Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2009)
172, *Part* 3, *pp.* 659–687

Prediction in multilevel generalized linear models

Anders Skrondal

Norwegian Institute of Public Health, Oslo, Norway

and Sophia Rabe-Hesketh

University of California, Berkeley, USA, and Institute of Education, London, UK

[Received February 2008. Final revision October 2008]

Summary. We discuss prediction of random effects and of expected responses in multilevel generalized linear models. Prediction of random effects is useful for instance in small area estimation and disease mapping, effectiveness studies and model diagnostics. Prediction of expected responses is useful for planning, model interpretation and diagnostics. For prediction of random effects, we concentrate on empirical Bayes prediction and discuss three different kinds of standard errors; the posterior standard deviation and the marginal prediction error standard deviation (comparative standard errors) and the marginal sampling standard deviation (diagnostic standard error). Analytical expressions are available only for linear models and are provided in an appendix. For other multilevel generalized linear models we present approximations and suggest using parametric bootstrapping to obtain standard errors. We also discuss prediction of expectations of responses or probabilities for a new unit in a hypothetical cluster, or in a new (randomly sampled) cluster or in an existing cluster. The methods are implemented in gllamm and illustrated by applying them to survey data on reading proficiency of children nested in schools. Simulations are used to assess the performance of various predictions and associated standard errors for logistic random-intercept models under a range of conditions.

Keywords: Adaptive quadrature; Best linear unbiased predictor (BLUP); Comparative standard error; Diagnostic standard error; Empirical Bayes; Generalized linear mixed model; gllamm; Mean-squared error of prediction; Multilevel model; Posterior; Prediction; Random effects; Scoring

1. Introduction

Multilevel generalized linear models are generalized linear models that contain multivariate normal random effects in the linear predictor. Such models are also known as hierarchical generalized linear models or generalized linear mixed (effects) models. A common special case is multilevel linear models for continuous responses. The random effects represent unobserved heterogeneity and induce dependence between units nested in clusters. In this paper we discuss prediction of random effects and expected responses, including probabilities, for multilevel generalized linear models.

There are several reasons why we may want to assign values to the random effects for individual clusters. Predicted random effects can be used for inference regarding particular clusters, e.g. to assess the effectiveness of schools or hospitals (e.g. Raudenbush and Willms (1995) and Goldstein and Spiegelhalter (1996)) and in small area estimation or disease mapping (e.g. Rao

Address for correspondence: Anders Skrondal, Division of Epidemiology, Norwegian Institute of Public Health, PO Box 4404 Nydalen, N-0403 Oslo, Norway.
E-mail: anders.skrondal@fhi.no

 0964–1998/09/172659

(2003)). Another important application is in model diagnostics, such as checking for violations of the normality assumption for the random effects (e.g. Lange and Ryan (1989)) or finding outlying clusters (e.g. Langford and Lewis (1998)).

There is a large literature on prediction of random effects and responses in multilevel *linear* models. Contributions from a frequentist stance include Swamy (1970), Rosenberg (1973), Rao (1975), Harville (1976), Ware and Wu (1981), Strenio *et al.* (1983), Kackar and Harville (1984), Reinsel (1984, 1985), Bondeson (1990), Candel (2004, 2007), Afshartous and de Leeuw (2005) and Frees and Kim (2006). References with a Bayesian perspective include Lindley and Smith (1972), Smith (1973), Fearn (1975) and Strenio *et al.* (1983). There are also relevant sections in the books by Searle *et al.* (1992), Vonesh and Chincilli (1997), Demidenko (2004), Jiang (2007) and McCulloch *et al.* (2008). A limitation of much of this work is a failure clearly to delineate different notions of uncertainty regarding predictions and to discuss which are appropriate for various purposes. Notable exceptions include Laird and Ware (1982) and in particular Goldstein (1995, 2003).

Compared with the linear case, there are few contributions regarding prediction of random effects in multilevel generalized linear models with other links than the identity. The reason may be that this case is considerably more challenging since results cannot be derived by matrix algebra and expressed in closed form. Insights from the literature on prediction of latent variables in the closely related item response models are hence useful. In this paper we briefly review various approaches to assigning values to random effects in multilevel generalized linear models, present different standard errors for empirical Bayes predictions of random effects and discuss the purposes for which each standard error is appropriate. We recommend using the posterior standard deviation as standard error for inferences regarding the random effects of specific clusters. We also suggest computationally efficient approximations for standard errors of empirical Bayes predictions in non-linear multilevel models as well as a computationally intensive parametric bootstrapping approach.

Predictions of expected responses, or response probabilities, are also often required. These are useful for interpreting and visualizing estimates for multilevel models using graphs. For example, in logistic regression models, the regression coefficients can be difficult to interpret, and we may want to explore the 'effects' of covariates on predicted probabilities. Furthermore, planning may require predictions of the responses of new units in existing clusters or in new clusters. For example, a credit card holder may apply for an extended limit on his credit card. In this case the financial institution may want to predict the probability that the applicant will default on his payment on the basis of his payment history. Regarding prediction of expected responses with non-linear link functions, we are not aware of any work apart from a few contributions in the literature on small area estimation (e.g. Farrell *et al.* (1997) and Jiang and Lahiri (2001)), a theoretical paper by Vidoni (2006) and some applied papers (e.g. Rose *et al.* (2006)). We point out that it is important to distinguish between different kinds of predictions, for instance whether a prediction concerns a new unit in a hypothetical cluster, or in a randomly sampled new cluster or in an existing cluster.

The plan of this paper is as follows. We start by introducing multilevel linear and generalized linear models in Section 2. In Section 3 we estimate a random-intercept model to investigate the contextual effect of socio-economic status (SES) on reading proficiency by using data from the 'Program for international student assessment' (PISA). We then discuss prediction of random effects in Section 4 and different kinds of standard errors that are associated with such predictions in Section 5. These methods are applied to the PISA data in Section 6. In Section 7 we describe prediction of different kinds of expected responses and their uncertainty and apply the methods to the PISA data in Section 8. In Section 9 we investigate the performance of some of

the proposed methods using Monte Carlo simulations. Finally, we close the paper with some concluding remarks.

The PISA data and the Stata 'do file' to perform the analysis that are presented in the paper can be obtained from

`http://www.blackwellpublishing.com/rss/SeriesA.htm`

2. Multilevel linear and generalized linear models

We restrict discussion to two-level models because the notation becomes unwieldy for higher level models. However, the ideas that are presented here can be extended to models with more than two levels. It is useful to introduce multilevel linear models briefly before discussing the generalized linear counterparts.

2.1. Multilevel linear models

For the response y_{ij} of unit i in cluster j, the two-level linear model can be expressed as

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\boldsymbol{\zeta}_j + \varepsilon_{ij},$$

where $\mathbf{x}_{ij}$ are covariates with fixed coefficients $\boldsymbol{\beta}$, $\mathbf{z}_{ij}$ are covariates with random effects $\boldsymbol{\zeta}_j$ and ε_{ij} are level 1 errors.

It is useful to write the model for all n_j responses $\mathbf{y}_j$ for cluster j as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\zeta}_j + \boldsymbol{\varepsilon}_j, \tag{1}$$

where $\mathbf{X}_j$ is an $n_j \times P$ matrix with rows $\mathbf{x}'_{ij}$, $\mathbf{Z}_j$ an $n_j \times Q$ matrix with rows $\mathbf{z}'_{ij}$ and $\boldsymbol{\varepsilon}_j = (\varepsilon_{1j}, \ldots, \varepsilon_{n_j j})'$. We allow the covariates $\mathbf{X}_j$ and $\mathbf{Z}_j$ to be random and assume that they are strictly exogenous (e.g. Chamberlain (1984)) in the sense that $E(\varepsilon_{ij}|\boldsymbol{\zeta}_j, \mathbf{X}_j, \mathbf{Z}_j) = E(\varepsilon_{ij}|\boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = E(\varepsilon_{ij}) = 0$, and $E(\boldsymbol{\zeta}_j|\mathbf{X}_j, \mathbf{Z}_j) = E(\boldsymbol{\zeta}_j) = \mathbf{0}$.

The random effects and level 1 errors are assumed to have multivariate normal distributions $\boldsymbol{\zeta}_j|\mathbf{X}_j, \mathbf{Z}_j \sim N(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\varepsilon}_j|\boldsymbol{\zeta}_j, \mathbf{X}_j, \mathbf{Z}_j \sim N(\mathbf{0}, \boldsymbol{\Theta}_j)$, both independent across clusters given the covariates. It is furthermore usually assumed that $\boldsymbol{\Theta}_j = \theta\mathbf{I}_{n_j}$. In this case, the responses for units i in cluster j are conditionally independent, given the covariates and random effects, and have constant variance θ.

For simplicity we shall sometimes consider the special case of a linear random-intercept model

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \zeta_j + \varepsilon_{ij},$$

where ζ_j is a cluster-specific deviation from the mean intercept β_0.

2.2. Multilevel generalized linear models

A two-level generalized linear model can be written as

$$h^{-1}\{E(y_{ij}|\boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij})\} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\boldsymbol{\zeta}_j \equiv \eta_{ij},$$

where $h^{-1}(\cdot)$ is a link function and η_{ij} is the linear predictor ('$\equiv$' denotes a definition). In other words, the conditional expectation of the response, given the covariates and random effects, is

$$\mu_{ij} \equiv E(y_{ij}|\boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = h(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\boldsymbol{\zeta}_j) = h(\eta_{ij}).$$

As for linear models, it is assumed that the random effects are multivariate normal and that the covariates are strictly exogenous. The responses are assumed to be conditionally independent,

given the covariates and random effects, and have conditional distributions from the exponential family. For this family of distributions, the conditional variance is given by

$$\mathrm{var}(y_{ij}|\mu_{ij}) = \phi_{ij} V(\mu_{ij}),$$

where ϕ_{ij} is a dispersion parameter and $V(\mu_{ij})$ is a variance function specifying the relationship between conditional variance and conditional expectation.

The multilevel linear model results when an identity link is specified, $\mu_{ij} = \eta_{ij}$, combined with a conditional normal distribution for the response $y_{ij}|\mu_{ij} \sim N(\mu_{ij}, \theta)$. In this case, the variance function is 1 and the dispersion parameter is a free parameter $\phi_{ij} = \theta$. Another important special case is a logistic regression model for dichotomous responses which combines a logit link, $\mathrm{logit}(\mu_{ij}) \equiv \log\{\mu_{ij}/(1-\mu_{ij})\} = \eta_{ij}$, with a conditional Bernoulli distribution for the response, $y_{ij}|\mu_{ij} \sim \mathrm{Bernoulli}(\mu_{ij})$. The variance function is now $V(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$ and the dispersion parameter is 1 (e.g. Skrondal and Rabe-Hesketh (2007a)).

We refer to Rabe-Hesketh and Skrondal (2008a) for a comprehensive discussion of multilevel generalized linear models.

2.2.1. *Relationship with item response and common factor models*

Item response models and common factor models can be written as

$$h^{-1}\{E(\mathbf{y}_j|\boldsymbol{\zeta}_j, \mathbf{X}_j)\} = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{\zeta}_j,$$

where the 'random effects' $\boldsymbol{\zeta}_j$ are called latent variables, common factors or latent traits, units i correspond to 'items' and clusters j correspond to subjects. The identity link produces common factor models (e.g. Lawley and Maxwell (1971)) and logit and probit links yield categorical factor models (e.g. Mislevy (1986)) or item response models (e.g. Embretson and Reise (2000)).

Note that the structure of these models is very similar to two-level generalized linear models. The difference is that the unknown parameter matrix $\boldsymbol{\Lambda}$ replaces the known cluster-specific covariate matrix $\mathbf{Z}_j$. Usually, but not necessarily, $\mathbf{X}_j\boldsymbol{\beta}$ is also replaced by intercepts $\mathbf{I}\boldsymbol{\beta} = \boldsymbol{\beta}$. Since parameters are usually treated as known when making predictions, the distinction between variables $\mathbf{Z}_j$ and parameters $\boldsymbol{\Lambda}$ becomes irrelevant.

See Skrondal and Rabe-Hesketh (2007b) for a recent review discussing the relationships between these and other models.

2.3. *Marginal likelihood*

Letting $\boldsymbol{\vartheta}$ denote the model parameters, the likelihood contribution for cluster j, $l_j(\boldsymbol{\vartheta}) \equiv g(\mathbf{y}_j|\mathbf{X}_j, \mathbf{Z}_j; \boldsymbol{\vartheta})$, becomes

$$l_j(\boldsymbol{\vartheta}) = \int \varphi(\boldsymbol{\zeta}_j; \boldsymbol{\Psi})\, f(\mathbf{y}_j|\boldsymbol{\zeta}_j, \mathbf{X}_j, \mathbf{Z}_j; \boldsymbol{\vartheta}^f)\, \mathrm{d}\boldsymbol{\zeta}_j = \int \varphi(\boldsymbol{\zeta}_j; \boldsymbol{\Psi}) \prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \boldsymbol{\vartheta}^f)\, \mathrm{d}\boldsymbol{\zeta}_j.$$

The first term in the integral is the random-effects density (multivariate normal with zero means and covariance matrix $\boldsymbol{\Psi}$) and the second term is the conditional density (or probability) of the responses given the random effects and covariates. We use the notation $\boldsymbol{\vartheta}^f$ to denote the vector of parameters appearing in the conditional response distribution, so that $\boldsymbol{\vartheta}$ consists of $\boldsymbol{\vartheta}^f$ and the unique elements of $\boldsymbol{\Psi}$. Since the clusters are assumed to be independent, the likelihood for the sample is $l(\boldsymbol{\vartheta}) = \Pi_{j=1}^{J} l_j(\boldsymbol{\vartheta})$.

