CHAPTER 9

BAYESIAN METHODS

9.1 Overview

Over the last two decades there has been an "MCMC revolution" in which Bayesian methods have become a highly popular and effective tool for the applied statistician. This chapter is a brief introduction to Bayesian methods and their applications in measurement error problems. The reader new to Bayesian statistics is referred to the bibliographic notes at the end of this chapter for further reading.

We will not go into the philosophy of the Bayesian approach, whether one should be an objective or a subjective Bayesian, and so forth. We recommend reading Efron (2005), who has a number of amusing comments on the differences between Bayesian and Frequentists, and also on the differences among Bayesians. Our focus here will be how to formulate measurement error models from the Bayesian perspective, and how to compute them. For those familiar with Bayesian software such as WinBUGS, a Bayesian analysis is sometimes relatively straightforward. Bayesian methods also allow one to use other sources of information, e.g., from similar studies, to help estimate parameters that are poorly identified by the data alone. A disadvantage of Bayesian methods, which is shared by maximum likelihood, is that, compared to regression calibration, computation of Bayes estimators is intensive. Another disadvantage shared by maximum likelihood is that one must specify a full likelihood and therefore one should investigate whether the estimator is robust to possible model misspecification.

9.1.1 Problem Formulation

Luckily, Bayesian methods start from a likelihood function, a topic we have already addressed in Chapter 8, and illustrated with a four-step approach in Figure 8.1.

In the Bayesian approach, there are five essential steps, see Figure 9.1.

- **Step 1**: This is the same as the first step in a likelihood approach. Specifically, one must specify a parametric model for every component of the data. Any likelihood analysis begins with the model one would use if $\mathbf{X}$ were observable.

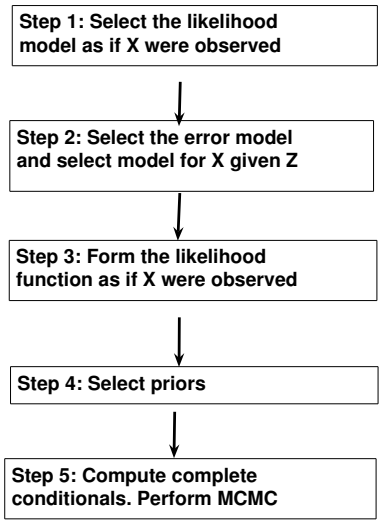

Figure 9.1 *Five basic steps in performing a Bayesian analysis of a measurement error problem. If automatic software such as WinBUGS is used, the complete conditionals, which often require detailed algebra, need not be computed.*

- **Step 2**: This step too agrees with the likelihood approach. The next crucial decision is the error model that is to be chosen. This could be a classical error model, a Berkson model, or a combination of the two. If one has classical components in the measurement error model, then typically one also needs to specify a distribution for the unobserved $\mathbf{X}$ given the observable covariates $\mathbf{Z}$.
- **Step 3**: The typical Bayesian approach treats $\mathbf{X}$ as missing data, and, in effect, imputes it multiple times by drawing from the conditional distribution of $\mathbf{X}$ given all other variables. Thus, at this step, the likelihood of all the data, including $\mathbf{W}$, is formed as if $\mathbf{X}$ were available.
- **Step 4**: In the Bayesian approach, parameters are treated as if they were random, one of the essential differences with likelihood methods. If one is going to treat parameters as random, then they need to be given distributions, called *prior distributions*. Much of the controversy among statisticians regarding Bayesian methods revolves around these prior distributions.
- **Step 5**: The final step is to compute Bayesian quantities, in particular the *posterior distribution* of parameters given all the observed data.

There are various approaches to doing this, most of them revolving around Markov Chain Monte Carlo (MCMC) methods, often based on the Gibbs Sampler. In some problems, such as with WinBUGS, users do not actually have to do anything but run a program, and the appropriate posterior quantities become available. In other cases though, either the standard program is not suitable to the problem, or the program does not work well, in which case one has to tailor the approach carefully. This usually involves detailed algebraic calculation of what are called the *complete conditionals*, the distribution of the parameters, and the **X** values, given everything else in the model. We give a detailed example of this process in Section 9.4.

9.1.2 Posterior Inference

Bayesian inference is based upon the posterior density, which is the conditional density of unobserved quantities (the parameters and unobserved covariates) given the observed data and summarizes all of the information about the unobservables. For example, the mean, median, or mode of the posterior density are all suitable point estimators. A region with probability $(1-\alpha)$ under the posterior is called a "credible set," and is a Bayesian analog to a confidence region. To calculate the posterior, one can take the joint density of the data and parameters and, at least in principle, integrate out the parameters to get the marginal density of the data. One can then divide the joint density by this marginal density to get the posterior density.

There are many "textbook examples" where the posterior can be computed analytically, but in practical applications this is often a non–trivial problem requiring high-dimensional numerical integration. The computational problem has been the subject of much recent research. The method currently receiving the most attention in the literature is the Gibbs sampler and related methods such as the Metropolis-Hastings algorithm (Hastings, 1970; Geman & Geman, 1984; Gelfand & Smith, 1990).

The Gibbs sampler, which is often called Markov Chain Monte Carlo (MCMC), generates a Markov chain whose stationary distribution is the posterior distribution. The key feature of the Gibbs sampler is that this chain can be simulated using only the joint density of the parameters, the unobserved **X**-values and the observed data, e.g., the product of the likelihood and the prior, and not the unknown posterior density which would require an often intractable integral. If the chain is run long enough, then the observations in a sample from the chain are approximately identically distributed with common distribution equal to

the posterior. Thus posterior moments, the posterior density, and other posterior quantities can be estimated from a sample from the chain.

The Gibbs sampler "fills-in" or imputes the values of the unobserved covariates $\mathbf{X}$ by sampling from their conditional distribution given the observed data and the other parameters. This type of imputation differs from the imputation of regression calibration in two important ways. First, the Gibbs sampler makes a large number of imputations from the conditional distribution of $\mathbf{X}$ whereas regression calibration uses a single imputation, namely the conditional expectation of $\mathbf{X}$ given $\mathbf{W}$ and $\mathbf{Z}$. Second, the Gibbs sampler conditions on $\mathbf{Y}$ as well as $\mathbf{W}$ and $\mathbf{Z}$ when imputing values of $\mathbf{X}$, but regression calibration does not use information about $\mathbf{Y}$ when imputing $\mathbf{X}$.

9.1.3 Bayesian Functional and Structural Models

We made the point in Section 2.1 that our view of functional and structural modeling is that in the former, we make no or at most few assumptions about the distribution of the unobserved $\mathbf{X}$-values. Chapters 5 and 7 describe methods that are explicitly functional, while regression calibration is approximately functional.

In contrast, likelihood methods (Chapter 8) and Bayesian methods necessarily must specify a distribution for $\mathbf{X}$ in one way or another, and here the distinction between functional and structural is blurred. Effectively, structural Bayesian/likelihood modeling imposes a simple model on $\mathbf{X}$, such as the normal model, while functional methods specify flexible distributions for $\mathbf{X}$. We use structural models in this chapter. Examples of this approach are given by Schmid & Rosner (1993), Richardson & Gilks (1993) and Stephens & Dellaportas (1992).

There are at least several ways to formulate a Bayesian functional model. One way would allow the distribution of $\mathbf{X}$ to depend on the observation number, i. Müller & Roeder (1997) use this idea for the case when $\mathbf{X}$ is partially observed. They assume that the $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ are jointly normally distributed with mean μ_i and covariance matrix Σ_i, where $\theta_i = (\mu_i, \Sigma_i)$ is modeled by a Dirichlet process distribution which itself has unknown hyperparameters. Lindley & El Sayyad (1968) is the first Bayesian paper on functional models, covering the linear regression case. Because of their complexity, we do not consider Bayesian functional models here.

A second possibility intermediate between functional and hard-core structural approaches is to specify flexible distributions, much as we suggested in Section 8.2.2. Carroll, Roeder, & Wasserman (1999) and Richardson, Leblond, Jaussent, & Green (2002) use mixtures of normal

distributions. Gustafson, Le, & Vallee (2002) use an approach based on approximating the distribution of $\mathbf{X}$ by a discrete distribution.

In this chapter, the $\mathbf{Z}_i$'s are treated as fixed constants as we have done before in non-Bayesian treatments. This makes perfect sense, since Bayesians only need to treat unknown quantities as random variables. Thus, the likelihood is the conditional density of the $\mathbf{Y}_i$'s, $\mathbf{W}_i$'s, and any $\mathbf{X}_i$'s that are observed, given the parameters and the $\mathbf{Z}_i$'s. The posterior is the conditional density of the parameters given all data, i.e., the $\mathbf{Z}_i$'s, $\mathbf{Y}_i$'s, $\mathbf{W}_i$'s, and any observed $\mathbf{X}_i$'s.

9.1.4 Modularity of Bayesian MCMC

The beauty of the Bayesian paradigm combined with modern MCMC computing is its tremendous flexibility. The technology is "modular" in that the methods of handling, e.g., multiplicative error, segmented regression and the logistic regression risk model can be combined easily. In effect, if one knows how to handle these problems separately, it is often rather easy to combine them into a single analysis and program.

9.2 The Gibbs Sampler

As in Chapter 8, especially equation (8.7), the first three steps of our Bayesian paradigm result in the likelihood computed as if $\mathbf{X}$ were observable. Dropping the second measure $\mathbf{T}$, this likelihood for an individual observation becomes

$$\begin{aligned} f(\mathbf{Y},\mathbf{W},\mathbf{X}|\mathbf{Z},\Omega) &= f_{Y|Z,X}(\mathbf{Y}|\mathbf{Z},\mathbf{X},\mathcal{B}) \\ &\quad \times f_{W|Z,X}(\mathbf{W}|\mathbf{Z},\mathbf{X},\widetilde{\alpha}_1) f_{X|Z}(\mathbf{X}|\mathbf{Z},\widetilde{\alpha}_2), \end{aligned}$$

where Ω is the collection of all unknown parameters. As in the fourth step of the Bayesian paradigm, we let Ω have a prior distribution $\pi(\Omega)$. The likelihood of all the "data" then becomes

$$\pi(\Omega)\prod_{i=1}^{n} f(\mathbf{Y}_i,\mathbf{W}_i,\mathbf{X}_i|\mathbf{Z}_i,\Omega).$$

To keep this section simple, we have not included the possibility of validation data here, but that could be done with only some additional effort, mostly notational. To keep notation compact, we will write the ensemble of $\mathbf{Y}$, $\mathbf{X}$, etc. as $\widetilde{\mathbf{Y}}$, $\widetilde{\mathbf{X}}$, etc. This means that the likelihood can be expressed as

$$\pi(\Omega) f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{X}}|\widetilde{\mathbf{Z}},\Omega).$$

The posterior distribution of Ω is then

$$f(\Omega\Big|\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{Z}}) = \frac{\pi(\Omega)\int f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{x}|\widetilde{\mathbf{Z}},\Omega)d\widetilde{x}}{\int \pi(\omega)f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{x}|\widetilde{\mathbf{Z}},\omega)d\widetilde{x}d\omega}. \tag{9.1}$$

The practical problem is that, even if the integration in $\widetilde{x}$ can be accomplished or approximated as in Chapter 8, the denominator of (9.1) may be very difficult to compute. Numerical integration typically fails to provide an adequate approximation even when there are as few as three or four components to Ω.

The Gibbs sampler is one solution to the dilemma. The Gibbs sampler is an iterative, Monte-Carlo method consisting of the following main steps, starting with initial values of Ω.

- Generate a sample of the unobserved $\mathbf{X}$-values by sampling from their posterior distributions given the current value of Ω, the posterior distribution of $\mathbf{X}_i$ being

$$f(\mathbf{X}_i|\mathbf{Y}_i,\mathbf{W}_i,\mathbf{Z}_i,\Omega) = \frac{f(\mathbf{Y}_i,\mathbf{W}_i,\mathbf{X}_i|\mathbf{Z}_i,\Omega)}{\int f(\mathbf{Y}_i,\mathbf{W}_i,x|\mathbf{Z}_i,\Omega)dx}. \tag{9.2}$$

  As we indicate below, this can be done without having to evaluate the integral in (9.2).

- Generate a new value of Ω from its posterior distribution given the observed data and the current generated $\mathbf{X}$-values, namely

$$f(\Omega\Big|\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{Z}},\widetilde{\mathbf{X}}) = \frac{\pi(\Omega)f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{X}}|\widetilde{\mathbf{Z}},\Omega)}{\int \pi(\omega)f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{X}}|\widetilde{\mathbf{Z}},\omega)d\omega}. \tag{9.3}$$

  Often, this is done one element of Ω at a time, holding the others fixed (as described below, here too we do not need to compute the integral). Thus, for example, if the j^{th} value of Ω is ω_j, and the other components of Ω are $\Omega_{(-j)}$, then the posterior in question is simply

$$\begin{aligned} &f(\omega_j|\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{Z}},\widetilde{\mathbf{X}},\Omega_{(-j)}) \\ &\qquad = \frac{\pi(\omega_j,\Omega_{(-j)})f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{X}}|\widetilde{\mathbf{Z}},\omega_j,\Omega_{(-j)})}{\int \pi(\omega_j^*,\Omega_{(-j)})f(\widetilde{\mathbf{Y}},\widetilde{\mathbf{W}},\widetilde{\mathbf{X}}|\widetilde{\mathbf{Z}},\omega_j^*,\Omega_{(-j)})d\omega_j^*}. \end{aligned} \tag{9.4}$$

- Repeat this many times. Discard the first few of the generated samples, the so-called burn-in period.

