Statistical Modelling 2003; **3**: 215–232

Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation

Sophia Rabe-Hesketh[1], **Andrew Pickles**[2] **and Anders Skrondal**[3]
[1]Department of Biostatistics and Computing, Institute of Psychiatry, King's College London, London, UK
[2]School of Epidemiology and Health Sciences and CCSR, The University of Manchester, Manchester, UK
[3]Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway

Abstract: When covariates are measured with error, inference based on conventional generalized linear models can yield biased estimates of regression parameters. This problem can potentially be rectified by using generalized linear latent and mixed models (GLLAMM), including a measurement model for the relationship between observed and true covariates. However, the models are typically estimated under the assumption that both the true covariates and the measurement errors are normally distributed, although skewed covariate distributions are often observed in practice. In this article we relax the normality assumption for the true covariates by developing nonparametric maximum likelihood estimation (NPMLE) for GLLAMMs. The methodology is applied to estimating the effect of dietary fibre intake on coronary heart disease. We also assess the performance of estimation of regression parameters and empirical Bayes prediction of the true covariate. Normal as well as skewed covariate distributions are simulated and inference is performed based on both maximum likelihood assuming normality and NPMLE. Both estimators are unbiased and have similar root mean square errors when the true covariate is normal. With a skewed covariate, the conventional estimator is biased but has a smaller mean square error than the NPMLE. NPMLE produces substantially improved empirical Bayes predictions of the true covariate when its distribution is skewed.

Key words: empirical Bayes prediction; factor model; generalized linear models; GLLAMM; logistic regression; measurement error; nonparametric maximum likelihood estimation

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received March 2002; **revised** October 2002, January 2003, April 2003; **accepted** April 2003

1 Introduction

Explanatory variables in generalized linear models are frequently subject to measurement error, for instance in epidemiology where the effects of lifetime exposure to pollutants, alcohol, exercise and so on are often of interest.

Address for correspondence: S Rabe-Hesketh, Senior Lecturer in Statistics, Department of Biostatistics and Computing, Institute of Psychiatry, King's College London, DeCrespigny Park, London SE5 8AF, UK. E-mail: spaksrh@iop.kcl.ac.uk

 10.1191/1471082X03st056oa

Explicit measurement modelling is crucial for at least two reasons: First, neglecting measurement error will often lead to biased estimates for the regression parameters. For instance, it is well known that ordinary logistic regression can lead to biased estimates of odds ratios when the covariates are subject to measurement error (Rosner *et al.*, 1990), a phenomenon known as regression dilution in the simple case of a single covariate. Joint modelling of the response and measurement process allows estimation of a 'disattenuated' odds ratio for the true covariate (see, for example, Carroll *et al.*, 1995). Secondly, measurement modelling facilitates prediction of the true covariate or exposure for an individual unit, utilizing not only the exposure measurements for the unit but also information from the outcome as well as 'borrowing strength' from the other units.

In conventional covariate measurement error models, it is assumed that both measurement errors and true covariate are normally distributed. The validity of these usually arbitrary assumptions is often questionable. Indeed, the assumptions are at odds with skewed covariate distributions, which are often observed in applications. Thus, it appears to be very useful to relax the assumption of a normal true covariate by instead using 'nonparametric maximum likelihood estimation' (NPMLE) (Laird, 1978). The NPMLE of the exposure distribution is discrete with nonzero probabilities at a finite set of locations (Simar, 1976; Laird, 1978; Lindsay, 1983).

Although the specification of nonparametric distributions for true covariates is theoretically attractive, research on evaluating the practical performance of NPMLE for covariate measurement error models is scarce and limited. First, little attention has been given to the evaluation of NPMLE for generalized linear models with covariate measurement error when the true covariate distribution is normal. In this case maximum likelihood under a correct model specification is known to be optimal and little is known regarding the finite sample performance of NPMLE. However, adequate performance of NPMLE under normality would appear to be a minimum requirement for NPMLE to be considered a robust alternative to the conventional method. The only simulation studies we are aware of are by Hu *et al.* (1998) and Schafer (2001). For Cox regression with covariate measurement error, Hu *et al.* claim that NPMLE produced regression estimates exhibiting some bias. However, instead of jointly estimating locations and masses as required for NPMLE, their 'pseudo NPMLE' method constrained the masses to lie on a subset of 20 prespecified locations. Hu *et al.* (1998) also found that the pseudo NPML estimates had a greater standard deviation than the estimates assuming a normal true covariate distribution. For logistic regression with covariate measurement error, Schafer (2001) states, without giving details, that the NPML estimates have similar characteristics to the maximum likelihood estimates based on the correct distributional assumptions. However, Schafer only simulates the somewhat unrealistic case where the measurement error variance is known. We compare the performance of the estimators for the nonparametric and correctly specified covariate distribution when the measurement error variance is estimated jointly with the other parameters and the mass locations for NPMLE are freely estimated. Secondly, we are aware of only two studies investigating the robustness of the conventional model to mis-specification of the true covariate distribution. Using the skewed chi-square distribution with 1 degree of freedom, Thoresen and Laake (2000) found that maximum likelihood estimation of the conventional model using 32-point

quadrature performed reasonably well both for small and large samples. This is in contrast to Schafer (2001), who found the regression coefficient estimates of the conventional model to be biased when the true covariate was simulated from a moderately skewed mixture of two normals with mixing probability 0.75, means 50 and 80 and variances 100 and 300. Again, the measurement error variance was taken as known and it is also not clear how the conventional model was estimated. Thirdly, the performance of NPMLE has, to our knowledge, only been assessed when the true covariate has a moderately skewed distribution given by the mixture of two normals mentioned above. Schafer (2001) found that the NPML estimates were unbiased in this case but had a greater root mean square error than those of the conventional model, where the measurement error variance was again taken as known. We will consider the case of a highly skewed covariate distribution as found in many applications since it is in these situations that NPMLE would be expected to have the greatest potential. We will assess the performance of maximum likelihood estimation for both the model assuming a normal and a nonparametric exposure distribution where the measurement error variance is estimated jointly with the other parameters. Fourthly, unlike previous simulation studies, we also evaluate the performance of empirical Bayes prediction of the true covariate. Prediction both under normality and using a nonparametric distribution is considered for normal as well as skewed true covariate distributions.

Estimation (using NPMLE or maximum likelihood (ML) under normality) and empirical Bayes prediction (based on NPMLE or normality) for generalized linear models with measurement error is implemented in `gllamm`, a Stata® program for generalized linear latent and mixed models (GLLAMM). The methods are used to estimate the effect of dietary fibre intake on heart disease. Using the terminology common in epidemiology, we will in the following sometimes refer to the true covariate as 'exposure,' the observed covariate as 'measured exposure' and the binary outcome as 'disease.'