Except for the case of multilevel linear models, the integrals usually do not have analytic solutions and must be evaluated numerically, typically by adaptive quadrature (e.g. Pinheiro and

Bates (1995) and Rabe-Hesketh *et al.* (2005)) or by Monte Carlo integration (e.g. McCulloch (1997)). Alternatives to maximum likelihood that do not require integration include penalized quasi-likelihood (e.g. Breslow and Clayton (1993)) and Markov chain Monte Carlo sampling (e.g. Clayton (1996)).

3. Application: contextual effect of socio-economic status on reading proficiency

It has been found in a large number of studies that various measures of the social composition of schools affect student achievement beyond the individual effects of student background characteristics (see Rumberger and Palardy (2005) for a recent literature review). In particular, it has been found that there is considerable variability in school mean SES in the UK and the USA and that school mean SES has a large effect on student achievement after controlling for individual SES (e.g. Willms (1986), Raudenbush and Bryk (2002), pages 135–141, and Rumberger and Palardy (2005)). Such findings have led to calls for comprehensive schooling or desegregation policies to narrow the gap in achievement between high and low SES students.

Here we shall estimate the *contextual* effects of SES on reading proficiency. We use the US sample from the PISA from 2000, an international educational survey funded by the Organisation for Economic Co-operation and Development that assesses reading and mathematical and scientific literacy among 15-year-old students (see `http://www.pisa.oecd.org`).

We define reading proficiency as achieving at least the second highest of five reading proficiency levels as defined in the PISA manual (Organisation for Economic Co-operation and Development, 2000). The motivation for this is that it is often easier to interpret changes in the proportion of children who are proficient than changes in mean reading scores. To derive the binary proficiency variable, we applied a threshold of 552.89 to the weighted maximum likelihood estimates (Warm, 1989) of reading ability derived from a partial credit item response model (see Adams (2002) for details). As a measure of SES, we use the international socio-economic index as defined in Ganzeboom *et al.* (1992).

We let the reading proficiency and SES of student i in school j be denoted y_{ij} and x_{ij} respectively and consider the random-intercept logistic regression model

$$\begin{aligned}\text{logit}\{\Pr(y_{ij}=1|\mathbf{x}_{ij},\zeta_j)\} &= \beta_0+\beta_1(x_{ij}-\bar{x}_{.j})+\beta_2\bar{x}_{.j}+\zeta_j \\ &= \beta_0+\beta_1 x_{ij}+(\beta_2-\beta_1)\bar{x}_{.j}+\zeta_j, \qquad \zeta_j|x_{ij}\sim N(0,\psi),\end{aligned}$$

where $\bar{x}_{.j}$ is the school mean SES and ζ_j is a school-specific random intercept. In this model, β_1 represents the within-school effect of SES and β_2 represents the between-school effect. The difference, $\beta_2-\beta_1$, represents the contextual effect: the additional effect of school mean SES on proficiency that is not accounted for by individual level SES. In research on school effects, the term *contextual effects* is often taken to refer to the effects of the 'hardware' of the school, such as location and resources, student body and teacher body, and not the 'software' of the school or *climate* (Ma *et al.*, 2008). However, the estimate of the 'contextual effect' $\beta_2-\beta_1$ will partially encompass the effects of all school level variables that are correlated with SES including school climate.

In the PISA data used here, there are 2069 students from 148 schools with between one and 28 students per school. The sample mean SES is 46.8. The sample standard deviation of individual SES is 17.6, the sample standard deviation of school mean SES (using one observation per school) is 9.0 and the sample standard deviation of the school mean-centred SES is 15.4. Thus, there is considerable socio-economic segregation between schools.

Maximum likelihood estimates of the model parameters and their standard errors are given in Table 1. These estimates were obtained using `gllamm` (e.g. Rabe-Hesketh and Skrondal (2008b))

Table 1. Maximum likelihood estimates for the random-intercept logistic regression model for the PISA data

Parameter	*Covariate*	*Estimate*	*Standard error*	*OR*	*95% confidence interval*
β_0		−4.785	0.427		
$10\beta_1$	$[(x_{ij}-\bar{x}_{.j})/10]$	0.184	0.031	1.2	(1.1,1.3)
$10\beta_2$	$[\bar{x}_{.j}/10]$	0.891	0.088	2.4	(2.1,2.9)
$10(\beta_1-\beta_2)$		0.707	0.092	2.0	(1.7,2.4)
ψ		0.280			

in Stata with 20-point adaptive quadrature. (For simplicity, we have ignored sampling weights here and refer to Rabe-Hesketh and Skrondal (2006) for pseudo-maximum-likelihood estimation taking the complex survey design of the PISA study into account.) Since the regression coefficients represent changes in the log-odds (logits), their exponentials represent odds ratios. The estimated odds ratios (ORs) are also given in Table 1 together with their approximate 95% confidence intervals. For a given school mean SES, every 10-unit increase in individual SES is associated with an estimated 20% increase in the odds of proficiency (within effect). The estimated odds ratio per 10-unit increase in school mean SES, for students whose individual SES equals the school mean, is 2.4 (between effect). The estimated odds double for every 10-unit increase in school mean SES for students with a given individual SES. This contextual effect is highly statistically significant ($z = 7.7$; $p < 0.001$) and may be due to direct peer influences, school climate, allocation of resources and organizational and structural features of schools.

ORs are difficult to interpret because they express multiplicative effects rather than additive effects and because odds are less familiar than proportions and probabilities. In Section 8 we therefore produce graphs of predicted probabilities to convey the magnitude of the estimated contextual, within-school and between-school effects of SES.

4. Prediction of random effects

We now discuss how to assign values to the random effects $\boldsymbol{\zeta}_j = (\zeta_{1j}, \ldots, \zeta_{Qj})'$ for individual clusters $j = 1, \ldots, J$. This assignment usually proceeds after the model parameters have been estimated, with the estimates $\hat{\boldsymbol{\vartheta}}$ treated as known parameters. When the model parameters are treated as known, the problem of assigning values to random effects can be approached from at least four different philosophical perspectives which we refer to as Bayesian, empirical Bayesian, frequentist prediction and frequentist estimation.

In the Bayesian approach, inference regarding $\boldsymbol{\zeta}_j$ for cluster j is based on the posterior distribution of $\boldsymbol{\zeta}_j$ given the known data for the cluster which are treated as observed values of random variables. However, some Bayesians also consider hypothetical replications of the data to validate Bayesian probability statements, which is referred to by Rubin (1984) as frequency calculations. Similarly, empirical Bayesians evaluate inferences with respect to joint sampling of $\boldsymbol{\zeta}_j$ and $\mathbf{y}_j$ (e.g. Morris (1983)). Robinson (1991) pointed out that this sampling model is also relevant for classical (i.e. frequentist) inference if the problem is viewed as assigning a value to the realization of a random variable. In this case, the random-effects distribution is viewed as representing the *variation* of $\boldsymbol{\zeta}_j$ (in the population), whereas Bayesians would view this prior distribution as representing *uncertainty* regarding $\boldsymbol{\zeta}_j$. Searle *et al.* (1992) also viewed the target of

inference as the unobserved realization of a random variable and used the word *prediction* to distinguish their approach from frequentist *estimation*. In frequentist estimation, $\boldsymbol{\zeta}_j$ are treated as fixed parameters, with only the responses viewed as random in the sampling model. In this case, inference regarding the random effects typically proceeds by maximum likelihood estimation.

In Sections 4.1 and 4.2, we use mostly Bayesian and empirical Bayesian reasoning, but it is useful to keep in mind that the difference from frequentist prediction is largely semantic (remembering that the model parameters are treated as known). We use the term *prediction* to avoid any confusion with frequentist estimation which is briefly described in Section 4.3.1.

4.1. Empirical posterior distribution

With the model parameters treated as known and equal to their maximum likelihood estimates $\hat{\boldsymbol{\vartheta}}$ we have two sources of information concerning the random effects. The first piece of information is the prior distribution $\varphi(\boldsymbol{\zeta}_j; \hat{\boldsymbol{\Psi}})$ of the random effects, representing our *a priori* knowledge about the random effects before 'seeing' the data for cluster j. The second piece of information is the data $\mathbf{y}_j$, $\mathbf{X}_j$ and $\mathbf{Z}_j$ for cluster j.

A natural way of combining the sources of information regarding the random effects is through the posterior distribution $\omega(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})$ of $\boldsymbol{\zeta}_j$, the distribution of $\boldsymbol{\zeta}_j$ updated with or *given* the data $\mathbf{y}_j$, $\mathbf{X}_j$ and $\mathbf{Z}_j$. Using Bayes theorem, we obtain

$$\omega(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \frac{\varphi(\boldsymbol{\zeta}_j; \hat{\boldsymbol{\Psi}})\ f(\mathbf{y}_j|\boldsymbol{\zeta}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}^f)}{g(\mathbf{y}_j|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})}.$$

The denominator is just the likelihood contribution $l_j(\hat{\boldsymbol{\vartheta}})$ of the jth cluster and usually does not have a closed form but can be evaluated numerically. Here the parameters are treated as known and equal to their estimates, so the posterior distribution is 'empirical' or 'estimated' (e.g. Carlin and Louis (2000a), page 58). In a fully Bayesian approach, prior distributions would be specified for the model parameters, and the posterior distribution of the random effects would be marginal with respect to these parameters. It should be noted that the estimated posterior distribution can also be derived from a frequentist perspective by treating $\boldsymbol{\zeta}_j$ as unobservable random variables and conditioning on the observed responses $\mathbf{y}_j$ (as well as $\mathbf{X}_j$ and $\mathbf{Z}_j$).

For linear models it follows from standard results on conditional multivariate normal densities that the posterior density is multivariate normal. For other response types, it follows from the Bayesian central limit theorem (e.g. Carlin and Louis (2000a), pages 122–124) that the posterior density tends to multivariate normality as the number of units n_j in the cluster increases (see Chang and Stout (1993) for asymptotic normality in binary response models).

4.2. Empirical Bayes prediction of the random effects

Empirical Bayes prediction is undoubtedly the most widely used method for assigning values to random effects. Empirical Bayes predictors (see Efron and Morris (1973, 1975), Morris (1983), Maritz and Lwin (1989) and Carlin and Louis (2000a,b)) of the random effects $\boldsymbol{\zeta}_j$ are the means of the empirical posterior distribution (with parameter estimates $\hat{\boldsymbol{\vartheta}}$ plugged in):

$$\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}} = E(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \int \boldsymbol{\zeta}_j\, \omega(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\, \mathrm{d}\boldsymbol{\zeta}_j. \tag{2}$$

Whenever the prior distribution is parametric, the predictor is denoted parametric empirical Bayes. Empirical Bayes prediction is usually referred to as 'expected *a posteriori*' (EAP) estimation in item response models (e.g. Bock and Aitkin (1981)) and as the 'regression method'

(e.g. Thurstone (1935) and Thomson (1938)) for factor scoring in factor analysis. The reason for the term 'empirical Bayes', which was coined by Robbins (1955), is that Bayesian principles are adapted to a frequentist setting by plugging in estimated model parameters. True Bayesians would obtain the posterior distribution of the random effects, assuming a prior distribution for $\boldsymbol{\vartheta}$, instead of simply plugging in estimates $\hat{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta}$.

The empirical Bayes predictor can be justified by considering the quadratic loss function

$$L^{\mathrm{EB}}(\tilde{\boldsymbol{\zeta}}_j, \boldsymbol{\zeta}_j) = (\tilde{\boldsymbol{\zeta}}_j - \boldsymbol{\zeta}_j)' \mathbf{W} (\tilde{\boldsymbol{\zeta}}_j - \boldsymbol{\zeta}_j),$$

where $\mathbf{W}$ is some arbitrary (usually symmetric) positive definite weight matrix. Treating the parameters as known, the empirical Bayes predictor minimizes the (estimated) posterior risk defined as the posterior expectation of the quadratic loss

$$R(\tilde{\boldsymbol{\zeta}}_j, \boldsymbol{\zeta}_j) = \int L^{\mathrm{EB}}(\tilde{\boldsymbol{\zeta}}_j, \boldsymbol{\zeta}_j)\, \omega(\boldsymbol{\zeta}_j | \mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\, \mathrm{d}\boldsymbol{\zeta}_j \tag{3}$$

(see proposition 5.2.(i) of Bernardo and Smith (1994)). In other words, the empirical Bayes predictor minimizes the posterior mean-squared error of prediction, given the responses and covariates.

The empirical Bayes predictor also minimizes the mean-squared error of prediction (MSEP) over the joint distribution of the random effects and the responses, giving it a frequentist motivation as the 'best predictor' (e.g. Searle *et al.* (1992), pages 261–262). The MSEP is the expectation of the posterior risk with respect to the distribution of $\mathbf{y}_j$ and is also called the empirical Bayes risk, Bayes risk or preposterior risk since this is the posterior loss one expects before having seen the data (Carlin and Louis (2000a), pages 332–334).

Apart from linear models, it is in general impossible to obtain empirical Bayes predictions by analytical integration, and numerical or simulation-based integration methods must be used. Note that empirical Bayes predictions are a by-product of maximum likelihood estimation of model parameters in the implementation of adaptive quadrature that was suggested by Rabe-Hesketh *et al.* (2005).