- Quantities such as the posterior mean and posterior quantiles are estimated by the sample mean and quantiles of $\Omega_1,\Omega_2,\ldots$, while kernel density estimates are used to approximate the entire posterior density or the marginal posterior density of a single parameter or subset of parameters.

An important point is that the first two steps do *not* require that one evaluates the integral in the denominator on the right hand sides of (9.2), (9.3) and (9.4).

Generating pseudo random observations from (9.4) is the heart of the Gibbs sampler. Often the prior on ω_j is conditionality conjugate so that the full conditional for ω_j is in the same parametric family as the prior, e.g., both are normal or both are inverse-gamma; see Section A.3 for a discussion of the inverse-gamma distribution. In such cases, the denominator of (9.4) can be determined from the form of the posterior and the integral need not be explicitly calculated.

If we do not have conditional conjugacy, then drawing from the full conditional of ω_j is more difficult. In this situation, we will use a Metropolis-Hastings step which will be described soon. The Metropolis-Hastings algorithm does not require that the integral in (9.4) be evaluated.

9.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm (MH algorithm) is a very versatile and flexible tool, and even includes the Gibbs sampler as a special case. Suppose we want to sample from a certain density, which in applications to Bayesian statistics is the posterior, and that the density is $Cf(\cdot)$, where f is known but the normalizing constant $C > 0$ is difficult to evaluate, see for example (9.3). The MH algorithm uses f without knowledge of C to generate a Markov chain whose stationary distribution is $Cf(\cdot)$.

To simplify the notation, we will subsume the unobserved $\mathbf{X}$ into Ω; this involves no loss of generality since a Bayesian treats all unknown quantities in the same way. Suppose that the current value of Ω is Ω_{curr}. The idea is to generate (see below) a "candidate" value Ω_{cand} and either accept it as the new value or reject it and stay with the current value. Over repeated application, this process results in random variables with the desired distribution.

Mechanically, one has to have a candidate distribution, which may depend upon the current value. We write this candidate density as $q(\Omega_{\text{cand}}|\Omega_{\text{curr}})$. Gelman, Stern, Carlin, & Rubin (2004) call $q(\cdot|\cdot)$ a "jumping rule" since it may generate the jump from Ω_{curr} to Ω_{cand}. Thus, a candidate Ω_{cand} is generated from $q(\cdot|\Omega_{\text{curr}})$. This candidate is accepted and becomes Ω_{curr} with probability

$$r = \min\left\{1, \frac{f(\Omega_{\text{cand}})q(\Omega_{\text{curr}}|\Omega_{\text{cand}})}{f(\Omega_{\text{curr}})q(\Omega_{\text{cand}}|\Omega_{\text{curr}})}\right\}. \tag{9.5}$$

More precisely, a uniform(0,1) random variable V is drawn, and then we set $\Omega_{\text{curr}} = \Omega_{\text{cand}}$ if $V \leq r$.

The popular "random-walk" MH algorithm uses $q(\Omega_{\text{cand}}|\Omega_{\text{curr}}) = h($

$\Omega_{\text{cand}} - \Omega_{\text{curr}}$) for some probability density h. Often, as in our examples, $h(\cdot)$ is symmetric so that

$$r = \min\left\{1, \frac{f(\Omega_{\text{cand}})}{f(\Omega_{\text{curr}})}\right\}. \tag{9.6}$$

The "Metropolis-Hastings within Gibbs algorithm" uses the MH algorithm at those steps in a Gibbs sampler where the full conditional is difficult to sample. Suppose sampling ω_j is one such step. If we generate the candidate $\omega_{j,\text{cand}}$ from $h(\cdot - \omega_{j,\text{curr}})$ where h is symmetric and $\omega_{j,\text{curr}}$ is the current value of ω_j, then r in (9.6) is

$$r = \min\left\{1, \frac{f(\omega_{j,\text{cand}}|\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}, \omega_{\ell,\text{curr}} \text{ for } \ell \neq j)}{f(\omega_{j,\text{curr}}|\widetilde{\mathbf{Y}}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{Z}}, \omega_{\ell,\text{curr}} \text{ for } \ell \neq j)}\right\}.$$

Often, h is a normal density, a heavy-tailed normal mixture, or a t-density. The scale parameter of this density should be chosen so that typical values of $\omega_{j,\text{cand}}$ are neither too close to nor too far from $\omega_{j,\text{curr}}$. If $\omega_{j,\text{cand}}$ is too close to $\omega_{j,\text{curr}}$ with high probability, then the MH algorithm takes mostly very small steps and does not move quickly enough. If $\omega_{j,\text{cand}}$ is generally too far from $\omega_{j,\text{curr}}$, then the probability of acceptance is small. To get good performance of the Metropolis within Gibbs algorithm, we might use a Normal$(0, \sigma^2)$ proposal density where σ^2 is tuned to the algorithm so that the acceptance probability is between 25% and 50%. Gelman, Carlin, Stern, & Rubin (2004, page 306) state that the optimal jumping rule has 44% acceptance in one dimension and about 23% acceptance probability in high dimensions when the jumping and target densities have the same shape. To allow for occasional large jumps, one might instead use a heavy-tailed normal mixture of 90% Normal$(0, \sigma^2)$ and 10% Normal$(0, L\sigma^2)$, where L might be 2, 3, 5, or even 10. This density is very easy to sample from, since we need only generate independent $Z \sim$ Normal$(0, 1)$ and $U \sim [0, 1]$. Then we multiply Z by σ or $\sqrt{L}\,\sigma$ according as $U \leq 0.9$ or $U > 0.9$. The Normal$(0, L\sigma^2)$ component gives the mixture heavy tails and allows the sampler to take large steps occasionally. One can experiment with the value of L to see which gives the best mixing, that is, the least autocorrelation in the sample.

More information on the Gibbs sampler and the MH algorithm can be found in Roberts, Gelman, & Gilks (1997), Chib & Greenberg (1995), Gelman et al. (2004), and in many other books and papers. See Roberts & Rosenthal (2001) for more discussion about scaling of MH jumping rules.

9.4 Linear Regression

In this section, an example is presented where the full conditionals are all conjugate. For those new to Bayesian computations, we will show in some detail how the full conditionals can be found. In the following sections, this example will be modified to models where some, but not all, full conditionals are conjugate.

Suppose we have a linear regression with a scalar covariate $\mathbf{X}$ measured with error and a vector $\mathbf{Z}$ of covariates known exactly. Then the first three steps in Figure 9.1 are as follows. The so–called "outcome model" for the outcome $\mathbf{Y}$ given all of the covariates (observed or not) is

$$\mathbf{Y}_i = \text{Normal}(\mathbf{Z}_i^t \beta_z + \mathbf{X}_i \beta_x, \sigma_\epsilon^2). \tag{9.7}$$

Suppose that we have replicates of the surrogate $\mathbf{W}$ for $\mathbf{X}$. Then the so–called "measurement model" is

$$\mathbf{W}_{i,j} = \text{Normal}(\mathbf{X}_i, \sigma_u^2),\ j = 1, \ldots, k_i. \tag{9.8}$$

Finally, suppose that the "exposure model" for the covariate measured with error, $\mathbf{X}$, given $\mathbf{Z}$ is

$$\mathbf{X}_i = \text{Normal}(\alpha_0 + \mathbf{Z}_i^t \alpha_z, \sigma_x^2). \tag{9.9}$$

The term "exposure model" comes from epidemiology where $\mathbf{X}$ is often exposure to a toxicant.

For this model it is possible to have conjugate priors for all of the full conditionals. The prior we will use is that independently

$$\beta_x = \text{Normal}(0, \sigma_\beta^2),\ \beta_z = \text{Normal}(0, \sigma_\beta^2 \boldsymbol{I})$$

$$\alpha_0 = \text{Normal}(0, \sigma_\alpha^2),\ \alpha_z = \text{Normal}(0, \sigma_\alpha^2 \boldsymbol{I}),$$

$$\sigma_\epsilon^2 = \text{IG}(\delta_{\epsilon,1}, \delta_{\epsilon,2}),\ \sigma_u^2 = \text{IG}(\delta_{u,1}, \delta_{u,2}),\ \sigma_x^2 = \text{IG}(\delta_{x,1}, \delta_{x,2}).$$

As discussed in Section A.3, this prior is conjugate for the full conditionals. Here $\text{IG}(\cdot, \cdot)$ is the inverse gamma density, and the hyperparameters σ_β and σ_μ are chosen to be "large" and the δ hyperparameters to be "small" so that the priors are relatively non-informative. In particular, because σ_β and σ_μ are large, using a mean of zero for the normal priors should not have much influence on the posterior. See Section A.3 for the definition of the inverse gamma distribution and discussion about choosing the hyperparameters of an inverse gamma prior. The unknowns in this model are $(\beta_x, \beta_z, \sigma_\epsilon, \sigma_u)$, $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, and $(\alpha_0, \alpha_z, \sigma_x)$.

Define

$$\boldsymbol{C}_i = \begin{pmatrix} \mathbf{Z}_i \\ \mathbf{X}_i \end{pmatrix},\ \mathbf{Y} = (\mathbf{Y}_1, ..., \mathbf{Y}_n)^t,\ \text{ and } \beta = \begin{pmatrix} \beta_z \\ \beta_x \end{pmatrix}.$$

The likelihood for a single observation is

$$f(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i|\mathbf{Z}_i, \Omega) = (2\pi)^{-3/2}\frac{1}{\sigma_x\sigma_\epsilon\sigma_u^{k_i}}$$
$$\times \exp\{-\left(\mathbf{Y}_i - \boldsymbol{C}_i^t\beta\right)^2/(2\sigma_\epsilon^2)\} \tag{9.10}$$
$$\times \exp\left\{-\textstyle\sum_{j=1}^{k_i}(\mathbf{W}_{i,j} - \mathbf{X}_i)^2/(2\sigma_u^2) - (\mathbf{X}_i - \alpha_0 - \mathbf{Z}_i^t\alpha_z)^2/(2\sigma_x^2)\right\}.$$

The joint likelihood is of course the product over index i of the terms (9.10). The joint density of all observed data and all unknown quantities (parameters and true $\mathbf{X}$'s for non-validation data) is the product of the joint likelihood and the joint prior.

In our calculations, we will use the following:

Rule: If for some p-dimensional parameter θ we have

$$f(\theta|\text{others}) \propto \exp\left\{-\left(\theta^t\boldsymbol{A}\theta - 2\boldsymbol{b}\theta\right)/2\right\}$$

where the constant of proportionality is independent of θ, then $f(\theta|\text{others})$ is Normal$(\boldsymbol{A}^{-1}\boldsymbol{b}, \boldsymbol{A}^{-1})$.

To find the full conditional for β, we isolate the terms depending on β in this joint density. We write the full conditional of β given the others as $f(\beta|\text{others})$. This gives us

$$f(\beta|\text{others}) \propto \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{C}_i^t\beta)^2 - \frac{1}{2\sigma_\beta^2}\beta^t\beta\right\}, \tag{9.11}$$

where the first term in the exponent comes from the likelihood and the second comes from the prior. Let $\mathcal{C}$ have i^{th} row $\boldsymbol{C}_i^t$ and let $\Delta = \sigma_\epsilon^2/\sigma_\beta^2$. Then (9.11) can be rearranged to

$$f(\beta|\text{others}) \propto \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{\beta^t\left(\mathcal{C}^t\mathcal{C} + \Delta\boldsymbol{I}\right)\beta + 2\mathcal{C}^t\mathbf{Y}\beta\right\}\right], \tag{9.12}$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^t$. Using the Rule,

$$f(\beta|\text{others}) = N\left(\left\{\mathcal{C}^t\mathcal{C} + \Delta\boldsymbol{I}\right\}^{-1}\mathcal{C}^t\mathbf{Y},\, \sigma_\epsilon^2\left(\mathcal{C}^t\mathcal{C} + \Delta\boldsymbol{I}\right)^{-1}\right) \tag{9.13}$$

Here we see how the Gibbs sampler can avoid the need to calculate integrals. The normalizing constant in (9.12) can be found from (9.13) simply by knowing the form of the normal distribution.

Result (9.13) is exactly what we would get without measurement error, except that for the non-validation data the $\mathbf{X}$'s in $\mathcal{C}$ are "filled-in" rather than known. Therefore, $\mathcal{C}$ will vary on each iteration of the Gibbs sampler. The parameters Δ and σ_ϵ will also vary even if there is no measurement error.

The full conditional for $\alpha = (\alpha_0, \alpha_z^t)^t$ can be found in the same way as for β. First, analogous to (9.11),

$$f(\alpha|\text{others}) \propto \exp\left\{-\frac{\sum_{i=1}^n\{\mathbf{X}_i - (\alpha_0 + \mathbf{Z}_i^t\alpha_z)\}^2}{2\sigma_x^2} - \frac{\alpha^t\alpha}{2\sigma_\alpha^2}\right\}.$$

Let $D_i = (1\ \mathbf{Z}_i^t)^t$ and let $\mathcal{D}$ be the matrix with i^{th} row equal to D_i^t. Also, let $\eta = \sigma_x^2/\sigma_\alpha^2$. Then, analogous to (9.13),

$$f(\alpha|\text{others}) = N\left\{\left(\mathcal{D}^t\mathcal{D} + \eta\boldsymbol{I}\right)^{-1}\mathcal{D}^t\mathbf{X},\ \sigma_x^2\left(\mathcal{D}^t\mathcal{D} + \eta\boldsymbol{I}\right)^{-1}\right\}, \quad (9.14)$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^t$.