2 Models

We will concentrate on the problem of estimating the association between a true exposure and disease in a logistic regression model, when exposure measurement is imperfect and replicate measurements are available for at least a subsample. This may be accomplished by introducing a latent variable for the unobserved true value of the exposure and specifying three submodels [following Clayton's (1992) terminology]; an exposure model, a measurement model and a disease model.

2.1 Exposure model

We will model true exposure F_i for unit i as

$$F_i = \gamma_0 + \gamma_1 x_i + u_i \tag{2.1}$$

where x_i is a covariate (there may be several), γ_0 and γ_1 are regression parameters and u_i is a latent variable representing the deviation of unit i's true exposure from the mean

exposure for covariate x_i. Traditionally, a normal exposure distribution has been assumed with $u_i \sim N(0, \tau^2)$, but we will also consider a nonparametric exposure distribution. Principally in the context of random intercept models, Simar (1976) and Laird (1978), and more generally Lindsay (1983), have shown that the NPMLE of the unspecified (possibly continuous) distribution is a discrete distribution with nonzero probabilities π_k at a finite set of locations z_k, $k = 1, \ldots, K$ (see also Lindsay *et al.*, 1991; Aitkin, 1996, 1999). The locations and probabilities can be estimated jointly with the other parameters using maximum likelihood estimation. Here the number of locations is increased until the largest maximized likelihood is achieved. However, there does not appear to exist a formal proof of boundedness of the likelihood with increasing K.

Maximum likelihood theory for this model is reviewed by Lindsay (1995) and Böhning (2000). Roeder *et al.* (1996) consider this approach for covariate measurement error in case–control studies with a validation sample, Aitkin and Rocci (2002) when there are neither replicates nor validation samples, Hu *et al.* (1998) for Cox regression with replicates and Schafer (2001) for linear, nonlinear and logistic regression.

An important advantage of NPMLE is that it is appropriate regardless of the latent variable distribution. True exposure could be continuous (normal or non-normal), as for instance blood pressure, or discrete as in medical diagnosis. Mixtures of discrete and continuous distributions are also a possibility, for example, nonsmokers within a cigarette exposure distribution. Relying on NPMLE, we can concentrate on the specification of other model components and need not worry about the nature of the latent variable distribution.

2.2 Measurement model

The classical measurement model assumes that the rth exposure measurement for unit i, f_{ir}, differs from the true exposure F_i by a normally distributed measurement error ϵ_{ir},

$$\begin{aligned} f_{ir} &= F_i + \epsilon_{ir}, \qquad \epsilon_{ir} \sim N(0, \sigma_f^2) \\ &= \gamma_0 + \gamma_1 x_i + u_i + \epsilon_{ir} \end{aligned} \tag{2.2}$$

where ϵ_{ir} and u_i are independent and we have substituted for F_i from (2.1). The repeated measurements on the same unit are therefore assumed to be conditionally independent given true exposure. We consider the case of (a) nontransformed measurement and (b) log transformed measurement.

The total variance of the measurements, conditional on x_i, is given by

$$\mathrm{var}(f_{ir}|x_i) = \sigma_f^2 + \mathrm{var}(u_i)$$

The conditional reliability R may therefore be estimated from the model parameters by substituting the relevant terms into

$$R = \frac{\mathrm{var}(F_i|x_i)}{\mathrm{var}(f_{ir}|x_i)} \tag{2.3}$$

(see, for example, Dunn (1989), p. 54).

Our framework allows very general measurement models, specified in terms of multivariate generalized linear models

$$g(\mu_{ir}) = \gamma_{0r} + \gamma_{1r}x_i + \lambda_r u_i$$

where λ_r is a factor loading. The measurements can therefore be of any type accommodated by generalized linear models including mixed types. For instance, the classical measurement model arises as the special case with identity links and conditionally normal measurements. If all measurements are dichotomous, we obtain a two-parameter item response model. Additional generality is allowed by including multidimensional latent variables.

2.3 Disease model

The disease model specifies the relationship between the outcome variable and true exposure and could also take the form of any generalized linear model. Often, a logistic regression on true exposure is used, that is,

$$\begin{aligned}\text{logit}(\text{P}[d_i = 1|F_i]) &= \alpha_0 + \alpha_1 x_i + \beta F_i \\ &= \delta_0 + \delta_1 x_i + \beta u_i \end{aligned} \tag{2.4}$$

where $\delta_0 = \alpha_0 + \beta\gamma_0$ and $\delta_1 = \alpha_1 + \beta\gamma_1$.

We have included x_i both in the disease and exposure models to allow for a direct effect of x_i on disease (α_1) as well as any indirect effects through affecting true exposure ($\beta\gamma_1$). If there is no direct effect, estimation of the reduced form parameters δ_0 and δ_1 requires nonlinear constraints since with $\alpha_1 = 0$, $\delta_1 = \beta\gamma_1$ is the product of the effect of true exposure on disease and the effect of x_i on true exposure. In `gllamm` this problem can be overcome by directly estimating the structural parameters (the αs, γs and β).

Note that the assumed relationship between true and measured exposures has implications for the interpretation of the odds ratio $\exp(\beta)$ in the disease model since this is measured per unit increase in F_i (or u_i). Using an identity link as in the classical measurement model has the advantage that the odds ratio can then be interpreted as the effect on the odds of an *absolute* difference in true exposure except when the log of the measured exposure is used in which case the effect relates to a *relative* difference in exposure.

2.4 Parameter restrictions

We will now consider in more detail the most common case of having two repeated exposure measurements as in the heart disease data analysed in Section 4. The model can in this case be presented graphically as in Figure 1, where circles represent latent variables, rectangles observed variables, arrows from explanatory to response variables represent regressions (linear or logistic) and short arrows to response variables represent residual variability (e.g., additive errors for linear regression).

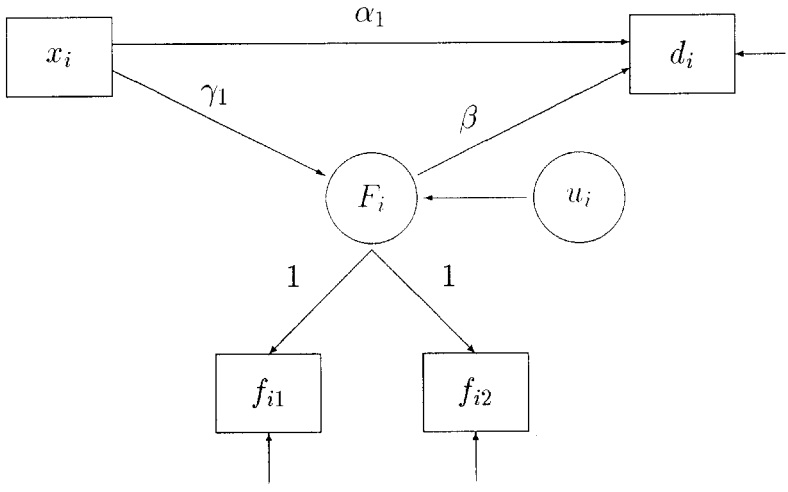

Figure 1 Logistic regression model with covariate measurement error and two replicates

Note that due to the absence of any direct effects of f_{ir} on d_i or vice versa, the model assumes that the risk of disease given true exposure is independent of measured exposure so that measurement error is nondifferential.

