J. R. Statist. Soc. B (1998)
60, *Part* 3, *pp.* 497–536

Some algebra and geometry for hierarchical models, applied to diagnostics

James S. Hodges†

University of Minnesota, Minneapolis, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 15th, 1997,* Professor P. J. Green *in the Chair*]

Summary. Recent advances in computing make it practical to use complex hierarchical models. However, the complexity makes it difficult to see how features of the data determine the fitted model. This paper describes an approach to diagnostics for hierarchical models, specifically linear hierarchical models with additive normal or *t*-errors. The key is to express hierarchical models in the form of ordinary linear models by adding artificial 'cases' to the data set corresponding to the higher levels of the hierarchy. The error term of this linear model is not homoscedastic, but its covariance structure is much simpler than that usually used in variance component or random effects models. The re-expression has several advantages. First, it is extremely general, covering dynamic linear models, random effect and mixed effect models, and pairwise difference models, among others. Second, it makes more explicit the geometry of hierarchical models, by analogy with the geometry of linear models. Third, the analogy with linear models provides a rich source of ideas for diagnostics for all the parts of hierarchical models. This paper gives diagnostics to examine candidate added variables, transformations, collinearity, case influence and residuals.

Keywords: Bayesian methods; Dynamic linear models; Multilevel models; Random effect models; Spatial data; Time varying regression; Variance components

1. Introduction

A research project had a data set describing 341 state-based health maintenance organizations (HMOs) serving US Government employees. Such HMOs operate in 42 states, the District of Columbia, Guam and Puerto Rico; the data set contains between 1 and 31 plans per jurisdiction with a median of 5. (I shall call these jurisdictions 'states'.) These data were analysed as part of estimating the cost of moving military retirees and dependents from a Defense Department health plan to plans serving US Government employees. One quantity of interest is each HMO's monthly premium (in US dollars) for individual subscribers. Fig. 1 summarizes the HMO data, with each dot representing a state's average monthly premium. (These data, current at September 1992, were obtained from the agency that administers health benefits for US Government employees.)

Plans in the same state are more likely to have similar premiums than plans in different states. Thus, if y_{ij} is the premium for plan j in state i, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, 45$, we might model these data as

$$y_{ij}|\theta_i, \sigma^2 \sim N(\theta_i, \sigma^2), \tag{1.1}$$

$$\theta_i|\mu, \tau^2 \sim N(\mu, \tau^2). \tag{1.2}$$

†*Address for correspondence*: Division of Biostatistics, University of Minnesota, Suite 200, 2221 University Avenue, Minneapolis, MN 55414, USA.
E-mail: hodges@gopher.ccbr.umn.edu

 1369–7412/98/60497

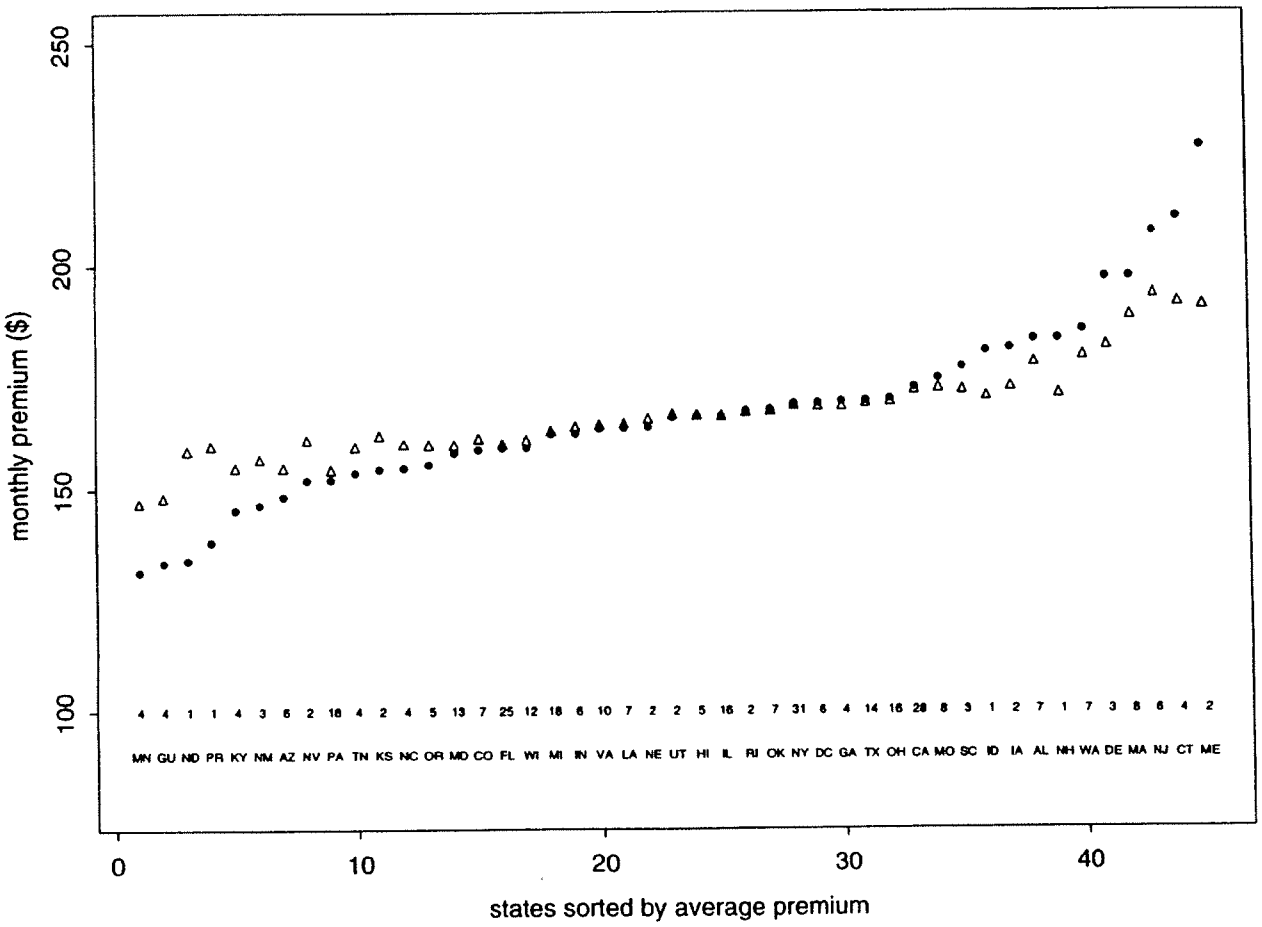

Fig. 1. Summaries and fitted values for the HMO data: ●, state average premium; △, posterior mean from the random effect model (Section 5.1) (each state's data are labelled at the bottom of the figure by the state's two-letter abbreviation and its number of plans)

In states currently without HMOs serving US Government employees, the mean premium would be modelled as a draw from distribution (1.2). A Bayesian analysis would add prior distributions for μ, σ^2 and τ^2. This is the simplest hierarchical model, 'hierarchical' because it is specified one level at a time: first, the model for the data, then the model for the θ_i. Bryk and Raudenbush (1992) give a thorough accessible introduction to hierarchical linear models.

This hierarchical model can be elaborated. If we add regressors describing plans and states, a model for the premium for plan j in state i would be

$$y_{ij}|\theta_i, \sigma^2 \sim N(\mathbf{x}_{ij}\theta_i, \sigma^2), \tag{1.3}$$

$$\theta_i|\mu, \Sigma \sim N(\mathbf{z}_i\mu, \Sigma), \tag{1.4}$$

where $\mathbf{x}_{ij}$ is a row p-vector of covariates for plan (i, j), θ_i is a column p-vector of regression coefficients for state i, $\mathbf{z}_i$ is a $p \times q$ matrix of covariates for state i and μ is a q-vector of regression coefficients. A Bayesian analysis would add priors for μ, σ^2 and Σ.

This paper proposes a method of examining hierarchical models that is analogous to the algebraic or geometric theory of linear models. The latter yields deep insight into linear models; by extending it to hierarchical models, similar insight and analytical power might be available. This paper's purpose is to extend the theory and to use it to produce specific diagnostic tools. The approach is applicable to random and mixed effect models, to dynamic linear models (time-varying regressions) used as smoothers and to some spatial models. The

common features of these models are a mean structure that is linear in unknown parameters and additive normal or t-errors. For simplicity, I refer to these all as hierarchical models, although some cannot be expressed as strict hierarchies.

Section 2 reformulates hierarchical models as linear models with simple error covariances, building a bridge to the theory of ordinary linear models. Section 3 provides some background for using the reformulation to derive diagnostics. Section 4 derives an added variable plot, a variable transformation diagnostic, collinearity measures, case influence diagnostics and residuals; Section 5 applies them to the HMO data. This reformulation has uses besides diagnostics; these are mentioned briefly in Section 6.

2. The reformulation: hierarchical models as linear models

2.1. Motivating examples

2.1.1. The one-way random effects model

For model (1.1)–(1.2), suppose that a Bayesian analysis uses an $N(M, s^2)$ prior for μ, with M and s^2 specified. Then the model's mean structure and associated prior are

$$y_{ij} = \theta_i + \epsilon_{ij}, \tag{2.1}$$

$$\theta_i = \mu + \delta_i, \tag{2.2}$$

$$\mu = M + \xi, \tag{2.3}$$

where $i = 1, \ldots, 45, j = 1, \ldots, n_i$, $\epsilon_{ij} \sim N(0, \sigma^2)$, $\delta_i \sim N(0, \tau^2)$ and $\xi \sim N(0, s^2)$. A Bayesian analysis would add prior distributions for σ^2 and τ^2. Rewrite equations (2.2) and (2.3) as

$$0 = -\theta_i + \mu + \delta_i, \tag{2.4}$$

$$M = \mu - \xi. \tag{2.5}$$

Equations (2.1), (2.4) and (2.5) have the form of a linear model; the left-hand side of each equation is known and the right-hand side is a linear function of unknown parameters with additive errors:

$$\left(\begin{array}{c} \mathbf{y} \\ \hline \mathbf{0}_{45\times 1} \\ \hline M \end{array}\right) = \left(\begin{array}{ccc|c} \mathbf{1}_{n_1} & \cdots & \mathbf{0}_{n_1} & \\ \vdots & \ddots & \vdots & \mathbf{0}_{341\times 1} \\ \mathbf{0}_{n_{45}} & \cdots & \mathbf{1}_{n_{45}} & \\ \hline & -\mathbf{I}_{45} & & \mathbf{1}_{45} \\ \hline & \mathbf{0}_{1\times 45} & & 1 \end{array}\right) \left(\begin{array}{c} \theta_1 \\ \vdots \\ \theta_{45} \\ \hline \mu \end{array}\right) + \left(\begin{array}{c} \epsilon \\ \hline \delta \\ \hline -\xi \end{array}\right), \tag{2.6}$$

where $\mathbf{I}_N$ is the identity matrix of dimension N, $\mathbf{1}_N$ is a column N-vector of 1s, $\mathbf{0}_{a\times b}$ is a matrix of 0s with dimension $a \times b$, $\mathbf{y}$ and ϵ are column 341-vectors of y_{ij} and ϵ_{ij} respectively, ordered with j changing most quickly, and δ is the column 45-vector of δ_j.

In the usual notation, equation (2.6) can be summarized as

$$\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E} \tag{2.7}$$

where $\mathbf{Y}$ and $\mathbf{X}$ are known, Θ is unknown and $\mathbf{E}$ is an unobserved error term. $\mathbf{E}$ has mean 0 and a diagonal covariance matrix Γ, the first 341 diagonal elements of Γ being σ^2, the next 45 τ^2 and the last s^2. A non-Bayesian analysis would drop the rows of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{E}$ corresponding to equation (2.5). A moment's inspection should convince the reader that a Bayesian

inference is the same if it uses equation (2.6) and particular priors for σ^2 and τ^2 or if it uses the same priors for σ^2 and τ^2 and the usual formulation in which equations (2.1), (2.4) and (2.5) are not combined into a single expression. The elaborated model (1.3)–(1.4) is easily reformulated as a linear model by analogy with equations (2.1), (2.4) and (2.5).

