CHAPTER 4

Adjusting for Mismeasured Continuous Variables

Having discussed at length the deleterious impact of treating mismeasured explanatory variables as if they are precisely measured, we now turn to methods of analysis which recognize and adjust for such mismeasurement. This chapter considers adjustments for the mismeasurement of continuous explanatory variables. Whereas the discussion in Chapters 2 and 3 about the performance of naive estimation in the presence of mismeasurement applies to different methods of statistical inference, we now turn to Bayesian inference implemented with MCMC algorithms as a route to adjust for mismeasurement. Readers with little background in the Bayes-MCMC approach are advised to consult the Appendix, and perhaps some of the references cited therein, before delving into what follows.

4.1 Posterior Distributions

In line with earlier chapters, let Y be the outcome or response variable, let X be the true value of a continuous explanatory variable which is subject to measurement error, let X^* be the noisy measurement of X, and let Z be a vector of other explanatory variables which are precisely measured. If the measurement error mechanism is viewed in terms of the distribution of the surrogate X^* given the true X, it is natural to factor the joint density of the relevant variables as

$$f(x^*, y, x, z) \quad = \quad f(x^*|y, x, z) f(y|x, z) f(x|z) f(z). \tag{4.1}$$

The first term on the right in (4.1), the conditional density of $(X^*|Y, X, Z)$, can be called the *measurement model*. This describes how the surrogate explanatory variable X^* arises from the true explanatory variable X, with Y and Z possibly influencing this process. The second term for $(Y|X, Z)$, which we refer to as the *outcome model* or the *response model*, describes the relationship between the response variable Y and the actual explanatory variables (X, Z). Typically the primary inferential goal is to uncover the form of this relationship. Taken together, the third and fourth terms give the joint distribution of (X, Z), which might be referred to as the *exposure model* in epidemiological applications. As we shall see, however, the third term usually plays a more pivotal role than the fourth term. The terminology of outcome, measurement, and exposure models, or slight variants of this, is fairly common. For instance, Richardson and Gilks (1993a,b) emphasize the trichotomy of disease, measure-

ment, and exposure models, in some of the first work to detail the application of MCMC techniques to Bayesian analysis in measurement error scenarios. Other 'early' work includes Stephens and Dellaportas (1992), Dellaportas and Stephens (1995), and Mallick and Gelfand (1996).

In a parametric modelling framework, specific distributions might be assumed for the components in (4.1), each involving unknown parameters. Then (4.1) guides the construction of a likelihood function, which could be used to make inferences about the unknown parameters if (X^*, Y, X, Z) were observed for each subject. Of course in reality only (X^*, Y, Z) is observed, so that

$$\begin{aligned} f(x^*, y, z) &= \int f(x^*, y, x, z)\, dx \\ &= \left\{ \int f(x^*|y, x, z) f(y|x, z) f(x|z)\, dx \right\} f(z) \end{aligned} \tag{4.2}$$

is required to form the likelihood. In some problems the integral involved is tractable so that the likelihood function is readily evaluated. In other problems, however, the integral will not have a closed-form. A strength of Bayes-MCMC analysis is that (4.2) is not required explicitly. Instead, we can work with (4.1) and let the MCMC algorithm do the required integration implicitly. A non-Bayes route to explicitly working with (4.1) but implicitly working with (4.2) is provided by the EM algorithm (Dempster, Laird and Rubin 1977), which we will comment upon later in this chapter.

As an example, say that the measurement model is both fully known and nondifferential, so that $f(x^*|y, x, z) = f(x^*|x)$ is a known distribution. This might occur if the measurement error mechanism has been investigated thoroughly in previous external studies. We note in passing, however, that the characteristics of mismeasurement processes are not always stable when the measurement scheme is transported from one population to another. For instance it would not be shocking for $Var(X^*|X)$ to differ when the same physical measurement scheme is applied in two populations with dissimilar characteristics. Thus it is not automatic that one could learn the measurement model with data from one population and use this knowledge when adjusting for mismeasurement in a second population. Another situation where one might want to take the measurement model distribution as known is when in fact the mismeasurement process is poorly understood, and one simply wishes to perform *sensitivity analysis* by doing the analysis under a number of different plausible choices for $f(x^*|x)$. Then interest focusses on the extent to which inferences change as the posited distribution changes.

While $f(x^*|x)$ is taken as known, say the response model $f(y|x, z, \theta_R)$ is only known up to a parameter vector θ_R, while the first component of the exposure model $f(x|z, \theta_{E1})$ is known up to a parameter vector θ_{E1}. It will turn out to be superfluous, but also say that the second component of the exposure model $f(z|\theta_{E2})$ is known up to a parameter vector θ_{E2}. Let $\theta = (\theta_R, \theta_{E1}, \theta_{E2})$ comprise all the unknown parameters involved. Bayesian inference requires a prior distribution for these parameters, which we encapsulate with a prior density $f(\theta_R, \theta_{E1}, \theta_{E2})$. Assuming a study with n subjects whose explanatory

variables and outcomes can be regarded as independent of one another, the joint distribution of all the relevant quantities is given as

$$f(x^*, y, x, z, \theta) = \left\{ \prod_{i=1}^{n} f(x_i^*|x_i) f(y_i|x_i, z_i, \theta_R) f(x_i|z_i, \theta_{E1}) f(z_i|\theta_{E2}) \right\} \times f(\theta_R, \theta_{E1}, \theta_{E2}). \quad (4.3)$$

The 'rough-and-ready' form of Bayes theorem states that the density of unobserved quantities U given observed quantities O is *proportional* to the joint density of U and O, where the proportionality is as a function U, for the actually observed O values. Applying this in the present context with $U = (x, \theta_R, \theta_{E1}, \theta_{E2})$ and $O = (x^*, y, z)$ causes us to recast (4.3) slightly as

$$f(x, \theta|x^*, y, z) \propto \left\{ \prod_{i=1}^{n} f(x_i^*|x_i) f(y_i|x_i, z_i, \theta_R) f(x_i|z_i, \theta_{E1}) f(z_i|\theta_{E2}) \right\} \times f(\theta_R, \theta_{E1}, \theta_{E2}). \quad (4.4)$$

If we want the actual normalized posterior density of the unobserved quantities given the observed quantities, then we must compute the integral of (4.4) over the unobserved quantities, given the fixed values of the observed quantities. As alluded to above, computing this integral explicitly may be problematic, or even nearly impossible. However, (4.4) is enough to implement Bayes-MCMC inference. In particular, so long as we can evaluate the right-side of (4.4), we can implement a MCMC algorithm to draw a Monte Carlo sample from the distribution of the unobserved quantities given the observed quantities. Moreover, such a sample for $(x, \theta|x^*, y, z)$ trivially leads to a sample for $(\theta|x^*, y, z)$ upon ignoring the sampled X values. That is, a Monte Carlo sample from the distribution of the unobserved *parameters* given all the observed data is obtained. All inferences about the parameters, and hence about the distributions they describe, stem from this Monte Carlo sample. This is a general strength of Bayes-MCMC inference in problems involving mismeasured, missing, or censored variables. The ideal but unobserved 'complete' data can be treated as unknowns described by the posterior distribution, in order to get at the posterior distribution of the unknown parameters given the observed data.

It is important to note that if θ_{E2} is independent of (θ_R, θ_{E1}) *a priori*, then the same will be true *a posteriori*, as (4.4) can be factored into a term not containing θ_{E2} and a term not containing the other unknowns $(x, \theta_R, \theta_{E1})$. That is,

$$f(x, \theta_R, \theta_{E1}, \theta_{E2}|x^*, y, z) = f(x, \theta_R, \theta_{E1}|x^*, y, z) f(\theta_{E2}|z),$$

where

$$f(x, \theta_R, \theta_{E1}|x^*, y, z) \propto \left\{ \prod_{i=1}^{n} f(x_i^*|x_i) f(y_i|x_i, z_i, \theta_R) f(x_i|z_i, \theta_{E1}) \right\} \times f(\theta_R, \theta_{E1}), \quad (4.5)$$

and

$$f(\theta_{E2}|z) \quad \propto \quad \left\{\prod_{i=1}^{n} f(z_i|\theta_{E2})\right\} f(\theta_{E2}). \tag{4.6}$$

We also note that prior independence of (θ_R, θ_{E1}) and θ_{E2} is commonly assumed, in the absence of any plausible reason why prior views about these different sets of parameters ought to be linked. Indeed, commonly a prior of the form

$$f(\theta_R, \theta_{E1}, \theta_{E2}) \quad = \quad f(\theta_R) f(\theta_{E1}) f(\theta_{E2})$$

is postulated. As a practical matter then, unless there is specific interest in the distribution of the precisely measured covariates Z, one can dispense with a model for them. One can simply work with (4.5) and not make any modelling assumptions about Z. There is no need to specify $f(z|\theta_{E2})$ or apply MCMC to (4.6). A *conditional exposure model* for the unobserved X given the observed Z suffices for making inferences about the parameters of the response model.

Before proceeding to a second example of determining a posterior distribution from disease, measurement, and exposure models, we pause to consider issues of parameter identifiability in such settings. If the measurement model also involves unknown parameters θ_M, then $\theta = (\theta_M, \theta_R, \theta_{E1})$ constitutes all the unknown parameters at play. Following the formal statistical definition of identifiability, the problem is said to be *identifiable* provided that different values of θ cannot correspond to the same distribution of *observable* data. Otherwise, the problem is said to be *nonidentifiable*. Usually in identifiable situations, *consistent* parameter estimation can be achieved, in the sense that Bayesian or maximum likelihood estimates will converge to true parameter values as the sample size increases. On the other hand, in nonidentifiable situations it may be that no amount of data can lead to the true parameter values, if in fact the true distribution of observable data corresponds to more than one set of parameter values.

In many measurement error scenarios one might like to start with quite general and flexible measurement, response, and exposure models. Typically this would lead to identifiability if all of (X^*, Y, X, Z) are to be observed, but result in nonidentifiability in the actual scenario where only (X^*, Y, Z) are observed. Roughly speaking, some additional knowledge or additional data are needed to gain identifiability. In the example above the extra information is complete knowledge of the measurement model (i.e., no unknown parameters are involved), and virtually any reasonable specification of response and conditional exposure models will yield consistent estimators of θ_R and θ_{E1} based on (4.5). It is not always so easy to verify this mathematically, however, as the distribution of the observable data may not have a closed-form expression.

Two other examples of how identifiability might be realized are considered below, namely having actual measurements of X and X^* for some subset of the study sample, or having repeated measurements of X^* for some subjects. Intuitively, if one wishes to adjust for mismeasurement then one must have some knowledge about how the mismeasurement arises in the sense of how X^*

is distributed given X and possibly (Y, Z), or have some means of learning about this distribution from data. Consideration of identifiability in modelling mismeasurement seems crucial, as in many practical situations one would like to adjust for mismeasurement without having a great deal of knowledge about the mechanisms leading to this mismeasurement. If one pushes too far in this direction, nonidentifiability will come into play.

An interesting facet of Bayesian inference is that in one sense it 'works' whether or not one has parameter identifiability. That is, the mechanics of forming a posterior distribution and obtaining parameter estimates from this distribution can be carried out equally well in nonidentifiable situations, modulo some technical concerns about MCMC which are alluded to in Chapter 5. Of course in another sense no form of inference can work in a nonidentifiable model, as estimators which tend to true parameter values do not result. One intuitive way of thinking about Bayesian inference in the absence of parameter identifiability is that the prior distribution plays more of a role than usual in determining the posterior belief about the parameters having seen the data. The question of whether Bayesian inferences based on nonidentifiable models might sometimes be useful has received scant attention in the literature (but see Neath and Samaniego 1997, and Gustafson, Le and Saskin 2001). A close look at this issue with reference to mismeasurement modelling appears in Section 6.3. In essence it seems that in some mismeasurement scenarios Bayesian inference from a nonidentifiable model can be worthwhile, as often the extent of nonidentifiability is modest, while the available prior information is sufficiently good.

As a second example of constructing a posterior distribution in a measurement error scenario, say that the nondifferential measurement model is only known up to a parameter vector θ_M, but for a randomly selected subset of the study subjects X itself is measured, in addition to (X^*, Y, Z). Commonly this is referred to as a *validation* study design. The subjects with (X^*, Y, X, Z) measurements are called the *complete* cases or the *validation substudy*, while subjects with only (X^*, Y, Z) measurements comprise the *reduced* cases. Validation designs are fairly widespread, as for some exposures it is possible to make *gold-standard* measurements, but this process is too expensive to be used for anything more than a small minority of study subjects. An obvious interesting question involving a cost-information tradeoff is what proportion of the subjects should be selected for gold-standard measurements. See Greenland (1988) and Spiegelman (1994) for discussion on this question.

As before, let θ_R parameterize the response model describing $(Y|X, Z)$, and let θ_{E1} parameterize the conditional exposure model describing $(X|Z)$. Let subscripts c and r denote complete and reduced cases respectively, so that $x = (x_1, \ldots, x_n)$ can be partitioned into x_c and x_r. The posterior density for the unobserved quantities given the observed quantities then takes the form

$$f(x_r, \theta_M, \theta_R, \theta_{E1}|x^*, y, x_c, z) \quad \propto \quad \prod_{i=1}^{n} f(x_i^*|x_i, \theta_M) \ \times$$

$$\prod_{i=1}^{n} f(y_i|x_i, z_i, \theta_R) \times$$
$$\prod_{i=1}^{n} f(x_i|z_i, \theta_{E1}) \times$$
$$f(\theta_M, \theta_R, \theta_{E1}). \tag{4.7}$$

Again this is derived from the rough-and-ready version of Bayes theorem which states that the conditional density of all the unobserved quantities given all the observed quantities is proportional to the joint density of the observed and unobserved quantities, viewed as a function of the unobserved quantities with the observed quantities fixed. While the right-hand side of (4.7) does not appear to distinguish between the observed and unobserved components of x, we shall see that MCMC sampling from (4.7) does provide a principled way to make simultaneous inferences about θ_M, θ_R, and θ_{E1}.

As a third scenario, say that no gold standard observations are available, but repeated measurements of X^* are made for at least some study subjects. In particular, say that m_i replicate measurements are made for the i-th subject. If we retain the nondifferential measurement error model parameterized by θ_M, and make the assumption that the replicate measurements of X^* are conditionally independent given the true value of X, then the posterior density of unobserved quantities given observed quantities takes the form

$$\begin{aligned} f(x, \theta_M, \theta_R, \theta_{E1}|x^*, y, z) \quad \propto \quad & \prod_{i=1}^{n} \prod_{j=1}^{m_i} f(x^*_{ij}|x_i, \theta_M) \times \\ & \prod_{i=1}^{n} f(y_i|x_i, z_i, \theta_R) \times \\ & \prod_{i=1}^{n} f(x_i|z_i, \theta_{E1}) \times \\ & f(\theta_M, \theta_R, \theta_{E1}). \end{aligned} \tag{4.8}$$

Here x^*_{ij} is the j-th of the m_i replicated X^* measurements for the i-th subject. It is worth noting that assessment of whether the replicates really are conditionally independent given X is a difficult task, given that X is unobserved. Typically this cannot be verified empirically, and thus devolves to consideration of scientific plausibility in the subject-area context at hand.

4.2 A Simple Scenario

We start with a simple example where the constituent models are based on the normal distribution. In particular, the nondifferential measurement model parameterized by $\theta_M = \tau^2$ is

$$X^*|X, Z, Y \sim N(X, \tau^2),$$

the response model parameterized by $\theta_R = (\beta_0, \beta_1, \beta_2, \sigma^2)$ is

$$Y|X, Z \quad \sim \quad N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2),$$

and the conditional exposure model parameterized by $\theta_{E1} = (\alpha_0, \alpha_1, \lambda^2)$ is

$$X|Z \quad \sim \quad N(\alpha_0 + \alpha_1 Z, \lambda^2).$$

We consider two possible study designs, each involving $n = 250$ subjects. In the first design (X^*, Y, Z) are observed for all subjects, while in addition a gold-standard measurement of X is made for 10 randomly-selected subjects. In the second design, no gold-standard measurements are available, but two replicate measurements of X^* are made for each subject, with the replicates being conditionally independent given the unobserved X.