The disease and measurement models jointly represent a factor model for three mixed responses where F_i is the common factor, β is the factor loading for the binary response d_i and the factor loadings for the continuous exposure measurements f_{ir} have been set to 1. In factor models at least one factor loading must be set to an arbitrary nonzero constant or the factor variance fixed. If no such restriction were used, multiplying the factor loadings by a positive constant a could be counteracted by dividing the factor variance by a^2, that is, the scale of the factor is not identified. Setting the factor loadings in the additive measurement model to 1 fixes the scale of the latent variable equal to that of the exposure measurements so that the odds ratio in the disease model can be interpreted as the effect of an increase in true exposure by one unit, in terms of the physical units (e.g., grams per day) in which exposure is measured. In addition, the mean of the factor u_i is set to 0 so that the intercept γ_0 in the measurement model is identified. In the case of NPMLE, a more common identification restriction is to set γ_0 or one of the locations to zero, but here we use the above restriction to make the estimates comparable across estimation methods. Because the exposure measurements are 'exchangeable' replicates, we have set both factor loadings equal and measurement error variances equal. This specification may be relaxed since we can identify a separate factor loading λ for one of the measurements as well as separate measurement error variances σ^2_{f1} and σ^2_{f2} in addition to var(F_i) and β. Differences in mean between measures are also allowed for, sample 'drift' being commonplace (Carroll *et al.*, 1995).

When there is only one exposure measurement it is important to note that var(F_i), β and σ^2_f are not jointly identified from the first and second order moments of the observed variables. This renders the conventional covariate measurement error models (where exposure, 'disease' and measurements are conditionally normal) nonidentified.

The reliability or measurement error variance is therefore usually, and often unrealistically, assumed to be known *a priori* in this case (typically estimated in another study). With a normal exposure distribution, theoretical identification may be achieved through information in the higher order moments by specifying a logistic disease model, but this is likely to be fragile (Rabe-Hesketh and Skrondal, 2001). Alternatively, identification can be obtained by using a non-normal exposure distribution (e.g., Reiersøl, 1950). Aitkin and Rocci (2002) argue that NPMLE leads to identification in models with conditionally normal 'disease' and measurements without replicates.

3 Estimation and prediction

3.1 Estimation

Owing to the conditional independence between the exposure measurements and disease status given true exposure, the likelihood is simply the product of the measurement, disease and exposure models, integrated over exposure or, equivalently u_i. If a normal distribution is assumed for u_i, the likelihood is

$$L(\boldsymbol{\theta}_{\mathrm{D}}, \boldsymbol{\theta}_{\mathrm{M}}, \tau) = \prod_i \int \mathrm{P}(d_i \mid u_i; \boldsymbol{\theta}_{\mathrm{D}}) \prod_{r=1}^{n_i} g(f_{ir} \mid u_i; \boldsymbol{\theta}_{\mathrm{M}}) g(u_i; \tau)\, \mathrm{d}u_i \tag{3.1}$$

where $\boldsymbol{\theta}_{\mathrm{D}}$ are parameters of the disease model, $\boldsymbol{\theta}_{\mathrm{M}}$ are parameters of the measurement model and the second product is over all n_i measurements available on unit i. If measurements or disease status are missing at random (MAR) or completely at random (MCAR), the maximum likelihood estimates will be consistent (e.g., Rubin, 1976). Note that replicates are often MCAR or MAR by design, in which case no assumption regarding missingness need be invoked for the measurements.

For logistic regression with a classical measurement model the terms in the likelihood become

$$\mathrm{P}(d_i \mid u_i; \boldsymbol{\theta}_{\mathrm{D}}) = \frac{\exp\{d_i[\alpha_0 + \beta\gamma_0 + (\alpha_1 + \beta\gamma_1)x_i + \beta u_i]\}}{1 + \exp[\alpha_0 + \beta\gamma_0 + (\alpha_1 + \beta\gamma_1)x_i + \beta u_i]}$$

$$g(f_{ir} \mid u_i; \boldsymbol{\theta}_{\mathrm{M}}) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{[f_{ir} - (\gamma_0 + \gamma_1 x_i + u_i)]^2}{2\sigma_f^2}\right)$$

and

$$g(u_i; \tau) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{u_i^2}{2\tau^2}\right)$$

The likelihood has no closed form but may be integrated numerically using Gauss–Hermite quadrature (e.g., Bock and Lieberman, 1970; Butler and Moffitt, 1982; Aitkin, 1999).

If the exposure distribution is not specified, the likelihood has the form

$$L(\boldsymbol{\theta}_{\mathrm{M}}^{K}, \boldsymbol{\theta}_{\mathrm{D}}^{K}, \boldsymbol{\pi}_{K}, \mathbf{z}_{K}) = \prod_{i} \sum_{k=1}^{K} \pi_k \mathrm{P}(d_i|u_i = z_k; \boldsymbol{\theta}_{\mathrm{D}}^{K}) \prod_{r=1}^{n_i} g(f_{ir}|u_i = z_k; \boldsymbol{\theta}_{\mathrm{M}}^{K}), \qquad (3.2)$$

where z_k are location parameters, π_k are mass parameters and $\mathbf{z}_K, \boldsymbol{\pi}_K$ are the corresponding K-dimensional parameter vectors. Note that the Gauss–Hermite quadrature approximation has the same form with z_k replaced by τv_k and with the crucial difference that the locations v_k and masses π_k are fixed *a priori*.

The parameters are estimated using a Newton–Raphson algorithm. Standard errors are estimated by inverting the observed information matrix, obtained by numerical differentiation of the log-likelihood. For NPMLE, the information matrix includes terms for the mass-point parameters so that standard errors for regression coefficients take account of the uncertainty regarding the 'true' locations and masses, but not regarding the number of mass-points. The standard errors are therefore conditional on the number of mass-points. To ensure positive measurement error variances, we estimate $\ln \sigma_f$ and use the delta method to derive standard errors of (back) transformed parameters. For NPMLE, $k-1$ log odds and $K-1$ locations are estimated to ensure valid probabilities and to constrain the mean of u_i to zero.