2.1.2. A spatial model: pairwise differences (Besag and Kempton, 1986)

Suppose that plots in a field experiment are arrayed in a long row and labelled $i = 1, \ldots, N$, plots i and $i+1$ being adjacent. Suppose that two treatments are allocated randomly to the plots, and let y_i be the yield of plot i. Let F_i be the (unobserved) fertility of plot i, for which the experimenter would like to account in the analysis. We might use the model

$$y_i = T_i\theta + F_i + \epsilon_i \tag{2.8}$$

where T_i is a treatment indicator, θ is the treatment effect and $\epsilon_i \sim N(0, \sigma^2)$. Suppose that the F_i change fairly smoothly, modelled as

$$F_i = F_{i-1} + \delta_i \tag{2.9}$$

for $i = 2, \ldots, N$, where $\delta_i \sim N(0, \tau^2)$, and that the experimenter adds a prior $\theta \sim N(M, s^2)$. By manipulating equation (2.9) as before, this pairwise differences model can be represented as

$$\begin{pmatrix} \mathbf{y} \\ \hline \mathbf{0}_{(N-1)\times 1} \\ \hline M \end{pmatrix} = \left(\begin{array}{cc} \mathbf{T} & \mathbf{I}_N \\ \hline \mathbf{0}_{(N-1)\times 1} & \Delta \\ \hline 1 & \mathbf{0}_{1\times N} \end{array}\right) \left(\begin{pmatrix} \theta \\ F_1 \\ \vdots \\ F_N \end{pmatrix} \right) + \left(\begin{array}{c} \epsilon \\ \hline \delta \\ \hline -\xi \end{array}\right) \tag{2.10}$$

where $\mathbf{T}$ is the N-vector of T_i and

$$\Delta = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & & 0 & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \tag{2.11}$$

is an $(N-1) \times N$ matrix.

2.1.3. Other models

Every hierarchical linear model in Bryk and Raudenbush (1992) can be reformulated as above. Duncan and Horn (1972) gave the reformulation for the Kalman smoother; this extends easily to every linear model with additive errors in West and Harrison (1989). Hilden-Minton (1995), section 2.1, gave the reformulation for mixed and random effect models; this extends easily to the models in chapters 1–9 of Searle *et al.* (1992).

2.2. The general form

The models shown so far have had three levels: one for the data, one modelling the parameters closest to the data and one for the mean structure's prior. A four-level example is a model with students within classrooms within schools. The reformulation in Section 2.1 extends trivially to such models.

In general, then, hierarchical models can be re-expressed as

$$\left(\begin{array}{c}\mathbf{y}\\\hline\mathbf{0}\\\hline\mathbf{M}\end{array}\right)=\left(\begin{array}{c|cccccc}\mathbf{X}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0}\\\hline \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}_3 & \mathbf{H}_4 & \cdots & \mathbf{0} & \mathbf{0}\\ \vdots & \vdots & & \ddots & & \vdots\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{H}_{r-1} & \mathbf{H}_r\\\hline \mathbf{G}_1 & \mathbf{G}_2 & \mathbf{G}_3 & \cdots & \mathbf{G}_{s-1} & \mathbf{G}_s\end{array}\right)\left(\begin{array}{c}\Theta_1\\\hline\Theta_2\end{array}\right)+\left(\begin{array}{c}\epsilon\\\hline\delta\\\hline\xi\end{array}\right), \tag{2.12}$$

where $\mathbf{0}$ represents suitably sized matrices of 0s. Equation (2.12) can be summarized by $\mathbf{Y}=\mathbf{X}\Theta+\mathbf{E}$, as in equation (2.7). The vector Θ_1 contains the smallest set of parameters such that, conditional on it, $\mathbf{y}$'s expectation is specified. $\mathbf{H}_1$ and $\mathbf{H}_2$ specify a model for Θ_1 as a function of second-level parameters in Θ_2. For the random effects model (2.6), $\mathbf{H}_1=-\mathbf{I}_{45}$ and $\mathbf{H}_2=\mathbf{1}_{45}$. Similarly, $\mathbf{H}_3$ and $\mathbf{H}_4$ specify a model for the second-level parameters as a function of third-level parameters in Θ_2, and so on. The $\mathbf{G}_j$ and $\mathbf{M}$ specify prior means for parameters in Θ that are not modelled as functions of higher level parameters.

The covariance matrix of $\mathbf{E}$, Γ, is block diagonal. Its uppermost block is Γ_1, the covariance for ϵ. The second block is Γ_2, the covariance for δ. The final block is $\text{cov}(\xi)=\Gamma_3$, the prior covariance, which is known. Γ is generally simpler than covariance matrices in random effects formulations of hierarchical models.

Some terminology will be helpful. In equation (2.12), the rows of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{E}$ corresponding to $\mathbf{X}_1$ are *data cases*; those corresponding to the $\mathbf{H}_j$ are *constraint cases*, because they specify stochastic constraints on the parameters in Θ_1, and those corresponding to the $\mathbf{G}_j$ are *prior cases*. Non-Bayesian analyses would omit the prior cases; Bayesian analyses would add prior distributions for unknown parameters in Γ_1 and Γ_2. 'Constraint case' describes non-data cases with unknown variances; prior cases have known (i.e. specified) variances.

Consider a model in the form (2.12), in which $\mathbf{G}_1$ is set to 0. If f denotes a generic probability density, the posterior density for Θ_1, Θ_2 and Γ is

$$f(\Theta_1, \Theta_2, \Gamma|\mathbf{y}) \propto f(\mathbf{y}|\Theta_1, \Theta_2, \Gamma)\, f(\Theta_1, \Theta_2, \Gamma) \tag{2.13}$$

by Bayes's theorem. By straightforward manipulations of expression (2.13),

$$\begin{aligned} f(\Theta_1, \Theta_2, \Gamma|\mathbf{y}) &\propto f(\mathbf{y}|\Theta_1, \Theta_2, \Gamma)\, f(\Theta_1, \Theta_2|\Gamma)\, f(\Gamma)\\ &= f(\mathbf{y}, \Theta_1, \Theta_2|\Gamma)\, f(\Gamma). \end{aligned} \tag{2.14}$$

Equations (2.7) and (2.12) represent $f(\mathbf{y}, \Theta_1, \Theta_2|\Gamma)$, so we might think of it as a pseudo-likelihood, 'pseudo' because Θ_1 and Θ_2 are unobserved (Nelder, 1972). Equation (2.14) also establishes that, for the models included, a Bayesian analysis of the reformulation (2.7), with $f(\Gamma)$ as Γ's prior, is identical with a Bayesian analysis of the original formulation using $f(\Gamma)$. This extends easily to all models covered by the general case.

It is convenient to use a reduced general form:

$$\left(\begin{array}{c}\mathbf{y}\\\hline\mathbf{0}\\\hline\mathbf{M}\end{array}\right)=\left(\begin{array}{c|c}\mathbf{X}_1 & \mathbf{0}\\\hline \mathbf{Z}_1 & \mathbf{Z}_2\\\hline \mathbf{W}_1 & \mathbf{W}_2\end{array}\right)\left(\begin{array}{c}\Theta_1\\\hline\Theta_2\end{array}\right)+\left(\begin{array}{c}\epsilon\\\hline\delta\\\hline\xi\end{array}\right) \tag{2.15}$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{W}_1$ and $\mathbf{W}_2$ are defined by matching matrix partitions in equations (2.12) and

(2.15), and **0** represents suitably sized matrices of 0s. This paper uses the reduced general form; extensions to the full general form are simple. Parts of this form can be null; for example, Θ_2 and the corresponding parts of **X** are absent in the pairwise differences model (2.10).

2.3. Antecedents

Hierarchical models were first discussed in generality by Lindley and Smith (1972), who used the hierarchical structure to specify a prior distribution for Θ_1 in stages. The reformulation has several antecedents. Henderson *et al.* (1959) used data and constraint cases ('mixed model equations') to compute best linear unbiased predictors. These equations have also been used in computing maximum likelihood estimates for variance components (Searle *et al.* (1992), section 7.6). Theil (1971) used prior cases to represent prior information about regression coefficients. In regression models, Salkever (1976) and Fuller (1980) obtained a frequentist predictive distribution for an observable y_f by adding y_f to the parameter vector and adding the artificial case $0 = \mathbf{x}'\beta - y_f + \epsilon$. Duncan and Horn (1972) re-expressed dynamic linear models as linear models to prove Gauss–Markov theorems about Kalman filters. Fellner (1986) used data and constraint cases to construct M-estimators for variance components. No doubt others have used the reformulation as well.

3. Diagnostics using the re-expression: background

3.1. Previous approaches to diagnostics

Bryk and Raudenbush (1992) gave a broad treatment of analysing data with a hierarchical structure, providing practical advice on building hierarchical models and checking some of their assumptions. Others have focused on specific problems, including outliers (Beckman *et al.*, 1987; Chaloner, 1994; Ho *et al.*, 1996; Hocking *et al.*, 1989; Langford and Lewis, 1998; Moulton, 1987; Sharples, 1990; Weiss, 1995, 1996; Weiss and Lazaro, 1992), normality of residuals (Dempster and Ryan, 1985; Lange and Ryan, 1989; Solomon, 1985), influential observations (Beckman *et al.*, 1987; Bradlow and Zaslavsky, 1997; Carlin and Polson, 1992; Christensen *et al.*, 1992; Harrison and West, 1991; Peña, 1991), selecting functional form or explanatory variables (Hall *et al.*, 1996; Wakefield, 1996; Wakefield and Bennett, 1996) and overall fit (Dey *et al.*, 1995).

Without judging the merit of these methods, it is noteworthy that none takes full advantage of the structure of hierarchical linear models. Diagnostics for ordinary linear models rely on the algebra and geometry of the column space of the design matrix (Cook and Weisberg, 1982; Atkinson, 1985), providing a powerful way to understand inferential summaries as functions of the data. Hierarchical models are more complicated because the variances at the different levels of the hierarchy make estimation non-linear in the observations: the geometry changes as the variances change. None-the-less, as will become clear, equation (2.7) provides a fruitful way to adapt linear model methods to hierarchical models.

Louis (1988) and Hilden-Minton (1995) did take advantage of the structure of hierarchical models. Each considered models in the form

$$\mathbf{Q}_i = \mathbf{C}_i\kappa + \mathbf{D}_i\zeta_i + \nu_i \tag{3.1}$$

where $\mathbf{Q}_i$ is the vector of observations for unit i, $\mathbf{C}_i$ is unit i's design matrix for the fixed effect vector κ, $\mathbf{D}_i$ is unit i's design matrix for the random effect vector ζ_i and ν_i is unit i's vector of

errors. Louis (1988) integrated out ζ_i to obtain a fixed effect model with a complex error covariance, then transformed to 'components with independent errors' and applied methods for ordinary linear models. This approach has two disadvantages. First, it focuses on the fixed effects κ; in many cases, the random effects are of primary interest (e.g. Bryk and Raudenbush (1992), section 1). Also, the method appears to be sensitive to estimates of the variance components. In the approach given here, equation (2.7) is often premultiplied by $\mathbf{\Gamma}^{-1/2}$ in deriving a diagnostic method. However, as will be shown, the resulting methods are generally not sensitive to $\mathbf{\Gamma}$.

Hilden-Minton's (1995) approach is closest to the method given here, in that it does not radically transform the geometry of the raw problem. Some diagnostics given below may not be derivable by using Hilden-Minton's approach, but otherwise his approach and that given here appear to be complementary.

3.2. Some general issues

3.2.1. Model fitting and computing

Model fitting methods that maximize the likelihood or pseudolikelihood are popular (e.g. Bryk and Raudenbush (1992)). However, the mode of the posterior or pseudolikelihood can be problematic because it often takes a value on the boundary of the parameter space ($\tau^2 = 0$ in the random effects model). Even when the mode is an interior point, approximating the posterior or pseudolikelihood by its curvature near the mode can misrepresent the function's shape (O'Hagan, 1976, 1985). However, integrating out the random effects in $\mathbf{\Theta}_1$ gives awkward equations for the posterior modes or maximum likelihood estimates (Searle *et al.*, 1992). Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990; Casella and George, 1992; Smith and Roberts, 1993) avoid the problems of modes and are straightforward for hierarchical models (e.g. Lange *et al.* (1992)). For some types of models, mode-based methods are not as subject to the problems noted above, but seem to have no advantage over MCMC methods.

Thus, before applying the diagnostics given below, we would begin with a sequence of draws from a suitable MCMC sampler; the discussion presumes that the sequence of draws is available. It appears necessary to use a proper prior for variances of constraint cases (τ^2 in the random effects model) to avoid numerical problems, improper posteriors and poor frequentist properties. (See DuMouchel and Waternaux (1992) for a brief discussion of this underexamined problem.) No problems are apparent with an improper prior for the data case variance (σ^2 in the random effects model) or with infinite $\mathbf{\Gamma}_3$.