To proceed we need a prior distribution for all the unknown parameters $(\alpha, \beta, \lambda^2, \sigma^2, \tau^2)$. Without a compelling reason to do otherwise, we assume prior independence of the parameters, so that

$$f(\alpha, \beta, \lambda^2, \sigma^2, \tau^2) \quad = \quad f(\alpha) f(\beta) f(\lambda^2) f(\sigma^2) f(\tau^2).$$

For the sake of convenience we assign improper 'locally uniform' priors to the regression coefficients β in the outcome model and α in the exposure model. That is, $f(\alpha) \propto 1$ and $f(\beta) \propto 1$. Intuitively, these represent 'prior ignorance' about the parameters, as they do not favour any particular value of the coefficients.

We note that improper priors are not without controversy, and some would prefer to represent prior ignorance with proper but very diffuse distributions. Indeed, improper priors can be problematic for variance parameters. Considerations of mathematical invariance suggest that if ν^2 is an unknown variance then the appropriate notion of a 'flat' prior is $f(\nu^2) \propto \nu^{-2}$, or equivalently $f(\nu) \propto \nu^{-1}$, rather than $f(\nu) \propto 1$. Roughly speaking, however, for such a prior to lead to a proper posterior distribution, it is necessary for the data to absolutely rule out the possibility that $\nu = 0$. Otherwise, the singularity in the prior at $\nu = 0$ will be transferred to the posterior distribution, making it improper. This is a fatal flaw, as it leaves no sense of posterior probability on which to base inferences. To be more specific, consider the second of the study designs. The observed data will definitely rule out $\tau^2 = 0$, as the replicate X^* measurements for a given subject will not be identical. However, since X is not observed, the data will not definitely rule out either $\sigma^2 = 0$ or $\lambda^2 = 0$. Thus improper priors $f(\sigma^2) \propto \sigma^{-2}$ and $f(\lambda^2) \propto \lambda^{-2}$ must be avoided. As a final remark along these lines, MCMC algorithms will not necessarily detect an improper posterior distribution. That is, one can proceed to implement MCMC sampling with a prior that produces an improper posterior, and perhaps obtain output that reasonably purports to represent the posterior distribution, even though the posterior distribution does not exist! Indeed, some early published papers on applications of MCMC methods made exactly this mistake. For more on this problem, see Hobert and Casella (1996).

In light of the foregoing discussion we use proper prior distributions for τ^2, σ^2, and λ^2. For the sake of simple MCMC it is convenient to assign Inverse

Gamma distributions as priors on the variances of normal distributions. We assign the $IG(0.5, 0.5)$ distribution as the prior for each of λ^2, σ^2, and τ^2. As discussed in the Appendix, this choice of hyperparameters can be viewed as giving a 'unit-information' prior, with a prior guess of one for each variance.

The chosen priors have the property of being *conditionally conjugate*, as discussed in the Appendix. For (4.7) in the case of the first study design and (4.8) in the case of the second study design, the *full conditional* distribution of each individual unobserved quantity given the other unobserved quantities and the observed data is easily identified as a standard distribution. For the sake of illustration, these distributions are identified in Section 4.11. The important point is that the ability to simulate draws from these distributions means that a simple MCMC algorithm—the Gibbs sampler—can be used to generate a sample from the joint posterior distribution of all unknowns, either (4.7) or (4.8).

We simulate a synthetic dataset under each study design, taking (X, Z) to have a bivariate normal distribution with standard normal marginal distributions and a 0.75 correlation coefficient. This implies $\alpha_0 = 0$, $\alpha_1 = 0.75$, and $\lambda^2 = 0.4375$. We set $\tau = 0.5$ in the measurement model, corresponding to substantial measurement error relative to $SD(X) = 1$. From Result 2.1 we note that this substantial measurement error will have a considerable impact; naive estimates of β_1 in these circumstances will be considerably biased, with an attenuation factor of 0.64. The parameters in the outcome model are set as $(\beta_0, \beta_1, \beta_2) = (0, 0.5, 0.25)$, and $\sigma = 1$. The datasets for the two designs are taken to be extensions of the same underlying (X^*, Y, Z) data set, with gold-standard measurements for some randomly chosen subjects added under the first design, and a second X^* realization for every subject added under the second design. Scatterplots of Y versus X and Y versus the first or only replicate of X^* appear in Figure 4.1. The superimposed least-squares fits (of Y to X and Y to X^*, ignoring Z in each case) show that attenuation is manifested, though by ignoring Z we do not see the considerable extent to which attenuation is actually at play.

For each study design a posterior sample of size 10000 is simulated, using the Gibbs sampler with 1000 burn-in iterations. That is, a total of 11000 iterations are used. As discussed in the Appendix, for successful Bayesian inference via MCMC sampling, one needs the sampler to both *converge* (i.e., burn-in) within a reasonable number of iterations and exhibit reasonable *mixing* in the sense of rapidly moving around the parameter space once convergence is attained. To garner some sense of how well the Gibbs sampler is working for the particular models and datasets at hand, Figure 4.2 gives traceplots of the Gibbs sampler output for some of the unknowns. These plots indicate both very quick convergence (well within the alloted 1000 burn-in iterations) and good mixing. Thus any posterior quantities of interest can be computed accurately from the MCMC output. Some general discussion of what are useful posterior quantities for inference and how to obtain them from MCMC output is given in the Appendix.

Figure 4.3 presents the posterior densities of β_0, β_1, β_2, and τ for the data

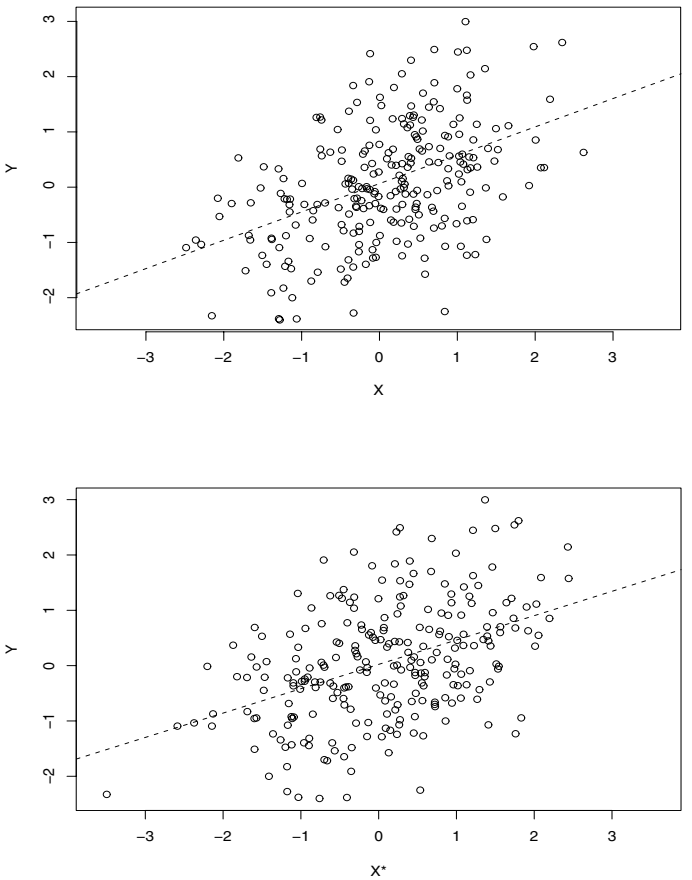

Figure 4.1 *Simulated data in the example of Section 4.2. The top panel is a scatterplot of Y versus the simulated but nominally unobserved X. The bottom panel is a scatterplot of Y versus the first or only X^* replicate. The sample size is $n = 250$. Each plot includes a least-squares fitted-line, to illustrate the attenuation that results from replacing X with X^*.*

sets arising from the two study designs. We also give posterior densities for $(\beta_0, \beta_1, \beta_2)$ under the 'naive' model which assumes the first (or only) X^* replicate is actually a precise measurement of X. Thus we compare inferences which are adjusted for measurement error to those which ignore it. For both study designs we see that the adjustment affects both the location and the width of the posterior densities for β_1 and β_2. As is predictable from the results of Chapter 2, the adjustment shifts the posterior distribution of β_1 away from zero, to correct for attenuation. Also, the effect of adjustment is to

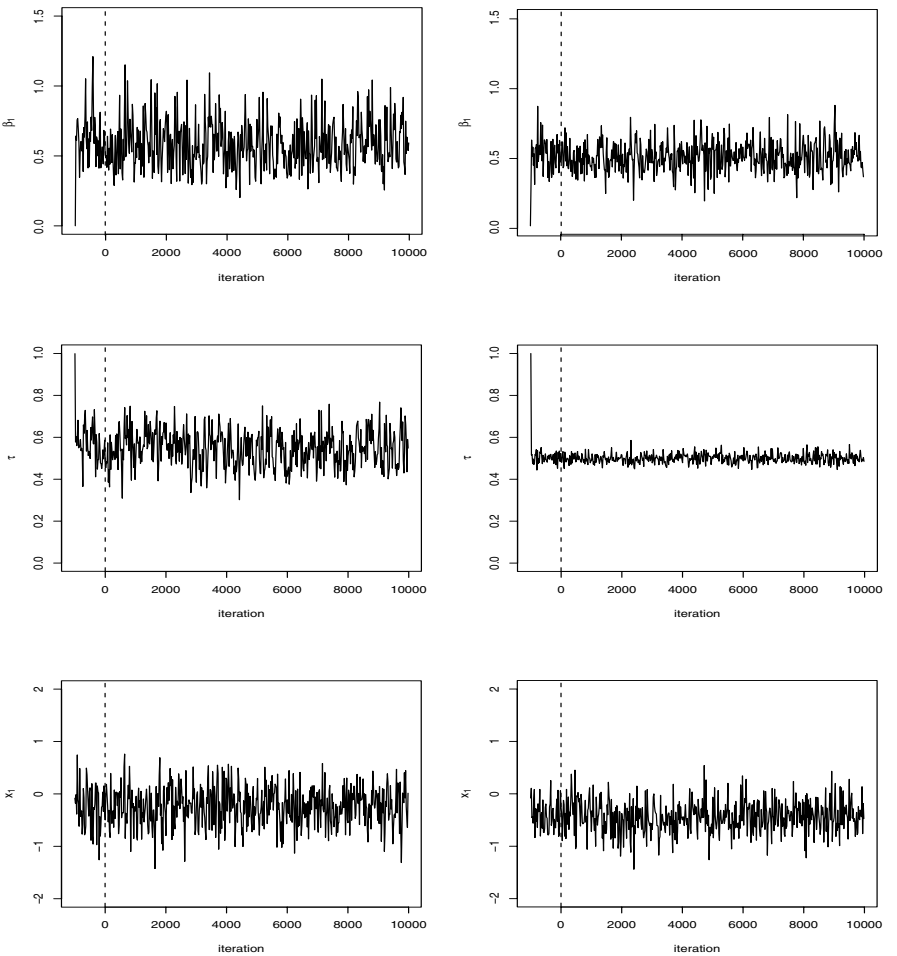

Figure 4.2 *Gibbs sampler output for unknowns β_1, τ, and x_1 in the example of Section 4.2. The plots on the left arise from the first study design, and those on the right from the second study design. The dashed vertical line indicates the end of the allotted burn-in iterations.*

shift the posterior distribution of β_2 toward zero, to compensate for the bias away from zero induced in the coefficient for Z when Z and X are positively correlated. In contrast, the corrected and naive posterior distributions for the intercept β_0 have a common centre. Result 2.1 indicates this is an artifact of X having a distribution which is centred at zero.

Figure 4.3 also shows that after accounting for measurement error the posterior distributions of β_1 and β_2 are wider than their naive counterparts. This

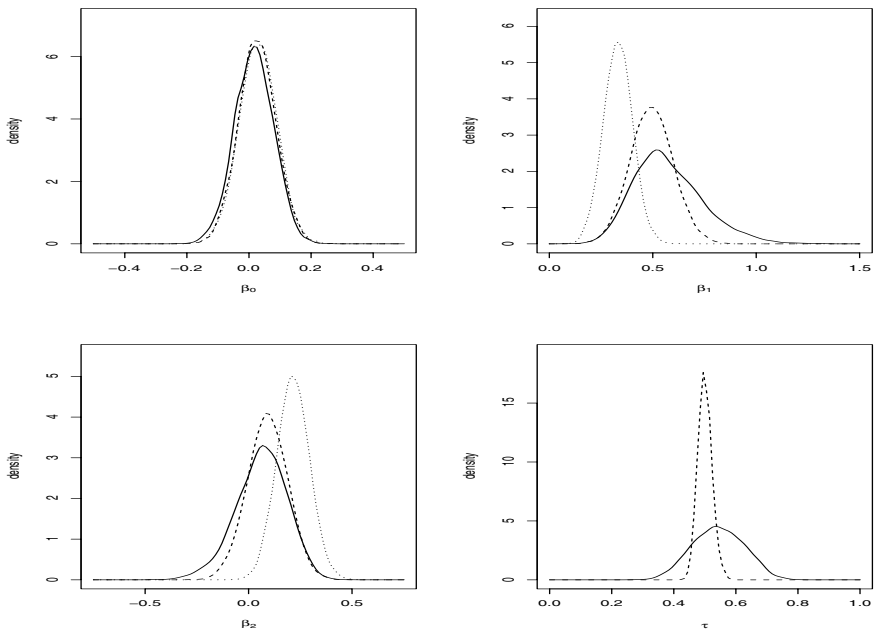

Figure 4.3 *Posterior distributions of β_0, β_1, β_2 and τ in the synthetic data example of Section 4.2. In each case the solid curve gives the posterior density arising from the validation design, the dashed curve gives the posterior density arising from the replication design, and the dotted curve gives the posterior density arising from the naive analysis which treats the first/only replicate of X^* as if it were X.*

extra width correctly reflects the information about X which is lost because only a noisy surrogate X^* is observed. Moreover, we see wider posterior distributions under the validation sample study design than under the replication study design. This arises because the former design, with gold-standard measurements for only 4% of the subjects, contains less information about the measurement model parameter τ than does the replication design. This is directly evidenced by comparing the posterior distribution of τ under the two scenarios. Since the magnitude of the attenuating bias depends on τ, more precise knowledge of τ leads to more precise knowledge of β_1 and β_2. The ability to automatically and appropriately propagate uncertainty from one parameter to another is a strength of the Bayesian approach to inference. An *adhoc* scheme to first determine a point estimate for τ and then use this estimate to derive a bias correction for estimates of β_1 and β_2 would not facilitate such a propagation.

Having carried out Bayes-MCMC analysis for simulated data under these designs, we now point out that in fact this is overkill to some extent. The normal and linear model structures in fact permit explicit determination of

the distribution of the outcome Y given the observed explanatory variables X^* and Z, as well as the distribution of X^* given Z. In particular, the jointly normal distribution of (X^*, Y, X) given Z can be identified, and in turn this yields

$$Y|X^*, Z \sim N\left(\tilde{\beta}_0 + \tilde{\beta}_1 X^* + \tilde{\beta}_2 Z, \tilde{\sigma}^2\right), \tag{4.9}$$

and

$$X^*|Z \sim N\left(\tilde{\alpha}_0 + \tilde{\alpha}_1 Z, \tilde{\lambda}^2\right), \tag{4.10}$$

where

$$\begin{aligned}
\tilde{\beta}_0 &= \beta_0 + \frac{\alpha_0}{1 + \lambda^2/\tau^2}, && (4.11)\\
\tilde{\beta}_1 &= \frac{\beta_1}{1 + \tau^2/\lambda^2}, && (4.12)\\
\tilde{\beta}_2 &= \beta_2 + \frac{\alpha_1\beta_1}{1 + \lambda^2/\tau^2}, && (4.13)\\
\tilde{\sigma}^2 &= \sigma^2 + \left(\frac{\lambda^2\tau^2}{\lambda^2 + \tau^2}\right)\beta_1^2, && (4.14)\\
\tilde{\alpha}_0 &= \alpha_0, && (4.15)\\
\tilde{\alpha}_1 &= \alpha_1, && (4.16)\\
\tilde{\lambda}^2 &= \lambda^2 + \tau^2. && (4.17)
\end{aligned}$$

We call (4.9) and (4.10) the *collapsed model*, as it describes only the observables (X^*, Y, Z) after having marginalized the unobservable X.

