# Bayesian Methods in Microeconometrics

Mingliang Li
SUNY-Buffalo
mli3@buffalo.edu

Justin L. Tobias
Purdue University
jltobias@purdue.edu

April, 2010

**Abstract**

This chapter reviews Bayesian analysis of models commonly employed in microeconometric applications. We begin with the model central to this literature - the linear regression model - and also explore some of its basic generalizations. We then turn to analysis of other common microeconometric models, including the probit, logit, tobit and ordered probit, and unite these within a common hierarchical structure. Finally, other prominent topics in microeconometrics, including problems of endogeneity, analysis of treatment effects, and the analysis of count and duration data, are also discussed and references to the literature are provided. Throughout we provide a description of Markov Chain Monte Carlo (MCMC) samplers for estimating the models and also provide several illustrative examples.

# 1 Introduction

This chapter is intended to serve as an introduction to Bayesian analysis of models commonly encountered in microeconomics. In what follows we cover much of the "how" to conduct Bayesian inference in microeconometric applications by discussing, in reasonable detail, the steps involved in posterior simulation via Markov Chain Monte Carlo (MCMC) methods in a wide array of models. To a lesser extent we also address issues of "why" one might choose to employ a Bayesian approach for estimating these models over well-established frequentist alternatives. Our answers to the latter types of questions tend to be pragmatic rather than grounded in theory, emphasizing the ease with which simulations from the posterior distribution can be used to calculate exact finite sample point estimates or complete posterior distributions for economically relevant quantities of interest.

The level of presentation of this chapter is quite similar to that provided in recent Bayesian textbooks, including (Koop 2003; Lancaster 2004; Geweke 2005; Koop, Poirier and Tobias 2007). The reader may, in fact, regard this chapter as an introduction to several of the more specialized chapters that appear elsewhere in this volume. For example, (Griffin, Quintana and Steel 2010) provide flexible alternatives to many of the more restrictive assumptions entertained here, all of which are presented under the assumption of normal sampling. Furthermore, (Rossi and Allenby 2010) also relax some of the distributional and prior assumptions made in the basic hierarchical and multinomial choice frameworks and illustrate the value of such models in marketing applications. Given the broad scope of a chapter like this one, however, our coverage of a representative model will typically be brief rather than fully detailed and, as is necessary, we will provide the reader with references to the literature that extend the basic methodology.

Our approach to this chapter is to teach by example. By this we mean that many of the models considered will contain an actual empirical example to illustrate the use of MCMC methods in practice. In most cases our examples employ data sets that are specific to the model being considered, although the estimation of several different types of models will be illustrated with a common data set on BMI (Body Mass Index) and wage outcomes. Finally, all of our applications are coded in MATLAB and our programs made available to the interested reader for inspection, refinement and additional modifications.[1]

[1]The code can be obtained from the website: http://web.ics.purdue.edu/~jltobias/handbook.html.

The outline of this chapter is as follows. Section 2 discusses linear models. We begin this presentation with a review of the normal linear regression model, deriving marginal, conditional and predictive posterior densities of interest. To illustrate how MCMC methods can be employed to accommodate interesting departures from the standard linear regression framework under "ideal" conditions, we also discuss several generalizations of the basic model, including allowing for parametric heteroscedasticity and incorporating a changepoint into the analysis. Our treatment of linear models then moves on to discuss hierarchical linear models and to review approaches to handling endogeneity problems in the context of a bivariate system of linear equations. Section 3 presents applications and posterior simulation strategies for univariate (nonlinear) latent variable models, including the probit, logit, tobit and ordered probit specifications. Section 4 extends these approaches to the multivariate case and considers the analysis of treatment effects models and multinomial and multivariate probit models. Finally, section 5 briefly reviews basic Bayesian approaches to the analysis of duration data, and the paper concludes with a summary in section 6.

# 2 Linear Models

The linear regression model is central to microeconometrics, and so it seems natural to begin our review of Bayesian microeconometrics with an investigation of this model. We start by discussing Bayesian inference in the linear regression model under ideal conditions. While a careful understanding of this model is surely useful in its own right, what is learned from analysis of the linear model will also prove useful when estimating generalized models, like those of sections 3 and 4, that will be linear in suitably defined latent data.

## 2.1 Bayesian Analysis of the Linear Regression Model

Before discussing such generalizations, we first consider a standard regression model of the form

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i, \qquad u_i|\boldsymbol{X}, \sigma^2 \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{x}_i$ is a $1 \times k$ vector of covariate data, $y_i$ is a scalar outcome of interest, $\boldsymbol{\beta}$ and $\sigma^2$ are a $k \times 1$ vector of regression parameters and a scalar variance parameter, respectively, and

$$\boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}.$$

The likelihood is derived from (1), and specification of the model is completed upon selecting a prior for the parameters $\boldsymbol{\beta}$ and $\sigma^2$. To this end, we choose proper priors of the forms:

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{V}_{\boldsymbol{\beta}}\right) \tag{2}$$

$$\sigma^2 \sim IG\left(\frac{a}{2}, b\right), \tag{3}$$

that is, a conditional normal prior for the regression coefficients and an inverse gamma prior for the variance parameter.[2] The hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_{\beta}, a$ and $b$ are known and selected by the researcher.

With respect to the prior, it is often selected to be conjugate (meaning that posteriors in the same distributional family are produced), as will be the case for (2)-(3), primarily for reasons of computational tractability [see, e.g., (Bernardo and Smith 1994; Poirier 1995)]. The adoption of conjugate priors can also be viewed as the addition of "fictitious" sample information to the analysis that is combined with the data in exactly the same way that additional (real) sample information would have been combined. Thus, conjugate priors enable the researcher to directly assess the informational content of the prior in terms of equivalent sample information - a useful result in practice, if for no other reason than to potentially mitigate concerns about the influence of the prior.

Within a given class of priors it remains, of course, to choose the hyperparameters. This decision can potentially be guided based upon the findings of past research, when available. While it may be difficult in general for the researcher to elicit her prior beliefs regarding unobservable parameters when such information does not exist, a useful exercise is to think about implications of the prior on the prior predictive, $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2$, which is something the researcher is probably informed about. Here, $\boldsymbol{y} = (\ y_1 \quad y_2 \quad \cdots \quad y_n\ )'$ and we will discuss more

[2] In this chapter we follow the general conventions of the Handbook and parameterize the inverse gamma as follows: $x \sim IG(c, d) \Rightarrow p(x) \propto x^{-(c+1)} \exp(-d/x)$, $c > 0, d > 0, x > 0$. For $c > 1$ it follows that $E(x) = d(c-1)^{-1}$ and, for $c > 2$, $\text{Var}(x) = d^2[(c-1)^2(c-2)]^{-1}$.

on the calculation of $p(\boldsymbol{y})$ below. Finally, it may also be tempting to simply use improper priors in practice, as they appear to be the closest approximation to letting the data completely speak for itself. Doing so is not without problems, however, as marginal likelihoods are generally no longer well-defined, improper priors can be unexpectedly informative for functions of the parameters, and marginalization paradoxes can occur. In this chapter, we simply employ conjugate (or conditionally conjugate) priors and in the limited space available focus on issues of implementation and posterior simulation rather than prior selection. Interested readers can specify their own priors, of course, and slightly modify the code provided to see how results change.

### 2.1.1 Marginal Posteriors, Conditional Posteriors and Posterior Predictive Distributions

The prior and likelihood combine via Bayes' Theorem to yield the joint posterior, up to a constant of proportionality. Applying this general result to our linear regression model, we obtain:

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2), \tag{4}$$

without the conditioning on $\boldsymbol{X}$ explicitly denoted. Though (4) will be tailored to the case of linear regression in this section, it also summarizes the general process of Bayesian learning, as priors combine with the likelihood to form posterior distributions for the model parameters. For low dimension problems, the right-hand side of (4) can be plotted to visualize how prior beliefs have been updated from the data. When the dimension of the parameter vector is very low, standard numerical integration routines such as Simpson's rule or Gaussian quadrature can be employed to approximate the normalizing constant of (4) and thereby plot a proper joint posterior density.

When the dimension of the parameter space is moderate or large and the structure of the posterior distribution is not simple, however, it becomes far more difficult to visualize interesting features of the posterior surface and direct calculation of the normalizing constant using standard methods is no longer possible. As many of the chapters of this volume describe, it is often possible, however, to simulate draws from (4) and use these draws to calculate posterior quantities of interest.

In the case of the linear regression model described here, such numerical methods are not required as the priors in (2)-(3) combine nicely with the likelihood function, and all of the requisite posterior densities are available in closed form. To see this, suppose that the objects of interest happen to

be the regression coefficients $\boldsymbol{\beta}$.[3] To this end, one would like to report posterior summary statistics such as posterior means or posterior standard deviations for the elements of the regression coefficient vector.

A step in this direction leads us to consider the posterior conditional $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}$. This is obtained upon noting that its density is proportional to that of the joint posterior $\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}$ in (4) and then completing the square in $\boldsymbol{\beta}$ to obtain [e.g., (Lindley and Smith 1972)]:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}, \sigma^2 \boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}}\right) \tag{5}$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}} = \left(\boldsymbol{X'X} + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\right)^{-1} \quad \text{and} \quad \boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}} = \boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}}\left(\boldsymbol{X'y} + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right). \tag{6}$$

The dependence of this density on $\sigma^2$ is rather undesirable, however, as we seek to report posterior statistics about $\boldsymbol{\beta}$ that do not require such conditioning on unobservables. The Bayesian prescription is clear: simply marginalize the nuisance parameter out of the problem, or obtain:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \int_0^\infty p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}) p(\sigma^2|\boldsymbol{y}) d\sigma^2. \tag{7}$$

The first term within the integral has been determined, as in (5), and it remains to calculate the marginal posterior for the variance parameter, $p(\sigma^2|\boldsymbol{y})$. This quantity can be obtained by starting with the joint posterior $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$ in (4), completing the square in $\boldsymbol{\beta}$, recognizing the resulting quadratic form in $\boldsymbol{\beta}$ as being part of a multivariate normal kernel for $\boldsymbol{\beta}$, and then integrating over the multivariate normal. Doing so gives:

$$\sigma^2|\boldsymbol{y} \sim IG\left(\frac{n+a}{2}, \tilde{b}\right), \tag{8}$$

where

$$\tilde{b} = \left[b + \frac{1}{2}\left(SSE + (\boldsymbol{\mu}_{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' \left[\boldsymbol{V}_{\boldsymbol{\beta}} + (\boldsymbol{X'X})^{-1}\right]^{-1} (\boldsymbol{\mu}_{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})\right)\right], \tag{9}$$

with

$$\hat{\beta} \equiv (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \quad \text{and} \quad SSE \equiv (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{10}$$

Equation (8), of course, is interesting in its own right as it can be used to calculate marginal posterior statistics for the variance parameter $\sigma^2$, using known properties of the inverse gamma

[3] Though it is convention in the profession to at least report a table of coefficient posterior means and standard deviations, and we tend to follow this convention in this chapter, parameters themselves rarely tell the whole story. The examples provided in the following subsections illustrate how quantities of interest are commonly functions of the parameters, and how posterior simulations can be easily used to make inference regarding such quantities.

distribution. Equations (5) and (8) also reveal that the prior in (2)-(3) is *conjugate*, as originally asserted, since the prior and posterior are of the same distributional family. For our purposes here, the result in (8) can also be substituted into (7) and the necessary integration performed to obtain the marginal posterior for $\boldsymbol{\beta}$. By doing so, we obtain:[4]

$$\boldsymbol{\beta}|\boldsymbol{y} \sim t(\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}, [2\tilde{b}\boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}}]^{-1}, n+a), \tag{11}$$

a multivariate $t$ density with mean $\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}$ (for $n+a>1$) and variance $(n+a-2)^{-1}2\tilde{b}\boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}} = E(\sigma^2|\boldsymbol{y})\boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}}$ (for $n+a>2$), with both $\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}$ and $\boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}}$ being defined in (6). The parameter $\tilde{b}$ is defined in (9) as the second parameter of the IG density. The density in (11) can be used to calculate posterior means, posterior standard deviations, optimal point and interval estimates [see, e.g., (Poirier 1995: Chapters 6 and 9)] or other desired quantities for the regression parameters $\boldsymbol{\beta}$.

Apart from estimation, we would also like to use our linear regression framework for two additional purposes: prediction and model comparison. In terms of the latter, marginal likelihoods [see, e.g., (Chib 2010) of this volume] are often calculated and Bayes factors and/or posterior model probabilities reported to compare models or average model-specific posterior predictions. With the priors employed in (2)-(3) together with our normal sampling model, the marginal density of the data $\boldsymbol{y}$ is also available analytically (e.g., Poirier 1995):

$$\boldsymbol{y} \sim t\left(\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}, \left[2b\left(\boldsymbol{I}_{\boldsymbol{n}} + \boldsymbol{X}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{X}'\right)\right]^{-1}, a\right). \tag{12}$$

Equation (12), when evaluated at the observed sample of data $\boldsymbol{y}^o$, provides the marginal likelihood, which provides a vehicle for model comparison, selection and averaging. (Chib 2010) of this volume provides many more details surrounding the calculation and use of marginal likelihoods in practice and we refer the interested reader there for further details.

In terms of prediction, consider a future, out-of-sample value $y_f$, presumed to be generated by the model in (1):

$$y_f = \boldsymbol{x}_f\beta + u_f, \quad u_f|\boldsymbol{X}, \boldsymbol{x}_f, \sigma^2 \sim \mathcal{N}(0, \sigma^2). \tag{13}$$

The posterior predictive density for $y_f$ (given values of the covariates $\boldsymbol{x}_f$) is obtained as:

$$p(y_f|\boldsymbol{x}_f, \boldsymbol{y}) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{0}^{\infty} p(y_f|\boldsymbol{x}_f, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})d\sigma^2 d\boldsymbol{\beta}, \tag{14}$$

[4]We, again, employ a parameterization of the multivariate $t$ distribution that is consistent with the usage in this Handbook. Specifically, for a $k\times 1$ vector $x$, writing $\boldsymbol{x} \sim t(\boldsymbol{\mu}, \boldsymbol{V}, \nu)$ implies $p(x) \propto [1+(\boldsymbol{x}-\boldsymbol{\mu})'V(\boldsymbol{x}-\boldsymbol{\mu})]^{-(k+\nu)/2}$.

noting that the future outcome $y_f$ is independent of the past outcomes $\boldsymbol{y}$, given the covariates $\boldsymbol{x}_f$ and parameters $\boldsymbol{\beta}$ and $\sigma^2$. Methods similar to those used in deriving (11) produce:

$$y_f|\boldsymbol{x}_f, \boldsymbol{y} \sim t\left(\boldsymbol{x}_f\boldsymbol{\mu}_{\boldsymbol{\beta}|\boldsymbol{y}}, \left[2\tilde{b}(1+\boldsymbol{x}_f\boldsymbol{V}_{\boldsymbol{\beta}|\boldsymbol{y}}\boldsymbol{x}_f')\right]^{-1}, n+a\right), \tag{15}$$

which can be used to make point, interval or other predictions regarding out-of-sample outcomes. The following example illustrates how such results can be used in practice.

### 2.1.2 An Illustrative Application with (Log) Wage Data

To illustrate how Bayesian calculations are carried out in this simplest specification, we obtain a sample of $n = 1,645$ observations on white males in the U.S. in 1993. Our data, taken from the National Longitudinal Survey of Youth (NLSY), contain information on wage outcomes for all of these respondents. The dependent variable we employ is the natural logarithm of hourly wages received in 1993, measured in 1993 dollars. Other demographic variables, included as right-hand-side covariates in our analysis, include years of schooling completed ($EDUCATION$), a test score variable ($SCORE$), years of schooling completed by the respondent's parents ($MOMED$ and $DADED$) and number of siblings ($NUMSIBS$) of the respondent. The variable $SCORE$ is constructed from a battery of tests administered to the NLSY participants during the fall and summer of 1980, and is standardized to have a sample mean of zero and sample variance equal to unity.

While analytic results are available for the $\boldsymbol{\beta}$ and $\sigma^2$ marginal posteriors, we employ simulation methods to generate draws from the joint posterior $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$ and use these simulations to calculate posterior quantities of interest. Sampling is conducted using the method of composition, first drawing from the marginal posterior for the variance parameter, $\sigma^2|\boldsymbol{y}$ in (8) and then sampling from the conditional posterior for the regression parameters, $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}$ in (5). This is repeated 25,000 times, producing 25,000 *iid* samples from the joint posterior $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$. Monte Carlo simulation methods are employed as their use will easily enable us to calculate posterior statistics for economically relevant (nonlinear) hourly wage gaps of the form:

$$\Delta(\boldsymbol{\beta}, \sigma^2; \boldsymbol{x}_l, \boldsymbol{x}_h) = \exp\left(\boldsymbol{x}_h\boldsymbol{\beta} + \frac{\sigma^2}{2}\right) - \exp\left(\boldsymbol{x}_l\boldsymbol{\beta} + \frac{\sigma^2}{2}\right) \tag{16}$$

in addition to simple posterior statistics regarding the parameters $\boldsymbol{\beta}$ and $\sigma^2$.

Equation (16) represents the expected hourly wage gap between persons of characteristics $\boldsymbol{x}_h$ and $\boldsymbol{x}_l$. In practice, the two sets of covariates are distinguished by evaluating $\boldsymbol{x}_h$ at a "higher" education category and $\boldsymbol{x}_l$ at a lower education category, while the remaining values in both covariate vectors are fixed at sample means. We focus in particular on the B.A./ High School wage gap, and the Ph.D./ High School wage gap. For the first of these cases, $\boldsymbol{x}_h$ sets $EDUCATION = 16$ while in the second, $\boldsymbol{x}_h$ sets $EDUCATION = 20$. For the High School comparison group, $\boldsymbol{x}_l$ sets education equal to 12 each time.

The ease with which our posterior simulations can be used to calculate quantities like (16) should not be overlooked, as classical approaches to inference, via delta-method asymptotics or the bootstrap, seem to be significantly more difficult to implement. For example, a point estimate (posterior mean) of (16) can be readily calculated as

$$\widehat{\Delta}(\boldsymbol{x}_l, \boldsymbol{x}_h) = \frac{1}{R}\sum_{r=1}^{R} \Delta(\boldsymbol{\beta}^{(r)}, \sigma^{2,(r)}; \boldsymbol{x}_l, \boldsymbol{x}_h), \tag{17}$$

with $\boldsymbol{\beta}^{(r)}$ and $\sigma^{2,(r)}$ denoting the $r^{th}$ simulation from the joint posterior and $R$ denoting the total number of simulations. A point estimate of the posterior standard deviation of (16) can be calculated in an analogous way, using the simulations produced from the Monte Carlo sampling scheme.

Posterior summary statistics associated with two different $\Delta$ parameters and the regression coefficients are reported in Table 1. These results are obtained upon setting $a = 6$, $b = 1$, $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}_k$ and $\boldsymbol{V}_{\boldsymbol{\beta}} = 10\boldsymbol{I}_k$, providing a proper yet reasonably uninformative prior.

Table 1: Posterior Statistics From Wage Data Application

| Variable / Parameter | Posterior Mean | Posterior Std. Dev. |
|---|---|---|
| Constant | 1.79 | .102 |
| EDUCATION | .044 | .007 |
| SCORE | .096 | .017 |
| MOMED | .003 | .006 |
| DADED | .007 | .005 |
| NUMSIBS | .004 | .006 |
| BA/HS GAP | 2.51 | .406 |
| Ph.D./HS GAP | 5.51 | .971 |

The entries of Table 1 have the expected signs and posterior means appear reasonable in magnitude, with education and test scores clearly having a meaningful impact on earnings. On average (and in

1993 dollars), those graduating with a four-year degree earn about $2.51 more per hour than high school graduates while those with a Ph.D. earn about $5.51 more per hour than their high school counterparts.

Rather than simply looking at differences in means, one can also obtain entire predictive wage distributions. To illustrate how this is done, Figure 1A plots posterior predictive hourly wage densities for two different hypothetical individuals - an individual with a high school degree who is otherwise "average" (that is, all other covariates are fixed at sample mean values), and a similarly average individual with a BA degree.

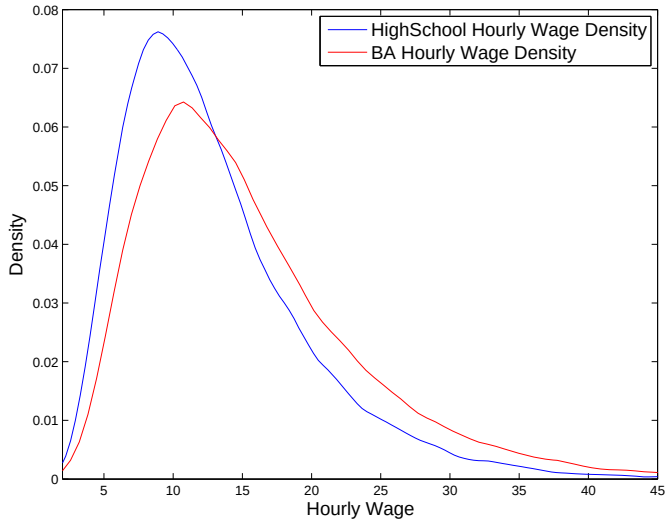


Figure 1A: Posterior Predictive Hourly Wage Densities

Letting $w_f = \exp(y_f)$ denote the hourly wage for a future or out-of-sample individual with characteristics $\boldsymbol{x}_f$, these figures can be obtained by noting:

$$p(w_f|\boldsymbol{x}_f, \boldsymbol{y}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{0}^{\infty} p(w_f|y_f) p(y_f|\boldsymbol{x}_f, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}) p(\sigma^2|\boldsymbol{y}) d\sigma^2 d\boldsymbol{\beta} \tag{18}$$

where superfluous information has been dropped from the conditioning above, when applicable.

Although evaluation of the multiple integral in (18) may seem like a daunting task, it is useful to pause and emphasize how simulation methods can be employed to generate draws from this predictive density. To begin, samples from $p(\sigma^2|\boldsymbol{y})$ and $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})$ are obtained by first drawing from (8) and then (5) immediately after, updating the conditioning in (5) to reflect the $\sigma^2$ just generated from (8). Let us denote this posterior sample at iteration $r$ as $(\sigma^{2,(r)}, \boldsymbol{\beta}^{(r)})$. A log

wage simulation $y_f^{(r)}$ is then produced by drawing from the normal sampling model in (13), setting $y_f^{(r)} = \boldsymbol{x}_f\boldsymbol{\beta}^{(r)} + \sigma^{(r)}z$, where $z \sim \mathcal{N}(0,1)$. Finally, an hourly wage simulation $w_f^{(r)}$ is produced by simply exponentiating the log wage simulation $y_f^{(r)}$, given that the conditional distribution is degenerate, i.e., $p(w_f|y_f) = I(w_f = \exp[y_f])$. This produces a set of draws from the posterior predictive density of hourly wages. Therefore, draws from (18) only require a few additional lines of code beyond what have already been written to fit the model, and obtaining them does not require change of variables analytics or large-sample approximations for inference.

In Figure 1A we generate 25,000 simulations from the posterior predictive density (18) in this manner and smooth these simulations via a kernel method to plot the posterior predictive hourly wage densities for the High School graduate and B.A. groups. In these calculations all covariates other than education are fixed at their sample mean values. The posterior simulations can also be used to directly calculate posterior predictive means, standard deviations, etc., as well as other economically relevant quantities such as the probability of being in poverty [e.g., (Geweke and Keane 2000)], and we briefly take up the last of these in the context of our application. Specifically, we calculate $\Pr(w_f < \$5 \mid \boldsymbol{y}, Ed = 12, X_{-Ed} = \overline{X}_{-Ed}) \approx .05$ and $\Pr(w_f < \$5 \mid \boldsymbol{y}, Ed = 16, X_{-Ed} = \overline{X}_{-Ed}) \approx .025$ where $Ed$ denotes education, $X_{-c}$ denotes all variables other than $c$ and $\overline{X}$ denotes the sample average. The value \$5 was chosen as an approximate hourly wage consistent with the poverty threshold at full-time employment.[5]

## 2.2 Heteroscedasticity in Linear Models

Despite the prominence of heteroscedasticity in the theory and practice of frequentist econometrics, and its ubiquitousness in classical graduate econometrics texts, the role of heteroscedasticity in Bayesian treatments is comparatively minor, and its appearance as a component of Bayesian empirical work seems the exception rather than the rule. Although several explanations exist for the differential treatment of this issue among frequentists and Bayesians, it remains rather strange, and potentially troubling, that heteroscedasticity seems to garner comparatively little attention.

Of course, the issue of heteroscedasticity has not been completely neglected among Bayesians.

[5]The U.S. Census Bureau reports a 1993 poverty threshold for a two person family with one child less than 18 years of age equal to \$9,960. Using 2,000 hours as the number of annual hours worked for someone engaged in full-time employment motivates our decision to use \$5 as the approximate hourly wage threshold. This is, admittedly, a simplified calculation and, among other things, assumes no other sources of income for the household.

(Poirier 2008), for example, builds upon (Lancaster 2003) and seeks to reconcile (White's 1980) heteroscedasticity-robust estimation of the OLS covariance matrix within the Bayesian framework.[6] In a different spirit, which seeks to employ Bayesian methods to flexibly model the variance function, (Yau and Kohn 2003) consider analysis of a normal linear regression model with splines employed for both the mean and variance functions and variable selection used to determine key terms in the variance function. (Leslie, Kohn and Nott 2007) present a related approach, where the error distribution is modeled nonparametrically, parameteric forms are specified for the mean and variance functions, and variable selection methods are used to select appropriate covariates in both sets of functions. Villani et al (2007), based upon an approach similar to the smoothly mixing regression model of Geweke and Keane (2007), describe a nonparametric-type approach to the modeling of heteroscedasticity.

Although these papers offer valuable contributions, the norm in applied work appears to remain one of conditional homoscedasticity. With this in mind we describe in the following section a simple generalization of this assumption which permits a multiplicative, parametric form of heteroscedasticity. When such a specification is not adequately flexible, the reader is invited to see the references listed above for more advanced alternatives.

### 2.2.1 Posterior Simulation in a Model of Parametric Heteroscedasticity

To provide some initial guidance and a simple first step toward handling linear models in the presence of heteroscedasticity, we consider analysis of the following specification:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\alpha} \overset{ind}{\sim} \mathcal{N}[0, \exp(\boldsymbol{z}_i\boldsymbol{\alpha})] \tag{19}$$

where it is understood that an intercept is included in $\boldsymbol{z}_i$ and that $\boldsymbol{z}_i$ can be the same as, or potentially different from $\boldsymbol{x}_i$. We complete the specification of our model in (19) with priors of the forms:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}) \tag{20}$$

$$\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{V}_{\boldsymbol{\alpha}}). \tag{21}$$

[6] Interestingly, the first sentence of (Poirier 2008) frames the exercise well, as it reads: "Often, researchers find it useful to recast frequentist procedures in Bayesian terms, so as to get a clear understanding of how and when the procedures work."