In order to achieve the NPML estimator, the maximum number of mass-points needs to be determined, so that if a further mass-point is added, it will either be estimated as having zero mass or as sharing a location (or very close location) with another mass. A common approach is to start estimation with a large number of mass-points and omit points that either merge with other points or whose mass approaches zero during maximization of the likelihood (e.g., Butler and Louis, 1992). In contrast, we introduce mass-points one by one using the concept of a directional derivative (e.g., Simar, 1976; Jewell, 1982; Böhning, 1982; Lindsay, 1983), referred to as the Gateaux derivative by Heckman and Singer (1984). Consider the maximized likelihood for K masses $L(\hat{\boldsymbol{\theta}}_{\mathrm{D}}^{K}, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^{K}, \hat{\boldsymbol{\pi}}_{K}, \hat{\mathbf{z}}_{K})$. To determine whether this is the NPMLE solution, we consider changing the discrete mass-point distribution along the path $((1-\lambda)\hat{\boldsymbol{\pi}}_K, \lambda)'$ with locations $(\hat{\mathbf{z}}_K, z_{K+1})'$, where $\lambda = 0$ corresponds to the current solution and $\lambda = 1$ places unit mass at a new location z_{K+1}. The directional derivative is then defined as

$$\Delta(z_{K+1}) = \lim_{\lambda \to 0} \frac{\ln L(\hat{\boldsymbol{\theta}}_{\mathrm{D}}^{K}, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^{K}, ((1-\lambda)\hat{\boldsymbol{\pi}}_K, \lambda)', (\hat{\mathbf{z}}_K, z_{K+1})') - \ln L(\boldsymbol{\theta}_{\mathrm{D}}^{K}, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^{K}, \hat{\pi}_K, \hat{\mathbf{z}}_K)}{\lambda} \qquad (3.3)$$

According to the general mixture maximum likelihood theorem (Lindsay, 1983; Böhning, 1982), the NPMLE has been found if and only if $\Delta(z_{K+1}) \leq 0$ for all z_{K+1}. In practice, our stopping rule is that for a small λ the numerator of (3.3) is nonpositive for all locations z_{K+1} on a fine grid spanning a wide range of values.

The algorithm can then be outlined as follows. Initially, the likelihood is maximized for a single mass-point, giving starting values for the regression coefficients and measurement error variance, but not β since it would not be identified. The likelihood is then maximized for $K=2$ mass-points. After maximizing the likelihood for K mass-points, a further mass-point is introduced if a location z_{K+1} can be found at

which introduction of a very small new mass λ increases the likelihood when all other parameters are held constant, giving a positive numerator of (3.3). If such a location can be found, this implies that a larger maximum likelihood is achievable with an extra mass-point. A new point is therefore introduced and the likelihood maximized using as starting values the parameters of the previous model with a new mass at the location yielding the greatest increase in likelihood. A mass greater than λ is used as starting value to avoid numerical problems. This procedure is repeated until no location can be found at which introduction of a small mass increases the likelihood.

Our approach is similar to the algorithm proposed by Simar (1976) and adapted by Heckman and Singer (1984), Follmann and Lambert (1989) (in one dimension), Davies and Pickles (1987) (in two dimensions), among others. The EM algorithm has also been used for maximizing the likelihood for a given number of mass points (e.g., Hinde and Wood, 1987; Butler and Louis, 1992; Aitkin, 1996, 1999; Schafer, 2001). This approach and algorithms for finding the NPMLE (both the number of masses and parameter estimates) are described in Lindsay (1995) and Böhning (2000).

3.2 Prediction of true exposure

The unit specific exposure can be predicted using empirical Bayes, the expected value of exposure given the units' observed disease status and exposure measurements, with parameter estimates plugged in. For NPMLE, the posterior probability that the deviation of unit i's exposure from the mean exposure is $u_i = \hat{z}_k$ becomes

$$\mathrm{P}(u_i = \hat{z}_k | f_{ir}, d_i; \hat{\boldsymbol{\theta}}_{\mathrm{D}}^K, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K, \hat{\boldsymbol{\pi}}_K, \hat{\mathbf{z}}_K) = \frac{\hat{\pi}_k \mathrm{P}(d_i | u_i = \hat{z}_k; \hat{\boldsymbol{\theta}}_{\mathrm{D}}^K) \prod_{r=1}^{n_i} g(f_{ir} | u_i = \hat{z}_k; \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K)}{\sum_{k=1}^K \hat{\pi}_k \mathrm{P}(d_i | u_i = \hat{z}_k; \hat{\boldsymbol{\theta}}_{\mathrm{D}}^K) \prod_{r=1}^{n_i} g(f_{ir} | u_i = \hat{z}_k; \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K)} \tag{3.4}$$

and the posterior mean of the deviation is

$$\tilde{u}_i \equiv \mathrm{E}(u_i | f_{ir}, d_i; \hat{\boldsymbol{\theta}}_{\mathrm{D}}^K, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K, \hat{\boldsymbol{\pi}}_K, \hat{\mathbf{z}}_K) = \sum_{k=1}^K \hat{z}_k P(u_i = \hat{z}_k | f_{ir}, d_i, \hat{\boldsymbol{\theta}}_{\mathrm{D}}^K, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K, \hat{\boldsymbol{\pi}}_K, \hat{\mathbf{z}}_K) \tag{3.5}$$

For Gaussian quadrature, simply replace $\hat{z}_k$ by $\hat{\tau} v_k$ and $\hat{\pi}_k$ by known π_k in the above equations. It follows from the exposure model that the empirical Bayes prediction of true exposure becomes

$$E(F_i \mid f_{ir}, d_i; \hat{\boldsymbol{\theta}}_D^K, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K, \hat{\boldsymbol{\pi}}_K, \hat{\mathbf{z}}_K) = \hat{\gamma}_0 + \hat{\gamma}_1 x_i + \mathrm{E}(u_i \mid f_{ir}, d_i; \hat{\boldsymbol{\theta}}_{\mathrm{D}}^K, \hat{\boldsymbol{\theta}}_{\mathrm{M}}^K, \hat{\pi}_K, \hat{\mathbf{z}}_K) \tag{3.6}$$

When a normal exposure distribution is assumed, the empirical Bayes predictions are shrunken compared to the raw measurements (see, for example, Plummer and Clayton, 1993) and will have smaller variance than the prior distribution. If an identity link is used in the measurement model, the measurement errors can be predicted by subtracting the empirical Bayes predictions from the measured exposures.

Note that empirical Bayes predictions are not confined to the estimated locations $\hat{\mathbf{z}}_K$. This is natural since NPMLE is usually interpreted as a nonparameteric estimator of a possibly continuous exposure distribution. In contrast, if the NPMLE were viewed as an estimate of a truly discrete distribution, each unit should be allocated to one of the locations, typically that with the largest posterior probability.