Armed with the collapsed model, the likelihood terms for the reduced cases in the validation design can be evaluated directly. Similarly, in the replication design scenario one can easily determine the normal distributions of $Y|X_1^*, X_2^*, Z$ and $X_1^*, X_2^*|Z$. In either design then, the likelihood function for $(\beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1, \tau^2, \sigma^2, \lambda^2)$ can be obtained in closed-form. Hence it is not strictly necessary to use a MCMC algorithm which includes the unobserved values of X as one of the unknown components described by the posterior distribution. The important point, however, is that the MCMC approach is still quite straightforward in problems where a closed-form likelihood function cannot be obtained, as is the situation for the further examples in this chapter.

4.3 Nonlinear Mixed Effects Model: Viral Dynamics

As an example of inferential adjustment for measurement error in the face of a more complex outcome model, we consider a dataset analyzed previously by Wu and Ding (1999), Ko and Davidian (2000) [henceforth KD2000], and Wu (2002) using non-Bayesian techniques. The data describe $n = 44$ HIV-infected patients treated with a combination therapy (ritonavir, 3TC, and zidovudine). Viral load (Plasma HIV-1 RNA) was measured over time for each

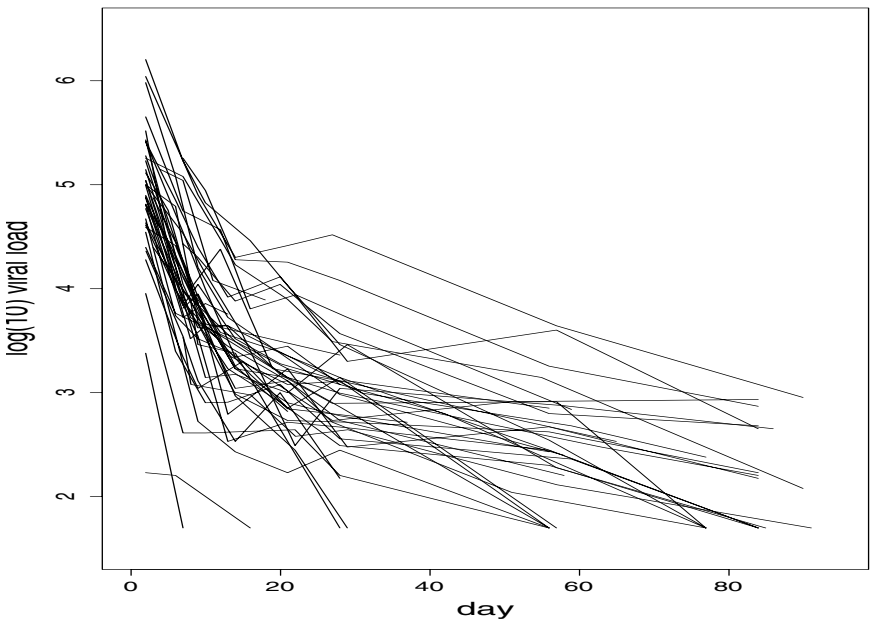

Figure 4.4 *Log_{10} viral load versus day post-treatment for all* 44 *subjects in the viral load example.*

patient. The number of measurements per patient ranges from two to eight, with the majority of subjects having seven or eight measurements. All but eight subjects had their first measurement on the second day post-treatment, with the exceptions having a first measurement sometime between days three and seven. The duration of follow-up ranges from 7 days to 91 days post-treatment, with a median follow-up of 62 days. To get some sense for the data, Figure 4.4 plots $\log_{10}$ viral load versus day post-treatment for all 44 subjects.

The two available covariates which may explain some of the viral dynamics are the baseline viral load and the baseline CD4 cell count. As discussed by KD2000, it is reasonable to view the baseline viral load as a precisely measured explanatory variable, but measured CD4 counts are known to be quite imprecise. There are no gold-standard or replicated measurement of baseline CD4 counts by which to formally gauge the magnitude of the measurement error. Therefore, following KD2000, we simply perform a sensitivity analysis. That is, multiple analyses are performed under different plausible assumptions about the magnitude of the measurement error.

Again following KD2000, the outcome model is

$$\log_{10} Y_{it} \quad \sim \quad N\left(\log_{10} \mu_{it}, \sigma^2\right),$$

where Y_{it} is the viral load for subject i measured on day t. The measurements are considered to be independent across time and across subjects given the underlying means μ_{it}. A typical viral dynamics model gives

$$\mu_{it} \quad = \quad P_{1i} \exp(-\lambda_{1i} t) + P_{2i} \exp(-\lambda_{2i} t),$$

where identifiability is achieved via the constraint $\lambda_{1i} > \lambda_{2i}$. That is, λ_{1i} governs the initial decline in viral load post-treatment, while λ_{2i} affects the longer-term behaviour.

The subject-specific parameters $(P_{1i}, P_{2i}, \lambda_{1i}, \lambda_{2i})$ are expressed in terms of fixed effects $\beta = (\beta_1, \ldots, \beta_7)$ and random effects b_{ki}, $k = 1, \ldots, 4$, $i = 1, \ldots, n$, via

$$\log P_{1i} \quad = \quad \beta_1 + \beta_2 Z_i + b_{1i}, \tag{4.18}$$

$$\log P_{2i} \quad = \quad \beta_3 + \beta_4 Z_i + b_{2i}, \tag{4.19}$$

$$\log \lambda_{1i} \quad = \quad \beta_5 + \beta_6 X_i + b_{3i}, \tag{4.20}$$

$$\log \lambda_{2i} \quad = \quad \beta_7 + b_{4i}. \tag{4.21}$$

Here Z_i and X_i are respectively the baseline $\log_{10}$ viral load and the baseline $\log(CD4/100)$ for the i-th subject. It is self-evident that P_{1i} and P_{2i}, which describe the initial viral load, will be strongly associated with the baseline measurement Z_i via the coefficients β_2 and β_4. The coefficient β_6 describes a hypothesized association between the first-phase decay and the baseline CD4 count.

Still following KD2000, inter-subject variation is modelled by normal distributions for the random effects. That is, all the random effects are considered to be independent of one another, with

$$b_{ki} \quad \sim \quad N(0, \omega_k^2),$$

for $k = 1, \ldots, 4$ and $i = 1, \ldots, n$. In all, the nonlinear mixed effects outcome model is parameterized by $(\beta_1, \ldots, \beta_7, \sigma^2, \omega_1^2, \ldots, \omega_4^2)$. For the purposes of Bayes-MCMC inference, however, we will also include the random effects themselves as unknown quantities described by the posterior distribution.

As mentioned above, the measurement error in Z is thought to be small, and can be ignored. The measurement of X, however, may be subject to substantial error. A nondifferential normal measurement error of the form

$$X^* | Y, X, Z, \quad \sim \quad N(X, \tau^2)$$

is posited. Replicate measurements of X^* are not available, so we simply consider different fixed values of τ in order to conduct a sensitivity analysis. Following KD2000, who in turn were guided by subject-area literature, we use $\tau^2 = 0$, $\tau^2 = 0.12$, $\tau^2 = 0.24$, and $\tau^2 = 0.36$. Since the sample variance of X^* is 0.49, we are entertaining the possibility of very substantial measurement error in X.

Finally, the model specification is completed with an exposure model. Along the lines of the previous example, we take

$$X | Z \quad \sim \quad N(\alpha_0 + \alpha_1 Z, \lambda^2).$$

Of course prior distributions are required for Bayesian analysis. As in the previous example we assign improper locally uniform priors to α and β, but proper priors to the variance components σ^2, ω_1^2, ..., ω_4^2, and λ^2. Both σ^2 and λ^2 are assigned unit-information inverse gamma prior distributions. Such priors are discussed in the Appendix. In the case of σ^2 the prior guess is $\sigma^2 \approx 4$. This is conservatively large, in the sense that $\sigma = 2$ on the $log_{10}Y_{it}$ scale corresponds to large multiplicative variation on the Y_{it} scale itself. In the case of λ^2 the prior guess is taken to be 0.49, which is the sample variance of X^*. Of course strictly speaking one ought not to use the data in order to choose the prior, though again we are using a conservatively large prior guess, as $Var(X|Z) < Var(X) < Var(X^*)$.

We could also assign inverse gamma priors to the random effect variances $\omega_1^2, \ldots, \omega_4^2$. Indeed, this is by far the most commonly used distribution for the variance of normally distributed random effects in Bayesian hierarchical models, as it leads to a simple Gibbs sampling update for the variance component. Arguably, however, the shape of an inverse gamma density is not appropriate for a random effect variance. In particular, the density goes to zero as the variance goes to zero, so that a variance equal to zero is ruled out *a priori*. However, because the random effects are unobserved, the data do not rule out the possibility that their variance is zero. (In contrast, the data do usually rule out the possibility that the variance of observed quantities given parameters is zero. In linear regression, for instance, if the data are not collinear then the population residual variance cannot be zero.) Indeed, one can show that in a simple random effect model with a random effect variance ω^2 being the only unknown parameter, the likelihood for $\omega^2 = 0$ is positive, and for some datasets the likelihood decreases monotonically as ω^2 increases. In such a case an inverse gamma prior will result in a posterior distribution for ω^2 that has an entirely different shape (increasing from zero to a mode, then decreasing) than the likelihood function. This deficiency of inverse gamma priors for variance components is discussed at more length by Gustafson and Wasserman (1996).

In light of this point, Gustafson (1997, 1998) advocates the use of a half-Cauchy prior for either a random effect variance ω^2 or standard deviation ω. The 'half' simply refers to the truncation of a Cauchy distribution centred at zero to positive values, given that a variance cannot be negative. This form lets the shape of the posterior distribution on ω adapt to the data; that is, the posterior density can either increase then decrease or be monotonically decreasing. In the present context we take

$$f(\omega_i) = \left(\frac{2}{\pi}\right) \frac{1}{\bar{\omega}_i} \frac{1}{1 + (\omega_i/\bar{\omega}_i)^2}, \tag{4.22}$$

where $\bar{\omega}_i$ is a hyperparameter which must be specified. A rough calibration is that $\omega_i = \bar{\omega}_i$ is half as likely as $\omega_i = 0$ according to the prior density (4.22). Based on this, setting each $\bar{\omega}_i = 1$ for $i = 1, \ldots, 4$ seems reasonable for the problem at hand. Given the form of (4.18) through (4.21), $\omega_i = 1$ corresponds to very substantial subject-to-subject variation in the viral dynamics, so the

choice of hyperparameters corresponds to quite diffuse prior distributions over *a priori* plausible values of the variance components.

On computational grounds a half-Cauchy prior for the variance of normally distributed random effects initially seems less appealing than an inverse gamma prior. In particular, the full conditional distribution does not correspond to a standard distribution. However, Gustafson (1997, 1998) shows that it 'almost' corresponds to a standard distribution, and gives efficient rejection sampling algorithms which take advantage of this fact. In fact, MCMC updating of a variance component is essentially no more difficult under a half-Cauchy prior than under a inverse gamma prior.

While σ^2, λ^2, ω^2, and α are amenable to Gibbs sampling updates, the remaining unknowns x, b, and β are not. Thus random walk Metropolis-Hastings updates are used for these parameters, though it is extremely tedious to tune the jump size for each parameter in order to obtain a reasonable acceptance rate bounded away from zero and one (see the Appendix for discussion of this point). Overall, the MCMC algorithm mixes quite well for some parameter but less well than others. In the case that $\tau^2 = 0.36$, traceplots from five independent runs of 5000 iterations after 1000 burn-in iterations are given in Figure 4.5. Clearly there is some cause for concern about MCMC error in the computation of posterior quantities, a point we return to shortly.

Table 4.1 reports posterior means and standard deviations of parameters under the $\tau^2 = 0$ and $\tau^2 = 0.36$ scenarios. Given the concern about how well the MCMC algorithm mixes in this example, the table also includes MCMC standard errors reported as a percentage of the computed posterior standard deviation. The MCMC standard error reflects the precision with which a parameter is estimated by the average of estimates from the five independent MCMC runs. It is reported as a percentage of the computed posterior standard deviation, as presumably error arising from the MCMC analysis is only worrisome if it is appreciable relative to the inherent statistical uncertainty about the parameter value. This point is discussed at more length in the Appendix. For most parameters Table 4.1 does indicate that the MCMC error is small relative to statistical uncertainty.

We focus on parameters β_5, β_6, and ω_3^2, as they describe the relationship between the baseline CD4 measurement and the viral dynamics. Table 4.2 gives the posterior mean and standard deviation of each parameter, under all four assumed τ^2 values. We see that the posterior mean of β_6 increases with the assumed value of τ^2. Of course this is to be expected, as a bigger assumed value of τ^2 implies a need for a bigger adjustment to the naive estimate of β_6. Also, the posterior standard deviation of β_6 increases with τ^2, to reflect the larger information loss associated with larger measurement error. Interestingly, the estimated variance of the random effects acting on λ_1 decreases with τ^2.

In comparing the present estimates to those of KD2000, we see that for β_6 both our posterior means and standard deviations tend to be larger than the point estimates and standard errors reported in their paper. Since their approach involves several approximations, we speculate that the present anal-

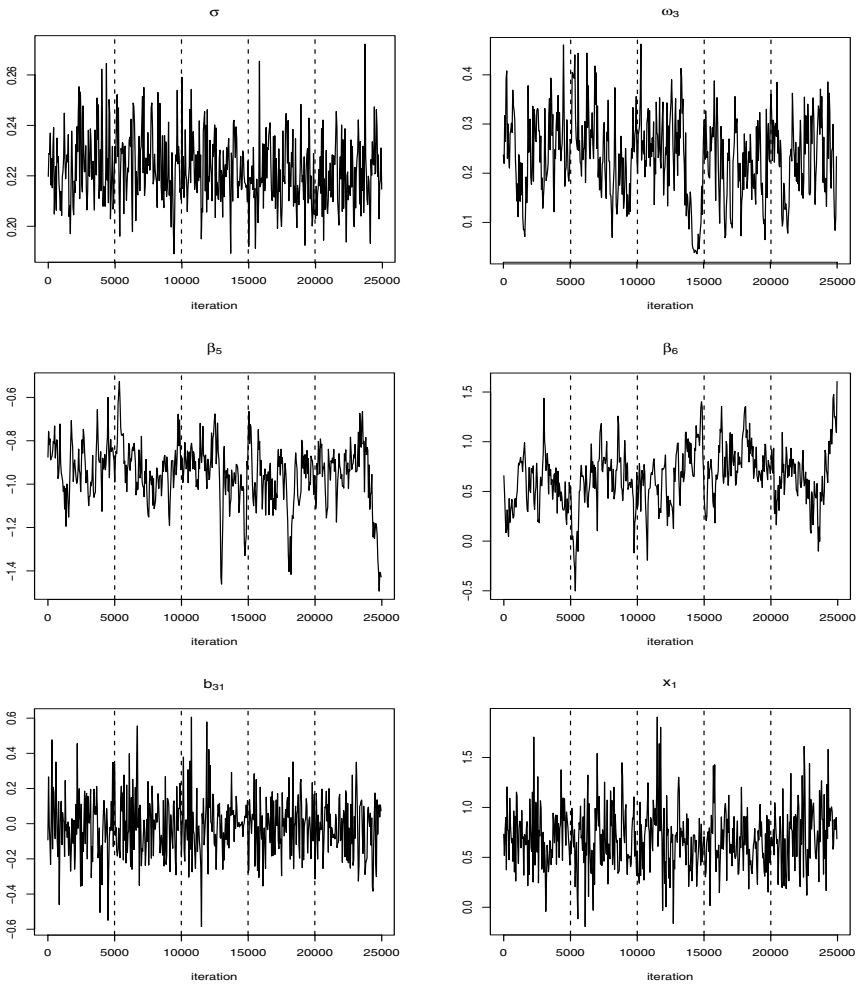

Figure 4.5 *MCMC output for selected quantities in the viral load example of Section 4.3, in the case of* $\tau^2 = 0.36$. *Output for* σ, ω_3, β_5, β_6, b_{31}, *and* x_1 *is given. The dashed horizontal lines separate the output of five independent MCMC runs, each of length 5000. Each run uses* 1000 *burn-in iterations (not displayed).*

	$\tau^2 = 0$			$\tau^2 = 0.36$		
	Posterior		sim.	Posterior		sim.
	mean	SD	error	mean	SD	error
β_1	12.28	0.15	7%	12.36	0.15	6%
β_2	2.10	0.20	3%	1.76	0.28	8%
β_3	8.12	0.16	14%	8.13	0.16	9%
β_4	1.69	0.20	6%	1.62	0.21	6%
β_5	-0.78	0.08	9%	-0.95	0.14	11%
β_6	0.12	0.10	7%	0.64	0.30	14%
β_7	-3.11	0.10	11%	-3.10	0.09	9%
σ^2	0.05	0.01	2%	0.05	0.01	6%
ω_1^2	0.16	0.12	5%	0.14	0.13	8%
ω_2^2	0.50	0.20	2%	0.52	0.21	6%
ω_3^2	0.10	0.04	2%	0.06	0.04	7%
ω_4^2	0.22	0.08	2%	0.21	0.08	6%
α_0				0.41	0.10	3%
α_1				-0.43	0.15	4%
λ^2				0.12	0.06	6%

Table 4.1 *Parameter estimates in the viral load example under the* $\tau^2 = 0$ *(no measurement error) and* $\tau^2 = 0.36$ *scenarios. In each case the posterior mean, posterior standard deviation, and MCMC standard error are given. In particular, the MCMC standard error is reported as a percentage of the posterior standard deviation.*

	$\tau^2 = 0$	$\tau^2 = 0.12$	$\tau^2 = 0.24$	$\tau^2 = 0.36$
$E(\beta_5\|\text{data})$	-0.776	-0.801	-0.893	-0.948
$SD(\beta_5\|\text{data})$	0.078	0.084	0.113	0.141
$E(\beta_6\|\text{data})$	0.122	0.191	0.448	0.642
$SD(\beta_6\|\text{data})$	0.089	0.135	0.236	0.287
$E(\omega_3^2\|\text{data})$	0.096	0.085	0.084	0.058
$SD(\omega_3^2\|\text{data})$	0.038	0.039	0.039	0.038

Table 4.2 *Selected parameter estimates in the viral load example for all four assumed* τ^2 *values. Posterior means and standard deviations for* β_5, β_6, *and* ω_3^2 *are given.*

ysis more fully reflects the various uncertainties at play in arriving at larger uncertainty assessments. This is quite a tentative conclusion, however, given that the MCMC-based estimates are susceptible to simulation error. It would be worthwhile to apply more specialized MCMC algorithms to this problem in an attempt to compute posterior quantities more precisely. Some discussion of more specialized algorithms is given in the Appendix.