These priors in (20) and (21) together with the likelihood implied by (19) yield a posterior of the form:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta})p(\boldsymbol{\alpha}) \left[\prod_{i=1}^{n} \exp(\boldsymbol{z}_i\boldsymbol{\alpha})\right]^{-1/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{\exp(\boldsymbol{z}_i\boldsymbol{\alpha})}\right). \tag{22}$$

We propose a Metropolis-within-Gibbs algorithm [see (Chib 2010) of this volume for further details] to generate draws from this posterior density.[7]

To this end, we first recognize that:

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{D}_{\boldsymbol{\beta}}\boldsymbol{d}_{\boldsymbol{\beta}}, \boldsymbol{D}_{\boldsymbol{\beta}}\right), \tag{23}$$

where

$$\boldsymbol{D}_{\boldsymbol{\beta}} = \left(\boldsymbol{X'W^{-1}X} + \boldsymbol{V_{\beta}^{-1}}\right)^{-1}, \qquad \boldsymbol{d}_{\boldsymbol{\beta}} = \boldsymbol{X'W^{-1}y} + \boldsymbol{V_{\beta}^{-1}\mu_{\beta}}, \tag{24}$$

and

$$\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\alpha}) \equiv diag\{\exp(\boldsymbol{z}_i\boldsymbol{\alpha})\}. \tag{25}$$

As for the sampling of $\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{y}$ we note from (22),

$$p(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{y}) \propto p(\boldsymbol{\alpha}) \left[\prod_{i=1}^{n} \exp(\boldsymbol{z}_i\boldsymbol{\alpha})\right]^{-1/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{\exp(\boldsymbol{z}_i\boldsymbol{\alpha})}\right), \tag{26}$$

which is not of a known form. We employ a random walk Metropolis-within-Gibbs step to generate draws from (26). Specifically, we first sample a candidate $\boldsymbol{\alpha}^*$ from a multivariate normal proposal density:

$$\boldsymbol{\alpha}^* \sim \mathcal{N}(\boldsymbol{\alpha}^{(r)}, c^2\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}) \tag{27}$$

where the $(r)$ superscript denotes the current value of the chain at iteration $r$. Implementation of our method requires choosing the scale matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ as well as the "tuning parameter" $c^2$, the latter of which will be chosen to optimize the mixing of the simulations within this scheme for variance parameter simulation. Following (Harvey 1976), we select the scale matrix by first noting

$$\frac{\epsilon_i^2}{\exp(\boldsymbol{z}_i\boldsymbol{\alpha})} \equiv \nu_i, \quad \nu_i \sim \chi_1^2, \tag{28}$$

suggesting that a point estimate of $\boldsymbol{\alpha}$ can be obtained from a regression of $\log \epsilon_i^2$ (which is known, given $\boldsymbol{\beta}$) on $\boldsymbol{z}_i$.[8] A reasonable choice for $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$, then, is the variance-covariance matrix from this

[7] (Tanizaki and Zhang 2001) provide a similar analysis to the one presented here.

[8] In practice, we start the MCMC chain by first running an OLS regression of $y_i$ on $\boldsymbol{x_i}$ to obtain the initial value of $\boldsymbol{\beta}$. We then run the above regression of $\log \epsilon_i^2$ on $\boldsymbol{z_i}$ to obtain an initial value of $\boldsymbol{\alpha}$, adding 1.27 to the intercept parameter. The adjustment to the intercept is due to the fact that the mean of $\log \nu_i$ is (approximately) -1.27 instead of zero.

regression, or,

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = 4.93(\boldsymbol{Z}'\boldsymbol{Z})^{-1} \tag{29}$$

where the value 4.93 denotes the approximate variance of $\log \nu_i$. The candidate $\boldsymbol{\alpha}^*$ generated from the proposal is then accepted with probability

$$\min\left\{1, \frac{p(y|\boldsymbol{\alpha}=\boldsymbol{\alpha}^*, \boldsymbol{\beta}=\boldsymbol{\beta}^{(r)})p(\boldsymbol{\alpha}^*)}{p(y|\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}=\beta^{(r)})p(\boldsymbol{\alpha}^{(r)})}\right\} \equiv \min\left\{1, \exp[g(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}^{(r)})]\right\}, \tag{30}$$

where

$$\begin{aligned} g(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}^{(r)}) = \; & -\frac{1}{2}\Big[\boldsymbol{\iota}_n'\boldsymbol{Z}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^{(r)}) + \sum_i (y_i - \boldsymbol{x}_i\boldsymbol{\beta}^{(r)})^2 \left(\exp(-\boldsymbol{z}_i\boldsymbol{\alpha}^*) - \exp(-\boldsymbol{z}_i\boldsymbol{\alpha}^{(r)})\right) \\ & + (\boldsymbol{\alpha}^* - \boldsymbol{\mu}_{\boldsymbol{\alpha}})'\boldsymbol{V}_{\boldsymbol{\alpha}}{}^{-1}(\boldsymbol{\alpha}^* - \boldsymbol{\mu}_{\boldsymbol{\alpha}}) - (\boldsymbol{\alpha}^{(r)} - \boldsymbol{\mu}_{\boldsymbol{\alpha}})'\boldsymbol{V}_{\boldsymbol{\alpha}}{}^{-1}(\boldsymbol{\alpha}^{(r)} - \boldsymbol{\mu}_{\boldsymbol{\alpha}})\Big], \end{aligned} \tag{31}$$

and $\boldsymbol{\iota}_n$ denotes an $n \times 1$ vector of ones. If the candidate $\boldsymbol{\alpha}^*$ is accepted, then $\boldsymbol{\alpha}^{(r+1)} = \boldsymbol{\alpha}^*$. Otherwise, the chain remains at its current value, setting $\boldsymbol{\alpha}^{(r+1)} = \boldsymbol{\alpha}^{(r)}$.

### 2.2.2 Adding Parametric Heteroscedasticity to the Log Wage Data Application

Using the algorithm above we estimate the heteroscedastic regression model employing the wage data of section 2.1.2. All covariates in $\boldsymbol{x}_i$ are included as covariates in $\boldsymbol{z}_i$ and we fit the model by sampling successively from the conditional posteriors in (23) and (26), discarding the first 10,000 of 100,000 simulations as the burn-in period. For our priors, we set $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\mu}_{\boldsymbol{\alpha}} = \mathbf{0}_k$ and $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\alpha}} = 100I_k$.

The tuning parameter $c^2$ is set equal to 1/2, which was chosen experimentally and in a rather ad hoc fashion, and results in an acceptance rate of approximately 23% in the M-H (Metropolis-Hastings) step. This roughly matches the rule of thumb of (Gelman, Roberts and Gilks 1996; Koop 2004: section 5.5.2), suggesting that random walk chain acceptance rates near 25 percent in large dimension problems may offer reasonable targets. Of course, interest should ultimately center on the numerical precision of the simulation-based estimate (which will vary with different choices of $c$) as well as its precision relative to what we could have obtained had iid draws been employed. To this end, we report in Table 2 inefficiency factors associated with alternate choices of $c^2$, and specifically consider $c^2 \in \{.1, .5, 1\}$ for illustration purposes. As shown in that table, the 25% target (which we come close to with $c^2 = .5$, whereas $c^2 = .1$ and $c^2 = 1$ yield acceptance rates of 56%

and 10%, respectively) has steered us in the right direction, and produces the lowest inefficiency factors among this set. In terms of the precision of posterior mean estimates of the regression parameters $\boldsymbol{\beta}$, these are largely unaffected by the choice of $c$ and the level of precision essentially equals what we would have obtained with an iid sample of equal size. The numerical precisions of our simulation-based estimates of $E(\alpha_j|\boldsymbol{y})$ are rather low relative to those obtained under iid sampling, however, as the inefficiency factors with $c^2 = .5$ are near 20.[9]

Posterior means and standard deviations of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are also provided in the table below. As the reader can see, posterior means and standard deviations of the regression parameters are only slightly changed relative to those reported in Table 1. In addition, we find some evidence that higher education increases conditional log wage variability, while the other covariates do not appear to play a strong role in explaining variation in log wages.

Table 2: Heteroscedastic Wage Application

| Variable | $\boldsymbol{\beta}$ Parameters | | | | | $\boldsymbol{\alpha}$ Parameters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Post. Mean | Post. Std. | Ineff. Factors ($c^2$) .1 | .5 | 1 | Post. Mean | Post. Std. | Ineff. Factors ($c^2$) .1 | .5 | = 1 |
| Constant | 1.79 | .103 | 1.12 | 1.18 | 1.19 | -1.97 | .287 | 33.83 | 21.52 | 28.27 |
| EDUCATION | .045 | .007 | 1.09 | 1.05 | 1.02 | .047 | .019 | 34.92 | 21.56 | 30.80 |
| SCORE | .094 | .017 | 1.10 | 1.16 | 1.15 | -.009 | .048 | 34.00 | 21.59 | 30.45 |
| MOMED | .002 | .006 | 1.03 | 1.04 | 1.05 | -.012 | .016 | 30.58 | 19.97 | 28.37 |
| DADED | .007 | .005 | 1.01 | 1.02 | 1.00 | .012 | .013 | 32.27 | 22.07 | 28.57 |
| NUMSIBS | .004 | .007 | 1.01 | 1.00 | 1.00 | .007 | .019 | 31.20 | 20.16 | 28.46 |
| BA / HS GAP | 2.97 | .47 | | | | | | | | |
| Ph.D. / HS GAP | 6.77 | 1.22 | | | | | | | | |

Like the homoscedastic version of this model in section 2.1.2, we can also examine posterior distributions of various hourly wage gaps, though now we additionally account for heteroscedasticity. Specifically, we consider:

$$\Delta(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{x}_l, \boldsymbol{x}_h, \boldsymbol{z}_l, \boldsymbol{z}_h) = \exp\left(\boldsymbol{x}_h\beta + \frac{\exp(\boldsymbol{z}_h\alpha)}{2}\right) - \exp\left(\boldsymbol{x}_l\beta + \frac{\exp(\boldsymbol{z}_l\alpha)}{2}\right) \tag{32}$$

and obtain posterior means and standard deviations of $\Delta$ for the same choices that were made in

[9]This implies that the numerical standard error of the MCMC-based estimate of $\boldsymbol{\alpha}$ is (approximately) $4.64 \approx \sqrt{21.5}$ times as large as the numerical standard error that would have been attained under iid sampling. Said differently, in order to achieve the sample level of numerical precision for the estimated mean of $\boldsymbol{\alpha}$ that we would get with $m$ iid draws, we would need to run the sampler for $M \approx 21.5m$ iterations.

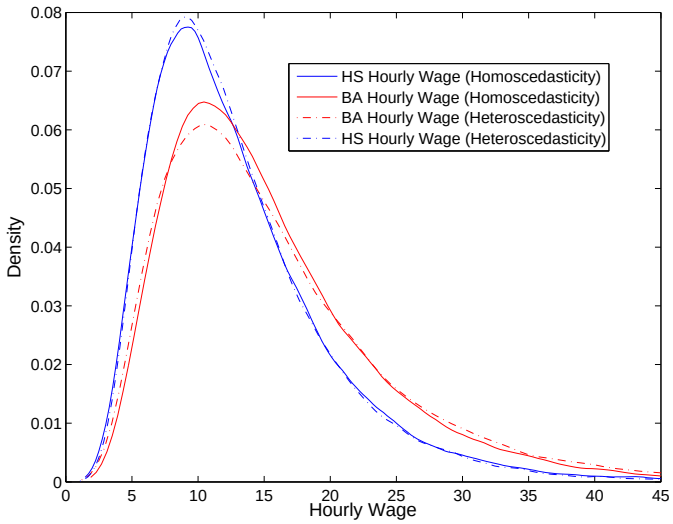


Figure 1B: Posterior Predictive Hourly Wage Densities

section 2.1.2. Table 2 reveals that point estimates of the wage gaps increase (relative to those of Table 1) when accounting for heteroscedasticity.

Lastly, we present in Figure 1B a plot of the hourly wage densities obtained within our heteroscedastic regression model along with those previously reported in Figure 1A under homoscedasticity. As the figure reveals, for high school graduates, little difference emerges between the posterior predictive densities. For the BA group, however, we begin to see the increased variance associated with the heteroscedastic predictive density, though these two curves are, again, rather similar. Analogous poverty probability calculations also remain nearly the same for the high school graduate group ( $\Pr(w_f < \$5 \mid \boldsymbol{y}, Ed = 12, X_{-Ed} = \overline{X}_{-Ed}) = .047$) and have slightly increased for the BA group ( $\Pr(w_f < \$5 \mid \boldsymbol{y}, Ed = 16, X_{-Ed} = \overline{X}_{-Ed}) = .032$).

## 2.3 A Linear Model with a Changepoint

The previous section illustrated the relative ease with which MCMC methods can be used to accommodate a departure from standard assumptions of the classical linear regression model. Other straightforward extensions include the imposition of inequality restrictions on the elements of $\boldsymbol{\beta}$ [e.g., (Geweke 1996b)] or analysis of multivariate linear outcomes, such as the Seemingly Unrelated Regressions (SUR) Model [e.g., (Zellner 1962; Percy 1992)]. We do not discuss these particular

extensions in this chapter, but turn our attention instead to an alternate generalization by considering a linear regression model with a single, unknown changepoint. The generalizations of such a model frequently appear in time series econometrics (Geweke and Terui 1993). Apart from simply keeping with our theme of generalizing the standard linear model, the methods described in this section also have considerable value in microeconomic applications - for example, in order to allow for jumps and nonlinearities in a regression function, or as a stepping stone toward understanding other related methods for nonparametric regression [e.g., (Smith and Kohn 1996)], as the resulting posterior simulators are highly similar to the one described here.

Let us switch notation slightly and suppose that a scalar outcome of interest, $y_t$, $t = 1, 2, \cdots, T$, can be expressed as

$$y_t|\boldsymbol{\alpha}, \boldsymbol{\theta}, \sigma^2, \lambda, \boldsymbol{X}_{1,(\lambda)}, \boldsymbol{X}_{2,(\lambda)} \sim \begin{cases} \mathcal{N}(\boldsymbol{x}_t\boldsymbol{\alpha}, \sigma^2) & \text{if } t \leq \lambda \\ \mathcal{N}(\boldsymbol{x}_t\boldsymbol{\theta}, \sigma^2) & \text{if } t > \lambda, \end{cases} \tag{33}$$

where $\boldsymbol{x}_t$ denotes a $1 \times k$ vector of characteristics at time $t$ and $\boldsymbol{X}_{j,(\lambda)}$, for $j = 1, 2$, assembles the covariate data for each "regime":

$$\boldsymbol{X}_{1,(\lambda)} \equiv \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_\lambda \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X}_{2,(\lambda)} \equiv \begin{bmatrix} \boldsymbol{x}_{\lambda+1} \\ \boldsymbol{x}_{\lambda+2} \\ \vdots \\ \boldsymbol{x}_T \end{bmatrix}. \tag{34}$$

The parameter $\lambda$ is a changepoint or breakpoint - for periods until $\lambda$, one regression is specified to generate $y$, and following $\lambda$, a new regression is specified to generate $y$. For simplicity, and with an eye toward our application in the following section, we suppose that the error variance in each regime is the same.

The priors for $\boldsymbol{\beta} = [\boldsymbol{\alpha}' \;\; \boldsymbol{\theta}']'$ and $\sigma^2$ are the same as those given in (2) and (3). The changepoint $\lambda$ is integer-valued and we choose a prior that specifies equal probability over each discrete value in the support. For example, one could specify

$$p(\lambda) \;\; = \;\; \frac{1}{T-1} I(\lambda \in \{1, 2, \ldots, T-1\}), \tag{35}$$

which imposes that at least one observation belongs to each regime.

### 2.3.1 Posterior Simulation

Stacking observations over $i$, we can write:

$$y = \boldsymbol{X}_{(\lambda)}\boldsymbol{\beta} + \epsilon, \quad \epsilon|\boldsymbol{X}, \lambda, \sigma^2 \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n), \tag{36}$$

where $\boldsymbol{X}_{(\lambda)}$ is a block diagonal matrix with $\boldsymbol{X}_{1,(\lambda)}$ in the upper block and $\boldsymbol{X}_{2,(\lambda)}$ in the lower block. The assumptions of our model imply:

$$p(\boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}) \propto p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)p(\lambda)\phi(\boldsymbol{y}|\boldsymbol{X}_{(\lambda)}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n). \tag{37}$$

We use the method of composition, as discussed in section 2.1.2, to directly generate samples from the joint posterior above, following the results of (Chin Choy and Broemeling 1980). This proceeds by drawing (in order) from $\lambda|\boldsymbol{y}$, $\sigma^2|\lambda, \boldsymbol{y}$ and $\boldsymbol{\beta}|\sigma^2, \lambda, \boldsymbol{y}$.

With respect to the first of these, one can show:

$$p(\lambda|\boldsymbol{y}) \propto p(\lambda)|\boldsymbol{D}_{(\lambda)}|^{-1/2}\left[b + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}_{(\lambda)}\boldsymbol{\mu}_{\boldsymbol{\beta}})'\boldsymbol{D}_{(\lambda)}^{-1}(\boldsymbol{y} - \boldsymbol{X}_{(\lambda)}\boldsymbol{\mu}_{\boldsymbol{\beta}})\right]^{-(n+a)/2} \tag{38}$$

with

$$\boldsymbol{D}_{(\lambda)} \equiv \boldsymbol{I}_n + \boldsymbol{X}_{(\lambda)}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{X}'_{(\lambda)}. \tag{39}$$

Since the prior for $\lambda$ is discrete-valued, one can calculate the (unnormalized) ordinates above for $\lambda \in \{1, 2, \ldots, T-1\}$, normalize these by dividing through by the sum of all such values, and then obtain a draw from the resulting discrete distribution.[10] The posterior conditional $\sigma^2|\lambda, \boldsymbol{y}$ is identical to that in (8), recognizing that $\boldsymbol{X}$ is now $\boldsymbol{X}_{(\lambda)}$ and must be re-calculated at each iteration of the algorithm. Similarly, $\boldsymbol{\beta}|\lambda, \sigma^2, \boldsymbol{y}$ is given in (5) and (6) where $\boldsymbol{X}$ is, again, replaced with $\boldsymbol{X}_{(\lambda)}$.

### 2.3.2 Example with U.S. Annual Temperature Data

In what follows we illustrate use of the changepoint model using a sample of annual U.S. temperature data. Specifically, we obtain information on annual temperatures in the United States over the period 1895-2006, providing $n = 112$ data points.

[10] To avoid calculating determinants and inverses of $n-$dimensional matrices in this step, note $\boldsymbol{D}_{(\lambda)}^{-1} = \boldsymbol{I}_n - \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{V}_{\boldsymbol{\beta}}{}^{-1}]^{-1}\boldsymbol{X}'$ and $|\boldsymbol{D}_{(\lambda)}| = |\boldsymbol{V}_{\boldsymbol{\beta}}||\boldsymbol{V}_{\boldsymbol{\beta}}{}^{-1} + \boldsymbol{X}'\boldsymbol{X}|$.

In providing this example we confess to know little (if anything) about the science of climate change. We do not introduce this example to either support or cast doubt on theories of global warming, but simply include it with the hope that the reader may appreciate the generality and usefulness of the model considered in this section as well as the relative ease with which simulation methods can be used to estimate its parameters.

We suppose that temperature patterns over this period may have a single break date and seek to learn about the location of this break as well as its magnitude. Although the simplicity of this model likely discredits it as an accurate descriptor of the evolution of U.S. temperature, it is worth noting that similar models with break points have been considered by others in the field, including (Ivanov and Evtimov 2009; Stockwell and Cox 2009) and that this issue has also been investigated by economists (e.g., Fomby and Vogelsang 2002) employing related models that potentially include breaks (Vogelsang and Franses 2005). We consider below a restricted version of the changepoint model discussed earlier in this section, tailored to our application, and specify:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 t + \beta_2 (t-\lambda)_+ + \epsilon_t, \quad t = 1, 2, \ldots, T \qquad (40) \\ &= \boldsymbol{x}_{t,\lambda}\boldsymbol{\beta} + \epsilon_t, \quad \epsilon_t | \boldsymbol{X}, \lambda, \sigma^2 \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \qquad (41) \end{aligned}$$

or stacked over $t$, we have, identical to (35),

$$\boldsymbol{y} = \boldsymbol{X}_{(\lambda)}\boldsymbol{\beta} + \epsilon, \quad \epsilon | \boldsymbol{X}, \lambda, \sigma^2 \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \qquad (42)$$

where

$$z_+ \equiv \max\{0, z\}, \quad \boldsymbol{x}_{t,\lambda} = [1 \;\; t \;\; (t-\lambda)_+], \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X}_{(\lambda)} \equiv \begin{bmatrix} \boldsymbol{x}_{1,\lambda} \\ \boldsymbol{x}_{2,\lambda} \\ \vdots \\ \boldsymbol{x}_{T,\lambda} \end{bmatrix}. \qquad (43)$$

This specification allows for different slopes before and after $\lambda$, but does not allow for a discrete jump in temperatures before and after the break. While such jumps may exist, we abstract from this possibility here and impose a smooth function relating year to expected temperature. Moreover, we do not consider the possibility of multiple breaks or potentially different variances across the regimes. Models of multiple breaks can be found, for example, in (Chib 1998; Koop and Potter 2007), among others. Finally, we restrict $\lambda \in \{3, 4, \ldots, T-3\}$, specify equal prior probability over each element of this set, and choose

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \sim \mathcal{N}\left[ \begin{pmatrix} 52 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} \right], \qquad (44)$$

$$\sigma^2 \sim IG(3, 1) \tag{45}$$

as our remaining priors. The sampling procedure of section 2.3.1 is employed to generate $10,000$ simulations from the joint posterior in (37), which are used to calculate the quantities in Figures 2 and 3.

Figure 2 plots the posterior mean of the regression function relating time to average temperature. To do this we calculate the conditional mean function $\boldsymbol{X}_{(\lambda)}\boldsymbol{\beta}$ for each simulation from the joint posterior. The collection of these functions are then averaged to obtain the posterior mean, which is presented in the figure along with the scatterplot of the raw data. Importantly, note that this approach accounts for uncertainty regarding the location of the changepoint and thus, unlike a classical method that would condition on a point estimate of the changepoint's location, is smooth and not necessarily "kinked." Again, it is worthwhile to emphasize how uncertainty in the location of the changepoint is easily accounted for when using our posterior simulations to calculate quantities of interest.

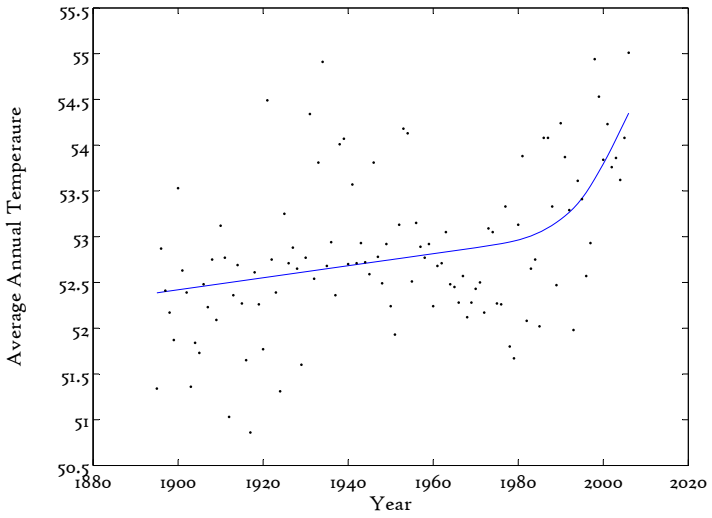


Figure 2: Raw temperature data and Posterior Mean

Figure 3 plots posterior simulations associated with the changepoint $\lambda$. The figure clearly shows an update of our uniform prior to a posterior suggesting that the changepoint has occurred since 1970 with a mode occurring in the late 1990's. This offers suggestive evidence that is broadly consistent with the global (or, at least U.S.) warming message.

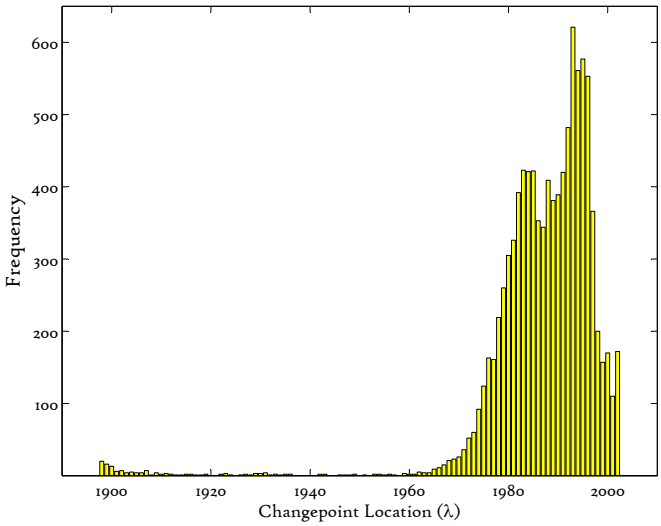


Figure 3: Posterior Changepoint Frequencies

## 2.4 Hierarchical Linear Models

Many data sets commonly used in applied work, including the National Longitudinal Survey of Youth (NLSY), Panel Study of Income Dynamics (PSID) and the Survey of Income and Program Participation (SIPP) are *longitudinal* in nature, tracking behaviors, outcomes and responses of a given set of individuals over a period of time. In a similar spirit, other data sets are characterized by (and the models employed should account for) *clustering* - where the outcomes of particular units are likely to be correlated with one another, given the sampling scheme or structure of the problem at hand. For example, wage outcomes are likely to be correlated across individuals within a given family, and student achievement scores are likely to be correlated across students within the same school. In this section we investigate hierarchical models for application to these types of data.

The literature on hierarchical models in Bayesian work is voluminous, originating with Lindley and Smith (1972) and Smith (1973) on the use of hierarchical priors. Recent Bayesian econometrics texts also highlight the importance of such models in applied work, (e.g., Geweke 2005, Section 3.1; Rossi, Allenby and McCulloch 2005, Chapter 5; Koop, Poirier and Tobias 2007, Chapter 12). (Geweke 2005) notes the connection between models with multilevel (i.e., hierarchical) priors

and latent variable specifications with conventional priors, while (Rossi, Allenby and McCulloch 2005) note that individual-level parameters in these types of analyses may be of primary interest and should not be simply regarded as nuisance parameters and marginalized out of the problem. Later in this chapter, following Geweke and Keane (2001), we unite a variety of popular univariate and multivariate nonlinear models within an encompassing hierarchial structure. In this section our goal is to briefly review how the linear regression framework of the previous sections can be generalized to accommodate the correlation patterns present in such data. We do so by considering the following basic hierarchical specification:

$$y_{it} = \beta_{0i} + \boldsymbol{x}_{it}\boldsymbol{\beta}_{1i} + \epsilon_{it}, \quad i = 1, 2, \ldots, n, \quad t = 1, 2, \ldots, T_i, \tag{46}$$

or

$$y_{it} = \boldsymbol{z}_{it}\boldsymbol{\beta_i} + \epsilon_{it} \tag{47}$$

with $\boldsymbol{z}_{it} = [1 \;\; \boldsymbol{x}_{it}]$ and $\boldsymbol{\beta}_i = [\beta_{0i} \;\; \boldsymbol{\beta}'_{1i}]'$. In the above, $i$ indexes outcomes for unit (or "individual") $i$ and $t$ is typically interpreted as time index. In this formulation we permit unit-level variation in both intercepts and slopes. This level of generality is seemingly reasonable yet somewhat uncommon, as much of the applied microeconometric literature finds it sufficient to impose homogeneity in slopes, (i.e., $\boldsymbol{\beta}_{1i} = \boldsymbol{\beta}_1$) and to permit variation only through unit-level intercepts. Such analyses follow as a restricted case of the analysis presented here. Finally, we also consider the general case of an unbalanced panel while simultaneously abstracting from related issues such as augmenting the model to include missing outcomes or covariate data.

We add structure to the model by supposing that the unit-level parameters are drawn from a common distribution. To this end, we specify a prior of the form:

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma} \overset{ind}{\sim} G(\boldsymbol{\beta}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{48}$$

where the common parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are of interest, and are objects we seek to learn about from the given data. The model is completed by adding priors for the common parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and a prior for the variance parameter $\sigma^2$.

This formulation of the model appears rather similar to classical random effects approaches where a distribution, like that in (48), is specified yet is typically designated as a "population distribution" characterizing variation in tastes. Unlike the Bayesian approach to this model, however, the unit-level parameters $\boldsymbol{\beta}_i$ from the frequentist perspective are commonly regarded as nuisance parameters

and are integrated out of the conditional likelihood. In many applications the $\boldsymbol{\beta}_i$ are, however, of primary interest and in such situations the Bayesian methodology seems particularly attractive, as the unit-level parameters are sampled in the course of implementing the posterior simulator. The adoption of the prior in (48) also imparts a form of shrinkage in the estimation of the $\boldsymbol{\beta}_i$ and helps to reduce concerns regarding overfitting in standard fixed effects modeling. Our representation of the model also assumes independence between the parameters of (48) and variables in $\boldsymbol{x}_{it}$, although such correlations can be modeled in generalizations of this specification.