Estimation and prediction for all models is carried out using `gllamm` (Rabe-Hesketh *et al.*, 2001a, b, 2002, 2003b), a general purpose program written in Stata® (StataCorp, 2003) for estimating 'generalized linear latent and mixed models' (e.g., Rabe-Hesketh *et al.*, 2003a; Skrondal and Rabe-Hesketh, 2004).

4 Analysis of heart disease data

We now estimate the effect of dietary fibre intake on heart disease. The dataset considered is on 337 middle-aged men, recruited between 1956 and 1966 and followed until 1976 (Morris *et al.*, 1977). There were two occupational groups, bank staff and London transport staff (drivers and conductors). At the time of recruitment, the men were asked to weigh their food over a seven-day period from which the total number of calories were derived as well as the amount of fat and fibre. Seventy-six bank staff had their diet measured in the same way again six months later. Coronary heart disease (CHD) was determined from personnel records and, after retirement, by direct communication with the retired men and by 'tagging' them at the Registrar General's Office. The latter ensured that deaths from CHD were recorded. However, the recording of nonfatal CHD after retirement was imperfect. There were 4601 man-years of observation and a total of 45 cases giving an overall incidence rate of 9.78 per 1000 man-years. Given the imperfect recording of CHD in later life, we will analyse merely the presence or absence of an event in the follow-up period using logistic regression rather than fitting a survival model.

Concerns arise in analysing the dataset using the standard model since the fibre measurements have a skewed distribution as shown in Figure 2. The model described in section 2 was estimated by maximum likelihood under normality as well as NPMLE. We considered fibre measurements on the original scale as well as log-transformed. Age, occupation and age by occupation were used as covariates. An indicator variable for the second measurement occasion was also included as a covariate in the measurement model to allow for a drift in the responses. The models assuming a normal exposure distribution were initially estimated with 20 quadrature points. Ten further points were repeatedly added until the resulting relative change in the maximized log-likelihood was less than 10^{-7}. After maximizing the likelihood for a given number of mass-points, a new mass with a log odds of -5 was moved in 1000 equal steps from the minimum to maximum deviation of measured fibre intake from its mean. If the increase in the log-likelihood exceeded 10^{-5} at any location a new mass was introduced. When no further points could be introduced, a smaller log odds of -7 was also tried.

Case (a) (nontransformed fibre measurements) required 50 points for Gaussian quadrature and 6 for NPMLE. Case (b) (log-transformed fibre measurements) required 60 points for quadrature and 8 for NPMLE.

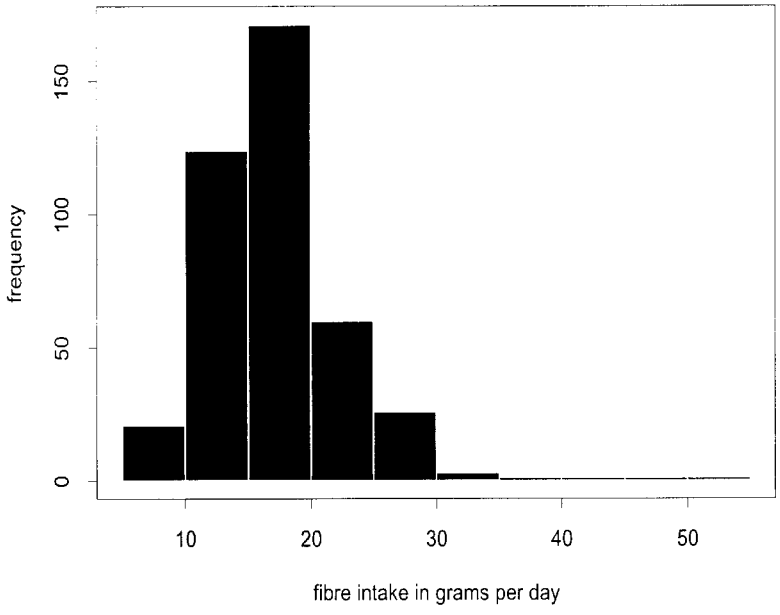

Figure 2 Histogram of fibre measurements

The estimates of the log odds ratio of heart disease β, the variance of true fibre intake conditional on covariates var(u_i), the measurement error variance σ_f^2, the reliability R conditional on covariates, and the locations and masses for NPMLE are given in Table 1. Both models suggest that increasing fibre intake reduces the odds of CHD. Ignoring the second fibre measurements and estimating β by ordinary logistic regression covarying for age, occupation and their interaction, gives estimates of -0.11 (0.04) for fibre and -1.54 (0.55) for log fibre, estimates that are attenuated relative to those of the full models as expected. The reliability estimates are larger using NPMLE than quadrature in both case (a) and (b), consistent with the findings of Hu *et al.* (1998).

The difference in log-likelihood and estimates of β between NPMLE and ML under normality was marked in case (a) but less so for (b). This may be because a normal

Table 1 Parameter estimates for heart disease data

	Case (a): Fibre		Case (b): Log fibre	
	Quadrature (60 points)	NPMLE (6 points)	Quadrature (90 points)	NPMLE (8 points)
β (SE)	-0.128 (0.051)	-0.146 (0.061)	-1.836 (0.740)	-1.753 (0.700)
Var (u_i) (SE)	23.655 (0.520)	24.926 (—)	0.068 (0.029)	0.072 (—)
σ_f^2 (SE)	6.948 (1.137)	6.132 (0.855)	0.022 (0.004)	0.018 (0.003)
R	0.773	0.803	0.759	0.796
Log-likelihood	-1372.354	-1319.787	-182.399	-172.255
Locations	-5.6, -3.0, 0.7, 7.6, 17.8, 32.6		-0.76, -0.52, -0.24, 0.04, 0.09, 0.37, 0.72, 1.10	
Masses (%)	1, 30, 43, 12, 1, 0.6		2, 4, 26, 40, 12, 15, 1, 0.6	

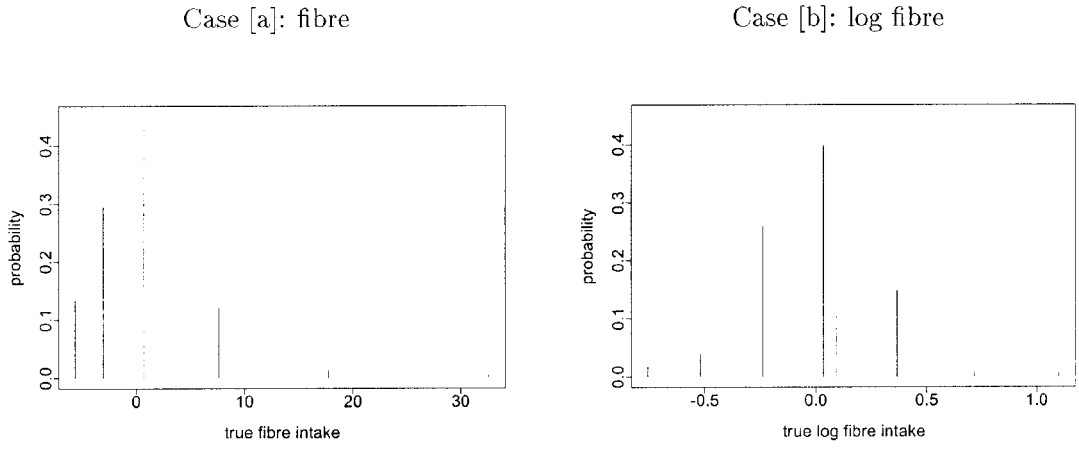

Figure 3 Estimated mass-point distributions using NPMLE for cases (a) and (b)

exposure distribution is less plausible for case (a) than (b) since the raw fibre measurements have a positively skewed distribution. As expected, the estimated mass-point distribution for case (a) is more skewed than for (b) as shown in Figure 3.

