22

234

Measurement error

Sylvia Richardson

22.1 Introduction

Errors in the measurement of explanatory variables is a common problem in statistical analysis. It is well known that ignoring these errors can seriously mislead the quantification of the link between explanatory and response variables, and many methods have been proposed for countering this. Measurement error was initially investigated in linear regression models (see for example Fuller, 1987), and recent research has led to its investigation in other regression models, motivated in particular by applications in epidemiology and other areas of biomedical research. Overviews of measurement error in epidemiology can be found in Caroll (1989), Armstrong (1990), Gail (1991), Liu and Liang (1991), Thomas *et al.* (1993) and Caroll *et al.* (1995). We too shall focus much of our discussion in the epidemiological context.

In epidemiological studies, it is rarely possible to measure all relevant covariates accurately. Moreover, recent work has shown that measurement error in one covariate can bias the association between other covariates and the response variable, even if those other covariates are measured without error (Greenland, 1980; Brenner, 1993). Apart from biasing the estimates, misspecification of explanatory variables also leads to loss of efficiency in tests of association between explanatory and response variables; this has recently been characterized in logistic regression models by Begg and Lagakos (1993).

Any method proposed for correcting parameter estimates in the presence of measurement error is dependent on some knowledge of the measurement-error process. It is often possible to build into the design of a study some assessment of the error process, either by the inclusion of a validation group, i.e. a subgroup of individuals for whom it is possible to obtain accurate measurements, or by performing repeated measurements on some of the

subjects (Willet, 1989; Marshall, 1989). How best to integrate this knowledge has been the subject of much research. Existing methods for dealing with measurement error differ according to:

- the type of covariate considered (continuous or categorical);
- the assumptions made on the measurement-error process (Berkson or classical, fully parametric or not);
- the estimation framework considered (maximum likelihood, quasi-likelihood, pseudo-likelihood or Bayesian); and
- whether or not approximations are used.

For continuous covariates, methods substituting an estimate of the expectation of the unobserved covariate given the measured one (also called the 'surrogate') have been discussed by Caroll *et al.* (1984), Rosner *et al.* (1989, 1990), Whittemore (1989) and Pierce *et al.* (1992). Some methods make a specific assumption of small error variance (Stefanski and Caroll, 1985; Whittemore and Keller, 1988; Chesher, 1991; Caroll and Stefanski, 1990). Semi-parametric methods have been considered by Pepe and Fleming (1991), Caroll and Wand (1991), Pepe *et al.* (1994), Robins *et al.* (1994, 1995) and Mallick and Gelfand (1995). While most of the research quoted above has been concerned with models appropriate for cohort studies, the estimation of logistic regression models for case-control studies with errors in covariates has recently been elaborated by Caroll *et al.* (1993) using pseudo-likelihood with non-parametric estimation of the marginal distribution of the unobserved covariates.

The formulation of measurement-error problems in the framework of a Bayesian analysis using graphical models, and the associated estimation methods using stochastic simulation techniques, have recently been developed (Thomas *et al.*, 1991; Stephens and Dellaportas, 1992; Gilks and Richardson, 1992; Richardson and Gilks, 1993a,b; Mallick and Gelfand, 1995). In this chapter, we recount this development, placing particular emphasis on two aspects: the flexibility of this approach, which can integrate successfully different sources of information on various types of measurement process; and the natural way in which all sources of uncertainty are taken account of in the estimation of parameters of interest. Outside the framework of graphical models, a Bayesian approach to logistic regression with measurement error has recently been proposed by Schmid and Rosner (1993).

The structure of the measurement-error problem in epidemiology can be formulated as follows. Risk factors (covariates) are to be related to the disease status (response variable) Y for each individual. However, for many or all individuals in the study, while some risk factors C are truly known, other risk factors X are unknown. It is sometimes possible to obtain information on the unknown risk factors X by recording one or several surrogate

236

measures Z of X for each individual. In other situations, ancillary risk factor information, gained by carrying out surveys on individuals outside the study, but related to the study individuals by known group characteristics, are used (Gilks and Richardson, 1992). To model this general situation, we shall distinguish three submodels (following the terminology introduced by Clayton, 1992):

- a disease model, which expresses the relationship between risk factors C and X and disease status Y;
- an exposure model which describes the distribution of the unknown risk factors X in the general population, or which relates the distribution of X to ancillary information; and
- a measurement model, which expresses the relationship between some surrogate information Z and the true risk factors X, or which links the observed survey to the ancillary risk-factor information.

We shall now detail the structure of two particular epidemiological designs; the first is widely encountered in epidemiology, for example in nutritional studies, whilst the second has arisen more prominently in occupational epidemiology.