To find the full conditional for $\mathbf{X}_i$, define $\overline{\mathbf{W}}_i = \sum_{J=1}^{k_i} W_{i,j}/k_i$. Then

$$\begin{aligned} f(\mathbf{X}_i|\text{others}) &\propto \exp\left[-(\mathbf{Y}_i - \mathbf{X}_i\beta_x - \mathbf{Z}_i^t\beta_z)^2/(2\sigma_\epsilon^2)\right] \quad (9.15)\\ &\times\exp\left\{-(\mathbf{X}_i - \alpha_0 - \mathbf{Z}_i^t\alpha_z)^2/(2\sigma_x^2) - k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right\}. \end{aligned}$$

After some algebra and applying the Rule again, $f(\mathbf{X}_i|\text{others})$ is seen to be normal with mean

$$\frac{(\mathbf{Y}_i - \mathbf{Z}_i^t\beta_z)(\beta_x/\sigma_\epsilon^2) + (\alpha_0 + \mathbf{Z}_i^t\alpha_z)/\sigma_x^2 + \overline{\mathbf{W}}_i/\sigma_{\overline{\mathbf{W}}}^2}{(\beta_x^2/\sigma_\epsilon^2) + (1/\sigma_x^2) + 1/\sigma_{\overline{\mathbf{W}}}^2}$$

and variance

$$\left\{(\beta_x^2/\sigma_\epsilon^2) + (1/\sigma_x^2) + (1/\sigma_{\overline{\mathbf{W}}}^2)\right\}^{-1}.$$

Notice that the mean of this full conditional distribution for $\mathbf{X}_i$ given everything else depends on $\mathbf{Y}_i$, so that, unlike in regression calibration, $\mathbf{Y}_i$ is used for imputation of $\mathbf{X}_i$.

Now we will find the full conditional for σ_ϵ^2. Recall that the prior is $\text{IG}(\delta_{\epsilon,1}, \delta_{\epsilon,2})$, where from Appendix A.3 we know that the $IG(\alpha, \beta)$ distribution has mean $\beta/(\alpha - 1)$ if $\alpha > 1$ and density proportional to $x^{-(\alpha+1)}\exp(-\beta/x)$. Isolating the terms depending on σ_ϵ^2 in the joint density of the observed data and the unknowns, we have

$$\begin{aligned} &f(\sigma_\epsilon^2|\text{others})\\ &\quad\propto (\sigma_\epsilon^2)^{-(\delta_{\epsilon,1}+n/2+1)}\exp\left\{\frac{-\delta_{\epsilon,2} + -\frac{1}{2}\sum_{i=1}^n(\mathbf{Y}_i - \mathbf{X}_i\beta_x - \mathbf{Z}_i^t\beta_z)^2}{\sigma_\epsilon^2}\right\} \end{aligned}$$

which implies that

$$f(\sigma_\epsilon^2|\text{others}) = \text{IG}\left[(\delta_{\epsilon,1} + n/2),\ \left\{\delta_{\epsilon,2} + (1/2)\sum_{i=1}^n(\mathbf{Y}_i - \mathbf{X}_i\beta_x - \mathbf{Z}_i^t\beta_z)^2\right\}\right].$$

By similar calculations,

$$f(\sigma_x^2|\text{others}) \propto (\sigma_x^2)^{-(\delta_{x,1}+n/2+1)}\exp\left\{\frac{-\delta_{x,2} - \frac{1}{2}\sum_{i=1}^n(\mathbf{X}_i - \mu_x)^2}{\sigma_x^2}\right\},$$

so that

$$f(\sigma_x^2|\text{others}) = \text{IG}\left[(\delta_{x,1} + (n/2)),\ \left\{\delta_{x,2} + (1/2)\sum_{i=1}^{n}(\mathbf{X}_i - \mu_x)^2\right\}\right].$$

Let $M_J = \sum_{i=1}^{n} k_i/2$. Then we have in addition that

$$f(\sigma_u^2|\text{others}) \propto (\sigma_u^2)^{-(\delta_{u,1}+M_J+1)} \exp\left\{\frac{-\delta_{u,2} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(\mathbf{W}_{i,j} - \mathbf{X}_i)^2}{\sigma_u^2}\right\},$$

whence

$$f(\sigma_u^2|\text{others}) = \text{IG}\left[(\delta_{u,1} + M_J),\ \left\{\delta_{u,2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k_i}(\mathbf{W}_{i,j} - \mathbf{X}_i)^2\right\}\right].$$

The Gibbs sampler requires a starting value for Ω. For β_x, β_z, and σ_ϵ, one can use estimates from the regression of $\mathbf{Y}_i$ on $\mathbf{Z}_i$ and $\mathbf{X}_i$ (validation data) or $\overline{\mathbf{W}}$ (non-validation data). Although there will be some bias, these naive estimators should be in a region of reasonably high posterior probability and bias should not be a problem since they are being used only as starting values. We start $\mathbf{X}_i$ at $\overline{\mathbf{W}}_i$. Also, μ_x and σ_x can be started at the sample mean and standard deviation of the starting values of the $\mathbf{X}_i$'s. The replication data can be used to find an analysis of variance estimate of σ_u^2 for use as a starting value, see equation (4.3).

9.4.1 Example

We simulated data with the following parameters: $n = 200$, $\beta^t = (\beta_0, \beta_x, \beta_z) = (1, 0.5, 0.3)$, $\alpha^t = (\alpha_0, \alpha_z) = (1, 0.2)$, $\mathbf{X}_i = \alpha_0 + \alpha_z\mathbf{Z}_i + \mathbf{V}_i$, where $\mathbf{V}_i \sim \text{Normal}(0, \sigma_x^2)$ with $\sigma_x = 1$. The $\mathbf{Z}_i$ were independent Normal$(1, 1)$, and since the analysis is conditioned on their values, their mean and variance are not treated as parameters. Also,

$$\mathbf{Y}_i = \beta_0 + \beta_x\mathbf{X}_i + \beta_z\mathbf{Z}_i + \epsilon_i, \tag{9.16}$$

where $\epsilon_i = \text{Normal}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.3$, and $\mathbf{W}_{i,j} = \text{Normal}(\mathbf{X}_i, \sigma_u^2)$, with $\sigma_u^2 = 1$. The observed data are $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_{i,1}, \mathbf{W}_{i,2})$.

We used Gibbs sampling with 10,000 iterations after a burn-in period of 2000 iterations. The prior parameters were $\sigma_\beta = \sigma_\alpha = 1000$, $\delta_{\epsilon,1} = 3, \delta_{\epsilon,2} = 1$, $\delta_{x,1} = 3, \delta_{x,2} = 1$, and $\delta_{u,1} = 3, \delta_{u,2} = 1$. As discussed in Section A.3, the choice of $\delta_{\epsilon,1} = 3$ and $\delta_{\epsilon,2} = 1$ suggests a prior guess at σ_ϵ^2 of $\delta_{\epsilon,2}/\delta_{\epsilon,1} = 1/3$ and that the prior has the amount of information that would be obtained from $2\delta_{\epsilon,1} = 6$ observations. The same is true of the other δ's. We experimented with other choices of these prior parameters, in particular, smaller values of the effective prior

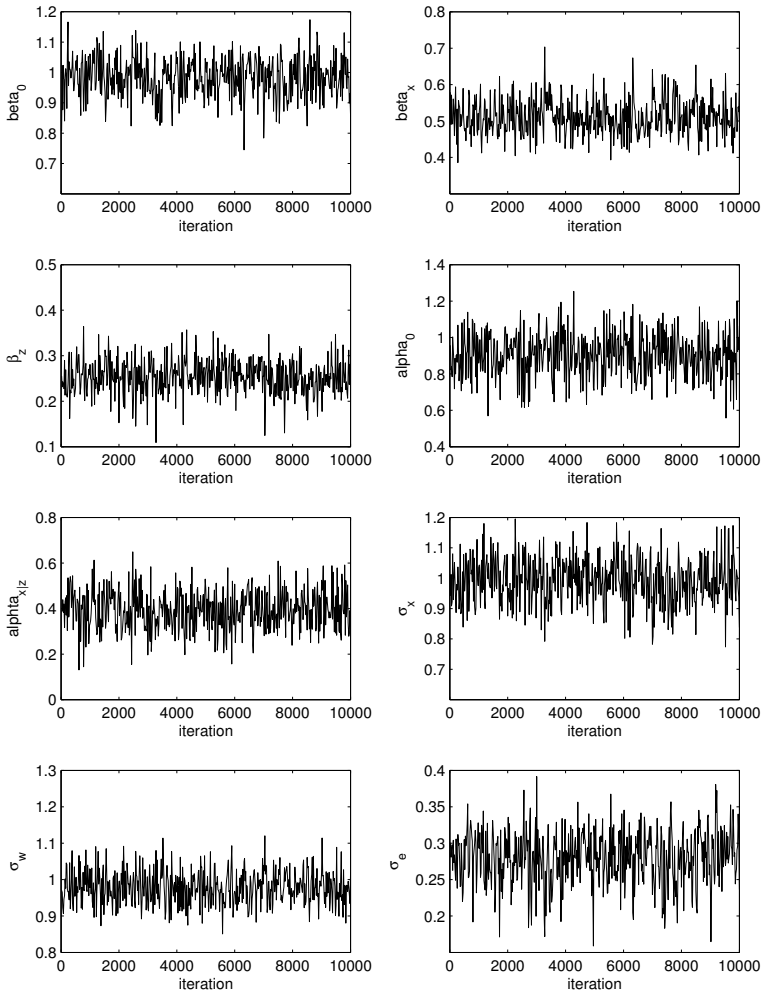

Figure 9.2 *Every 20th iteration of the Gibbs sampler for the linear regression example.*

sample size, and found that the posterior was relatively insensitive to the priors provided that $\delta_{\epsilon,2}$ is not too large.

Starting values for the unobserved covariates were $\mathbf{X}_i = \overline{\mathbf{W}}_i = (\mathbf{W}_{i,1} + \mathbf{W}_{i,2})/2$. The starting values of the parameters were chosen independently: $\sigma_x, \sigma_u, \sigma_\epsilon \sim \text{Uniform}(0.05, 3)$. The starting value for β and α were generated from (9.13) and (9.14).

Figure 9.2 shows every 20th iteration of the Gibbs sampler. These

are the so-called trace plots that are used to monitor convergence of the Gibbs sampler, i.e., at convergence, they should have no discernible pattern. No patterns are observed, and thus the sampler appears to have mixed well. This subset of the iterations was used to make the plots clearer; for estimation of posterior means and variance, all iterates were used. Using all iterates, the sample autocorrelation for β_x looks like an AR(1) process with a first-order autocorrelation of about 0.7. We used a large number (10,000) of iterations, to reduce the potentially high Monte Carlo variability due to autocorrelation.

To study the amount of Monte Carlo error from Gibbs sampling and to see if 10,000 iterations is adequate, the Gibbs sampler was repeated four more times on the same simulated data set but with new random starting values for σ_x, σ_u, and σ_ϵ. The averages of the five posterior means and standard deviations for β_x were 0.4836 and 0.0407. The standard deviation of the five posterior means, which estimates Monte Carlo error, was only 0.00093. Thus, the Monte Carlo error of the estimated posterior means was small relative to the posterior variances, and of course this error was reduced further by averaging the five estimates. The results for the other parameters were similar.

It is useful to compare this Bayesian analysis to a naive estimate that ignores measurement error. The naive estimate from regressing $\mathbf{Y}_i$ on $\overline{\mathbf{W}}_i$ and $\mathbf{Z}_i$ was $\widehat{\beta}_x = 0.346$ with a standard error of 0.0233, so the naive estimator is only about half as variable as the Bayes estimator, but the mean square error of the naive estimator will be much larger and due almost entirely to bias. The estimated attenuation was 0.701 and so the bias-corrected estimate was $0.346/0.701 = 0.494$. Ignoring the uncertainty in the attenuation, the standard error of the bias-corrected estimate is $0.0233/0.701 = 0.0322$. This standard error is smaller than the posterior standard deviation but is certainly an underestimate of variability, and if we wanted to use the bias-corrected estimator we would want to use the bootstrap or the sandwich formula to get a better standard error.

In summary, in this example the Bayes estimate of β_x is similar to the naive estimate corrected for attenuation, which coincides with the regression calibration estimate. The Bayes estimator takes more work to program but gives a posterior standard deviation that takes into account uncertainty due to estimating other parameters. The estimator corrected for attenuation would require bootstrapping or some type of asymptotic approximation, e.g., the delta-method or the sandwich formula from estimating equations theory, to account for this uncertainty. However, for linear regression, Bayesian MCMC is a bit of overkill. The real strength of Bayesian MCMC is the ability to handle more difficult problems, e.g., segmented regression with multiplicative errors, a prob-

lem which appears not to have been discussed in the literature but which can be tackled by MCMC in a straightforward manner; see Section 9.1.4.

9.5 Nonlinear Models

The ideas in Section 9.4 can be generalized to complex regression models in $\mathbf{X}$.