5 Simulation study

Parameters similar to the estimates in Table 1 were used to simulate measured exposure and disease status for the same number of units and repeated measurements of exposure as in the study data (see Table 2 for parameter values used). We carried out two sets of

Table 2 Simulation results for normal exposure distribution: mean and standard deviations (SD) of parameter estimates, mean standard error (SE), bias with 95% confidence interval and root mean square error (RMSE)

		Quadrature (60 points)					
Parameter	True value	Mean	SD	Mean SE	Bias	95% CI for bias	RMSE
β	−0.127	−0.127	0.041	0.041	0.000	(−0.009, 0.008)	0.041
var(u_i)	23.805	23.35	2.64	2.49	−0.46	(−0.99, 0.07)	2.67
σ_f^2	6.864	7.04	1.08	1.14	0.17	(−0.04, 0.39)	1.09
R	0.776	0.767	0.040	—	−0.009	(−0.017, −0.001)	0.041
		NPMLE					
Parameter	True value	Mean	SD	Mean SE	Bias	95% CI for bias	RMSE
β	−0.127	−0.125	0.041	0.041	0.002	(−0.006, 0.010)	0.040
var(u_i)	23.805	23.82	2.63	—	0.01	(−0.52, 0.54)	2.62
σ_f^2	6.864	6.59	1.11	0.99	−0.27	(−0.50, −0.05)	1.14
R	0.776	0.782	0.040	—	0.006	(−0.002, 0.014)	0.040

100 simulations. The first set used a normal true exposure distribution and the second set used a skewed distribution. For each simulated dataset, the parameters were estimated using both Gaussian quadrature and NPMLE. The empirical Bayes predictions of u_i were compared with the simulated 'true' values. Sixty quadrature points were used.

5.1 Normal true exposure distribution

The numbers of mass-points required for NPMLE were 4 (twice), 5 (23 times), 6 (48 times), 7 (16 times), 8 (9 times), 9 (once) and 11 (once). Table 2 shows the mean estimates and standard errors of β, var(u_i), σ_f^2 and R together with the standard deviation, bias and root mean square errors of these parameter estimates. Both methods give good results with no substantial bias in any of the parameter estimates for either method, except for a small negative bias in σ_f^2 for NPMLE. The root mean square errors are similar for both methods. The mean standard errors are close to the standard deviations of the parameter estimates for both quadrature and NPMLE. As previously noted with the real study data, NPMLE tended to attribute a smaller portion of the total variance to measurement error, giving larger reliability estimates. The finding that NPMLE works very well when the true distribution is normal is consistent with Schafer's (2001) results, but inconsistent with the results of Hu *et al.* (1998), possibly because Hu *et al.* used an approximate NPMLE method with a fixed set of locations.

Figure 4(a) shows a plot of the differences between the empirical Bayes predictions using quadrature and the true (simulated) values of u_i for the first five simulations and includes a reference line for zero difference. Lowess curves based on all 100 simulations are superimposed, representing the mean bias and the root mean square error. The latter was obtained by taking the square root of the lowess curve fitted to the squared differences between the estimated and true latent variables. For bias, the negative slope and zero intercept are consistent with shrinkage of the form $\tilde{u}_i = \lambda u_i$, $(\lambda < 1)$ since $E(\tilde{u}_i - u_i) = (\lambda - 1)u_i$. The root mean square error increases as the absolute value of u_i increases. The bias and root mean square error curves for NPMLE are identical to those using quadrature (not shown).

5.2 Skewed true exposure distribution

The exposure distribution was simulated by counting the number of 'thresholds' (0.2, 0.9, 1.3, 1.6, 1.8, 2.0, 2.2, 2.4, 2.7, 3.0) exceeded by a standard normal variable, mimicking commonly observed distributions of questionnaire sum scores; see Figure 4(b). The resulting discrete exposure variable was rescaled to have the same expectation and variance as the normal exposures in the previous simulation. The simulation results are summarized in Table 3. When the model is estimated by Gaussian quadrature, the regression coefficient β is underestimated, in absolute value, by about 20% (95% CI from 14 to 28%). The mean standard error of the exposure variance estimates is much lower than the standard deviation and root mean square error of these estimates (see Tables 3).

The NPMLE method required 5 masses (9 times), 6 masses (27 times) 7 masses (37 times), 8 masses (18 times), 9 masses (7 times) and 10 masses (2 times). Two of the NPMLE solutions gave outlying parameter estimates (large negative values of β) and were omitted from the table. There is a slight negative bias for the measurement error