### 2.4.1 Posterior Simulation in the Gaussian Hierarchical Model

To fix ideas, let us suppose that $\epsilon_{it}|\boldsymbol{Z}, \sigma^2 \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Extensions of the model to allow for unit-specific variance parameters or non-normality can also be handled, when desired. In addition, we employ specific priors of the forms:

$$\begin{aligned}
\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &\overset{iid}{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right), \quad i = 1, 2, \ldots, n && (49)\\
\boldsymbol{\beta} &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}\right) && (50)\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} &\sim W\left([\kappa \boldsymbol{R}]^{-1}, \kappa\right) && (51)\\
\sigma^2 &\sim IG\left(\frac{a}{2}, b\right), && (52)
\end{aligned}$$

with $W$ denoting a Wishart distribution.[11] The joint posterior distribution of all parameters of the model is then:

$$p\left(\{\boldsymbol{\beta}_i\}_{i=1}^n, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2|\boldsymbol{y}\right) \propto p(\sigma^2)p(\boldsymbol{\beta})p(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \prod_{i=1}^{n} \left[\phi\left(\boldsymbol{y}_i|\boldsymbol{Z}_i\boldsymbol{\beta}_i, \sigma^2 \boldsymbol{I}_{T_i}\right) \phi(\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})\right], \tag{53}$$

where $\boldsymbol{y}_i = [y_{i1} \quad y_{i2} \quad \cdots \quad y_{iT_i}]'$, $\boldsymbol{y} = [\ \boldsymbol{y}_1' \quad \boldsymbol{y}_2' \quad \cdots \quad \boldsymbol{y}_n'\ ]'$, $\boldsymbol{Z}_i$, in a similar manner, stacks the $\{\boldsymbol{z}_{it}\}_{t=1}^{T_i}$ into a $T_i \times k$ matrix, and $\boldsymbol{\beta}_i$ is a $k \times 1$ vector of parameters.

We generate samples from this joint posterior by employing a blocking step, as described in (Chib and Carlin 1999). That is, we propose a scheme to sample from the joint conditional $p(\{\boldsymbol{\beta}_i\}_{i=1}^n, \boldsymbol{\beta}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2, \boldsymbol{y})$ by first drawing from $p(\boldsymbol{\beta}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2, \boldsymbol{y})$ and then drawing (independently) from the series of conditional posteriors: $p(\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2, \boldsymbol{y})$. As such, the sampling of $\boldsymbol{\beta}$ and $\{\boldsymbol{\beta}_i\}$ makes uses of the marginal-conditional decomposition, takes place in a single block and, importantly, the sampling of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n$ must occur immediately following the sampling of $\boldsymbol{\beta}$, with

[11] We parameterize the Wishart as follows: $\boldsymbol{H} \sim W_k(\boldsymbol{A}, \nu) \Rightarrow p(\boldsymbol{H}) \propto |\boldsymbol{H}|^{(\nu-k-1)/2} \exp[-(1/2)tr(\boldsymbol{A}^{-1}\boldsymbol{H})]$ where $\boldsymbol{H}$ is a $k \times k$ matrix.

no other simulation steps intervening. The sampler is completed by drawing from the complete posterior conditionals for the variance parameter and inverse covariance matrix.

Noting

$$\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2 \overset{ind}{\sim} \mathcal{N}\left(\boldsymbol{Z}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{T_i} + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{Z}_i'\right), \quad i = 1, 2, \ldots, n, \tag{54}$$

it follows that

$$\boldsymbol{\beta}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{D}_{\boldsymbol{\beta}}\boldsymbol{d}_{\boldsymbol{\beta}}, \boldsymbol{D}_{\boldsymbol{\beta}}), \tag{55}$$

where

$$\boldsymbol{D}_{\boldsymbol{\beta}} \equiv \left[\left(\sum_i \boldsymbol{Z}_i'[\sigma^2\boldsymbol{I}_{T_i} + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{Z}_i']^{-1}\boldsymbol{Z}_i\right) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-\mathbf{1}}\right]^{-1} \tag{56}$$

and

$$\boldsymbol{d}_{\boldsymbol{\beta}} = \left(\sum_i \boldsymbol{Z}_i'[\sigma^2\boldsymbol{I}_{T_i} + \boldsymbol{Z}_i\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{Z}_i']^{-1}\boldsymbol{y}_i\right) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{\beta}}. \tag{57}$$

A sample from the conditional density $p(\{\boldsymbol{\beta}_i\}|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2, \boldsymbol{y})$ will then produce the desired draw from our joint conditional posterior distribution. Inspection of (53) shows that each of the $\boldsymbol{\beta}_i$ are conditionally independent *a posteriori*. Thus, we can independently sample from each conditional posterior, or specifically, we can independently draw, for $i = 1, 2, \ldots, n$:

$$\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \sigma^2, \boldsymbol{y} \overset{ind}{\sim} \mathcal{N}\Bigg(\left[\boldsymbol{Z}_i'\boldsymbol{Z}_i/\sigma^2 + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\left(\boldsymbol{Z}_i'\boldsymbol{y}_i/\sigma^2 + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right), \left[\boldsymbol{Z}_i'\boldsymbol{Z}_i/\sigma^2 + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\Bigg). \tag{58}$$

Finally, letting $T = \sum_{i=1}^n T_i$ we obtain:

$$\sigma^2|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_n, \boldsymbol{y} \sim IG\left(\frac{T+a}{2}, \left[b + \frac{1}{2}\sum_{i=1}^n\sum_{t=1}^{T_i}(y_{it} - \boldsymbol{z}_{it}\boldsymbol{\beta}_i)^2\right]\right) \tag{59}$$

and the conditional posterior density for the inverse covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-\mathbf{1}}$ is

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}|\boldsymbol{\beta}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_n, \boldsymbol{y} \sim W\left(\left[\sum_{i=1}^n(\boldsymbol{\beta}_i - \boldsymbol{\beta})(\boldsymbol{\beta}_i - \boldsymbol{\beta})' + \kappa\boldsymbol{R}\right]^{-1}, n + \kappa\right). \tag{60}$$

A posterior simulator for this normal hierarchical linear model proceeds by successively sampling from (55), (58), (59) and (60).

### 2.4.2 Two Applications of Hierarchical Linear Modeling

To show how hierarchical models can be estimated in practice we provide two illustrative examples. The first application has become something of a classic in the MCMC literature (though the application itself is, perhaps, rather unappealing), making use of the data set employed in the seminal study of (Gelfand et al. 1990) that helped to illuminate the benefits afforded by Gibbs sampling in empirical work. The second application, hopefully more engaging, involves the impact of class size on student achievement.

#### Application #1

In our initial application, each of $n = 30$ rats are weighed at five different points in time, specifically 8, 15, 22, 29 and 36 days since birth. The outcome, $y_{it}$, denotes the weight of rat $i$ in grams at date $t$, while $x_{it}$ simply denotes the time of measurement, and as such, $x_{it} = x_{jt}$ for all $i, j$.

For our priors, we set

$$\boldsymbol{\mu_\beta} = \begin{bmatrix} 100 \\ 15 \end{bmatrix}, \quad \boldsymbol{V_\beta} = \begin{bmatrix} 40^2 & 0 \\ 0 & 100 \end{bmatrix}, \quad a = 6, \quad b = 40, \quad \kappa = 5 \quad \text{and} \quad R = \begin{bmatrix} 100 & 0 \\ 0 & .25 \end{bmatrix}. \tag{61}$$

Posterior means and standard deviations for a selection of parameters are provided in Table 3 below.

Table 3: Posterior Quantities for a Selection of Parameters

| Parameter | Post Mean | Post Std. |
|---|---|---|
| $\beta_{Intercept}$ | 106.6 | 2.33 |
| $\beta_{Slope}$ | 6.18 | .108 |
| $\boldsymbol{\Sigma_\beta}(1,1)$ | 124.7 | 42.00 |
| $\boldsymbol{\Sigma_\beta}(2,2)$ | .277 | .087 |
| $\boldsymbol{\Sigma_\beta}(1,2)/\sqrt{\boldsymbol{\Sigma_\beta}(1,1) \times \boldsymbol{\Sigma_\beta}(2,2)}$ | -.126 | .210 |
| $\beta_{0,5}$ | 90.73 | 5.50 |
| $\beta_{0,30}$ | 106.62 | 5.15 |
| $\beta_{1,5}$ | 6.43 | .229 |
| $\beta_{1,30}$ | 6.13 | .214 |

The results suggest that the birth weight of an "average" rat is about 106.6 grams, and the average weight gained per day is about 6.2 grams. The $5^{th}$ entry of the table gives the correlation between

the unit-level intercept and slope parameters. The posterior mean of this correlation is negative, suggesting a degree of "catching up"- rats that are large at birth tend to have lower growth rates than rats that are comparably small at birth. The last four entries of the table provide posterior means and standard deviations of parameters for the $5^{th}$ and $30^{th}$ rats. The results here are consistent with the general pattern in the "population": rat 5 is smaller at birth than rat 30, but also has a higher growth rate.

It is also worth noting that our posterior simulator produces draws for each $\boldsymbol{\beta}_i$ and therefore interesting quantities involving comparisons of unit-level parameters can be easily calculated. For example, we can state: $\Pr(\beta_{1,5} > \beta_{1,30}|\boldsymbol{y}) \approx .84$, suggesting reasonably strong evidence that the $5^{th}$ rat possesses a faster rate of growth than the $30^{th}$. Quantities like these can be quite useful in practice, particularly when using a hierarchical model for other, more interesting pursuits. For example, (Geweke, Gowrisankaran and Town 2003) use a hierarchical model and unit-level parameter estimates to evaluate and rank the performances of hospitals while (Aitken and Longford 1986; Laird 1989; Li and Tobias 2005) have used them to compare the performances of schools. Stochastic frontier models [e.g., (Koop et al. 1997; Koop and Steel 2001)] also share a very similar structure and goal, with one-sided distributions for unit-level parameters commonly employed to gauge the "efficiency" of cross-sectional units. In these applications, posterior simulations of the unit-level parameters are quite valuable and can be used to address interesting and relevant economic questions.

**Application #2**

In our second application we follow (Krueger 1998; Krueger and Whitmore 2001) and apply our model to analyze data from Project STAR (Student/Teacher Achievement Ratio). Project STAR was an experiment in Tennessee that randomly assigned students to one of three types of classes - small class, regular size class, and regular size class with a teacher's aide (regular/aide class). In order to be eligible for participation in the experiment, each school had to be large enough to have at least three classes per grade, thus enabling all three types of classes to be represented in every school. The panel used in our application is unbalanced, as some schools have more than three classes per grade (though all have at least three), and moreover, the number of students within a given class type is not always constant in the data (for example, some small classes have 15 students while others have 16). The dependent variable we specify is a measure of student achievement and, specifically, is the average of a reading percentile score and math percentile score of a Project STAR

student. There are two treatment variables - a dummy variable indicating whether a student is assigned to a small class and another indicating assignment to a regular/aide class. The default category, therefore, is assignment to regular class.

The Project STAR data we use contains 79 participating schools with a total of 5,726 students who entered the project during kindergarten. The panel is unbalanced, as each school is not represented by the same number of students. We focus on the achievement measure taken at the end of the kindergarten year and consider heterogeneity of treatment impacts across schools. Therefore, in this application of the model in (47), $i$ denotes the school and $t$ no longer represents a time index but, instead, denotes the student within a school.

As one can see from the estimation results in Table 4, being in a small class is associated with an expected increase of 5.48 percentile points in the average test score. Furthermore, being assigned to a regular sized class with a teacher's aide does not appear to provide any large improvement on average over assignment to a regular classroom. Importantly, the effects of class size reductions appear to vary greatly across schools, as reflected in the posterior mean of the square root of the (2,2) element of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ in (49), which is 10.6. As (49) suggests with these values, and our posterior simulations directly reveal, several schools even show a negative small class effect.

We also note that the correlation among elements of $\boldsymbol{\beta}_i$ are quite strong. The positive correlation between the small class and regular / aide parameters suggests that schools most inclined to benefit from smaller classes are also the ones with relatively large benefits to adding an aide to regular sized classrooms. We also see a rather strong, negative correlation between the school specific intercepts (regular-sized class parameters) and small and regular / aide parameters. One interpretation of this result, which seems to be sensible, is that schools whose students score low (high) in regular-sized classes are the ones that benefit the most (least) from class size reductions. Adoption of the hierarchical specification to this data reveals not only a sizeable amount of heterogeneity across schools, but also sheds light on the schools that would be most impacted by changes to class size.

Table 4: Posterior means, standard deviations and probabilities of being positive of the parameters

| Parameter | Post. Mean | Post. Std. | $\Pr(\cdot > 0 \mid \boldsymbol{y})$ |
|---|---|---|---|
| $\beta_0$ (intercept) | 51 | 1.82 | 1 |
| $\beta_1$ (small class) | 5.48 | 1.44 | 1 |
| $\beta_2$ (regular/aide class) | 0.311 | 1.26 | 0.596 |
| $\sqrt{\sigma^2}$ | 22.9 | 0.221 | 1 |
| $\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(1,1)}$ | 15.2 | 1.32 | 1 |
| $\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(2,2)}$ | 10.6 | 1.24 | 1 |
| $\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(3,3)}$ | 8.93 | 1.14 | 1 |
| $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(1,2)/\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(1,1) \times \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(2,2)}$ | -0.454 | 0.111 | 0.000125 |
| $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(1,3)/\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(1,1) \times \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(3,3)}$ | -0.483 | 0.111 | 0.000125 |
| $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(2,3)/\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(2,2) \times \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(3,3)}$ | 0.548 | 0.118 | 1 |

The foregoing examples and discussion were intended to introduce rather than fully describe Bayesian approaches to hierarchical linear models. Such methods can be easily extended to non-linear models, including the binary and multiple choice models we will discuss in later sections. Moreover, recent work has sought to relax many of the distributional assumptions made, particularly in modeling the unit-level parameters (see, e.g., (Rossi and Allenby 2010) of this volume). Finally, we note that time-invariant covariates can be included in the middle stage of the hierarchy, or the modeling of $\boldsymbol{\beta}_i$, and more general error structures can be considered, for example, allowing for autocorrelation among the $\epsilon_{it}$ [see, e.g., (Chib and Jeliazkov 2006) for handling this and other issues in a more complex dynamic binary choice setting].

## 2.5 Endogeneity in Linear Models

The problem of endogeneity plays a central role in the practice of microeconometrics. While most textbook discussions of and applications involving endogeneity are classical in nature, centered upon or employing IV, 2SLS or other approaches for estimation, studies such as (Drèze 1976; Drèze and Richard 1983; Geweke 1996a; Kleibergen and Zivot 2003; Hoogerheide, Kleibergen and van Dijk 2007; Sims 2007; Conley et al 2008) mark important Bayesian advances to this literature. The importance of this issue is also suggested by the rather prominent and detailed treatment it receives in many current Bayesian textbooks (Lancaster 2004: Chapter 8; Rossi, Allenby and McCulloch 2005: Chapter 7; Koop, Poirier and Tobias 2007: Chapter 14) and even elsewhere in this volume [e.g., (Rossi and Allenby 2010; Sims 2010)]. Furthermore, numerous applications

have been tackled from a Bayesian point of view, often highlighting the ease with which MCMC methods can be adapted to deal with endogeneity problems in many different kinds of models [e.g., (Li 1998; Geweke, Gowrisankaran and Town 2003; Munkin and Trivedi 2003; Deb, Munkin and Trivedi 2006b; Kline and Tobias 2008; Chib et al. 2009)].

We frame our discussion of endogeneity within the context of a linear regression model, where one of the right-hand side variables is endogenous. While this is somewhat restrictive, it is not terribly so, as simple generalizations can accommodate higher dimension endogeneity problems. Moreover, a recent study by (Chernozhukov and Hansen 2008) suggests that this is the modal model entertained in the literature[12] and thus serves as a natural starting point for our analysis.

Consider the model:

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 x_i + \boldsymbol{\alpha}_2 \boldsymbol{w}_i + \epsilon_i \quad (62) \\ x_i &= \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{z}_i + u_i, \quad (63) \end{aligned}$$

where

$$\left[ \begin{array}{c} \epsilon_i \\ u_i \end{array} \right] \Bigg| \boldsymbol{W}, \boldsymbol{Z} \overset{iid}{\sim} \mathcal{N}\left[ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{array} \right) \right] \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

The exogenous variables $\boldsymbol{w}_i$ are covariates entering the $y$-outcome equation while $\boldsymbol{z}_i$ enter the reduced form equation for $x$. As shown below, there can be (and always is) overlap between these two sets of variables, yet identification will require the appearance of at least one column of $\boldsymbol{Z}$ that is not contained in $\boldsymbol{W}$.

Letting $\boldsymbol{\theta}$ denote all the parameters of the model, we can write

$$p(\epsilon_i, u_i|\boldsymbol{\theta}) = p(\epsilon_i|u_i, \boldsymbol{\theta}) p(u_i|\boldsymbol{\theta}). \quad (64)$$

Noting that the Jacobian of the transformation from $(\epsilon_i, u_i)$ to $(y_i, x_i)$ is unity, we obtain

$$\begin{aligned} p(y_i, x_i|\boldsymbol{\theta}) &= \phi\left( y_i \left| \alpha_0 + \alpha_1 x_i + \boldsymbol{\alpha}_2 \boldsymbol{w}_i + \frac{\sigma_{\epsilon u}}{\sigma_u^2}(x_i - \beta_0 - \boldsymbol{\beta}_1 \boldsymbol{z}_i), \sigma_\epsilon^2(1 - \rho_{\epsilon u}^2) \right. \right) \quad (65) \\ &\quad \times \phi(x_i|\beta_0 + \boldsymbol{\beta}_1 \boldsymbol{z}_i, \sigma_u^2), \end{aligned}$$

where $\rho_{\epsilon u} \equiv \sigma_{\epsilon u}/[\sigma_\epsilon \sigma_u]$.

[12](Chernozhukov and Hansen 2008) find 108 articles in AER/QJE/JPE over the period 1999-2004 that employ linear IV, and 91 of these report results with one endogenous right-hand side variable.

It is useful to pause and discuss identification in the context of this system of equations. To this end, first consider the case where the set of exogenous covariates are common to both equations, i.e., $\boldsymbol{z}_i = \boldsymbol{w}_i$. In this case, (65) becomes:

$$\begin{aligned} p(y_i, x_i|\boldsymbol{\theta}) &= \phi\left(y_i \left| \left[\alpha_0 - \beta_0 \frac{\sigma_{\epsilon u}}{\sigma_u^2}\right] + \left[\alpha_1 + \frac{\sigma_{\epsilon u}}{\sigma_u^2}\right] x_i + \left[\boldsymbol{\alpha}_2 - \boldsymbol{\beta}_1 \frac{\sigma_{\epsilon u}}{\sigma_u^2}\right] \boldsymbol{w}_i, \sigma_\epsilon^2(1-\rho_{\epsilon u}^2)\right.\right) \\ &\quad \times \phi(x_i|\beta_0 + \boldsymbol{\beta}_1 \boldsymbol{z}_i, \sigma_u^2). \end{aligned} \tag{66}$$

Some quick accounting, then, shows that the likelihood is a function of just 7 (blocks of) parameters:

$$\beta_0,\ \boldsymbol{\beta}_1,\ \sigma_u^2,\ \psi_0 = [\alpha_0 - \beta_0 \frac{\sigma_{\epsilon u}}{\sigma_u^2}],\ \psi_1 = [\alpha_1 + \frac{\sigma_{\epsilon u}}{\sigma_u^2}],\ \boldsymbol{\psi}_2 = [\boldsymbol{\alpha}_2 - \boldsymbol{\beta}_1 \frac{\sigma_{\epsilon u}}{\sigma_u^2}] \text{ and } \psi_3 = \sigma_\epsilon^2(1-\rho_{\epsilon u}^2), \tag{67}$$

whereas we seek to recover the 8 "structural" parameters of $\boldsymbol{\theta}$:

$$\alpha_0, \quad \alpha_1, \quad \boldsymbol{\alpha}_2, \quad \beta_0, \quad \boldsymbol{\beta}_1, \quad \sigma_u^2, \quad \sigma_\epsilon^2, \text{ and } \sigma_{\epsilon u}. \tag{68}$$

As a result, the quantities in (67) are identified by the likelihood whereas the full set of structural parameters in (68) are not identifiable. Importantly, note that the "causal effect" $\alpha_1$ - the object that garners most attention in practice - is among the parameters that are not identifiable when the set of covariates appearing in (62) and (63) are the same.

While several assumptions regarding the model can be used to achieve identification in a specification like (62)-(63),[13] the most common one is to assume the presence of at least one element of $\boldsymbol{Z}$ that is not contained in $\boldsymbol{W}$. That is, a careful understanding of the problem at hand leads to the determination of a set of variables (or "instruments") in $\boldsymbol{z}_i$ that are not contained in $\boldsymbol{w}_i$ and can be exploited for purposes of identification and estimation. Indeed, (65) shows how such exclusion restrictions can be exploited for identification purposes: The parameter $\boldsymbol{\beta}_1$ is identifiable from the marginal (reduced form) density of $x_i$, and the coefficient on the elements of $\boldsymbol{z}$ *not contained in* $\boldsymbol{w}$ in the conditional density $y|x$ becomes $-[\sigma_{\epsilon u}/\sigma_u^2]\boldsymbol{\beta}_1$. Together, these two pieces of information enable identification of the ratio $\sigma_{\epsilon u}/\sigma_u^2$, which is attributable to the role of unobserved confounding.

[13] The "kitchen sink" approach represents such an alternative, where a host of covariates are included in (62), and the rich set of employed observables is argued to be sufficient to render $\epsilon$ and $x$ uncorrelated, or at least approximately so. Common sense or an inspection of (65) reveals that (62) can then be estimated as a single equation when $\sigma_{\epsilon u} = 0$, leading to a recursive system. [The implications of this assumption were first noted by the Cowles Commission, with (Christ 1994) offering a nice overview of their early econometric contributions in this (and other) settings. The restriction $\sigma_{\epsilon u} = 0$ and its implications remain relevant today, e.g., in identification of VAR models]. While this identification strategy typically does not sit well with the majority of practitioners, who have come to view IV as *the* solution to the identification problem and *the* device enabling the extraction of causal impacts, it does occasionally find a sympathetic referee (or two or three). (Dearden, Ferri and Meghir 2006) is a prominent, well-crafted example. Other alternatives for identification, even less widely used in practice, include the imposition of cross-equation parameter restrictions.

Once this ratio is known, the causal effect $\alpha_1$ as well as the remaining parameters of the model clearly become identifiable, as is evident from (65). This simple argument illustrates the value of instruments as vehicles for identification, and also suggests potential difficulties in separating $\alpha_1$ from $\sigma_{\epsilon u}/\sigma_u^2$ when the instruments are poor (weak). We will revisit this issue in the analysis of section 2.5.3.

### 2.5.1 Posterior Simulation

Stack the variables into vectors and matrices by writing:

$$\begin{bmatrix} y_i \\ x_i \end{bmatrix} = \begin{bmatrix} 1 & x_i & \boldsymbol{w}_i & 0 & \mathbf{0} \\ 0 & 0 & \mathbf{0} & 1 & \boldsymbol{z}_i \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \boldsymbol{\alpha}_2 \\ \beta_0 \\ \boldsymbol{\beta}_1 \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \tag{69}$$

or

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_i, \tag{70}$$

with $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{X}}_i$, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\epsilon}}_i$ defined in the obvious ways. Furthermore, suppose we continue to employ priors of the forms:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}) \tag{71}$$

$$\boldsymbol{\Sigma}^{-1} \sim W\left[(\kappa \boldsymbol{R})^{-1}, \kappa\right]. \tag{72}$$

With this done, posterior simulation in our linear model with an endogeneity problem follows in a straightforward way. In particular, a simple two-block Gibbs algorithm can be employed that iteratively samples from the following two conditional posterior distributions:

$$\boldsymbol{\beta}|\boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{D}_{\boldsymbol{\beta}} \boldsymbol{d}_{\boldsymbol{\beta}}, \boldsymbol{D}_{\boldsymbol{\beta}}), \tag{73}$$

where

$$\boldsymbol{D}_{\boldsymbol{\beta}} = \left( \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i{}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i \right)^{-1}, \quad \boldsymbol{d}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^{n} \left( \tilde{\boldsymbol{X}}_i{}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{y}}_i \right) \tag{74}$$

and

$$\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{x} \sim W\left( \left[ \sum_{i=1}^{n} \tilde{\boldsymbol{\epsilon}}_i \tilde{\boldsymbol{\epsilon}}_i{}' + \kappa \boldsymbol{R} \right]^{-1}, n + \kappa \right). \tag{75}$$

A posterior simulator for this model proceeds by sampling from (73) and (75). The reader may note, and perhaps be puzzled by, the connection of the above sampling scheme to what one would obtain from a standard SUR analysis where no endogenous variables appear as right-hand side covariates. That is, one may rightfully ask: why does the simulator for this model with an endogeneity problem reduce to essentially the same simulator for a bivariate SUR without an endogeneity concern? The connection here critically relies on the Jacobian of transformation being equal to one; such a result would not be obtained for a purely simultaneous equations model that is not triangular.

### 2.5.2 Application: BMI Data

To illustrate how such methods are applied in practice, we consider a restricted application of the model of (Kline and Tobias 2008), who employ data from the British Cohort study, a longitudinal survey of the cohort of all people born in Great Britain between April 5 and April 11, 1970. The data set contains the usual set of demographic variables and wage outcomes, along with heights and weights of the survey participants. Furthermore, one of the survey waves also obtains information on the heights and weights of the respondent's parents. These variables enable us to calculate the respondent's Body Mass Index (BMI), defined as weight (in kilograms) divided by the square of height (in meters), as well as the BMI of his/her parents.

We use hourly wages as our outcome of interest, which are observed when the respondents are approximately 29 years of age. Furthermore, we consider the analysis for males only. Our application is designed with the primary intent to estimate the "causal" impact of BMI on (log) hourly wages, a question that has received rather significant attention within the labor literature. As additional controls, we include family income (when the respondent was 10 years of age), and whether or not the respondent has a college degree. The constructed parental BMI variables, denoted MomBMI and DadBMI are used as our instruments (exclusion restrictions) for child BMI. Our final sample consists of $n = 2,561$ observations.[14]

Coefficient posterior means and standard deviations are reported in Table 5 below, setting $\boldsymbol{\mu_\beta} = 0$ $\boldsymbol{V_\beta} = 10 I_k$, $\kappa = 5$, and $\boldsymbol{R} = I_2$ as our prior hyperparameters.

[14]This data set is, unfortunately, restricted access and therefore can not be made available on the website accompanying this chapter.

Table 5: Parameter Posterior Means and Standard Deviations from BMI Application

| Log Wage Equation | | |
|---|---|---|
| Variable | Post. Mean | Post Std. |
| Constant | 2.96 | .215 |
| BMI | -.041 | .008 |
| FamInc | .001 | .0001 |
| Degree | .244 | .024 |
| **BMI Equation** | | |
| Variable | Post. Mean | Post Std. |
| Constant | 15.10 | .670 |
| FamInc | .0003 | .001 |
| Degree | -.599 | .172 |
| MomBMI | .176 | .018 |
| DadBMI | .259 | .020 |
| **Other Parameters** | | |
| Variable | Post. Mean | Post Std. |
| $\sigma^2_\epsilon$ | .200 | .011 |
| $\sigma^2_u$ | 11.22 | .305 |
| $\rho_{\epsilon u}$ | .342 | .057 |

The results presented in Table 5 are generally consistent with the findings of (Kline and Tobias 2008). First, our instruments are important variables in explaining variation in BMI, clearly suggesting that higher parental BMI leads to higher child BMI, as our simulations would show $\Pr(\beta_{1,MomBMI} > 0, \beta_{1,DadBMI} > 0 | \boldsymbol{y}, \boldsymbol{x}) = 1$. Furthermore, the relationship between BMI and wages is negative, as a one point increase in BMI leads to an approximate 4.1 percent reduction in hourly wages. Finally, the role of unobservables is also important and strong evidence is provided that $\rho_{\epsilon u} > 0$. (Kline and Tobias 2008) argue that this is consistent with a tradeoff between work effort and health - individuals unobservably dedicated to their job (thus, presumably, earning higher wages) do so at the expense of investments in health (regular exercise, maintaining a well-balanced diet, etc.), leading to a positive correlation between $\epsilon$ and $u$.