9.5.1 A General Model

The models we will study are all special case of the following general outcome model

$$[\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta, \sigma_\epsilon] = \text{Normal}\{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta), \sigma_\epsilon^2\}, \tag{9.17}$$

where

$$m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta) = \phi(\mathbf{X}_i, \mathbf{Z}_i)^t\beta_1 + \psi(\mathbf{X}_i, \mathbf{Z}_i, \theta)^t\beta_2 \tag{9.18}$$

is a linear function in β_1, β_2 and nonlinear in θ. The functions ϕ, and ψ may include nonlinear terms in $\mathbf{X}$ and $\mathbf{Z}$ as well as interactions and may be scalar or vector valued. When $\psi \equiv 0$ particular cases of model (9.17) include linear and polynomial regression, interaction models, and multiplicative error models. An example of nonlinear component is $\psi(\mathbf{X}_i, \mathbf{Z}_i, \theta) = |\mathbf{X}_i - \theta|_+$ that appears in segmented regression with an unknown break point location. We assume that the other components of the linear model in Section 9.4 remain unchanged and that $\mathbf{X}_i$ is scalar, though this assumption could easily be relaxed. The unknowns in this model are $(\beta, \theta, \sigma_\epsilon, \sigma_u)$, $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, $(\alpha_0, \alpha_z, \sigma_x)$.

In addition to the priors considered in Section 9.4 we consider a general prior $\pi(\theta)$ for θ and assume that all priors are mutually independent. It is easy to check that the full conditionals $f(\alpha|\text{others})$, $f(\sigma_x^2|\text{others})$ and $f(\sigma_u^2|\text{others})$ are unchanged, and that

$$f(\sigma_\epsilon^2|\text{others}) = \text{IG}\left[\delta_{\epsilon,1} + (n/2),\ \delta_{\epsilon,2} + (1/2)\sum_{i=1}^{n}\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2\right].$$

Denoting by $\mathcal{C}(\theta)$ the matrix with i^{th} row

$$\boldsymbol{C}_i^t(\theta) = [\phi(\mathbf{X}_i, \mathbf{Z}_i), \psi(\mathbf{X}_i, \mathbf{Z}_i, \theta)],$$

letting $\beta = (\beta_1^t, \beta_2^t)^t$, and letting $\Delta = \sigma_\epsilon^2/\sigma_\beta^2$, the full conditional for β becomes normal with mean $\{\mathcal{C}(\theta)^t\mathcal{C}(\theta) + \Delta\boldsymbol{I}\}^{-1}\mathcal{C}(\theta)^t\mathbf{Y}$ and covariance matrix $\{\mathcal{C}(\theta)^t\mathcal{C}(\theta) + \Delta\boldsymbol{I}\}^{-1}$.

By grouping together all terms that depend on θ one obtains

$$f(\theta|\text{others}) \propto \exp\left[-\sum_{i=1}^{n}\frac{\{\mathbf{Y}_i^{(1)} - \psi(\mathbf{X}_i, \mathbf{Z}_i, \theta)\beta_2\}^2}{2\sigma_\epsilon^2}\right]\pi(\theta), \tag{9.19}$$

where $\mathbf{Y}_i^{(1)} = \mathbf{Y}_i - \phi(\mathbf{X}_i, \mathbf{Z}_i)\beta_1$. Since ψ is a nonlinear function in θ this full conditional is generally not in a known family of distributions regardless of how $\pi(\theta)$ is chosen. One can update θ using a random walk MH step using Normal$(\theta, B\sigma_\theta^2)$ as the proposal density, where B is tuned to get a moderate acceptance rate.

The full conditional for $\mathbf{X}_i$ is

$$f(\mathbf{X}_i|\text{others}) \propto \exp\left[-\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2/(2\sigma_\epsilon^2)\right] \quad (9.20)$$
$$\times\exp\left\{(\mathbf{X}_i - \alpha_0 - \alpha_z\mathbf{Z}_i)^2/(2\sigma_x^2) + k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right].$$

To update $\mathbf{X}_i$, we use a random walk MH step with Normal$(\mathbf{X}_i, B\,\sigma_u^2/k_i)$ with the "dispersion" factor, B, chosen to provide a reasonable acceptance rate.

We now discuss the details of implementation for polynomial, multiplicative measurement error, and segmented regression.

9.5.2 Polynomial Regression

A particular case of the outcome model (9.17) is the polynomial regression in $\mathbf{X}$

$$\mathbf{Y}_i = \mathbf{Z}_i^t\beta_z + \mathbf{X}_i\beta_{x,1} + \cdots + \mathbf{X}_i^p\beta_{x,p} + \epsilon_i, \quad (9.21)$$

for some $p > 1$, where ϵ_i are independent Normal$(0, \sigma_\epsilon^2)$, obtained by setting $\phi(\mathbf{X}_i, \mathbf{Z}_i) = (\mathbf{Z}_i^t, \mathbf{X}_i, \ldots, \mathbf{X}_i^p)$ and $\psi(\mathbf{X}_i, \mathbf{Z}_i, \theta) = 0$. The i^{th} row of $\mathcal{C} := \mathcal{C}(\theta)$ is $\boldsymbol{C}_i^t = \phi(\mathbf{X}_i, \mathbf{Z}_i)$ and $\beta = (\beta_z^t, \beta_{x,1}, \ldots, \beta_{x,p})^t$. With this notation, all full conditionals are as described in Section 9.5.1. In particular, the full conditional of θ in (9.19) is not necessary because $\psi = 0$. In this example, the full conditional for $\mathbf{X}_i$ is the only non–standard distribution and can be obtained as a particular case of (9.20) as

$$f(\mathbf{X}_i|\text{others}) \propto \exp\left\{-(\mathbf{Y}_i - \boldsymbol{C}_i^t\beta)^2/(2\sigma_\epsilon^2)\right\} \quad (9.22)$$
$$\times\exp\left\{-(\mathbf{X}_i - \alpha_0 - \mathbf{Z}_i^t\alpha_z)^2/(2\sigma_x^2) - k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right\}.$$

The full conditional for $\mathbf{X}_i$ is non-standard because $\boldsymbol{C}_i$ contains powers of $\mathbf{X}_i$.

To illustrate these ideas consider the quadratic regression in $\mathbf{X}$

$$\mathbf{Y}_i = \beta_0 + \beta_{x,1}\mathbf{X}_i + \beta_{x,2}\mathbf{X}_i^2 + \beta_z\mathbf{Z}_i + \epsilon_i, \quad (9.23)$$

with $\beta_{x,2} = 0.2$ and the other parameters unchanged. To update $\mathbf{X}_i$ the proposal density was Normal$(\mathbf{X}_i, B\,\sigma_u^2/k_i)$. After some experimentation, the "dispersion" factor B was chosen to be 1.5 to get approximately 25% acceptance. We found that the performance of the Gibbs sampler was not particularly sensitive to the value of B and B equal to 1 or 2.5 also worked well.

As in the linear example, we used five runs of the Gibbs sampler, each

with 10,000 iterations, and with the same starting value distribution as before. The posterior means of β_0, $\beta_{x,1}$, $\beta_{x,2}$, and β_z were 1.015, 0.493 0.191, and 0.348, close to the true values of the parameters which were 1.0, 0.5, 0.2, and 0.3. In contrast, the naive estimates obtained by fitting (9.23) with $\mathbf{X}_i$ replaced by $\overline{\mathbf{W}}_i$ were 1.18, 0.427, 0.104, and 0.394, so, in particular, the coefficient of $\mathbf{X}^2$ was biased downward by nearly 50%. The posterior standard deviations were 0.057, 0.056, 0.027, and 0.040, while the standard errors of the naive estimates were 0.079, 0.052, 0.021, and 0.049.

9.5.3 Multiplicative Error

We now show that a linear regression model (9.7) with multiplicative measurement error is a particular case of model (9.17). As discussed in Section 4.5, this model is relatively common in applications. Indeed, if $\mathbf{X}_i^* = \log(\mathbf{X}_i)$ and $\mathbf{W}_{i,j}^* = \log(\mathbf{W}_{i,j})$ then the outcome model becomes

$$Y_i = \mathbf{Z}_i^t \beta_z + e^{\mathbf{X}_i^*} \beta_x + \epsilon_i,$$

which can be obtained from (9.17) by setting $\phi(\mathbf{X}_i^*, \mathbf{Z}_i) = (\mathbf{Z}_i^t, e^{\mathbf{X}_i^*})$ and $\psi(\mathbf{X}_i^*, \mathbf{Z}_i, \theta) = 0$. The i^{th} row of $\mathcal{C} := \mathcal{C}(\theta)$ is $\boldsymbol{C}_i^t = \phi(\mathbf{X}_i^*, \mathbf{Z}_i)$ and $\beta = (\beta_z^t, \beta_x)^t$.

We replace the exposure model (9.9) by a lognormal exposure model where (9.24) holds with $\mathbf{X}_i$ replaced by , i.e.,

$$\mathbf{X}_i^* \sim \text{Normal}(\alpha_0 + \mathbf{Z}_i^t \alpha_z, \sigma_x^2). \tag{9.24}$$

The measurement model is

$$[\mathbf{W}_{i,j}^* | \mathbf{X}_i] \sim \text{Normal}(\mathbf{X}_i^*, \sigma_u^2),\ j = 1, \ldots, k_i,\ i = 1, \ldots, n. \tag{9.25}$$

With this notation the full conditionals for this model are the same as in Section 9.5.1. One trivial change is that $\mathbf{X}_i$ is replaced everywhere by $\mathbf{X}_i^*$ and the full conditional of θ is not needed because $\psi = 0$.

To illustrate these ideas we simulated 200 observations with $\beta_0 = 1$, $\beta_x = 0.3$, $\beta_z = 0.3$, $\alpha_0 = 0$, $\alpha_z = 0.2$, $\sigma_x = 1$, and $\sigma_u = 1$. The $\mathbf{Z}_i$ were Normal$(-1, 1)$. We ran the Gibbs sampler with tuning parameter $B = 2.5$ which gave a 30% acceptance rate. Figure 9.3 shows the output from one of five runs of the Gibbs sampler. There were 10,500 iterations of which the first 500 were discarded. One can see that β_0 and, especially, β_x mix more slowly than the other parameters, yet their mixing seems adequate. In particular, the standard deviation of the five posterior means for β_x was 0.0076 giving a Monte Carlo standard error of $0.0078/\sqrt{5} = 0.0034$. while the posterior standard deviation of that parameter was 0.0377 about ten times larger than the Monte Carlo standard error.

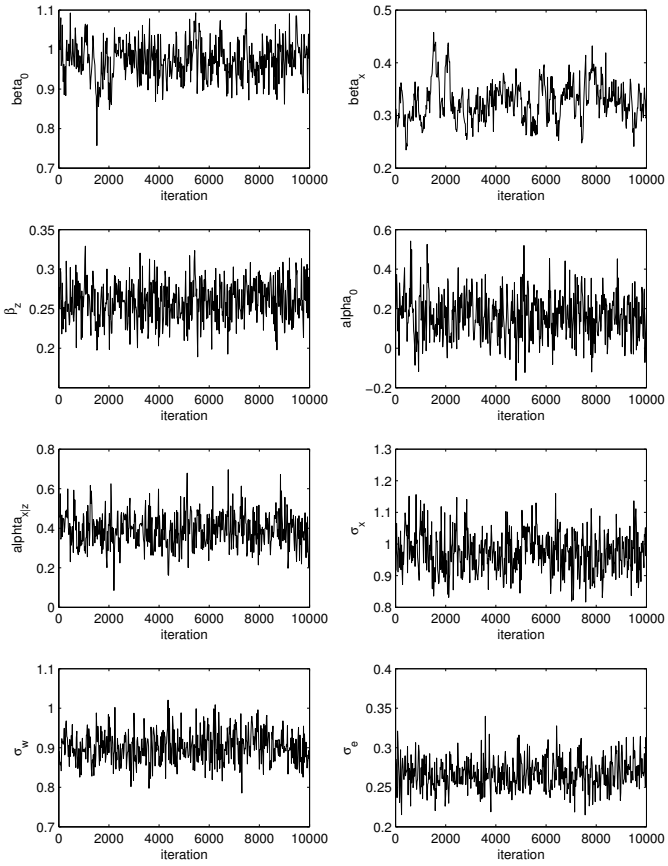

Figure 9.3 *Every 20th iteration of the Gibbs sampler for the linear regression example with multiplicative error.*

9.5.4 Segmented Regression

A commonly used regression model is a segmented line, that is, two lines joined together at a knot. This model can be written as

$$\mathbf{Y}_i = \mathbf{Z}_i^t \beta_z + \beta_{x,1}\mathbf{X}_i + \beta_{x,2}(\mathbf{X}_i - \theta)_+ + \epsilon_i, \tag{9.26}$$

where we use the notation $a_+ = \min(0,\ a)$, θ is the knot, $\beta_{x,1}$ is the slope of $\mathbf{Y}$ on $\mathbf{X}$ before the knot and $\beta_{x,2}$ is the change in this slope at the knot. An intercept could be included in $\mathbf{Z}_i^t\beta_z$.