In a linear random-intercept model, the empirical Bayes predictor is

$$\tilde{\zeta}_j^{\mathrm{EB}} = \hat{R}_j \left\{ \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - \mathbf{x}_{ij}' \hat{\boldsymbol{\beta}}) \right\}, \tag{4}$$

where

$$0 < \hat{R}_j \equiv \frac{\hat{\psi}}{\hat{\psi} + \hat{\theta}/n_j} < 1.$$

The term in curly brackets in equation (4) is the mean 'raw' or total residual for cluster j, which is sometimes called the 'ordinary least squares estimator' or maximum likelihood estimator of ζ_j (see Section 4.3.1). $\hat{R}_j$ is a *shrinkage* factor which pulls the empirical Bayes prediction towards 0, the mean of the prior distribution. The shrinkage factor can be interpreted as the estimated reliability of the mean raw residual as a 'measurement' of ζ_j (the variance of the 'true score' divided by the total variance). The reliability decreases when n_j decreases and when $\hat{\theta}$ increases compared with $\hat{\psi}$; the conditional density of the responses $\prod_{i=1}^{n_j} f(y_{ij} | \zeta_j, \mathbf{x}_{ij}; \hat{\boldsymbol{\vartheta}})$ then becomes flat and uninformative compared with the prior density $\varphi(\zeta_j; \hat{\psi})$.

For a linear random-intercept model the conditional expectation of the empirical Bayes predictor, given the random intercept, is

$$E_{\mathbf{y}}(\tilde{\zeta}_j^{\mathrm{EB}} | \zeta_j, \mathbf{X}_j; \hat{\boldsymbol{\vartheta}}) = \hat{R}_j \zeta_j.$$

The conditional bias $(\hat{R}_j - 1)\zeta_j$ is 'inward' or towards zero. Such inward bias is also found in logistic and probit random-intercept models (e.g. Hoijtink and Boomsma (1995)). In all multilevel generalized linear models, the empirical Bayes predictor is unconditionally unbiased since $E_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = E_{\mathbf{y}}\{E(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\} = E(\boldsymbol{\zeta}_j|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \mathbf{0}$.

For linear models, the posterior mean (assuming known model parameters) is the best linear unbiased predictor (BLUP) (e.g. Goldberger (1962) and Robinson (1991)) because it is linear in $\mathbf{y}_j$, unconditionally unbiased and best in the sense that it minimizes the marginal sampling variance of the prediction error. With parameter estimates plugged in, the posterior mean is sometimes referred to as the empirical best linear unbiased predictor (EBLUP). Note that in contrast with parametric empirical Bayes prediction, the concept of best linear unbiased prediction does not rely on distributional assumptions (e.g. Searle *et al.* (1992)).

Deely and Lindley (1981) argued that substitution of estimated parameters in the empirical Bayes predictor is purely pragmatic and has limited statistical rationale. For special cases of linear mixed models, Morris (1983) derived a correction that was designed to counteract the bias that is incurred by substituting estimates for parameters and Rao (1975) proposed a correction that minimizes the mean-squared error when analysis-of-variance or moment estimators are used to estimate the model parameters (see also Reinsel (1984)). However, whenever $\hat{\boldsymbol{\vartheta}}$ is consistent, the effect of substituting estimates for parameters is expected to be small when the sample size is large.

4.3. Alternative methods

4.3.1. Maximum likelihood estimation

After estimation of $\boldsymbol{\vartheta}$, the random effects $\boldsymbol{\zeta}_j$ are sometimes treated as the only unknown *parameters* to be *estimated* by maximizing the likelihood

$$L(\boldsymbol{\zeta}_j) = \prod_{i=1}^{n_j} f(y_{ij}|\boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \hat{\boldsymbol{\vartheta}}^f).$$

As would be expected, the estimates for a cluster become asymptotically unbiased as the number of units in the cluster tends to ∞, although this result is of limited practical utility when cluster sizes are small. Unlike the empirical Bayes predictor, the maximum likelihood estimator for linear models is conditionally unbiased, given the values of the random effects $\boldsymbol{\zeta}_j$.

An advantage of maximum likelihood estimation is that no distributional assumptions need to be invoked for the random effects. However, maximum likelihood estimates have a large mean-squared error when the clusters are not large, which was described as the 'bouncing beta problem' by Rubin (1980). Furthermore, the likelihood does not have a maximum in models for binary data if all responses for a cluster are the same, or in random-coefficient models if the cluster size is less than the number of random effects. Neither example poses any problems for empirical Bayes prediction owing to the information that is provided by the prior distribution. A more fundamental problem with maximum likelihood estimation is that $\boldsymbol{\zeta}_j$ are treated as unknown parameters or fixed effects, which is at odds with the model specification where $\boldsymbol{\zeta}_j$ are random effects.

In logistic random-intercept models or item response models, the maximum likelihood estimator is biased 'outwards' or away from zero for finite cluster sizes, the opposite phenomenon of shrinkage (see Hoijtink and Boomsma (1995)). For such models an unbiased 'weighted maximum likelihood estimator' was proposed by Warm (1989). In factor analysis, maximum likelihood estimation of factor scores is referred to as Bartlett's method (e.g. Bartlett (1938)).

4.3.2. Empirical Bayes modal prediction

Instead of using the posterior mean as in empirical Bayes prediction, we could use the posterior mode. The posterior mode minimizes the posterior expectation of the 0–1 loss function

$$L^{\mathrm{BM}}(\boldsymbol{\zeta}_j, \tilde{\boldsymbol{\zeta}}_j) = \begin{cases} 0 & \text{if } |\boldsymbol{\zeta}_j - \tilde{\boldsymbol{\zeta}}_j| \leqslant \boldsymbol{\varepsilon}, \\ 1 & \text{if } |\boldsymbol{\zeta}_j - \tilde{\boldsymbol{\zeta}}_j| > \boldsymbol{\varepsilon}, \end{cases}$$

where $\boldsymbol{\varepsilon}$ is a vector of minute numbers such that $L^{\mathrm{BM}}(\boldsymbol{\zeta}_j, \tilde{\boldsymbol{\zeta}}_j)$ is 0 when $\tilde{\boldsymbol{\zeta}}_j$ is in the close vicinity of $\boldsymbol{\zeta}_j$ and 1 otherwise. This kind of prediction is called 'maximum *a posteriori*' (MAP) prediction in item response theory (e.g. Bock and Aitkin (1981)).

Generally, there is no analytical expression for the empirical Bayes modal predictor in multilevel generalized linear models and we must resort to numerical methods. Since the denominator of the posterior distribution does not depend on $\boldsymbol{\zeta}_j$ as seen in Section 4.1, we can obtain empirical Bayes modal predictions as solutions to the estimating equations

$$\frac{\partial}{\partial \boldsymbol{\zeta}_j} \ln\{\varphi(\boldsymbol{\zeta}_j; \hat{\boldsymbol{\Psi}})\} + \frac{\partial}{\partial \boldsymbol{\zeta}_j} \ln\{f(\mathbf{y}_j | \boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \hat{\boldsymbol{\vartheta}}^f)\} = \mathbf{0}, \tag{5}$$

assuming that standard second-order conditions for maximization are fulfilled. If $f(\mathbf{y}_j|\boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \hat{\boldsymbol{\vartheta}}^f)$ is viewed as the likelihood (see Section 4.3.1), the empirical Bayes modal predictor can be viewed as a penalized maximum likelihood estimator where the penalty term serves to shrink the predictions towards the prior mode.

In contrast with empirical Bayes, empirical Bayes modal predictions can be obtained by using computationally efficient gradient methods and do not require numerical integration. For this reason, empirical Bayes modal prediction is often used as an approximation to empirical Bayes prediction. Indeed, for linear models the posterior is multivariate normal so the empirical Bayes and empirical Bayes modal predictors coincide.

The version of adaptive quadrature that was suggested for maximum likelihood estimation of model parameters by Pinheiro and Bates (1995) and Schilling and Bock (2005) yields empirical Bayes modal predictions as a by-product.

5. Empirical Bayes standard errors

We now present different kinds of covariance matrices for empirical Bayes predictions. In practice, standard deviations are often called standard errors in this context. There are two principal uses of empirical Bayes standard errors; either for inferences regarding the 'true' realized values of $\boldsymbol{\zeta}_j$ for individual clusters (comparative standard errors) or for model diagnostics (diagnostic standard errors). Posterior standard deviations and prediction error standard deviations serve the former purpose, and marginal sampling standard deviations serve the latter purpose. Closed form expressions for the special case of linear multilevel models are presented in Appendix A.

5.1. Comparative standard errors

Here we consider standard errors that are appropriate for inferences regarding the realized values of $\boldsymbol{\zeta}_j$. One important use of such standard errors is for making comparisons between clusters, and for this reason Goldstein (1995) used the term 'comparative standard error'.

5.1.1. Posterior standard deviations

The empirical Bayesian posterior covariance matrix of the random effects is given by

$$\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \int (\boldsymbol{\zeta}_j - \tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}})(\boldsymbol{\zeta}_j - \tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}})' \, \omega(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) \, \mathrm{d}\boldsymbol{\zeta}_j.$$

(The posterior risk, which was discussed in Section 4.2, is just a weighted sum of the elements of this covariance matrix.) The corresponding variances can also be viewed as the conditional mean-squared error of prediction (CMSEP), given $\mathbf{y}_j$, when the parameters $\boldsymbol{\vartheta}$ are assumed known (Booth and Hobert, 1998).

Assuming approximate normality of the empirical posterior distribution (and known model parameters), Bayesian credible intervals can be formed by using the posterior mean and posterior standard deviation. Bayesian credible intervals have a known probability of containing the random effects for given observed data and are thus conditional on the data, which was referred to as *conditional* empirical Bayes coverage by Carlin and Louis (2000a), page 79. Interestingly, Rubin's (1984), page 1160, frequency calibration argument implies that correct credible intervals should have correct *unconditional* empirical Bayes coverage (at the same level of confidence), i.e. coverage with respect to joint sampling of $\boldsymbol{\zeta}_j$ and $\mathbf{y}_j$. Therefore, the intervals are also appropriate for frequentist prediction. The posterior standard deviation is commonly used as a standard error of prediction in multilevel generalized linear models (e.g. Ten Have and Localio (1999)) and item response theory (e.g. Bock and Mislevy (1982) and Embretson and Reise (2000)).

In general, there is no closed form for the posterior covariance matrix and the integrals must be approximated for instance by adaptive quadrature. For a linear random-intercept model, the posterior variance is

$$\mathrm{var}(\zeta_j|\mathbf{y}_j, \mathbf{X}_j; \hat{\boldsymbol{\vartheta}}) = (1 - \hat{R}_j)\hat{\psi}.$$

As expected, the posterior variance is smaller than the prior variance owing to the information that is gained regarding the random intercept by knowing the responses $\mathbf{y}_j$.

To account for parameter uncertainty, Booth and Hobert (1998) considered the CMSEP over the distribution of $\hat{\boldsymbol{\vartheta}}$ and $\boldsymbol{\zeta}_j$, for given $\mathbf{y}_j$. In a random-intercept model, their approximation amounts to adding a Taylor series expansion of $E\{(\tilde{\zeta}_j^{\mathrm{EB}} - \ddot{\zeta}_j)^2|\mathbf{y}_j\}$ as a correction term to the empirical posterior variance, where $\ddot{\zeta}_j$ is the posterior mean based on the true parameters $\boldsymbol{\vartheta}$ instead of on the estimates $\hat{\boldsymbol{\vartheta}}$. If a consistent estimator $\hat{\boldsymbol{\vartheta}}$ is used, the correction term will become small when there are a large number of clusters. Using flat priors for the model parameters, Kass and Steffey (1989) suggested a very similar approximation for the Bayesian posterior covariance matrix. For the CMSEP, Booth and Hobert (1998) also obtained a correction term by parametric bootstrapping.

Ten Have and Localio (1999) used numerical integration to evaluate the Kass and Steffey approximation for multilevel logistic regression. In a related setting, Tsutakawa and Johnson (1990) adopted a Bayesian approach, taking parameter uncertainty into account by specifying prior distributions for $\boldsymbol{\vartheta}$ and using Bayesian approximations to obtain the posterior mean and variance of $\boldsymbol{\zeta}_j$. Laird and Louis (1987) suggested using bootstrapping to estimate the posterior covariance matrix taking parameter uncertainty into account. Their type III parametric bootstrap consists of repeatedly simulating new data from the estimated model and re-estimating the parameters to generate replicates of the empirical Bayes predictions and their posterior standard deviations. The posterior variance, taking parameter uncertainty into account, is then estimated by the mean of the posterior variance plus the variance of the posterior means (see Rao (2003), page 187, for a discussion of bias correction for this estimator).

5.1.2. Prediction error standard deviations

The (marginal) prediction error covariance matrix is the covariance matrix of the prediction errors $\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}} - \boldsymbol{\zeta}_j$ under repeated sampling of the responses from their marginal distribution,

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j | \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta}) = \int (\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j)(\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j)' g(\mathbf{y}_j | \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta}) \, \mathrm{d}\mathbf{y}_j,$$

where we have omitted the term involving $E_{\mathbf{y}}(\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j | \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta})$ because this expectation is 0 owing to the unconditional unbiasedness of the empirical Bayes predictor. The corresponding variance can also be viewed as the unconditional MSEP when the parameters ϑ are treated as known (Booth and Hobert, 1998). Weighted sums of the elements of the prediction error covariance matrix give the (empirical) Bayes risk or preposterior risk that was discussed in Section 4.2.