### 2.5.3 A Few Comments on Weak Instruments

As the reader may be aware, there has been a great deal of attention given recently to the problem of weak / many instruments [an excellent recent treatment of this issue from the Bayesian perspective is offered by (Hoogerheide, Kaashoek and van Dijk 2007)]. Much interest in this issue developed

subsequent to (Bound, Jaeger and Baker's 1995) critique of the study by (Angrist and Krueger 1991), the latter of which paved the way for a host of instrumental variable-based studies and accounted for an increased emphasis on natural experiments in economics. While it is not our intent to review the history of this literature, or to document any of the recent frequentist developments within it, we do wish to note that Bayesian approaches are not immune to the weak instruments "problem" and that the presence of such instruments has potentially important consequences for posterior inference.

We demonstrate this point via two generated data experiments. Specifically we first generate $n = 1,000$ observations from a simplified version of (62)-(63):[15]

$$y_i = \alpha_0 + \alpha_1 x_i + \epsilon_i \tag{76}$$

$$x_i = \beta_0 + \beta_1 z_i + u_i \tag{77}$$

where $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $\sigma_\epsilon = \sigma_u = 1$ and $\rho_{xy} = .5$ to introduce a reasonable degree of unobserved confounding.[16] We generate two different data sets under two different values of $\beta_1$, setting $\beta_1 = .01$ or $\beta_1 = 1$, to investigate what happens to aspects of the joint posterior when the instrument is "weak" or "strong," respectively. To justify these labels, note that the population $R$-squared for the reduced form (marginal density for $x$) equation in (77), given that $\text{Var}(z) = 1$ and $\sigma_\epsilon = \sigma_u = 1$, is

$$\frac{\beta_1^2}{\beta_1^2 + 1}. \tag{78}$$

Thus, when the instrument is weak in this design, the population R-squared is (approximately) .0001[17] while the strong instrument gives a population $R$-squared value of 1/2.

Results from this experiment are provided in Figures 4-5. Before discussing these details, we first note that inference regarding the "total effect" of $x$ on $y$, obtained from the conditional density $y|x$ in (66) as $\alpha_1 + \sigma_{\epsilon u}/\sigma_u^2$, is not affected by the quality of the instrument. To illustrate, we note that the posterior mean (and standard deviation) for this total effect parameter are 1.23 (.027) and 1.26 (.028) for the weak and strong instrument cases, respectively. Thus, with an equal sample

[15] The intention here is not to reproduce sampling properties of the procedure by generating numerous data sets of the same size. The points we seek to make can be illustrated with one realization of data from this (adequately large) sample.

[16] A complete description of the parameters used to generate the data is given in Table 6.

[17] Though this seems small, it is very similar to the R-squared obtained from a regression of educational attainment on quarter of birth using the (Angrist and Krueger 1991) data.

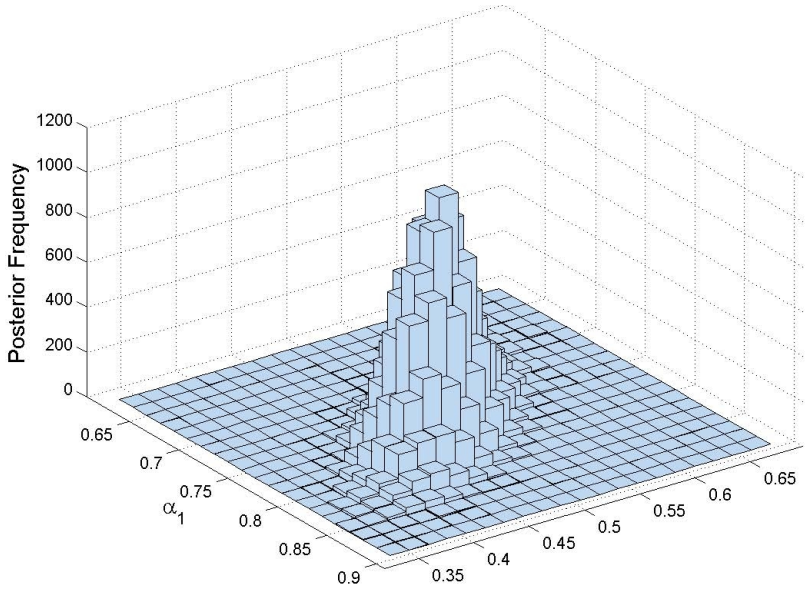


Figure 4: Joint Posterior, Strong Instrument

size, our ability to assess the overall or total impact of $x$ on $y$ is independent of the strength of the instrument.

Figures 4 and 5 show how the strength of the instrument does, however, aid in separating the "causal" effect $\alpha_1$ from the effect attributable to unobserved confounding, $\sigma_{\epsilon u}/\sigma_u^2$. In Figure 4 results for the strong instrument case are presented. Importantly, note how this joint posterior has nearly collapsed around the parameter values $\alpha_1 = .75$ [left-side axis] and $\sigma_{\epsilon u}/\sigma_u^2 = .5$ [right-side axis] that were used to generate the data.

Figure 5 presents a similar set of results for the weak instruments case. First, we note the very diffuse axes over which this posterior surface is plotted, which are vastly more spread out than those of Figure 4. Second, we observe the ridge in the likelihood surface along the line (approximately) given by $\alpha_1 + \sigma_{\epsilon u}/\sigma_u^2 = 1.25$. In the presence of weak instruments, we fare equally well in identifying the total effect of $x$ on $y$, but our ability to separate this impact into a "causal" effect and an effect arising from unobserved confounding suffers substantially.

Our final generated data experiment illustrates the role of the prior when weak instruments are present. For this purpose we generate 2 different data sets, each with $n = 200$ observations. The

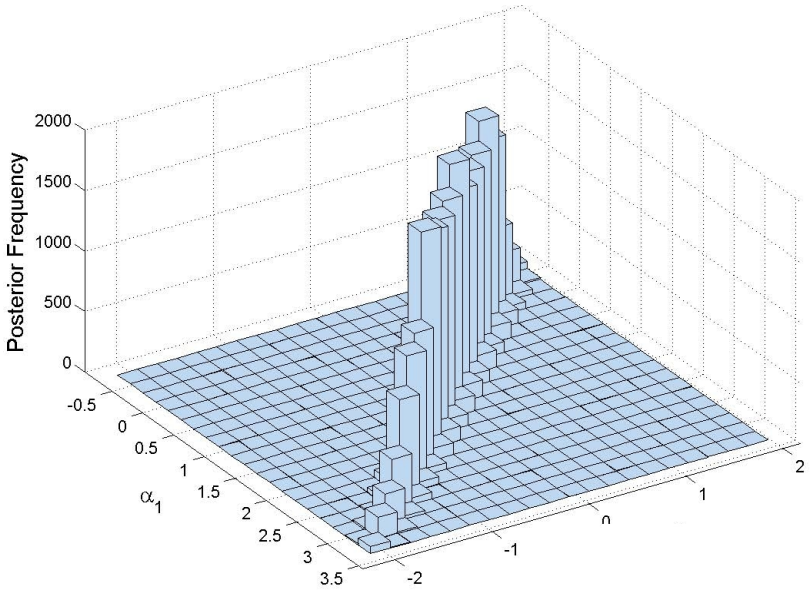


Figure 5: Joint Posterior, Weak Instrument

process used to generate each data set follows the same design as the one previously employed in this section, leading to separate data sets created for analysis of the "weak" and "strong" IV cases. Unlike the previous analysis, however, we also consider two different priors and employ both in the estimation of the weak IV and strong IV generated data sets. The hyperparameters $\kappa = 3$, $\boldsymbol{R} = \boldsymbol{I}_2$ and $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{I}_4$ are constant in all experiments. However, for one prior (denoted P1), we set the prior mean of $\boldsymbol{\beta}$ to be the zero vector: $\boldsymbol{\mu}_{\boldsymbol{\beta}} = [0\ 0\ 0\ 0]'$ and in the second (denoted P2), we set $\boldsymbol{\mu}_{\boldsymbol{\beta}} = [0\ \ -3\ 0\ 0]'$. Thus, the two priors alter the mean of $\alpha_1$, with P1 centering the $\alpha_1$ prior over zero, and P2 centering it over -3. Our goal is to examine how this change in prior impacts our posterior results, and to assess its differential impact across the weak and strong IV data sets in particular. The fact that the prior should matter when the IV is weak was already suggested by Figure 5.

Table 6: Posterior Means From Generated Data Experiment

| | True Value | Strong, P1 | Strong, P2 | Weak, P1 | Weak, P2 |
|---|---|---|---|---|---|
| $\alpha_0$ | 2.00 | 2.02 | 2.00 | 1.59 | .553 |
| $\alpha_1$ | .750 | .728 | .710 | .338 | -.725 |
| $\beta_0$ | -1.00 | -1.07 | -1.07 | -.984 | -.982 |
| $\beta_1$ | 1.00 or .010 | .968 | .945 | -.016 | .004 |
| $\sigma_\epsilon^2$ | 1.00 | 1.05 | 1.08 | 1.79 | 4.53 |
| $\sigma_u^2$ | 1.00 | 1.07 | 1.08 | .912 | .913 |
| $\rho_{\epsilon u}$ | .500 | .500 | .514 | .601 | .886 |

As shown in Table 6, with strong instruments, parameter posterior means are quite close to the values of the parameters used to generate the data. As expected, the shift in the prior mean of $\alpha_1$ does lower the posterior mean of $\alpha_1$ when the instruments are strong, though not tremendously so. On the other hand, when the instruments are weak, posterior means do not closely match the values of the parameters used to generate the data and the prior has a sizeable impact on our calculations, as the posterior mean of $\alpha_1$ even changes sign and remains far away from .75.

Although not shown in the tables above, the mixing of the posterior simulations is also strongly affected by the quality of the instruments; in the strong IV case, inefficiency factors [see (Chib 2010) of this volume for additional information on these factors] were less than 4 for all parameters, and approximately unity for some, suggesting that our Gibbs calculations are essentially of the same quality as those that would be obtained under iid sampling from the posterior. Furthermore, in the weak instrument case, inefficiency factors for regression and variance parameters of (76) as well as $\rho_{\epsilon u}$ were in excess of 1,000, suggesting that more than 1,000$n$ Gibbs simulations would be required to achieve the numerical accuracy afforded by $n$ iid posterior draws.

# 3 Nonlinear Hierarchical Models

In this section we review posterior simulation in several univariate nonlinear models that are linear in suitably defined latent data [(Geweke and Keane 2001; Geweke 2005) take up a wide range of latent variable models and offer a similar type of unifying treatment]. The structure of and resulting posterior simulators for many simple and popular models of this type can be united with a hierarchical structure and we present such a general description here. Let $\boldsymbol{\theta} = [\boldsymbol{\beta}' \ \boldsymbol{\alpha}' \ \sigma^2]'$ and

consider a representative *univariate latent linear model*, expressed as

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\sigma^2) && (79)\\
z_i|\boldsymbol{X},\boldsymbol{\theta} &\overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta},\sigma^2), \quad i=1,2,\ldots,n && (80)\\
y_i &= g(z_i,\boldsymbol{\alpha}), \quad i=1,2,\ldots,n. && (81)
\end{aligned}
$$

Equation (79) involves a prior for the model's parameters $\boldsymbol{\theta}$, and throughout we specify prior independence among these parameters. Equation (80) describes the generation of a latent variable, only partially observed by the econometrician. To fix ideas we specify that this latent variable is conditionally normally distributed throughout our discussion. The normality assumption can be relaxed (and frequently is in practice), and below we will reference sources that generalize this assumption in the context of particular models. Finally, $y_i$ represents the observed outcome, connected to the latent data and parameters through the function $g(z_i,\boldsymbol{\alpha})$.

With the use of data augmentation [e.g., (Tanner and Wong 1987; Albert and Chib 1993a)], posterior simulation in these models typically proceeds by first characterizing the joint posterior distribution of latent data and parameters:

$$
\begin{aligned}
p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\theta}) && (82)\\
&= p(\boldsymbol{\theta})\prod_{i=1}^{n}\phi(z_i|\boldsymbol{x}_i\boldsymbol{\beta},\sigma^2)I\big[y_i=g\left(z_i,\boldsymbol{\alpha}\right)\big], && (83)
\end{aligned}
$$

with $\boldsymbol{z}=[z_1\ z_2\ \cdots\ z_n]'$, $\boldsymbol{y}=[y_1\ y_2\ \cdots\ y_n]'$, $I(\cdot)=1$ if the statement in the parentheses is true and $I(\cdot)=0$ otherwise. A posterior simulator produced via the Gibbs sampler, then, successively draws from complete posterior conditionals for $\boldsymbol{\alpha}$, $\sigma^2$, $\boldsymbol{\beta}$ and $\boldsymbol{z}$. Assuming that the prior for $\boldsymbol{\beta}$ is of the form: $\boldsymbol{\beta}\sim\mathcal{N}(\boldsymbol{\mu_\beta},\boldsymbol{V_\beta})$, $\sigma^2\sim IG(a/2,b)$ and not making any further assumptions regarding the prior for $\boldsymbol{\alpha}$, we obtain:

$$
\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{z},\sigma^2,\boldsymbol{y} &\sim \mathcal{N}\left(\left[\boldsymbol{X}'\boldsymbol{X}/\sigma^2+\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\left[\boldsymbol{X}'\boldsymbol{z}/\sigma^2+\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu_\beta}\right],\left[\boldsymbol{X}'\boldsymbol{X}/\sigma^2+\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\right) && (84)\\
\sigma^2|\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{y} &\sim IG\left(\frac{n+a}{2},\left[b+\frac{1}{2}\sum_{i=1}^{n}(z_i-\boldsymbol{x}_i\boldsymbol{\beta})^2\right]\right) && (85)\\
p(z_i|\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{y}) &\propto \phi(z_i|\boldsymbol{x_i}\boldsymbol{\beta},\sigma^2)I\big[z_i\in\{z_i:y_i=g(z_i,\boldsymbol{\alpha})\}\big] && (86)\\
p(\boldsymbol{\alpha}|\boldsymbol{z},\boldsymbol{y}) &\propto p(\boldsymbol{\alpha})\prod_{i=1}^{n}I\big[y_i=g\left(\boldsymbol{z_i},\boldsymbol{\alpha}\right)\big]. && (87)
\end{aligned}
$$

The computational value of the latent data becomes apparent from expressions (84) and (85), as the conditional posteriors for the parameters $\boldsymbol{\beta}$ and $\sigma^2$ closely mimic those in section 2.1, given that the model is essentially a linear regression model in the latent data $\boldsymbol{z}$. The third posterior conditional is also easily sampled, as it amounts to drawing each $z_i$, independently from a $\mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2)$ distribution, truncated to a region (interval) defined by $\boldsymbol{\alpha}$ and $y_i$.

In the following sections we review how this structure unites Bayesian approaches to several popular econometric models, including the probit model, the tobit model and the ordered probit. We begin with models for binary choice.

## 3.1 Models for Binary Choice

For models of binary choice the equations of (79)-(81) apply with $\boldsymbol{\alpha}$ being null (empty) and $\sigma^2$ restricted to unity for identification purposes. Therefore, posterior simulation in binary choice problems only involves sampling from (84) and (86). Furthermore, equation (81) specializes to:

$$y_i = I(z_i > 0), \quad i = 1, 2, \ldots, n. \tag{88}$$

### 3.1.1 The Probit Model

The probit model emerges under the assumption of conditionally normally distributed latent data, as in (80). Therefore, posterior simulation proceeds, as noted in (Albert and Chib 1993a), by first drawing $\boldsymbol{\beta}$ from the normal posterior conditional in (84) (with $\sigma^2 = 1$) and then independently sampling the latent data as follows:

$$z_i|\boldsymbol{y}, \boldsymbol{\beta} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\boldsymbol{x}_i\boldsymbol{\beta}, 1) & \text{if } y_i = 1 \\ \mathcal{TN}_{(-\infty,0]}(\boldsymbol{x}_i\boldsymbol{\beta}, 1) & \text{if } y_i = 0 \end{cases}, \quad i = 1, 2, \ldots, n, \tag{89}$$

where, notationally, $x \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes that $x$ is a normally distributed random variable with (untruncated) mean $\mu$ and (untruncated) variance $\sigma^2$ which is then truncated to the interval $(a, b)$. This truncated density retains the shape of the normal density over $(a, b)$, is zero outside this interval, and is simply scaled up to be proper. While one can generate draws from the truncated normal above by repeatedly drawing from a $\mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, 1)$ distribution and simply waiting for a draw

that falls in the desired half-line, this process is quite inefficient, and sometimes prohibitively so. Draws from the desired truncated normals in (89) can, however, be *directly* produced using the method of inversion. To this end, let

$$u \sim U(0,1)$$

be a draw from the uniform distribution on the unit interval. We can then form the variable $w$, where

$$w = \mu + \sigma\Phi^{-1}\left(\Phi\left(\frac{a-\mu}{\sigma}\right) + u\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]\right), \tag{90}$$

and simple derivations show that $w \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2)$.

When applying this result for posterior simulation of latent data in the probit model, note when $y_i = 1$, $a = 0$ and $b = \infty$ and thus $\Phi([b - \boldsymbol{x}_i\boldsymbol{\beta}]/\sigma) = 1$. Likewise, when $y_i = 0$, $a = -\infty$ and $b = 0$ so that $\Phi([a - \boldsymbol{x}_i\boldsymbol{\beta}]/\sigma) = 0$. In this way a Gibbs sampler for the probit proceeds in two steps: multivariate normal sampling from the $\boldsymbol{\beta}$ posterior conditional and independent truncated normal sampling for the posterior conditional for the latent data.[18]

### 3.1.2 The Logit Model

The model that is (arguably) most commonly employed with binary data is the logit model, which specifies

$$\Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})}. \tag{91}$$

There are several Bayesian alternatives for estimating the logit model and we describe a few possibilities here.

For purposes of continuity, we should first review how the framework of (79) - (81) can be extended to accommodate the logit. This can be done by, again, first setting $\sigma^2 = 1$ and noting $\boldsymbol{\alpha}$ is null for the logit. Furthermore, we expand the parameter vector $\boldsymbol{\theta}$ to $\boldsymbol{\theta} = [\boldsymbol{\beta}' \ \boldsymbol{\lambda}']'$. The new set of parameters $\boldsymbol{\lambda}$ will be regarded as *scale mixing variables*, and the addition of such variables to the error variance will aid in expanding the normal sampling model. To illustrate the role of these

[18] (Holmes and Held 2006) actually suggest using a blocking step, first marginalizing $\boldsymbol{\beta}$ out of the latent variable equation, producing a multivariate normal for the latent data vector $\boldsymbol{z}$. Each element of $\boldsymbol{z}$ must then be sampled from a univariate truncated normal, whose mean depends on all other elements of $\boldsymbol{z}$ and must be updated each time a new element of $\boldsymbol{z}$ is sampled. While this may afford some improved mixing properties, its use remains rather uncommon in practice, as the simpler alternative in (84) and (89) displays adequate mixing performance.

variables, we could consider the linear model

$$y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon_i|\boldsymbol{X}, \boldsymbol{\lambda}, \sigma^2, \overset{ind}{\sim} \mathcal{N}(0, \lambda_i\sigma^2),$$

where

$$\lambda_i|\boldsymbol{\lambda_0} \overset{iid}{\sim} G(\boldsymbol{\lambda_0})$$

for some prior distribution $G$ and set of hyperparameters $\boldsymbol{\lambda_0}$. As demonstrated in previous work [e.g., (Andrews and Mallows 1974; Carlin and Polson 1991; Geweke 1993; Koop, Poirier and Tobias 2007, chapter 15)], different choices of $G$ give rise to sampling models (marginalzed over $\boldsymbol{\lambda}$) other than the normal. Specifically, assuming $\lambda_i \sim IG(\nu/2, \nu/2)$ produces a Student-t sampling model for $y|\boldsymbol{\beta}, \sigma^2$ while assuming $\lambda_i \sim Exp(2)$ (an exponential distribution with mean 2) produces a double exponential sampling model for $y|\boldsymbol{\beta}, \sigma^2$.

A similar structure, with $\lambda_i$ specified to follow the asymptotic distribution of the Kolmogorov distance statistic, produces a logistic distribution for $\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2$. Applied to our latent variable representation of the binary choice model, we write:

$$z_i|\boldsymbol{\beta}, \lambda_i \overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, 4\lambda_i^2), \quad i = 1, 2, \ldots, n, \tag{92}$$

where the $\lambda_i$ are independently distributed, with priors that follow the asymptotic distribution of the Kolmogorov distance statistic:

$$p(\lambda_i) = 8\sum_{k=1}^{\infty}(-1)^{k+1}k^2\lambda_i \exp(-2k^2\lambda_i^2), \quad \lambda_i > 0, \quad i = 1, 2, \ldots, n. \tag{93}$$

We sketch in general detail why this strategy reproduces the logit model.[19] We do so by obtaining the density for the latent $z_i$ marginalized over the mixing variable $\lambda_i$. To this end note that, provided it is permissible to interchange the order of integration and summation (and dropping the subscript $i$ in $\lambda_i$ for notational ease):

$$\begin{aligned} p(z_i|\boldsymbol{x}_i, \boldsymbol{\beta}) &= \int_0^{\infty}(2\pi 4\lambda^2)^{-1/2}\exp\left(-\frac{1}{8\lambda^2}(z_i - \boldsymbol{x}_i\boldsymbol{\beta})^2\right) 8\sum_{k=1}^{\infty}(-1)^{k+1}k^2\lambda\exp(-2k^2\lambda^2)d\lambda \quad (94) \\ &= 4(2\pi)^{-1/2}\sum_{k=1}^{\infty}(-1)^{k+1}k^2\int_0^{\infty}\exp\left[-\frac{1}{2}\left(\frac{1}{4}(z_i - \boldsymbol{x_i}\boldsymbol{\beta})^2\lambda^{-2} + 4k^2\lambda^2\right)\right]d\lambda. \quad (95) \end{aligned}$$

The integral above can be simplified by observing [e.g. (Andrews and Mallows 1974: equation 2.2)]:

$$\int_0^{\infty}\exp\left(-\frac{1}{2}[a^2u^2 + b^2u^{-2}]\right)du = \left(\frac{\pi}{2a^2}\right)^{1/2}\exp(-|ab|). \tag{96}$$

[19] See (Andrews and Mallows 1974; Stefanski 1991) for further details based on Laplace transforms.

Making use of the formula above, and then simplifying (95) gives the alternating series representation:

$$p(z_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \sum_{k=1}^{\infty} (-1)^{k+1} k \exp\left(-k|z_i - \boldsymbol{x}_i\boldsymbol{\beta}|\right). \tag{97}$$

To evaluate this quantity, recall that $(1+x)^{-1}$ can be represented in series form as $\sum_{k=0}^{\infty}(-1)^k x^k$ for $|x| < 1$. Thus, by differentiation,

$$(1+x)^{-2} = \sum_{k=1}^{\infty} (-1)^{k+1} k x^{k-1}. \tag{98}$$

Applying this result to our formula in (97) with $x = \exp(-|z_i - \boldsymbol{x}_i\boldsymbol{\beta}|)$, we obtain:

$$p(z_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(-|z_i - \boldsymbol{x}_i\boldsymbol{\beta}|)}{[1+\exp(-|z_i - \boldsymbol{x}_i\boldsymbol{\beta}|)]^2} = \frac{\exp[-(z_i - \boldsymbol{x}_i\boldsymbol{\beta})]}{(1+\exp[-(z_i - \boldsymbol{x}_i\boldsymbol{\beta})])^2}, \tag{99}$$

producing a logistic density for the latent $z_i$.

The above shows how the logistic distribution can be represented as a scale mixture of normals and thus how MCMC methods can be employed to estimate the model using the general structure of (79) - (81). While sampling from the $\boldsymbol{\beta}$ and $\boldsymbol{z}$ posterior conditionals proceeds similarly to the probit model, calculations for the logit also require sampling from the conditional distribution of the mixing variables $\boldsymbol{\lambda}$. (Chen and Dey 1998) propose to use a Student-t approximation to the logistic distribution and sample $\lambda_i^2$, $i = 1, 2, \ldots, n$ from an optimally chosen inverse gamma proposal density. Moreover, they discuss procedures for efficient calculation of the infinite sum in (93) that is required in the M-H step. (Holmes and Held 2006) also pursue this approach for estimating the logit, using rejection sampling with a generalized inverse Gaussian proposal density to sample the mixing variables $\boldsymbol{\lambda}$. In either approach, however, the calculations involved remain reasonably non-trivial, and such techniques appear infrequently in practice.

A second approach for fitting the binary logit (as well as the multinomial logit) has been suggested by (Frühwirth-Schnatter and Frühwirth 2007), with extensions for variable selection provided by (Tüchler 2008). They begin by noting, as shown by (McFadden 1974), that the logit likelihood can be derived from a latent variable framework based on type I extreme value assumptions on the disturbances. Specifically, the utility afforded by the $y_i = 0$ choice (denoted $z_{i0}$) is assumed to follow a type I extreme value distribution (with covariates omitted for identification purposes), while the latent utility afforded by the $y_i = 1$ option (which includes covariates) is given as

$$z_{i1} = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad p(\epsilon_i) = \exp\left[-\epsilon_i - \exp(-\epsilon_i)\right]. \tag{100}$$

The idea of (Frühwirth-Schnatter and Frühwirth 2007), similar in spirit to that of (Chib et al. 2002), is to replace the computationally troublesome type I extreme value distribution in (100) with a nearly identical - yet computationally more appealing - normal mixture approximation (see, e.g. (Griffin, Quintana and Steel 2010) of this volume for more discussion of mixture models). Specifically, they write

$$p(\epsilon_i) \approx \sum_{r=1}^{10} w_r \phi(\epsilon_i; m_r, s_r^2) \tag{101}$$

where the weights $\{w_r\}_{r=1}^{10}$, component means $\{m_r\}_{r=1}^{10}$ and component variances $\{s_r^2\}_{r=1}^{10}$ are chosen to minimize the Kullback-Leibler distance between the mixture approximation and the extreme value distribution. The optimal values of these parameters are enumerated in Table 1 of (Frühwirth-Schnatter and Frühwirth 2007: 3511) and are not repeated here for the sake of brevity.

Making this replacement, the (approximate) latent utility for the $y_i = 1$ choice becomes

$$p(z_{i1}|\boldsymbol{\beta}) = \sum_{r=1}^{10} w_r \phi(z_{i1}; \boldsymbol{x}_i\boldsymbol{\beta} + m_r, s_r^2). \tag{102}$$

When fitting mixture models like these, it is helpful to augment the mixture density with a set of component indicator variables, $\{r_i\}_{i=1}^n$ where $r_i = j$ denotes that $z_{i1}$ is "drawn from" the $j^{th}$ component of the mixture (with mean $\boldsymbol{x}_i\boldsymbol{\beta} + m_j$ and variance $s_j^2$). In this case, $j = 1, 2, \ldots, 10$ given the ten component approximation to the Type I extreme value distribution. Formally, we write

$$z_{i1}|\boldsymbol{\beta}, r_i \overset{ind}{\sim} \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta} + m_{r_i}, s_{r_i}^2), \quad \Pr(r_i = j) = w_j, \quad j = 1, 2, \ldots, 10.$$

The above pieces yield the following augmented posterior distribution for the logit

$$\begin{aligned} p(\boldsymbol{z}_1, \boldsymbol{z}_0, \boldsymbol{\beta}, \boldsymbol{r}|\boldsymbol{y}) &\propto p(\boldsymbol{\beta}, \boldsymbol{r}) p(\boldsymbol{z}_1, \boldsymbol{z}_0|\boldsymbol{\beta}, \boldsymbol{r}) p(\boldsymbol{y}|\boldsymbol{z}_1, \boldsymbol{z}0, \boldsymbol{\beta}, \boldsymbol{r}) \\ &= p(\boldsymbol{\beta}) \prod_{i=1}^{n} \Bigg[ \left( \sum_{j=1}^{10} I(r_i = j) w_j \right) \phi(z_{i1}; \boldsymbol{x}_i\boldsymbol{\beta} + m_{r_i}, s_{r_i}^2) p(z_{i0}) \\ &\quad \times [I(y_i = 0) I(z_{i0} \geq z_{i1}) + I(y_i = 1) I(z_{i0} < z_{i1})] \Bigg], \end{aligned} \tag{103}$$

where $p(z_{i0})$ is type I extreme value and represents the latent utility of the $y_i = 0$ option. In practice, this latent variable does not need to be simulated in the course of the sampler, though its value does indirectly affect the sampling of $\boldsymbol{z}_1$.