Formulating a hierarchical model as in equation (2.15) allows a simple Gibbs sampler. Conditionally on $\mathbf{\Gamma}$, the posterior distribution of $\mathbf{\Theta}$ is multivariate normal with mean and covariance that are formally identical with the generalized least squares estimate of $\mathbf{\Theta}$ and its covariance respectively. Conditionally on $\mathbf{\Theta}$, the posterior distributions of unknown parameters in $\mathbf{\Gamma}$ have simple forms if these parameters have conjugate priors. The formulation in equation (2.15) also makes it easy to use t-errors as in, for example, Lange *et al.* (1989). By introducing a latent gamma variable for each element of $\mathbf{E}$ (Evans and Swartz (1995), p. 267), the conditional distribution of $\mathbf{\Theta}$ given the scale parameters and latent gamma variables is again multivariate normal.

3.2.2. Bayesian or frequentist?

The methods given below are in the spirit of Weisberg (1983), emphasizing four considerations: a diagnostic should be aimed at detecting a specific problem; diagnostics should

compute quickly; a diagnostic should have a corresponding graphic, so the effect of individual observations can be assessed, and that graphic should allow users to look at the data as directly as possible.

I have used the Bayesian formalism and Bayesian terminology to derive inferential procedures and to phrase inferential summaries. However, explicitly Bayesian diagnostics (as in Ho *et al.* (1996) or Weiss (1995, 1996)) would violate Weisberg's fourth principle, because Bayesian diagnostics are all posterior distributions or moments and thus place an extra layer of mathematics between the analyst and the data. Combining the Bayesian formalism with non-Bayesian diagnostics is not as paradoxical as it may seem. Prominent Bayesians (Smith, 1986; Berger, 1992, 1993) have declared that, although the Bayesian formalism is the only suitable way to draw inferences, exploratory analysis and diagnosis are more informal and non-Bayesian methods are acceptable. Among non-Bayesians, it is becoming more common to view the Bayesian formalism as a means for generating procedures whose frequentist properties can then be assessed, usually by simulation. From this view, the present paper takes a particular approach to fitting hierarchical models and derives diagnostics for that approach, some of which can be used with other approaches. All overtly frequentist methods for hierarchical models use large sample approximations for distribution theory; a frequentist view of a procedure derived using Bayes's theorem also treats that procedure as an approximation. In the simple problems that have been examined, Bayesian procedures — using low information proper priors — tend to have good frequentist properties (Carlin and Louis (1996), section 4.4).

4. Some diagnostics for hierarchical models

This section uses the reformulation from Section 2 to derive several diagnostics. It is not intended to be a full toolkit, but to illustrate how familiar diagnostics can be adapted for hierarchical models.

4.1. Added variable plots

Added variable plots allow a visual check of whether a variable should be added to a linear model. A common added variable plot for linear models with homoscedastic errors is derived as follows (Cook and Weisberg (1982), p. 44, and Atkinson (1985), section 5.2). Suppose that the explanatory variables currently in the model fill the columns of $\mathbf{A}$ and that $\mathbf{B}$ is a candidate variable. In the usual matrix notation, the model is

$$\mathbf{y} = \mathbf{A}\beta + \mathbf{B}\phi + \epsilon. \tag{4.1}$$

Premultiplying both sides of equation (4.1) by $\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ yields

$$\hat{\mathbf{e}} = \{\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\}\mathbf{B}\phi + \{\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\}\epsilon \tag{4.2}$$

where $\hat{\mathbf{e}}$ contains the residuals from the least squares fit of $\mathbf{y}$ to $\mathbf{A}$. If $E(\epsilon) = 0$,

$$E(\hat{\mathbf{e}}) = \{\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\}\mathbf{B}\phi. \tag{4.3}$$

If $\mathbf{B}$ should enter the model, a plot with $\hat{\mathbf{e}}$ on the vertical axis and $\{\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\}\mathbf{B}$ on the horizontal axis will show points clustered around a line through the origin with slope ϕ.

In a hierarchical model, variables can be added at any level. The added variable plot can be applied directly to such variables, using four steps (with explanations of steps 1, 3 and 4 to follow): step 1, reformulate the candidate variable and call it $\mathbf{B}$; step 2, write the hierarchical

model with the candidate variable as in equation (4.4) or (4.1), with $\mathbf{X}$ taking the place of $\mathbf{A}$; step 3, premultiply the resulting equation by $\Gamma^{-1/2}$, where Γ takes a suitable value, to obtain a homoscedastic errors problem; step 4, draw the usual added variable plot.

4.1.1. Step 1

$\mathbf{B}$'s format depends on the level to which the candidate variable belongs. Consider adding plan enrolment to the data case model for the HMO data. If $\mathbf{B}_1$ is the column vector of plan enrolments, $\mathbf{B}$ is $(\mathbf{B}_1', \mathbf{0}_{1\times 45}, 0)'$, conforming to equation (2.6). To add state average expenses per hospital admission to the constraint case model, let $\mathbf{B}_2$ be the column 45-vector of average expenses per admission; then $\mathbf{B}$ is $(\mathbf{0}_{1\times 341}, \mathbf{B}_2', 0)'$.

4.1.2. Step 3

The added variable plot is not sensitive to the choice of Γ. To see this, write the hierarchical model and candidate variable as

$$\mathbf{Y} = \mathbf{X}\Theta + \mathbf{B}\phi + \mathbf{E} \tag{4.4}$$

and premultiply by $\Gamma^{-1/2}$ to yield

$$\mathsf{Y} = \mathsf{X}\Theta + \mathsf{B}\phi + \mathsf{E} \tag{4.5}$$

where the sanserif fount indicates premultiplication by $\Gamma^{-1/2}$. Now premultiply equation (4.5) by $\mathbf{I} - \mathbf{V}$, for $\mathbf{V} = \mathsf{X}(\mathsf{X}'\mathsf{X})^{-1}\mathsf{X}'$, to yield

$$\hat{\mathsf{E}} = (\mathbf{I} - \mathbf{V})\mathsf{Y} = (\mathbf{I} - \mathbf{V})\mathsf{B}\phi + (\mathbf{I} - \mathbf{V})\mathsf{E}. \tag{4.6}$$

Thus, for any Γ, if $\mathbf{B}$ should enter the model as in equation (4.4), the added variable plot will estimate the correct slope ϕ. Changing Γ only affects the plot by changing the scales of Y and B and by changing the shrinkage via $\mathbf{I} - \mathbf{V}$, i.e. by changing the relative position of the data and constraint case points along the plot's regression line. It seems reasonable, then, to use the posterior mean of Γ.

4.1.3. Step 4

The residuals in this plot are scaled: the vertical axis is $\hat{\mathsf{E}} = \Gamma^{-1/2}\hat{\mathbf{E}} = \Gamma^{-1/2}(\mathbf{Y} - \mathbf{X}\hat{\Theta})$. This puts data, constraint and prior case residuals on the same scale.

It may appear odd to use all cases to judge a variable appearing in just one level of the model, but the data and constraint cases convey distinct information. The distinct contributions are easily seen by fitting lines separately to the data and constraint cases in an added variable plot: the two lines will differ. None-the-less, the data and constraint case residuals are linearly related. For any choice of Γ, write $\hat{\mathsf{E}}$ as $(\hat{\mathsf{E}}_d, \hat{\mathsf{E}}_c)'$, where the subscripts indicate data cases and constraint or prior cases respectively, and note that $\mathsf{X}'\hat{\mathsf{E}} = 0$, so

$$\left(\begin{array}{c} \mathbf{X}_1' \\ \hline \mathbf{0} \end{array}\right)\Gamma_1^{-1/2}\hat{\mathsf{E}}_d = -\left(\begin{array}{c|c} \mathbf{Z}_1' & \mathbf{W}_1' \\ \hline \mathbf{Z}_2' & \mathbf{W}_2' \end{array}\right)\left(\begin{array}{c|c} \Gamma_2^{-1/2} & \mathbf{0} \\ \hline \mathbf{0} & \Gamma_3^{-1/2} \end{array}\right)\hat{\mathsf{E}}_c. \tag{4.7}$$

Equation (4.7) also holds if $\hat{\mathsf{E}}_d$ and $\hat{\mathsf{E}}_c$ are replaced with the analogous partitions of $(\mathbf{I} - \mathbf{V})\mathsf{B}$. The data and constraint case residuals vary subject to these constraints, and thus each conveys information that might affect the interpretation of the added variable plot.

Hilden-Minton (1995), section 3.1.1, gave an added variable plot for random and mixed effect models in the form (3.1).

4.2. Transforming variables

Transforming the outcome or explanatory variables can remove curvature or interactions and can make residuals look normally distributed. For the usual linear model in equation (4.1), it is common to consider Box–Cox transformations for the outcome variable:

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0. \end{cases} \tag{4.8}$$

A graphical method to determine likely values of λ is as follows: expand $y^{(\lambda)}$ as a linear Taylor series around $\lambda = 1$ to create a new explanatory variable $\mathbf{B}$ whose coefficient is linear in λ; then draw an added variable plot for $\mathbf{B}$. One possibility for $\mathbf{B}$ has ith co-ordinate $\mathbf{B}_i = \hat{y}_i \log(\hat{y}_i) - \hat{y}_i + 1$, where $\hat{y}_i$ is the fitted value from the least squares fit of the untransformed y. If $\mathbf{B}$ is added to the linear model, it has coefficient $1 - \lambda$. (Cook and Weisberg (1982), section 2.4.3, and Atkinson (1985), section 6.4, describe this $\mathbf{B}$, due to Andrews, and alternatives.) A similar constructed variable and plot are easily derived to examine transformations of explanatory variables; this will not be discussed here.

This method is readily applied to hierarchical models in the form (2.7). To draw the graphic, construct $\mathbf{B}$ as above for the data cases; then treat it like any other candidate variable and draw an added variable plot as in Section 4.1.

4.3. Collinearity and estimability

In ordinary linear models, collinear columns in the design matrix can disturb coefficient estimates, inflate their standard errors and create numerical problems. Stewart (1987) reviewed collinearity measures for linear models; Belsley (1991) is an extended discussion of an influential view. Some issues are unresolved, e.g. when to centre columns of the design matrix before computing collinearity measures (Belsley, 1984). Everyone agrees that a large condition number—the square root of the ratio of the largest and smallest eigenvalues of their favourite matrix—indicates a problem. The eigenstructure of the design matrix is also the key to estimability (identifiability) of linear functions of parameters.

The theories of collinearity and estimability can be applied to hierarchical models formulated as in equation (2.7) to address three distinct concerns: numerical stability, the speed of convergence of MCMC algorithms and estimability. For the Gibbs sampler described in Section 3.2, in each iteration $\mathbf{\Gamma}$ is fixed and $\mathbf{\Theta}$'s conditional posterior is obtained by generalized least squares computations. Thus, the usual collinearity measures can be applied to $\mathbf{\Gamma}^{-1/2}\mathbf{X}$ to determine whether the sampler is numerically stable.

If the parameters are highly correlated *a posteriori*, MCMC algorithms that update one parameter at a time may converge slowly because they make tiny steps in the elongated region of high posterior probability. Reparameterization can improve convergence by changing posterior correlations (Gelfand *et al.*, 1995). At any given draw of an element of $\mathbf{\Theta}$, $\mathbf{\Gamma}$ is held fixed, so high posterior correlations can be seen in the eigenstructure of

$$\mathrm{cov}(\mathbf{\Theta}|\mathbf{Y}, \mathbf{\Gamma}) = (\mathbf{X}'\mathbf{\Gamma}^{-1}\mathbf{X})^{-1}, \tag{4.9}$$

i.e. by the usual collinearity measures. Reparameterizations can be compared by comparing their eigenstructures.

Finally, a linear function of $\mathbf{\Theta}$, $l'\mathbf{\Theta}$, is estimable (identified) if and only if there is a vector ρ such that $\mathbf{X}'\mathbf{X}\rho = l$ (Graybill (1976), section 13.2). The estimability of $l'\mathbf{\Theta}$ is not affected by the $\mathbf{\Gamma}$ used to compute $\mathbf{X}'\mathbf{X}$, as long as it is not extreme, e.g. neither $\mathbf{\Gamma}_2$ nor its inverse is the zero matrix. $\mathbf{\Gamma}$ will, however, determine how near $l'\mathbf{\Theta}$ is to inestimability.

4.4. Case influence

Case influence diagnostics show how inference or prediction summaries change when cases are deleted. For hierarchical models, constraint and prior case influence are of interest. Deleting a prior case is like setting the corresponding parameter's prior variance to infinity. Deleting constraint cases is somewhat more complicated. For the random effects model applied to the HMO data set, deleting Minnesota's constraint case means that Minnesota's posterior mean is no longer shrunk towards μ. If this has a large effect on the estimate of Minnesota's θ_i, we might review the choice of a random effects model, but otherwise this is not an especially useful fact. However, deleting Minnesota's constraint case may also affect the posterior distributions of θ_i from other states, because their shrinkage may change if they no longer borrow strength (or weakness!) from Minnesota. When the state level model is more complicated, deleting Minnesota's constraint case is equivalent to using a mean shift outlier model (Cook and Weisberg (1982), section 2.2.2, and Atkinson (1985), section 5.5) in the constraint cases. The influence of Minnesota's constraint case on Minnesota's θ_i is now somewhat more interesting, as it reflects on the entire state level model and not simply on the decision to use random effects.