22.2 Conditional-independence modelling

22.2.1 Designs with individual-level surrogates

The structure of the three submodels described above can be characterized through the following conditional-independence assumptions:

$$\text{disease model} \qquad [Y_i \mid X_i, C_i, \beta] \tag{22.1}$$

$$\text{measurement model} \qquad [Z_i \mid X_i, \lambda] \tag{22.2}$$

$$\text{exposure model} \qquad [X_i \mid C_i, \pi] \tag{22.3}$$

where subscript i denotes the individual, and β, λ and π are model parameters. Variables in (22.1–22.3) can be scalar or vector. Equations (22.1–22.3) are called *model conditional distributions* ('model conditionals' for short). Since we work in a Bayesian framework, we require prior distributions for β, λ and π.

Conditional-independence assumptions

By asserting (22.1–22.3) as model conditionals, we imply far more than the conditional dependencies made explicit in those equations: we also imply conditional independence relationships which follow from the directed Markov assumption (Lauritzen *et al.*, 1990). This states that the joint distribution of all the variables can be written as the product of the model

conditionals:

$$[\beta]\,[\lambda]\,[\pi]\prod_i [X_i \mid C_i, \pi]\,[Z_i \mid X_i, \lambda]\,[Y_i \mid X_i, C_i, \beta], \tag{22.4}$$

where $[a]$ generically denotes the distribution of a, and $[a \mid b]$ generically denotes the conditional distribution of a given b. Thus, in particular, the following apply:

- (22.1) states that the disease status of individual i, Y_i, is only dependent on its true exposure X_i, on known covariates C_i and on unknown parameters β. We are thus in the classical case where, conditionally on the true exposure being known, the surrogate measures Z_i do not add any information on the disease status. This is a fundamental assumption made in most of the work on measurement error in epidemiology.
- (22.2) states that by conditioning on appropriately defined parameters λ and the true exposure X_i, the surrogate measures Z_i are independent among individuals. The construction of λ will be detailed in an example.
- (22.3) models the population distribution of unknown risk factors among individuals in terms of parameters π. Dependence between the different components of vector X_i can be accommodated through parameters contained in π but the risk factors X_i are assumed independent between individuals given C_i and π.

By specifying the conditional distribution of the surrogate Z given the true exposure X as in (22.2), we are placing ourselves in the Bayesian analog of what is traditionally referred to as the 'classical error model', where measurement error is independent of X. Another type of error model which has been considered in the literature is the Berkson error model, where (22.2) is replaced by $[X_i \mid Z_i, \lambda]$. With the Berkson error model, usually no model need be specified for the marginal distribution of Z.

Conditional-independence graph

An influence diagram or conditional-independence graph corresponding to (22.1–22.3), encompassing several epidemiological designs, is shown in Figure 22.1. We use squares to denote observed quantities and circles to denote unobserved quantities. See Spiegelhalter *et al.* (1995: this volume) for further discussion of conditional independence graphs. Figure 22.1 identifies six groups of individuals, grouped according to which variables are recorded. For example, for individuals in Part 1 of Figure 22.1, variables X_i, Y_i, C_i and Z_i are recorded, whilst for individuals in Part 2 only variables Y_i, C_i and Z_i are recorded. Designs which record X_i on some individuals presume the availability of a 'gold standard', i.e. an error-free method for measuring X.

Parts 1 and 4 of Figure 22.1 are validation studies. In a validation study, both X_i and Z_i are recorded on each individual, providing information

