting draws. To estimate the denominator, we fix $(\mathbf{D}, \boldsymbol{\zeta})$ at $(\mathbf{D}^*, \boldsymbol{\zeta}^*)$ and continue the MCMC iterations; in each cycle of this run, we also draw ζ_j from $q(\zeta_j|\mathbf{y}, \mathbf{x}, \mathbf{D}^*, \boldsymbol{\beta}, \mathbf{b})$. We then average $\alpha_j(\zeta_j^*, \zeta_j|\mathbf{y}_j, \mathbf{x}_j, \boldsymbol{\beta}, \mathbf{D}^*, \mathbf{z}_{-(\mathbf{x}^*,\lambda)})$ over the draws on $(\boldsymbol{\beta}, \zeta_0, \zeta_1, \mathbf{b})$ from this run. Simultaneously, from the output of the latter run we estimate $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{D}^*, \boldsymbol{\zeta}^*)$ as

$$G^{-1}\sum_{g=1}^{G}\mathscr{N}_p\left(\boldsymbol{\beta}^*|\mathbf{B}^{(g)}\mathbf{B}_0^{-1}\boldsymbol{b}_0 + \mathbf{B}^{(g)}\sum_{i=1}^{n}\mathbf{V}_{j,i}'\boldsymbol{\Lambda}_{j,i}^{-1^{(g)}}\mathbf{y}_{j,i}^{*(g)}, \mathbf{B}^{(g)}\right), \tag{4.4}$$

where $\mathbf{B}^{(g)} = \{\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{V}'_{j,i}\mathbf{\Lambda}_{j,i}^{-1^{(g)}}\mathbf{V}_{j,i}\}^{-1}$ and

$$\mathbf{\Lambda}_{j,i}^{(g)} = \begin{pmatrix} \lambda_i^{-1^{(g)}}\mathbf{\Sigma}_j^* + \mathbf{W}_i\mathbf{D}^*\mathbf{W}'_i & \lambda_i^{-1^{(g)}}\boldsymbol{\omega}_j^* \\ \lambda_i^{-1^{(g)}}\boldsymbol{\omega}_j^{*'} & \lambda_i^{-1^{(g)}} \end{pmatrix}.$$

Our estimate of the marginal likelihood now appears by substituting (4.1)–(4.4) into (4.1).

5. The effect of high school athletics on earnings

In this section, we apply our modeling framework to examine the relationship between participation in high school athletics and future labor market earnings. Barron et al. (2000) studied this relationship with the help of data from the National Longitudinal Survey of Youth (NLSY). Based on Becker's (1965) allocation-of-time model and human capital theory, the paper identified three unobservable factors that could confound the effect of participation in high school athletics on earnings: low returns to human capital, high capability, and lower preference for leisure. From a cross-section model, and data on weekly earnings of males for 1992, Barron et al. (2000) estimated their model by 2SLS methods but were unable to find a causal link between participation and earnings.

In our reanalysis of this problem we depart by first involving panel data (covering the period 1989–1992) and second by considering data on both males and females. As above, our data is drawn from the NLSY but is restricted to individuals who had graduated from high school no later than 1984, the year when a special survey was conducted to gather information about participation in high school athletics. By that time about 2140 men and 1,940 women in the data base had graduated from high school. After excluding individuals with invalid information on variables of interest such as income, work history, education, demographics and family background, we obtain a sample of 2113 individuals. This raw data reveals that the participation rate in high school athletics is about 50% for male graduates and 40% for female graduates; those who participated tend to come from more educated families and have higher subsequent incomes. The average weekly income ranges from \$383.75 in 1989 to \$468.72 in 1992, with individuals who participated earning around \$90 more in weekly wages than non-participants.

5.1. Model specifications

We borrow on, for example, Card (1999) to specify the outcome and intake models. The outcome $\mathbf{y}_{j,i} : 4 \times 1$ is the log of weekly earnings, while the covariates in $\mathbf{V}_i$ consist of age in years minus 17, educational attainment in years, job tenure in years, indicator of marital status, indicator of male, indicator of black, indicator variables for (i) parent's high school graduation status; (ii) some years of college; (iii) college graduation status and (iv) graduate degree status, followed by indicator variables of the years 1990–1992. To model dependence in the outcomes we assume that the constant term is heterogenous across subjects. We did attempt to fit models with further heterogeneity but because these models were not supported by the data we do not report on them below.