4.4 Logistic Regression I: Smoking and Bladder Cancer

The template of measurement model, response model, and exposure model applies equally well to scenarios with a dichotomous response. We illustrate this in the context of a particular medical study, using a subset of the data reported on by Band *et al.* (1999). These data were collected in an occupational oncology research program, with the aim of identifying cancer risk factors. Relevant information, including cigarette smoking history, was obtained using self-administered questionnaires from 15463 male cancer patients aged 20 years or older. These patients were identified via a population-based cancer registry. In addition to analyzing occupational information, Band *et al.* also examined the relationship between smoking and cancer for various tumor sites using the case–control approach. For bladder cancer, the cases are the 1038 patients diagnosed with this disease. The controls are those patients diagnosed with other cancers, except those known to be strongly associated with cigarette smoking as described in Wynder and Hoffman (1982). These exceptions include cancers of the lung, lip, oral cavity and pharynx, esophagus, stomach, pancreas, larynx, and kidney. As a result, 7006 controls were identified. The strategy of choosing patients with other cancers, rather than healthy individuals, as controls, has both drawbacks and advantages. It does limit the generalizability of study results. However, in the context of an exposure measured with considerable error it may limit the potential for differential measurement error than can arise in case-control studies. Since the controls and cases are similar in the sense of having a cancer diagnosis, attitudes towards questions about smoking exposure are less likely to vary greatly between the two groups.

The cigarette pack-year, defined as the number of years of smoking twenty cigarettes a day, is used as the exposure variable. For the purposes of illustration we consider only the data on self-reported current or former smokers, i.e., those who report a positive pack-year value on the questionnaire. In doing so we focus on the association between the extent of exposure and disease amongst current and former smokers. This reduction yields $n_0 = 5348$ controls and $n_1 = 914$ cases. For an analysis which includes the non-smokers, see Gustafson, Le and Vallée (2002). In a related vein, we note that Schmid and Rosner (1993) consider measurement and exposure models under which a self-reported zero exposure may or may not correspond to an actual zero exposure. Their work is amongst the first to employ a Bayes-MCMC approach for dealing with mismeasurement.

We consider the response model

$$\text{logit}Pr(Y = 1|X, Z) \quad = \quad \beta_0 + \beta_1 X + \beta_2 Z, \tag{4.23}$$

where Y is the disease indicator, taking the value zero for controls and one for cases, X is the logarithm of the actual smoking exposure in pack-years, and Z is the patient's age in years. In thinking of such a model we are taking the common approach of analyzing data collected in a *retrospective* case-control study as if they were collected in a *prospective* cohort study. An asymptotic justification for this pretense is given by Prentice and Pyke (1979), although strictly speaking this does not apply to Bayes inference, or to situations involving mismeasured covariates. We will gloss over this point for now, but return to it at length in Section 4.7. In particular, we will introduce a way to check the validity of the pretense in the Bayesian context.

Clearly a self-reported pack-year value will be subject to measurement error. Unfortunately, there is no mechanism by which to rigourously estimate the nature or magnitude of this error. Thus we perform a sensitivity analysis, and draw inferences under different measurement error scenarios. Since the self-reported pack-year value requires a mental assessment of exposure over past years, it is plausible that the magnitude of the measurement error is proportional to the actual exposure. Such multiplicative error on the pack-year scale corresponds to additive error after the logarithmic transform is employed. Letting X^* be the logarithm of the self-reported pack-year value, a simple measurement model would be

$$X^*|Y, X, Z \quad \sim \quad N(X, \tau^2). \tag{4.24}$$

Of course this model assumes the measurement error is nondifferential, unbiased, and normally distributed. More complicated scenarios that relax one or more of these assumptions could be developed; for instance, Gustafson, Le, and Vallée consider scenarios involving a bias, so that $E(X^*|X = x) \neq x$. However, (4.24) is an obvious initial scenario to consider. We try different values of τ in (4.24) to see how inferences vary with the assumed magnitude of the measurement error.

Finally, we require an exposure model. A normal model is convenient. In light of the earlier discussion we require only a model for $X|Z$, rather than (X, Z) jointly. Thus we take

$$X|Z \quad \sim \quad N(\alpha_0 + \alpha_1 Z, \lambda^2).$$

Following the template from Section 4.1, the three component models lead to a posterior distribution of the form

$$\begin{aligned} f(x, \beta, \alpha, \lambda^2|x^*, y, z) \quad \propto \quad & \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\frac{(x_i^* - x_i)^2}{\tau^2}\right) \times \\ & \prod_{i=1}^{n} \frac{\exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2 z_i)\}}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)} \times \end{aligned}$$

τ	β_1	β_2
0	0.253 (0.039)	0.0177 (0.0031)
0.25	0.267 (0.041)	0.0176 (0.0031)
0.50	0.312 (0.049)	0.0166 (0.0031)

Table 4.3 *Inferences in the smoking and bladder cancer example, under different assumed measurement error magnitudes. The estimates of β_1 and β_2 are posterior means, with posterior standard deviations in parentheses.*

$$\left(\frac{1}{\lambda^2}\right)^{n/2} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\frac{(x_i - \alpha_0 - \alpha_1 z_i)^2}{\lambda^2}\right\} \times f(\beta, \alpha, \lambda^2). \tag{4.25}$$

We take locally uniform priors for β and α, but use an inverse gamma prior for λ^2. In particular, we again use a unit-information prior for λ^2, with a conservative prior guess taken to be the sample variance of X^*.

While the components of x and β do not have standard full conditional distributions, MCMC simulation from the posterior distribution given by (4.25) is still relatively straightforward. Some implementation details are given in Section 4.11. Posterior means and standard deviations for the coefficients β_1 and β_2 appear in Table 4.3, assuming $\tau = 0$, $\tau = 0.25$, and $\tau = 0.5$ respectively. Referring back to the original pack-year exposure scale, these values can be regarded as corresponding to no measurement error, 25% multiplicative error, and 50% multiplicative error respectively. Qualitatively, inferences about the exposure coefficient β_1 behave as expected. The posterior mean of β_1 increases with the assumed value of τ, to compensate for a bigger assumed attenuation. Also, the posterior standard deviation of β_1 increases with τ, to reflect less information about the actual exposures. Note, however, that the increase in the estimate of β_1 under $\tau = 0.25$ relative to $\tau = 0$ is quite slight. Moreover, even when $\tau = 0.5$ the posterior mean of β_1 is only one posterior standard deviation larger than the posterior mean when $\tau = 0$. Finally, the posterior mean of β_2 is quite insensitive to the assumed value of τ. In light of the results in Chapter 2, this suggests that X and Z are not strongly correlated. Indeed, the sample correlation between X^* and Z is only 0.16.

Overall, since $\tau = 0$ to $\tau = 0.5$ spans a range from no measurement error to a very substantial multiplicative measurement error in pack-year exposure, we have some confidence that inferences arrived at assuming no measurement error are not badly biased by the measurement error. This is comforting in the face of an exposure such as smoking where this is little hope of ascertaining the precise magnitude of the measurement error via gold-standard measurements, replicated measurements, or other techniques.

4.5 Logistic Regression II: Framingham Heart Study

As a second example of incorporating measurement error with a binary outcome variable, we consider data from an epidemiological cohort study. The Framingham Heart study involves a large cohort followed over a long period of time (Dawber, Kannel and Lyell 1963). We consider the data accrued from the initial exam through the first ten biennial follow-up exams, taking the outcome variable to be twenty-year mortality from any cause. Explanatory variables, as recorded on initial exam, include age, gender, weight, smoking status, serum cholesterol level, diastolic blood pressure, and systolic blood pressure. Given the choice of twenty-year mortality as the outcome variable, we restrict attention to the subset of the cohort aged 55 or younger at the initial exam, yielding a sample of size $n = 4526$.

Several of the explanatory variables are measured with error. We focus on serum cholesterol, for which measurements at both the initial exam and the first follow-up exam are available, at least for the majority of subjects. The 'true' level of serum cholesterol is assumed constant across the initial and first follow-up exams, with the two measurements taken as pure replicates. Also, examination of the measured values suggests that a normality assumption is more appropriate for the logarithm of the measurements. Thus if X represents the logarithm of the true serum cholesterol level (in mg/100ml), while X_1^* and X_2^* are the logarithms of the measured values at the initial and first follow-up exams respectively, we entertain the nondifferential measurement model

$$\begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} |X, Z, Y \quad \sim \quad N\left\{ \begin{pmatrix} X \\ X \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \right\}.$$

Here Y is the response variable (coded as zero/one for alive/dead twenty years after initial exam), while Z is a vector of other explanatory variables. In line with the previous examples, the response and conditional exposure models are taken to be

$$logit\{Pr(Y = 1|X, Z)\} \quad = \quad \beta_0 + \beta_1 X + \beta_2' Z,$$

and

$$X|Z \quad \sim \quad N(\alpha_0 + \alpha_1' Z, \lambda^2).$$

The precisely measured explanatory variables Z include both linear and quadratic terms for age at initial exam, specifically $(AGE - 42)$ and $(AGE - 42)^2$, where 42 is the average initial age in the sub-cohort being considered. Clearly age at initial exam will be a very important predictor in a twenty-year mortality study, so both linear and quadratic terms are included in an attempt to fully account for this explanatory variable. Other components of Z include gender (coded as zero/one for male/female), smoking status at initial exam (coded as zero/one for no/yes), 'metropolitan relative weight' at initial exam (coded as percentage of 'normal'), diastolic blood pressure as measured at initial exam (in mm-Hg), and systolic blood pressure as measured at initial exam (in mm-Hg).

	$\hat{\beta}$	(PSD)	$\exp\left(\hat{\beta}\right)$
AGE_1	0.083	(0.008)	1.09
AGE_2	-0.001	(0.001)	1.00
GENDER	-0.475	(0.091)	0.62
SMOKING	0.756	(0.095)	2.13
LOG-CHOL	0.131	(0.120)	1.14
RELWHT	-0.009	(0.046)	0.99
DBP	0.140	(0.071)	1.15
SBP	0.316	(0.068)	1.37

Table 4.4 *Estimated coefficients in the Framingham analysis. For each coefficient the estimate* $\hat{\beta}$ *is the posterior mean, with the posterior standard deviation (PSD) given in parentheses. The corresponding adjusted odds-ratio* $\exp\left(\hat{\beta}\right)$ *is also given. The* AGE_1 *and* AGE_2 *coefficients are reported as log odds-ratios with respect to* $(age-42)$ *and* $(age-42)^2$*. The* $LOG-CHOL$ *coefficient is reported with respect to a change of* $\log(1.5)$ *in log serum-cholesterol; hence the odds-ratio can be interpreted in relation to a* 50% *increase in the serum-cholesterol level. Each of the RELWHT, DBP, and SBP coefficients are reported with respect to a one standard-deviation change in the explanatory variable.*

The approach to prior construction and MCMC fitting used in the previous example (and discussed in Section 4.11) is easily adapted to the present setting. We again employ locally uniform prior distributions for the response model coefficients β and the conditional exposure model coefficients α, and unit-information inverse gamma priors for τ^2 and λ^2. To be conservative, the prior guesses for both τ^2 and λ^2 are taken to be the sample variance of X_1^*. One twist in the present scenario is that one or both X^* measurements are missing for some subjects. It is easy to see, however, that the conditional distribution of say X_{1i}^* given all other quantities is simply given by the measurement model as $N(X_i, \tau^2)$. Thus it is trivial to extend the MCMC algorithm to also update the missing values of X^*.

Examination of a posterior sample of size 5000 reveals fast convergence and good mixing. This output yields the posterior means and standard deviations for the components of β which appear in Table 4.4. Note that the LOG-CHOL coefficient β_1 is reported with respect to a change of $\log(1.5)$ units on the log-cholesterol scale, so that the corresponding odds-ratio can be interpreted with respect to a 50% increase in cholesterol level. Note also that the posterior mean of τ is 0.10, which is quite large in relation to the sample standard deviations of $SD(X_1^*) = 0.2$ and $SD(X_2^*) = 0.19$. Thus the data suggest substantial measurement error is manifested.

By way of contrast, an analogous analysis which ignores measurement error is also considered. This analysis treats the first cholesterol measurement as exact when it is available, and substitutes the second measurement for the first in the 1686 cases where the first is missing but the second is not. The 107 subjects with both measurements missing are omitted. Again parameterized

relative to an additive change of log(1.5) in log-cholesterol, a posterior mean and standard deviation of 0.120 and 0.093 are obtained for the corresponding logistic regression coefficient. In particular, these values are obtained from direct MCMC fitting of the outcome model only, with observed X values. Relative to this naive analysis then, adjusting for measurement error leads to both a bigger estimated coefficient and a larger assessment of statistical uncertainty about the coefficient. The changes are relatively slight in both cases, however, despite indications of substantial measurement error. The MCMC output also yields the posterior probability that the log-cholesterol coefficient is positive given the data. Specifically, this is computed to be 0.86 in the model which adjusts for measurement error and 0.89 in the model which does not. Thus the adjustment does *not* lead to stronger evidence of a cholesterol effect on mortality, since the modest increase in the estimated coefficient is in fact more than offset by the modest increase in the posterior standard deviation of the coefficient.

4.6 Issues in Specifying the Exposure Model

A common criticism of the parametric approach to measurement error problems is that one runs the risk of specifying a distributional form for the exposure distribution that is far from correct. Of course this can be construed as a criticism of all parametric statistical modelling, though one might argue that the problem is particularly acute for exposure models. In particular, it is hard to check empirically whether a postulated exposure model seems reasonable, simply because the exposure variable itself is not observed. Some advocate trying to avoid the specification of an exposure model all together. We will comment on such methods in Section 4.9. Others advocate the use of very flexible models for exposure distributions. For instance, in a non-Bayesian context Roeder, Carroll and Lindsay (1996) use semiparametric mixture modelling for the exposure distribution. In the Bayesian context Muller and Roeder (1997), Carroll, Roeder and Wasserman (1999), and Richardson, Leblond, Jaussent and Green (2002) use mixture models for the exposure distribution, while Mallick, Hoffman and Carroll (2002) consider a different approach based on the Pólya tree prior (Lavine 1992).