It has been shown by Searle *et al.* (1992), page 263, among others, that

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j | \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta}) = E_{\mathbf{y}}\{\mathrm{cov}(\zeta_j | \mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta})\}.$$

Approximating the expected posterior covariance matrix by the posterior covariance matrix given the observed data, we propose the approximation

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j | \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta}) \approx \mathrm{cov}(\zeta_j | \mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\vartheta}). \tag{6}$$

For linear models, the posterior covariance matrix does not depend on the responses $\mathbf{y}_j$ so the approximation becomes exact.

If the sampling distributions of the prediction errors are approximately normal, the (marginal) prediction error standard deviations could be used to construct confidence intervals for realized random effects. Under normality of the prediction errors, such Wald-type confidence intervals have correct *unconditional* empirical Bayes and frequentist prediction coverage. However, unlike intervals that are based on the posterior standard deviations, the intervals have no *conditional* interpretation, given the data for a cluster.

In multilevel linear models, Goldstein (1995, 2003) defined the comparative standard error as the marginal prediction error standard deviation. This equals the posterior standard deviation in the linear case. However, in multilevel *generalized linear* models, the marginal prediction error standard deviation is not identical to the posterior standard deviation. For these models, we suggest using the posterior standard deviation as comparative standard error because the corresponding confidence intervals should have correct conditional and unconditional coverage (under normality). Booth and Hobert (1998) made an analogous point, advocating the CMSEP in favour of the unconditional MSEP that is usually used in small area estimation. In Section 9.1.2 we compare the standard errors using simulations.

Note that the prediction error covariances are not fully frequentist since the sampling variability of $\hat{\vartheta}$ is ignored. In linear models, it is easy to take uncertainty in the estimated regression parameters into account (see Appendix A), and Kackar and Harville (1984) gave approximations also taking the uncertainty of the estimated variance parameters into account for two-level linear models.

We could also use parametric bootstrapping to estimate the prediction error variances, first drawing random effects from their prior distribution and subsequently responses from their conditional distribution given the random effects. The true random effects are then just the simulated effects and, subtracting these from the empirical Bayes predictions, we can estimate the prediction error variances. To reflect the imprecision of the parameter estimates, the parameters should be re-estimated in each bootstrap sample. However, the resulting bootstrap estimator of the prediction error variance is still biased because the bootstrap samples are generated by using estimated parameters (Hall and Maiti, 2006). Hall and Maiti (2006) suggested a double-bootstrap procedure to correct this bias. An alternative approach is to use bootstrapping to

correct the bias of analytical expansions for the prediction error variance (see the references in Hall and Maiti (2006)).

5.2. Diagnostic standard errors

For model diagnostics, it is useful to consider the marginal sampling covariance matrix of the empirical Bayes predictor

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \mathrm{cov}_{\mathbf{y}}\{E(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\} = \int \tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}} \tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}\prime} g(\mathbf{y}_j|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\, \mathrm{d}\mathbf{y}_j,$$

where we have again used the fact that $E_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \mathbf{0}$. This is the covariance matrix of the predictions under repeated sampling of the responses from their marginal distribution, keeping the covariates fixed and plugging in parameter estimates $\hat{\boldsymbol{\vartheta}}$.

The marginal sampling standard deviation can be used for detecting clusters that appear inconsistent with the model (e.g. Lange and Ryan (1989) and Langford and Lewis (1998)). For this reason, Goldstein (1995) referred to this quantity as the 'diagnostic standard error'.

Unfortunately there is no closed form expression for multilevel generalized linear models with non-linear links. However, it is shown in Appendix B that

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \hat{\boldsymbol{\Psi}} - E_{\mathbf{y}}\{\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\}.$$

This led Skrondal (1996) to suggest the approximation

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) \approx \hat{\boldsymbol{\Psi}} - \mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}). \tag{7}$$

For linear models, this approximation holds perfectly, so the marginal sampling variance is $\hat{R}_j\hat{\psi}$ for linear random-intercept models.

Because of shrinkage, the sampling variance is smaller than the prior variance. This has led some researchers (e.g. Louis (1984)) to suggest adjusted empirical Bayes predictors with the same covariance matrix as the prior distribution. This predictor minimizes the posterior expectation of the quadratic loss function (for given parameter estimates) in equation (3) subject to the side-condition that the predictions match the estimated first- and second-order moments of the prior distribution.

The sampling covariances are not fully frequentist since the sampling variability of $\hat{\boldsymbol{\vartheta}}$ is ignored. However, for linear models, it is quite straightforward to take the uncertainty of the estimation of the regression parameters $\boldsymbol{\beta}$ (but not the uncertainty due to estimation of the variance parameters $\boldsymbol{\Psi}$ and θ) into account (see Appendix A).

We could also estimate the sampling variance by using parametric bootstrapping, first sampling the random effects from the prior distribution and then the responses from their conditional distribution given the random effects and the covariates. (See Section 3 for an example and Section 9 for a comparison of sampling standard deviations based on the approximation and based on bootstrapping.) An advantage of the bootstrapping approach is that uncertainty in the parameter estimates $\hat{\boldsymbol{\vartheta}}$ is easily accommodated by re-estimating the parameters in each bootstrap sample.

6. Application continued: prediction of school-specific intercepts

We selected 10 schools from the US PISA data with a range of sample sizes n_j and with large, small and intermediate values of the empirical Bayes predictions $\tilde{\zeta}_j^{\mathrm{EB}}$ based on the parameter estimates for the random-intercept logistic regression model that are presented in Table 1.

Table 2. Predictions of random intercepts and associated standard errors for 10 schools from the PISA data

School	n_j	$\tilde{\zeta}_j^{\mathrm{EB}}$	$\tilde{\zeta}_j^{\mathrm{EBM}}$	*Comparative standard error*		*Diagnostic standard error*	
				$SD(\zeta_j\|\mathbf{y}_j)$ *(approximation (6))*	$SD(\tilde{\zeta}_j^{\mathrm{EB}}-\zeta_j)$ *(bootstrap†)*	$SD(\tilde{\zeta}_j^{\mathrm{EB}})$ *(approximation (7))*	$SD(\tilde{\zeta}_j^{\mathrm{EB}})$ *(bootstrap†)*
105	1	−0.043	−0.040	0.520	0.506	0.097	0.131
85	3	0.132	0.140	0.501	0.496	0.171	0.181
33	4	−0.433	−0.428	0.474	0.463	0.236	0.262
6	10	−0.473	−0.456	0.451	0.422	0.276	0.306
42	12	−0.005	0.001	0.397	0.394	0.350	0.346
35	13	0.800	0.792	0.394	0.379	0.354	0.352
2	17	0.478	0.478	0.363	0.371	0.386	0.379
67	21	0.031	0.039	0.349	0.347	0.398	0.393
54	22	−0.325	−0.319	0.341	0.333	0.405	0.407
19	25	0.861	0.852	0.332	0.323	0.412	0.419

†Bootstrapping using 1000 replicates.

Table 2 gives the school identifier, cluster size n_j, empirical Bayes prediction (using `gllamm` with 20-point adaptive quadrature), empirical Bayes modal prediction (using `xtmelogit` in Stata with 20-point adaptive quadrature), comparative standard errors and diagnostic standard error $\mathrm{SD}(\tilde{\zeta}_j^{\mathrm{EB}})$. For the comparative standard errors, both the posterior standard deviation $\mathrm{SD}(\zeta_j|\mathbf{y}_j)$ and the prediction error standard deviation $\mathrm{SD}(\tilde{\zeta}_j^{\mathrm{EB}}-\zeta_j)$ are given. The latter is obtained by using parametric bootstrapping with 1000 replications, and $\mathrm{SD}(\zeta_j|\mathbf{y}_j)$ also represents the approximation in expression (6). For the diagnostic standard error, results from both the approximation in expression (7) and parametric bootstrapping are reported. Note that none of the standard errors incorporate parameter uncertainty.

We see that the modes and means of the posterior distributions are quite close (compared with the magnitude of the posterior standard deviations), indicating that the posterior distributions are quite symmetric. The posterior standard deviation (or approximate comparative standard error) is lower than the estimated prior standard deviation $\sqrt{\hat{\psi}}=0.53$ and tends to decrease with increasing cluster size n_j, reflecting the increasing accuracy with which ζ_j can be predicted. The sampling standard deviations of the empirical Bayes predictions (or diagnostic standard errors) are lower than the prior standard deviation because of shrinkage and, as expected, this is less so for larger cluster sizes. The approximations for the standard errors work reasonably well.

If the empirical Bayes predictions have approximately normal sampling distributions, the diagnostic standard error can be used to identify outlying schools. For example, schools 35 and 19 might be considered outlying because the empirical Bayes predictions exceed two diagnostic standard errors (ignoring the multiple-testing problem; see Longford (2001) and Afsharthous and Wolf (2007)). However, as we shall see in Section 9.1.1, the normal approximation works only for large cluster sizes combined with a small random-intercept variance.

If the sampling distributions of the prediction errors are approximately normal, the posterior standard deviation could be used to form confidence intervals for the realized random intercepts or form confidence intervals for differences. For instance, the difference in school-specific intercepts between schools 35 and 42 is predicted as 0.805 with an approximate standard error of $\sqrt{(0.394^2+0.397^2)}=0.559$, so an approximate 95% confidence interval for the difference in realized intercepts is $0.805\pm 1.96\times 0.559$, giving confidence limits −0.29 and 1.90.

7. Prediction of expected responses and probabilities

In this section we consider prediction of different kinds of expectations of the responses y_{ij} for covariate values $\mathbf{x}_{ij} = \mathbf{x}^0$ and $\mathbf{z}_{ij} = \mathbf{z}^0$. In the longitudinal setting this kind of prediction is usually called forecasting. For categorical responses, the expectations of interest are probabilities.

7.1. Conditional expectation: prediction for a unit in a hypothetical cluster

The conditional mean response, or probability, for a unit with covariate values $\mathbf{x}^0$ and $\mathbf{z}^0$ in a hypothetical cluster with random effects $\boldsymbol{\zeta}_j = \boldsymbol{\zeta}_j^0$ is given by

$$\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j^0) \equiv E_y(y_{ij}|\boldsymbol{\zeta}_j^0, \mathbf{x}^0, \mathbf{z}^0; \hat{\boldsymbol{\beta}}) = \int_{-\infty}^{\infty} y_{ij} f(y_{ij}|\boldsymbol{\zeta}_j^0, \mathbf{x}^0, \mathbf{z}^0; \hat{\boldsymbol{\beta}}) \, \mathrm{d}y_{ij} = h(\mathbf{x}^{0\prime}\hat{\boldsymbol{\beta}} + \mathbf{z}^{0\prime}\boldsymbol{\zeta}_j^0).$$

The conditional variance of the linear predictor due to parameter uncertainty (given $\boldsymbol{\zeta}_j = \boldsymbol{\zeta}_j^0$) is $\mathbf{x}^{0\prime}\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{x}^0$. In linear models the linear predictor becomes the prediction of the conditional mean response, and therefore $\sqrt{\{\mathbf{x}^{0\prime}\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{x}^0\}}$ becomes the standard error of the prediction and can be used to form confidence intervals. For multilevel generalized linear models we can use the delta method to obtain the standard error of prediction, or form confidence intervals for the linear predictor and apply the inverse link function to the limits of the confidence interval.

Instead of using particular values $\boldsymbol{\zeta}_j^0$ of the random effects, we can consider the distribution of $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)$ in the population of clusters. For example, Duchateau and Janssen (2005) used the random-effects density $\varphi(\boldsymbol{\zeta}_j; \hat{\boldsymbol{\Psi}})$ to derive the density function of the conditional probability in a logistic regression model, giving a 'prevalence density'. Since the inverse link function $h(\cdot)$ is a monotonic function, substituting given percentiles of $\mathbf{z}^{0\prime}\boldsymbol{\zeta}_j$ (for fixed $\mathbf{z}^0$) gives the corresponding percentiles of $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)$ (given the covariates). In random-intercept models, it is natural to consider the median by substituting $\zeta_j = 0$, and perhaps a 95% range by substituting $\zeta_j = \pm 1.96\sqrt{\hat{\psi}}$; see Section 8 and Fig. 3 there for examples.

An alternative to using the prior distribution of the random effects $\varphi(\boldsymbol{\zeta}_j; \hat{\boldsymbol{\Psi}})$ to derive a distribution of $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)$ would be to use the posterior distribution $\omega(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})$. The expectations of these two types of distributions of $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)$ are discussed in Section 7.2 and Section 7.3 respectively.

7.2. Population-averaged expectation: prediction for a unit in a new cluster

We now consider the predicted mean response for the population of clusters. Using the double-expectation rule, the (predicted) population average of the conditional mean response, or probability, $\bar{\mu}(\mathbf{x}^0, \mathbf{z}^0)$ is obtained by integrating $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)$ over the (prior) random-effects distribution,

$$\bar{\mu}(\mathbf{x}^0, \mathbf{z}^0) \equiv E_y(y_{ij}|\mathbf{x}^0, \mathbf{z}^0; \hat{\boldsymbol{\vartheta}}) = \int_{-\infty}^{\infty} \hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j) \varphi(\boldsymbol{\zeta}_j; \hat{\boldsymbol{\Psi}}) \, \mathrm{d}\boldsymbol{\zeta}_j.$$

This population-averaged or marginal expectation can be used to make a prediction for a unit in a new cluster, assuming that the new cluster is sampled randomly.