Under the normal prior $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta})$, the sampling of $\boldsymbol{\beta}$ from the (approximate) conditional posterior follows immediately:

$$\boldsymbol{\beta}|\boldsymbol{z}_1, \boldsymbol{z}_0, \boldsymbol{r}, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{D_\beta d_\beta}, \boldsymbol{D_\beta}) \tag{104}$$

where

$$\boldsymbol{D_\beta} = \left(\boldsymbol{X'\Sigma^{-1}X} + \boldsymbol{V_\beta^{-1}}\right)^{-1} \quad \text{and} \quad \boldsymbol{d_\beta} = \boldsymbol{X'\Sigma^{-1}\tilde{y}} + \boldsymbol{V_\beta^{-1}\mu_\beta},$$

with $\boldsymbol{\Sigma} \equiv diag\{s_{r_i}^2\}$, $\boldsymbol{\tilde{y}} \equiv \boldsymbol{y} - \boldsymbol{m}$, $\boldsymbol{m} \equiv [m_{r_1} \quad m_{r_2} \quad \cdots \quad m_{r_n}]'$.

In a similar way, the component indicator variables are sampled independently from their discrete conditional posterior distributions,

$$\Pr(r_i = j|\boldsymbol{z}_1, \boldsymbol{z}_0, \boldsymbol{\beta}, \boldsymbol{y}) \propto \frac{w_j}{s_j}\phi\left(\frac{z_{i1} - \boldsymbol{x}_i\beta - m_j}{s_j}\right), \quad j = 1, 2, \ldots, 10. \tag{105}$$

For the latent data $\boldsymbol{z}_1$, Frühwirth-Schnatter and Frühwirth go back to the exact representation of the logit and note, by a simple change of variables, $\exp(-z_{i0}) \sim Exp(1)$ and likewise, $\exp(-z_{i1}) \sim Exp(\lambda_i)$ where "*Exp*" denotes an exponential distribution, $\lambda_i \equiv \exp(\boldsymbol{x}_i\boldsymbol{\beta})$ and $\exp(\cdot)$ denotes the exponential function.

When $y_i = 1$, the observed outcome imposes the restriction $z_{i0} < z_{i1}$ on the latent data or, equivalently, $\exp(-z_{i0}) > \exp(-z_{i1})$. In this instance, therefore, $\exp(-z_{i1})$ can be sampled as the minimum of two exponential random variables, which turns out to imply

$$\exp(-z_{i1}) \sim Exp(1 + \lambda_i), \quad \text{when } y_i = 1. \tag{106}$$

When $y_i = 0$, we obtain the restriction $\exp(-z_{i0}) \leq \exp(-z_{i1})$. Similar algebra can be employed to verify that $\exp(-z_{i1})$ can be sampled as the sum of two exponential random variables in this case:

$$\exp(-z_{i1}) = x_{1i} + x_{2i}, \quad x_{1i} \sim Exp(1 + \lambda_i), \quad x_{2i} \sim Exp(\lambda_i), \text{ when } y_i = 0. \tag{107}$$

Posterior simulation in the logit via auxiliary variable augmentation proceeds by sampling from (104) and (105) and then using (106) and (107) to sample the latent utility vector $\boldsymbol{z}_1$. It is useful to note that each of the steps only requires sampling from standard distributions, making this an attractive algorithm for the practitioner.

Yet another approach for posterior simulation in the logit involves the M-H algorithm. To this end we first note that the Hessian for the logit is obtained as

$$\boldsymbol{\mathcal{H}} = -\boldsymbol{X'AX} = -\sum_{i=1}^{n} \boldsymbol{x}_i'\Lambda_i(1 - \Lambda_i)\boldsymbol{x}_i, \tag{108}$$

where $\boldsymbol{X}$ is the $n \times k$ matrix of stacked covariate data, $\Lambda_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i\boldsymbol{\beta}))$ and $\boldsymbol{A}$ is an $n \times n$ diagonal matrix with $\Lambda_i(1-\Lambda_i)$ assembled on the main diagonal. This Hessian can be used to scale the proposal density in the M-H step. A random-walk M-H algorithm, for example, would proceed by sampling

$$\boldsymbol{\beta}^* \sim \mathcal{N}\left(\boldsymbol{\beta}^{(r)}, -c^2\boldsymbol{\mathcal{H}}^{-1}\right) \tag{109}$$

where $c^2$ is a tuning parameter chosen to minimize the relative inefficiency of the M-H procedure.

Given a current parameter vector $\boldsymbol{\beta}^{(r)}$, and a $\boldsymbol{\beta}^*$ sampled from (109), the chain will move from $\boldsymbol{\beta}^{(r)}$ to $\boldsymbol{\beta}^*$ with probability

$$\min\{1, p_r\} \tag{110}$$

where

$$p_r = \exp\left[\log p(\boldsymbol{\beta}^*) - \log p(\boldsymbol{\beta}^{(r)}) + \sum_{i=1}^{n}\left(y_i \log\left[\frac{\Lambda_i^*}{\Lambda_i^{(r)}}\right] + (1-y_i)\log\left[\frac{1-\Lambda_i^*}{1-\Lambda_i^{(r)}}\right]\right)\right], \tag{111}$$

and $\Lambda_i^*$ and $\Lambda_i^{(r)}$ denote the logit predicted probability for person $i$ evaluated at the candidate and current value of the chain, respectively, and the first two terms in $p_r$ denote the (log) difference in prior ordinates.

Implementation of the M-H algorithm requires an initial estimate of the parameter vector $\boldsymbol{\beta}$ in order to calculate $\boldsymbol{A}$ and thus the Hessian $\boldsymbol{\mathcal{H}}$. One possibility in this regard is to simply perform MLE of the logit, which is cheaply obtained and widely available in most software packages, and use the MLE estimate to calculate $\boldsymbol{A}$ and thus the Hessian.

An alternative that would not require MLE calculation is to simply to start with a guess for $\boldsymbol{\beta}$, perhaps setting $\boldsymbol{\beta} = 0$, producing $\Lambda_i = 1/2\ \forall i$, and run the M-H algorithm above. Once a desired number of simulations have been produced, the algorithm can be terminated, a posterior mean of $\boldsymbol{\beta}$ calculated, and the matrix $\boldsymbol{A}$ and Hessian $\boldsymbol{\mathcal{H}}$ can then be updated. This process could be repeated until the calculated posterior means have stabilized. We take this route in the example provided in this section. Since the likelihood function of the logit model is strictly concave, the above Metropolis-Hastings approach can work well, especially for samples sufficiently large.

### 3.1.3 Other link functions

While the probit and logit are the most widely used among the binary choice models, they are not the only possibilities, and indeed, can be inappropriately restrictive. The complementary log-log link model, for example, specifying

$$\Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = 1 - \exp[-\exp(\boldsymbol{x}_i\boldsymbol{\beta})] \tag{112}$$

offers an asymmetric link, and can be estimated using methods similar to our final M-H algorithm for the logit. Other skewed link models, including skew-normal links, are described by (Chen, Dey and Shao 1999) and further possibilities are explored by (Basu and Mukhopadhyay 2000). (Geweke and Keane 2000) describe a binary choice model based on finite normal mixtures.

### 3.1.4 Application

Our application again makes use of the British Cohort Study data of section 2.5.2. In this case our binary outcome of interest is whether or not the respondent is obese, whose clinical definition is having a BMI in excess of 30. The covariates we employ include parental BMI and indicators denoting whether or not the respondent is married, has a college degree, or exercises regularly. For our priors, we choose $\boldsymbol{\beta} \sim \mathcal{N}(0, 100\boldsymbol{I}_6)$.

We estimate the probit model using the algorithm discussed in section 3.1.1 and also estimate the logit and complementary log-log specifications for this data. For the logit, we start out with $\boldsymbol{\beta} = \mathbf{0}, c^2 = 2$ and calculate the Hessian at $\boldsymbol{\beta} = \mathbf{0}$. The posterior simulator is then run for $10,000$ iterations and posterior means are calculated based on the final 5,000 simulations. The Hessian is then recalculated at the updated posterior mean, $c$ is set to unity and the simulator is run for an additional 25,000 iterations. This process is then repeated one final time, with final posterior statistics calculated from the last 24,000 iterations of this third run. We approach the complementary log-log model in the same way, sampling candidates from our proposal density as:

$$\boldsymbol{\beta}^* \sim \mathcal{N}\left(\boldsymbol{\beta}^{(r)}, c^2(\boldsymbol{X'DX})^{-1}\right), \tag{113}$$

where $\boldsymbol{D}$ is a diagonal matrix with

$$\left(y_i \frac{\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1-\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))} - (1-y_i)\exp(\boldsymbol{x}_i\boldsymbol{\beta})\right)^2 \tag{114}$$

as the $(i, i)$ entry. Therefore, for the complementary log-log model, the covariance matrix of our proposal density is chosen as a scaled Berndt, Hall, Hall and Hausman (BHHH) estimate (Berndt et. al 1974) of the inverse information matrix.

In Table 7 below we provide posterior means and standard deviations associated with model-specific marginal effects for each of these variables. When required, we evaluate these at the sample average of the continuous parental BMI variables and set the marriage, college degree, and regular exercise variables to unity. For binary covariates, marginal effects are calculated as the difference in predicted probabilities upon setting the binary variable to zero, and the remaining covariates fixed at means (or unity). To compare these models, we also calculate log marginal likelihoods. For the probit model, these are calculated using the method of (Chib 1995) while marginal likelihoods for the logit and complementary log-log models, which use the M-H algorithm, are calculated using (Chib and Jeliazkov 2001).

Table 7: Marginal Effect Posterior Means and Posterior Standard Deviations
From Binary Choice Application

| Variable | Probit | | Logit | | Complementary-Log-Log | |
|---|---|---|---|---|---|---|
| | Post. Mean | Post Std. | Post Mean | Post Std. | Post Mean | Post Std |
| MomBMI | .010 | .002 | .009 | .002 | .008 | .002 |
| DadBMI | .011 | .002 | .010 | .002 | .009 | .002 |
| Married | .022 | .012 | .022 | .011 | .021 | .010 |
| Degree | -.016 | .016 | -.017 | .016 | -.018 | .016 |
| ExerciseReg | -.003 | .016 | -.004 | .015 | -.003 | .014 |
| Log ML | -928.93 | | -936.55 | | -936.22 | |

The results in the table are sensible and operate in the direction that we might expect. A one point increase in either maternal or paternal BMI increases the likelihood of child obesity by about 1 percent; married individuals are about 2 percent more likely to be obese, and those with a college degree are nearly 2 percent less likely to be obese. The results obtained are quite consistent across models, and the data favors the probit specification among these three alternatives.

### 3.2 The Tobit Model

The tobit model is a widely used specification for censored data and (Chib 1992) marks the first MCMC-based Bayesian procedure for inference in the tobit. The basic tobit specification, with a

single censoring point at zero, can be mapped into the framework of (79) - (81) with $\boldsymbol{\alpha}$ being null and (81) specializing to:

$$y_i = \max\{0, z_i\}. \tag{115}$$

As such, a posterior simulator for the tobit is remarkably simple. The regression parameters $\boldsymbol{\beta}$ are sampled directly from (84) and the variance parameter $\sigma^2$ is drawn directly from (85). As for the latent data, let $D_i = I(y_i > 0)$. Equation (86) together with the rule in (115) then imply the latent $z_i$ can be sampled by setting

$$z_i = D_i y_i + (1 - D_i) w_i, \tag{116}$$

where

$$w_i \overset{ind}{\sim} \mathcal{TN}_{(-\infty,0)}(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma^2), i = 1, 2, \ldots, n. \tag{117}$$

In other words, $w_i$ only needs to be simulated for the set of observations with $y_i = 0$. Other generalizations of the tobit, such as allowing for an unknown censoring point, or allowing for two-sided censoring, offer straightforward extensions of this basic model.

## 3.3 Models for Ordinal Outcomes

The ordered probit represents another commonly encountered microeconometric model that fits within the structure described by (79) - (81). For the ordered probit, $\sigma^2 = 1$, and the parameters $\boldsymbol{\alpha}$ are *threshold* or *cutpoint* parameters to be estimated within the model. Specifically, we assume $y_i \in \{1, 2, \ldots, J\}$, where the discrete outcomes have a natural ordinal interpretation, such as degrees of agreement / disagreement with a given statement.

For the ordered probit, (81) becomes

$$y_i = j \quad \text{if} \quad \alpha_j < z_i \leq \alpha_{j+1}, \quad j = 1, 2, \ldots, J. \tag{118}$$

An intercept parameter is presumed to be in $\boldsymbol{x}_i$, and standard identification conditions are imposed on the cutpoints, namely: $\alpha_1 = -\infty$, $\alpha_2 = 0$ and $\alpha_{J+1} = \infty$.

### 3.3.1 Posterior Simulation

In terms of posterior simulation, a standard Gibbs sampler can be applied, as in (Albert and Chib 1993a). The regression parameters $\boldsymbol{\beta}$ are simulated as in (84), with $\sigma^2 = 1$, the latent data are

drawn from a truncated normal distribution implied by (86) and (118):

$$z_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y} \overset{ind}{\sim} \mathcal{TN}_{(\alpha_{y_i}, \alpha_{y_i+1}]}(\boldsymbol{x}_i\boldsymbol{\beta}, 1), \tag{119}$$

and the elements of the cutpoint vector $\alpha_j$, under an improper prior of the form $p(\boldsymbol{\alpha}) \propto c$, can be sampled from their conditional posterior distributions:

$$\alpha_j|\boldsymbol{\alpha_{-j}}, \boldsymbol{z}, \boldsymbol{y} \sim U\left[\max\left\{\alpha_{j-1}, \{z_i : y_i = j-1\}\right\}, \min\left\{\alpha_{j+1}, \{z_i : y_i = j\}\right\}\right]. \tag{120}$$

Unfortunately the algorithm above does not mix well in practice. (Cowles 1996) investigates this issue and suggests sampling $\boldsymbol{\alpha}$ and $\boldsymbol{z}$ in a blocking step by first integrating out the latent $\boldsymbol{z}$ (i.e., working directly with the ordered probit likelihood), sampling $\boldsymbol{\alpha}$, and then sampling from the complete posterior conditional distribution for the latent data. This is done in an M-H step, where a series of truncated normal densities are used to sample the cutpoints. (Nandram and Chen 1996) also investigate this issue and discuss posterior simulation based on a rescaling transformation.

To motivate their reparameterization, suppose $J = 3$ to fix ideas (so that there is only one unknown cutpoint, denoted as $\alpha$) and let

$$\boldsymbol{\delta} = \frac{\boldsymbol{\beta}}{\alpha}, \quad \sigma = \frac{1}{\alpha}, \quad \text{and} \quad \tilde{z}_i = \frac{z_i}{\alpha}. \tag{121}$$

This reparameterization leads to an equivalent model:

$$\tilde{z}_i = \boldsymbol{x}_i\boldsymbol{\delta} + \nu_i, \quad \nu_i|\boldsymbol{X}, \sigma \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{122}$$

with

$$y_i = \begin{cases} 1 & \text{if } \tilde{z}_i \leq 0 \\ 2 & \text{if } 0 < \tilde{z}_i \leq 1 \\ 3 & \text{if } \tilde{z}_i > 1 \end{cases}. \tag{123}$$

If we additionally specify that a diffuse, improper prior on $\boldsymbol{\beta}, \boldsymbol{\alpha}$ is employed: $p(\boldsymbol{\beta}, \boldsymbol{\alpha}) \propto c$, we obtain the following joint posterior for the reparameterized model:

$$p(\boldsymbol{\delta}, \sigma^2, \tilde{\boldsymbol{z}}|\boldsymbol{y}) \propto \sigma^{-n} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}[\tilde{z}_i - \boldsymbol{x}_i\boldsymbol{\delta}]^2\right) I(\tilde{\alpha}_{y_i} < \tilde{z}_i \leq \tilde{\alpha}_{y_i+1}) \tag{124}$$

where all of the $\tilde{\alpha}_{y_i}$ are *known* upon reparameterization. The conditionals for $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{z}}$ are like those in (84) and (86), with $\boldsymbol{V}_{\boldsymbol{\beta}}^{-1} = 0$ and the conditional support in (86) defined by the intervals $I(\tilde{\alpha}_{y_i} < \tilde{z}_i^* \leq \tilde{\alpha}_{y_i+1})$, all of which do not depend on unknown parameters. The (reparameterized) variance parameter is sampled as:

$$\sigma^2|\boldsymbol{\delta}, \tilde{\boldsymbol{z}}, \boldsymbol{y} \sim IG\left(\frac{n}{2}, \left[\frac{1}{2}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\boldsymbol{\delta})'(\tilde{\boldsymbol{z}} - \boldsymbol{X}\boldsymbol{\delta})\right]\right). \tag{125}$$

For each post-convergence iteration, the original parameters $\alpha$ and $\boldsymbol{\beta}$ can be calculated. When the outcome variable takes on more than three possible values, (Nandram and Chen 1996) suggest the use of a M-H algorithm to sample the unknown cutpoints where all parameters are drawn in a single step based on a Dirichlet proposal density for differences in cutpoint values. Other contributions to this literature include (Chen and Dey 2000; Albert and Chib 2001; Graves, Jeliazkov and Kutzbach 2008). The first and last of these references consider multivariate ordinal outcomes, departures from normality and a series of related issues.

### 3.3.2 Ordered Probit: Application

To illustrate estimation of the ordered probit in practice, we revisit our data from the British Cohort Study. In this case we refine our classification of weight categories into "normal" weight ($y = 1$), "overweight" ($y = 2$) and "obese" ($y = 3$). The first of these is defined as a BMI less than twenty five,[20] the second represents a BMI between 25 and 30, and obesity denotes BMI in excess of 30.

We employ the same set of covariates as those used in section 3.1.4 and make use of the reparameterization technique described above, employing improper priors for $\boldsymbol{\beta}$ and $\alpha$. The sampler is run for 2,500 iterations, and the first 500 of these are discarded as the burn-in period.

Table 8: Posterior Statistics From Ordinal BMI Application

| Parameter / Variable | | | Probability Change: Parental BMI | | |
|---|---|---|---|---|---|
| | Post. Mean | Post. Std. | Category | Post Mean. | Post. Std. |
| Constant | -2.89 | .220 | | | |
| MomBMI | .051 | .006 | $y = 1$ | .150 | .011 |
| DadBMI | .067 | .008 | | | |
| Married | .214 | .041 | $y = 2$ | -.097 | .008 |
| Degree | -.169 | .051 | | | |
| ExerciseReg. | .026 | .058 | $y = 3$ | -.053 | .005 |
| $\alpha$ | 1.29 | .034 | | | |

In addition to coefficient posterior means and standard deviations, we also report posterior means and posterior standard deviations for a particular effect of interest, as summarized in the rightmost

[20] While it is indeed possible to be "underweight," we abstract from this issue, and note that approximately one percent of our sample had a BMI less than 19.

3 columns of Table 8. In particular, we consider how probabilities associated with each BMI classification change in response to a one standard deviation decrease in parental BMI values. Specifically, we first calculate the predicted probability of each BMI category at sample means of MomBMI and DadBMI, with the remaining covariates fixed at unity. This process is repeated, but this time the three probabilities are calculated upon setting the MomBMI and DadBMI values at one standard deviation below their respective sample means (with the remaining covariates still fixed at unity). Posterior means and standard deviations of the probability changes resulting from these 1 standard deviation decreases in parental BMI are then reported in the final two columns of Table 8. The results show that such a reduction in parental BMI leads to a 15 percent increase in the probability that the child will be normal weight, a 9.7 percent decrease in the probability of being overweight and a 5.3 percent decrease in the probability of obesity.

# 4 Multivariate Latent Models

We build upon the topics of the last section to now discuss multivariate latent variable models. We proceed in a similar fashion by first introducing a general latent multivariate framework and then discussing particular models that emerge from this specification. The basic model that we have in mind is a straightforward multivariate generalization of (79) - (81), where $\boldsymbol{\theta} = [\boldsymbol{\beta}' \; vec(\boldsymbol{\Sigma})']'$, and[21]

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\beta})p(\boldsymbol{\Sigma^{-1}}) && (126) \\ \boldsymbol{z}_i|\boldsymbol{X},\boldsymbol{\theta} &\overset{ind}{\sim} \mathcal{N}\left(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}\right), \quad i = 1,2,\ldots,n && (127) \\ \boldsymbol{y}_i|\boldsymbol{z}_i &= g(\boldsymbol{z}_i), \quad i = 1,2,\ldots,n. && (128) \end{aligned}$$

For our priors in (126), we continue to use a multivariate normal prior for $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta})$ and a Wishart prior for $\boldsymbol{\Sigma^{-1}}$: $\boldsymbol{\Sigma^{-1}} \sim W([\kappa \boldsymbol{R}]^{-1}, \kappa)$.

With an eye toward implementation of the Gibbs sampler in a model of this form, we obtain the following conditional posterior distributions:

$$\boldsymbol{\beta}|\boldsymbol{\Sigma},\boldsymbol{Z},\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{D_\beta d_\beta}, \boldsymbol{D_\beta}\right) \quad (129)$$

[21] We do not discuss multivariate ordinal models here. See (Graves, Jeliazkov and Kutzbach 2008), for example, for further discussion.

where

$$\boldsymbol{D_\beta} \equiv \left[ \left( \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i \right) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \right]^{-1} \quad \text{and} \quad \boldsymbol{d_\beta} \equiv \left( \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{z}_i \right) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu_\beta}, \tag{130}$$

$$\boldsymbol{\Sigma^{-1}} | \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{y} \sim W \left( \left[ \sum_{i=1}^{n} (\boldsymbol{z}_i - \boldsymbol{X}_i \boldsymbol{\beta})(\boldsymbol{z}_i - \boldsymbol{X}_i \boldsymbol{\beta})' + \kappa \boldsymbol{R} \right]^{-1}, n + \kappa \right) \tag{131}$$

and

$$p(\boldsymbol{z}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y}) \propto \phi(\boldsymbol{z}_i | \boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}) I[\boldsymbol{z}_i \in \{\boldsymbol{z}_i : \boldsymbol{y}_i = g(\boldsymbol{z}_i)\}], \quad i = 1, 2, \ldots, n. \tag{132}$$

Rather trivially, we note that this framework also describes a Gibbs algorithm for sampling in the Seemingly Unrelated Regressions (SUR) model. For the standard SUR specification, outcomes are fully observed so that there is no latent data (i.e., $\boldsymbol{y}_i = \boldsymbol{z}_i$) and posterior simulation simply involves the sampling of $\boldsymbol{\beta}$ from (129) and $\boldsymbol{\Sigma^{-1}}$ from (131). Though the SUR model is free of any latent data, it is useful nonetheless to note that a limiting version of this framework also describes posterior simulation within this system of linear equations.

As shown in the following discussion, the model above is also sufficiently general to include generalized tobit,[22] the multinomial probit and the multivariate probit models as special cases. Each of these important microeconometric models, however, will impose different restrictions on $\boldsymbol{\Sigma}$ for identification purposes. Therefore, procedures for sampling $\boldsymbol{\Sigma^{-1}}$ (or its constituent elements) will differ across the models. Furthermore, since the mapping in (128) will also change with the model, procedures for sampling the latent $\boldsymbol{z_i}$ will also differ among the specifications enumerated below. The sampling of $\boldsymbol{\beta}$ from its posterior conditional, however, will proceed as described in (129) in all cases, with the latent and covariate data suitably defined within the context of each model.

### 4.1 The Hurdle / Sample Selection Model

The simple tobit specification of section 3.2 is often criticized for its inability to simultaneously account for the incidence of zeros and the density of non-zero outcomes. Such concerns can be mitigated, and the performance of the model generally improved, when elaborating the structure

[22] (Amemiya 1985) enumerates 5 different types of generalized tobit models and discusses classical estimation and inference for each. In what follows we discuss posterior simulation in Amemiya's Type 2 (hurdle model) and Type 5 (potential outcomes model) specifications, noting that similar methods apply to the estimation of all 5 generalized tobit variants.

with a separate process for modeling the zero outcome. The following specification, commonly termed the *hurdle* or *sample selection model*, studied at length from a Bayesian perspective by (van Hasselt 2008), adds this level of generality:

$$z_{i1} = \boldsymbol{r}_i\boldsymbol{\alpha} + u_{i1} \tag{133}$$

$$z_{i2} = \boldsymbol{w}_i\boldsymbol{\delta} + u_{i2} \tag{134}$$

where

$$\boldsymbol{u_i} = \left[ \begin{array}{c} u_{i1} \\ u_{i2} \end{array} \right] \Bigg| \boldsymbol{R}, \boldsymbol{W}, \boldsymbol{\Sigma} \overset{iid}{\sim} \mathcal{N}\left(0, \left[ \begin{array}{cc} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array} \right]\right) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{135}$$

and

$$y_i = \exp[z_{i2}] I(z_{i1} > 0). \tag{136}$$

The above specification contains two latent variable equations, unlike the single equation of the tobit model. Equation (136) establishes the connection between the latent and observed data and clarifies the roles of each latent variable in the hurdle model. Specifically, $z_{i1}$ in (133) models the $y_i = 0$ or $y_i \neq 0$ event. If the latent $z_{i1}$ is positive, then the observed non-zero outcome is presumed to be generated from (134) and is given as $\exp[z_{i2}]$.[23] Similar to the standard tobit, if $z_{i1}$ is non-positive, then the observed $y_i$ is set to zero.

The model above fits exactly within the multivariate framework described in (126) - (128) with $\boldsymbol{z}_i = [z_{i1}\ z_{i2}]'$, $\boldsymbol{\beta} = [\boldsymbol{\alpha}'\ \boldsymbol{\delta}']'$ and

$$\boldsymbol{X}_i = \left[ \begin{array}{cc} \boldsymbol{r}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{w}_i \end{array} \right]. \tag{137}$$

A Gibbs sampling algorithm, then, follows from (129) - (132), with the sampling of $\boldsymbol{\Sigma^{-1}}$ and $\boldsymbol{z}_i$ tailored to the hurdle model. As for the regression parameters, posterior simulation of $\boldsymbol{\beta}$ proceeds exactly as in (129), given our definitions of the covariate and latent data. As for the sampling of $\boldsymbol{\Sigma^{-1}}$, a slight complication is introduced, given a restriction on the (1,1) element of $\boldsymbol{\Sigma}$; the posterior conditional for $\boldsymbol{\Sigma^{-1}}$ conditioned on this prior restriction is no longer Wishart.

One alternative to the sampling of $\boldsymbol{\Sigma^{-1}}$ in this instance (and, in fact, to the sampling of the covariance matrix in all models is this section), is to ignore this restriction, implement a sampler that traverses through a non-identified parameter space, and simply post-process the posterior simulations to focus on the identifiable quantities of interest: $\boldsymbol{\alpha}/\boldsymbol{\Sigma}_{(1,1)}, \boldsymbol{\delta}, \rho_{12}$ and $\sigma_2^2$ (where $\rho_{12}$

[23] The exponential term is introduced to guarantee that the observed outcome is positive, and is also consistent with a majority of applied work, where the *potential* outcome $z_{i2}$ in (134) is modeled log-linearly.

denotes the correlation between $u_1$ and $u_2$). This approach has been advocated by (McCulloch and Rossi 1994; Rossi, Allenby and McColluch 2005), who argue that navigating through the non-identified parameter space simplifies the posterior computations and also improves mixing of the posterior simulations. Under this approach, sampling of $\mathbf{\Sigma^{-1}}$ proceeds identically to (131).