Notwithstanding these approaches, typically the inferential focus is on parameters in the response model, which is one level removed from the exposure model. Intuitively one might imagine that this separation would act as a buffer, so that a considerable misspecification of the exposure distribution might yield only a mild bias in estimators of the response model parameters. Indeed, the emphasis in the literature on misspecification at the level of exposure model seems slightly odd, as one might imagine that misspecification of the measurement model, or of the response model itself, would be a bigger problem in practice. This point is returned to later in Section 6.2.

As an initial investigation of the role of the exposure model, consider the normal model discussed in Section 4.2, in the further restricted scenario that τ^2 is known (say from previous studies) and only a single X^* measurement is

made for each subject. Thus a joint distribution for $(X^*, Y, X|Z)$ is assumed to be of the form

$$\begin{aligned} X^*|Y, X, Z &\sim N\left(X, \tau^2\right), \\ Y|X, Z &\sim N\left(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2\right), \\ X|Z &\sim N\left(\alpha_0 + \alpha_1 Z, \lambda^2\right), \end{aligned}$$

with a total of seven unknown parameters. As discussed in Section 4.2, this implies a collapsed model for $(Y, X^*|Z)$ with the normal forms (4.9) for $(Y|X^*, Z)$ and (4.10) for $(X^*|Z)$. In particular, the one-to-one mapping between the initial parameterization $\theta = (\beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1, \sigma^2, \lambda^2)$ and the parameterization of the collapsed model $\tilde{\theta} = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\sigma}^2, \tilde{\lambda}^2)$ is given by (4.11) through (4.17). We focus on inference about β_1 in the outcome model, which can be expressed in terms of collapsed model parameters as

$$\beta_1 = \left(1 + \frac{\tau^2}{\tilde{\lambda}^2 - \tau^2}\right) \tilde{\beta}_1. \tag{4.26}$$

In studying the large-sample behaviour of estimators we write $AVAR$ to denote the asymptotic variance of an estimator and $ACOV$ to denote the asymptotic covariance between two estimators. That is, for consistent estimators $(\hat{\psi}_1, \hat{\psi}_2)$ of parameters (ψ_1, ψ_2)

$$n^{1/2} \begin{pmatrix} AVAR[\psi_1] & ACOV[\psi_1, \psi_2] \\ ACOV[\psi_1, \psi_2] & AVAR[\psi_2] \end{pmatrix}^{-1/2} \begin{pmatrix} \hat{\psi}_1 - \psi_1 \\ \hat{\psi}_2 - \psi_2 \end{pmatrix} \Rightarrow N_2(0, I),$$

where $N_2(0, I)$ denotes the bivariate standard normal distribution, and the convergence is in distribution, as the sample size n tends to infinity.

We can express the asymptotic variance incurred in estimating β_1 in terms of asymptotic variances and covariances in the collapsed model, using the well-known 'Delta method' based on a Taylor series expansion. Applied to (4.26) it yields

$$\begin{aligned} AVAR[\beta_1] = &\left(1 + \frac{\tau^2}{\tilde{\lambda}^2 - \tau^2}\right)^2 AVAR\left[\tilde{\beta}_1\right] + \\ &\left\{\frac{\tau^4 \tilde{\beta}_1^2}{(\tilde{\lambda}^2 - \tau^2)^4}\right\} AVAR\left[\tilde{\lambda}^2\right] + \\ &\left\{\frac{-2\tilde{\lambda}^2 \tau^2 \tilde{\beta}_1}{(\tilde{\lambda}^2 - \tau^2)^3}\right\} ACOV\left[\tilde{\beta}_1, \tilde{\lambda}^2\right]. \end{aligned} \tag{4.27}$$

Now the readily computed Fisher information matrix for the collapsed model gives

$$\begin{aligned} AVAR\left[\tilde{\beta}_1\right] &= \tilde{\sigma}^2/\tilde{\lambda}^2, \\ AVAR\left[\tilde{\lambda}^2\right] &= 2\tilde{\lambda}^4, \\ ACOV\left[\tilde{\beta}_1, \tilde{\lambda}^2\right] &= 0. \end{aligned}$$

Plugging these expressions into (4.27), and expressing the result in terms of the original parameterization, gives

$$AVAR\left[\beta_1\right] = \frac{\sigma^2}{\lambda^2}\left(1+\frac{1}{1-\rho^2}\Delta+\frac{2\rho^2}{1-\rho^2}\Delta^2\right), \tag{4.28}$$

where $\Delta = \tau^2/\lambda^2 = Var(X^*|X,Z)/Var(X|Z)$ is the measurement error magnitude expressed as a proportion of the conditional variance of X given Z, while $\rho^2 = \{Cor(Y,X|Z)\}^2 = \beta_1^2\lambda^2/(\sigma^2+\beta_1^2\lambda^2)$ indicates the strength of the relationship between X and Y given Z. Note that (4.28) increases with Δ, to reflect the loss of information incurred as the magnitude of the measurement error increases.

Hypothetically, if the parameters of the exposure model were known rather than estimated, then the asymptotic variance of $\hat{\beta}_1$ would be given by only the first of the three terms in (4.27), and consequently by only the first two of the three terms in (4.28). That is, the proportion of $AVAR\left[\hat{\beta}_1\right]$ due to uncertainty about the exposure model depends only on Δ and ρ, and can be expressed as

$$w(\Delta,\rho) = \frac{2\rho^2\Delta^2}{1-\rho^2+\Delta+2\rho^2\Delta^2}. \tag{4.29}$$

This proportion is plotted as a function of Δ in Figure 4.6, for some specific values of ρ. We see that unless the outcome-exposure relationship is very strong (ρ large), or the measurement error is quite large (Δ large), inference about β_1 is *not* substantially improved by knowing the correct parameter values in the exposure model. This provides a sense in which the exposure model is not particularly influential in the estimation of an outcome model parameter.

Given that in many circumstances the exposure model seems not to contribute very much to inferences about the outcome model parameters, one might posit that using an incorrect exposure model will not be terribly damaging. In fact, in situations where the measurement, response, and exposure models are all normal linear models, estimators of the response model parameters are remarkably robust in the face of an actual exposure distribution which differs from the postulated exposure model. Quite general results along these lines are given by Fuller (1987). To shed light on this we present two results which are proved in Section 4.11.

Result 4.1 *Assume that (Y,X,Z) jointly have finite second moments, and say X^* and (Y,Z) are conditionally independent given X, with $E(X^*|X)=X$, and $Var(X^*|X)=\tau^2$, where τ^2 is known. Let l_0 be the large-sample limit of the X coefficient from a linear regression of Y on $(1,X,Z)$, with independent realizations of (Y,X,Z). Now consider fitting the model which assumes (i) $X^*|Y,X,Z \sim N(X,\tau^2)$, (ii) $Y|X,Z \sim N(\beta_0+\beta_1X+\beta_2Z,\sigma^2)$, and (iii) $X|Z \sim N(\alpha_0+\alpha_1Z,\lambda^2)$, to independent realizations of (Y,X^*,Z). Let l_1 be the large-sample limit of a Bayes or maximum likelihood estimator of β_1. This estimator is consistent, in the sense that $l_1=l_0$.*

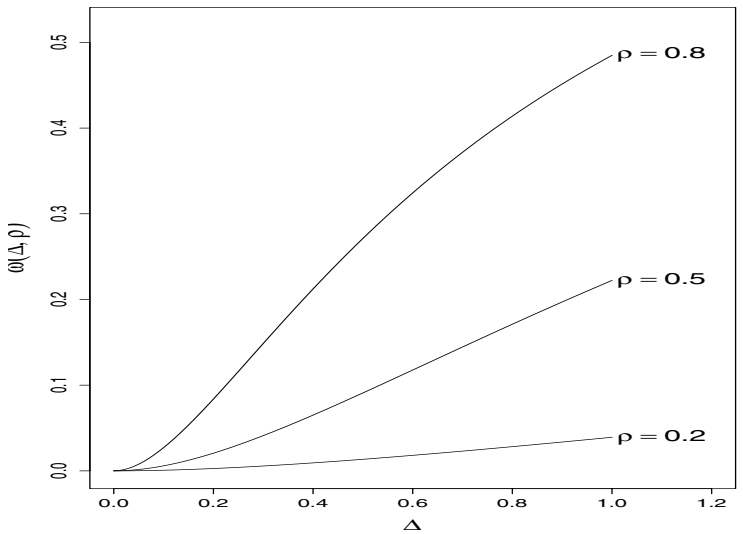

Figure 4.6 *Proportion of* $AVAR[\beta_1]$ *due to uncertainty about the exposure model parameters. Specifically,* $w(\Delta, \rho)$ *defined by (4.29) is plotted as a function of* $\Delta = Var(X^*|X)/Var(X)$*, for selected values of* $\rho = Cor(X, Y|Z)$*.*

We emphasize that here nothing is being assumed about the correctness or incorrectness of the response model, as was also stressed in many of the Chapter 2 results. Also, the measurement model is only assumed to be correct up to first and second moments. Thus the result is somewhat more general than that given by Fuller (1987). The important content of the result is that regardless of the real shape of the exposure distribution, the postulated normal linear exposure distribution will suffice to give the same limiting estimate as would be obtained by applying the response model alone to (Y, X, Z) data. Put succinctly, incorrectly postulating a normal linear exposure model does not induce any asymptotic bias in the estimation of β_1.

While this robustness to exposure model misspecification in the sense of consistency may seem somewhat remarkable in its own right, in fact an even stronger statement obtains. Specifically, the asymptotic variance of the estimator is unaffected by the extent to which the exposure model is misspecified. For simplicity we give this result in the simpler context of no additional precisely measured covariates. Fuller (1987) gives such a result in a more general context.

Result 4.2 *Consider a Bayesian or maximum likelihood estimator* $\hat{\beta}_1$ *of* β_1 *based on fitting three-part model (i)* $X^*|Y, X \sim N(X, \tau^2)$*, (ii)* $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$*, and (iii)* $X \sim N(\alpha_0, \lambda^2)$*, to n independent realizations of* (X^*, Y)*. Here* τ^2 *is known, while* $(\beta_1, \beta_1, \alpha_0, \sigma^2, \lambda^2)$ *are unknown. Now say that in fact (i) and (ii) are correct specifications, so that we can write* β_0*,* β_1*, and* σ^2

unambiguously as 'true' parameter values. Also say that X has an arbitrary distribution with finite variance. Of course Result 4.1 implies that $\hat{\beta}_1$ is a consistent estimator of β_1. But now the asymptotic variance of $\hat{\beta}_1$ is given by

$$AVAR[\beta_1] = \frac{\sigma^2}{Var(X)}\left(1 + \frac{1}{1-\rho^2}\Delta + \frac{2\rho^2}{1-\rho^2}\Delta^2\right), \qquad (4.30)$$

where

$$\begin{aligned}\rho &= Cor(Y, X) \\ &= \frac{\beta_1}{\{\sigma^2 + \beta_1^2 Var(X)\}^{1/2}},\end{aligned}$$

and $\Delta = \tau^2/Var(X)$ is the measurement error variance expressed as a proportion of $Var(X)$. In particular, the asymptotic variance of the estimator depends only on the variance of X, not on the shape of distribution. Thus the asymptotic distribution of the estimator is entirely unaffected by the extent to which the true exposure distribution does or does not match the postulated normal distribution.

We note in passing that (4.30) matches (4.28) in the case that an additional precisely measured covariate is involved. Taken together, Results 4.1 and 4.2 support the use of a normal exposure model regardless of whether this is viewed as plausible in the scenario at hand.

It is folklore in statistical science that phenomena arising in normal linear models tend to be approximately manifested in generalized linear models as well. Thus the extreme insensitivity to exposure distribution misspecification arising with normal response models might be expected to carry over to binary response models. Some recent work, however, speaks against this to some extent. Richardson and Leblond (1997) consider normal linear measurement and exposure models in conjunction with a logistic regression model for the binary outcome Y given imprecisely measured exposure variable X and precisely measured covariate Z. They apply Bayes-MCMC fitting to simulated data, and demonstrate some deterioration in the estimation of response model coefficients as the true exposure distribution deviates from a normal distribution. They summarize their findings as having shown "some sensitivity to misspecification, the overall picture being that of a moderate bias in the estimates and increased posterior standard deviations." We note in passing that Richardson and Leblond consider an exposure model for X alone, while the outcome model is for Y given both X and Z. This is tantamount to assuming that X and Z are independent of one another, so that the requisite conditional exposure model for $X|Z$ can be reduced to a model for X alone. However, they explicitly simulate datasets which involve positive correlation between X and Z. Thus there is a second aspect to the misspecification that is not acknowledged by the authors—both the shape of the postulated X distribution and the postulated independence of X and Z are violated in their simulated scenarios.

In their investigation of mixture modelling for the exposure distribution,

Richardson, Leblond, Jaussent and Green (2002) also make reference to the use of misspecified exposure models. The scenario is that of a nondifferential normal measurement model for $X^*|X$, a logistic regression model for $Y|X$, and a mixture-of-normals model for X. Additional precisely measured covariates are not considered. The exposure model is quite flexible, as the number of components in the mixture, the weight of each component, and the mean and variance of each component are all treated as unknown parameters. Reversible jump MCMC (Green 1995) is used to deal with the varying dimensional parameter space, as exemplified for mixtures by Richardson and Green (1997). On simulated data, Richardson, Leblond, Jaussent and Green compare estimates arising under their flexible exposure model to those arising from a simple normal exposure model. For several scenarios where the true exposure distribution is grossly nonnormal, they report mean-squared errors for outcome model coefficients that are considerably larger under the normal exposure model than under the flexible exposure model. In light of these findings it seems relevant to consider more flexible exposure models.

4.7 More Flexible Exposure Models

Some Bayesian approaches to flexible exposure models in tandem with nonnormal response models have been considered recently in the literature. Much of this work involves mixture models, as in Carroll, Roeder and Wasserman (1999), Muller and Roeder (1997), and Richardson, Leblond, Jaussent and Green (2002). The approach of Roeder, Carroll and Lindsay (1996), while not Bayesian, pursues a similar tack.

Some of the suggested techniques for flexible exposure distribution modelling are quite demanding to implement and execute. Gustafson, Le and Vallée (2002) suggest the arguably simpler approach of modelling the exposure distribution as a discrete distribution on a suitably large set of support points. To illustrate this in a simple context, say there is a single imprecisely measured exposure variable with no additional precisely measured covariates. Let $f_M(x^*|x,y)$ be the completely known measurement model and let $f_R(y|x,\theta_R)$ be the response model. Let $g[1] < g[2] < \ldots < g[m-1] < g[m]$ denote the m fixed support points of the exposure distribution for X. It is convenient to introduce an indexing variable $a_i \in \{1,\ldots,m\}$ for each subject, such that $x_i = g[a_i]$. Then $a = (a_1,\ldots,a_n)$ rather than $x = (x_1,\ldots,x_n)$ can be viewed as describing the unobserved exposures. In particular, an exposure model parameterized by θ_E is specified as $f_E(a|\theta_E)$, giving the joint posterior density of interest as

$$\begin{aligned} f(a,\theta_R,\theta_E|x^*,y) \quad \propto \quad & \prod_{i=1}^{n} f_M(x_i^*|g[a_i],y_i) f_R(y_i|g[a_i],\theta_R) \quad \times \\ & \prod_{i=1}^{n} f_E(a_i|\theta_E) \quad \times \\ & f(\theta_R,\theta_E). \end{aligned} \tag{4.31}$$

In fact the obvious exposure model simply assigns a probability to each gridpoint, which we express as $Pr(a_i = j|\gamma) = \gamma_j$. Put more succinctly, $a_1, \ldots, a_n$ are modelled as independent and identically distributed from the Multinomial$(1, \gamma)$ distribution, where $\gamma = (\gamma_1, \ldots, \gamma_m)$. A simple noninformative prior for γ would be a uniform distribution over the m-dimensional probability simplex, which is equivalently the Dirichlet$(1, \ldots, 1)$ distribution. Since the Dirichlet distribution is a conjugate prior for a Multinomial model, MCMC updating of γ is easily accomplished via Gibbs sampling. More specifically, the joint posterior density (4.31) specializes to

$$\begin{aligned} f(a, \theta_R, \gamma|x^*, y) \quad \propto \quad & \prod_{i=1}^{n} f_M(x_i^*|g[a_i], y_i) f_R(y_i|g[a_i], \theta_R) \quad \times \\ & \prod_{j=1}^{m} \gamma_j^{c_j(a)} \quad \times \\ & f(\theta_R), \end{aligned} \tag{4.32}$$

where $c_j(a) = \sum_{i=1}^{n} I\{a_i = j\}$ is the number of x values currently assigned to the j-th support point. In terms of MCMC updating, whatever strategy might be used to update θ_R under a parametric exposure model can be used here without modification. Moreover, as mentioned above, updating γ is straightforward. The remaining issue is how to update the components of a. A simple approach is to use a discrete version of the random walk MH algorithm.