In linear models, the population average is obtained by simply plugging in the mean of the random effects (which is $\mathbf{0}$) in the expression for the conditional expectation, $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \mathbf{0})$, and the corresponding sampling variance is $\mathbf{x}^{0\prime}\,\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{x}^0$. In this case the predicted marginal expectation can also be used to predict a *response* y_{ij} for a unit in a new cluster j with covariates values $\mathbf{x}^0$ and $\mathbf{z}^0$. The variance of the prediction error $y_{ij} - \hat{y}_{ij}$, treating $\hat{\boldsymbol{\Psi}}$ and $\hat{\theta}$ as known, becomes

$$\mathrm{var}_{y,\hat{\boldsymbol{\beta}}}(y_{ij} - \hat{y}_{ij}|\mathbf{x}^0, \mathbf{z}^0; \hat{\boldsymbol{\Psi}}, \hat{\theta}) = \mathbf{x}^{0\prime}\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{x}^0 + \mathbf{z}^{0\prime}\hat{\boldsymbol{\Psi}}\mathbf{z}^0 + \hat{\theta}.$$

Afshartous and de Leeuw (2005) called this method of predicting responses the 'prior prediction method'. They showed that the population-averaged expectation is also the posterior expectation for a new unit in a new cluster (when parameters are assumed known), making it a Bayes rule under squared error loss. This predictor therefore also minimizes the unconditional MSEP.

In most models with non-linear link functions, we cannot obtain population-averaged expectations or probabilities by simply plugging in the mean of the random effects in the expression for conditional expectation. The integral that is involved in the expectation must generally be evaluated numerically or by simulation, a notable exception being probit models (e.g. Rabe-Hesketh and Skrondal (2008a, b)). For a two-level complementary log–log-discrete-time survival model, Rose *et al.* (2006) nevertheless predicted the probability of survival for a new unit in a new cluster by using the conditional predicted probability $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \mathbf{0})$ with random effects set to zero instead of the population-averaged probability.

The fact that population-averaged and conditional expectations differ, $\bar{\mu}(\mathbf{x}^0, \mathbf{z}^0) \neq \hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \mathbf{0})$, leads to the important distinction between marginal (or population-averaged) effects and conditional (or cluster-specific) effects in multilevel generalized linear models. Briefly, marginal effects express comparisons of population strata defined by covariate values, whereas conditional effects express comparisons holding the cluster-specific random effects (and covariates) constant.

Approximate confidence intervals for predicted marginal expectations can be obtained by simulating parameters from their estimated asymptotic sampling distribution (see Section 8 and Fig. 2 there for examples).

7.3. Cluster-averaged expectation: prediction for a new unit in an existing cluster

We now consider the mean response for a particular cluster, which we call *cluster-averaged* in contrast with population-averaged expectation. Since the random effects for the cluster are unknown, we cannot use the conditional mean that was discussed in Section 7.1. Instead, we average over the posterior distribution which represents all our knowledge about the random effects for the cluster.

The cluster-averaged expectation $\tilde{\mu}_j(\mathbf{x}^0, \mathbf{z}^0)$ is obtained by integrating $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)$ over the posterior distribution of the random effects for the cluster

$$\tilde{\mu}_j(\mathbf{x}^0, \mathbf{z}^0) \equiv E_{\zeta}\{\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}\} = \int_{-\infty}^{\infty} \hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \boldsymbol{\zeta}_j)\, \omega(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}})\, \mathrm{d}\boldsymbol{\zeta}_j.$$

This posterior expectation can be used to make predictions for a new unit in the existing cluster j, exploiting the information that we already have about the cluster. The posterior expectation is a Bayes rule under squared error loss and is the empirical best predictor (EBP) that was suggested by Jiang and Lahiri (2001) for small area estimation of proportions. For non-linear link functions, $\tilde{\mu}_j(\mathbf{x}^0, \mathbf{z}^0) \neq \hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}})$, so the posterior expectation should be obtained by using, for instance, numerical integration (see Section 8 and Fig. 3 there for examples). Simply plugging in the empirical Bayes predictions of the random effects $\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}$ in non-linear functions is nevertheless not uncommon (e.g. Gibbons *et al.* (1994) and Farrell *et al.* (1997)).

It is also sometimes useful to obtain 'post-dictions' ('predictions' after the fact) for an existing unit in an existing cluster. For example, in longitudinal binary data, 'post-dicted' probabilities can be used to plot individual growth trajectories for visualizing aspects of the model and the data (e.g. Rabe-Hesketh and Skrondal (2008b), pages 269–271). It may appear odd to use the observed response for a unit (within the posterior distribution of $\boldsymbol{\zeta}_j$ given $\mathbf{y}_j$) to make a prediction for the same unit, but it is the unknown probability that we are predicting, not the observed response.

For linear models, the posterior expectation of the conditional mean response simply becomes $\hat{\mu}(\mathbf{x}^0, \mathbf{z}^0, \tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}})$ and can be used as a predicted response $\tilde{y}_{ij}^{\mathrm{p}}$ for a new unit in an existing cluster. The variance of the prediction error $y_{ij} - \tilde{y}_{ij}^{\mathrm{p}}$, treating $\hat{\boldsymbol{\Psi}}$ and $\hat{\theta}$ as known, is

$$\begin{aligned}\mathrm{var}_{y,\hat{\boldsymbol{\beta}}}(y_{ij} - \tilde{y}_{ij}^{\mathrm{p}} | \mathbf{x}^0, \mathbf{z}^0; \hat{\boldsymbol{\Psi}}, \hat{\theta}) = {} & \mathbf{x}^{0\prime} \mathrm{cov}(\hat{\boldsymbol{\beta}}) \mathbf{x}^0 + \mathbf{z}^{0\prime} \mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}} - \boldsymbol{\zeta}_j | \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) \mathbf{z}^0 \\ & - \mathbf{x}^{0\prime} \mathrm{cov}(\hat{\boldsymbol{\beta}}) \mathbf{X}_j' \hat{\boldsymbol{\Sigma}}_j^{-1} \mathbf{Z}_j \hat{\boldsymbol{\Psi}} \mathbf{z}^0 - \mathbf{z}^{0\prime} \hat{\boldsymbol{\Psi}} \mathbf{Z}_j' \hat{\boldsymbol{\Sigma}}_j^{-1} \mathbf{X}_j \mathrm{cov}(\hat{\boldsymbol{\beta}}) \mathbf{x}^0 + \hat{\theta}.\end{aligned}$$

As pointed out by Afshartous and de Leeuw (2005), this 'multilevel prediction method' minimizes the conditional and unconditional MSEP (for known parameters) since it is a Bayes rule under squared error loss. Not surprisingly, therefore, their simulations for linear multilevel models show that this method produces a smaller MSEP for predicting responses for a new unit in an *existing* cluster, compared with the population-averaged expectation that was discussed in Section 7.2.

8. Application continued: predicting probabilities of reading proficiency

Returning to the PISA data on reading proficiency and SES, we now demonstrate how graphs of predictions can be used to convey complex estimated relationships and their uncertainty. This graphical approach is especially poignant when communicating the results of statistical modelling to non-statistical audiences such as educators and policy makers. All predictions are obtained by using `gllapred`, the prediction command of `gllamm`.

We first consider three kinds of effect (the between, within and contextual effect) of SES on the population-averaged probability of reading proficiency. We calculated predicted population-averaged probabilities $\bar{\mu}(\mathbf{x}_j^0)$ for covariate values $\mathbf{x}_j^0 = (x_{ij}^0 - \bar{x}_{\cdot j}^0, \bar{x}_{\cdot j}^0)'$ chosen to represent the three kinds of effects of SES (note that the random part of the model contains a random intercept only, so for simplicity $\mathbf{z}^0$ is omitted from the notation that was introduced in Section 7):

(a) between effect, $\mathbf{x}_j^0 = (0, \bar{x}_{\cdot j})'$, where $\bar{x}_{\cdot j}$ ranges from 25 to 68;
(b) contextual effect, $\mathbf{x}_j^0 = (45 - \bar{x}_{\cdot j}, \bar{x}_{\cdot j})'$, where $\bar{x}_{\cdot j}$ ranges from 25 to 68;
(c) within effect, $\mathbf{x}_j^0 = (x^0 - 45, 45)'$, where x^0 ranges from 25 to 68.

The corresponding curves are shown in Fig. 1. The broken curve (between effect) represents the expected proportion of students who are proficient as a function of school mean SES for students whose SES equals the school mean. The full curve (contextual effect) represents the expected proportion of students who are proficient as a function of school mean SES for students whose individual SES is 45. Finally, the dotted curve (within effect) represents the proportion of students who are proficient as a function of individual SES for a school whose mean SES is 45. We see that the within effect is quite small compared with the between-school and contextual effects, with the expected proportion proficient increasing by less than 0.1 when individual SES increases from 25 to 68. The contextual effect is very pronounced, with the expected proportion proficient ranging from about 0.1 to about 0.7 as school mean SES increases from the lowest to the highest value in the sample and when individual SES is held constant at 45.

Unfortunately, plots such as Fig. 1 ignore the uncertainty that is involved in making predictions using estimated model parameters. To address this problem, Fig. 2 shows approximate pointwise 95% confidence bands for the predicted population-averaged probability $\bar{\mu}(\mathbf{x}_j^0)$ for the contextual effect with $\mathbf{x}_j^0 = (45 - \bar{x}_{\cdot j}, \bar{x}_{\cdot j})'$. To produce the confidence bands, we randomly drew 1000 parameter vectors from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\vartheta}}$ and covariance matrix $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\vartheta}})$, the estimated asymptotic sampling distribution of the estimates. For each randomly drawn parameter vector $\boldsymbol{\vartheta}^k$, $k = 1, \ldots, 1000$, we computed the predicted mar-

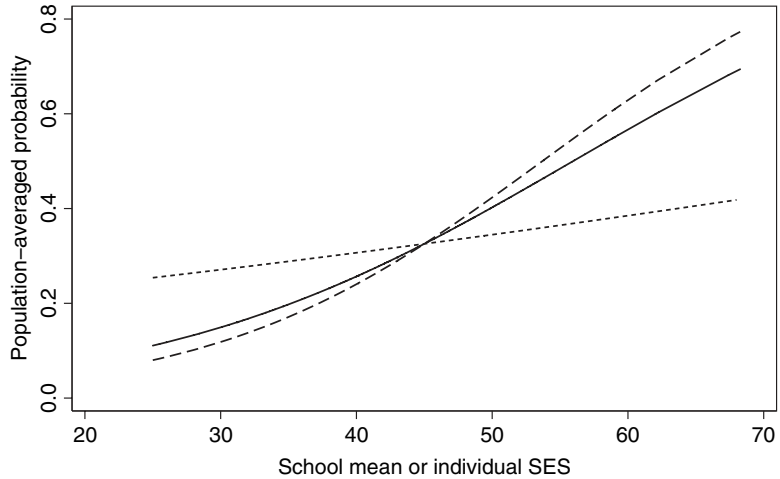

Fig. 1. Between-school (– – –), contextual (———) and within-school (-------) effects of SES on the predicted population-averaged probability of proficiency, with individual SES set to 45 for the contextual effect and school mean SES set to 45 for the within-school effect

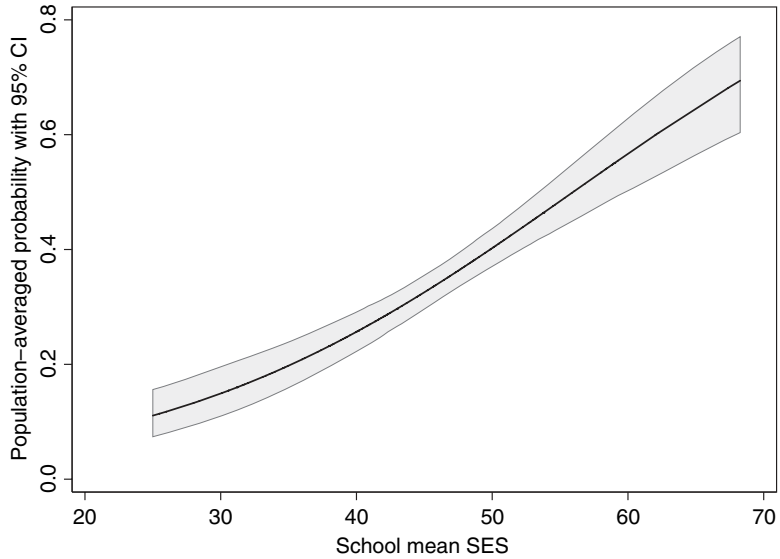

Fig. 2. Contextual effect of SES: predicted population-averaged probabilities of reading proficiency as a function of school mean SES for students with SES equal to 45, with pointwise 95% confidence intervals representing parameter uncertainty (by simulation with 1000 replicates)

ginal mean $\bar{\mu}^k(\mathbf{x}_j^0)$ for each school and then identified the 25th- and 976th-largest values for each school.