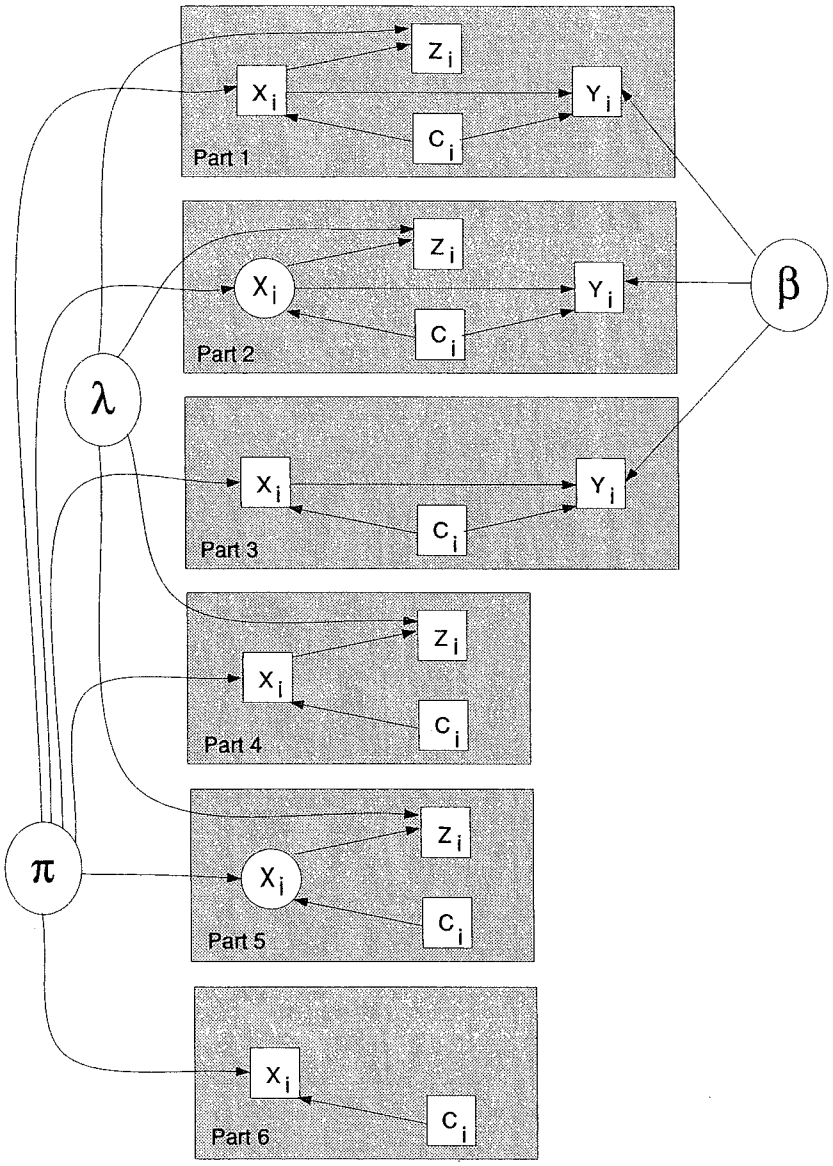

Figure 22.1 *A graph corresponding to equations (22.1–22.3).*

on the measurement-process parameters λ. A validation study is either internal, if disease status Y_i is also recorded (Part 1), or external, if there is no information on disease status (Part 4). Part 2 represents the common situation where only surrogates and disease status are known. Part 3 represents a subgroup in which only the true exposure and disease status are known. Clearly if all the individuals were in this group, there would be no measurement-error problem. In general, the number of individuals included in Parts 1 and 3 will be small compared to Part 2. Finally, Parts 5 and 6 represent 'survey' situations in which information is obtained only on surrogates or on the true exposure. This global influence diagram illustrates how information can flow from one part of the graph to another. For example, Part 6 contributes information on π which in turn provides information on X_i in Part 5, so that Part 5 can yield some information on λ.

When no gold standard is available, validation groups are ruled out and the estimation of measurement model parameters must rely on other sources of information. Designs might then include several measuring instruments with possible repeated determinations. Let Z_{ihr} denote the r^{th} repeated measurement of instrument h for individual i. The model conditional for the measurement process (22.2) now becomes:

$$[Z_{ihr} \mid X_i, \lambda_h]. \tag{22.5}$$

This equation states that, conditionally on the true value of the covariate X_i and on λ_h, there is independence of the surrogates Z_{ihr} between repeats and between instruments.

22.2.2 Designs using ancillary risk-factor information

In occupational or environmental epidemiology, risk-factor information for each individual is often not directly available and must be obtained from ancillary, aggregate-level information, such as a job-exposure matrix in an industrial-epidemiological application. Job-exposure matrices provide information on exposures to each of many industrial agents in each of many finely subdivided categories of occupation. They are commonly constructed by industrial experts from detailed job descriptions obtained in a specially conducted survey. Thus the exposure to industrial agents of each individual in the disease study can be assessed using only his job title, by referring to a job-exposure matrix. The measurement-error model implied by this design is different from those considered above, as imprecision in exposure information provided by the job-exposure matrix must be taken into account.

240

Model conditionals

We consider the case of a dichotomous exposure. Let π_{jk} denote the underlying (unobserved) probability of being exposed to agent k, for individuals in job j. We assume that π_{jk} is the same in the disease study and in the job-exposure survey. Let m_{jk} denote the number of people in the job-exposure survey with job j who were considered by the experts to be exposed to agent k. The model conditional for the aggregate-level survey data $\{m_{jk}\}$ is then:

$$[m_{jk} \mid \pi_{jk}, n_j] = \text{Binomial}\,(\pi_{jk}, n_j), \tag{22.6}$$

where n_j is the number of people with job j included in the survey. Equation (22.6) represents the measurement model in our general formulation in Section 22.1.