That is, candidates of the form $a_i^* = a_i + V$ are generated, where V is a symmetric distribution on $\{-k, \ldots, -1, 1, \ldots, k\}$ for a suitably chosen value of k. As a minor point, care must be taken at either end of the grid, perhaps by using 'reflection' to maintain candidate moves for which the reverse moves are equally likely. That is, the precise candidate generation scheme is

$$a_i^* \quad = \quad \begin{cases} 1 + \{(1 - (a_i + V)\} & \text{if } a_i + V < 1, \\ m - \{(a_i + V) - m\} & \text{if } a_i + V > m, \\ a_i + V & \text{otherwise.} \end{cases}$$

Then the MH algorithm acceptance probability is based simply on the ratio of (4.32) at the candidate value to (4.32) at the current value. Since (4.32) factors into separate terms for each a_i, these updates can be carried out efficiently in a parallel fashion.

As noted by Gustafson, Le and Vallée (2002), there are competing desiderata in selecting the fixed support points for the exposure distribution. Say the measurement model is simply $X^*|Y, X \sim N(X, \tau^2)$ where τ^2 is known, as would be the case in a sensitivity analysis, for instance. On the one hand, we want the number of support points m to be small, relative to the sample size n in particular. This is needed if one hopes to estimate the probabilities γ well. On the other hand, the spacing of the grid points cannot be too large in relation to τ if one wants the discrete approximation to the distribution of X given X^* and parameters to be faithful to the underlying 'true' smooth distribution. Viewing the second desideratum as more pressing than the first, we choose an

equally-spaced grid, with $g[1] = min_i\{x_i^*\} - 2.5\tau$, $g[m] = max_i\{x_i^*\} + 2.5\tau$, and a spacing between adjacent gridpoints of $\tau/4$. This spacing can be viewed as giving a 16-point approximation to the main body $(x \pm 2\tau)$ of the conditional distribution for $X^*|X = x$ viewed as a function of x. Of course the method is expected to perform better when the resulting m is small compared to the sample size n. In turn, this suggests better performance when τ is larger. Gustafson, Le and Vallée (2002) suggest using the same choice of gridpoints based on an estimated value of τ in validation designs where X is observed for some subjects and τ is not known in advance. In general, grid-based statistical models are often unattractive because it is unclear how to choose the coarseness of the grid. In measurement error models, however, either complete or rough knowledge of the measurement error magnitude provides guidance for the choice of grid spacing.

Of course most studies do involve additional precisely measured covariates, so the exposure model requires a conditional distribution for $(X|Z)$ rather than a marginal distribution for X. It is a formidable challenge to construct flexible models for a conditional distribution of X given p other covariates Z. Indeed, none of the mixture model based approaches mentioned earlier have been extended to this more general scenario. Similarly, direct adaptation of the grid approach above would require each probability γ_j to be a function of Z. To avoid such a complicated structure, Gustafson, Le and Vallée (2002) resort to assuming that X and Z are independent. Clearly such an assumption is grossly violated in many applications.

A middle ground approach is to use what we term a *reverse-exposure* model, which postulates a joint distribution for (X, Z) in terms of the grid-based model for the marginal distribution of X along with a simple parametric model for Z given X. Of course this is literally a 'reverse' situation, as we have already illustrated that generally if a model for $X|Z$ is postulated then it is not necessary to postulate a model for Z. But particularly in situations where the association between X and Z is not suspected to be strong, the overall effect of using a very flexible model for X and a parametric model for $Z|X$ will be a reasonably flexible model for $X|Z$.

In particular, with a parametric model $f_{RE}(z|x, \theta_{RE})$ for $Z|X$ (RE standing for Reverse Exposure), the posterior density (4.32) generalizes to

$$\begin{aligned} f(a, \theta_R, \theta_{RE}, \gamma|x^*, y) \quad \propto \quad & \prod_{i=1}^{n} f_M(x_i^*|g[a_i], y_i) f_R(y_i|g[a_i], \theta_R) \quad \times \\ & \prod_{j=1}^{m} \gamma_j^{c_j(a)} \quad \times \\ & \prod_{i=1}^{n} f_{RE}(z_i|g[a_i], \theta_{RE}) \quad \times \\ & f(\theta_R, \theta_{RE}). \end{aligned} \tag{4.33}$$

Note that MCMC fitting is only slightly more complicated when using (4.33)

as opposed to (4.32). The updating of θ_R remains unchanged, as does the Dirichlet updating of γ. The discretized random-walk MH algorithm can again be used to update a, with the only extension being the need to include the f_{RE} terms in the calculation of acceptance probabilities. Finally, updating of θ_{RE} would typically be straightforward, based on the observed z values and the current x values implied by the current a values.

We try the grid-based reverse exposure model for the smoking and bladder cancer example of Section 4.4. Recall that this example involves log pack-years as the X variable and age as the Z variable. The normal linear exposure model for $X|Z$ is replaced by the grid model for X and a normal linear model for $Z|X$. Note that the selection of gridpoints outlined above leads to $m = 155$ gridpoints when $\tau = 0.25$, and $m = 88$ gridpoints when $\tau = 0.5$. Given the sample size is $n = 6262$, we do have scenarios where m/n is small, as desired.

The MCMC mixing of the outcome model parameters seems comparable to that observed with the simpler model of Section 4.4. The resulting posterior means of $\hat{\beta}_1 = 0.270$ in the $\tau = 0.25$ scenario and $\hat{\beta}_1 = 0.315$ in the $\tau = 0.5$ scenario are essentially indistinguishable from those given in Table 4.3 for the normal exposure model, particularly in light of the posterior standard deviations reported there. In fact, breaking the present MCMC output of 5000 iterations into ten batches of size 500 yields simulation standard errors of 0.002 when $\tau = 0.25$ and 0.003 when $\tau = 0.5$, so that the present posterior means are within 'numerical tolerance' of those arising from the simpler exposure model. In the present example, then, there is no apparent effect of moving to a more flexible exposure model. Of course it is always comforting when inferences are essentially unchanged upon moving to a more flexible model.

It is interesting to examine the estimated exposure distribution that arises from fitting the grid-based reverse exposure model. In particular, the posterior mean of γ corresponds to the estimated distribution of X displayed in the middle panels of Figure 4.7, for $\tau = 0.25$ and $\tau = 0.5$ respectively. While there is some roughness, as one might expect from a grid-based model, there is also a clear suggestion of nonnormality in the sense of a long left tail. As a check on the plausibility of the estimated X distribution, replicated X^* distributions are created. That is a sample of n replicated X^* values is created by first sampling X values from the estimated discrete distribution of X, and then adding $N(0, \tau^2)$ noise to these values. Histograms of these replicated X^* samples appear in the bottom panels of Figure 4.7, using the gridpoints as midpoints of the histogram bins. These can then be compared to the histograms of the actual X^* values in the top panels of Figure 4.7. Generally the actual and replicated X^* samples are quite similar, suggesting the procedure has done a reasonable job of estimating the X distribution. Of course, even though the estimated X distribution is decidedly nonnormal, it still yields estimates of the response model parameters that are essentially the same as those arising from a normal model for the X distribution.

In general it seems that consideration of when flexible exposure models are needed and how best to achieve such flexibility are ripe areas for further research. In the case of logistic-regression response models, the lack of

applicable asymptotics means the evidence on how well misspecified normal exposure models work in estimating response model parameters is quite limited. As mentioned, Richardson and Leblond (1997) and Richardson, Leblond, Jaussent and Green (2002) do provide some evidence in favour of using more complex models. Conversely, in the example above a more flexible model provides evidence of a nonnormal exposure distribution, yet estimates of the response model parameter are essentially the same as under a normal exposure model. In light of these findings, further investigation is warranted. A second avenue for investigation involves the comparison of Bayesian methods allowing a flexible exposure model to non-Bayesian counterparts. Outwardly it seems that a flexible exposure model is more easily accommodated within the Bayesian paradigm, but careful comparisons have not been drawn.

4.8 Retrospective Analysis

We now return to a point which was glossed over earlier in this chapter. In Section 4.4 we analyzed the retrospectively collected data on smoking and bladder cancer as if it were collected prospectively. This is common practice in biostatistics, based primarily on a justification given by Prentice and Pyke (1979). This work, however, does not apply to Bayesian analysis, nor to situations involving measurement error. This point has received recent attention from Roeder, Carroll and Lindsay (1996), Muller and Roeder (1997), Seaman and Richardson (2001), and Gustafson, Le and Vallée (2002).

One possibility for analyzing retrospectively collected data subject to measurement error is to use simple parametric models for the distribution of explanatory variables given the outcome variable. This approach has received attention from Armstrong, Whittemore and Howe (1989), Buonaccorsi (1990), and also Gustafson, Le and Vallée (2000) in a Bayesian context. There is some concern, however, that these methods will be particularly lacking in robustness to model misspecification.

Gustafson, Le and Vallée (2002) consider Bayesian modelling which formally accounts for the retrospective way in which the data are collected, while taking advantage of the simpler posterior distribution arising under a prospective model. In particular, they present a scheme to represent the posterior distribution arising from retrospective analysis by reweighting a Monte Carlo sample from the posterior distribution based on prospective data collection. For simplicity we review this approach in the context of a single mismeasured predictor without additional precisely measured covariates.

Say that in the population of interest the prospective relationship between outcome Y and predictor X is governed by a logistic regression model

$$\text{logit}Pr(Y = 1|X) \quad = \quad \tilde{\beta}_0 + \beta_1 X.$$

The fundamental link between this prospective relationship for Y given X and the retrospective relationship for X given Y is succinctly expressed as

$$\log \frac{f(x|y = 1)}{f(x|y = 0)} \quad = \quad \text{logit}\{Pr(Y = 1|X = x)\} - \text{logit}\{Pr(Y = 1)\}$$

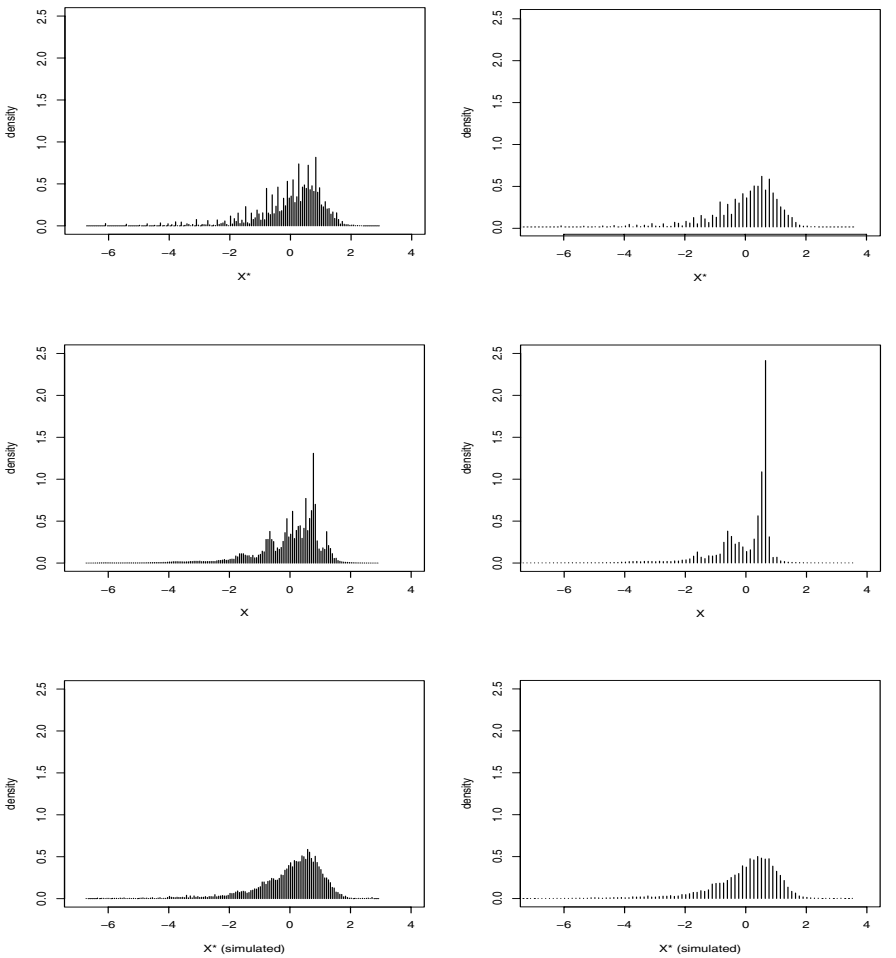

Figure 4.7 *Actual distribution of X^* and estimated distribution of X in the example of Section 4.7. The left panels correspond to assumed measurement error of magnitude $\tau = 0.25$, while the right panels correspond to $\tau = 0.5$. In either case the gridpoint spacing is $\tau/4$. The top panels give histograms of the observed (and centred) X^* with respect to the grid spacing. The middle panels give the estimated distribution of X corresponding to the posterior mean of γ. The bottom panels give histograms of a 'replicated' sample of X^* obtained by adding simulated measurement error to a sample from the estimated distribution of X.*

$$= \quad \beta_0^* + \beta_1 x, \tag{4.34}$$

where

$$\beta_0^* \quad = \quad \tilde{\beta}_0 - \text{logit}\{Pr(Y=1)\}. \tag{4.35}$$

Note that (4.35) underscores the well-known fact that the prospective intercept $\tilde{\beta}_0$ cannot be estimated from retrospective data without knowledge of the outcome prevalence $Pr(Y=1)$.

One approach is to assume $f(x|y=0)$ belongs to some parametric family indexed by ω. Then (β_1, ω) would completely parameterize the problem, as β_0^* would be a deterministic function of (β_1, ω) in order to ensure that

$$f(x|y=1) \quad = \quad \exp(\beta_0^* + \beta_1 x) f(x|y=0)$$

is a normalized density function. However, to mimic the role of an exposure model, it is useful to instead assume that

$$h(x) = r f(x|y=0) + (1-r) f(x|y=1) \tag{4.36}$$

belongs to some parametric family of densities indexed by ω. In particular, we fix $r = n_0/(n_0+n_1)$. Then the case-control sampling scheme can be thought of as first selecting at random $Y=0$ with probability r or $Y=1$ with probability $1-r$, and then sampling X given Y. Under such a sampling scheme (4.36) is the marginal distribution of X, and thereby is akin to the exposure distribution. Thus we henceforth refer to (4.36) as the *pseudo-exposure* distribution.

If we do take $h(x) = h(x|\omega)$ to belong to a parametric family of densities indexed by ω, then the density of X given Y is parameterized by (β_1, ω) according to

$$f(x|y) \quad = \quad \frac{\exp\{y\,(\beta_0^* + \beta_1 x)\}}{r + (1-r)\exp\,(\beta_0^* + \beta_1 x)}\, h(x|\omega).$$

Here β_0^* is a deterministic function of (β_1, ω) given as the solution to

$$E_\omega \left\{ \frac{1}{\beta_0^* + \beta_1 X} \right\} \quad = \quad 1,$$

where the expectation is with respect to X having density $h(\cdot|\omega)$.