The outcome model (9.26) is a particular case of model (9.17) with $\phi(\mathbf{X}_i, \mathbf{Z}_i) = (\mathbf{Z}_i^t, \mathbf{X}_i)$ and $\psi(\mathbf{X}_i, \mathbf{Z}_i, \theta) = (\mathbf{X}_i - \theta)_+$. The i^{th} row of $\mathcal{C}(\theta)$

is $\boldsymbol{C}_i^t(\theta) = \{\mathbf{Z}_i^t, \mathbf{X}_i, (\mathbf{X}_i - \theta)_+\}^t$ and $\beta = (\beta_z^t, \beta_{x,1}, \beta_{x,2})^t$. With this notation, all full conditionals are as described in Section 9.5.1.

To illustrate segmented regression with measurement error and unknown knot location we simulated data with $n = 200$, $J = 2$, $\beta_0 = 1$, $\beta_x = 1$, $\beta_{x,2} = 0.8$, $\beta_z = 0.1$, $\theta = 1$, $\alpha_0 = 1$, $\alpha_z = 0$, $\sigma_\epsilon = 0.15$, $\sigma_x = 1$, and $\sigma_u = 1$. The $\mathbf{Z}_i$ were Normal$(1, 1)$. Since $\alpha_z = 1$, the $\mathbf{X}_i$ were Normal$(1, 1)$ independently of the $\mathbf{Z}_i$.

We ran the Gibbs sampler five times, each with 10,000 iterations. Starting values for θ were Uniform$(0.5,\ 1.5)$. In the prior for θ, we used the Normal$(\mu_\theta, \sigma_\theta^2)$ distribution with $\mu_\theta = \overline{\mathbf{W}}$ and $\sigma_\theta = 5\, s(\overline{\mathbf{W}})$ where $s(\overline{\mathbf{W}})$ was the sample standard deviation of $\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_n$. This prior was designed to have high prior probability over the entire range of observed values of $\mathbf{W}$. In the proposal density for θ, we used $B = 0.01$. This value was selected by trial and error and gave an acceptance rate of 36% and adequate mixing. The posterior mean and standard deviation of θ were 0.93 and 0.11, respectively. The Monte Carlo standard error of the posterior mean was only 0.005.

Figure 9.4 reveals how well the Bayesian modeling imputes the $\mathbf{X}_i$ and leads to good estimates of θ. The top, left plot shows the true $\mathbf{X}_i$ plotted with the $\mathbf{Y}_i$. The bottom, right plot is similar, except that instead of the unknown $\mathbf{X}_i$ we use the imputed $\mathbf{X}_i$ from the 10,000th iteration of the fifth run of the Gibbs sampler. Notice that the general pattern of $\mathbf{X}$ versus $\mathbf{Y}$ is the same for the true and the imputed $\mathbf{X}_i$. In contrast, a plot of $\mathbf{Y}_i$ and either $\overline{\mathbf{W}}_i$ or $\widehat{E}(\mathbf{X}_i|\overline{\mathbf{W}}_i) = (1-\widehat{\lambda})\overline{\mathbf{W}}_i + \widehat{\lambda}\overline{\mathbf{W}}_i$ shows much less similarity with the $(\mathbf{X}_i, \mathbf{Y}_i)$ plot. Here $\widehat{\lambda}$ is the estimated attenuation and $\overline{\mathbf{W}}$ is the mean of $\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_n$.

The plot of the imputed $\mathbf{X}_i$ versus $\mathbf{Y}_i$ shows the existence and location of the knot quite clearly, and it is not surprising that θ can be estimated with reasonably accuracy. Of course, this "feedback" of information about the $\mathbf{X}_i$ to information about θ works both ways. Accurate knowledge of θ well helps impute the $\mathbf{X}_i$. One estimates both the $\mathbf{X}_i$ and θ well in this example because their joint posterior has highest probability near their true values.

9.6 Logistic Regression

In this section, we assume the same model with nonlinear measurement error as in Section 9.5 but with a binary outcome. We use the logistic regression model

$$\log\left\{\frac{P(\mathbf{Y}_i = 1|\mathbf{X}_i, \mathbf{Z}_i)}{P(\mathbf{Y}_i = 0|\mathbf{X}_i, \mathbf{Z}_i)}\right\} = m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)$$

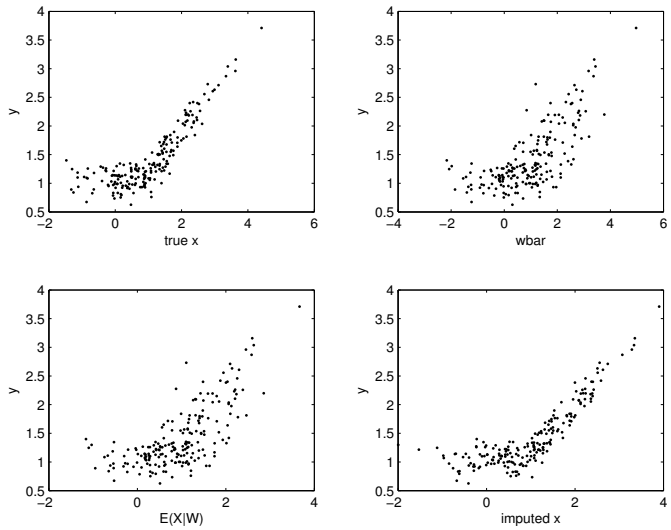

Figure 9.4 *Segmented regression. Plots of* $\mathbf{Y}_i$ *and* $\mathbf{X}_i$ *and three estimator of* $\mathbf{X}_i$*. Top left:* $\mathbf{Y}$ *plotted versus the true* $\mathbf{X}$*. Top right:* $\mathbf{Y}$ *plotted versus the mean of the replicated* $\mathbf{W}$*-values. Bottom left:* $\mathbf{Y}$ *plotted versus the regression calibration estimates of* $\mathbf{X}$*. Bottom right:* $\mathbf{Y}$ *plotted versus the imputed* $\mathbf{X}$ *in a single iteration of the Gibbs sampler. Note how the Gibbs sampler more faithfully reproduces the true* $\mathbf{X}$*-values.*

so the outcome likelihood is proportional to

$$\exp\left[\sum_{i=1}^{n}\mathbf{Y}_i m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)-\sum_{i=1}^{n}\log\left\{1+e^{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)}\right\}\right],$$

$$[\beta,\theta|\text{others}]\propto\exp\left[\sum_{i=1}^{n}\mathbf{Y}_i m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)-\sum_{i=1}^{n}\log\left\{1+e^{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)}\right\}-\frac{\beta^t\beta}{2\sigma_\beta^2}\right]\pi(\theta), \tag{9.27}$$

and

$$[\mathbf{X}_i|\text{others}]\propto\exp\left[\sum_{i=1}^{n}\mathbf{Y}_i m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)-\sum_{i=1}^{n}\log\left\{1+e^{m(\mathbf{X}_i,\mathbf{Z}_i,\beta,\theta)}\right\}+\frac{(\mathbf{X}_i-\alpha_0-\alpha_z\mathbf{Z}_i)^2}{\sigma_x^2}+\frac{(\overline{\mathbf{W}}_i-\mathbf{X}_i)^2}{\sigma_{\overline{\mathbf{W}}}^2}\right]. \tag{9.28}$$

To update $\mathbf{X}_i$ we use a random-walk MH step with the same Normal($\mathbf{X}_i$, $B\,\sigma^2_{\overline{\mathbf{W}}}$) proposal density as for polynomial regression. To update β we use

a random-walk MH step with proposal density $N\{\beta, B'\text{var}(\widehat{\beta})\}$ where $\text{var}(\widehat{\beta})$ is the covariance matrix of the naive logistic regression estimator using $\overline{\mathbf{W}}$ in place of $\mathbf{X}$ and B' is another tuning constant. A similar strategy may be applied to update θ when ψ in (9.18) is not identically zero.

To illustrate the fitting algorithms for logistic regression with measurement error we simulated data from a quadratic regression similar to the one in Section 9.5.2 but with a binary response following the logistic regression model. The intercept β_0 was changed to -1 so that there were roughly equal numbers of 0's and 1's among the $\mathbf{Y}_i$. Also, the sample size was increased to $n = 1500$ to ensure reasonable estimation accuracy for β. Otherwise, the parameters were the same as the example in Section 9.5.2. The tuning parameters in the MH steps were $B = B' = 1.5$. This gave acceptance rates of about 52% for the $\mathbf{X}_i$ and about 28% for β.

Figure 9.5 show the output from one of the five runs of the Gibbs sampler. The samplers appear to have converged and to have mixed reasonably well. The posterior mean of β was $(-1.18, 0.55, 0.24, 0.30)$ which can be compared to $\beta = (-1, 0.5, 0.2, 0.3)$. The posterior standard deviations were $(0.13, 0.17, 0.09, 0.06)$. The Monte Carlo error, as measured by the between-run standard deviations of the posterior means, was less than one-tenth as large as the posterior standard deviations.

9.7 Berkson Errors

The Bayesian analysis of Berkson models is similar to, but somewhat simpler than, the Bayesian analysis of error models. The reason for the simplicity is that we only need a Berkson error model for $[\mathbf{X}|\mathbf{W}]$ or $[\mathbf{X}|\mathbf{W}, \mathbf{Z}]$. If instead we had an error model $[\mathbf{W}|\mathbf{X}, \mathbf{Z}]$ then, as we have seen, we would also need a structural model $[\mathbf{X}|\mathbf{Z}]$.

We will consider nonlinear regression with a continuously distributed $\mathbf{Y}$ first and then logistic regression.

9.7.1 Nonlinear Regression with Berkson Errors

Suppose that we have outcome model (9.17), which for the reader's convenience is

$$[\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta, \sigma_\epsilon] = \text{Normal}\{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta), \sigma_\epsilon^2\}, \tag{9.29}$$

but now with Berkson error so that we observe $\mathbf{W}_i$ where

$$\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i, \ E(\mathbf{U}_i|\mathbf{Z}_i, \mathbf{W}_i) = 0.$$

Model (9.29) is nonlinear in general, but includes linear models as a special case. The analysis in Section 9.5.1, which was based upon repli-

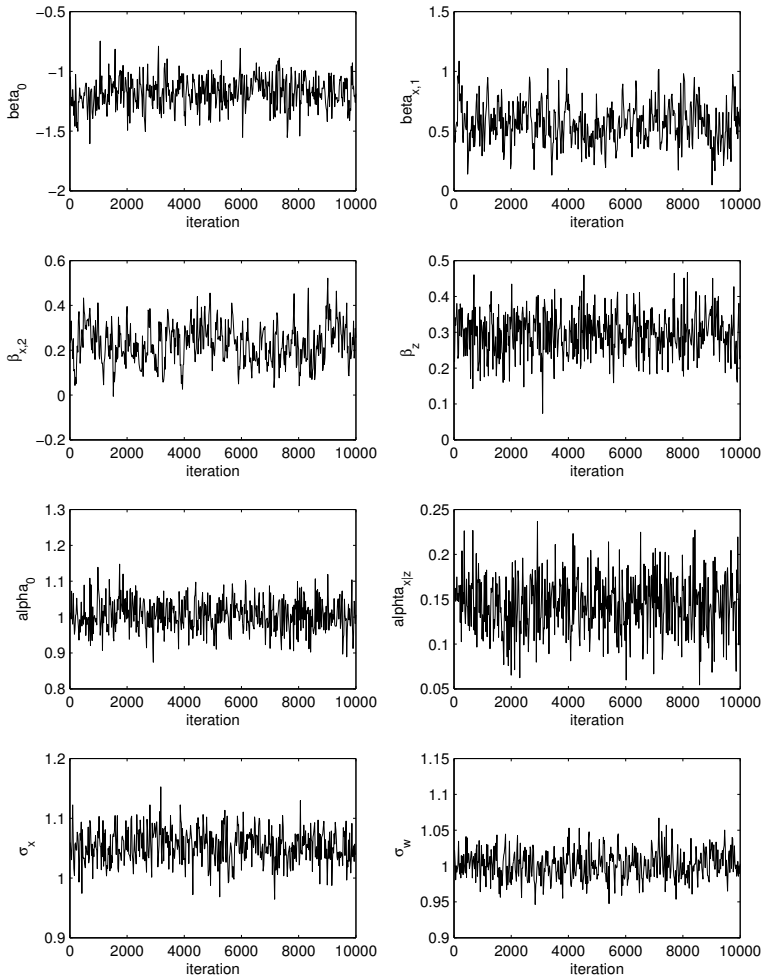

Figure 9.5 *Every 20th iteration of the Gibbs sampler for the quadratic logistic regression example.*

cated classical measurement error and a structural model that says that $\mathbf{X}|\mathbf{Z} \sim \text{Normal}(\alpha_0 + \alpha_z\mathbf{Z})$, must be changed slightly because of the Berkson errors. The only full conditionals that change are for the $\mathbf{X}_i$. Specifically, equation (9.20), which is

$$f(\mathbf{X}_i|\text{others}) \propto \exp\left[-\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2/(2\sigma_\epsilon^2)\right] \times \exp\left\{-(\mathbf{X}_i - \alpha_0 - \alpha_z\mathbf{Z}_i)^2/(2\sigma_x^2) - k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right\},$$

is modified to

$$f(\mathbf{X}_i|\text{others}) \propto \exp\left[-\{\mathbf{Y}_i - m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)\}^2/(2\sigma_\epsilon^2)\right] \quad (9.30)$$
$$\times \exp\left\{-(\mathbf{W}_i - \mathbf{X}_i)^2/(2\sigma_u^2)\right] .$$

Thus, we see two modifications. The term $-(\mathbf{X}_i - \alpha_0 - \alpha_z\mathbf{Z}_i)^2/(2\sigma_x^2)$ in (9.20), which came from the structural assumption, is not needed and $k_i(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2$ is replaced by $(\mathbf{W}_i - \mathbf{X}_i)^2$ since there are no replicates in the Berkson model. That's it for changes—everything else is the same!