For the intake model, we let $\mathbf{v}_{i0}$ consist of an indicator of male, indicator of black, and the four indicator variables on parent's educational attainment in the outcome model. Our instruments are largely those in Barron et al. (2000). In some specifications the instruments are the size of the school enrollment divided by 100, an indicator of bad health, and the weight and height of the subject in pounds, each divided by 10. In other specifications, we exclude the school enrollment variable from the set of instruments.

Our analysis of these data is based on eight different models, derived from the unrestricted and restricted forms of $\boldsymbol{\Omega}_j$, the two different choices of instruments, and two different values of v (namely 5 and 20). Each of these models is fit under the prior where $\boldsymbol{b}_0 = 0\mathbf{i}_{10}$ and $\mathbf{B}_0$ is a diagonal matrix with 10 as the variance for the intercept and 1 as the variance for the remaining coefficients. Further, we let $\mathbf{m}_{j0} = 0\mathbf{i}_T$, $\mathbf{M}_{j0} = 1\mathbf{I}_T$, $\mathbf{c}_{j0} = 0\mathbf{i}_T$ and $\mathbf{C}_{j0} = 1\mathbf{I}_T$. Finally, we set $\mathrm{E}(D) = 0.5$ with $v_0 = 6$ degrees of freedom.

5.2. Estimation results

The fitting of each of our models is based on 20,000 MCMC iterations following a burn-in of a 1000 cycles. We begin by comparing the various models via the log-marginal likelihoods and the implied Bayes factors. The resulting log marginal likelihoods are given in Table 3. It is clear from this table that model 6 is the model favored by the data. This model has an unrestricted covariance matrix with $v = 20$, and contains the full set of instruments. Model 8, which differs from model 6 only in terms of the set of instruments, is a close second. It can also be seen from the table that models that are alike except for the degrees of freedom (for example models 1 and 2) produce smaller support for the lower degrees of freedom. In this connection we note that from these results we can actually derive the posterior distribution of v without having to run a model in which v is unknown. This is because under a uniform prior on v, the marginal posterior probability that $v = 5$ is simply proportional to the marginal likelihood $m(\mathbf{y}, \mathbf{x}|M, v = 5)$.

For the favored model 6, Table 4 gives the posterior mean and variance of the coefficients in the outcome equations and the treatment-intake equation (except for the coefficients on the year indicators) along with some summary statistics of the included covariates. The estimates of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ in columns 3 and 5 agree with the general findings in

Table 3
Log-marginal likelihoods for different model specifications

Model	Estimates of marginal likelihoods						
	Covariance matrix		Instruments		$v = 5$	$v = 20$	ln(marg. lik.)
	Restricted	Unrestricted	With enrollment	Without enrollment			
1	x	—	x	—	x	—	−2301.71
2	x	—	x	—	—	x	−1891.53
3	x	—	—	x	x	—	−2324.93
4	x	—	—	x	—	x	−1914.62
5	—	x	x	—	x	—	−2287.76
6	—	x	x	—	—	x	−1847.68
7	—	x	—	x	x	—	−2299.46
8	—	x	—	x	—	x	−1851.04

A cross indicates if the model feature in that column is present in the model.

Table 4
Model 6: Posterior means and standard deviations of $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ (without year indicators), γ and $\boldsymbol{\delta}$

Summary statistics and estimation results for model 6

	Sample mean	$\boldsymbol{\beta}_0$		$\boldsymbol{\beta}_1$		$\gamma, \boldsymbol{\delta}$	
		Mean	SD	Mean	SD	Mean	SD
Const.		4.466	0.131	4.393	0.138	−0.807	0.337
Age-17	11.07	0.004	0.007	0.016	0.007		
Education	13.89	0.083	0.008	0.085	0.007		
Job tenure (years)	3.45	0.024	0.003	0.021	0.003		
Married	0.53	0.004	0.018	0.007	0.018		
Male	0.51	0.361	0.032	0.356	0.032	0.124	0.054
Black	0.22	−0.170	0.038	−0.100	0.040	0.202	0.133
Parents Hs	0.43	0.070	0.037	0.066	0.043	0.200	0.055
Parents scol	0.14	0.061	0.053	0.018	0.053	0.310	0.074
Parents col	0.11	0.153	0.062	0.118	0.057	0.380	0.080
Parents grad	0.07	0.000	0.082	0.085	0.065	0.405	0.097
Enrollment/100	13.58					−0.014	0.003
Bad health	0.04					−0.191	0.110
Height/10	57.92					0.010	0.006
Weight/10	14.78					0.012	0.009