One might object that data cases are part of the data whereas the constraint cases are part of the model, so deleting one is not like deleting the other. But any modelling choice inserts information into the analysis, just as a data point inserts information. Thus, although these may be qualitatively distinct kinds of information, the reformulation allows us conveniently to examine the effect of each.

Common influence measures may have problems if they are applied directly to hierarchical models using the reformulation. First, the local influence methods of Cook (1986) and Kass *et al.* (1989) use the mode of the likelihood or posterior respectively which, as noted above, can be misleading. Williams's (1987) partial influence measure (extended by Davison and Tsai (1992)) has the same difficulty. Second, Cook's distance and similar measures (Atkinson (1985), section 2.2, and Cook and Weisberg (1982), section 3.5) describe a case's influence on the entire parameter vector. For hierarchical models, the parameter vector has many elements and the influence of a given case is often felt strongly by one parameter and weakly by the others. Thus, Cook's distance can miss important effects (Bradlow and Zaslavsky, 1997). Finally, hierarchical models have at least two unknown error variances; deleting a case changes the information about $\mathbf{\Gamma}$ and thus affects the posterior mean of $\mathbf{\Theta}$ non-linearly. If Cook's distance is used with $\mathbf{\Gamma}$ fixed, case deletion affects the parameter estimates linearly.

These considerations might suggest rerunning the Gibbs sampler for each deleted case, but the computing would generally be prohibitive. Thus I consider two less computer-intensive methods: the first a single-parameter analogue to Cook's distance (henceforth 'the linear approximation method') and the second an importance sampling method.

To derive the linear approximation method, fix $\mathbf{\Gamma}$ at its posterior mean and apply the updating formula for ordinary linear models (Weisberg (1985), section 5A.1, and Atkinson (1985), section 2.2). The approximate change in $\mathbf{\Theta}$ from deleting the rth case is

$$\hat{\mathbf{\Theta}}_{(-r)} - \hat{\mathbf{\Theta}} \approx -(\mathbf{X}'\mathbf{X})^{-1} x_r' \hat{\mathbf{E}}_r / (1 - v_{rr}) \tag{4.10}$$

where $\hat{\mathbf{\Theta}}$ is the posterior mean of $\mathbf{\Theta}$ using the full data set, $\hat{\mathbf{\Theta}}_{(-r)}$ is the analogous quantity with the rth case deleted, $\hat{\mathbf{E}}_r$ is the rth row of $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{\Theta}}$, x_r is the rth row of $\mathbf{X}$ and v_{rr} is the rth diagonal element of $\mathbf{V}$. The change in the ith element of $\hat{\mathbf{\Theta}}$ from deleting the rth case is the ith element of expression (4.10). Equation (4.10) is approximate because deleting the rth case changes the information about $\mathbf{\Gamma}$, but expression (4.10) keeps $\mathbf{\Gamma}$ fixed.

The second approach, suggested by Gelfand *et al.* (1992) (see also Bradlow and Zaslavsky (1997), MacEachern and Peruggia (1995) and Peruggia (1997)) uses the Gibbs sequence as an importance sample. Label the Gibbs sequence by $k = 1, \ldots, m$. Let $\mathbf{Y}_{(r)}$ denote $\mathbf{Y}$ with the rth case deleted, let $\Psi = (\Theta, \Gamma)$ and let $g(\Psi)$ be a function of Ψ. Then

$$\begin{aligned} E\{g(\Psi)|\mathbf{Y}_{(r)}\} &= \int g(\Psi) f(\Psi|\mathbf{Y}_{(r)})\, d\Psi \\ &= \int g(\Psi) \frac{f(\Psi|\mathbf{Y}_{(r)})}{f(\Psi|\mathbf{Y})} f(\Psi|\mathbf{Y})\, d\Psi \\ &\approx \frac{1}{m} \sum_{k=1}^{m} g(\Psi_k) \frac{f(\Psi_k|\mathbf{Y}_{(r)})}{f(\Psi_k|\mathbf{Y})} \end{aligned} \tag{4.11}$$

where Ψ_k, $k = 1, \ldots, m$, is the Gibbs sequence. The proportionality constants of $f(\Psi|\mathbf{Y})$ and $f(\Psi|\mathbf{Y}_{(r)})$ are not necessary if the weights $f(\Psi_k|\mathbf{Y}_{(r)})/m\, f(\Psi_k|\mathbf{Y})$ are rescaled to sum to 1. For data, constraint and prior cases, $f(\Psi_k|\mathbf{Y}_{(r)})/f(\Psi_k|\mathbf{Y})$ is simple, corresponding to a single row of equation (2.15), and will not be given.

For either the linear approximation method or the importance sampling method, the change in $g(\Psi)$ induced by a case deletion can be calibrated using the relative change:

$$\mathrm{RC}\{g(\Psi); r\} = [E\{g(\Psi)|\mathbf{Y}_{(r)}\} - E\{g(\Psi)|\mathbf{Y}\}]/\mathrm{psd}\{g(\Psi)|\mathbf{Y}\}, \tag{4.12}$$

where psd means 'posterior standard deviation' and, for the linear approximation, $g(\Psi)$ is an element of Θ. By the usual folk wisdom, $|\mathrm{RC}| > 2$ suggests an influential case.

To assess these case influence methods, consider an artificial example based on the HMO data. This example is intended to bend both methods until they break; I discuss its relevance to practice below. Suppose that each state has eight plans and that the state averages are the same as in the actual data set except for Maine, which has the highest average in the actual data set. Set the prior variance for μ to ∞ and put low information priors on σ^2 and τ^2. (I used gamma priors with mean 11 and variance 110 for $1/\sigma^2$ and $1/\tau^2$, but any continuous proper prior will give the same qualitative result.) Label Maine as $i = 1$ and assume that one of its HMOs, labelled $j = 1$, has premium y_{11} equal to Maine's average. Now increase y_{11} to show the effect of an outlying plan. Fig. 2 shows the posterior means of the θ_i as y_{11} increases; the abscissa is the number of within-state standard deviations ($\eta = 22.4$) y_{11} is increased above the average of Maine's other plans. The uppermost line is Maine's average, the dotted line is Maine's posterior mean and the cluster of lines describes the posterior means of the other 44 states. As y_{11} increases, Maine's posterior mean increases at a damped rate while the other posterior means shrink slowly towards μ. When y_{11} reaches 33η from its starting point, the state posterior means approach μ rapidly. If y_{11} were originally 40η greater than the average of Maine's other plans and were removed, Maine's posterior mean would *increase* to the value at the extreme left-hand side of the plot.

Now consider the importance sampling method, and suppose that y_{11} is about 40η from the average of Maine's other HMOs. The θ_i have small posterior variances because each state borrows ample strength from the others. The Gibbs draws for θ_1 will thus be within a few posterior standard deviations of $E(\theta_1|\mathbf{Y})$, and so will the importance sampling computation of $E(\theta_1|\mathbf{Y}_{(r)})$, which reweights the Gibbs draws. If we delete the constraint case for θ_1, the true value of $E(\theta_1|\mathbf{Y}_{(r)})$, given by the diagonal line in Fig. 2, is far from $E(\theta_1|\mathbf{Y})$, but the importance sampling method cannot detect this because the Gibbs draws are all near $E(\theta_1|\mathbf{Y})$.

The linear approximation's performance is illustrated in Fig. 3. The diagonal line is, as in Fig. 2, the average of Maine's HMOs including y_{11}. It is also the posterior mean of θ_1 if the

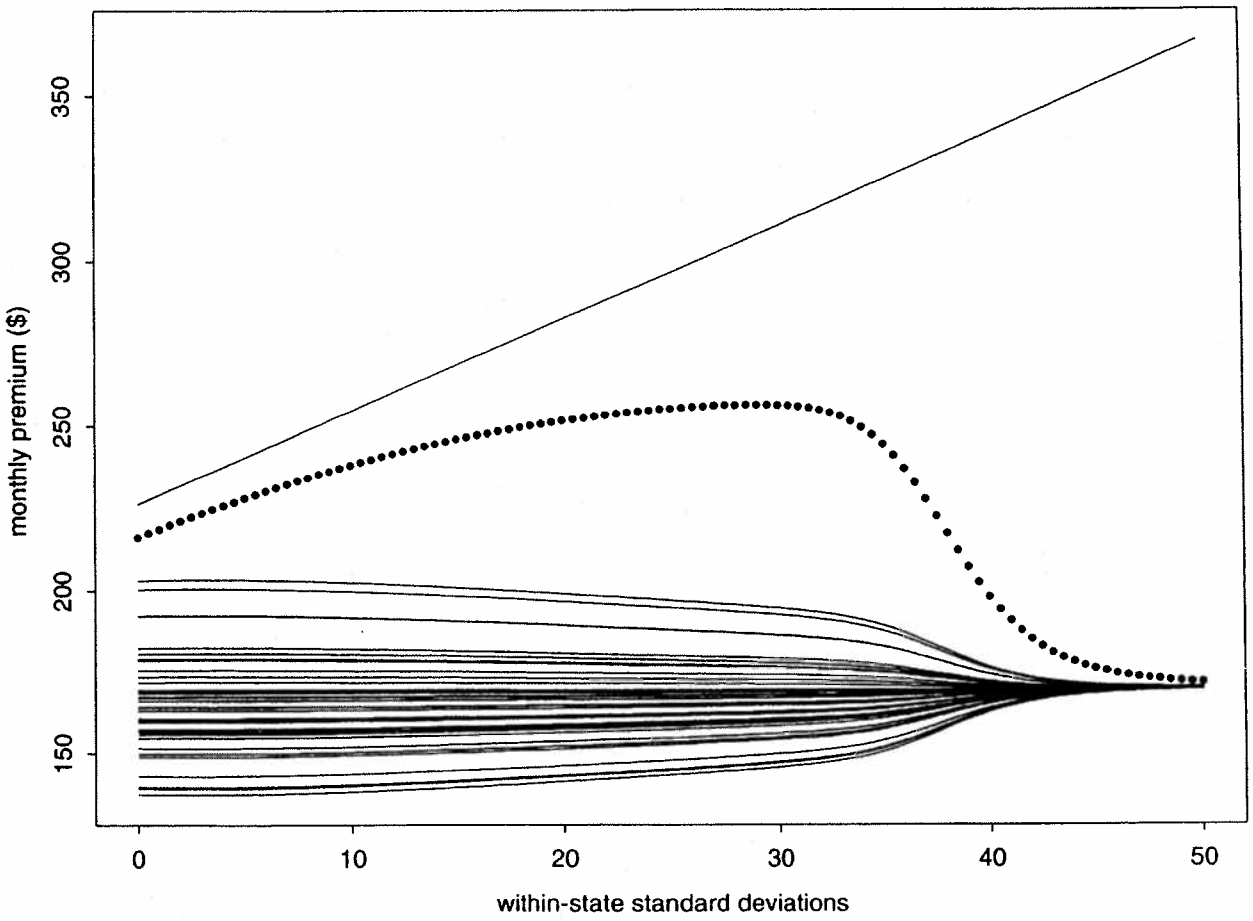

Fig. 2. Effect of increasing y_{11} in the artificial data set (Section 4.4) (the horizontal axis is the number of within-state standard deviations by which y_{11} is greater than the average of the other y_{1j}): diagonal line, average of y_{1j}; dotted curve, posterior mean of θ_1; curves in the lower half of the figure, posterior means of θ_i for the other states

constraint case for θ_1 is deleted. The adjacent broken curve is the linear approximation to the posterior mean of θ_1 under the same deletion. The linear approximation is quite good until y_{11} reaches about 25η, after which it overstates the effect of deleting Maine's constraint case.

The horizontal line in Fig. 3 is the posterior mean of θ_1 when y_{11} is deleted. The bold dotted curve is $E(\theta_1|\mathbf{Y})$, as in Fig. 2. Twined around these is a dotted curve indicating the linear approximation to the posterior mean of θ_1 when y_{11} is deleted. This approximation is also quite good until y_{11} reaches about 25η.