The unknown dichotomous exposure X_{ik} of disease-study individual i to agent k is linked to the job-exposure matrix through his job title $j = j(i)$. Since individual i is exposed to agent k ($X_{ik} = 1$) with probability $\pi_{j(i)k}$, the model conditional for exposure in the disease study is given by:

$$[X_{ik} \mid \pi_{j(i)k}] = \text{Bernoulli}\,(\pi_{j(i)k}). \tag{22.7}$$

This represents the exposure model in our general formulation in Section 22.1. The disease model is given as before by equation (22.1).

Note that the job-exposure survey does not provide information directly on X, but rather on the prior distribution of X. We are thus in neither the classical nor the Berkson measurement-error situation. Gilks and Richardson (1992) demonstrate good performance of Bayesian modelling for analysing designs of this kind.

Suppose, in addition to the job-exposure matrix, that direct surrogate dichotomous measures Z_{ik} of X_{ik} are available for some or all disease-study individuals. These might be provided by expert assessment, as in the job-exposure survey. The previous set-up can easily be generalized to include both sources of risk-factor information. We need only specify one additional model conditional:

$$[Z_{ik} \mid X_{ik}, \delta_k], \tag{22.8}$$

where δ_k represents misclassification parameters corresponding to errors in the coding of exposure to agent k. The measurement model now consists of both (22.6) and (22.8). The graph associated with this model is represented in Figure 22.2.

22.2.3 Estimation

Estimation of the models described above can be carried out straightforwardly by Gibbs sampling; see Gilks *et al.* (1995: this volume) for a general description of this method, and Gilks and Richardson (1992) and

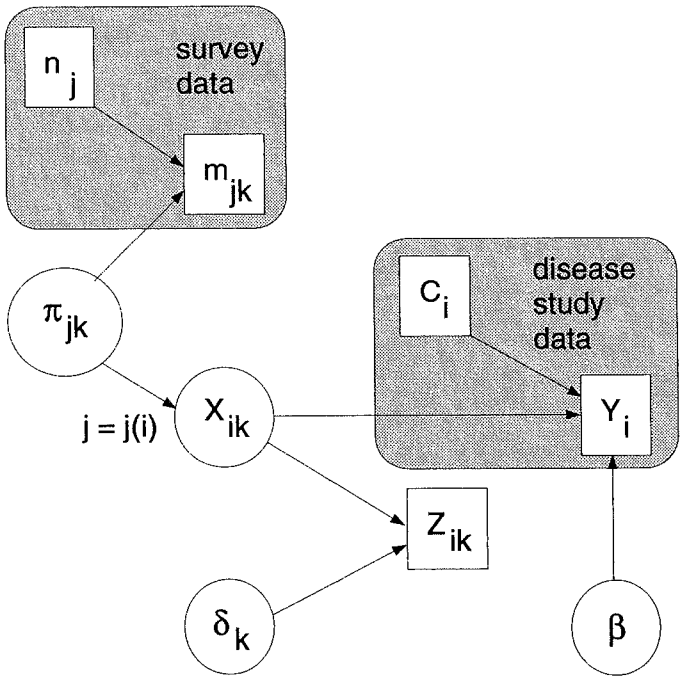

Figure 22.2 *A graph corresponding to equations (22.1) and (22.6–22.8).*

Richardson and Gilks (1993a) for computational details including full conditional distribtions for the above models. Sampling from full conditional distributions was performed using adaptive rejection sampling (Gilks and Wild, 1992; Gilks, 1995: this volume).

22.3 Illustrative examples

In this section, we present a series of examples. Our aims are to illustrate different types of measurement-error situation, to discuss the performance of our approach and to outline areas for further research. We have used simulated data sets throughout to evaluate the performance of our method of analysis.

22.3.1 Two measuring instruments with no validation group

In our first example, we present the analysis of a simulated data set reproducing a design where information on the measurement parameters is obtained through the combination of two measuring instruments.

242

Design set-up

Two risk factors are involved in the disease model. The first one, X, is measured with error and the second one, C, is known accurately. We consider the case of a logistic link between risk factors and disease status. Specifically, we suppose that Y_i follows a Bernoulli distribution with parameter α_i, where logit $\alpha_i = \beta_0 + \beta_1 X_i + \beta_2 C_i$. We suppose that the exposure vector (X, C) follows a bivariate normal distribution, with mean μ and variance-covariance matrix Σ. Thus the vector β in (22.1) comprises parameters $\beta_0, \beta_1, \beta_2$; and π in (22.3) comprises μ and Σ.