It is also useful to convey the variability between clusters due to the random part of the model. Fig. 3 considers the contextual effect for $\mathbf{x}_j^0 = (45 - \bar{x}_{.j}, \bar{x}_{.j})'$ and shows the school-specific posterior mean probabilities $\tilde{\mu}_j(\mathbf{x}_j^0)$ for the schools in the sample (dots), together with the corresponding estimated median probability $\hat{\mu}(\mathbf{x}_j^0, \zeta_j) = \hat{\mu}(\mathbf{x}_j^0, 0)$ (full curve) and the 2.5- and 97.5-percentiles $\hat{\mu}(\mathbf{x}_j^0, \pm 1.96\sqrt{\hat{\psi}})$ (broken curves), as a function of school mean $\bar{x}_{.j}$ when student SES is 45. Fig. 3 shows the conditional effect of school mean SES and the

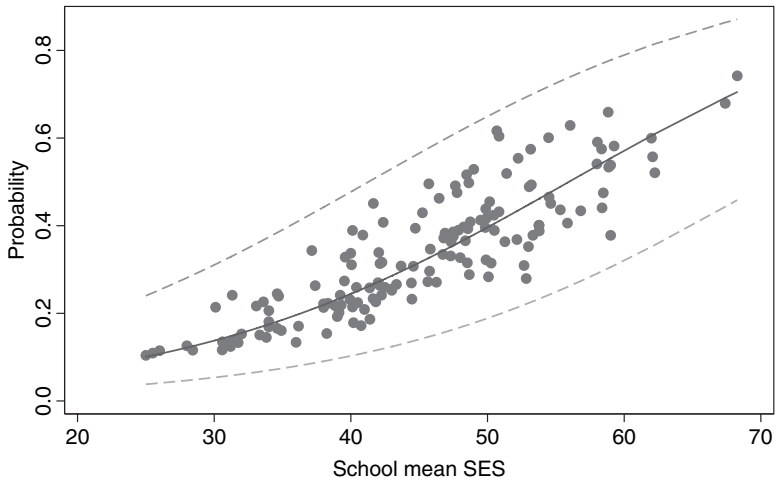

Fig. 3. Contextual effect of SES: predicted median probability of reading proficiency (———) and 95% range of probabilities (– – –) as a function of school mean SES for students with SES equal to 45; predicted school-specific posterior mean probabilities (∘) for students with SES equal to 45

variability between schools keeping student SES constant. Whereas the 95% range conveys the estimated variability in the population, the school-specific predictions can be useful for identifying schools that do remarkably well or badly taking into account the school mean SES (with student SES held constant). The school-specific predictions all lie within the 95% range, and this is probably due to shrinkage. The effect of another covariate, such as gender, could also be considered by producing separate curves for boys and girls. If gender had a school level random coefficient, displaying posterior mean probabilities by gender would also be informative.

Table 3 presents various predicted probabilities for the same schools as in Table 2. Since these probabilities depend on $\mathbf{x}_j^0 = (45 - \bar{x}_{\cdot j}, \bar{x}_{\cdot j})$, the cluster-mean SES $\bar{x}_{\cdot j}$ is provided as well. The population-averaged probabilities $\bar{\mu}(\mathbf{x}_j^0)$ are closer to 0.5 than the median probabilities $\hat{\mu}(\mathbf{x}_j^0, 0)$, but they do not differ dramatically here because the estimated random-intercept variance is quite small. To help to interpret the cluster-averaged or posterior mean probabilities $\tilde{\mu}_j(\mathbf{x}_j^0)$ and the conditional probabilities $\hat{\mu}(\mathbf{x}_j^0, \tilde{\zeta}_j^{\mathrm{EB}})$, we present the empirical Bayes predictions

Table 3. Different kinds of predicted probabilities of reading proficiency for 10 schools from the PISA data (with student SES set to 45)

School	n_j	$\bar{x}_{\cdot j}$	$\bar{\mu}(\mathbf{x}_j^0)$	$\hat{\mu}(\mathbf{x}_j^0, 0)$	$\tilde{\zeta}_j^{\mathrm{EB}}$	$\tilde{\mu}_j(\mathbf{x}_j^0)$	$\hat{\mu}(\mathbf{x}_j^0, \tilde{\zeta}_j^{\mathrm{EB}})$
105	1	34.000	0.187	0.175	−0.043	0.181	0.169
85	3	34.000	0.187	0.175	0.132	0.206	0.195
33	4	53.000	0.451	0.448	−0.433	0.352	0.345
6	10	40.200	0.259	0.247	−0.473	0.179	0.170
42	12	49.833	0.400	0.393	−0.005	0.396	0.392
35	13	50.846	0.416	0.411	0.800	0.604	0.608
2	17	47.765	0.367	0.359	0.478	0.475	0.475
67	21	47.333	0.361	0.352	0.031	0.363	0.359
54	22	53.318	0.456	0.454	−0.325	0.378	0.375
19	25	50.680	0.413	0.408	0.861	0.617	0.620

$\tilde{\zeta}_j^{\mathrm{EB}}$ again in Table 3. We see that the school-specific probabilities $\tilde{\mu}_j(\mathbf{x}_j^0)$ differ more from the population-averaged (or median) probabilities $\bar{\mu}(\mathbf{x}_j^0)$ (or $\hat{\mu}(\mathbf{x}_j^0, 0)$) when the posterior distribution has its mean further from 0 as would be expected. As discussed, plugging the empirical Bayes prediction into the conditional response probability does not give the posterior mean probability. The latter is closer to 0.5, and the difference is greater for smaller cluster sizes where the posterior standard deviations are larger (see Table 2), but none of the differences are very pronounced.

9. Monte Carlo simulations

We now use simulations to assess the performance of methods for obtaining diagnostic and comparative standard errors for empirical Bayes predictions of random effects and to assess the performance of approximations that are sometimes used for predicted response probabilities.

We consider one of the most common types of multilevel generalized linear model, a random-intercept logistic regression model,

$$\mathrm{logit}\{\Pr(y_{ij}=1|\zeta_j)\} = \underbrace{\beta_0}_{0} + \zeta_j,$$

where $\zeta_j \sim N(0, \psi)$. The model can alternatively be written as a latent response model

$$y_{ij}^* = \underbrace{\beta_0}_{0} + \zeta_j + \varepsilon_{ij},$$

$$y_{ij} = \begin{cases} 1 & \text{if } y_{ij}^* > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\zeta_j \sim N(0, \psi)$ and ε_{ij} has a standard logistic distribution which has zero mean and variance $\pi^2/3$. The intra-class correlation ICC* between different latent responses y_{ij}^* and $y_{i'j}^*$ in the same cluster becomes

$$\mathrm{ICC}^* = \frac{\psi}{\psi + \pi^2/3}.$$

To investigate the effects of the cluster size n_j and intraclass correlation ICC*, we use a full factorial design with $n_j \in \{3, 10, 20, 100\}$ and $\mathrm{ICC}^* \in \{0.1, 0.2, 0.5, 0.8\}$, corresponding to $\sqrt{\psi} \in \{0.60, 0.91, 1.81, 3.62\}$. With cluster sizes ranging from 1 to 28 and an estimated ICC* of 0.08, the PISA data are most similar to the conditions $n_j = 3$, $n_j = 10$ and $n_j = 20$ combined with $\mathrm{ICC}^* = 0.1$. For each condition we simulate responses for $J = 10000$ clusters of the same size $n_j = n$ from the logistic random-intercept model.

We obtain predictions that are based on true parameter values, imitating the situation where naive parametric bootstrapping is performed without re-estimating the model parameters in each bootstrap sample so that parameter uncertainty is ignored. The 10000 clusters can therefore be viewed as independent bootstrap samples.

9.1. Empirical Bayes predictions of random effects

9.1.1. Diagnostic standard errors

Posterior means and standard deviations of ζ_j are obtained by 30-point adaptive quadrature. The standard deviation of the empirical Bayes predictions across the 10000 clusters is a simulation-based estimate of the diagnostic standard error of the empirical Bayes predictions. For

each cluster, we also obtain an approximate squared diagnostic standard error as shown in expression (7), by using the posterior variance for the second term in the following equality, instead of its expectation:

$$\mathrm{var}_{\mathbf{y}}(\tilde{\zeta}_j^{\mathrm{EB}};\hat{\boldsymbol{\vartheta}}) = \hat{\psi} - E_{\mathbf{y}}\{\mathrm{var}(\zeta_j|\mathbf{y}_j;\hat{\boldsymbol{\vartheta}})\} \approx \hat{\psi} - \mathrm{var}(\zeta_j|\mathbf{y}_j;\hat{\boldsymbol{\vartheta}}).$$

The mean of this approximation across the 10000 clusters is an alternative simulation-based estimate of the squared diagnostic standard error. Both simulation-based estimates of the diagnostic standard error were very close in our experiment, never differing from each other by more than 2% (we report the latter estimate in Fig. 4).

The most likely use of the diagnostic standard error is for the detection of unusual clusters based on a normal approximation of the sampling distribution of the empirical Bayes predictions. We therefore consider the null hypothesis that the model is correct and perform z-tests for each cluster using

(a) the simulation-based diagnostic standard error and
(b) the approximate diagnostic standard error.

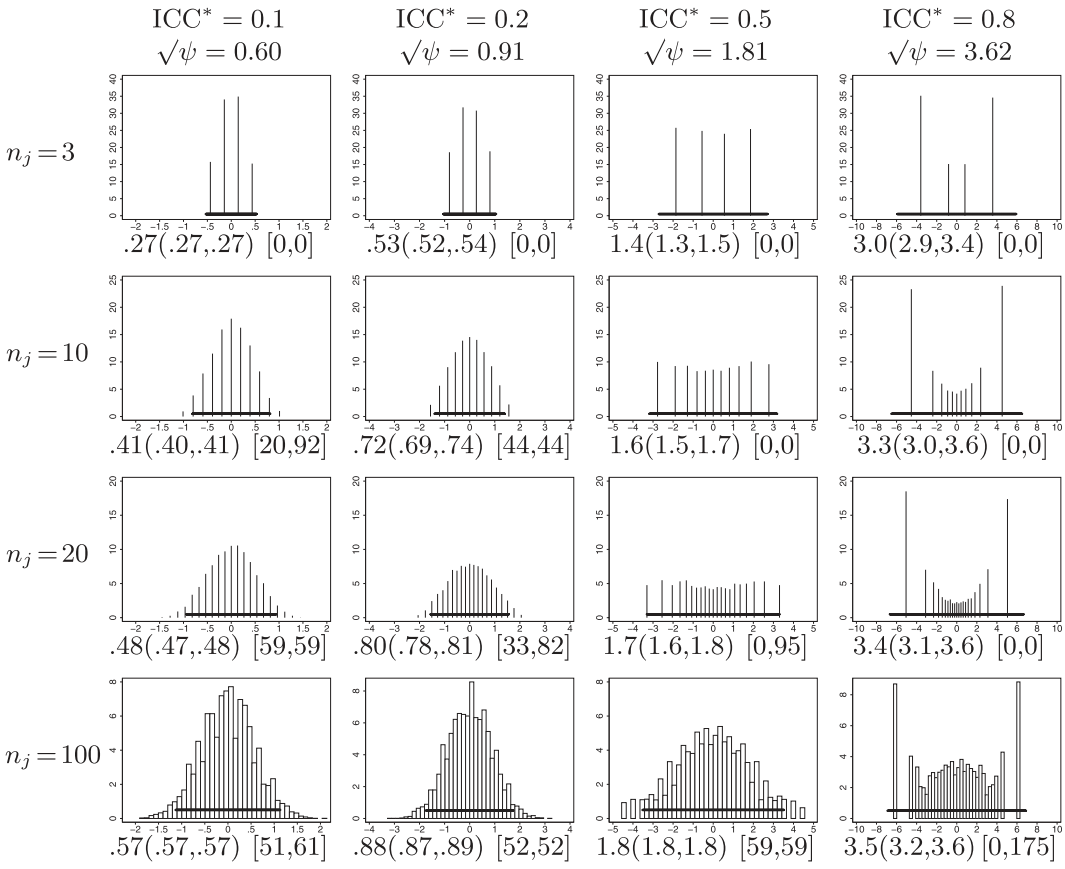

Fig. 4. Empirical sampling distributions of empirical Bayes predictions for various intraclass correlations and cluster sizes: below each graph we report SD($\tilde{\zeta}_j^{\mathrm{EB}}$) (from parametric bootstrapping), followed by the 10th and 90th percentiles of the approximations of this standard error in parentheses, followed in square brackets by type I error rates (per thousand) using the simulation-based SD($\tilde{\zeta}_j^{\mathrm{EB}}$) and the approximation $\sqrt{\{\hat{\psi} - \mathrm{var}(\zeta_j|\mathbf{y}_j)\}}$ for each cluster, where the nominal rate is 50 per thousand; the horizontal bars represent the intervals ±1.96 SD($\tilde{\zeta}_j^{\mathrm{EB}}$)

The results are presented in Fig. 4. The graphs show the empirical sampling distribution of the empirical Bayes predictions for each of the 16 conditions, together with the interval $\pm 1.96\,\mathrm{SD}(\tilde{\zeta}_j^{\mathrm{EB}})$. For predictions outside this interval, the null hypothesis is rejected. The distributions of the empirical Bayes predictions look markedly non-normal for most conditions. This is partly because the predictions are discrete with $n_j + 1$ unique values, corresponding to all possible cluster totals of the responses, $0, 1, \ldots, n_j$. This will also be true if the model includes covariates, because in a logistic regression model

$$f(\mathbf{y}_j|\mathbf{X}_j) = \{1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}+\zeta_j)\}^{-n_j} \prod_{i=1}^{n_j} \exp(\mathbf{x}_{ij}'\boldsymbol{\beta}+\zeta_j)^{y_{ij}}$$
$$= \exp\left(\zeta_j \sum_{i=1}^{n_j} y_{ij}\right) \{1+\exp(\mathbf{x}_{ij}'\boldsymbol{\beta}+\zeta_j)\}^{-n_j} \prod_{i=1}^{n_j} \exp(\mathbf{x}_{ij}'\boldsymbol{\beta} y_{ij}),$$

so the cluster total $\Sigma_i y_{ij}$ is a sufficient statistic for ζ_j.

For $\mathrm{ICC}^* = 0.8$, the distributions are very non-normal with large proportions of extremely large and small empirical Bayes predictions. The distributions look increasingly normal as the intraclass correlation ICC^* decreases and the cluster size n_j increases (towards the bottom left of Fig. 4).