For deleting θ_1's constraint case, the linear approximation is, from expression (4.10),

$$E(\theta_1|\mathbf{Y}_{(r)}) \approx (1+\nu)\,E(\theta_1|\mathbf{Y}) - \nu\,E(\mu|\mathbf{Y}), \tag{4.13}$$

where $\nu = \sigma^2/n\tau^2 \geqslant 0$. To use this, replace σ^2 and τ^2 by their full data posterior means. If an outlier in Maine causes $E(\tau^2|\mathbf{Y})$ to be small and $E(\sigma^2|\mathbf{Y})$ to be large, then deleting θ_1's constraint case increases the former and decreases the latter, so expression (4.10) overshoots $E(\theta_1|\mathbf{Y}_{(r)})$. The degree of inaccuracy depends on the amount that the deleted case changes the posterior distributions of σ^2 and τ^2, which can be substantial.

For deleting y_{11} itself, the linear approximation for the posterior mean of θ_1 is

$$E(\theta_1|\mathbf{Y}_{(r)}) \approx (1+\nu)\,E(\theta_1|\mathbf{Y}) - \nu y_{11} \tag{4.14}$$

for

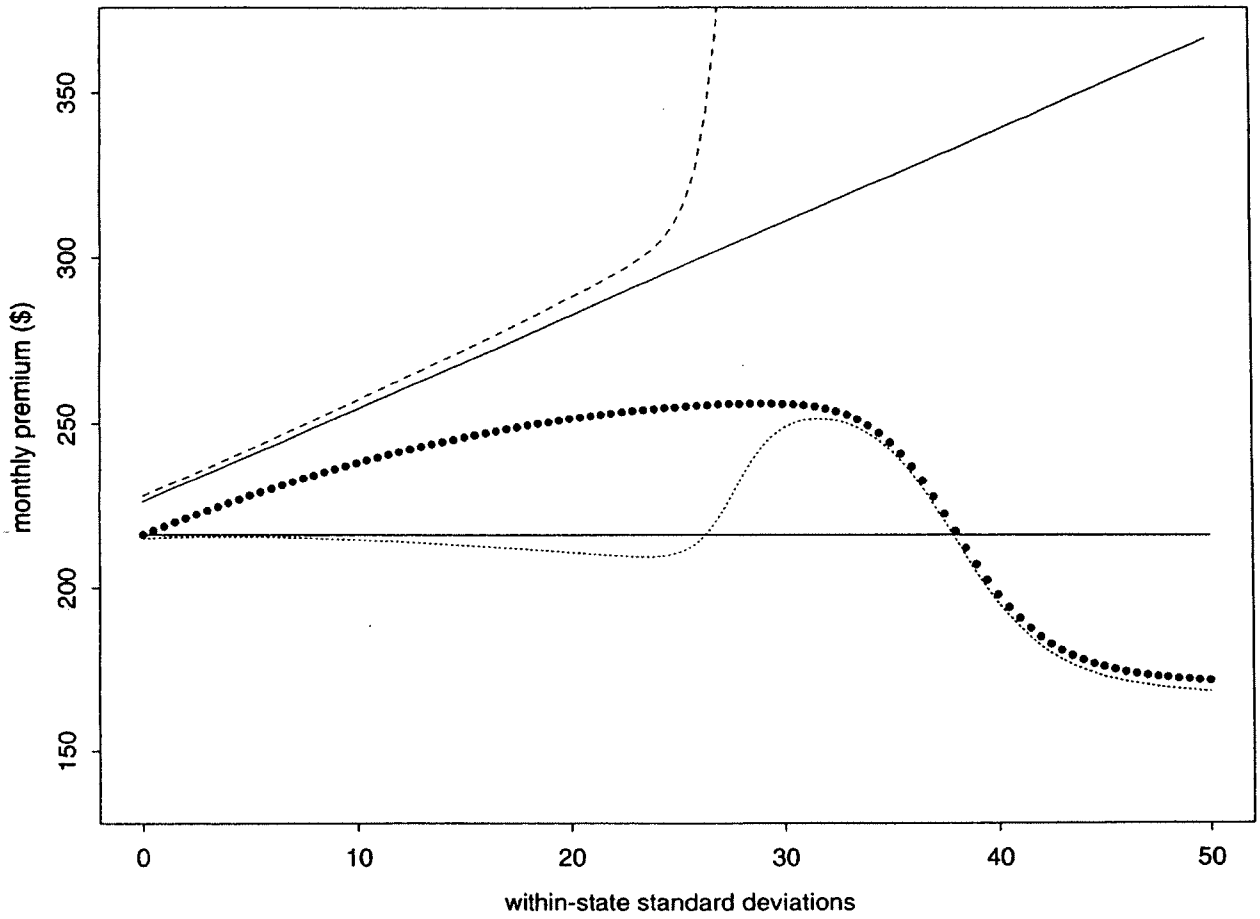

Fig. 3. Performance of the linear approximation influence measure (Section 4.4) (the horizontal axis is the number of within-state standard deviations by which y_{11} is greater than the average of the other y_{1j}): diagonal line, average of y_{1j} and also the true posterior mean of θ_1 if the constraint case for θ_1 is deleted; bold dotted curve, posterior mean of θ_1; horizontal line, posterior mean of θ_1 if y_{11} is deleted; small dotted curve, linear approximation to the posterior mean of θ_1 if y_{11} is deleted

$$\nu = \frac{Nn + h}{Nn(n - 1 + h) - h},$$

where $h = \sigma^2/\tau^2$. Because $\nu \geqslant 0$, the linear approximation moves $E(\theta_1|\mathbf{Y}_{(r)})$ away from $E(\theta_1|\mathbf{Y})$ in the direction opposite y_{11}. This works well until y_{11} becomes very extreme.

The example used an extreme outlier to exercise the two influence methods. However, constraint case deletions readily produce the effect shown above provided that the posterior variance of the affected θ_i is small. Section 5.2 shows this using the actual HMO data and a model suggested by diagnostics. Thus each method has a weakness, but the linear approximation method seems less flawed because it overstates influence when it fails whereas the importance sampling method understates influence when it fails.

For data cases, the effects shown above only occur when an outlier is extreme relative to the data case error variance. It is difficult to imagine an outlier that is sufficiently extreme to make the linear approximation fail but difficult to detect on a casual examination of the data. Thus, the linear approximation seems adequate for data cases.

MacEachern and Peruggia (1995) gave a modified importance sampler intended to avoid the difficulty noted here. However, their method uses posterior modes and thus would be problematic for hierarchical models.

4.5. Residuals

Consider the familiar plot with internally Studentized residuals on the vertical axis and fitted values on the horizontal axis (Cook and Weisberg (1982), section 2.3.1). For simplicity, suppose that Γ is fixed so the posterior mean for Θ, $\hat{\Theta}$, is the generalized least squares estimate. If we proceed as for the four previous diagnostics, plotting data and constraint cases on the same plot, the points for the constraint cases are uninformative because the residuals for the constraint cases are $0 - (\mathbf{Z}_1\hat{\Theta}_1 + \mathbf{Z}_2\hat{\Theta}_2)$, whereas the fitted values are $\mathbf{Z}_1\hat{\Theta}_1 + \mathbf{Z}_2\hat{\Theta}_2$. These *are* the correct residuals; rather, the appropriate fitted values are $\mathbf{Z}_2\hat{\Theta}_2$. Thus, unlike the first four diagnostics, the data and constraint cases need separate residual plots.

Now consider the usual residual plot for the data cases only. Another problem surfaces, this one intrinsic to hierarchical models. In ordinary linear models, using the notation of Section 4.1, the vector of residuals $\hat{\epsilon}$ is orthogonal to the vector of fitted values $\mathbf{A}\hat{\beta}$ by construction. In hierarchical models, such orthogonality is generally not present. From equation (4.7), recalling that the sanserif fount indicates premultiplication by $\Gamma^{-1/2}$,

$$\begin{aligned}\hat{\mathsf{y}}'\hat{\mathsf{E}}_d &= \hat{\Theta}_1'\mathsf{X}_1'\hat{\mathsf{E}}_d \\ &= -\hat{\Theta}_1'(\mathsf{Z}_1'|\mathsf{W}_1')\hat{\mathsf{E}}_c \\ &= \hat{\Theta}_1'(\mathsf{Z}_1'\mathsf{Z}_1 + \mathsf{W}_1'\mathsf{W}_1)\hat{\Theta}_1 + \hat{\Theta}_1'(\mathsf{Z}_1'\mathsf{Z}_2 + \mathsf{W}_1'\mathsf{W}_2)\hat{\Theta}_2. \end{aligned} \tag{4.15}$$

For the pairwise differences model, Z_2 and W_2 are null, so $\hat{\mathsf{y}}'\hat{\mathsf{E}}_d \geqslant 0$, i.e. the data case residuals and fitted values are positively correlated. For the one-way random effects model, $\tau^2\hat{\mathsf{y}}'\hat{\mathsf{E}}_d = \hat{\Theta}_1'\hat{\Theta}_1 - N\hat{\mu}\bar{\Theta}_1$ where $\bar{\Theta}_1$ is the simple average of the $\hat{\theta}_i$. Usually $\hat{\mu} \approx \bar{\Theta}_1$, in which case $\tau^2\hat{\mathsf{y}}'\hat{\mathsf{E}}_d \approx (\hat{\Theta}_1 - \bar{\Theta}_1)'(\hat{\Theta}_1 - \bar{\Theta}_1)$ which is non-negative. For both models, $\hat{\mathsf{y}}'\hat{\mathsf{E}}_d \geqslant 0$ for each Γ, so it is non-negative marginally as well.

The data case plot of Studentized residuals *versus* fitted values has a complex structure. In the one-way random effects model, let $n_i = n$ and suppose that the constraint case errors δ_i are all 0, i.e. that each state has mean $\theta_i = \mu$. For simplicity, let $M = \mu = 0$, set s^2 to ∞ and set Γ to a non-extreme value. If $\bar{y}_{i.}$ is the average of y_{ij} over j, $\bar{y}_{..}$ is the average of y_{ij} over i and j, and $\bar{\epsilon}_{i.}$ and $\bar{\epsilon}_{..}$ are defined analogously,

$$\hat{y}_{ij} = \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{y}_{i.} + \frac{\sigma^2}{n\tau^2 + \sigma^2}\bar{y}_{..} = \bar{\epsilon}_{..} + \alpha(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}) \tag{4.16}$$

for $\alpha = n\tau^2/(n\tau^2 + \sigma^2)$ and

$$\hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij} = \epsilon_{ij} - \bar{\epsilon}_{i.} + (1 - \alpha)(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}). \tag{4.17}$$

The term $\epsilon_{ij} - \bar{\epsilon}_{i.}$ is the usual residual variation; the term $(1 - \alpha)(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})$ is peculiar to hierarchical models. Over repeated realizations of ϵ_{ij}, $\hat{y}_{ij}$ and $\bar{\epsilon}_{i.} - \bar{\epsilon}_{..}$ will have mean 0 because $E(\epsilon_{ij}) = 0$. However, for any given realization of ϵ_{ij}, the $\bar{\epsilon}_{i.}$ will be spread around $\bar{\epsilon}_{..}$ and because shrinkage is proportional the $\hat{y}_{ij}$ that are largest in absolute value will be shrunk the most and will tend to have the largest $\hat{\epsilon}_{ij}$. Thus, the plot of Studentized residuals *versus* fitted values will show a positive slope, a kind of regression effect.

Now suppose that at least one δ_i is not 0, and fix Γ. The mean of the data case residuals, across repeated realizations of ϵ_{ij} holding δ_i fixed, is

$$E(\hat{\mathbf{E}}_d|\delta, \xi) = (\mathbf{I} - \mathbf{V})\boldsymbol{\Gamma}^{-1/2}\begin{pmatrix} \mathbf{0} \\ \hline \delta \\ \hline \xi \end{pmatrix} \tag{4.18}$$

which is non-zero in general, i.e. the residuals are biased. (Hilden-Minton (1995), section 4.1, gives an alternative derivation.)

Thus, the data case residuals display three effects: the usual residual variation, bias and a regression effect. The last two can produce strange patterns in a plot of Studentized residuals *versus* fitted values. For the pairwise differences model (2.10), the residuals within each treatment group fall along a sloping line, but the treatment groups have distinct, roughly parallel lines. More complex models will yield more complex structure.

These residual patterns are not an artefact of the reformulation but are intrinsic to hierarchical models. Empirical Bayes residuals (e.g. Hilden-Minton (1995)) will display the same patterns because they are computed as above with a particular $\boldsymbol{\Gamma}$. The joint posterior distribution of the ϵ_{ij} (Chaloner, 1994) will also show the regression and bias effects: $E(\epsilon_{ij}|\mathbf{Y})$ is simply $E(\epsilon_{ij}|\mathbf{Y}, \boldsymbol{\Gamma})$ integrated against the posterior distribution for $\boldsymbol{\Gamma}$, but $E(\epsilon_{ij}|\mathbf{Y}, \boldsymbol{\Gamma})$ is identical with $\hat{\epsilon}_{ij}$ computed with the same $\boldsymbol{\Gamma}$.