Concerning the measurement process, we consider the case of two measuring instruments. The first instrument has low precision but is unbiased. In contrast, the second instrument has a higher precision but is known to be biased. Since Instrument 2 has a higher precision, it is used in preference to Instrument 1 on the entire study population. However, we aim to correct the bias of Instrument 2 by including in the design a subgroup of relatively small size in which both instruments are measured, Instrument 1 being administered twice.

We suppose that the model conditional for the r^{th} repeat of Instrument 1 is a normal distribution with mean X_i and variance θ_1^{-1}:

$$[Z_{i1r} \mid X_i, \theta_1] = \mathrm{N}(X_i, \theta_1^{-1}), \qquad r = 1, 2.$$

Here the only measurement parameter is θ_1, the precision (the inverse of the variance) of Instrument 1. Parameter θ_1 corresponds to λ_1 in (22.5). For the biased Instrument 2, the model conditional is also normal:

$$[Z_{i2} \mid X_i, \phi_2, \psi_2, \theta_2] = \mathrm{N}(\phi_2 + \psi_2 X_i, \theta_2^{-1}).$$

Here the measurement parameters are the intercept ϕ_2 and slope ψ_2 (expressing the linear relationship between the true exposure and its surrogate), and the precision θ_2. These parameters correspond to λ_2 in equation (22.5).

The study is thus designed to include two parts (i.e. two subgroups of individuals) with 1 000 individuals in Part 1 and $n = 200$ or 50 individuals in Part 2. In Part 1, only Instrument 2 has been recorded. In Part 2, Instrument 1 has been measured twice and Instrument 2 has been recorded once on each individual.

A data set was generated using 'true' values of β_0, β_1, β_2, θ_1, ϕ_2, ψ_2 and θ_2 given in Table 22.1 (column: 'true values') and with

$$\mu = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1.02 & 0.56 \\ 0.55 & 0.96 \end{pmatrix}.$$

Thus we have simulated a situation with two detrimental risk factors X and C having relative risks 2.46 and 3.32 respectively, with a positive correlation of 0.56 between X and C. Note that Instrument 2 is three times more accurate than Instrument 1 ($\theta_2 = 3 \times \theta_1$).

parameter	true value	Gibbs sampling analysis n=200 mean ± s.d.		Gibbs sampling analysis n=50 mean ± s.d.		classical analysis with Instrum. 2 mean	± s.e.
β_0	-0.8	-0.81	0.32	-0.77	0.40	-0.17	0.11
β_1	0.9	1.03	0.36	0.98	0.42	0.14	0.07
β_2	1.2	1.25	0.14	1.36	0.21	1.57	0.11
θ_1	0.3	0.31	0.03	0.25	0.04	–	–
ϕ_2	0.8	0.81	0.11	0.84	0.14	–	–
ψ_2	0.4	0.44	0.09	0.45	0.14	–	–
θ_2	0.9	0.91	0.06	0.96	0.13	–	–

Table 22.1 *Gibbs sampling analysis of a design with 2 measuring instruments*

Results

Table 22.1 presents the results from a Gibbs sampling analysis of the simulated data set. Hyperparameters were chosen to specify only vague prior information (for details see Richardson and Gilks, 1993a). We have summarized marginal posterior distributions of parameters of interest by reporting posterior means and standard deviations. In the last two columns of Table 22.1, we have given classical estimates of log relative risks and their standard errors which would be obtained if the analysis relied solely on the data of Part 1, i.e. if no correction for measurement error was performed. We see that these estimates are quite hopeless and that the simulated situation is one in which measurement error, if not taken into account, substantially influences the results.

The results show that our estimation method has performed satisfactorily with all the estimated values lying within one posterior standard deviation of the values set in the simulation. As expected, the posterior standard deviation for β_2 which corresponds to the covariate C measured without error is smaller than that for β_1 which corresponds to X. Note the influence of the size n of Part 2 on all posterior standard deviations. The measurement parameters for Instrument 2 have been well estimated, even though our design did not include a validation subgroup. This highlights how information has been naturally propagated between the two measuring instruments and between the two parts of the design.

In this particular design, even though there is no gold standard, the data still contain information on ϕ_2 and ψ_2 because there is information on X_i from the repeats of the unbiased Instrument 1, information which represents in effect a 'simulated gold standard'. The size n of Part 2 is clearly crucial in this process: with a smaller validation subgroup, the results deteriorate.

22.3.2 Influence of the exposure model

At each step of the approach we have outlined so far, conditional distributions need to be explicitely specified in a parametric way. Whilst some of these parametric distributions arise naturally, such as the choice of a logistic model for the disease risk, other assumed distributional forms are more arbitrary. In particular, there are some cases where little is known about the distribution of the exposure X and an appropriate model for it (Thomas *et al.*, 1991).