Below each graph in Fig. 4 we report the simulation-based diagnostic standard error, together with the 10th and 90th percentiles (in parentheses) of the approximate diagnostic standard error across the 10000 clusters. The approximation works poorly for $\mathrm{ICC}^* = 0.8$ where both percentiles tend to be quite different from the simulation-based diagnostic standard error. The rejection rates (per thousand) by using the simulation-based and approximate diagnostic standard error together with a normal approximation are given in square brackets and should be compared with the nominal rate of 50 (per thousand). The test seems to work for $\mathrm{ICC}^* \leqslant 0.2$ and $n_j = 100$, where the distributions appear to be approximately normal, performs reasonably for the neighbouring conditions of $\mathrm{ICC}^* = 0.1$ and $n_j = 20$, and $\mathrm{ICC}^* = 0.5$ and $n_j = 100$, but fails for the other conditions.

9.1.2. Comparative standard errors

For the same simulated data as above, we consider both the posterior standard deviation $\mathrm{SD}(\zeta_j|\mathbf{y}_j)$ and the parametric bootstrap estimate of the prediction error standard deviation $\mathrm{SD}(\tilde{\zeta}_j^{\mathrm{EB}} - \zeta_j)$. The bootstrap estimate can be obtained either as the standard deviation of the prediction errors across the 10000 clusters, or as the square root of the mean of the squared posterior standard deviations. The two simulation-based estimates agree very closely, and we use the latter. We assessed the performance of the standard errors by forming a confidence interval for the realized random intercept and checking whether the actual realized random intercept falls outside the interval ('non-coverage').

Table 4 gives results in the same format as in Fig. 4. We do not present graphs of the empirical prediction error distributions because they all looked approximately normal. The non-coverage rates are fairly close to the nominal rates and appear to be somewhat better for the posterior standard deviation than for the prediction error standard deviation.

9.2. Predicted response probabilities

9.2.1. Prediction for a unit in a new cluster

We compare our recommended method, the population-averaged or marginal probability, $\hat{\pi}^{\mathrm{M}} = \bar{\mu}(\mathbf{x}^0)$, with the conditional probability $\hat{\pi}^{\mathrm{C}} = \hat{\mu}(\mathbf{x}^0, 0)$ given that $\zeta_j = 0$. The latter predictor, which is also the median probability, is easier to compute.

Table 4. Prediction error standard deviation $SD(\tilde{\zeta}_j^{EB} - \zeta_j)$ by parametric bootstrapping, 10th and 90th percentiles of $SD(\zeta_j|\mathbf{y}_j)$ (in parentheses) and non-coverage (per thousand) of confidence interval for ζ_j based on $SD(\tilde{\zeta}_j^{EB} - \zeta_j)$ and $SD(\zeta_j|\mathbf{y}_j)$ respectively (in square brackets)

n_j	*Results for the following values of ICC**			
	$ICC^* = 0.1$	$ICC^* = 0.2$	$ICC^* = 0.5$	$ICC^* = 0.8$
3	0.54 (0.54,0.54) [51,54]	0.74 (0.73,0.74) [52,52]	1.2 (1.1,1.2) [51.54]	2.0 (1.3,2.2) [49.49]
10	0.45 (0.44,0.45) [53,52]	0.55 (0.53,0.58) [53,52]	0.79 (0.63,1.1) [53,50]	1.5 (0.67,2.0) [57,49]
20	0.37 (0.36,0.38) [51,50]	0.43 (0.41,0.47) [48,48]	0.61 (0.44,0.78) [53,52]	1.3 (0.46,1.9) [61,50]
100	0.20 (0.19,0.21) [51,53]	0.22 (0.20,0.24) [50,46]	0.32 (0.20,0.47) [50,51]	0.83 (0.21,1.7) [66,50]

The ratio of the MSEP for the two methods depends on the intraclass correlation of the latent responses ICC^* and on the fixed part of the linear predictor, $\mathbf{x}^{0\prime}\hat{\boldsymbol{\beta}}$. We considered $\mathbf{x}^{0\prime}\hat{\boldsymbol{\beta}}$ ranging from 0 to 3 and computed both probabilities for the four values of the intraclass correlation of the latent responses that were used previously. Since the population-averaged probability gives the expected proportion of new units with $y_{ij} = 1$, the expectation of the squared error of prediction $(y_{ij} - \hat{\pi}_{ij})^2$ is

$$\bar{\mu}(\mathbf{x}^0)(1 - \hat{\pi}_{ij})^2 + \{1 - \bar{\mu}(\mathbf{x}^0)\}(0 - \hat{\pi}_{ij})^2.$$

Fig. 5 shows the ratio of the MSEP using the median probability *versus* the population-averaged probability as a function of $\mathbf{x}^{0\prime}\hat{\boldsymbol{\beta}}$ for the four values of the intraclass correlation. We see that the MSEP is never more than 5% greater for the median compared with the population-averaged probability if the intraclass correlation is 0.5 or less. However, for higher intraclass correlations the difference becomes more substantial, exceeding 15% for an intraclass correlation of 0.8 when the fixed part of the linear predictor exceeds 1.72. (For $ICC^* = 0.8$ and $\mathbf{x}^{0\prime}\boldsymbol{\beta} = 1.72$ we obtain $\hat{\pi}^M = 0.66$ and $\hat{\pi}^C = 0.85$.)

9.2.2. Prediction for a new unit in an existing cluster

We now compare the cluster-averaged or posterior mean probability $\tilde{\mu}_j(\mathbf{x}^0)$ with the conditional probability $\hat{\mu}(\mathbf{x}^0, \tilde{\zeta}_j^{EB})$ given that the random intercept equals its posterior mean. This is useful since the former is preferred but the latter can be obtained in most standard software. For the simulated data that were considered in the previous section, we deleted one response per cluster and subsequently predicted it by using the two methods. The ratio of the MSEP (across the 10000 clusters) was very close to 1 across conditions, the largest ratios being 1.02 for $ICC^* = 0.5$ and $n_j = 3$ and 1.05 for $ICC^* = 0.8$ and $n_j = 3$, when the posterior standard deviations tend to be large. Simulations with the fixed part of the linear predictor set to 1 and 2 also gave ratios close to 1, the largest ratios being 1.05 for conditions with $n_j = 3$. Substituting the empirical Bayes prediction into the expression for the conditional probability therefore is a reasonable approach for the range of conditions that are considered here.

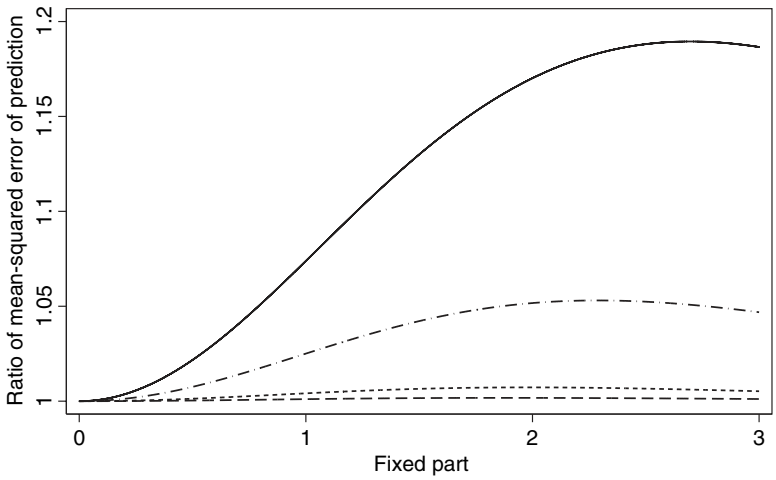

Fig. 5. Ratio of the MSEP comparing the median with population-averaged response probabilities for four values of the intraclass correlation of the latent responses when the fixed part $\mathbf{x}^{0\prime}\hat{\beta}$ ranges from 0 to 3 with ICC* = 0.1 (– – –), ICC* = 0.2 (-------), ICC* = 0.5 (· · · · · ·) and ICC* = 0.8 (———)

10. Concluding remarks

We have investigated prediction of random effects and of expected responses, including probabilities, in multilevel generalized linear models.

For prediction of random effects, we have concentrated on empirical Bayes prediction and discussed three different kinds of standard errors for the predictions: posterior standard deviations, prediction error standard deviations (comparative standard errors) and marginal sampling standard deviations (diagnostic standard errors). We have discussed the interpretation of these different notions of uncertainty and suggested approximations for some of the standard errors. For prediction of expected responses, or response probabilities, we have considered three different kinds of expectations: conditional expectations, population-averaged (or marginal) expectations and cluster-averaged (or posterior mean) expectations. We have discussed their use and shown how to obtain them. The methods have been illustrated by applying them to survey data on children nested in schools.

Our simulations for a random-intercept logistic regression model suggest that the sampling distribution of the empirical Bayes predictions is too discrete and non-normal for the diagnostic standard error to be used in the usual way for identifying outliers, except for cluster sizes of 100 or more combined with intraclass correlations of 0.5 or less, or cluster sizes of 20 or more combined with intraclass correlations of 0.1 or less. In these situations, the proposed approximation for the diagnostic standard error works well.

The sampling distribution of the prediction errors is quite normal across the range of intraclass correlations and cluster sizes that were considered, and using the marginal prediction error standard deviation as standard error produces adequate inferences based on the normal approximation. However, the posterior standard deviation is preferred from a theoretical perspective and performed somewhat better in the simulations. We therefore recommend using the posterior standard deviation as comparative standard error.

For predicting the response of a new unit in the random-intercept logistic regression model, we recommend using the population-averaged probability if the prediction is for a new cluster and the cluster-averaged probability if the prediction is for an existing cluster. A simpler alternative to the population-averaged probability is the conditional probability given that the random

intercept is 0. Our simulations showed that this alternative increases the MSEP substantially compared with the marginal probability if the intraclass correlation is high and the fixed part of the linear predictor is large. A simpler alternative to the posterior mean probability is the conditional probability given that the random intercept is equal to its posterior mean. This approach worked well for the range of situations that was considered.

Simulation results for predictions in linear mixed models were reported in Afshartous and de Leeuw (2005). Further work would be useful to investigate the performance of different types of predictions for response types other than continuous and dichotomous.

A great advantage of specifying statistical models is that they can be used for prediction. For instance, many of the predicted probabilities that were discussed in this paper could not be obtained by using generalized estimating equations. However, the quality of the predictions hinges on the appropriateness of the model specification. In particular, it has been found that a misspecified random-effects distribution can lead to poor performance of empirical Bayes prediction of the random effects (e.g. Rabe-Hesketh *et al.* (2003) and McCulloch and Neuhaus (2007)). To safeguard against such misspecification one might leave the distribution of the random effects unspecified and use non-parametric maximum likelihood estimation (see Clayton and Kaldor (1987) and Rabe-Hesketh *et al.* (2003) and the references therein).

Although we have focused on multilevel generalized linear models in this paper, the ideas extend directly to generalized latent variable models such as those described in Rabe-Hesketh *et al.* (2004) and Skrondal and Rabe-Hesketh (2004, 2007b). For these general models, as well as multilevel generalized linear models, almost all of the methods are implemented in `gllapred` and `gllasim`, the prediction and simulation commands of `gllamm` (e.g. Rabe-Hesketh and Skrondal, 2008b).

Acknowledgements

We are very grateful to the Guest Associate Editors and two reviewers for constructive comments that have helped to improve the paper considerably. We also thank the Research Council of Norway for a grant supporting our collaboration.

Appendix A

Here we give analytical results for linear multilevel or mixed models, $\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\zeta}_j + \boldsymbol{\varepsilon}_j$. The *empirical Bayes predictor* is

$$\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}} = \hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}), \tag{8}$$

where $\hat{\boldsymbol{\Sigma}}_j \equiv \mathbf{Z}_j\hat{\boldsymbol{\Psi}}\mathbf{Z}_j' + \hat{\boldsymbol{\Theta}}_j$ is the estimated residual covariance matrix of $\mathbf{y}_j$. The *maximum likelihood estimator* is

$$\tilde{\boldsymbol{\zeta}}_j^{\mathrm{ML}} = (\mathbf{Z}_j'\hat{\boldsymbol{\Theta}}_j^{-1}\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\hat{\boldsymbol{\Theta}}_j^{-1}(\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}). \tag{9}$$

The empirical *posterior covariance matrix* and *marginal prediction error covariance matrix* are (e.g. Searle *et al.* (1992))

$$\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j, \mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}} - \boldsymbol{\zeta}_j|\mathbf{X}_j, \mathbf{Z}_j; \hat{\boldsymbol{\vartheta}}) = \hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{Z}_j\hat{\boldsymbol{\Psi}}. \tag{10}$$

For fixed $\hat{\boldsymbol{\Psi}}$ and $\hat{\theta}$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is just the generalized least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{J}\mathbf{X}_j'\boldsymbol{\Sigma}_j^{-1}\mathbf{X}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{X}_j'\boldsymbol{\Sigma}_j^{-1}\mathbf{y}_j.$$

It therefore follows from results derived in Harville (1976) that the posterior covariance matrix and marginal prediction error covariance matrix, taking the uncertainty of the estimated regression parameters

into account, become

$$\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\Psi}},\hat{\theta})=\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}-\boldsymbol{\zeta}_j|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\Psi}},\hat{\theta})=\hat{\boldsymbol{\Psi}}-\hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{Z}_j\hat{\boldsymbol{\Psi}}+\hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{X}_j\,\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{X}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{Z}_j\hat{\boldsymbol{\Psi}},$$

where

$$\mathrm{cov}(\hat{\boldsymbol{\beta}})=\left(\sum_{j=1}^{J}\mathbf{X}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{X}_j\right)^{-1}$$

is the covariance matrix of the generalized least squares estimator.