Sample means of the raw variables are also included.

the earnings literature: positive effects of age, education, job tenure and male gender on earnings; negative effect of being black. Note that the results in Table 4 point towards interactions between some of the covariates and the treatment state as captured by different coefficient estimates across treatment states. For example, we observe a higher negative earnings effect of being black in the control group. The estimates of the treatment-intake equation (γ) in the last two columns indicate that males and black students are more likely to participate in high school athletics, as are students with more highly educated parents. Finally, the instruments school enrollment and bad health are associated with a negative effect on participation, while height and weight have a positive effect on participation.

As mentioned above, we model dependence in the outcomes through a random effect on the constant term. In Fig. 4(a) we provide a plot of the posterior distribution of the variance of the random effect. The posterior mean of D is centered around 0.23. This is large compared to the variances of the outcome errors. For example, for the year 1992, the posterior density of the two error variances are centered 0.12 and 0.11 (as shown in Fig. 4(b) and (c)). The posterior mean of the variances in the previous three years vary between 0.06 and 0.10. In other words, our results suggest that there is considerable individual heterogeneity in these data.

We now turn to the causal effect of participation of high school athletics on weekly earnings. First we consider the posterior distributions of the correlation coefficients ρ_0 and ρ_1 from Model 6. These distributions, which are summarized in Fig. 5, reveal the pattern of confounding across time and across the treatment states. From Fig. 5 we see that there is considerable confounding, that there is some variation in the extent of confounding by year and that the confounding patterns differ by treatment state. Specifically, there is

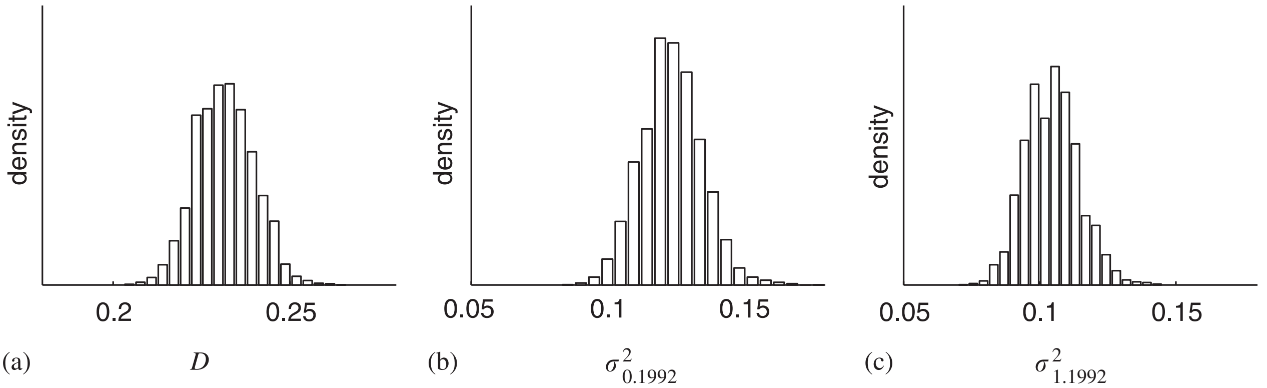


Fig. 4. Model 6: Posterior distributions of the variance of the random effects, and outcome variances for the year 1992 in the control group and the treatment group.