The exposure model (22.3) we have implemented in our recent work is that of a simple Gaussian distribution. A natural generalization of the exposure model would be a mixture of Gaussians or other distributions. As a first step in that direction, we have assessed the effect of misspecifying a simple Gaussian variable for exposure when the true exposure is a mixture of Gaussians or a χ^2 distribution, as we now describe.

Simulation set-up

We consider a study with one known risk factor C and another risk factor X measured with error by a surrogate Z on 1 000 individuals. The study also includes a validation group containing 200 individuals where both X and Z are measured. As in the previous example, we assume a logistic link between risk factor and disease status. Three datasets were simulated, differing in the generating distribution for true exposures X:

$$\begin{aligned}
&\text{(a)} \quad X_i \sim \frac{1}{2}\mathrm{N}(-1.0, 1.0) + \frac{1}{2}\mathrm{N}(2.0, 1.0);\\
&\text{(b)} \quad X_i \sim \frac{1}{2}\mathrm{N}(-3.0, 1.0) + \frac{1}{2}\mathrm{N}(4.0, 1.0);\\
&\text{(c)} \quad X_i \sim \chi^2_1.
\end{aligned}$$

In each dataset the surrogate Z_i was generated by:

$$[Z_i \mid X_i, \theta] = \mathrm{N}(X_i, \theta^{-1}).$$

Each of the three simulated datasets was analysed by Gibbs sampling, with the assumption that the exposure model for (X, C) was misspecified as a bivariate normal distribution with mean μ and variance-covariance matrix Σ, with a vague prior distribution for μ centred around $(0.5, 0.5)$ and a Wishart prior distribution for Σ with 5 degrees of freedom and identity scale matrix.

The results are shown in Table 22.2. With a bimodal, symmetric true exposure distribution (Datasets (a) and (b)), the parameters are still adequately estimated, although there is some deterioration when the modes are well-separated (Dataset (b)). However in Dataset (c), where the true exposure distribution is skewed, misspecification has led to attenuation of the estimate of β_1, and its 95% credible interval no longer contains the

true value. Hence misspecification of the exposure model can influence the regression results. Further studies of the influence of misspecification are warranted.

parameter	true value	Dataset (a) mean ± *s.d.*		Dataset (b) mean ± *s.d.*		Dataset (c) mean ± *s.d.*	
β_0	-0.8	-0.82	0.11	-0.72	0.13	-0.68	0.12
β_1	0.9	0.91	0.08	0.77	0.07	0.56	0.08
β_2	1.2	1.20	0.08	1.03	0.08	1.34	0.13
θ	0.9	0.87	0.08	0.87	0.09	0.86	0.08
μ_X	0.5	0.50	0.06	0.50	0.11	1.35	0.07
μ_C	0.5	0.55	0.05	0.63	0.10	1.24	0.05

Table 22.2 *Gibbs sampling analysis with misspecified exposure distribution*

Mixture models provide great flexibility in modelling distributions with a variety of shapes. The next stage in the development of our graphical model will be to employ a mixture model for the exposure distribution (22.3). Gibbs sampling analysis of mixtures is discussed by Robert (1995: this volume). A difference between our set-up and the set-up usually considered in mixture problems is that, in our case, we do not observe a fixed sample from the mixture distribution; rather, this sample (corresponding to the unknown risk factor X) is generated anew at each iteration of the Gibbs sampler. It will be thus interesting to see how convergence of the Gibbs sampler is modified by the additional level of randomness.

22.3.3 Ancillary risk-factor information and expert coding

In our last example, we illustrate how ancillary information can be combined with expert coding to provide estimates of both regression coefficients and probabilities of misclassification. This extends the work of Gilks and Richardson (1992). We first describe how the job-exposure survey data and the disease-study data were generated.

Generating disease-study data

Each of 1 000 individuals were randomly and equiprobably assigned to one of four occupations ($j = 1, \ldots, 4$). For each individual, exposure status ($X_i = 1$: exposed; $X_i = 0$: not exposed) to a single industrial agent was randomly assigned according to true job-exposure probabilities $\{\pi_j\}$:

246

j	π_j
1	0.9
2	0.3
3	0.5
4	0.6

Disease status ($Y_i = 1$: diseased; $Y_i = 0$: not diseased) was assigned with probability of disease α_i, where logit $\alpha_i = \beta_0 + \beta_1 X_i$.