This analysis illustrates a general principle, which may have been obvious to the reader, but should be emphasized. When we have a Berkson model which gives $[\mathbf{X}|\mathbf{Z}, \mathbf{W}]$, we do not need a model for marginal density $[\mathbf{W}]$ of $\mathbf{W}$— the $\mathbf{W}_i$ are observed so that we can condition upon them. In contrast, if we have a error model for $[\mathbf{W}|\mathbf{Z}, \mathbf{X}]$, we cannot do a conditional analysis given the $\mathbf{X}_i$ since these are unobserved, and therefore a structural model for $[\mathbf{X}]$ or, perhaps, $[\mathbf{X}|\mathbf{Z}]$ is also needed.

9.7.2 Logistic Regression with Berkson Errors

When errors are Berkson, the analysis of a logistic regression model described in Section 9.6 changes in a way very similar to the changes just seen for nonlinear regression. In particular, equation (9.28), which is

$$[\mathbf{X}_i|\text{others}] \propto \exp\left[\sum_{i=1}^n \mathbf{Y}_i m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta) - \sum_{i=1}^n \log\left\{1 + e^{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)}\right\}\right.$$
$$\left. + \frac{(\mathbf{X}_i - \alpha_0 - \alpha_z\mathbf{Z}_i)^2}{\sigma_x^2} + \frac{(\overline{\mathbf{W}}_i - \mathbf{X}_i)^2}{\sigma_{\overline{\mathbf{W}}}^2}\right],$$

becomes

$$[\mathbf{X}_i|\text{others}] \propto \exp\left[\sum_{i=1}^n \mathbf{Y}_i m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta) - \sum_{i=1}^n \log\left\{1 + e^{m(\mathbf{X}_i, \mathbf{Z}_i, \beta, \theta)}\right\}\right.$$
$$\left. + \frac{(\mathbf{W}_i - \mathbf{X}_i)^2}{\sigma_u^2}\right]. \quad (9.31)$$

As before, the term $(\mathbf{X}_i - \alpha_0 - \alpha_z\mathbf{Z}_i)^2/\sigma_x^2$ in (9.28) came from the structural model and is not needed for a Berkson analysis and $\overline{\mathbf{W}}_i$ is replaced by $\mathbf{W}_i$ because there is no replication.

9.7.3 Bronchitis Data

We now continue the analysis of the bronchitis data described in Section 8.7. Recall, that in that section we found that the MLE of the Berkson measurement error standard deviation, σ_u, was zero. Our Bayesian

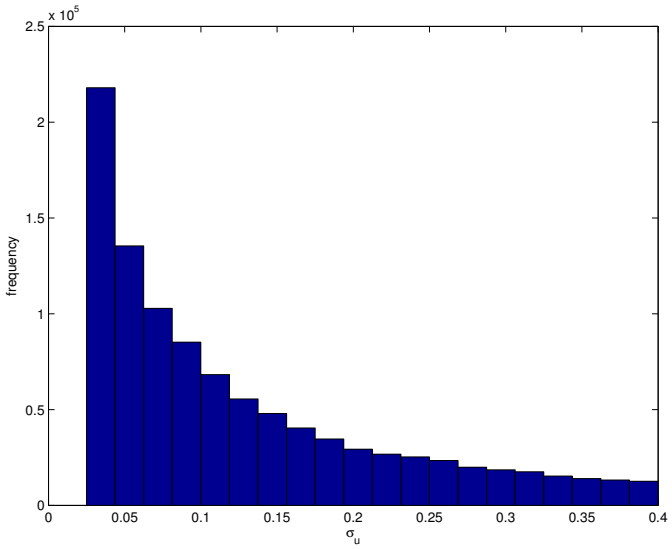

Figure 9.6 *Munich bronchitis data. Histogram of 1,250,000 samples from the posterior for* σ_u.

analysis will show that σ_u is poorly determined by the data. Although σ_u is theoretically identifiable, for practical purposes it is not identified. Gustafson (2005) has an extensive discussion of non-identified models. He argues in favor of using informative priors on non-identified nuisance parameters, such as σ_u here. The following analysis applies Gustafson's strategy to σ_u.

We will use a Uniform $(0.025, 0.4)$ prior for σ_u. This prior seems reasonable, since σ_w is 0.72, so the lower limit of the prior implies very little measurement error. Also, the upper limit is over twice the value, 0.187, assumed in previous work by Gössi and Küchenhoff (2001). We will use a Uniform $\{1.05 \min(W_i),\ 0.95 \max(W_i)\}$ prior for $\beta_{x,2}$. This prior is reasonable since $\beta_{x,2}$ is a TLV (threshold limiting value) within the range of the observed data. The prior on β, the vector of all regression coefficient, is Normal$(0, 10^6 I)$.

There were five MCMC runs, each of 250,000 iterations excluding a burn-in of 1000 iterations. Figure 9.6 is a histogram of the 1,250,000 values of σ_u^2 from the five runs combined. The posterior is roughly proportional to the likelihood, since there are uniform priors on σ_u and $\beta_{x,2}$ and a very diffuse prior on β. The histogram is monotonically decreasing, in agreement with the MLE of 0 for σ_u. However, the posterior is very diffuse and much larger values of σ_u are plausible under the poste-

rior. In fact, the posterior mean, standard deviation, 0.025 quantile, and 0.975 quantile of σ_u were 0.13, 0.098, 0.027, and 0.37, respectively. The 95% credible interval of (0.027, 0.37) is not much different than (0.0344, 0.3906), the interval formed by the 2.5 and 97.5 percentiles of the prior. Thus, the data provide some, but not much, information about σ_u.

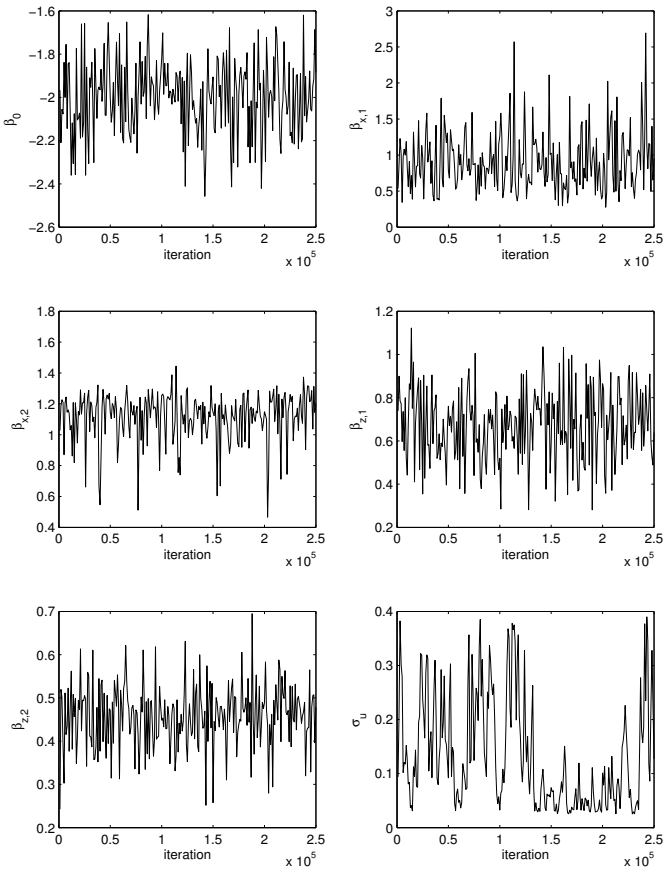

Figure 9.7 *Trace plots for the Munich bronchitis data.*

Figure 9.7 shows trace plots for the first of the five MCMC runs. Trace plots for the other runs are similar. The mixing for σ_u is poor, but the mixing for the other parameters is much better. The poor mixing of σ_u was the reason we used 250,000 iterations per run rather than a smaller value such as 10,000 that was used in previous examples.

We experimented with a Uniform$(0, 10)$ prior for σ_u and encountered difficulties. On some runs, the sampler would get stuck at $\sigma_u = 0$ and

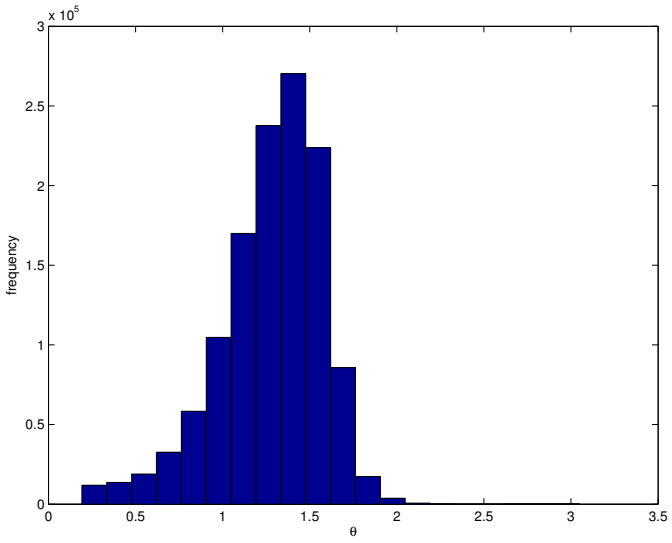

Figure 9.8 *Munich bronchitis data. Histogram of 1,250,000 samples from the posterior for TLV,* $\beta_{x,2}$.

$\mathbf{X}_i = \mathbf{W}_i$ for all i. On runs where this problem did not occur the mixing was very poor for σ_u and fair to poor for the other parameters. We conclude that a reasonably informative prior on σ_u is necessary. However, fixing σ_u at a single value, as Gössi and Küchenhoff (2001) have done, is not necessary.

Figure 9.8 is a histogram of the 1,250,000 value of $\beta_{x,2}$ from the combined runs with burn-ins excluded. The posterior mean of $\beta_{x,2}$ was 1.28, very close to the naive of 1.27 found in Section 8.7. This is not surprising since the simulations in Section 8.7.3 showed that the naive estimator had only a slight negative bias. The 95% highest posterior density credible interval was (0.53, 1.73).

9.8 Automatic implementation

Bayesian analysis for complex models with covariates measured with error needs to be based on carefully constructed prior, full conditional and proposal distributions combined with critical examination of the convergence and mixing properties of the Markov Chains. The MATLAB programs used in the previous sections are specially tailored and optimized to address these issues. However, standard software such as WinBUGS may prove to be a powerful additional tool in applications

where many models are explored. We now show how to use WinBUGS for fitting models introduced in Sections 9.4 and 9.5.

9.8.1 Implementation and simulations in WinBUGS

We describe in detail the implementation of the linear model in Section 9.4 and note only the necessary changes for the more complex models. The complete commented code presented in Appendix B.8.1 follows step–by–step the model description in Section 9.4.

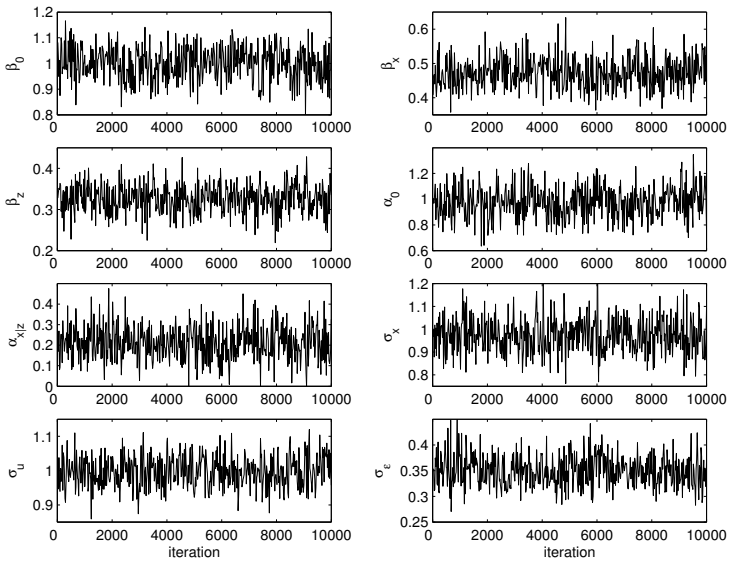

Figure 9.9 *Every* 20^{th} *iteration for the WinBUGS Gibbs sampler for the linear regression example.*

The first `for` loop specifies the outcome, measurement and exposure model (9.7), (9.8), and (9.9). Note that `Nobservations` is the sample size and that the # sign indicates a comment. The code is structured and intuitive. For example, the two lines in the outcome model

```
Y[i]~dnorm(meanY[i],taueps)
meanY[i]<-beta[1]+beta[2]*X[i]+beta[3]*Z[i]
```

specify that the outcome of the i^{th} subject, $\mathbf{Y}_i$, has a normal distribution with mean $m_Y(i) = \beta_1 + \beta_2\mathbf{X}_i + \beta_3\mathbf{Z}_i$ and precision parameter $\tau_\epsilon = 1/\sigma^2_\epsilon$. It is quite common in Bayesian analysis to specify the normal distribution in terms of its precision instead of its variance.