Now say that X^* is observed as a surrogate for X, with a completely specified nondifferential measurement model. That is, $X^*|X, Y \equiv X^*|X$ has a known distribution. The retrospective data collection then implies a posterior distribution for the unknown values of X along with the unknown parameters (β_1^*, ω) as

$$\begin{aligned} f(x, \beta_1, \omega | x^*, y) \quad &\propto \quad f(x^*, x | y, \beta_1, \omega) f(\beta_1, \omega) \\ &\propto \quad \left\{ \prod_{i=1}^n f(x_i^*|x_i) f(x_i|y_i, \beta_1, \omega) \right\} f(\beta_1, \omega) \\ &\propto \quad \prod_{i=1}^n f(x_i^*|x_i) \quad \times \end{aligned}$$

$$\prod_{i=1}^{n} \frac{\exp\left[y_i\left\{\beta_0^*(\beta_1,\omega)+\beta_1 x_i\right\}\right]}{r+(1-r)\exp\left\{\beta_0^*(\beta_1,\omega)+\beta_1 x_i\right\}} h(x_i|\omega) \quad \times$$
$$f(\beta_1,\omega).$$

Upon reparameterizing from β_0^* to $\beta_0 = \beta_0^* + log(1-r) - log(r)$, and ignoring multiplicative constants not depending on the unknowns (x, β_1, ω), this can be expressed as

$$\begin{aligned} f(x,\beta_1,\omega|x^*,y) \quad \propto \quad & \prod_{i=1}^{n} f(x_i^*|x_i) \quad \times \\ & \prod_{i=1}^{n} \frac{\exp\left[y_i\left\{\beta_0(\beta_1,\omega)+\beta_1 x_i\right\}\right]}{1+\exp\left\{\beta_0(\beta_1,\omega)+\beta_1 x_i\right\}} h(x_i|\omega) \quad \times \\ & f(\beta_1,\omega). \end{aligned} \tag{4.37}$$

Here the constraint which determines β_0 as a function of (β_1, ω) is now expressed as

$$E_\omega\left\{\frac{1}{\beta_0+\beta_1 X}\right\} \quad = \quad r,$$

which is readily solved numerically.

The important point is that (4.37) looks a lot like the posterior density which would arise in a prospective setting with $h(\cdot|\omega)$ as an exposure model, namely

$$\begin{aligned} f(x,\beta_0,\beta_1,\omega|x^*,y) \quad \propto \quad & \prod_{i=1}^{n} f(x_i^*|x_i) \quad \times \\ & \prod_{i=1}^{n} \frac{\exp\{y_i(\beta_0+\beta_1 x_i)\}}{1+\exp(\beta_0+\beta_1 x_i)} h(x_i|\omega) \quad \times \\ & f(\beta_0,\beta_1,\omega). \end{aligned} \tag{4.38}$$

The difference is that β_0 is also an unknown parameter under (4.38), whereas it is a known function of the unknown parameters under (4.37). To see if this is a meaningful difference in practice we propose drawing a posterior sample of the parameters under (4.38) and then using *importance sampling* to make this sample represent the posterior (4.37). See Evans and Swartz (2000) for general discussion on the use of importance sampling for Bayesian inference. This approach is particularly attractive in the present scenario, as it is much easier to apply MCMC algorithms to (4.38) than to (4.37) given the constraint involved in the latter case.

A technical difficulty in applying importance sampling is that (4.38) is a distribution over $(x, \beta_0, \beta_1, \omega)$ whereas (4.37) is a distribution over (x, β_1, ω) only. To overcome this we artificially extend (4.37) to the larger space by appending the conditional distribution $\beta_0|\beta_1, \omega, x \sim N(a_0 + a_1\beta_1, a_2^2)$ to (4.37) for some fixed (a_0, a_1, a_2). To minimize extra 'noise' in the importance weights we choose a as follows. We generate a preliminary sample from (4.38), and fit

scenario	post. mean	post. SD	rel. change
(i)	0.570 / 0.509	0.229 / 0.210	26.6%
(ii)	0.290 / 0.283	0.088 / 0.086	8.3%
(iii)	0.363 / 0.363	0.048 / 0.048	0.1%

Table 4.5 *Posterior means and standard deviations for β_1 in the example of Section 4.8. In each instance the first entry is based on the posterior arising from the prospective analysis, while the second entry corresponds to the retrospective analysis. Thus the first entry is computed from the MCMC sample while the second entry reflects the importance sampling adjustment. The reported relative change is the absolute difference between the two posterior means expressed as a percentage of the (prospective) posterior standard deviation. Scenarios (i), (ii), and (iii) involve increasing sample sizes, as described in the text.*

the postulated linear model for $\beta_0|\beta_1$ to the sampled values. Then (a_0, a_1) are taken to be the fitted coefficients and a_2 is taken to be the residual standard deviation. The idea here is to make the contrived distribution of $\beta_0|\beta_1$ under (4.37) as close as possible to the actual distribution of $\beta_0|\beta_1$ under (4.38). This limits the amount of extra noise introduced into the importance weights.

We try this procedure on a subset of the smoking and bladder cancer data from Section 4.4. As before, X^* and X are the self-reported and actual smoking exposures on the log pack-years scale, and we consider the measurement model $X^*|X \sim N(X, \tau^2)$ where $\tau = 0.5$. The pseudo-exposure model is simply taken to be normal with $\omega = (\alpha_0, \lambda^2)$ being the mean and variance. In the first instance we create dataset (i) by randomly selecting $n_0 = 100$ of the 5348 controls and $n_1 = 100$ of the 914 cases. Given these data, a posterior sample of size 500 is generated from (4.38) using MCMC as in Section 4.4. In fact this sample is obtained from every tenth of 5000 post burn-in MCMC iterations, in order to obtain a high-quality posterior sample with low serial correlation. Importance weights are determined by taking the ratio of the extended version of (4.37) to (4.38) for each sample point and the normalizing. The upper-left panel of Figure 4.8 plots the importance weights against the values of β_1. There does appear to be a substantial negative correlation between the weights and the values, implying that the posterior mean of β_1 under the retrospective analysis is smaller than under the prospective analysis. Indeed, Table 4.5 indicates the difference is substantial, as the two posterior means differ by about one-quarter of a posterior standard deviation. Thus the importance sampling is providing a nonnegligible adjustment to the prospective posterior distribution, to account for the fact that the data were collected retrospectively.

The same analysis is performed on dataset (ii) comprised of $n_0 = n_1 = 500$ randomly selected controls and cases, and on the full dataset (iii) of $n_0 = 5348$ controls and $n_1 = 914$ cases. Figure 4.8 and Table 4.5 indicate that the importance sampling adjustment becomes less pronounced as the sample sizes increase, to the point of being negligible for the full dataset. This is in keeping

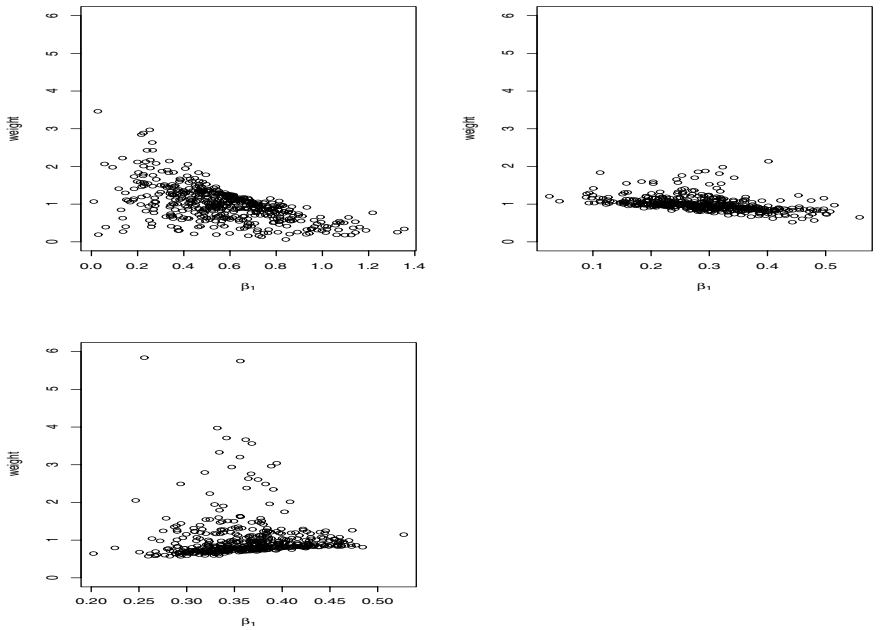

Figure 4.8 *Importance weights in the example of Section 4.8. In each scenario the importance weights for the posterior sample of* 500 *realizations are plotted against the sample values of* β_1*. The weights are normalized to sum to* 500*, so that one is the mean weight. Scenarios (i), (ii), and (iii) involve increasing sample sizes, as described in the text.*

with the aforementioned large-sample justifications for analyzing retrospective data as if it were collected prospectively.

4.9 Comparison with Non-Bayesian Approaches

There is a vast literature on methods to adjust for mismeasurement of continuous explanatory variables, and it is beyond the scope of this book to provide a comprehensive review of this work. Nevertheless, it seems prudent to remark on the relationship between Bayes-MCMC methods as described in this chapter and other methods.

For inference in measurement error scenarios, the nearest cousin to Bayesian estimation is maximum likelihood estimation. Both approaches require a likelihood function arising from fully specified measurement, outcome, and exposure models. In particular, the likelihood function is obtained by integrating over the true but unobserved explanatory variable. As alluded to in Section 4.2, often this integration cannot be done explicitly in closed-form. In such cases, one possibility is to evaluate the likelihood function using a numeri-

cal integration scheme, so that iterative schemes to maximize the likelihood function can be employed. For instance, Crouch and Spiegelman (1990) suggest a particular numerical integration scheme to handle a logistic regression outcome model with normal measurement and exposure models.

As we have seen, for Bayesian inference MCMC techniques provide a general approach to doing the requisite integration 'behind the scenes.' For a likelihood analysis the *expectation-maximization* (EM) algorithm can play a similar role, with an E-step that updates an expected log-likelihood by averaging over the unobserved explanatory variable given the observed quantities and the current parameter values, and a M-step that updates parameter values by maximizing the expected log-likelihood. In general, by iterating back and forth between E and M steps the algorithm converges to parameter values corresponding to at least a local maxima of the likelihood function.

One issue in applying the EM algorithm to measurement error problems is that the expectations to be computed in the E-step may themselves not have closed-form expressions, so that approximations must be used. Schafer (1987) discusses approximations specific to a measurement error model, while more generally there is a considerable literature on Monte Carlo approximations to the E-step. Also, it requires extra effort to obtain standard errors when maximum likelihood estimates are obtained via the EM algorithm (Louis 1982). These limitations likely contribute to the dearth of literature on likelihood-based inference in measurement error problems noted by Carroll, Ruppert and Stefanski (1995). Since Bayes-MCMC inference does not require approximations (or more precisely it can be made as exact as desired by increasing the Monte Carlo sample size), and since uncertainty assessments are obtained without any extra effort, one can argue that the Bayes-MCMC approach is the easiest route to likelihood-based inference in measurement error models.

There are a number of inferential approaches to measurement error problems that are designed to avoid the full specification of measurement, outcome, and exposure models. In particular, much of the literature addresses methods that avoid the explicit construction of an exposure model. Sometimes such approaches are referred to as *functional* methods, whereas *structural* methods do make assumptions about the exposure distribution. Two quite general and popular functional approaches involve *regression calibration* and *simulation extrapolation*, both of which we describe briefly.

While special cases of regression calibration appear earlier (notably Rosner, Willett and Spiegelman 1989 for a logistic regression outcome model), the general idea is discussed by Carroll and Stefanski (1990) and Gleser (1990). Consider outcome variable Y, actual exposure X, mismeasured exposure X^*, and other precisely measured explanatory variables Z. The central idea of regression calibration is quite straightforward. Naively using (X^*, Z) instead of the desired but unobservable (X, Z) as explanatory variables in fitting the outcome model leads to biased inference in general. However, at least approximately unbiased inference can arise if the explanatory variables are taken to be $\{\hat{m}(X^*, Z), Z\}$, where $\hat{m}(X^*, Z)$ is an estimate of $m(X^*, Z) = E(X|X^*, Z)$. That is, a 'best-guess' X value is imputed for each subject, given

the values of X^* and Z. Moreover, in some scenarios $m(\cdot)$ can be estimated without the specification of an exposure model. In particular, with a validation subsample one can directly regress X^* on (X, Z) and use the fitted relationship as $\hat{m}(\cdot)$, which is then applied to impute best-guess X values for the main sample subjects. It is interesting to note, however, that there is no obvious analogue of regression calibration in the ostensibly simpler scenario of a known measurement model but no validation subsample.

It is straightforward to see how regression calibration works in the simple setting of a linear outcome model. Say $E(Y|X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$, and X^* is conditionally independent of (Y, Z) given X. Then

$$\begin{aligned} E(Y|X^*, Z) &= E\{E(Y|X^*, X, Z)|X^*, Z)\} \\ &= E\{\beta_0 + \beta_1 X + \beta_2 Z|X^*, Z) \\ &= \beta_0 + \beta_1 E(X|X^*, Z) + \beta_2 Z, \end{aligned}$$

so that in principle one can estimate $(\beta_0, \beta_1, \beta_2)$ by regression of Y on $m(X^*, Z)$ and Z. Rosner, Willett and Spiegelman (1989) and Rosner, Spiegelman and Willett (1990) demonstrate that the same general approach works reasonably with logistic regression models for binary outcomes, provided the measurement error is relatively modest in magnitude.

A clear difference between regression calibration and Bayes-MCMC inference is that the former is a two-stage procedure: first guessed values for X are imputed using only (X^*, Z), then the imputed values are used as regressors in the outcome model for Y. In contrast, the Bayes analysis involves simultaneous consideration of (X^*, Y, X, Z). In general the two-stage procedure will be somewhat inefficient relative to a Bayes or likelihood procedure, though this does not seem to have been looked at closely in the literature. On the other hand, of course, the lack of need for an exposure model does make the regression calibration approach less susceptible to bias arising from model misspecification.

Another issue with the two-stage nature of regression-calibration involves the propagation of uncertainty. The usual standard errors obtained when fitting the outcome model will be too small, as they will neglect the uncertainty arising because the imputed X values are not the true X values. There is some scope for correcting this via appropriate mathematical expansions, or more generally via bootstrapping (Carroll, Ruppert and Stefanski 1995). In contrast, a nice feature of Bayes-MCMC analysis is that appropriate uncertainty assessments are built-in to the analysis, with no special effort required to extract them. In particular, Bayes-MCMC inference averages over plausible values of X in light of the data, rather than imputing a single best-guess and then proceeding as if this guess is correct.

Another quite general approach to measurement error correction is simulation extrapolation (SIMEX), as first suggested by Cook and Stefanski (1995), and discussed at length by Carroll, Ruppert and Stefanski (1995). The basic idea is very simple and intuitively appealing. Say that for the actual dataset at hand the measurement error variance $Var(X^*|X)$ is estimated or known to

be τ_0^2, and let $\hat{\theta}(\tau_0^2)$ be the naive (i.e., unadjusted for measurement error) estimate of the parameter of interest θ. It is simple to add further measurement error to an already noisy predictor, so one can create a new dataset in which the measurement error variance is $\tau_1^2 > \tau_0^2$ simply by adding $N(0, \tau_1^2 - \tau_0^2)$ noise to the original X^* values. Thus the naive estimate in the new dataset is $\hat{\theta}(\tau_1^2)$. One can continue to add further noise and thereby determine $\hat{\theta}(\tau^2)$ at numerous values of $\tau^2 > \tau_0^2$. In particular, one could consider fitting a smooth line or curve to these points in order to represent $\hat{\theta}(\tau^2)$ as a function of τ^2. By extrapolating this curve to values of τ *smaller* than τ_0^2, one considers estimates of θ under hypothetical reductions in measurement error. In particular, the extrapolated value of $\hat{\theta}(0)$ is taken as an estimate of θ which is fully corrected for measurement error.