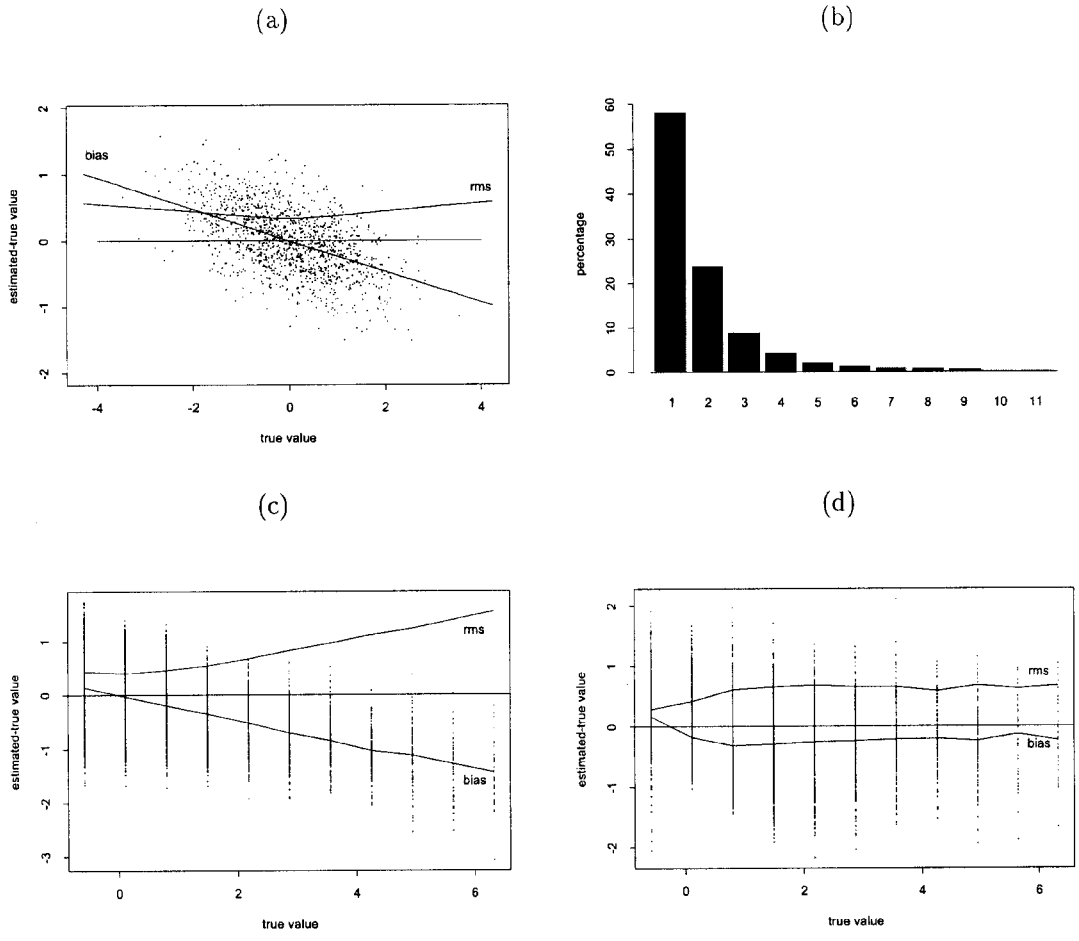

Figure 4 (a) Normal exposure distribution: comparison of realized values of u and empirical Bayes predictions using quadrature; (b) Skewed exposure distribution; (c) Skewed exposure distribution: comparison of realized values of u and empirical Bayes predictions using quadrature; (d) Skewed exposure distribution: comparison of realized values of u and empirical Bayes predictions using NPMLE

variance. Although there is no evidence of bias for β, the estimates have a greater root mean square error than those assuming a normal exposure distribution. Fortunately, the larger variance of the β estimates is correctly reflected in the mean standard error. The finding that the NPML estimates of β have a greater root mean square error than the estimates assuming a normal exposure distribution is consistent with Schafer (2001).

Plots comparing the predicted and true values of u_i are given in Figures 4(c) and (d) for Gaussian quadrature and NPMLE, respectively. In the case of quadrature, shrinkage is again apparent, leading to very large biases when the true exposure is large. In contrast, the bias and root mean square error for NPMLE are approximately constant (in absolute value) over the range of true values of u_i and much lower than for quadrature for large values of u_i.

Table 3 Simulation results for skewed exposure distribution: mean and standard deviations (SD) of parameter estimates, mean standard error (SE), bias with 95% confidence interval and root mean square error (RMSE)

		Quadrature (60 points)					
Parameter	True value	Mean	SD	Mean SE	Bias	95% CI for bias	RMSE
β	−0.127	−0.101	0.044	0.052	0.026	(0.018, 0.035)	0.051
var(u_i)	23.805	23.00	4.83	2.49	−0.80	(−1.78, 0.17)	4.88
σ_f^2	6.864	7.06	1.13	1.15	0.20	(−0.03, 0.43)	1.14
R	0.776	0.759	0.058	—	−0.017	(−0.029, −0.006)	0.060
		NPMLE					
Parameter	True value	Mean	SD	Mean SE	Bias	95% CI for bias	RMSE
β	−0.127	−0.143	0.079	0.083	−0.016	(−0.032, 0.003)	0.080
var(u_i)	23.805	23.63	4.71	—	−0.18	(−1.13, −0.78)	4.69
σ_f^2	6.864	6.42	0.98	0.91	−0.44	(−0.64, −0.24)	1.07
R	0.776	0.781	0.051	—	0.005	(−0.006, 0.015)	0.051

6 Discussion

We have considered generalized linear models with possibly non-normal exposures measured with error. Nonparametric maximum likelihood estimation (NPMLE) was developed for this setting and implemented in the `gllamm` software. For an application with skewed measures of exposure, NPMLE was used for estimating logistic regression, correcting for measurement error. Finally, a Monte Carlo experiment was conducted to investigate the performance of NPMLE.

Our simulation study suggests that NPML estimation yields unbiased estimates of the odds ratio and other parameters of interest except for a possible downward bias of the measurement error variance. When the true exposure distribution is normal, the NPML estimates do not appear to be less efficient than the quadrature estimates assuming a normal exposure distribution. When the true exposure distribution is highly skewed, assuming a normal exposure distribution gives biased estimates of the odds ratio for exposure whereas there is little evidence of bias for NPMLE. However, the NPML estimates unfortunately have a larger root mean square error. If predictions of true exposure are required, NPMLE is the preferred method since the predictions are substantially closer to the true values than those of the conventional model when the exposure distribution is skewed. When the true exposure distribution is normal, both methods provide nearly identical predictions.

An alternative to NPMLE would be to use smooth, flexible estimates of the mixing distribution (see, for example, Gallant and Nychka, 1987; Stefanski and Carroll, 1990; Zhang, 1990; Davidian and Gallant, 1992; Magder and Zeger, 1996). An advantage of these approaches is that the estimated mixing distribution may resemble the true (continuous) distribution more closely than the discrete mass-point distribution estimated by NPMLE. However, the smoothness of the estimated distribution often depends on an arbitrary parameter. Furthermore, it should be noted that despite the discreteness of NPMLE, the method does not assume a truly discrete distribution.

In an extensive simulation study comparing NPMLE with a mixture of normals, Magder and Zeger (1996) show that both methods give similar mean square errors of the fixed effects in linear mixed models with different true mixing distributions. Therefore if, as in the present study, interest is focused mostly on the fixed effects, NPMLE appears to be as suitable and arguably easier to estimate than, for example, a mixture of normals.

Our implementation of NPMLE performed well both in the application and simulations, typically requiring less than 15 Newton–Raphson iterations to maximize the likelihood for a given number of mass-points. This could be due to the likelihood being less flat than in mixture problems without replicates. Estimation of the conventional model assuming a normal exposure distribution required a large number of quadrature points. This is likely to be due to the integrands having sharp peaks that can easily fall between adjacent quadrature locations (e.g., Lesaffre and Spiessens, 2001). Adaptive quadrature (Naylor and Smith, 1982) can overcome these problems and has been implemented in `gllamm` (Rabe-Hesketh *et al.*, 2002, 2003b). However, we have only described ordinary quadrature in this article since it closely parallels NPMLE.