We also supposed that, for each of the 1 000 individuals, an expert was able to code their exposure with a variable Z_i. This was generated from X_i with the following misclassification probabilities:

$$P(Z_i = 1 \mid X_i = 0) = \gamma_1; \qquad P(Z_i = 0 \mid X_i = 1) = \gamma_2.$$

Exposures X_i were thenceforth assumed unknown for each individual, so the analysis was based only on Y_i, Z_i and the job-exposure survey data, described next. Three datasets were generated:

(a) Z_i not available;

(b) $\gamma_1 = 0.2, \gamma_2 = 0.5$;

(c) $\gamma_1 = 0.1, \gamma_2 = 0.3$.

Generating the job-exposure survey data

Each of 150 individuals were assigned an occupation and an exposure status, as for the disease study individuals. The observed job-exposure matrix $\{n_j, m_j\}$ (where n_j is the number of surveyed individuals in job j and m_j is the number 'coded by the expert' as 'exposed') was compiled from these 150 individuals, and the job-exposure probabilities π_j were thenceforth assumed unknown. The job-exposure matrix was assumed to be available for each of the three datasets described above.

Analysing the simulated data

For the graphical model, we assumed a normal $N(0, 9)$ prior distribution for the regression parameter β, and almost flat priors for logit π_j, logit γ_1 and logit γ_2. We ran the Gibbs sampler for 7 000 iterations, discarding the first 100 iterations before analysis. The results are presented in Table 22.3.

By comparing the results for Datasets (b) and (c) with Dataset (a), one can clearly see that the information given by the coding of the expert leads to improved estimates of the regression coefficients, with smaller posterior standard deviations. The improvement is more marked in (c) as the misclassification probabilities are smaller than in (b). Morevover good estimates of the misclassification probabilities are also obtained.

Dataset	parameters	β_0	β_1	γ_1	γ_2
(a)	true values	-1.00	1.50	–	–
	posterior mean	-0.99	1.70	–	–
	posterior *s.d.*	0.30	0.43	–	–
(b)	true values	-1.00	1.50	0.20	0.50
	posterior mean	-1.20	1.63	0.16	0.51
	posterior *s.d.*	0.27	0.34	0.04	0.03
(c)	true values	-1.00	1.50	0.10	0.30
	posterior mean	-1.11	1.56	0.08	0.32
	posterior *s.d.*	0.18	0.24	0.04	0.03

Table 22.3 *Gibbs sampling analysis of designs with ancillary risk-factor information and expert coding*

22.4 Discussion

In this chapter, we have presented a unifying representation through conditional independence models of measurement-error problems with special reference to epidemiological applications. There are several advantages of this approach over methods previously proposed which are extensively discussed in Richardson and Gilks (1993a). Of paramount importance is its flexibility, which enables the modelling of an extensive range of measurement-error situations without resorting to artificial simplifying assumptions. This has important design implications for future studies. Now that analyses of complex measurement-error designs can be carried out successfully, there is more freedom at the design stage. An important area for future research is thus to give guidelines for complex designs.

The key to the construction of such models is the stipulation of suitable conditional independence assumptions. Careful thought has to be given to the implications of each of these assumptions in any particular context. For example, in (22.1–22.4) we have assumed independence between the Y conditional on the X, C and β. This is an appropriate assumption in chronic disease epidemiology but is likely to be violated when considering infectious diseases. Indeed, the disease status of an individual is influenced, through contagious contacts, by the disease status of other individuals. As another example, the conditional independence between repeated measures of surrogates given the true risk factor, assumed by (22.5), would not hold if there is a systematic bias in the measurement.

The approach we have developed is fully parametric. The influence on regression parameter estimates of misspecification of the measurement error or exposure distributions gives cause for concern. The use of flexible mixture distributions is a natural way to relax the fully parametric set-up. This

248

approach has been taken by Mallick and Gelfand (1995). By using a mixture of cumulative beta distribution functions to model unknown cumulative distribution functions, they formulate a semi-parametric Bayesian approach to a measurement-error problem, implemented through a single-component Metropolis–Hastings algorithm.

References

Armstrong, B. G. (1990) The effects of measurement errors on relative risk regression. *Am. J. Epidem.*, **132**, 1176–1184.

Begg, M. D. and Lagakos, S. (1993) Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. *J. Am. Statist. Ass.*, **88**, 166–170.

Brenner, H. (1993) Bias due to non-differential misclassification of polytomous confounders. *J. Clin. Epidem.*, **46**, 57–63.

Caroll, R. J. (1989) Covariance analysis in generalized linear measurement error models. *Statist. Med.*, **8**, 1075–1093.

Caroll, R. J. and Stefanski, A. (1990) Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Am. Statist. Ass.*, **85**, 652–663.

Caroll, R. J. and Wand, M. P. (1991) A semiparametric estimation in logistic measurement error models. *J. R. Statist. Ass.* B, **53**, 573–585.