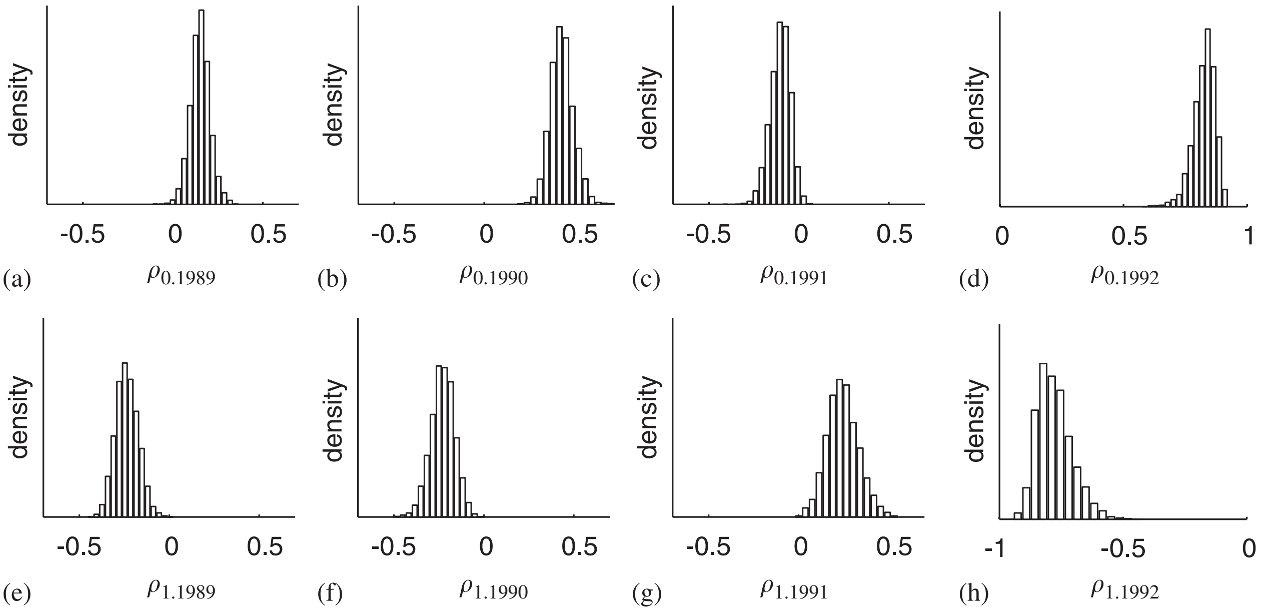


Fig. 5. Model 6: Posterior distributions of the correlation coefficients for years 1989–1992 for the control group (upper panel) and the treatment group (lower panel).

evidence of positive confounding in the control group (except for the year 1991) with posterior means of the correlation coefficients between 0.15 and 0.83, and negative confounding in the treatment group (except for the year 1991), with posterior means of the correlation coefficients between −0.23 and −0.78. A possible explanation for these results is that students with lower returns to human capital are more likely to participate in high school athletics, while students with lower preference for leisure and higher ability tend to decide against participation.

To estimate the causal effect of participation in high school athletics (after controlling for confounding on unobserved factors and individual heterogeneity), we focus on the predictive treatment effects. Table 5 provides the estimates of the predictive ATE and predictive quantile treatment effects (for the 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles) under Model 6. As can be seen from this table, the average treatment effect is positive in all years, with the estimated ATE pointing to an increase in weekly earnings between \$15.32 and \$44.60. In terms of percentages of the average weekly income, the ATEs for the four years are 11%, 4%, 8% and 10% respectively. Further, from columns (4) through (6) of the

Table 5
Model 6: Estimates of the predictive average and the 0.05, 0.25, 0.5, 0.75 and 0.95 quantile treatment effects for the years 1889–1992

Predictive treatment effects from model 6 ($ per week)

Year	Mean	Quantiles				
		0.05	0.25	0.5	0.75	0.95
1989	41.26	14.30	25.45	40.35	47.98	79.08
1990	15.32	10.11	11.52	16.42	21.25	−5.69
1991	36.31	−1.53	11.75	20.13	45.12	105.39
1992	44.60	24.00	32.60	44.42	52.69	79.36