The nested `for` loop corresponding to the replication model

```
for (j in 1:Nreplications) {W[i,j]~dnorm(X[i],tauu)}
```

specifies that, conditional on the unobserved exposure, $\mathbf{X}_i$, of the i^{th} subject the proxies $\mathbf{W}_{i,j}$ are normally distributed with mean $\mathbf{X}_i$ and precision $\tau_u = 1/\sigma_u^2$. Here `Nreplications` is the number of replications and it happened to be the same for all subjects. A different number of replications could easily be accommodated by replacing the scalar `Nreplications` by a vector `Nreplications[]`.

The code corresponding to the measurement error model

```
X[i]~dnorm(meanX[i],taux)
meanX[i]<-alpha[1]+alpha[2]*Z[i]
```

specifies that the exposure of the i^{th} subject, $\mathbf{X}_i$, has a normal distribution with mean $\alpha_1 + \alpha_2 \mathbf{Z}_i$ and precision parameter $\tau_x = 1/\sigma_x^2$.

The code for prior distributions

```
tauu~dgamma(3,1)
taueps~dgamma(3,1)
taux~dgamma(3,1)
```

specifies that the precision parameters $\tau_u, \tau_\epsilon, \tau_x$ have independent Gamma priors with parameters 3 and 1. The `dgamma(a,b)` notation in WinBUGS specifies a Gamma distribution with mean a/b and variance a/b^2. The code for prior distributions

```
for (i in 1:nalphas){alpha[i]~dnorm(0,1.0E-6)}
for (i in 1:nbetas){beta[i]~dnorm(0,1.0E-6)}
```

specifies that the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3$ have independent normal priors with mean zero and precision 10^{-6}. Here `nalphas` and `nbetas` denote the number of α and β parameters.

The last part of the code contains only definitions of explicit functions of the model parameters. For example

```
sigmaeps<-1/sqrt(taueps)
sigmau<-1/sqrt(tauu)
sigmax<-1/sqrt(taux)
```

define the standard deviations $\sigma_\epsilon = 1/\sqrt{\tau_\epsilon}$, $\sigma_u = 1/\sqrt{\tau_u}$ and $\sigma_x = 1/\sqrt{\tau_x}$ for the outcome, replication and exposure models respectively and

```
lambda<-tauu/(tauu+taux)
```

defines the reliability ratio $\lambda = \tau_u/(\tau_u + \tau_x) = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$.

To assess the quality of inference based on the WinBUGS program, we simulated $2,000$ data sets from the linear model with measurement

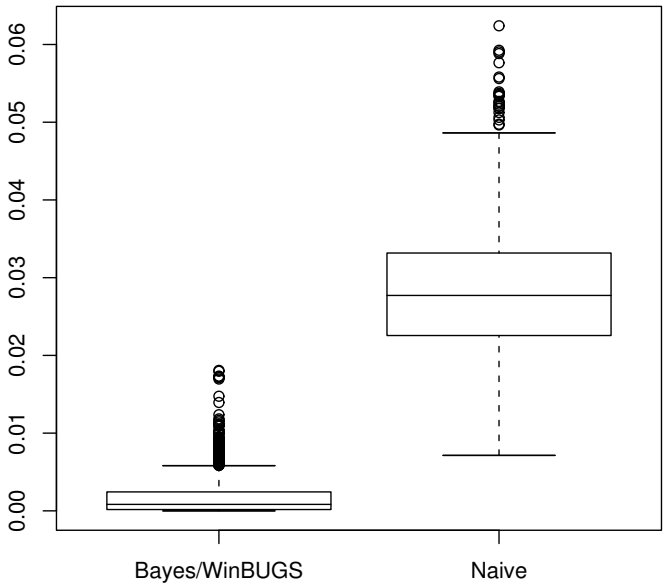

Figure 9.10 *Squared error for the Bayes and Naive methods for estimating the exposure effect β_x in the linear model with measurement error in the model (9.16).*

error described in Section 9.4.1. For each data set we used $10,500$ simulations based on the WinBUGS program and we discarded the first 500 simulations as burn in.

Figure 9.9 shows every 20th iteration of the Gibbs sampler for one data set indicating that the mixing properties are comparable to those shown in Figure 9.2. However, this is not always the case and WinBUGS programs typically need 10 to 100 times more simulations than expert programs to achieve comparable estimation accuracy. Of course, the time saved by using WinBUGS instead of writing a program often compensates for the extra computational time.

Figure 9.10 displays the squared error of the posterior mean of the exposure effect β_x using Bayes and naive estimators for the linear model with measurement error introduced in Section 9.4. More precisely, for the

d^{th} data set, $d = 1, \ldots, 2000$, denote by $\widehat{\beta}_{x,d}^{(B)}$ the posterior mean of β_x using the WinBUGS program and by $\widehat{\beta}_{x,d}^{(N)}$ the MLE of β_x in a standard linear regression where $\mathbf{X}_i$ is replaced by $\overline{\mathbf{W}}_i = (\mathbf{W}_{i1} + \mathbf{W}_{i2})/2$. Then, the two boxplots in Figure 9.9 correspond to $(\widehat{\beta}_{x,d}^{(B)} - \beta_x)^2$ and $(\widehat{\beta}_{x,d}^{(N)} - \beta_x)^2$ respectively.

We also calculated the coverage probabilities of β_x by the 90% and 95% equal-tail probability credible intervals obtained from the Bayesian analysis based on MCMC simulations implemented in WinBUGS. The true value of the parameter β_x was covered for 89.5% and 94.6% of the data sets by the 90% and 95% credible intervals respectively. In contrast, the true value of β_x was never covered by the 95% confidence interval of the naive analysis because of its bias.

9.8.2 More complex models

Only minor changes are necessary to fit the quadratic polynomial regression model in Section 9.5.2. Indeed, the only change is that the specification of the mean function of the outcome model becomes

```
meanY[i]<-beta[1]+beta[2]*X[i]+beta[3]*pow(X[i],2)
         +beta[4]*Z[i]
```

while the number of β parameters in the data `nbetas` is changed from 3 to 4. Here `pow(X[i],2)` represents $\mathbf{X}_i^2$.

As discussed in Section 9.5.3, the multiplicative measurement error model is equivalent with an additive measurement error model using a log exposure scale. This can be achieved by the transformations $\mathbf{W}_{i,j}^* = \log(\mathbf{W}_{i,j})$ and $\mathbf{X}_i^* = \log(\mathbf{X}_i)$. From a notational perspective in WinBUGS, there is no need to use the $\mathbf{X}_i^*$ notation instead of the $\mathbf{X}_i$ as long as the data is transformed accordingly. Therefore, the only necessary change is that the mean function of the outcome model becomes

```
meanY[i]<-beta[1]+beta[2]*exp(X[i])+beta[3]*Z[i]
```

where `exp(X[i])` represents $e^{\mathbf{X}_i^*}$ and `W[i,j]` represents $\mathbf{W}_{i,j}^*$.

To fit the segmented regression model in Section 9.5.4 one needs to change the mean function of the outcome model to

```
meanY[i]<-beta[1]+beta[2]*X[i]
         +beta[3]*(X[i]-theta)*step(X[i]-theta)
         +beta[4]*Z[i]
```

where `(X[i]-theta)*step(X[i]-theta)` represents $(\mathbf{X}_i - \theta)_+$ because `step(a)` in WinBUGS is equal to a if $a > 0$ and 0 otherwise. One needs only add the prior for θ

```
theta~dnorm(barWbar,prec.theta)
```

where `barWbar` represents the average of all W_{ij} observations and `prec.theta` represents $1/(25\sigma^2_{\overline{\mathbf{W}}})$ and are part of the data.

WinBUGS uses a rather inefficient simulation algorithm for fitting complex measurement error models. This is most probably due to the sampling scheme which updates one parameter at a time and does not take advantage of the explicit full conditionals of groups of parameters. For example, if $\gamma = (\gamma_1, \gamma_2)^t$ has a full conditional Normal$(\mu_\gamma, \Sigma_\gamma)$ with a very strong posterior correlation it is much more efficient to sample directly from Normal$(\mu_\gamma, \Sigma_\gamma)$ then to sample γ_1 given γ_2 and the others and then γ_2 given γ_1 and the others.

Therefore, the mixing properties of the Markov Chains generated by WinBUGS should be carefully analyzed using multiple very long chains. We also found that simple reparameterizations such as centering and orthogonalization of covariates can substantially improve mixing.

While we encourage development, when feasible, of expert programs along the lines described in Sections 9.4 and 9.5, WinBUGS can be a valuable additional tool. The main strengths of WinBUGS are:

1. Flexibility – moderate model changes correspond to simple program changes.
2. Simplicity – program follows almost literally the statistical model.
3. Robustness – program is less prone to errors.
4. Operability – programs can be called from different environments, such as R or MATLAB.

The main weakness of WinBUGS is that chains may exhibit very poor mixing properties when parameters have high posterior correlations. This problem may be avoided by expert programs through the careful study of full conditional distributions.

9.9 Cervical Cancer and Herpes

So far in this chapter, we have assumed that a continuously distributed covariate is measured with error. However, Bayesian analysis is straightforward when a discrete covariate is misclassified.

In this section, we continue the analysis given in Section 8.4 of the cervical cancer data discussed in Section 1.6.10. In particular, we continue the retrospective parameterization in Section 8.4 using $\alpha_{xd} = \Pr(\mathbf{W} = 1|\mathbf{X} = x, \mathbf{Y} = d)$ and $\gamma_d = \Pr(\mathbf{X} = 1|\mathbf{Y} = d)$, $x = 0, 1$ and $d = 0, 1$.

We use beta priors with parameters (a_{xd}, b_{xd}) for the α's and (a^*_d, b^*_d) for the γ's, with the α's and γ's being mutually independent. If we impose the constraints, $\alpha_{x0} = \alpha_{x1}$ for $x = 0, 1$, then we have a four-parameter,

nondifferential measurement error model. The log-odds ratio is related to the γ's by

$$\beta = \log\left[\{\gamma_1/(1-\gamma_1)\}/\{\gamma_0/(1-\gamma_0)\}\right].$$

Thus, the posterior distribution of β can be found via transformation from the posterior distribution of the γ's.

If we could observe all the $\mathbf{X}$'s, the joint density of the parameters and all the data would be proportional to

$$\prod_{x=0}^{1}\prod_{d=0}^{1}\left[\alpha_{xd}^{a_{xd}-1}\left(1-\alpha_{xd}\right)^{b_{xd}-1}\right. \tag{9.32}$$
$$\left.\times\prod_{i=1}^{n}\left\{\alpha_{xd}^{\mathbf{W}_i}\left(1-\alpha_{xd}\right)^{1-\mathbf{W}_i}\right\}^{I(\mathbf{X}_i=x,\mathbf{Y}_i=d)}\right]$$
$$\times\prod_{d=0}^{1}\left[\gamma_d^{a_d^*-1}\left(1-\gamma_d\right)^{b_d^*-1}\prod_{i=1}^{n}\left\{\gamma_d^{\mathbf{X}_i}\left(1-\gamma_d\right)^{1-\mathbf{X}_i}\right\}^{I(\mathbf{Y}_i=d)}\right].$$

We can use (9.4) and (9.32) to note that the posterior distribution of γ_d is a beta distribution with parameters $\sum_{i=1}^{n}\mathbf{X}_i I(\mathbf{Y}_i=d)+a_d^*$ and $\sum_{i=1}^{n}(1-\mathbf{X}_i)I(\mathbf{Y}_i=d)+b_d^*$. The posterior distribution of α_{xd} is also a beta distribution but with parameters $\sum_{i=1}^{n}\mathbf{W}_i I(\mathbf{X}_i=x,\mathbf{Y}_i=d)+a_{xd}$ and $\sum_{i=1}^{n}(1-\mathbf{W}_i)I(\mathbf{X}_i=x,\mathbf{Y}_i=d)+b_{xd}$. The conditional distribution of a missing $\mathbf{X}_i$, given the $(\mathbf{W}_i,\mathbf{Y}_i)$ and the parameters, is Bernoulli with success probability $p_{1i}/(p_{0i}+p_{1i})$, where

$$p_{xi}=\gamma_{\mathbf{Y}_i}^{x}\left(1-\gamma_{\mathbf{Y}_i}\right)^{1-x}\alpha_{x\mathbf{Y}_i}^{\mathbf{W}_i}\left(1-\alpha_{x\mathbf{Y}_i}\right)^{1-\mathbf{W}_i}.$$

Thus, in order to implement the Gibbs sampler, we need to simulate observations from the Bernoulli and beta distributions, both of which are easy to do using standard programs, so the Metropolis-Hastings algorithm was not needed.

For nondifferential measurement error, the only difference in these calculations is that $\alpha_{x0}=\alpha_{x1}=\alpha_x$, which have a beta prior with parameters (a_x,b_x) and a beta posterior with parameters $\sum_{i=1}^{n}\mathbf{W}_i I(\mathbf{X}_i=x)+a_x$ and $\sum_{i=1}^{n}(1-\mathbf{W}_i)I(\mathbf{X}_i=x)+b_d$.