Caroll, R. J., Gail, M. H and Lubin, J. H. (1993) Case-control studies with errors in covariates. *J. Am. Statist. Ass.*, **88**, 185–199.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Non-linear Measurement Error Models*. London: Chapman & Hall.

Caroll, R. J., Spiegelman, C., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. (1984) On errors in variables for binary regression models. *Biometrika*, **71**, 19–26.

Chesher, A. (1991) The effect of measurement error. *Biometrika*, **78**, 451–462.

Clayton, D. G. (1992) Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health* (eds J. H. Dwyer, M. Feinleib and H. Hoffmeister), pp. 301-331. Oxford: Oxford University Press.

Fuller, W. A. (1987) *Measurement Error Models*. New York: Wiley.

Gail, M. H. (1991) A bibliography and comments on the use of statistical models in epidemiology in the 1980s. *Statist. Med.*, **10**, 1819–1885.

Gilks, W. R. (1995) Full conditional distributions. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 75–88. London: Chapman & Hall.

Gilks, W. R. and Richardson, S. (1992) Analysis of disease risks using ancillary risk factors, with application to job-exposure matrices. *Statist. Med.*, **11**, 1443–1463.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1995) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 1–19. London: Chapman & Hall.

Greenland, S. (1980) The effect of misclassification in the presence of covariates. *Am. J. Epidem.*, **112**, 564–569.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H. G. (1990) Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Liu, X. and Liang, K. J. (1991) Adjustment for non-differential misclassification error in the generalized linear model. *Statist. Med.*, **10**, 1197–1211.

Mallick, B. K. and Gelfand, A. E. (1995) Semiparametric errors-in-variables models: a Bayesian approach. Technical report, Imperial College, London University.

Marshall, J. R. (1989) The use of dual or multiple reports in epidemiologic studies. *Statist. Med.*, **8**, 1041–1049.

Pepe, M. S. and Fleming, T. R. (1991) A non-parametric method for dealing with mismeasured covariate data. *J. Am. Statist. Ass.*, **86**, 108–113.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994) Auxiliary outcome data and the mean score method. *J. Statist. Plan. Inf.*, **42**, 137–160.

Pierce, D. A., Stram, D. O., Vaeth, M. and Schafer, D. W. (1992) The errors in variables problem: considerations provided by radiation dose-response analyses of the A-bomb survivor data. *J. Am. Statist. Ass.*, **87**, 351–359.

Richardson, S. and Gilks, W. R. (1993a) Conditional independence models for epidemiological studies with covariate measurement error. *Statist. Med.*, **12**, 1703–1722.

Richardson, S. and Gilks, W. R. (1993b) A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am. J. Epidem.*, **138**, 430–442.

Robert, C. P. (1995) Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 441–464. London: Chapman & Hall.

Robins, J. M., Hsieh, F. and Newey, W. (1995) Semiparametric efficient estimates of a conditional density with missing or mismeasured covariates. *J. R. Statist. Soc.* B, **57**, 409–424.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.

Rosner, B., Spiegelman, D. and Willett, W. C. (1989) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist. Med.*, **8**, 1051–1069.

Rosner, B., Spiegelman, D. and Willett, W. C. (1990) Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am. J. Epidem.*, **132**, 734–745.

Schmid, C. H. and Rosner, B. (1993) A Bayesian approach to logistic regression models having measurement error following a mixture distribution. *Statist. Med.*, **12**, 1141–1153.

Spiegelhalter, D. J., Best, N. G., Gilks, W. R. and Inskip, H. (1995) Hepatitis B: a case study in MCMC methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 21–43. London: Chapman & Hall.

Stefanski, L. A. and Caroll, R. J. (1985) Covariate measurement error in logistic regression. *Ann. Statist.*, **13**, 1335–1351.

Stephens, D. A. and Dellaportas, P. (1992) Bayesian analysis of generalised linear models with covariate measurement error. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 813–820. Oxford: Oxford University Press.

Thomas, D. C., Gauderman, W. J. and Kerber, R. (1991) A non-parametric Monte Carlo approach to adjustment for covariate measurement errors in regression problems. Technical report, Department of Preventive Medicine, University of Southern California.

Thomas, D., Stram, D. and Dwyer, J. (1993) Exposure measurement error: influence on exposure-disease relationship and methods of correction. *Annual Rev. Pub. Health*, *14*, 69–93.

Whittemore, A. S. (1989) Errors in variables regression using Stein estimates. *Am. Statist.*, **43**, 226–228.

Whittemore, A. S. and Keller, J. B. (1988) Approximations for regression with covariate measurement error. *J. Am. Statist. Ass.*, **83**, 1057–1066.

Willett W. (1989) An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statist. Med.*, **8**, 1031–1040.