Thus SIMEX provides a route to measurement error adjustment which is very easy to understand and implement. In addition, it does not require the specification of an exposure model. On the other hand, it does require the choice of a functional form for $\hat{\theta}$ as a function of τ^2 in the requisite curve-fitting procedure. There is always a danger in extrapolating a function beyond the range where function values are observed, so that the extrapolated value of $\hat{\theta}(0)$ can be sensitive to the choice of functional form. Kuchenhoff and Carroll (1997) provide an example of this, in a relatively complex setting involving segmented regression. Thus the tradeoff is that SIMEX avoids the potential pitfall of bias due to a poorly specified exposure model, but introduces the potential pitfall of bias due to a poor extrapolation. Neither problem is readily diagnosed from the observed data, so it is hard to make a definitive comparison between SIMEX and likelihood-based inference. Of course with correctly specified models likelihood-based inference will necessarily be more efficient than SIMEX estimation. In the specialized context of Kuchenhoff and Carroll (1997), maximum likelihood estimators are found to have considerably smaller variance than both SIMEX and regression calibration estimators.

While regression calibration and SIMEX are relatively general functional methods for measurement error, there are numerous other approaches which are typically aimed at quite specific modelling scenarios. Carroll, Ruppert and Stefanski (1995) discuss a number of such methods. Historically structural approaches involving an exposure model have not been as popular, though they have received increased attention recently, perhaps in large part due to the advent of MCMC techniques. While the potential for exposure model misspecification is an issue, arguably it has been exaggerated as a rationale for using functional approaches. In addition to the transparent straightforwardness of working with fully specified (i.e., structural) probability models, there is always the possibility of using more flexible exposure models, as discussed earlier in the chapter. It also seems strange that exposure model misspecification receives so much attention when both structural and functional approaches are prone to misspecification of the measurement and outcome models. We return to this point at greater length in Section 6.2.

4.10 Summary

We hope this chapter illustrates that the Bayes-MCMC machinery provides a general and powerful way to adjust inferences to account for measurement error in explanatory variables. In particular, the recipe is quite clear: make the best specifications of measurement, outcome, and exposure models one can, along with appropriate prior distributions for the unknown parameters in these models. Then apply MCMC to sample from the posterior distribution of unknown parameters and unobserved variables given the observed variables. Inferential summaries obtained from such a posterior sample will be adjusted in a principled way to account for the measurement error.

As with most statistical techniques, of course, the devil can be in the details. On the technical side, it is not always easy to choose or devise a particular MCMC algorithm that works well on the problem at hand. On the modelling side, sometimes the need for three fully specified models and priors may seem arduous, particularly in relation to some non-Bayesian techniques which seem to entail fewer assumptions. On balance the author contends that the advantages of the Bayes-MCMC approach more than outweigh the disadvantages, though admittedly there is scope for further research on the comparative utility of various methods.

4.11 Mathematical Details

4.11.1 Full Conditional Distributions in the Example of Section 4.2

Consider the first study design in the example of Section 4.2. With the assumed model and prior distributions, the posterior distribution (4.7) for the unobserved quantities $U = (x_R, \beta, \alpha, \tau^2, \sigma^2, \lambda^2)$ given the observed quantities $O = (x^*, x_C, y, z)$ becomes

$$\begin{aligned} f(U|O) \quad \propto \quad & \left(\frac{1}{\tau^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i^* - x_i)^2}{2\tau^2}\right) \times \\ & \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{2\sigma^2}\right) \times \\ & \left(\frac{1}{\lambda^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \alpha_0 - \alpha_1 z_i)^2}{2\lambda^2}\right) \times \\ & \left(\frac{1}{\tau^2}\right)^{(0.5+1)} \exp\left(-\frac{(0.5)}{\tau^2}\right) \times \\ & \left(\frac{1}{\sigma^2}\right)^{(0.5+1)} \exp\left(-\frac{(0.5)}{\sigma^2}\right) \times \\ & \left(\frac{1}{\lambda^2}\right)^{(0.5+1)} \exp\left(-\frac{(0.5)}{\lambda^2}\right). \end{aligned}$$

By viewing this expression as a function of one individual unobserved quantity at a time we can easily 'read off' the full conditional distributions. We use a

complement notation, so that generically ν^C would denote all the unobserved quantities other than ν along with all the observed quantities. For instance,

$$\alpha|\alpha^C \sim N_2\left\{(A'A)^{-1}A'x, \lambda^2(A'A)^{-1}\right\},$$

where A is the $n \times 2$ design matrix with i-th row $(1, z_i)$. Similarly,

$$\beta|\beta^C \sim N\left\{(B'B)^{-1}B'y, \sigma^2(B'B)^{-1}\right\},$$

where B is the $n \times 3$ design matrix with i-th row $(1, x_i, z_i)$. Note in particular that B depends on x_R. For the variance components we have,

$$\lambda^2|\lambda^{2\,C} \sim IG\left\{(n+1)/2, (\|x - A\alpha\|^2 + 1)/2\right\},$$

$$\sigma^2|\sigma^{2\,C} \sim IG\left\{(n+1)/2, (\|y - A\beta\|^2 + 1)/2\right\},$$

and

$$\tau^2|\tau^{2\,C} \sim IG\left\{(n+1)/2, (\|x^* - x\|^2 + 1)/2\right\}.$$

Finally, we also need full conditional distributions for the components of x_R. If i indexes a reduced case then viewing the posterior density as a function of x_i alone gives the conditional distribution of $x_i|x_i^C$ as normal, with

$$E\left(x_i|x_i^C\right) = \frac{(1/\tau^2)x_i^* + (\beta_1^2/\sigma^2)\{(y - \beta_0 - \beta_2 z_i)/\beta_1\}}{(1/\tau^2) + (\beta_1^2/\sigma^2)}$$

and

$$Var\left(x_i|x_i^C\right) = \frac{1}{(1/\tau^2) + (\beta_1^2/\sigma^2)}.$$

Thus to implement the Gibbs sampler in this example we simply need to sample repeatedly from these full conditional distributions. Thus standard routines to generate realizations from normal and inverse gamma distributions are all that is required.

In the case of the second study design, the full conditionals for α, β, σ^2, and λ^2 are exactly as above. But now

$$\tau^2|\tau^{2\,C} \sim IG\left\{0.5 + n, 0.5 + \|x_1^* - x\|^2 + \|x_2^* - x\|^2\right\},$$

and for reduced cases $x_i|x_i^C$ is normal with

$$E\left(x_i|x_i^C\right) = \frac{(2/\tau^2)\bar{x}_i^* + (\beta_1^2/\sigma^2)\{(y - \beta_0 - \beta_2' z_i)/\beta_1\}}{(2/\tau^2) + (\beta_1^2/\sigma^2)},$$

and

$$Var\left(x_i|x_i^C\right) = \frac{1}{(2/\tau^2) + (\beta_1^2/\sigma^2)},$$

where $\bar{x}_i^* = (x_{1i}^* + x_{2i}^*)/2$ is simply the average of the two X^* measurements for the i-th subject.

4.11.2 MCMC for the Models of Sections 4.4 and 4.5

In applying MCMC to logistic regression models we must typically update a vector of coefficients β having a full conditional density of the form

$$f(\beta|\beta^C) \quad \propto \quad \prod_{i=1}^{n} \frac{\exp\{y_i(\beta_0 + \beta_1 x_i + \beta_2' z_i)\}}{1 + \exp\{\beta_0 + \beta_1 x_i + \beta_2' z_i\}} f(\beta), \tag{4.39}$$

where we have followed the notation of (4.23). We elucidate both a good but expensive update and a poor but cheap update. If we are willing to use an iterative numerical routine to maximize the logistic regression log-likelihood, then we can sample a candidate β vector from the normal distribution with mean and variance chosen to match the log-likelihood's mode and curvature at the mode. Note that by doing so we are generating a candidate value that does not depend on the current value, which might be referred to as an *independence MH algorithm*. Standard asymptotic theory dictates that this normal distribution will typically be an excellent approximation to (4.39), as long as the sample size is moderately large. In turn this implies a very high Metropolis-Hastings acceptance rate for the update. The combination of a high acceptance rate and a candidate-generating distribution which does not depend on the current value of β makes for a very efficient MCMC algorithm. However, maximizing the likelihood will take considerable CPU time, and since x is updated in between the updates to β, it is necessary to re-maximize every time β is to be updated. Thus this is a good but slow update.

A much cheaper update not involving maximization is the random-walk MH algorithm applied to one component of β at a time. While such an update is fast, it has several drawbacks. First, it is necessary to set a tuning parameter (the jump size) for each component of β. Second, this algorithm may mix quite slowly. There is some scope for improving the mixing by centering the explanatory variables (i.e., forcing them to have mean zero), as discussed by Gilks and Roberts (1996) and Roberts and Sahu (1997), although we have not found this to be a panacea. As well, centering is slightly less straightforward in the present scenario where x changes from update to update of β.

In implementing updates to logistic regression parameters we tend to compromise and use both slow and fast updates. In particular, we preserve the desired stationary distribution by randomly choosing the update method immediately prior to each update. That is, with some small probability a slow update is performed, but otherwise a fast update is carried out. Empirically this seems to be a decent overall strategy which achieves both reasonable mixing and reasonable execution time.

4.11.3 Proof of Result 4.1

Upon expressing β_1 in terms of the collapsed model parameterization we have

$$l_1 \quad = \quad lsl(\tilde{\beta}_1)\left(1 + \frac{\tau^2}{lsl(\tilde{\lambda}^2) - \tau^2}\right), \tag{4.40}$$

where $lsl(\tilde{\beta}_1)$ and $lsl(\lambda^2)$ are the large-sample limits of Bayes or maximum likelihood estimators of $\tilde{\beta}_1$ and $\tilde{\lambda}^2$ respectively. Since $lsl(\tilde{\beta}_1)$ will be the large-sample limit of the X^* coefficient in regressing Y on $(1, X^*, Z)$, Result 2.1 gives immediately

$$lsl(\tilde{\beta}_1) = \frac{l_0}{1 + \tau^2 / \{Var(X)(1 - Cor(X,Z)^2)\}}. \quad (4.41)$$

On the other hand, standard large-sample theory gives the limit of the estimated residual variance for $X^*|Z$ under possible model misspecification as

$$\begin{aligned} lsl(\tilde{\lambda}^2) &= Var(X^*)\left\{1 - Cor(X^*, Z)^2\right\} \\ &= Var(X)\left\{1 - Cor(X,Z)^2\right\} + \tau^2, \end{aligned} \quad (4.42)$$

where the second equality follows from conditional independence of X^* and (Y, Z) given X, along with $E(X^*|X) = X$ and $Var(X^*|X) = \tau^2$. Plugging (4.41) and (4.42) into (4.40) gives $l_1 = l_0$, as desired.

4.11.4 Proof of Result 4.2

Here we use an alternate notation for the parameters in the collapsed model, for the sake of more compact mathematical expressions. In particular, the collapsed model corresponding to this full model can be parameterized by $\theta = (\gamma_0, \gamma_1, \nu, \phi_0, \omega)$, with $Y|X^* \sim N(\gamma_0 + \gamma_1 X^*, \nu)$ and $X^* \sim N(\phi_0, \omega)$. Moreover, β_1 in the full model can be estimated as $\hat{\beta}_1 = g\left(\hat{\theta}\right)$, where

$$g(\theta) = \gamma_1 \left(1 + \frac{\tau^2}{\omega - \tau^2}\right).$$

Straightforward calculation gives the score vector for the collapsed model as

$$s(\theta; X^*, Y) = \begin{pmatrix} \nu^{-1}(Y - \gamma_0 - \gamma_1 X^*) \\ \nu^{-1}(Y - \gamma_0 - \gamma_1 X^*)X^* \\ -(0.5)\nu^{-1} + (0.5)\nu^{-2}(Y - \gamma_0 - \gamma_1 X^*)^2 \\ \omega^{-1}(X^* - \phi_0) \\ -(0.5)\omega^{-1} + (0.5)\omega^{-2}(X^* - \phi_0)^2 \end{pmatrix}. \quad (4.43)$$

For the time being assume that X has been scaled so that $E(X) = 0$ and $Var(X) = 1$. The value $\theta = \theta^*$ solving $E\{s(\theta; X^*, Y)\} = 0$ is given by

$$\begin{aligned} \gamma_0 &= \beta_0 \\ \gamma_1 &= \beta_1/(1 + \tau^2) \\ \nu &= \sigma^2 + \beta_1^2 \tau^2/(1 + \tau^2) \\ \phi_0 &= 0 \\ \omega &= 1 + \tau^2, \end{aligned}$$

which does not depend on the actual distribution of X (above and beyond the fixed first and second moments). Standard large-sample theory for misspecified models (e.g., White 1982) implies that Bayesian and maximum likelihood

estimators of θ converge to θ^*. Moreover, the asymptotic variance of such an estimator is given as $A^{-1}BA^{-1}$, where

$$A = -E\{s'(\theta^*; X^*, Y)\}$$

and

$$B = E\left[\{s(\theta^*; X^*, Y)\}\{s(\theta^*; X^*, Y)\}'\right].$$

Under the assumed scaling of X it is simple to compute

$$A = \begin{pmatrix} \nu^{-1} & 0 & 0 & 0 & 0 \\ & \nu^{-1}(1+\tau^2) & 0 & 0 & 0 \\ & & (0.5)\nu^{-2} & 0 & 0 \\ & & & (1+\tau^2)^{-1} & 0 \\ & & & & 0.5(1+\tau^2)^{-2} \end{pmatrix}.$$

Since A is diagonal and $g(\theta)$ depends only on (γ_1, ω), the only elements of B that must be computed are B_{22}, B_{55}, and B_{25}. Towards this end, note that we can write

$$Y - \gamma_0 - \gamma_1 X^* = \beta_1 \left(\frac{\tau^2}{1+\tau^2}\right) X + \epsilon_1$$

and

$$X^* = X + \epsilon_2,$$

where

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 + \gamma_1^2\tau^2 & -\gamma_1\tau^2 \\ -\gamma_1\tau^2 & \tau^2 \end{bmatrix}\right)$$

independently of X. Starting from (4.43) this leads to

$$\begin{aligned} B_{22} &= A_{22} + \frac{(\beta_1 - \gamma_1)^2}{\nu^2}\delta \\ B_{55} &= A_{55} + \frac{1}{4(1+\tau^2)^4}\delta \\ B_{25} &= A_{25} + \frac{\beta_1 - \gamma_1}{2\nu\omega^2}\delta, \end{aligned}$$

where

$$\delta = E\left(X^4\right) - 3.$$

Bearing in mind the standardization, if X has a normal distribution then $\delta = 0$. Of course theory dictates that $B = A$ in this case of a correctly specified model, as witnessed by the above equations.

Now from the Delta method we have the asymptotic variance of $\hat{\beta}_1 = g(\hat{\theta})$ as

$$\begin{aligned} Avar\left[g(\theta)\right] = &\left(1+\tau^2\right)^2 (A^{-1}BA^{-1})_{22} + \\ &\gamma_1^2\tau^4 (A^{-1}BA^{-1})_{55} + \\ &-2\gamma_1(1+\tau^2)\tau^2 (A^{-1}BA^{-1})_{25}. \end{aligned}$$

This specializes to

$$\begin{aligned} Avar\left[g(\theta)\right] &= (1+\tau^2)^2\left\{\frac{\sigma^2}{1+\tau^2}+\frac{\beta_1^2\tau^2}{(1+\tau^2)^2}+\frac{\beta_1^2\tau^4}{(1+\tau^2)^4}\delta\right\}+ \\ &\quad \frac{\beta_1^2\tau^4}{(1+\tau^2)^2}\left\{2(1+\tau^2)^2+\delta\right\}+ \\ &\quad (-2)\beta_1\tau^2\left\{\frac{\beta_1\tau^2}{(1+\tau^2)^2}\delta\right\}. \end{aligned}$$

Clearly the terms involving δ cancel, so that the asymptotic variance of the estimator does not depend on the actual distribution of the standardized X. In particular,

$$Avar\left[g(\theta)\right] = (1+\tau^2)\sigma^2+\beta_1^2\tau^2+2\beta_1^2\tau^4. \tag{4.44}$$

Finally, it remains to generalize this expression to the case where X is not standardized. From invariance considerations it is straightforward to determine that the more general expression can be determined from (4.44) by first substituting $\beta_1 Var(X)$ for β_1 and $\tau^2/Var(X)$ for τ^2, and then dividing the entire expression by $Var(X)$. This results in (4.30), as desired.

It is interesting to note that misspecification does affect the asymptotic variance of estimators of γ_1 and ω, as these variances do depend on δ. However misspecification also induces dependence between the two estimators, so that when they are combined to estimate β_1 there is a seemingly remarkable cancellation of the terms involving δ.