The *marginal sampling covariance matrix* of the empirical Bayes predictions is

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})=\hat{\boldsymbol{\Psi}}-\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})=\hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{Z}_j\hat{\boldsymbol{\Psi}}. \tag{11}$$

If $\hat{\boldsymbol{\beta}}$ is estimated by maximum likelihood for fixed $\hat{\boldsymbol{\Psi}}$ and $\hat{\theta}$ (generalized least squares) the marginal sampling covariance matrix, taking the uncertainty of the estimated regression parameters into account, becomes

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\Psi}},\hat{\theta})=\hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{Z}_j\hat{\boldsymbol{\Psi}}-\hat{\boldsymbol{\Psi}}\mathbf{Z}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{X}_j\,\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{X}_j'\hat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{Z}_j\hat{\boldsymbol{\Psi}}. \tag{12}$$

Appendix B

Proposition 1.

$$\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})=\hat{\boldsymbol{\Psi}}-E_{\mathbf{y}}\{\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})\}.$$

Proof.

$$\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})=E_{\mathbf{y}}\{\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})\}+\mathrm{cov}_{\mathbf{y}}\{E(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})\}$$

$$\Updownarrow$$

$$\underbrace{\mathrm{cov}_{\mathbf{y}}\{E(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})\}}_{\mathrm{cov}_{\mathbf{y}}(\tilde{\boldsymbol{\zeta}}_j^{\mathrm{EB}}|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})}=\underbrace{\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})}_{\hat{\boldsymbol{\Psi}}}-E_{\mathbf{y}}\{\mathrm{cov}(\boldsymbol{\zeta}_j|\mathbf{y}_j,\mathbf{X}_j,\mathbf{Z}_j;\hat{\boldsymbol{\vartheta}})\}.$$

We first use a useful identity for covariance matrices and the equivalence then follows from rearranging the terms. Finally, we use the definition of the empirical Bayes predictor and the symbol for the covariance matrix of the random effects. The proposition was used by Skrondal (1996).

References

Adams, R. (2002) Scaling PISA cognitive data. In *PISA 2000 Technical Report* (eds R. Adams and M. Wu), pp. 99–108. Paris: Organisation for Economic Co-operation and Development.

Afshartous, D. and de Leeuw, J. (2005) Prediction in multilevel models. *J. Educ. Behav. Statist.*, **30**, 109–139.

Afshartous, D. and Wolf, M. (2007) Avoiding 'data snooping' in multilevel and mixed effects models. *J. R. Statist. Soc.* A, **170**, 1035–1059.

Bartlett, M. S. (1938) Methods of estimating mental factors. *Nature*, **141**, 609–610.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. New York: Wiley.

Bock, R. D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Bock, R. D. and Mislevy, R. J. (1982) Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Measmnt*, **6**, 431–444.

Bondeson, J. (1990) Prediction in random coefficient regression models. *Biometr. J.*, **32**, 387–405.

Booth, J. G. and Hobert, J. P. (1998) Standard errors of prediction in generalized linear mixed models. *J. Am. Statist. Ass.*, **93**, 262–272.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Candel, M. J. J. M. (2004) Performance of empirical bayes estimators of random coefficients in multilevel analysis: some results for the random intercept-only model. *Statist. Neerland.*, **58**, 197–219.

Candel, M. J. J. M. (2007) Empirical bayes estimators of the random intercept in multilevel analysis: performance of the classical, Morris and Rao version. *Computnl Statist. Data Anal.*, **51**, 3027–3040.

Carlin, B. P. and Louis, T. A. (2000a) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
Carlin, B. P. and Louis, T. A. (2000b) Empirical Bayes: past, present and future. *J. Am. Statist. Ass.*, **95**, 1286–1289.
Chamberlain, G. (1984) Panel data. In *Handbook of Econometrics*, vol. II (eds Z. Griliches and M. D. Intriligator), pp. 1247–1318. Amsterdam: North-Holland.
Chang, H. and Stout, W. (1993) The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, **58**, 37–52.
Clayton, D. G. (1996) Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 275–301. London: Chapman and Hall.
Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.
Deely, J. J. and Lindley, D. V. (1981) Bayes empirical Bayes. *J. Am. Statist. Ass.*, **76**, 833–841.
Demidenko, E. (2004) *Mixed Models: Theory and Applications*. New York: Wiley.
Duchateau, L. and Janssen, P. (2005) Understanding heterogeneity in mixed, generalized mixed and frailty models. *Am. Statistn*, **59**, 143–146.
Efron, B. and Morris, C. (1973) Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Am. Statist. Ass.*, **68**, 117–130.
Efron, B. and Morris, C. (1975) Data analysis using Stein's estimator and its generalizations. *J. Am. Statist. Ass.*, **70**, 311–319.
Embretson, S. E. and Reise, S. P. (2000) *Item Response Theory for Psychologists*. Mahwah: Erlbaum.
Farrell, P. J., MacGibbon, B. and Tomberlin, T. J. (1997) Bootstrap adjustments for empirical Bayes interval estimates of small-area proportions. *Can. J. Statist.*, **25**, 75–89.
Fearn, T. (1975) A Bayesian approach to growth curves. *Biometrika*, **62**, 89–100.
Frees, E. W. and Kim, J.-S. (2006) Multilevel model prediction. *Psychometrika*, **71**, 79–104.
Ganzeboom, H. G. B., De Graaf, P., Treiman, D. J. and de Leeuw, J. (1992) A standard international socio-economic index of occupational status. *Socl Sci. Res.*, **21**, 1–56.
Gibbons, R. D., Hedeker, D., Charles, S. C. and Frisch, P. (1994) A random-effects probit model for predicting medical malpractice claims. *J. Am. Statist. Ass.*, **89**, 760–767.
Goldberger, A. S. (1962) Best linear unbiased prediction in the generalized linear regression model. *J. Am. Statist. Ass.*, **57**, 369–375.
Goldstein, H. (1995) *Multilevel Statistical Models*, 2nd edn. London: Arnold.
Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn. London: Arnold.
Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *J. R. Statist. Soc.* A, **159**, 385–409.
Hall, P. and Maiti, T. (2006) On parametric bootstrap methods for small area prediction. *J. R. Statist. Soc.* B, **68**, 221–238.
Harville, D. A. (1976) Extension of the Gauss-Markov theorem to include the estimation of random effects. *Ann. Statist.*, **2**, 384–395.
Hoijtink, H. and Boomsma, A. (1995) On person parameter estimation in the dichotomous Rasch model. In *Rasch Models: Foundations, Recent Developments, and Applications* (eds G. H. Fischer and I. W. Molenaar), pp. 53–68. New York: Springer.
Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.
Jiang, J. and Lahiri, P. (2001) Empirical best prediction for small area inference with binary data. *Ann. Inst. Statist. Math.*, **53**, 217–243.
Kackar, R. N. and Harville, D. A. (1984) Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Am. Statist. Ass.*, **79**, 853–862.
Kass, R. E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Statist. Ass.*, **84**, 717–726.
Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Am. Statist. Ass.*, **82**, 739–757.
Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.
Lange, N. and Ryan, L. M. (1989) Assessing normality in random effects models. *Ann. Statist.*, **17**, 624–642.
Langford, I. H. and Lewis, T. (1998) Outliers in multilevel data (with discussion). *J. R. Statist. Soc.* A, **161**, 121–160.
Lawley, D. N. and Maxwell, A. E. (1971) *Factor Analysis as a Statistical Method*. London: Butterworth.
Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.
Longford, N. T. (2001) Simulation-based diagnostics in random-coefficient models. *J. R. Statist. Soc.* A, **164**, 259–273.
Louis, T. A. (1984) Bayes and empirical Bayes estimates of a population of parameter values. *J. Am. Statist. Ass.*, **79**, 393–398.
Ma, X., Ma, L. and Bradley, K. D. (2008) Using multilevel modeling to investigate school effects. In *Multilevel Modelling of Educational Data* (eds A. A. O'Connell and D. B. McCoach), pp. 59–110. Charlotte: Information Age Publishing.

Maritz, J. S. and Lwin, T. (1989) *Empirical Bayes Methods*. London: Chapman and Hall.
McCulloch, C. E. (1997) Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–170.
McCulloch, C. E. and Neuhaus, J. (2007) Prediction of random effects and effects of misspecification of their distribution. *West Coast Stata Users Group Meet.*, *Marina Del Rey*. (Available from `http://repec.org/wcsug2007/12.html`.)
McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008) *Generalized, Linear and Mixed Models*, 2nd edn. New York: Wiley.
Mislevy, R. J. (1986) Recent developments in the factor analysis of categorical variables. *J. Educ. Statist.*, **11**, 3–31.
Morris, C. (1983) Parametric empirical Bayes inference: theory and applications. *J. Am. Statist. Ass.*, **78**, 47–65.
Organisation for Economic Co-operation and Development (2000) *Manual for the PISA 2000 Database*. Paris: Organisation for Economic Co-operation and Development. (Available from `http://www.pisa.oecd.org/dataoecd/53/18/33688135.pdf`.)
Pinheiro, J. C. and Bates, D. M. (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Computnl Graph. Statist.*, **4**, 12–35.
Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2003) Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statist. Modllng*, **3**, 215–232.
Rabe-Hesketh, S. and Skrondal, A. (2006) Multilevel modelling of complex survey data. *J. R. Statist. Soc.* A, **169**, 805–827.
Rabe-Hesketh, S. and Skrondal, A. (2008a) Generalized linear mixed effects models. In *Longitudinal Data Analysis* (eds G. M. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), pp. 79–106. Boca Raton: Chapman and Hall–CRC.
Rabe-Hesketh, S. and Skrondal, A. (2008b) *Multilevel and Longitudinal Modeling using Stata*, 2nd edn. College Station: Stata Press.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004) Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167–190.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005) Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econometr.*, **128**, 301–323.
Rao, C. R. (1975) Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics*, **31**, 545–554.
Rao, J. N. K. (2003) *Small Area Estimation*. New York: Wiley.
Raudenbush, S. W. and Bryk, A. S. (2002) *Hierarchical Linear Models*. Thousand Oaks: Sage.
Raudenbush, S. W. and Willms, J. D. (1995) Estimation of school effects. *J. Educ. Behav. Statist.*, **20**, 307–335.
Reinsel, G. C. (1984) Estimation and prediction in a multivariate random effects generalized linear model. *J. Am. Statist. Ass.*, **79**, 406–414.
Reinsel, G. C. (1985) Mean squared error properties of empirical Bayes estimators in a multivariate random effects general linear model. *J. Am. Statist. Ass.*, **80**, 642–650.
Robbins, H. (1955) An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability* (ed. J. Neyman), pp. 157–164. Berkeley: University of California Press.
Robinson, G. K. (1991) That BLUP is a good thing: the estimation of random effects. *Statist. Sci.*, **6**, 15–51.
Rose, C. E., Hall, D. B., Shiver, B. D., Clutter, M. L. and Borders, B. (2006) A multilevel approach to individual tree survival prediction. *For. Sci.*, **52**, 31–43.
Rosenberg, B. (1973) Linear regression with randomly dispersed parameters. *Biometrika*, **60**, 65–72.
Rubin, D. B. (1980) Using empirical Bayes techniques in the law school validity studies. *J. Am. Statist. Ass.*, **75**, 801–827.
Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.
Rumberger, R. W. and Palardy, G. J. (2005) Does segregation still matter? The impact of student composition on academic achievement in high school. *Teach. Coll. Rec.*, **107**, 1999–2045.
Schilling, S. G. and Bock, R. D. (2005) High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, **70**, 533–555.
Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. New York: Wiley.
Skrondal, A. (1996) *Latent Trait, Multilevel and Repeated Measurement Modelling with Incomplete Data of Mixed Measurement Levels*. Oslo: UiO.
Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman and Hall–CRC.
Skrondal, A. and Rabe-Hesketh, S. (2007a) Redundant overdispersion parameters in multilevel models. *J. Educ. Behav. Statist*, **32**, 419–430.
Skrondal, A. and Rabe-Hesketh, S. (2007b) Latent variable modelling: a survey. *Scand. J. Statist.*, **34**, 712–745.
Smith, A. F. M. (1973) A general Bayesian linear model. *J. R. Statist. Soc.* B, **35**, 67–75.
Strenio, J. L. F., Weisberg, H. I. and Bryk, A. S. (1983) Empirical Bayes estimation of individual growth curve parameters and their relations to covariates. *Biometrics*, **39**, 71–86.

Swamy, P. A. V. B. (1970) Efficient inference in a random coefficient regression model. *Econometrica*, **38**, 311–323.
Ten Have, T. R. and Localio, A. R. (1999) Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics*, **55**, 1022–1029.
Thomson, G. H. (1938) *The Factorial Analysis of Human Ability*. London: University of London Press.
Thurstone, L. L. (1935) *The Vectors of Mind*. Chicago: University of Chicago Press.
Tsutakawa, R. K. and Johnson, J. C. (1990) The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, **55**, 371–390.
Vidoni, P. (2006) Response prediction in mixed effects models. *J. Statist. Planng Inf.*, **136**, 3948–3966.
Vonesh, E. F. and Chinchilli, V. M. (1997) *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Dekker.
Ware, J. H. and Wu, M. C. (1981) Tracking: prediction of future values from serial measurements. *Biometrics*, **37**, 427–437.
Warm, T. A. (1989) Weighted likelihood estimation of ability in item response models. *Psychometrika*, **54**, 427–450.
Willms, J. D. (1986) Social class segregation and its relationship to pupils' examination results in Scotland. *Am. Sociol. Rev.*, **51**, 224–241.