An alternate approach, which works directly with the identified parameters, is to first express $u_{i2}$ conditionally on $u_{i1}$ as:

$$u_{i2} = \sigma_{12} u_{i1} + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \sigma_\nu^2), \tag{138}$$

where $\sigma_\nu^2 = \sigma_2^2 - \sigma_{12}^2$ and $\nu_i$ and $u_{1i}$ are independent. Thus, we can re-write our model as

$$z_{i1} = \boldsymbol{r}_i \boldsymbol{\alpha} + u_{i1} \tag{139}$$

$$z_{i2} = \boldsymbol{w}_i \boldsymbol{\delta} + \sigma_{12} u_{i1} + \nu_i. \tag{140}$$

Aside from the latent data $\boldsymbol{z}$, (the sampling of which has not yet been addressed in either the identified or non-identified approaches), posterior simulation in the model with this parameterization cycles through three blocks of parameters: $\boldsymbol{\beta}$, $\sigma_{12}$ and $\sigma_\nu^2$. To this end, we adopt the following priors for these quantities: $\sigma_\nu^2 \sim IG(a/2, b)$, $\sigma_{12} \sim \mathcal{N}(\mu_{12}, V_{12})$, and $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta})$. The first two of these effectively "replace" the Wishart prior for $\mathbf{\Sigma^{-1}}$ in (126). Finally, note that the inverse gamma prior on $\sigma_\nu^2$ imposes $\sigma_\nu^2 > 0$ and thus $\mathbf{\Sigma}$ is restricted to be positive definite.

As stated previously, the regression parameters $\boldsymbol{\beta}$ are sampled as in (129). The covariance and variance parameters are then sampled by drawing:

$$\sigma_\nu^2 | \boldsymbol{\beta}, \boldsymbol{z}, \sigma_{12}, \boldsymbol{y} \sim IG\left(\frac{n+a}{2}, \left[b + \frac{1}{2}\sum_{i=1}^{n}(z_{i2} - \boldsymbol{w}_i\boldsymbol{\delta} - \sigma_{12}u_{i1})^2\right]\right) \tag{141}$$

and

$$\sigma_{12} | \boldsymbol{\beta}, \boldsymbol{z}, \sigma_\nu^2, \boldsymbol{y} \sim \mathcal{N}(D_{12} d_{12}, D_{12}), \tag{142}$$

where

$$D_{12} = \left(\boldsymbol{u}_1'\boldsymbol{u}_1/\sigma_\nu^2 + V_{12}^{-1}\right)^{-1} \quad \text{and} \quad d_{12} = \boldsymbol{u}_1'\boldsymbol{u}_2/\sigma_\nu^2 + V_{12}^{-1}\mu_{12}. \tag{143}$$

In the expression above, $\boldsymbol{u}_j = [u_{1j}\ u_{2j}\ \cdots\ u_{nj}]',\ j = 1, 2$, denote the error vectors which are known given $\boldsymbol{z}$ and $\boldsymbol{\beta}$.

It remains to discuss the sampling of the latent data $\boldsymbol{z}_i$, and the following describes how this is accomplished under either sampling scheme. First, suppose that $y_i > 0$. This observed outcome

implies the restrictions $z_{i1} > 0$, $z_{2i} = \log(y_i)$. In this case, the sampling of $z_{2i}$ is trivial (as its posterior is degenerate given $\boldsymbol{y}$) , and $z_{1i}$ can be drawn from the univariate truncated normal:

$$z_{i1}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y} \sim \mathcal{TN}_{(0,\infty)}\left[\boldsymbol{r}_i\boldsymbol{\alpha} + \frac{\sigma_{12}}{\sigma_\nu^2 + \sigma_{12}^2}(\log y_i - \boldsymbol{w}_i\boldsymbol{\delta}), \left(1 - \frac{\sigma_{12}^2}{\sigma_\nu^2 + \sigma_{12}^2}\right)\right], \quad i \in \{i : y_i > 0\}. \quad (144)$$

On the other hand, consider the case when $y_i = 0$. This produces the restriction $z_{i1} \leq 0$, while no restriction is placed upon the potential (log) outcome $z_{i2}$. In this case we can sample directly from the bivariate conditional posterior distribution of the latent data $\boldsymbol{z_i} = [z_{i1}\ z_{i2}]'$ by first drawing $z_{i1}$ from its truncated normal conditional posterior:

$$z_{i1}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y} \sim \mathcal{TN}_{(-\infty,0]}\big(\boldsymbol{r}_i\alpha, 1\big), \quad i \in \{i : y_i = 0\}, \quad (145)$$

and then drawing $z_{i2}$ from its conditional normal posterior distribution:

$$z_{i2}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y} \sim \mathcal{N}\left[\boldsymbol{w}_i\boldsymbol{\delta} + \sigma_{12}(z_{i1} - \boldsymbol{r}_i\boldsymbol{\alpha}), \sigma_2^2 - \sigma_{12}^2\right], \quad i \in \{i : y_i = 0\}. \quad (146)$$

Posterior simulation in the hurdle model proceeds by sampling from (129), (144) - (146), and either (131) or (141)- (142), depending on whether one chooses to work in the identified or non-identified space. In this case, relatively little is lost in terms of the complexity of the algorithm, or its mixing properties, by working directly with identified parameters. Finally we note that posterior simulations from the joint posterior of the hurdle model can be used to easily calculate policy-relevant parameters and therefore move beyond the narrow goal of parameter estimation. For example, these simulations can be used to summarize how hypothetical changes in the value of a covariate or set of covariates impact the probability that $y_i$ is zero, or how such changes impact the entire distribution of $\boldsymbol{y}$. Such calculations, as they involve nonlinear functions of the model parameters, are, in our view, comparably difficult to carry out from the classical perspective, perhaps allaying concerns that Bayesian approaches are "harder" [see, e.g., (Sims 2010) of this volume]. It is worth noting, though potentially underemphasized in this chapter, that this point applies to all of the nonlinear models previously discussed as well as the multivariate nonlinear models that follow.

## 4.2 Endogeneity in Nonlinear Models

The analysis of section 2.5 considered the endogeneity of a right-hand side variable in a continuous (linear) framework. In practice, of course, endogeneity concerns are not limited to models

with continuous variables and indeed, this problem often arises with discrete, censored, or ordinal outcome data.[24] In practice, unfortunately, researchers analyzing such data sometimes abandon nonlinear specifications in favor of linear models, shunning appropriate econometrics in favor of the familiarity and ease of linear methods such as IV or 2SLS.

In this section we review posterior simulation in a particular nonlinear model with an endogenous right-hand side variable. As the reader will see, the posterior simulator for this model is only slightly more complicated than the one described in section 2.5 for linear outcomes. As such, the techniques described here are really no more involved than those employed with linear models.

### 4.2.1 Posterior Simulation with an Endogenous Binary Variable

Though there are many different possibilities to consider here, let us fix ideas on a specific model, noting that the methods to follow clearly generalize to cases where the outcome is also latent or the endogenous variable is censored or ordered. Below, we consider a standard representation of a dummy endogenous variable model:

$$z_{i1} = \boldsymbol{r}_i\boldsymbol{\alpha} + u_{i1} \tag{147}$$

$$y_i = \alpha_0 + \alpha_1 D_i + \boldsymbol{s}_i\boldsymbol{\alpha}_2 + u_{i2} \tag{148}$$

where

$$D_i = I(z_{i1} > 0). \tag{149}$$

The observed responses consist of a continuous outcome $y_i$ and a binary "treatment" variable $D_i$, and the latter of these is specified to be generated by the latent variable in (147). In practice, $\alpha_1$ is commonly the object of interest as the "causal" impact of the binary variable $D_i$ on $y_i$ [see (Chamberlain 2010) of this volume for more on Bayesian estimation of treatment impacts]. However, with observational data, determining this causal impact is not a trivial exercise, as individuals self-select into treatment regimes, thereby producing a correlation between $u_1$ and $u_2$. In frequentist parlance, the presence of this correlation leads to biased and inconsistent OLS-based estimation of $\alpha_1$ using (148) only. Such theoretical concerns, as well as our understanding of the problem being studied, require that we allow for correlation among the unobservables. Therefore, we continue to make a bivariate normality assumption as in (135), with $\sigma_{12}$ capturing the role of unobserved confounding in the model.

[24] See, for example, (Li 1998; Geweke et al. 2003; Chib 2003) for Bayesian examples.

In terms of posterior simulation, the model in (147) - (149) is nearly identical to that for the hurdle model presented in the previous section. To make the connection between the two models explicit, simply define $\boldsymbol{w}_i = [1 \; D_i \; \boldsymbol{s}_i]$ and $\boldsymbol{\delta} = [\alpha_0 \; \alpha_1 \; \boldsymbol{\alpha}_2]'$. The link between the latent and observed data, as in (128) for this model, reduces to:

$$y_i = z_{i2}, \qquad D_i = I(z_{i1} > 0). \tag{150}$$

Posterior sampling of $\boldsymbol{\beta}$ then proceeds as in (129). For a sampler that navigates through the identified parameter space, elements of the (reparameterized) $\boldsymbol{\Sigma}$ can be sampled as in (141) - (142). The latent variables $z_{i1}$ are sampled independently, $i = 1, 2, \ldots, n$ from:

$$z_{i1}|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{D} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}\left(\boldsymbol{r}_i\boldsymbol{\alpha} + \frac{\sigma_{12}}{\sigma_\nu^2+\sigma_{12}^2}[y_i - \alpha_0 - \alpha_1 - \boldsymbol{s}_i\boldsymbol{\alpha}_2], 1 - \rho_{12}^2\right) & \text{if } D_i = 1 \\ \mathcal{TN}_{(-\infty,0]}\left(\boldsymbol{r}_i\boldsymbol{\alpha} + \frac{\sigma_{12}}{\sigma_\nu^2+\sigma_{12}^2}[y_i - \alpha_0 - \alpha_1 - \boldsymbol{s}_i\boldsymbol{\alpha}_2], 1 - \rho_{12}^2\right) & \text{if } D_i = 0 \end{cases}. \tag{151}$$

A posterior simulator for the dummy variable treatment effects model is given by (129), (141) - (142) and (151).

### 4.2.2 Application: A Count Data Model with Endogeneity

In practice, many models with endogeneity problems fit conveniently within the framework described by (126) - (128). One notable exception is in the analysis of count outcomes which, heretofore, has been ignored within this chapter, yet is an important specification in the analysis of microeconometric data. For this reason we pause to provide an application involving count data and choose to do so within the framework of a count outcome with an endogenous explanatory variable. Posterior simulation in such a case, unfortunately, does not proceed just by standard Gibbs steps, but instead makes use of several Metropolis-Hastings substeps to conduct the necessary sampling.

Our example comes from the study of (Lakdawalla et al. 2006).[25] These authors study how the receipt of Highly Active AntiRetroviral Therapy (HAART) for HIV-positive patients impacts the subsequent number of sexual partners. The authors note that the presence of such treatment generally improves the health and longevity of its recipients, which might then, in turn, potentially impact the sexual behavior of the treated. Furthermore, if sexual activity increases significantly in

[25] We kindly thank the authors for supplying us with their data.

response to treatment, the availability of HAART could even *reduce* social welfare, as the increased level of sexual activity upon receiving the health-improving treatment may leave the HIV-negative community at increased risk for infection.

We propose the following Bayesian framework, which is slightly different from the specification of Lakdawalla et al:

$$y_i|\boldsymbol{\beta}, \epsilon_i \overset{ind}{\sim} Po[\exp(d_i\beta_0 + \tilde{\boldsymbol{x}}_i\boldsymbol{\beta}_1 + \epsilon_i)] \tag{152}$$

$$z_i = \boldsymbol{w}_i\boldsymbol{\gamma} + u_i \tag{153}$$

$$d_i = I(z_i > 0) \tag{154}$$

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \Big| \boldsymbol{X}, \boldsymbol{W} \overset{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 = 1 \end{pmatrix}\right] \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{155}$$

The model we employ specifies that, conditional on individual $i$'s idiosyncratic term $\epsilon_i$, the number of sexual partners $y_i$ follows a Poisson distribution with mean $\exp(\boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i)$, $\boldsymbol{x}_i = [d_i \ \tilde{\boldsymbol{x}}_i]$, $\boldsymbol{\beta} = [\beta_0 \ \boldsymbol{\beta}_1']'$. The covariates $\boldsymbol{x}_i$ help to explain variation in the number of sexual partners across individuals, and contains the binary $d_i$ denoting the decision to receive the HAART regimen, which is regarded as potentially endogenous. To motivate this potential endogeneity concern, it may be the case that people in failing or poor health will be more likely to seek out and receive the therapy but less likely to participate in risky sexual activities. This possibility is handled by permitting correlation among $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ through $\sigma_{\epsilon u}$. Furthermore, we note that the addition of $\epsilon_i$ relaxes the restrictive assumption of the Poisson that the variance and mean are the same, and thus we permit overdispersion through the adoption of a Poisson-lognormal mixture.

Since the receipt of the HAART treatment is a binary outcome, we augment the likelihood function by introducing a latent variable $z_i$ such that $z_i > 0$ if $d_i = 1$ and $z_i \leq 0$ otherwise. Although most of the covariates affecting the decision to receive the therapy also affect the count outcome, we, following (Lakdawalla et al. 2006), exploit state-level variation in the availability / generosity of public insurance for HIV positive individuals. This includes accounting for two variables capturing the "medically needy threshold" set by the state, expressed as a percentage of the federal poverty line, and an indicator variable denoting whether the state's income eligibility threshold for Medicare through Supplemental Social Security Income (SSI) was less than 10 percentage points lower than the federal guideline. These two variables are included in $\boldsymbol{w}_i$ but omitted from $\boldsymbol{x}_i$. To complete the Bayesian analysis, we specify that $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}})$, $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{\gamma}})$ and $p(\boldsymbol{\Sigma}^{-1}) \propto f_W(\boldsymbol{\Sigma}^{-1}|[\kappa\boldsymbol{R}]^{-1}, \kappa)I(\sigma_u^2 = 1)$ where the indicator function notes that the $(2,2)$ element

of $\boldsymbol{\Sigma}$ is restricted to unity for identification purposes.

Following (Chib et al. 1998; Munkin and Trivedi 2003), we simulate from the joint posterior distribution of $\epsilon_i$, $\boldsymbol{\beta}$, $\boldsymbol{z}_i$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma^{-1}}$ by sampling these parameters iteratively from a Gibbs sampler, with Metropolis substeps used as needed. The conditional posterior density of $\epsilon_i$ is proportional to

$$p(\epsilon_i|\boldsymbol{\Xi}_{-\epsilon_i}, \boldsymbol{y}, \boldsymbol{d}) \propto \quad \exp[-\exp(\boldsymbol{x}_i\boldsymbol{\beta}+\epsilon_i)][\exp(\epsilon_i)]^{y_i} \exp\{-\frac{1}{2(\sigma_\epsilon^2-\sigma_{\epsilon u}^2)}[\epsilon_i - \sigma_{\epsilon u}(z_i - \boldsymbol{w}_i\boldsymbol{\gamma})]^2\}, \quad (156)$$

where $\boldsymbol{\Xi}_{-x}$ denotes the parameters other than $x$. Although $\epsilon_i$ cannot be sampled directly, as the form of (156) is uncommon, an M-H step can be employed. Specifically, a candidate draw can be sampled from a $t$ distribution centered around the mode of $\ln p(\epsilon_i|\boldsymbol{\Xi}_{-\epsilon_i}, \boldsymbol{y}, \boldsymbol{d})$, with scale parameter and degrees of freedom parameter equal to $(\nu\omega V_{\hat{\epsilon}_i})^{-1}$ and $\nu$, respectively. In practice, we choose $V_{\hat{\epsilon}_i}$ as the negative inverse Hessian of $\ln p(\epsilon_i|\boldsymbol{\Xi}_{-\epsilon_i}, \boldsymbol{y}, \boldsymbol{d})$ evaluated at the mode and both $\nu$ and $\omega$ are tuning parameters. The candidate draw $\epsilon_i^*$ is then accepted with the probability

$$\min\left\{\frac{p(\epsilon_i^*|\boldsymbol{\Xi}_{-\epsilon_i}, \boldsymbol{y}, \boldsymbol{d})q(\epsilon_i^{(t-1)})}{p(\epsilon_i^{(t-1)}|\boldsymbol{\Xi}_{-\epsilon_i}, \boldsymbol{y}, \boldsymbol{d})q(\epsilon_i^*)}, 1\right\},$$

where the $(t-1)$ superscript denotes the current value of the chain and $q(\cdot)$ denotes the proposal density. We use a similar Metropolis step to sample the parameter vector $\boldsymbol{\beta}$ whose conditional posterior is proportional to

$$p(\boldsymbol{\beta}|\boldsymbol{\Xi}_{-\boldsymbol{\beta}}, \boldsymbol{y}, \boldsymbol{d}) \propto \exp[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})'\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})]\prod_{i=1}^{n}\exp[-\exp(\boldsymbol{x}_i\boldsymbol{\beta}+\epsilon_i)][\exp(\boldsymbol{x}_i\boldsymbol{\beta}+\epsilon_i)]^{y_i}. \quad (157)$$

The proposal density used in the sampling of $\boldsymbol{\beta}$ is a multivariate $t$ distribution with location parameter $\hat{\boldsymbol{\beta}} = \arg\max \ln p(\boldsymbol{\beta}|\boldsymbol{\Xi}_{-\boldsymbol{\beta}}, \boldsymbol{y}, \boldsymbol{d})$, scale parameter of $(\mu\tau\boldsymbol{V}_{\hat{\boldsymbol{\beta}}})^{-1}$ and degrees of freedom parameter $\mu$. Again, we select $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ as the negative inverse Hessian of $\ln p(\boldsymbol{\beta}|\boldsymbol{\Xi}_{-\boldsymbol{\beta}}, \boldsymbol{y}, \boldsymbol{d})$ evaluated at the mode and both $\tau$ and $\mu$ are tuning parameters.

The latent data $\boldsymbol{z}$ are sampled independently from

$$z_i|\boldsymbol{\Xi}_{-z_i}, \boldsymbol{y}, \boldsymbol{d} \sim \begin{cases} \mathcal{TN}_{(-\infty,0]}\left(\boldsymbol{w}_i\boldsymbol{\gamma} + \sigma_{\epsilon u}\sigma_\epsilon^{-2}\epsilon_i, 1-\sigma_{\epsilon u}^2\sigma_\epsilon^{-2}\right) & \text{if } d_i = 0 \\ \mathcal{TN}_{(0,\infty)}\left(\boldsymbol{w}_i\boldsymbol{\gamma} + \sigma_{\epsilon u}\sigma_\epsilon^{-2}\epsilon_i, 1-\sigma_{\epsilon u}^2\sigma_\epsilon^{-2}\right) & \text{if } d_i = 1 \end{cases}. \quad (158)$$

The conditional posterior of the parameter vector $\boldsymbol{\gamma}$ is normal:

$$\boldsymbol{\gamma}|\boldsymbol{\Xi}_{-\boldsymbol{\gamma}}, \boldsymbol{y}, \boldsymbol{d} \sim \mathcal{N}\left(\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{d}_{\boldsymbol{\gamma}}, \boldsymbol{D}_{\boldsymbol{\gamma}}\right), \quad (159)$$

where

$$\boldsymbol{D}_{\boldsymbol{\gamma}} = [\boldsymbol{W}'\boldsymbol{W}(1 - \sigma_{\epsilon u}^2 \sigma_{\epsilon}^{-2})^{-1} + \boldsymbol{V}_{\boldsymbol{\gamma}}^{-1}]^{-1}, \tag{160}$$

$$\boldsymbol{d}_{\boldsymbol{\gamma}} = [\boldsymbol{W}'(\boldsymbol{z} - \sigma_{\epsilon u}\sigma_{\epsilon}^{-2}\boldsymbol{\epsilon})(1 - \sigma_{\epsilon u}^2 \sigma_{\epsilon}^{-2})^{-1} + \boldsymbol{V}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}] \tag{161}$$

and $\boldsymbol{W}$, $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ have been stacked over $i$ in the obvious way.

In terms of the sampling of $\boldsymbol{\Sigma^{-1}}$ we could, again, employ the reparameterization as described in (141)-(142). In this instance, however, we use the algorithm of (Nobile 2000), who provides a method for directly sampling from an inverse Wishart, conditional on a fixed diagonal element of $\boldsymbol{\Sigma}$. We express this as

$$p(\boldsymbol{\Sigma^{-1}}|\boldsymbol{\Xi}_{-\boldsymbol{\Sigma}}, \boldsymbol{y}, \boldsymbol{d}) \propto f_W\left(\boldsymbol{\Sigma^{-1}} \left| \left[\kappa R + \begin{bmatrix} \boldsymbol{\epsilon} & \boldsymbol{z} - \boldsymbol{W}\boldsymbol{\gamma} \end{bmatrix}'\begin{bmatrix} \boldsymbol{\epsilon} & \boldsymbol{z} - \boldsymbol{W}\boldsymbol{\gamma} \end{bmatrix}\right]^{-1}, n + \kappa \right.\right) I(\sigma_u^2 = 1). \tag{162}$$

Estimation results using this algorithm are listed in Table 9. Non-whites, females and less educated persons are less likely to receive the therapy and have fewer sexual partners. The HAART regimen is found to have a positive and strong impact on sexual behavior. Specifically, the results suggest that the receipt of the regimen increases the mean number of sexual partners by about $[\exp(1.31) - 1] \times 100\% = 271\%$. Since the mean number of partners in the sample is 2.16, the marginal effect suggests that treated individuals have 5.85 additional partners. Among the two instrumental variables employed, the "medically needy threshold" proves to be empirically important, and the associated coefficient has a probability of being positive near one, indicating that individuals who are eligible for Medicaid through a medically needy program are more likely to receive the therapy. Finally, the variance of the error term $\epsilon_i$ is about 1.74, indicating some overdispersion for the conditionally Poisson-distributed number of partners outcome. The covariance estimate between $\epsilon_i$ and $u_i$ also shows that unobservables affecting the number of partners and the treatment are negatively correlated, consistent with the notion that people who are less healthy are less involved in risky sexual behavior but more active in seeking the treatment. These results are qualitatively quite similar to those reported in (Lakdawalla et al. 2006), although our model and approach differ slightly from theirs.

Table 9: Posterior means, standard deviations and probabilities of being positive of the parameters

| Variable | $E(\beta\|D)$ | $Std(\beta\|D)$ | $Pr(\beta > 0\|D)$ |
|---|---|---|---|
| Partners equation | | | |
| Age | -0.0464 | 0.00549 | 0 |
| Non-white | -0.133 | 0.1 | 0.0913 |
| Female | -0.584 | 0.107 | 0 |
| Less than HS degree | -0.608 | 0.154 | 0.00015 |
| High school degree | -0.676 | 0.14 | 0 |
| Some college or AA degree | -0.4 | 0.139 | 0.00321 |
| State per capita income | 0.063 | 0.035 | 0.965 |
| Percent living in urban areas | -0.0225 | 0.0129 | 0.0411 |
| Abortion rate | 0.0228 | 0.00954 | 0.991 |
| Percent thinking homosexuality wrong | -5.05 | 1.44 | 0.000513 |
| Percent praying several times a week | 7.03 | 1.96 | 1 |
| HAART | 1.31 | 0.313 | 0.999 |
| HAART equation | | | |
| Age | 0.000701 | 0.00413 | 0.567 |
| Non-white | -0.212 | 0.0745 | 0.00206 |
| Female | -0.15 | 0.0766 | 0.025 |
| Less than HS degree | -0.263 | 0.119 | 0.0128 |
| High school degree | -0.136 | 0.112 | 0.113 |
| Some college or AA degree | -0.126 | 0.111 | 0.128 |
| State per capita income | -0.0534 | 0.027 | 0.0234 |
| Percent living in urban areas | 0.0267 | 0.0098 | 0.997 |
| Abortion rate | -0.0172 | 0.00731 | 0.00925 |
| Percent thinking homosexuality wrong | 2.64 | 1.4 | 0.969 |
| Percent praying several times a week | -2.95 | 1.84 | 0.0546 |
| Medically needy threshold | 0.00416 | 0.00173 | 0.992 |
| SSI threshold $> 65\%$ of FPL | 0.115 | 0.156 | 0.772 |
| Covariance matrix | | | |
| Variance $\sigma_\epsilon^2$ | 1.74 | 0.228 | 1 |
| Covariance $\sigma_{\epsilon u}$ | -0.912 | 0.19 | 0 |

### 4.3 Treatment Effects Models

A generalization of the model in the previous section is to explicitly consider the *counterfactual* or *potential* outcome. This represents the outcome the agent would have experienced had he / she made a different treatment decision than the one actually made. Consistent with the specification

in (126) - (128), we write this model as a system of three latent variable equations:[26]

$$z_{i2} = \boldsymbol{w}_i\boldsymbol{\theta} + u_{i2} \tag{163}$$

$$z_{i1} = \boldsymbol{x}_i\boldsymbol{\beta}_1 + u_{i1} \tag{164}$$

$$z_{i0} = \boldsymbol{x}_i\boldsymbol{\beta}_0 + u_{i0} \tag{165}$$

where

$$D_i = I(z_{i2} > 0) \tag{166}$$

$$y_i = D_i z_{i1} + (1 - D_i) z_{i0} \tag{167}$$

and

$$\left[ \begin{array}{c} u_{i2} \\ u_{i1} \\ u_{i0} \end{array} \right] \Bigg| \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\Sigma} \sim \mathcal{N}\left( \left[ \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right], \left[ \begin{array}{ccc} 1 & \sigma_{21} & \sigma_{20} \\ \sigma_{21} & \sigma_1^2 & \sigma_{10} \\ \sigma_{20} & \sigma_{10} & \sigma_2^2 \end{array} \right] \right) \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{168}$$

Equations (166) and (167) represent the mapping between the observed and latent data in the potential outcomes model. Equation (163) describes the treatment decision, whose marginal analysis is identical to the probit analysis of section 3.1.1. Equations (164) and (165) describe the outcome (or potential outcome) in each treatment regime. For example, if $D_i = 1$, then the treated outcome $z_{i1}$ is observed, while the untreated outcome $z_{i0}$ is not. Conversely, when $D_i = 0$, the untreated outcome $z_{i0}$ is observed while the treated outcome is not.

This model, just like the previous models of this section, can be stacked into vector/matrix form for each individual, letting $\boldsymbol{z}_i = [z_{i2}\ z_{i1}\ z_{i0}]'$, $\boldsymbol{\beta} = [\boldsymbol{\theta}'\ \boldsymbol{\beta}_1'\ \boldsymbol{\beta}_0']'$ and

$$\boldsymbol{X}_i = \left[ \begin{array}{ccc} \boldsymbol{w}_i & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{x}_i \end{array} \right]. \tag{169}$$

Therefore, we find ourselves in a familiar situation when faced with the task of posterior simulation in the potential outcomes model. The parameter vector $\boldsymbol{\beta}$ will be sampled from (129) and the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ can be sampled from (131) using the method of (Nobile 2000) if the (1,1) element of $\boldsymbol{\Sigma}$ is set to unity. Alternatively, we can choose to work in the non-identified parameter space and post-process the draws to restrict our focus on identifiable parameters. In terms of the latent data, two latent quantities must be drawn for each individual: First, latent values of $z_{i2}$ will be drawn for each individual from a univariate truncated normal, with conditional

[26] Covariates can also change with the regime, though we do not consider this in the notation below.

support restricted by the observed value of $D_i$. Second, the potential (missing) outcome will also be sampled for each individual from the corresponding conditional normal defined by (127). We omit the details of this procedure here, as it follows similarly to those described earlier, and complete details can be found in (Koop, Poirier and Tobias 2007: 225-229).