We used uniform priors throughout, so that $a_{xd}=b_{xd}=a_d^*=b_d^*=1$. We ran the Gibbs sampling with an initial burn-in period of $2,000$ simulations, and then recorded every 50th simulation thereafter. The posterior modes were 0.623 and 0.927, respectively, these being very close to the maximum likelihood estimates. Note the large difference between the estimates for $d=1$ and for $d=0$, indicating the critical nature of whether or not the error is assumed to be nondifferential.

This example shows the value of validation data—without it, one is forced to assume nondifferential error and may, unwittingly, reach erro-

neous conclusions because this assumption does not hold. If at all feasible, the collection of validation is worth the extra effort and expense.

9.10 Framingham Data

As an illustration, we consider only those males ages 45+ whose cholesterol values at Exam #3 ranged from 200 to 300, giving a data set of $n = 641$ observations. Recall that $\mathbf{Y}$ is the indicator of coronary heart disease. Initial frequentist analysis of this data set showed no evidence of age or cholesterol effects, so we work only with two covariates, smoking status, $\mathbf{Z}$, and $\mathbf{X} = \log(\text{SBP}-50)$, where SBP is long-term average systolic blood pressure. The main surrogate $\mathbf{W}$ is the measurement of $\log(\text{SBP}-50)$ at Exam #3, while the replicate $\mathbf{T}$ is $\log(\text{SBP}-50)$ measured at Exam #2. Given $(\mathbf{Z}, \mathbf{X})$, $\mathbf{W}$ and $\mathbf{T}$ are assumed independent and normally distributed with mean $\mathbf{X}$ and variance σ_u^2; $\sigma_u^2 = \widetilde{\alpha}_1$ in the general notation of Chapter 8. The distribution of $\mathbf{X}$ given $\mathbf{Z}$ is assumed to be normal with mean $\alpha_0 + \alpha_z \mathbf{Z}$ and variance $\sigma_{x|z}^2$ ($\widetilde{\alpha}_2$ in the general notation). We also assume that $\sigma_{x|z}^2$ is constant, i.e., independent of $\mathbf{Z}$. Let $\Theta = (\sigma_u^2, \alpha_0, \alpha_z, \sigma_{x|z}^2)$.

Previous analysis suggested that the measurement error variance is less than 50% of the variance of the true long-term SBP given smoking status. We define $\Delta = \sigma_u^2/\sigma_{x|z}^2$ to be the ratio of these variances and assume $\Delta \in (0, 0.5)$. Restricting the range here makes sense, and we would not credit an analysis that suggested that the measurement error variance is larger than the variance of true long-term SBP given smoking status.

The Bayesian analysis will be based on the original model, so that $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z})$ is treated as being logistic with mean

$$H(\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z}) \, .$$

The unknown parameters are $(\beta_0, \beta_x, \beta_z, \alpha_0, \alpha_z, \sigma_{x|z}^2, \Delta)$. The first five of these are given diffuse (noninformative) locally uniform priors, the next-to-last has a diffuse inverse Gamma prior, the density functions being proportional to $1/\sigma_{x|z}^2$, and Δ has a uniform prior on the interval between zero and one half.

We use WinBUGS to implement the Bayesian logistic regression model. The WinBUGS model together with an R file used for data and output manipulation are provided as part of the software files for this book.

Mixing was very good for β_z, α_0, α_z, $\sigma_{x|z}^2$, σ_u^2 and λ. For these parameters $1,000$ burn-in and $10,000$ simulations were enough for accurate estimation. However, the chains corresponding to β_0 and β_x were mixing very slowly and we ran $310,000$ iterations of the Gibbs algorithm and discarded the first $10,000$ as burn-in. Figure 9.11 displays every 600^{th}

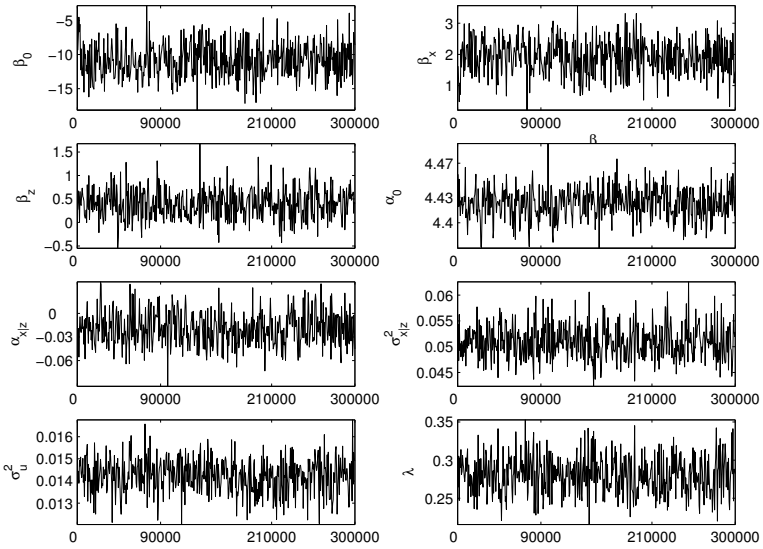

Figure 9.11 *Every 600th iteration of the Gibbs sampler for Framingham example.*

iteration for the model parameters with similar, but less clear patterns, for the un–thinned chains.

Table 9.1 compares the inference results for the maximum likelihood analysis based on the regression calibration approximation with the Bayesian inference based on Gibbs sampling. Clearly, the two types of inferences agree reasonably closely on most parameters. The Bayesian analysis estimates an 8.5% higher effect of SBP $\beta_x = 1.91$ for Gibbs sampling compared to $\beta_x = 1.76$ for Maximum Likelihood, but the difference is small relative to the standard errors. Results in Table 9.1 are similar to the likelihood and regression calibration results given in Section 8.5, and the differences are easily due to our use here of only 641 out of the 1,615 subjects analyzed in Section 8.5.

9.11 OPEN Data: A Variance Components Model

The OPEN Study was introduced in Section 1.2 and Section 1.5, see Subar, Kipnis, Troiano, et al. (2003) and Kipnis, Midthune, Freedman, et al. (2003)indexsLongitudinal data. Briefly, each participant completed up to two food frequency questionnaires (FFQ) which measured reported Protein intake, and also up to two biomarkers for Protein intake (urinary nitrogen). Letting $\mathbf{Y}$ denote the logarithm of the FFQ, $\mathbf{W}$ the logarithm of the biomarker and $\mathbf{X}$ the logarithm of usual intake, the variance com-

Parameter	ML. est.	Boot. se	Bayes p. mean	Bayes p. std.
β_0	-10.10	2.400	-10.78	2.542
β_x	1.76	0.540	1.91	0.562
β_z	0.38	0.310	0.40	0.302
α_0	4.42	0.019	4.42	0.019
$10 \times \alpha_z$	-0.19	0.210	-0.20	0.217
$10 \times \sigma^2_{x\|z}$	0.47	0.033	0.51	0.032
$10 \times \sigma^2_u$	0.14	0.011	0.16	0.008
λ	0.30	0.031	0.28	0.025

Table 9.1 *Framingham data. The effects of SBP and smoking are given by* β_x *and* β_z*, respectively. The measurement error variance is* σ^2_u*. The mean of long-term SBP given smoking status is linear with intercept* α_0*, slope* α_z *and variance* $\sigma^2_{x|z}$*. Also,* $\lambda = \sigma^2_u/\sigma^2_{x|z}$*. "ML" = maximum likelihood, "se" = standard error, "Boot." = bootstrap, "Bayes" =Bayesian inference based on Gibbs sampling implemented in WinBUGS, "p. mean" = posterior mean, and "p. std" = posterior standard deviation.*

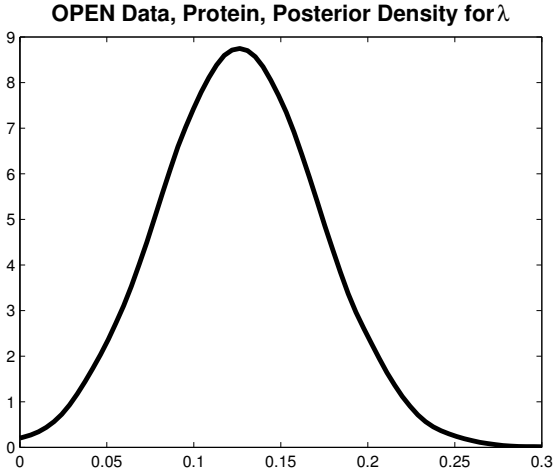

Figure 9.12 *Results of the OPEN Study for Protein intake for females. Plotted is the posterior density of the attenuation* λ*, defined in this case as the slope of the regression of true intake on a single food frequency questionnaire. The posterior mean is* 0.13*, with* 95% *credible interval* $[0.04, 0.21]$*, roughly in line with results reported previously.*

ponents model used is

$$\begin{aligned} \mathbf{Y}_{ij} &= \beta_0 + \beta_x \mathbf{X}_i + r_i + \epsilon_{ij}, \\ \mathbf{W}_{ij} &= X_{ij} + U_{ij}, \end{aligned} \tag{9.33}$$

where $\epsilon_{ij} = \text{Normal}(0, \sigma^2_\epsilon)$, $U_{ij} = \text{Normal}(0, \sigma^2_u)$ and $r_i = \text{Normal}(0, \sigma^2_r)$: the terms r_i is a person-specific bias or equation error, see Section 1.5. In Chapter 11, we note that (9.33) is a linear mixed model with repeated measures. We used a subset of the women in the OPEN study for this analysisindexsLongitudinal data.

The purpose of the OPEN study was to investigate the properties of the FFQ for use in large cohort studies. In regression calibration, Chapter 4, in a cohort study we use the regression of usual intake on the FFQ as the predictor of disease outcome. The slope of this regression is simply

$$\lambda_{\text{regcal}} = \text{cov}(Q, X)/\text{var}(Q).$$

Kipnis, Subar, Midthune, et al. (2003) describe λ_{regcal} as the attenuation factor and note that the regression calibration approximation says that if the true relative risk is R, then the observed relative risk from the use of the FFQ will be $R^{\lambda_{\text{regcal}}}$. For example, a true relative risk of 2 would appear as $2^{\cdot 4} = 1.32$ if the attenuation factor were 0.4 and as $2^{\cdot 2} = 1.15$ if the attenuation factor were 0.2. It is thus of considerable interest to estimate λ_{regcal}. The WinBUGS code along with the prior distributions used is given in Appendix B.8.2.

We plot the posterior density of λ_{regcal} in Figure 9.12. The posterior mean is 0.13, with 95% credible interval $[0.04, 0.21]$, roughly in line with results reported by Kipnis, Subar, Midthune, et al. (2003). This means that a true relative risk of 2 for Protein intake will be attenuated to a relative risk of $2^{0.13} = 1.09$ when using the FFQ. As Kipnis, et al. state: *"Our data clearly document the failure of the FFQ to provide a sufficiently accurate report of absolute protein ... intake to allow detection of their moderate associations with disease.*

Bibliographic Notes

Since the first edition of this book, the literature on Bayesian computation has exploded. The reader is referred to Gelman, Carlin, Stern, & Rubin, Gelman, (2004), Carlin & Louis (2000), and Gilks, Richardson, & Spiegelhalter, (1996) for a thorough introduction. Other important references include two classics, Box & Tiao (1973) and Berger (1985). The latter has an extensive and excellent theoretical treatment. There is also now a statistical package for Bayesian computation, called WinBUGS: we will illustrate the use of WinBUGS in this chapter. The literature

now even includes an excellent book devoted exclusively to the Bayesian approach to measurement error modeling, especially for categorical data, see Gustafson (2004).

Good introductions to MCMC are given by Gelman, Carlin, Stern, & Rubin (2004), Carlin & Louis (2003), and Gilks, Richardson, & Spiegelhalter (1996).

The mechanics of stopping the Gibbs sampler and whether one should use one long sequence or a number of shorter sequences are matters of some controversy and not discussed here; however, we note that Gelman & Rubin (1992) and Geyer (1992) give exactly opposite recommendations. There is a large literature on diagnostics for convergence; see Cowles & Carlin (1996), Polson (1996), Brooks & Gelman (1998), Kass, Carlin, Gelman, & Neal (1998), and Mengersen, Robert, & Guihenneuc-Jouyaux (1999). Kass et al. (1998) is an interesting panel discussion of what is actually done in practice by three Bayesian experts, Carlin, Gelman, and Neal: Kass, though also an expert, is the moderator so we do not learn about his views or experiences. This discussion is quite interesting and well worth reading, unless you are already a Bayesian expert yourself, and probably even in that case. It seems that the experts do not use sophisticated convergence diagnostics, because they feel that these can be misleading. However, they all look at trace plots of various parameters, such as Figure 9.2. Carlin and Gelman monitor $\widehat{R}$ (Gelman & Rubin, 1992), which compares the estimated posterior variance from several chains combined to the average posterior variance from the individual chains. $\widehat{R}$ close to 1 means that the chains have mixed. Carlin and Neal also compute autocorrelations of various parameters; high autocorrelations are a sign of slow mixing. Neal also suggests looking at the log posterior density, which will be neither steadily increasing nor steadily decreasing if the chain has converged.

Alternatives to the Metropolis-Hastings algorithm have been proposed, though they seem less used in practice. For example, Smith & Gelfand (1992) discuss the rejection method and the weighted bootstrap method. Ritter & Tanner (1992) and references therein discuss ways of drawing samples from (9.4), including the griddy Gibbs sampler, which effectively discretizes the components of Ω in a clever way; this can be useful since sampling from a multinomial distribution is trivial.