Hilden-Minton (1995), sections 4.1 and 4.2, offered two ways around this problem. The first is to fit a separate model for each state and to use the residuals from these fits. This avoids bias and regression effects, producing residuals with well-known properties. Unfortunately, it cannot be applied to all the models covered in the present paper, in particular to models that are not truly hierarchical, such as the pairwise differences model. Also, the separate models in each state may have unstable fits. None-the-less, this approach can be useful for truly hierarchical models because it allows the lowest level (e.g. plan level) model to be evaluated free from the effects of deficiencies in the higher level model.

Hilden-Minton's second solution was to construct linear transforms of the data case residuals, $l'\hat{\mathbf{E}}_d$, that are independent of each other and have no bias. This construction is easily adapted to the reformulation given here and, in the cases that I have examined, the bias-free residuals are also free of the regression effect. Unfortunately, a transformation to $l'\hat{\mathbf{E}}_d$ can obliterate indications such as multiple outliers which are manifest in a plot of the untransformed residuals. Also, $l'\hat{\mathbf{y}}$ is necessarily 0 for such l. Finally, effects in the $l'\hat{\mathbf{E}}_d$ are not easy to attribute to individual observations, which is usually an important goal in residual analyses.

How, then, do we proceed? We routinely ignore structure in residuals from ordinary linear models, e.g. that estimated residuals are correlated and supernormal (Cook and Weisberg (1982), section 2.3.4). If we are to use the data case plot of Studentized residuals *versus* fitted values, we must learn to ignore more structure than we ignore in residual plots for ordinary linear models, i.e. we must be content with only finding grosser outliers and heteroscedasticity than we expect to find in ordinary linear regressions.

Variances for residuals are easily derived. Chaloner (1994) gave the posterior variance of data case errors for the one-way random effects model; her method is easily adapted to the reformulation. Those desiring a frequentist covariance matrix (over repeated realizations of ϵ and δ) can use the fact that $\text{cov}(\mathbf{E}) = \mathbf{I}$, so

$$\text{cov}(\hat{\mathbf{E}}_d) = E\{\text{cov}(\hat{\mathbf{E}}_d|\boldsymbol{\Gamma})\} + \text{cov}\{E(\hat{\mathbf{E}}_d|\boldsymbol{\Gamma})\} = [E(\mathbf{I} - \mathbf{V})]_d \tag{4.19}$$

where the latter expectation is with respect to the posterior distribution of $\boldsymbol{\Gamma}$, $\mathbf{V}$ was given

in Section 4.1, and $[\cdot]_d$ selects the data case rows and columns. These moments can be computed from the MCMC draws or approximated by computing $\mathbf{V}$ with $\boldsymbol{\Gamma}$ set to its posterior expectation. It is also straightforward to derive $\text{cov}(\hat{\mathbf{E}}_d)$ over repeated realizations of ϵ with δ held fixed.

The foregoing considerations apply to constraint case residuals as well. They, also, are subject to bias and regression effects induced by the next highest level of the hierarchy and, as in equation (4.19), their covariance is $\text{cov}(\hat{\mathbf{E}}_c) = [E(\mathbf{I} - \mathbf{V})]_c$ where $[\cdot]_c$ selects the constraint case rows and columns.

5. Example: the health maintenance organizations data set

Section 5.1 applies the diagnostics to the HMO data and the random effects model, whereas Section 5.2 gives a variant analysis.

5.1. The random effects model fitted to the raw health maintenance organizations data

For the random effects fit, I used $M = 0$ and $s^2 = 10^6$, a flat prior for $1/\sigma^2$ and a low information gamma prior for $1/\tau^2$, with mean 11 and variance 110. I ran a Gibbs sampler for 1000 iterations for various starting values. Convergence was immediate for all starting values and Gelman and Rubin (1992) diagnostics were all close to 1, indicating that the chains with different starting values mixed adequately. The calculations below used all 1000 iterations based on the following starting values: for θ_i, the average of state i's premia; for μ, the average of the starting values for the θ_i (166.9); for $1/\sigma^2$, the reciprocal of the within-state mean square (0.00204); for $1/\tau^2$, the reciprocal of (sample variance of y − within-state mean square) (0.00707). A run of 5000 iterations yielded results that were identical to two significant digits.

Fig. 1 shows the posterior means for the 45 states as triangles and the state averages as dots. $E(\mu|\mathbf{Y}) = 167$, with posterior standard deviation 2.5. Between-state variance remains: the posterior mean of τ^2, the between-state variance, was 179, the posterior standard deviation was 70 and the 95% equal-tailed highest posterior density (HPD) region was (61, 339), reflecting some skewness. The analogous quantities for σ^2, the error variance, were 502, 43 and (424, 592) respectively, indicating a more symmetric distribution.

The scaled design matrix $\mathbf{X}$ has condition number 6.8, giving no suggestion of collinearity problems. (All scalings used the posterior means of σ^2 and τ^2.) Moreover, $\mathbf{X}'\mathbf{X}$ is of full rank for finite non-zero σ^2 and τ^2, so μ and the θ_i are identified.

Fig. 4 is the data case residual plot; I Studentized the residuals by setting $\boldsymbol{\Gamma}$ to its posterior mean. As expected, the residuals lie along the sloping line. Ignoring this slope, the plot shows an outlier with $\hat{y}$ about 180, an HMO in Washington state. There is no strong suggestion of a funnel opening to the right. However, Fig. 4 shows somewhat more points in the range (2, 4) than in the range (−4, −2), and a normal quantile plot of the Studentized data case residuals (not shown) also suggests a slightly long upper tail. Fig. 5 shows Studentized residuals from a model that simply fits a mean for each state, as suggested by Hilden-Minton (1995); it adds nothing to Fig. 4, although some may prefer to say that Fig. 4 adds nothing to Fig. 5. Fig. 6 shows a normal quantile plot of the Studentized constraint case residuals, which also suggests a long upper tail.

Fig. 7 gives a check for transforming the premiums y_{ij}. The line is the ordinary least squares fit to this plot, with intercept −0.002 (standard error 0.04) and slope 0.33 (standard error 0.04), suggesting a transformation with $\lambda = 0.67$. This finding is consistent with the residual

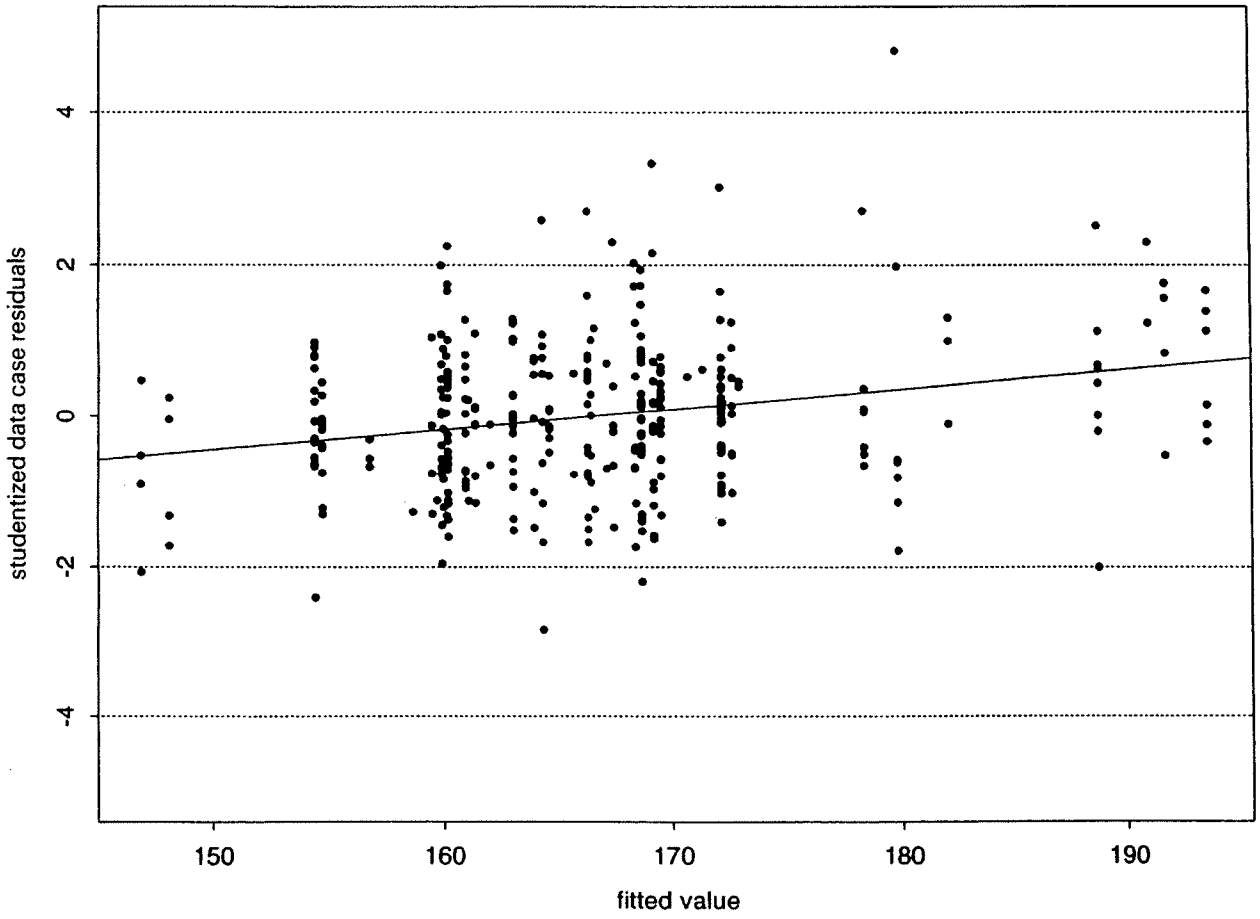

Fig. 4. Studentized residuals *versus* fitted values, for the random effects model fit to the HMO data (Section 5.1): ———, ordinary least squares regression of residuals on fitted values

plot. The outlying plan in Washington state has abscissa near 0 and thus has little effect on the estimated λ.

Candidates for added variables at the plan level were log(enrolment) and at the state level were expenses per admission (an index of the state's health care costs, obtained from the 1990 US census) and indicator variables for each of six regions of the USA. Added variable plots indicated that expenses per admission and the New England indicator should enter; Fig. 8 shows the former plot. In Fig. 8 the ordinary least squares slope is 0.007 (standard error 0.002), whereas the New England indicator had ordinary least squares slope 35 (standard error 8). Fig. 8 suggests that the effect of expenses per admission may be caused by a few points, particularly Guam's (GU) constraint case at the lower left-hand side. Although expenses per admission are higher in New England than elsewhere, the two variables appear to convey distinct information: the ordinary least squares fit to Fig. 8 without the New England plans or constraint cases is nearly identical with the full data fit. Section 5.2 considers a fit that includes both these state level variables.