In the potential outcomes framework, parameters of interest often center around the outcome gain (or loss) from receipt of treatment: $z_{i1} - z_{i0}$. Parameters that garner the most attention in this literature include the Average Treatment Effect ($ATE$):

$$ATE(\boldsymbol{\beta}, \boldsymbol{x}) = \boldsymbol{x}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \tag{170}$$

the effect of Treatment on the Treated ($TT$):

$$TT(\boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{z}, D(\boldsymbol{z}) = 1) = \boldsymbol{x}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\rho_{21}\sigma_1 - \rho_{20}\sigma_0)\frac{\phi(\boldsymbol{z\theta})}{\Phi(\boldsymbol{z\theta})} \tag{171}$$

and the Local Average Treatment Effect (LATE):

$$LATE(\boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{z}, \tilde{\boldsymbol{z}}, D(\boldsymbol{z}) = 0, D(\tilde{\boldsymbol{z}}) = 1) = \boldsymbol{x}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + (\rho_{21}\sigma_1 - \rho_{20}\sigma_0)\left(\frac{\phi(\tilde{\boldsymbol{z}}\boldsymbol{\theta}) - \phi(\boldsymbol{z\theta})}{\Phi(\tilde{\boldsymbol{z}}\boldsymbol{\theta}) - \Phi(\boldsymbol{z\theta})}\right), \tag{172}$$

where $\rho_{jk}$ denotes the correlation parameter between $\boldsymbol{u_j}$ and $\boldsymbol{u_k}$.

ATE summarizes the average gain (or loss) from treatment, TT represents the expected gain (or loss) from treatment for those actually taking the treatment (at a given set of characteristics $\boldsymbol{z}$), and LATE denotes the expected gain (or loss) from treatment for those that would receive the treatment at $\tilde{\boldsymbol{z}}$ but would not receive the treatment at $\boldsymbol{z}$. (Imbens and Angrist 1994) introduce the LATE parameter and interpret it as a treatment effect for a subgroup of "compliers" - individuals whose treatment behavior can be manipulated through the presence (or absence) of the instrument. For example, in the (Angrist 1990) study of the Vietnam-era draft, LATE will recover the average earnings gain (or loss) from military service for those who are induced to join the military because of a low draft lottery number, but otherwise would not serve. When treatment effects differ across individuals, different instruments define different LATE parameters - in the Angrist example, the results would not speak to the impact of military service on the post-service earnings of those who joined the military voluntarily. (Heckman, Tobias and Vytlacil 2001; Heckman, Tobias and Vytlacil 2003) provide further results, discuss mean treatment effect parameters that are not covariate-dependent and provide asymptotic derivations under a variety of distributional assumptions. (Chamberlain 2010) of this volume also discusses more details relating to Bayesian approaches to treatment effect modeling.

The expressions above can be evaluated at particular values of $\boldsymbol{x}$ and $\boldsymbol{z}$ and the posterior simulations of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ used to estimate and characterize the posterior uncertainty regarding these average treatment impacts. Alternatively, one can average over the covariates' values to eliminate the dependence of these expressions on $\boldsymbol{x}$ and $\boldsymbol{z}$ [e.g., (Chib and Jacobi 2007)].

The mean effects listed above are certainly interesting, but also rather limiting: they summarize various mean treatment impacts for different subpopulations. Other quantities, such as $\text{Var}(z_1 - z_0)$ or quantiles of $z_1 - z_0$ are also of interest, but receive minimal attention in this literature.

The reason motivating this restricted focus lies with the cross-regime covariance parameter $\sigma_{10}$. When we consider the likelihood function for this model,

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \boldsymbol{y}, \boldsymbol{D}) = \prod_{\{i:D_i=1\}} p(D_i = 1, y_i^{(1)}) \prod_{\{i:D_i=0\}} p(D_i = 0, y_i^{(0)}), \tag{173}$$

with $y_i^{(j)} = z_{ij}$, it is clear that this parameter does not enter the likelihood and thus is not identified. That is, observations will *either* belong to regime 0 or regime 1, but never both. As such, the likelihood does not directly inform us about $\sigma_{10}$. However, many features of the outcome gain $z_1 - z_0$ will depend on the cross-regime covariance $\sigma_{10}$, leaving the researcher wanting to do and say more, but typically resigning herself to focus on identifiable quantities like the mean effects listed above.

When conducting simulation-based posterior inference using the model described above, however, posterior simulations regrading $\sigma_{10}$ are produced, potentially enabling an expanded focus beyond conventional treatment effect parameters. (Vijverberg 1993) first noticed the possibility of learning about $\sigma_{10}$. These ideas were refined and a Gibbs sampling algorithm for the normal model produced by (Koop and Poirier 1997). (Chib and Hamilton 2000) address this issue by setting the cross-regime parameter to zero, and derive and apply posterior simulators for a variety of non-normal sampling models under this restriction. (Poirier and Tobias 2003; Li, Poirier and Tobias 2004) further describe the nature of learning that takes place regarding $\sigma_{10}$. Since $\boldsymbol{\Sigma}$ must be positive definite, they show that the *conditional* support of the non-identified correlation $\rho_{10}$ is the interval:

$$\rho_{10}|\rho_{21}, \rho_{20} \in \left( \rho_{21}\rho_{20} - [(1-\rho_{21}^2)(1-\rho_{20}^2)]^{1/2}, \rho_{21}\rho_{20} + [(1-\rho_{21}^2)(1-\rho_{20}^2)]^{1/2} \right). \tag{174}$$

Thus, as the data pins down the values of the identified correlations $\rho_{21}$ and $\rho_{20}$, learning about $\rho_{10}$ takes place given the support restrictions above. The extent of this learning, however, is seriously

limited, as the shape of the posterior within the bounds above is simply the conditional prior for the non-identified correlation. Nonetheless, the bounds above can be informative, particularly when unobserved confounding is large, and can serve to update our prior beliefs about the cross-regime correlation, potentially enabling the researcher to characterize something beyond mean treatment parameters. In the most recent statement on this issue, (Chib 2007) suggests working with the likelihood for the observed data rather than the potential outcomes, noting that such an approach improves the mixing of the posterior simulations and also frees the researcher from dealing with the non-identified correlation parameter.

## 4.4 The Multinomial Probit Model

The multinomial probit (MNP) model (see, e.g., (Geweke, Keane and Runkle 1994,1997) or (Train 2003, chapter 5) and the references cited therein) also maps directly into the framework given by (126)- (128). In the MNP model, an agent makes a single choice among $J$ alternatives. We let $y_i$ represent the observed choice made by agent $i$, and enumerate the alternatives so that $y_i \in \{0, 1, \ldots, J-1\}$.

The MNP model is derived from a random utility framework where a multivariate latent variable, generated as in (127), is specified to describe the utility afforded by each alternative. In practice, of course, utility needs to be normalized both for level and scale, as the observed choices made by agents would be unaffected by the addition of a common constant to each utility level or by multiplication of utility by a constant. The issue of level normalization is typically accomplished by considering differences in utility relative to some base alternative, which, here, we treat as alternative zero. This focus on utility differences reduces the dimension of the model to $J-1$ rather than $J$, and we assume (126) - (128) applies to the analysis of such differences in utility. (Rossi and Allenby 2010) of this volume provide more details on analysis of the MNP models as well as the multivariate probit model considered in the next section. Relatedly, multimomial logit (MNL) and mixed logit models are commonly used to analyze this type of data. We do not consider these models here, unfortunately, but refer the reader to (Rossi, Allenby and McCulloch 2005: sections 3.11 and chapter 5; Train 2003) for more details.

As for scale normalization in the MNP model, we again have several possibilities, which already have been discussed. First, we can normalize a diagonal element of the $(J-1) \times (J-1)$ covariance

matrix $\mathbf{\Sigma}$. The posterior conditional distribution of $\mathbf{\Sigma^{-1}}$ is then a Wishart, with a diagonal entry restricted to unity. (McCulloch, Polson and Rossi 2000) discuss a reparameterization of the restricted covariance matrix that enables sampling of the elements of $\mathbf{\Sigma}$ based on simple Gibbs steps. (Nobile 2000) also provides an algorithm that enables direct sampling from a Wishart given such a diagonal restriction. Either of these approaches can be applied to form a sampler that navigates through the identified parameter space.

Alternatively, the restriction can be ignored, a standard Wishart prior employed for $\mathbf{\Sigma^{-1}}$, leading to a standard posterior conditional for $\mathbf{\Sigma^{-1}}$. Identified functions of parameters can then be calculated from such a sampler. In this approach, "off-the-shelf" routines can be used to perform the sampling, yielding computational simplicity and improved mixing properties.

It remains to discuss the sampling of the latent data within the MNP model. To this end, let $\boldsymbol{z}_i = [z_{i1}\ z_{i2}\ \cdots\ z_{iJ-1}]'$ represent the latent differenced utility vector in (127). The link in (128) for the multinomial probit reduces to:

$$y_i = \begin{cases} 0 & \text{if } \max\{z_{il}\}_{l=1}^{J-1} \leq 0 \\ j & \text{if } \max\{z_{il}\}_{l=1}^{J-1} = z_{ij} \end{cases} . \tag{175}$$

Therefore, the posterior conditional distributions of each $\boldsymbol{z}_i$ are independent across individuals, and the posterior conditional density for each latent vector $\boldsymbol{z}_i$ is a normal distribution, truncated to a "cone" defined by the restrictions above.

While it is not possible to draw directly from this multivariate truncated distribution, (Geweke 1991) provides an alternative, specialized for use in the MNP model by (McCulloch and Rossi 1994). They note that the posterior conditional distributions for each $z_{ij}$ are univariate truncated normal, with conditional supports defined through the restrictions in (175). As such, we can apply methods like those described for the probit model to generate a series of univariate truncated normal draws for each element of the latent variable vector $\boldsymbol{z}_i$. For example, if alternative 0 is chosen by agent $i$, then each of the $z_{ij}$ are restricted to be non-positive. Similarly, if $j \neq 0$ is chosen by agent $i$, then $z_{ij}$ is restricted to be positive and at least as large as all of the other $z_{il}$. Thus, sampling of the latent data involves first calculating the conditional mean and variance of each $z_{ij}$ from the $J-1$ dimensional multivariate normal in (127), determining the support restrictions on $z_{ij}$ given the observed choice made by agent $i$, and then sampling from the resulting univariate truncated normal. This process is repeated for each element of $\boldsymbol{z}_i$ (and then for all $i$), noting that the most recent simulations of the $z_{il}$ are used when calculating the conditional mean and

variance for successive elements of $\boldsymbol{z}_i$ Finally, apart from the sampling mechanism for the MNP model discussed so far, it is worthwhile to note from (Keane 1992) that identification of the MNP model can be quite fragile.

## 4.5 The Multivariate Probit Model

The multivariate probit (MVP) model [e.g., (Chib and Greenberg 1998)] is quite similar in structure to the multinomial probit of the previous section and shares the hierarchical, latent representation that unites the models in this section. In the MVP, agents continue to face a choice among $J$ different alternatives, yet are not restricted to choose a single element among the set. Furthermore, factors not observed by the econometrician may generate correlation among these choices, motivating the desire to consider outcomes jointly rather than individually.

For analysis of the MVP model, we let $\boldsymbol{y}_i = [y_{i1}\ y_{i2}\ \cdots\ y_{iJ}]'$, $y_{ij} \in \{0, 1\}\ \forall\ i, j$ and

$$\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_{iJ} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_J \end{bmatrix}. \tag{176}$$

In the above notation we specify that the covariates vary both with the agent and the alternative, and we also allow for alternative-specific slope parameters. The former of these assumptions may or may not be true in practice, though such data differences do not significantly affect the development of the model or posterior simulator. The restrictions in (128) for the MVP model reduce to a series of probit-like restrictions:

$$y_{ij} = I(z_{ij} > 0), \quad j = 1, 2, \ldots, J. \tag{177}$$

The identification problem in the MVP model is slightly different from that in the MNP model, given the nature of the observed responses in (177). In particular, if we were to multiply the latent equation in (127) by a diagonal matrix $\boldsymbol{C}$, then one can show that

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{\beta} = \boldsymbol{\beta}_0, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0) = \Pr\big(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{\Sigma}_0\boldsymbol{C}'\big), \tag{178}$$

where $\tilde{\boldsymbol{\beta}}$ is constructed by multiplying each $\boldsymbol{\beta}_{0j}$ by the $(j, j)$ element of $\boldsymbol{C}$. In other words, each latent variable equation can be rescaled yet leave the likelihood function for the observed data unchanged.

(Chib and Greenberg 1998) present a Bayesian analysis of the MVP model, working with the covariance matrix $\boldsymbol{\Sigma}$ in restricted correlation form (and thus work directly with identifiable parameters) and propose a tailored Metropolis-Hastings step for sampling this restricted correlation matrix. Alternatively, (Edwards and Allenby 2003; Rossi, Allenby and McCulloch 2005) advocate working in the non-identified parameter space and adopting a sampler that only requires simulation from standard distributions. Specifically, they suggest working with an unrestricted $\boldsymbol{\Sigma}$. The regression parameters $\boldsymbol{\beta}$ are sampled from (129), the inverse covariance matrix is sampled from (131) and each latent vector $\boldsymbol{z}_i$ is again multivariate truncated normal, with conditional support restrictions given by (177). As such, the algorithm of (Geweke 1991) can be applied, and each component $z_{ij}, j = 1, 2, \ldots, J$ drawn from a univariate truncated normal.

To deal with the identification problem in the (Edwards and Allenby 2003) approach, the posterior simulations are post-processed to report identifiable quantities. In the general case, this requires setting $\boldsymbol{C}$ to be a $J \times J$ diagonal matrix with diagonal entries $\{\sigma_{j,j}^{-1/2}\}$ and calculating $\boldsymbol{C\beta}$ and $\boldsymbol{C\Sigma C}'$ as the identifiable regression parameters and covariance matrix, respectively. This approach to posterior simulation of the MVP model is attractive in that it only requires sampling from standard distributions and offers improved mixing properties in practice.

# 5 Duration Models

Microeconometric applications often involve the analysis of duration data, with Bayesian applications including the investigation of unemployment duration (Lancaster 1979; Li 2006), employment duration (Campolieti 1997), and time spent in bankruptcy (Li 1999), among others. A particularly salient feature of economic applications is the possibility (or probability) that such variables exhibit state dependence, meaning that the probability of exiting a spell may depend on the length of time the person has remained within that spell.

One approach to duration modeling specifies that the duration random variable of interest, $T$ is continuous, with probability density function denoted as $f(t)$. In this case, the cumulative distribution function, denoted $F(t)$ is obtained as $F(t) = \text{Prob}(T < t) = \int_0^t f(u)du$, while the survivor function (or the probability that $T$ continues at time $t$), is simply one minus the cumulative

distribution function:

$$S(t) = \text{Prob}(T \geq t) = 1 - \text{Prob}(T < t) = 1 - F(t).$$

In many applications of duration modeling, interests center around the hazard function (Cox 1972), $\lambda(t) = f(t)/S(t)$. This is the instantaneous probability of $T$ ending exactly at $t$, conditional on the event that it has last until $t$. The special interest in the hazard often leads researchers to embrace it as the "primitive" and from it the implied survivor and density functions are derived. Letting $\Lambda(t) = \int_0^t \lambda(u)du$ denote the integrated hazard, it can be easily shown that $S(t) = \exp[-\Lambda(t)]$ and $f(t) = \exp[-\Lambda(t)]\lambda(t)$.

To conduct a Bayesian duration analysis, we assume that we have $n$ observations on the duration random variable $T$ that are (conditionally) independent from one another, and denote these as $t_1$, $t_2$, $\cdots$, $t_n$. The likelihood function of the model is

$$p(\boldsymbol{y}|\boldsymbol{\Xi}) = \prod_{i=1}^{n} \exp[-\Lambda(t_i)]\lambda(t_i),$$

where $\boldsymbol{y} = [t_1 \ t_2 \ \cdots \ t_n]'$ denotes the data and $\boldsymbol{\Xi}$ all the parameters. If we specify that the hazard function remains constant over time, i.e., $\lambda(t) = \lambda$, the integrated hazard function reduces to $\Lambda(t) = \lambda t$, and the likelihood function becomes $p(\boldsymbol{y}|\boldsymbol{\Xi}) = \prod_{i=1}^{n} \exp(-\lambda t_i)\lambda$. In this case only one parameter, $\lambda$, appears in the likelihood function. Recognizing this as an exponential sampling likelihood, the gamma prior:

$$p(\lambda) = f_G(a, b) = b^{-a}\Gamma(a)^{-1}\lambda^{a-1}\exp(-\lambda b^{-1}),$$

is known to be conjugate. Specifically, if this prior is employed, we obtain:

$$p(\lambda|\boldsymbol{y}) = f_G(a + n, (b^{-1} + \sum_{i=1}^{n} t_i)^{-1}).$$

The foregoing approach, based on the assumption of a constant hazard, is quite restrictive. A more flexible alternative is to specify that the hazard is constant over suitably short intervals, but potentially different across intervals. To this end we divide the time horizon into $K$ shorter periods and specify that the hazard function remains constant within each period, but varies across different periods [e.g., (Holford 1976)]. In other words, $\lambda = \lambda_1$ within the first period, $\lambda = \lambda_2$ within the second period, and so on.

Correspondingly, we also divide duration $t_i$ into $K$ parts, $t_{i1}$, $t_{i2}$, $\cdots$, $t_{iK}$, and use $d_{i1}$, $d_{i2}$, $\cdots$, $d_{iK}$ to indicate whether $t_i$ ends in a particular period. For example, if $t_i$ ends in the middle of the fourth period, $t_{i1} = t_{i2} = t_{i3} = 1$, $t_{i4} = \frac{1}{2}$, $t_{i5} = t_{i6} = \cdots = t_{iK} = 0$, only $d_{i4} = 1$ and all other $d_{ik} = 0$. The likelihood function for the piecewise constant baseline hazard is

$$p(\boldsymbol{y}|\boldsymbol{\Xi}) = \prod_{i=1}^{n} \exp(-\sum_{k=1}^{K} \lambda_k t_{ik}) \prod_{k=1}^{K} \lambda_k^{d_{ik}}.$$

As with the constant hazard case, we can specify a common gamma prior for the piecewise hazards:

$$\lambda_k \overset{iid}{\sim} G(a, b), \quad k = 1, 2, \ldots, K,$$

producing

$$\lambda_k|\boldsymbol{y} \overset{ind}{\sim} G(a + \sum_{i=1}^{n} d_{ik}, (b^{-1} + \sum_{i=1}^{n} t_{ik})^{-1}), \quad k = 1, 2, \ldots, K.$$

## 5.1 Discrete Time Approaches

Sometimes we do not know when duration $t_i$ ends exactly, but know that it ends within one of the $K$ periods, e.g., $k = 4$. A discrete model can be used for this type of data (Campolieti 1997). Denote $\Phi(\gamma_k)$ as the probability $t_i$ continues in period $k$, conditional on the fact that it lasts until period $k-1$, where $\Phi(\cdot)$ stands for the standard normal cumulative distribution function and $\gamma_k$ is a period specific parameter, similar to $\lambda_k$ in the continuous time model. Let $s_{i1}$, $s_{i2}$, $\cdots$, $s_{iK}$ be indicator variables denoting whether a duration continues in period $k$. For instance, if duration $t_i$ ends in the fourth period, $s_{i1} = s_{i2} = s_{i3} = 1$ and $s_{i4} = s_{i5} = \cdots = s_{iK} = 0$. The likelihood function of the discrete model is

$$p(\boldsymbol{y}|\boldsymbol{\gamma}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \Phi(\gamma_k)^{s_{ik}} [1 - \Phi(\gamma_k)]^{d_{ik}},$$

where, as defined in the continuous model, $d_{ik}$ indicates whether duration $t_i$ ends in period $k$. A normal prior such as $\phi(\mu_\gamma, V_\gamma)$ is commonly used for the period-specific rate parameter $\gamma_k$.

Following (Albert and Chib 1993a; Albert and Chib 1993b; Campolieti 1997), one can use a Gibbs sampler with data augmentation to simulate draws from the posterior distribution of the parameters. For each duration $t_i$, and for each period $k$ within which duration $t_i$ continues or ends, the

likelihood function is augmented with a latent variable $y_{ik} = \gamma_k + \epsilon_{ik}$, where $\epsilon_{ik} \overset{iid}{\sim} \mathcal{N}(0,1)$, such that $y_{ik} > 0$ if $t_i$ continues in period $k$ and $y_{ik} \leq 0$ if it ends in period $k$.

The Gibbs sampler consists of two steps. In the first step, $y_{ik}$ is drawn from a $\mathcal{TN}_{(0,\infty)}(\gamma_k, 1)$ distribution if $t_i$ continues in period $k$ and from $\mathcal{TN}_{(-\infty,0]}(\gamma_k, 1)$ if it ends in period $k$. The second step draws $\gamma_k$ from $\mathcal{N}(D_k d_k, D_k)$, where

$$D_k = [V_\gamma^{-1} + \sum_{i=1}^{n} I(s_{ik} = 1 \text{ or } d_{ik} = 1)]^{-1} \quad \text{and} \quad d_k = V_\gamma^{-1}\mu_\gamma + \sum_{i=1}^{n} I(s_{ik} = 1 \text{ or } d_{ik} = 1) y_{ik}.$$

## 5.2 Other Generalizations

Duration models can be extended in various directions to accommodate additional features. Sometimes it is reasonable to specify that the baseline hazards in neighboring periods are similar to each other, and therefore it is natural to impose a smoothing prior on the piecewise constant baseline hazards [see (Campolieti 2000) for an application in the discrete time model]. For example, we can impose the following prior on the first difference in the baseline hazards of adjacent periods: $\lambda_{k+1} - \lambda_k \sim \mathcal{N}(0, \eta)$. Smaller values of $\eta$ place stronger prior information on the baseline hazards and make the estimates of baseline hazards smoother. The smoothness of the baseline hazard estimates also depends on the order of differencing, and it is possible to specify a prior using a higher order of differencing among the baseline hazards.

Right censoring represents another common feature of duration data. Assume that the censoring occurs at the end of period $K$ so that we only observe duration $t_i$ up to that point. If $t_i$ ends in a period beyond $K$, the timing of the termination will not be observed. Importantly, the likelihood functions discussed previously automatically take into account this issue. Note that if duration $t_i$ is censored at the end of period $K$ and $t_i$ continues beyond that point, the indicator variables $d_{i1}$, $d_{i2}$, $\cdots$, $d_{iK}$ will all be zero. For the continuous model, the likelihood function of $t_i$ reduces to $\exp(-\sum_{k=1}^{K} \lambda_k t_{ik})$, corresponding to the probability of survival at the end of period $K$. In the discrete model, the likelihood function becomes $\prod_{k=1}^{K} \Phi(\gamma_k)^{s_{ik}}$, which is also the probability of surviving until period $K$.

Another well-established result in duration analysis is that the failure to account for the heterogeneity in hazards results in identification problems in duration dependence estimation. Variation

in hazard rates across agents can be explained by observable characteristics that change over time. Following the proportional hazard analysis framework (Cox 1972), we incorporate a $1 \times j$ vector of time-varying covariates $\boldsymbol{x}_{ik}$ into the duration model. In the continuous time model, the hazard that $t_i$ ends in period $k$ is proportional to the baseline hazard $\lambda_k$: $\lambda_{ik} = \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\lambda_k$, where $\boldsymbol{\beta}$ is a $j \times 1$ vector representing the impacts of covariates $\boldsymbol{x}_{ik}$ on the hazard rate. The likelihood function accommodating such time-varying covariates is

$$p(\boldsymbol{y}|\boldsymbol{\Xi}) = \prod_{i=1}^{n} \exp[-\sum_{k=1}^{K} \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\lambda_k t_{ik}] \prod_{k=1}^{K} [\exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\lambda_k]^{d_{ik}}.$$

In the discrete time model,time-varying covariates are incorporated by replacing $\Phi(\gamma_k)$ with $\Phi(\gamma_k + \boldsymbol{x}_{ik}\boldsymbol{\beta})$ in the likelihood function.

The modeling of time-varying covariates is an attempt to deal with the above identification problem, but it is unlikely that the heterogeneity in hazard can be captured entirely by observables. As such, it is also important to model the unobserved heterogeneity in hazard across individuals. Consistent with the proportional hazard specification, we add unobserved heterogeneity to the hazard function in the same way as we introduced observed heterogeneity in the hazard. In the continuous model we change the hazard function to $\lambda_{ik} = \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\eta_i\lambda_k$, where $\eta_i$ represents the individual level unobserved heterogeneity.

Researchers [e.g., (Lancaster 1979)] often specify that $\eta_i$ follows a Gamma distribution, $\eta_i|v \overset{iid}{\sim} G(v, v^{-1})$, so that it has unit mean and variance equal to $v^{-1}$. In the discrete model, we change the continuation probability to $\Phi(\gamma_k + \boldsymbol{x}_{ik}\boldsymbol{\beta} + \alpha_i)$ to reflect the influence of the individual specific random effect on the hazard function. A standard, simple parametric assumption for the heterogeneity terms would be $\alpha_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Alternatively, we could be more flexible and allow a nonparametric specification of the unobserved heterogeneity through the use of a Dirichlet process prior as in (Campolieti 2001). Finally, duration data sometimes appear in a hierarchical form (Guo and Rodriguez 1992; Sastry 1997; Bolstad and Manda 2001; Li 2007). For example, we may observe durations for individuals clustered within the same household, the same city, and so on. In such cases, we can capture unobserved heterogeneity in the hazard at the various levels via the proportional hazard approach. The following example provides an illustration.

## 5.3 An Example

We modify the application in (Li 2007) who studies the timing of high school dropout decisions, using data from the High School and Beyond Longitudinal survey. The full model estimated by Li accounts for the heterogeneity in hazard rates at the individual, school and state levels. To simplify our discussions, we consider here a simpler version and model only sources of individual level heterogeneity, ignoring any possible correlations that may take place for individuals within the same school or state. Adopting the continuous time model, we define the hazard of dropping out of high school during month $k$ for individual $i$ as $\lambda_{ik} = \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\eta_i\lambda_k$, where $\boldsymbol{x}_{ik}$ is a $j \times 1$ vector of individual level covariates, $\eta_i \overset{iid}{\sim} G(v, v^{-1})$ represents individual $i$'s random effect in the hazard function, and $\lambda_k$ corresponds to the piecewise constant baseline hazard in month $k$. The likelihood function for the model is

$$p(\boldsymbol{y}|\boldsymbol{\Xi}) = \prod_{i=1}^{n} v^v \Gamma(v)^{-1} \eta_i^{v-1} \exp(-\eta_i v) \exp[-\sum_{k=1}^{K} \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\eta_i\lambda_k t_{ik}] \prod_{k=1}^{K} [\exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\eta_i\lambda_k]^{d_{ik}}.$$

We specify the following priors: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta})$, $\lambda_k \overset{iid}{\sim} G(a_\lambda, b_\lambda)$, $v \sim G(a_v, b_v)$, where $\boldsymbol{\beta_0} = \mathbf{0}$, $\boldsymbol{V_\beta} = 1000\boldsymbol{I_j}$, $a_\lambda = a_v = 0.01$ and $b_\lambda = b_v = 100$.

A Metropolis-within-Gibbs algorithm is used to generate samples from the joint posterior distribution. The parameters $\{\lambda_k\}_{k=1}^K$, $\{\eta_i\}_{i=1}^n$, are sampled using Gibbs steps by drawing, in order, from:

$$\lambda_k|\boldsymbol{y}, \{\eta_i\}_{i=1}^n, \boldsymbol{\beta}) \overset{ind}{\sim} G(a_\lambda + \sum_{i=1}^{n} d_{ik}, [b_\lambda^{-1} + \sum_{i=1}^{n} \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\eta_i t_{ik}]^{-1}),$$

and,

$$\eta_i|\boldsymbol{y}, \{\lambda_k\}_{k=1}^K, v, \boldsymbol{\beta} \overset{ind}{\sim} G(v + \sum_{i=1}^{n} d_{ik}, [v + \sum_{i=1}^{n} \exp(\boldsymbol{x}_{ik}\boldsymbol{\beta})\lambda_k t_{ik}]^{-1}).$$

The conditional posterior distributions of $v$ and $\boldsymbol{\beta}$ are not of a known form and therefore cannot be sampled directly. For these parameters, we employ M-H steps.