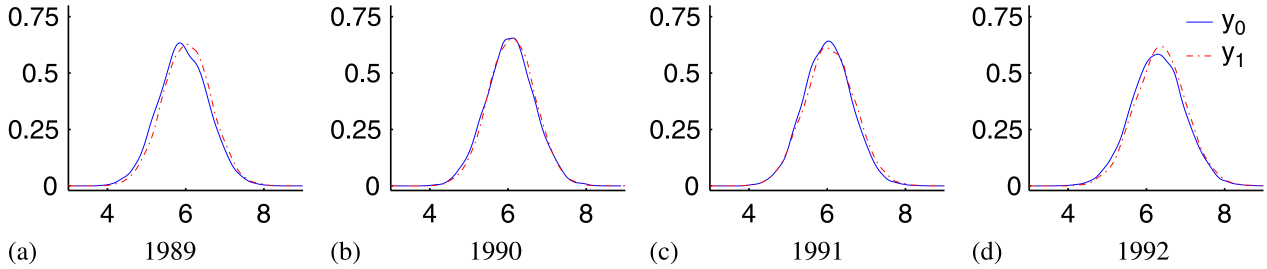


Fig. 6. Model 6: Predictive densities of the potential outcomes for the years 1989–1992.

table we see that the predictive treatment effect is positive for all time periods at the 25%, 50% and 75% quantiles. With two exceptions, we also observe a positive effect at the 5% and 95% quantiles. The positive effect on earnings can also be seen from Fig. 6 where we plot the predictive density of the potential outcomes.

6. Conclusion

We have considered a Bayesian analysis of a new model that can be used to find the effect of a treatment applied non-randomly at baseline given a panel of outcomes observed subsequently. For this model we have developed tuned MCMC methods to sample the posterior distribution, to compute the marginal likelihood, and to find various treatment effects of interest. One important feature of our modeling and fitting approach is that it does not require the unknowable joint distribution of the potential outcomes or the simulation of the counterfactuals. That the analysis can proceed in this way comes from Chib (2004). We have examined the performance of the proposed methodology in simulation experiments and have found that the inferences are sufficiently well behaved for the model and methods to be useful in practical applications. Our simulation experiments also reveal the gains from exploiting panel information versus fitting a sequence of cross-section models. Finally, we have applied the proposed techniques to study the effect of participation in high school athletics on future earnings for a sample of subjects drawn from the NLSY. On the basis of these empirical studies, both with simulated and real data, it appears that the model and estimation framework have promise for applied work.

Acknowledgments

The authors thank the referees and the editor for their constructive and extremely helpful comments and suggestions.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–779.

Barron, J.M., Ewing, B.T., Waddell, G.R., 2000. The effect of high school athletic participation on education and labor market outcomes. The Review of Economics and Statistics 82, 409–421.

Becker, G., 1965. A theory of allocation of time. Economic Journal 75, 493–517.

Card, D., 1999. The causal effect of education on earnings. Handbook of Labor Economics, vol. 3. Elsevier, Amsterdam.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

Chib, S., 2001. Markov Chain Monte Carlo methods: computation and inference. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. North-Holland, Amsterdam, pp. 3569–3649.

Chib, S., 2004. Analysis of treatment response data without the joint distribution of counterfactuals. Journal of Econometrics, in press.

Chib, S., Carlin, B., 1999. On MCMC sampling in hierarchical longitudinal models. Statistics and Computing 9, 17–26.

Chib, S., Greenberg, E., 1994. Bayes Inference in Regression Models with ARMA (p,q) Errors. Journal of Econometrics 64, 183–206.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. American Statistician 49, 327–335.

Chib, S., Hamilton, B.H., 2002. Semiparametric Bayes analysis of longitudinal data treatment models. Journal of Econometrics 110, 67–89.

Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis–Hastings output. Journal of the American Statistical Association 96, 270–281.

Hirano, K., Imbens, G., Rubin, D.B., Zhao, X.H., 2000. Estimating the effect of an influenza vaccine in an encouragement design. Biostatistics 1, 69–88.

Imbens, G.W., Rubin, D.B., 1997. Bayesian inference for causal effects in randomized experiments with noncompliance. The Annals of Statistics 25, 305–327.

Lee, L.F., 1978. Unionism and wage rates: a simultaneous equation model with qualitative and limited dependent variables. International Economic Review 19, 415–433.

Li, M., Poirier, D.J., Tobias, J., 2004. Do dropouts suffer from dropping out? Estimation and prediction of outcome gains in generalized selection models. Journal of Applied Econometrics 9, 203–225.

Yau, L., Little, R.J.A., 2001. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data with application to a job training assessment for the unemployed. Journal of the American Statistical Association 96, 1232–1244.