In the influence analysis, the importance sampling and linear approximation measures were in reasonable agreement, with two groups of exceptions. First, the linear approximation found near-zero influences of constraint cases on the θ_i for other states, and of individual plans on the θ_i for other states. However, for these cases the importance sampling method also found $|RC| < 1$, which is well within the range of the Gibbs draws for each θ_i, so these measures can be assumed to be accurate. The two methods also disagreed for the most

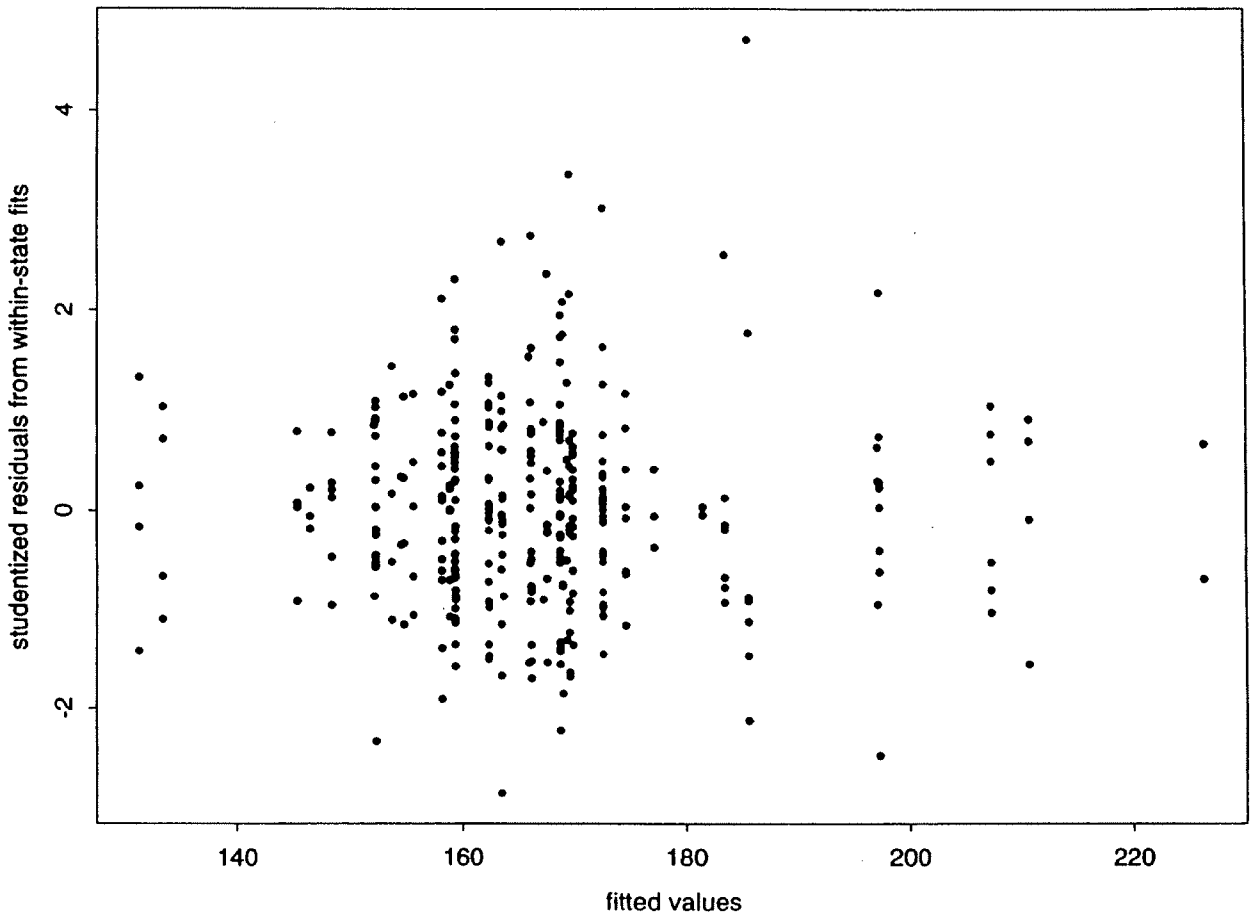

Fig. 5. Studentized residuals *versus* fitted values, for within-state models fit to the HMO data (Section 5.1)

extreme influences, each of which was the influence of deleting a constraint case on the posterior mean of its own θ_i. For these cases, the true RCs are easily computed and the linear approximation was effectively exact. When $|RC| < 1$, the importance sampling method was very good, but for more extreme true RCs the importance sampling RCs were somewhat damped. For these constraint cases, the importance weights were dominated by a single Gibbs iterate; this was also true when 5000 iterations were used instead of 1000.

Four cases had true $|RC| > 2$, each involving the effect of deleting a constraint case on its own θ_i: Maine ($RC = 3.4$), Connecticut ($RC = 2.1$), Puerto Rico ($RC = -2.1$) and North Dakota ($RC = -2.4$). Deleting Maine's constraint case, for example, increases Maine's posterior mean from \$191 to \$231. Each of these states has few plans and an extreme average premium and thus was shrunk considerably: hence the large constraint case influence. These influences prompt a review of the decision to use a hierarchical model; I would stick by that decision.

Cook's distance, computed with Γ fixed at its full data posterior mean, had maximum values of 0.38 and 0.15 for Maine's and North Dakota's constraint cases respectively, so it identified the two cases with the largest $|RC|$ but grossly understated their influence.

5.2. Variant analysis: including two variables at the state level

The model to be fitted here is

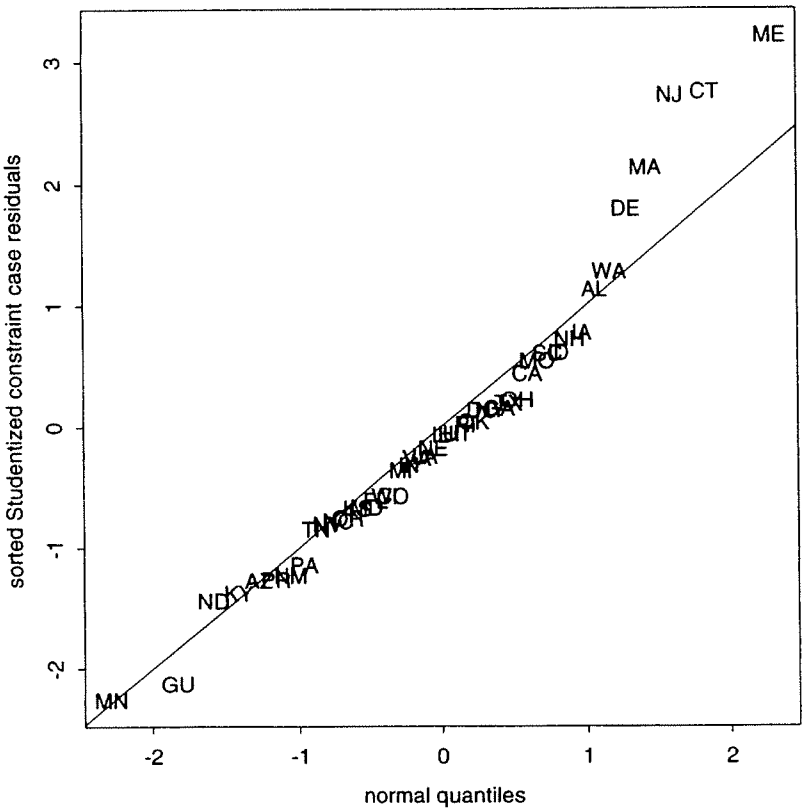

Fig. 6. Normal quantile plot of Studentized constraint case residuals, for the random effects model fit to the HMO data (Section 5.1)

$$\begin{aligned} y_{ij} &= \theta_i + \epsilon_{ij}, \\ \theta_i &= \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \delta_i, \end{aligned} \tag{5.1}$$

where z_{1i} is expenses per admission for the ith state, z_{2i} is an indicator for New England, the γs are parameters to be estimated, $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\delta_i \sim N(0, \tau^2)$.

To avoid collinearity, I centred and scaled expenses per admission. Taking account of this, the added variable plots in Section 5.1 suggested Gibbs starting values of 6.7 and 35 for γ_1 and γ_2 respectively. Otherwise, I used the same starting values as in Section 5.1. The Gibbs sampler was unstable through about 250 iterations; from iteration 251 to 1000 the sampler provided satisfactory convergence diagnostics, and these iterations were used for the computations that follow. Table 1 gives the posterior means and standard deviations and 95% HPD regions for the three state level parameters and the two variances.

The posterior mean of τ^2 is 0.33 compared with 179 for the random effects model in Section 5.1. Although striking, this reduction is not outlandish: the two new explanatory variables need only to account for enough between-state variation to 'argue' that the remaining between-state variation is attributable to ϵ. The reduction in τ^2 does explain the Gibbs sampler's initial instability: the sampler used a wildly inaccurate starting value for τ^2. The starting value's effect can be seen by tracing $\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}$'s condition number through the Gibbs iterates. Before roughly iteration 250, the condition number is between 10 and 20, but around iteration 250 it climbs sharply and is soon in the mid-hundreds. Evaluated at the posterior mean of $\boldsymbol{\Gamma}$ (computed using the last 750 iterations), $\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X}$ has condition number 315, which is high. A further examination shows that $\boldsymbol{\Theta}$'s posterior density lies very near to a subspace of dimension 3, which is a consequence of the evidence favouring small τ^2.

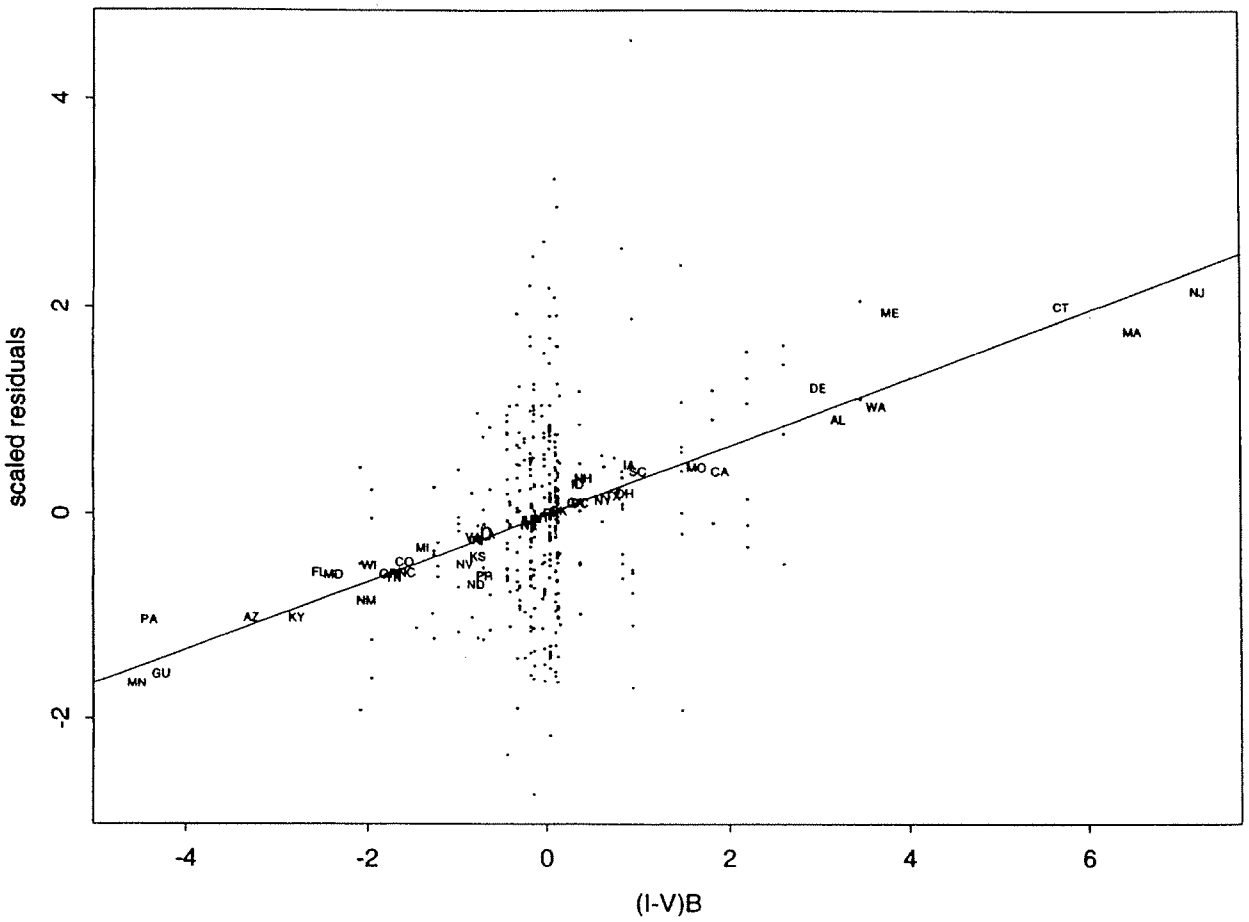

Fig. 7. Added variable plot checking for a transformation, for the random effects model fit to the HMO data (Section 5.1): data cases are represented by dots and constraint cases by their state's two-letter abbreviation, and the line is the ordinary least squares fit

The Gibbs fit upholds the added variable plots: for both new variables, the 95% HPD region excludes 0 and the posterior means are reasonably close to the slopes in the added variable plots. The residual plot shows little regression effect and the same outlying plan in Washington state. The residual plot and the normal quantile plot of the data case residuals again give some suggestion of a skewed residual distribution.

For case influence, the linear approximation had RCs ranging from -70 to 98, with the large |RC|s all being influences on a θ_i from deleting its own constraint case. These approximate RCs were large for two reasons. First, the corresponding exact RCs really are large: because τ^2 is small, the θ_i have small posterior variances. Second, the linear approximation RCs are all 3.43 times the exact values. This is an example of where the linear approximation breaks down, as described in Section 4.4. The actual range of the RCs is from -20 to 29; by comparison, the importance sampling method gave RCs ranging between -4.5 and 1.7. The most extreme values of the importance sampling RC all arose from deleting the Pennsylvania constraint case, which gave one Gibbs iterate an importance weight of 0.93.