Table 10: Posterior estimates and marginal effects of the coefficients

| Variable/Parameter | E($\beta|D$) | Std($\beta|D$) | P($\beta > 0|D$) | Marginal effect |
|---|---|---|---|---|
| Female | -0.14 | 0.0709 | 0.0201 | -12.8 |
| Minority | -0.186 | 0.0824 | 0.00962 | -16.7 |
| Family income ($10,000) | -0.0524 | 0.0368 | 0.0674 | -5.04 |
| Base year test score | -0.838 | 0.0446 | 0 | -56.7 |
| Father's education (year) | -0.0505 | 0.0107 | 0 | -4.92 |
| Mother's education (year) | -0.0771 | 0.0122 | 0 | -7.41 |
| Number of siblings | 0.0947 | 0.0206 | 1 | 9.95 |
| Dropout eligibility | 0.654 | 0.112 | 1 | 93.6 |
| Variance parameter ($v^{-1}$) | 0.929 | 0.13 | 1 | |

In Table 10 we list summary posterior statistics and also calculate the marginal effect of a covariate which corresponds to the percentage change in the dropout hazard due to a one-unit increase in the covariate $x_j$, or $[\exp(\beta_j) - 1] \times 100$. Our results show that being eligible to drop out of high school under compulsory schooling laws increases the dropout hazard of an individual by 93.6 percent. An increase of $10,000 in parental income decreases the dropout hazard by 5.04 percent, while variance parameter estimate $v^{-1}$ indicates considerable unobserved variation across individuals in the dropout hazard.

# 6 Conclusion

We have reviewed Bayesian approaches to estimation in many models commonly encountered in microeconomics. While not exhaustive, the models considered in this chapter are among the most widely-used in practice and can serve to accommodate many of the data types and some of the econometric problems that the practitioner will face. While not completely flexible, as nearly all the posterior simulators have been presented under conditionally normal sampling assumptions, we have provided references to the literature for extensions of the basic framework, and noted the "modularity" of MCMC methods. That is, existing computational techniques can be employed to expand the sampling window to the class of scale mixtures or finite mixtures of normals, for example, and implementation of these steps proceeds in largely the same way regardless of the model employed (Geweke and Keane 2001). Finally, examples using real data for many different models have been provided and code is made available to the interested practitioner for inspection, refinement, or further modification.

# References


[1] Aitken, M. and N. Longford (1986). 'Statistical Modeling Issues in School Effectiveness Studies (with discussion)'. *Journal of the Royal Statistical Society, Series A*, 149: 1-42.

[2] Albert, J. and S. Chib (1993a). 'Bayesian Analysis of Binary and Polychotomous Response Data'. Journal of the American Statistical Association, 88: 669-79.

[3] Albert, J. H. and S. Chib (1993b). 'A Practical Bayes Approach for Longitudinal Probit Regression Models with Random Effects'. Technical Report, Department of Mathematics and Statistics, Bowling Green State University.

[4] Albert, J. and S. Chib (2001). 'Sequential Ordinal Modeling with Applications to Survival Data'. *Biometrics*, 57: 829-36.

[5] Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.

[6] Andrews, D.F. and C.L. Mallows (1974). 'Scale Mixtures of Normal Distributions'. Journal of the Royal Statistical Society, Series B, 36: 99-102.

[7] Angrist, J.D. (1990). 'Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records'. *American Economic Review*, 80: 313-36.

[8] Angrist, J. and A. Krueger (1991). 'Does Compulsory School Attendance Affect Schooling and Earnings?'. *Quarterly Journal of Economics*, 106: 979-1014.

[9] Basu, S. and S. Mukhopadhyay (2000). 'Bayesian Analysis of Binary Regression Using Symmetric and Asymmetric Links'. *Sankhya*, 62, Series B, Pt. 3: 372-87.

[10] Bernardo, J. and A.F.M. Smith (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.

[11] Berndt, E., B. Hall, R. Hall and J. Hasuman (1974). 'Estimation and Inference in Nonlinear Structural Models'. *Annals of Social Measurement*, 3:653-665.

[12] Bolstad, W.M. and S.O. Manda (2001). 'Investigating Child Mortality in Malawi Using Family and Community Random Effects: A Bayesian Analysis'. *Journal of the American Statistical Association*, 96: 12-19.

[13] Bound, J., D. Jaeger and R. Baker (1995). 'Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Regressors is Weak'. *Journal of the American Statistical Association*, 90: 443-50.

[14] Campolieti, M. (1997). 'Bayesian Estimation of Duration Models: An Application of the Multiperiod Probit Model'. *Empirical Economics*, 22: 461-80.

[15] Campolieti, M. (2000). 'Bayesian Estimation and Smoothing of the Baseline Hazard in Discrete Time Duration Models'. *Review of Economics and Statistics*, 82/4: 685-701.

[16] Campolieti, M. (2001). 'Bayesian Semiparametric Estimation of Discrete Duration Models: An Application of the Dirichlet Process Prior'. *Journal of Applied Econometrics*, 16/1: 1-22.

[17] Campolieti, M. (2003). 'On the Estimation of Hazard Models with Flexible Baseline Hazards and Nonparametric Unobserved Heterogeneity'. *Economics Bulletin*, 3/24: 1-10.

[18] Carlin, B.P. and N. G. Polson (1991). 'Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler'. *The Canadian Journal of Statistics*, 19: 399-405.

[19] Chamberlain, G. (2010). 'Bayesian Aspects of Treatment Choice', in J. Geweke, G. Koop and H. van Dijk, (eds.), *Handbook of Bayesian Econometrics*. Oxford: Oxford University Press.

[20] Chen, M-H. and D. Dey (1998). 'Bayesian Modeling of Correlated Binary Responses via Scale Mixture of Multivariate Normal Link Functions'. *Sankhya*, 60, Series A, Pt. 3: 322-43.

[21] Chen, M-H. and D. Dey (2000). 'Bayesian Analysis for Correlated Ordinal Data Models', in D.K. Dey, S.K. Ghosh and B.K. Mallick, eds., *Generalized Linear Models: A Bayesian Perspective*, Marcel Dekker, 133-57.

[22] Chen, M-H, D. Dey and Q. Shao (1999). 'A New Skewed Link Model for Dichotomous Quantal Response Data'. *Journal of the American Statistical Association*, 94: 1172-86.

[23] Chernozhukov, V. and C. Hansen (2008). 'The Reduced Form: A Simple Approach to Inference with Weak Instruments'. *Economics letters*, 100: 68-71.

[24] Chib, S. (1992), 'Bayes Inference in the Tobit Censored Regression Model'. *Journal of Econometrics*, 51: 79-99.

[25] Chib, S. (1995). 'Marginal Likelihood from the Gibbs Output'. *Journal of the American Statistical Association*, 90: 1313-21.

[26] Chib, S. (1998). 'Estimation and Comparison of Multiple Change-Point Models', *Journal of Econometrics*, 86: 221-41.

[27] Chib, S. (2003). 'On Inferring Effects of Binary Treatments with Unobserved Confounders (with discussion)', in J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, (eds.), *Bayesian Statistics 7*, Oxford: Oxford University Press, 66-84.

[28] Chib, S. (2007), 'Analysis of Treatment Response Data Without the Joint Distribution of Potential Outcomes'. *Journal of Econometrics*, 140: 401-412.

[29] Chib, S. (2010). 'MCMC Methods', in J. Geweke, G. Koop and H. van Dijk, (eds.), *Handbook of Bayesian Econometrics*. Oxford: Oxford University Press.

[30] Chib, S. and B.P. Carlin (1999). 'On MCMC Sampling in Hierarchical Longitudinal Models', *Statistics and Computing*, 9: 17-26.

[31] Chib, S. and E. Greenberg (1998). 'Analysis of Multivariate Probit Models'. *Biometrika*, 85, 347-361.

[32] Chib, S. and E. Greenberg (2007). 'Semiparametric Modeling and Estimation of Instrumental Variable Models'. *Journal of Computational and Graphical Statistics*, 16: 86-114.

[33] Chib, S., E. Greenberg and I. Jeliazkov (2009). 'Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection', *Journal of Computational and Graphical Statistics*, 18: 321-48.

[34] Chib, S., E. Greenberg and R. Winkelmann (1998). 'Posterior Simulation and Bayes Factors in Panel Count Data Models'. *Journal of Econometrics*, 86: 33-54.

[35] Chib, S. and B. Hamilton (2000). 'Bayesian Analysis of Cross Section and Clustered Data Treatment Models'. *Journal of Econometrics*, 97: 25-50.

[36] Chib, S. and B. Hamilton (2002). 'Semiparametric Bayes Analysis of Longitudinal Data Treatment Models'. *Journal of Econometrics*, 110, 67-89.

[37] Chib, S. and Jacobi, L. (2007). 'Modeling and Calculating the Effect of Treatment at Baseline from Panel Outcomes'. *Journal of Econometrics*, 140, 781-801.

[38] Chib, S. and Jacobi, L. (2008a). 'Causal Effects from Panel Data in Randomized Experiments with Partial Compliance', in S. Chib, W. Griffiths, G. Koop and D. Terrell, (eds.), *Advances in Econometrics, Volume 23*, Bingley: Jai Press, 183-215.

[39] Chib, S. and Jacobi, L. (2008b). 'Analysis of Treatment Response Data from Eligibility Designs'. *Journal of Econometrics*, 144: 465-78.

[40] Chib, S. and Jeliazkov, I. (2001). 'Marginal Likelihood from the Metropolis-Hastings Output'. *Journal of the American Statistical Association*, 96: 270-81.

[41] Chib, S. and I. Jeliazkov (2006). 'Inference in Semiparametric Dynamic Models for Binary Longitudinal Data'. *Journal of the American Statistical Association*, 101: 685-700.

[42] Chib, S., F. Nardair and N. Shephard (2002). 'Markov Chain Monte Carlo methods for stochastic volatility models'. *Journal of Econometrics*, 108: 281-316.

[43] Chib, S. and R. Winkelmann (2001). 'Markov Chain Monte Carlo Analysis of Correlated Count Data'. *Journal of Business and Economic Statistics*, 19: 428-435.

[44] Chin Choy, J.H. and L.D. Broemeling (1980). 'Some Bayesian Inferences for a Changing Linear Model'. *Technometrics*, 22/1 : 71-8.

[45] Christ, C.F. (1994). 'The Cowles Commission's Contributions to Econometrics at Chicago, 1939-1955'. *Journal of Economic Literature* 32: 30-59.

[46] Cox, D. R. (1972). 'Regression Models and Life-Tables'. *Journal of the Royal Statistical Society, Series B*, 34/2: 187-220.

[47] Conley, T., C. Hansen, R. McCulluch and P. Rossi (2008). 'A Semi-Parametric Bayesian Approach to the Instrumental Variables Problem'. *Journal of Econometrics*, 144: 276-305.

[48] Cowles, M.K. (1996). 'Accelerating Monte Carlo Markov Chain convergence for Cumulative-link Generalized Linear Models'. *Statistics and Computing*, 6: 101-111.

[49] Dearden, L. J. Ferri and C. Meghir (2002). 'The Effect of School Quality on Educational Attainment and Wages'. *Review of Economics and Statistics*, 84: 1-20.

[50] Deb, P., M. Munkin and P.K. Trivedi (2006a). 'Private Insurance, Selection, and the Health Care Use: A Bayesian Analysis of a Roy-type Model'. *Journal of Business and Economic Statistics*, 24: 403-15.

[51] Deb, P., M. Munkin and P.K. Trivedi (2006b). 'Bayesian Analysis of the Two-Part Model with Endogeneity: Application to Health Care Expenditure'. *Journal of Applied Econometrics*, 21: 1081-99.

[52] Drèze, J.H. (1976). 'Bayesian Limited Information Analysis of the Simultaneous Equations Model'. *Econometrica*, 44: 1045-75.

[53] Drèze, J.H. and J-F. Richard (1983). 'Bayesian Analysis of Simultaneous Equation Systems', in Z. Grilliches and M.D. Intrilligator, (eds.), *Handbook of Econometrics, Vol. 1.*

[54] Edwards, Y.D. and G.M. Allenby (2003). 'Multivariate Analysis of Multiple Response Data'. *Journal of Marketing Research*, 40: 321-334.

[55] Fomby, T.M. and T.J. Vogelsang (2002). 'The Application of Size Robust Trend Analysis to Global Warming Temperature Series'. *Journal of Climate*, 15: 117-123.

[56] Frühwirth-Schnatter, S. and R. Früwirth (2007). 'Auxiliary Mixture Sampling with Applications to Logistic Models,' *Computational Statistics & Data Analysis*, 51: 3509-28.

[57] Gelfand, A.E., S.E. Hills, A. Racine-Poon and A.F.M. Smith (1990). 'Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling'. *Journal of the American Statistical Association*, 412: 972-85.

[58] Gelman, A., G.O. Roberts and W.R. Gilks (1996). 'Efficient Metropolis Jumping Rules', in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 5*, Oxford: Oxford University Press, 599-607.

[59] Geweke, J. (1991). 'Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints', in E.M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, Fairfax: Interface Foundation of North America, Inc., 571-78, working paper version in PDF format at http://www.biz.uiowa.edu/faculty/jgeweke/papers.html.

[60] Geweke, J. (1993). 'Bayesian Treatment of the Independent Student-t Linear Model'. *Journal of Applied Econometrics*, 8: S19-S40.

[61] Geweke, J. (1996a). 'Bayesian Reduced Rank Regression in Econometrics'. *Journal of Econometrics*, 75: 121-46.

[62] Geweke, J. (1996b). 'Bayesian Inference for Linear Models Subject to Linear Inequality Constraints', in W.O Johnson, J.C. Lee and Z. Zellner (eds.), *Modeling and Prediction: Honoring Seymour Geisser*, Springer-Verlag, 248-63.

[63] Geweke, J. (2004). 'Getting it Right: Joint Distribution Tests of Posterior Simulators'. *Journal of the American Statistical Association*, 99: 799-804.

[64] Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, New York: Wiley.

[65] Geweke, J., G. Gowrisankaran and R.J. Town (2003). 'Bayesian Inference for Hospital Quality in a Selection Model'. *Econometrica*, 71: 1215-38.

[66] Geweke, J. and M. Keane (1999). 'Mixture of Normals Probit Models', in C. Hsiao, K. Lahiri, L-F Lee and M.H. Pesaran (eds.), *Analysis of Panels and Limited Dependent Variables: A Volume in Honor of G. S. Maddala*, Cambridge: Cambridge University Press, 49-78, working paper version in PDF format at http://www.biz.uiowa.edu/faculty/jgeweke/papers.html.

[67] Geweke, J. and M. Keane (2000). 'An Empirical Analysis of Income Dynamics Among Men in the PSID: 1968-1989'. *Journal of Econometrics*, 96: 293-356.

[68] Geweke, J. and M. Keane (2001). 'Computationally Intensive Methods for Integration in Econometrics', in J.J. Heckman and E. Leamer, (eds.), *Handbook of Econometrics, Volume 5*, Elsevier.

[69] Geweke, J. and M. Keane (2007). 'Smoothly Mixing Regressions'. *Journal of Econometrics*, 138: 252-91.

[70] Geweke, J., M. Keane and D. Runkle (1994). 'Alternative Computational Approaches to Inference in the Multinomial Probit Model'. *Review of Economics and Statistics*, 76: 609-32.

[71] Geweke, J., M. Keane and D. Runkle (1997). 'Statistical Inference in the Multinomial Multiperiod Probit Model'. *Journal of Econometrics*, 80: 125-66.

[72] Geweke, J. and N. Terui (1993). 'Bayesian Threshold Autoregressive Models for Nonlinear Time Series'. *Journal of Time Series Analysis*, 14: 441-54.

[73] Giordani, P. and R. Kohn (2008). 'Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models'. *Journal of Business and Economic Statistics*, 26: 66-77.

[74] Graves, J., I. Jeliazkov and M. Kutzbach (2008). 'Fitting and Comparison of Models for Multivariate Ordinal Outcomes', in S. Chib, W. Griffith, G. Koop and D. Terrell, (eds.), *Advances in Econometrics, Volume 23*, 115-56.

[75] Griffin, J., F. Quintana and M.F.J. Steel (2010). 'Flexible and Nonparametric Modelling', in J. Geweke, G. Koop and H. van Dijk, (eds.), *Handbook of Bayesian Econometrics*. Oxford: Oxford University Press.

[76] Gu, Y., D.G. Fiebig, E. Cripps and R. Kohn (2009). 'Bayesian Estimation of a Random Effects Heteroscedastic Probit Model'. *Econometrics Journal*, 12: 324-339.

[77] Guo, G., and G. Rodriguez (1992). 'Estimating a Multivariate Proportional Hazards Model for Clustered Data Using the EM Algorithm, with an Application to Child Survival in Guatemala'. *Journal of the American Statistical Association*, 87: 969-76.

[78] Harvey, A.C. (1976). 'Estimating Regression Models with Multiplicative Heteroscedasticity'. *Econometrica*, 44: 461-65.

[79] Heckman, J.J., J.L. Tobias and E. Vytlacil (2001). 'Four Parameters of Interest in the Evaluation of Social Programs'. *Southern Economic Journal*, 68: 210-33.

[80] Heckman, J.J., J.L. Tobias and E. Vytlacil (2003). 'Simple Estimators for Treatment Parameters in a Latent Variable Framework'. *Review of Economics and Statistics*, 85: 748-55.

[81] Holford, T.R. (1976). 'Life Tables with Concomitant Information'. *Biometrics*, 32/3: 587-97.

[82] Holmes, C.C. and L. Held (2006). 'Bayesian Auxiliary Variable Models for Binary and Multinomial Regression'. *Bayesian Analysis*, 1: 146-68.

[83] Hoogerheide, L.F., F. Kleibergen and H.K. van Dijk (2007). 'Natural Conjugate Priors for the Instrumental Variables Regression Model Applied to the Angrist-Krueger Data'. *Journal of Econometrics*, 138: 63-103.

[84] Hoogerheide, L.F., J.F. Kaashoek and H.K. van Dijk (2007). 'On the Shape of Posterior Densities and Credible Sets in Instrumental Variable Regression Models with Reduced Rank: An Application of Flexible Sampling Methods Using Neural Networks'. *Journal of Econometrics*, 139: 154-80.

[85] Imbens, G.W. and J.D. Angrist (1994). 'Identification and Estimation of Local Average Treatment Effects'. *Econometrica*, 62: 467-75.

[86] Ivanov, M.A. and S. N. Evtimov (2009). '1963: The Break Point of the Northern Hemisphere Temperature Trend During the Twentieth Century'. *International Journal of Climatology*, forthcoming

[87] Keane, M. (1992). 'A Note on Identification in the Multinomial Probit Model'. *Journal of Business and Economics Statistics*, 10: 193-200.

[88] Kliebergen, F. and E. Zivot (2003). 'Bayesian and Classical Approaches to Instrumental Variable Regression'. *Journal of Econometrics*, 114: 29-72.

[89] Kline, B. and J.L. Tobias (2008). 'The Wages of BMI: Bayesian Analysis of a Skewed Treatment Response Model with Nonparametric Endogeneity'. *Journal of Applied Econometrics*, 23: 767-93.

[90] Koop, G. (2003). *Bayesian Econometrics*, John Wiley and Sons.

[91] Koop, G., Osiewalski, J. and M. Steel (1997). 'Bayesian Efficiency Analysis through Individual Effects: Hospital Cost Frontiers'. *Journal of Econometrics*, 76: 77-105.

[92] Koop, G. and D.J. Poirier (1997). 'Learning about the Across-Regime Correlation in Switching Regression Models'. *Journal of Econometrics*, 78: 217-227.

[93] Koop, G. and M. Steel (2001). 'Bayesian Analysis of Stochastic Frontier Models', in B. Baltagi, (ed.), *A Companion to Theoretical Econometrics*, Blackwell Publishers, 520-37.

[94] Koop, G., D. Poirier and J. Tobias (2007). *Bayesian Econometric Methods*, Cambridge: Cambridge University Press.

[95] Koop, G. and S. Potter (2007). 'Estimation and Forecasting in Models with Multiple Breaks'. *Review of Economic Studies*, 74: 763-89.

[96] Krueger, A. (1998). 'Reassessing the View That American Schools are Broken'. *FRBNY Economic Policy Review*, March, 29-43.

[97] Krueger, A. and D. Whitmore (2001). 'The Effect of Attending a Small Class in the Early Grades on College Test-Taking and Middle School Test Results: Evidence from Project STAR'. *Economic Journal*, 111: 1-28.

[98] Laird, N.M. (1989). 'Empirical Bayes Ranking Methods'. *Journal of Educational and Behavioral Statistics*, 14: 29-46.

[99] Lakdawalla, D., N. Sood and D. Goldman (2006). 'HIV Breakthroughs and Risky Sexual Behavior'. *Quarterly Journal of Economics*, 121: 1063-1102.

[100] Lancaster, T. (1979). 'Econometric Methods for the Duration of Unemployment'. *Econometrica*, 47/4: 939-56.

[101] Lancaster, T. (2003). 'A Note on Bootstraps and Robustness', unpublished manuscript, Brown University.

[102] Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing.

[103] Leslie, D.S., R. Kohn and D.J. Nott (2007). 'A General Approach to Heteroscedastic Linear Regression'. *Statistics and Computing*, 17: 131-46.

[104] Li, K. (1998). 'Bayesian Inference in a Simultaneous Equations Model with Limited Dependent Variables'. *Journal of Econometrics*, 85: 387-400.

[105] Li, K. (1999). 'Bayesian Analysis of Duration Models: An Application to Chapter 11 Bankruptcy'. *Economics Letters*, 63/3: 305-12.

[106] Li, M. (2006). 'High School Completion and Future Youth Unemployment: New Evidence from High School and Beyond'. *Journal of Applied Econometrics*, 21/1: 23-53.

[107] Li, M. (2007). 'Bayesian Proportional Hazard Analysis of the Timing of High School Dropout Decisions'. *Econometric Reviews*, 26/5: 529-56.

[108] Li, M. and J.L. Tobias (2005). 'Bayesian Modeling of School Effects Using Hierarchical Models with Smoothing Priors'. *Studies in Nonlinear Dynamics and Econometrics*, 9/3, Article 4.

[109] Li, M. , D.J. Poirier and J.L. Tobias (2004). 'Do Dropouts Suffer from Dropping Out? Estimation and Prediction of Outcome Gains in Generalized Selection Models'. *Journal of Applied Econometrics*, 19: 203-25.

[110] Lindley, D.V. and A.F.M. Smith (1972). 'Bayes Estimates for the Linear Model'. *Journal of the Royal Statistical Society, Series B*, 34: 1-41.

[111] McCulloch, R. E., N. G. Polson and P. E. Rossi (2000). 'A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters'. *Journal of Econometrics*, 99: 173-93.

[112] McCulloch, R. and P. Rossi (1994). 'An Exact Likelihood Analysis of the Multinomial Probit Model'. *Journal of Econometrics*, 64: 207-40.

[113] McFadden, D. (1974). 'Conditional Logit Analysis of Qualitative Choice Behavior', in P. Zarembka, (ed.), *Frontiers of Econometrics*, New York: Academic Press, 105-142.

[114] Munkin, M.K. and P.K. Trivedi (2003). 'Bayesian Analysis of Self-Selection Model with Multiple Outcomes Using Simulation-Based Estimation: An Application to the Demand for Healthcare'. *Journal of Econometrics*, 114: 197-220.

[115] Munkin, M. and P.K. Trivedi (2008). 'Bayesian Analysis of the Ordered Probit Model with Endogenous Selection'. *Journal of Econometrics*, 143: 334-48.

[116] Nandram, B. and M-H Chen (1996). 'Accelerating Gibbs Sampler Convergence in the Generalized Linear Models via a Reparameterization'. *Journal of Statistical Computation and Simulation*, 54: 129-44.

[117] Nobile, A. (2000). 'Comment: Bayesian Multinomial Probit Models with a Normalization Constraint'. *Journal of Econometrics*, 99: 335-45.

[118] Percy, D. (1992). 'Prediction for Seemingly Unrelated Regressions'. *Journal of the Royal Statistical Society, Series B*, 54: 243-52.

[119] Poirier, D. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press.

[120] Poirier, D. (2008). 'Bayesian Interpretations of Heteroskedastic Consistent Covariance Estimators Using the Informed Bayesian Bootstrap', working paper, University of California, Irvine.

[121] Poirier, D.J. and J.L. Tobias (2003). 'On the Predictive Distributions of Outcome Gains in the Presence of an Unidentified Parameter'. *Journal of Business and Economic Statistics*, 21: 258-68.

[122] Rossi, P.E. and G. M. Allenby (2010). 'Bayesian Applications in Marketing', in J. Geweke, G. Koop and H. van Dijk (eds.), *Handbook of Bayesian Econometrics*, Oxford: Oxford University Press.

[123] Rossi, P.E., G.M. Allenby and R. McCulloch (2005). *Bayesian Statistics and Marketing*, Wiley.

[124] Sastry, N. (1997). 'A Nested Frailty Model for Survival Data, With an Application to the Study of Child Survival in Northeast Brazil'. *Journal of the American Statistical Association*, 92: 426-35.

[125] Sims, C. (2007). 'Thinking about Instrumental Variables', available online at http://sims.princeton.edu/yftp/IV/.

[126] Sims, C. (2010). 'Understanding Non-Bayesians', in J. Geweke, G. Koop and H. van Dijk (eds.), *Handbook of Bayesian Econometrics*, Oxford: Oxford University Press.

[127] Smith, A.F.M. (1973). 'A General Bayesian Linear Model'. *Journal of the Royal Statistical Society Series B*, 35: 67-75.

[128] Smith, M.D. (2005). 'Using Copulas to Model Switching Regimes with Application to Child Labour'. *The Economic Record*, 81: S47-57.

[129] Stefanski, L.A. (1991). 'A Normal Scale Mixture Representation of the Logistic Distribution'. *Statistics and Probability Letters*, 11/1, 69-70.

[130] Smith, M. and R. Kohn (1996). 'Nonparametric Regression Using Bayesian Variable Selection'. *Journal of Econometrics*, 75: 317-43.

[131] Stockwell, D.R.B. and A. Cox (2009). 'Structural Break Models of Climatic Regime-Shifts: Claims and Forecasts,' working paper available at http://arxiv.org/abs/0907.1650.

[132] Tanizaki, H. and X. Zhang (2001). 'Posterior Analysis of the Multiplicative Heteroscedasticity Model'. *Communications in Statistics, Theory and Methods*, 30: 855-74.

[133] Tanner, M. and W. Wong (1987). 'The Calculation of Posterior Distributions by Data Augmentation'. *Journal of the American Statistical Association*, 82: 528-49.

[134] Train, K.E. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press.

[135] Tüchler, R. (2008). 'Bayesian Variable Selection for Logistic Models using Auxiliary Mixture Sampling'. *Journal of Computational and Graphical Statistics*, 17: 76-94.

[136] van Hasselt, M. (2008). 'Bayesian Inference in a Sample Selection Model', working paper, Department of Economics, University of Western Ontario.

[137] van den Broeck, J., G. Koop, J. Osiewalski and M. Steel (1994). 'Stochastic Frontier Models: A Bayesian Perspective'. *Journal of Econometrics*, 61: 273-303.

[138] Vijverberg, W. (1993). 'Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity'. *Journal of Econometrics*, 57: 69-89.

[139] Villani, M., R. Kohn and P. Giordani (2007). 'Nonparametric Regression and Density Estimation Using Smoothly Varying Normal Mixtures', Sveriges Riksbank Working Paper Series, No. 211.

[140] Vogelsang, T.J. and P.H. Frances (2005). 'Are Winters Getting Warmer?'. *Environmental Modelling & Software*, 20: 1449-1455.

[141] White, H. (1980). 'A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity'. *Econometrica*, 48: 817-38.

[142] Yau, P. and R. Kohn (2003). 'Estimation and Variable Selection in Nonparametric Heteroscedastic Regression'. *Statistics and Computing*, 13: 191-208.

[143] Zellner, A. (1962). 'An Efficient Method of Estimating Seemingly Unrelated Regressions and Test for Aggregation Bias'. *Journal of the American Statistical Association*, 57: 348-68.