`gllamm` can also be used for the case where several exposures, for example fat and fibre, are subject to measurement error. When the joint exposure distribution is assumed to be multivariate normal, (adaptive) Gaussian product quadrature is used to approximate the marginal log-likelihood. NPMLE is in this case implemented by using multidimensional locations. Measurement error variances can be allowed to differ between methods, for example if a 'gold standard' is available in a validation sample. Different links and distributions can be specified for the response and measurement models, for example where measured exposures are a mixture of continuous and categorical variables. Currently, `gllamm` allows continuous, dichotomous, ordered and unordered categorical responses to be modelled, as well as rankings (Skrondal and Rabe-Hesketh, 2003), counts and continuous or discrete durations (Rabe-Hesketh *et al.*, 2001c), see also Skrondal and Rabe-Hesketh (2004) for application. The program can be downloaded from www.iop.kcl.ac.uk/IoP/Departments/BioComp/programs/gllamm.html.

Acknowledgements

We would like to thank Colin Taylor for help in the design of the estimation program and David Clayton for both invaluable discussion and for providing us with the example dietary data. We are also grateful for helpful comments from the reviewers and an associate editor. The work was partially supported by grant H519255050 (AP).

References

Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–62.

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–28.

Aitkin M, Rocci R (2002) A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, **12**, 163–74.

Bock RD, Lieberman M (1970) Fitting a response model for n dichotomously scored items. *Psychometrika*, **33**, 179–97.

Butler JS, Moffitt R (1982) A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, **50**, 761–64.

Butler SM, Louis TA (1992) Random effects models with nonparametric priors. *Statistics in Medicine*, **11**, 1981–2000.

Böhning D (1982) Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Annals of Statistics*, **10**, 1006–1008.

Böhning D (2000) *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others.* London: Chapman & Hall.

Carroll RJ, Ruppert D, Stefanski LA (1995) *Measurement Error in Nonlinear Models.* London: Chapman & Hall.

Clayton DG (1992) Models for the analysis of cohort and case–control studies with inaccurately measured exposures. In Dwyer JH, Feinlieb M, Lippert P, Hoffmeister H eds. *Statistical models for longitudinal studies on health.* New York: Oxford University Press.

Davidian M, Gallant AR (1992) Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quindine. *Journal of Pharmacokinetics and Biopharmaceutics*, **20**, 529–56.

Davies R, Pickles A (1987) A joint trip timing store-type choice model for grocery shopping, including inventory effects and nonparametric control for omitted variables. *Transportation Research, A*, **21**, 345–61.

Dunn G (1989) *Design and analysis of reliability studies: statistical evaluation of measurement errors.* Sevenoaks: Edward Arnold.

Follmann DA, Lambert D (1989) Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, **84**, 295–300.

Gallant AR, Nychka DW (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica*, **55**, 363–90.

Heckman J, Singer B (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.

Hinde JP, Wood ATA (1987) Binomial variance component models with a non-parametric assumption concerning random effects. In Crouchley R ed. *Longitudinal data analysis.* Aldershot: Avebury.

Hu P, Tsiatis AA, Davidian M (1998) Estimating the parameters in the Cox model when the covariate variables are measured with error. *Biometrics*, **54**, 1407–1419.

Jewell NP (1982) Mixtures of exponential distributions. *Annals of Statistics*, **10**, 479–84.

Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805–11.

Lesaffre E, Spiessens B (2001) On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics*, **50**, 325–35.

Lindsay BG (1983) The geometry of mixture likelihoods. Part I: a general theory. *Annals of Statistics*, **11**, 783–92.

Lindsay BG (1995) *Mixture models: theory, geometry and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Hayward, CA: Institute of Mathematical Statistics.

Lindsay BG, Clogg CC, Grego J (1991) Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, **86**, 96–107.

Magder SM, Zeger SL (1996) A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, **11**, 1141–51.

Morris JN, Marr JW, Clayton DG (1977) Diet and heart: postscript. *British Medical Journal*, **2**, 1307–14.

Naylor JC, Smith, AFM (1982) Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214–25.

Plummer M, Clayton D (1993) Measurement error in dietary assessment: an investigation using covariance structure models. Part II. *Statistics in Medicine*, **12**, 937–48.

Rabe-Hesketh S, Skrondal A (2001) Parameterization of multivariate random effects models for categorical data. *Biometrics*, **57**, 1256–64.

Rabe-Hesketh S, Pickles A, Skrondal A (2001a) GLLAMM: A general class of multilevel models and a Stata program. *Multilevel Modelling Newsletter*, **13**, 17–23.

Rabe-Hesketh S, Pickles A, Skrondal A (2001b) *GLLAMM Manual.* Technical report. 2001/01. Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London. Downloadable from http://www.iop.kcl.ac.uk/iop/departments/biocomp/programs/gllamm.html (accessed 20 May 2003).

Rabe-Hesketh S, Yang S, Pickles A (2001c) Multilevel models for censored and latent responses. *Statistical Methods in Medical Research*, **10**, 409–27.

Rabe-Hesketh S, Skrondal A, Pickles A (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, **2**, 1–21.

Rabe-Hesketh S, Skrondal A, Pickles A (2003a) Generalized multilevel structural equation modeling. *Psychometrika*, in press.

Rabe-Hesketh S, Skrondal A, Pickles A (2003b) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Submitted for publication.*

Reiersøl O (1950) Identifiability of a linear relation between variables which are subject to error. *Econometrica*, **18**, 375–89.

Roeder K, Carroll RJ, Lindsay BG (1996) A semiparametric mixture approach to case–control studies with errors in covariables. *Journal of the American Statistical Association*, **91**, 722–32.

Rosner B, Spiegelman D, Willett WC (1990) Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, **132**, 734–45.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581–92.

Schafer DW (2001) Semiparametric maximum likelihood for measurement error regression. *Biometrics*, **57**, 53–61.

Simar L (1976) Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics*, **4**, 1200–209.

Skrondal A, Rabe-Hesketh S (2003) Multilevel logistic regression for polytomous data and rankings. *Psychometrika*, **68**, 267–87.

Skrondal A, Rabe-Hesketh S (2004) *Generalized latent variable modeling: multilevel, longitudinal and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

StataCorp (2003) *Stata Statistical Software: Release 8.0.* College Station, TX: Stata Corporation.

Stefanski L, Carroll RJ (1990) Deconvoluting kernel density estimators. *Statistics,* **21**, 169–84.

Thoresen M, Laake P (2000) A simulation study of measurement error correction methods in logistic regression. *Biometrics*, **56**, 868–72.

Zhang C (1990) Fourier methods for estimating mixing densities and distributions. *Annals of Statistics*, **18**, 806–31.