For the state level regression parameters, all data cases had $|RC| < 0.57$ from the linear approximation. Among the constraint cases, Guam had the largest linear approximation |RC|, with an RC for expenses per admission of -2.8: without the Guam constraint case, the posterior mean of the coefficient of expenses per admission would be approximately -0.3

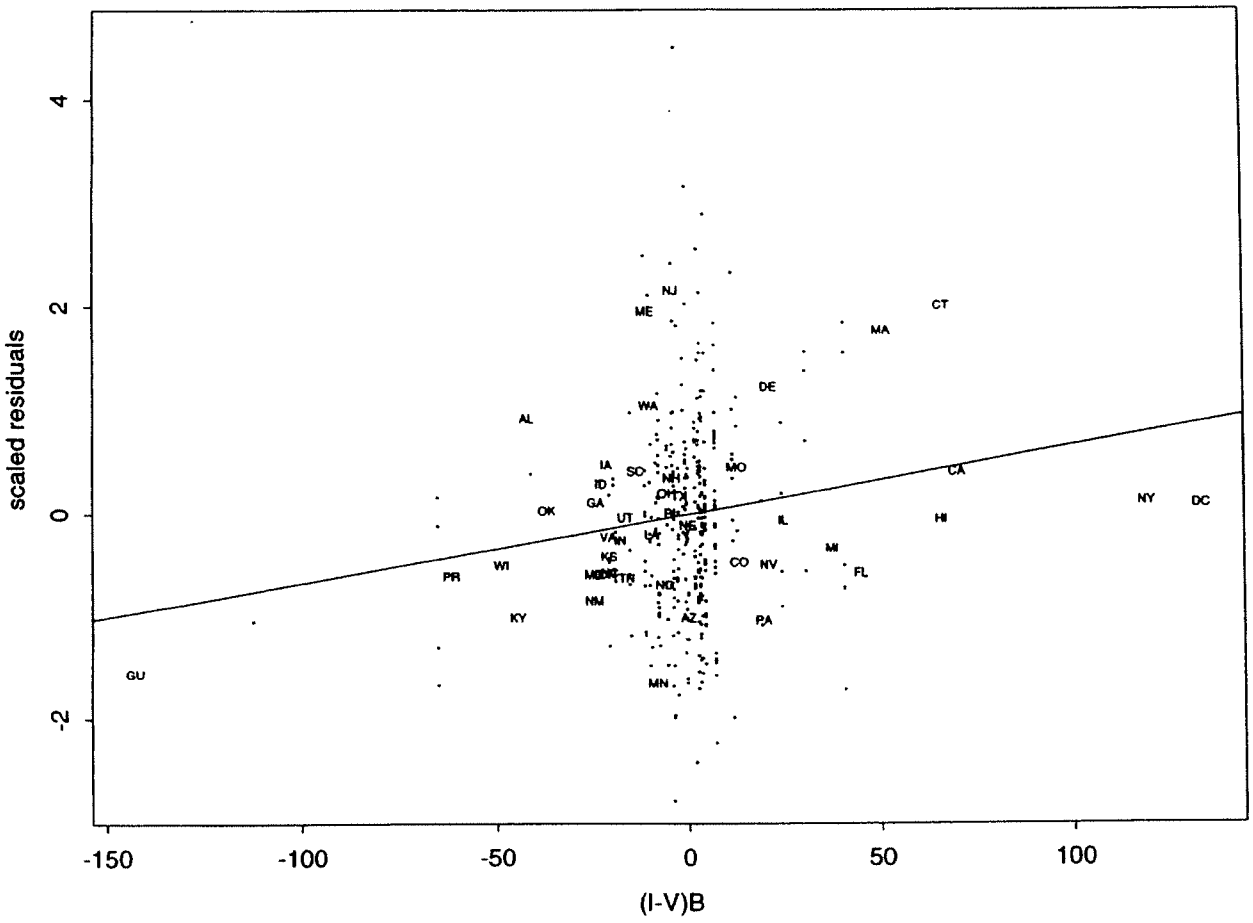

Fig. 8. Added variable plot for expenses per admission, for the random effects model fit to the HMO data (Section 5.1): data cases are represented by dots and constraint cases by their state's two-letter abbreviation, and the line is the ordinary least squares fit

Table 1. Posterior means and standard deviations for the model with expenses per admission and New England

Parameter	*Posterior mean*	*Posterior standard deviation*	*95% HPD region*
Intercept	163.5	1.4	(160.6, 166.2)
Expenses per admission	3.9	1.5	(0.92, 6.8)
New England	32.8	6.0	(21.0, 44.4)
σ^2	565	44	(484, 654)
τ^2	0.33	0.42	(0.022, 1.5)

instead of 3.9. If Guam's constraint case and four plans are deleted from Fig. 8, the slope becomes barely non-significant by an ordinary least squares analysis. No other |RC| is sufficiently large to change the sign of a parameter estimate.

6. Some further uses of the reformulation

This paper has shown how the reformulation of Section 2 can be used to derive diagnostics for a broad class of models. The reformulation has other uses. Hodges and Sargent (1996)

used it to provide a rigorous foundation for counting degrees of freedom in the same class of models. This theory is related to Hilden-Minton's (1995) transformed residuals, which were discussed in Section 4.5. The reformulation suggests a way to speed computation for the same class of models; this is under development. Constraint cases have been used to unify a broad class of survival models based on proportional hazards regression (Sargent, 1996a) and applied to frailty modelling (Sargent, 1995), time-varying hazards (Sargent, 1997) and to smoothing analyses of variance and subgroup analyses (Sargent, 1996b; Sargent and Hodges, 1997). Further applications are forthcoming.

Acknowledgements

This paper was greatly improved by the comments of Dan Sargent, Tom Louis, the RAND Statistics Group and three referees. The work was supported by RAND-sponsored research, the Division of Biostatistics, School of Public Health, University of Minnesota, the Community Programs for Clinical Research on AIDS Statistical Center (National Institutes of Health–National Institute of Allergy and Infectious Diseases grant NO 1-41-55228) and the Minnesota Oral Health Clinical Research Center (National Institutes of Health–National Institute of Dental Research grant P30-DE09737).

References

Atkinson, A. C. (1985) *Plots, Transformations, and Regression.* Oxford: Clarendon.

Beckman, R. J., Nachtsheim, C. J. and Cook, R. D. (1987) Diagnostics for mixed-model analysis of variance. *Technometrics*, **29**, 412–426.

Belsley, D. A. (1984) Demeaning conditioning diagnostics through centering (with discussion). *Am. Statistn*, **38**, 73–93.

———(1991) *Conditioning Diagnostics.* New York: Wiley.

Berger, J. O. (1992) Discussion of a paper by JS Hodges. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 256–258. Oxford: Clarendon.

———(1993) Contributed discussion. In *Case Studies in Bayesian Statistics* (eds C. Gatsonis, J. S. Hodges, R. E. Kass and N. D. Singpurwalla), pp. 302–303. New York: Springer.

Besag, J. and Kempton, R. (1986) Statistical analysis of field experiments using neighboring plots. *Biometrics*, **42**, 231–251.

Bradlow, E. T. and Zaslavsky, A. M. (1997) Case influence analysis in Bayesian inference. *J. Comput. Graph. Statist.*, **6**, 1–18.

Bryk, A. S. and Raudenbush, S. W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park: Sage.

Carlin, B. P. and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman and Hall.

Carlin, B. P. and Polson, N. G. (1992) Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 577–586. Oxford: Clarendon.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *Am. Statistn*, **46**, 167–174.

Chaloner, K. (1994) Residual analysis and outliers in Bayesian hierarchical models. In *Aspects of Uncertainty* (eds A. F. M. Smith and P. R. Freeman), pp. 153–161. Chichester: Wiley.

Christensen, R., Pearson, L. M. and Johnson, W. O. (1992) Case-deletion diagnostics for mixed models. *Technometrics*, **34**, 38–45.

Cook, R. D. (1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression.* New York: Chapman and Hall.

Davison, A. C. and Tsai, C.-L. (1992) Regression model diagnostics. *Int. Statist. Rev.*, **60**, 337–353.

Dempster, A. P. and Ryan, L. M. (1985) Weighted normal plots. *J. Am. Statist. Ass.*, **80**, 845–850.

Dey, D. K., Gelfand, A. E., Swartz, T. B. and Vlachos, P. K. (1995) Simulation-based model checking for hierarchical models. Unpublished.

DuMouchel, W. and Waternaux, C. (1992) Discussion of a paper by CN Morris and SL Normand. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 338–341. Oxford: Clarendon.

Duncan, D. B. and Horn, S. D. (1972) Linear dynamic recursive estimation from the viewpoint of regression analysis. *J. Am. Statist. Ass.*, **67**, 815–821.

Evans, M. and Swartz, T. (1995) Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist. Sci.*, **10**, 254–272.

Fellner, W. H. (1986) Robust estimation of variance components. *Technometrics*, **28**, 51–60.

Fuller, W. A. (1980) The use of indicator variables in computing predictions. *J. Econometr.*, **12**, 231–243.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Clarendon.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–472.

Graybill, F. A. (1976) *Theory and Application of the Linear Model.* North Scituate: Duxbury.

Hall, C. B., Zeger, S. L. and Bandeen-Roche, K. J. (1996) Adjusted variable plots for regression with dependent data. Unpublished.

Harrison, J. and West, M. (1991) Dynamic linear model diagnostics. *Biometrika*, **78**, 797–808.

Henderson, C. R., Kempthorne, O., Searle, S. R. and von Krosigk, C. N. (1959) Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

Hilden-Minton, J. A. (1995) Multilevel diagnostics for mixed and hierarchical linear models. *PhD Dissertation.* University of California, Los Angeles.

Ho, Y.-Y., Peruggia, M. and Santner, T. J. (1996) Stage-wise outlier detection in hierarchical Bayesian repeated measures models. Unpublished.

Hocking, R. R., Green, J. W. and Bremer, R. H. (1989) Variance-component estimation with model-based diagnostics. *Technometrics*, **31**, 227–239.

Hodges, J. S. and Sargent, D. J. (1996) Counting degrees of freedom in hierarchical and other richly parameterized models. *University of Minnesota 25th Anniversary Meet., Minneapolis, June 22nd.*

Kass, R. E., Tierney, L. and Kadane, J. B. (1989) Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, **76**, 663–674.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the t distribution. *J. Am. Statist. Ass.*, **84**, 881–896.

Lange, N., Carlin, B. P. and Gelfand, A. E. (1992) Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *J. Am. Statist. Ass.*, **87**, 615–626.

Lange, N. and Ryan, L. (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.

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.

Louis, T. A. (1988) General methods for analyzing repeated measures. *Statist. Med.*, **7**, 29–45.

MacEachern, S. N. and Peruggia, M. (1995) Importance link function estimation for Markov chain Monte Carlo methods. Unpublished.

Moulton, B. R. (1987) Diagnostics for group effects in regression analysis. *J. Bus. Econ. Statist.*, **5**, 275–282.

Nelder, J. A. (1972) Discussion on Bayes estimates for the linear model (by D. V. Lindley and A. F. M. Smith). *J. R. Statist. Soc.* B, **34**, 18–20.

O'Hagan, A. (1976) On posterior joint and marginal modes. *Biometrika*, **63**, 329–333.

———(1985) Shoulders in hierarchical models. In *Bayesian Statistics 2* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 697–710. Amsterdam: North-Holland.

Peña, D. (1991) Measuring influence in dynamic regression models. *Technometrics*, **33**, 93–101.

Peruggia, M. (1997) On the variability of case-deletion importance sampling weights in the Bayesian linear model. *J. Am. Statist. Ass.*, **92**, 199–207.

Salkever, D. S. (1976) The use of dummy variables to compute predictions, prediction errors, and confidence intervals. *J. Econometr.*, **4**, 393–397.

Sargent, D. J. (1995) A general framework for random effects survival analysis in the Cox proportional hazards setting. *Report 95-004.* Division of Biostatistics, University of Minnesota, Minneapolis.

———(1996a) A general framework for hierarchical survival models in the Cox proportional hazards regression setting. *PhD Dissertation.* Division of Biostatistics, University of Minnesota, Minneapolis.

———(1996b) A hierarchical model method for subgroup analysis of time-to-event data in the Cox regression setting. *Joint Statistical Meet. Chicago, Aug.*

———(1997) A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Anal.*, **3**, 13–25.

Sargent, D. J. and Hodges, J. S. (1997) Smoothed ANOVA with application to subgroup analysis. *Research Report 97-002*. Division of Biostatistics, University of Minnesota, Minneapolis.
Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. New York: Wiley.
Sharples, L. D. (1990) Identification and accommodation of outliers in general hierarchical models. *Biometrika*, **77**, 445–452.
Smith, A. F. M. (1986) Some Bayesian thoughts on modelling and model choice. *Statistician*, **35**, 97–102.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Solomon, P. J. (1985) Transformations for components of variance and covariance. *Biometrika*, **72**, 233–239.
Stewart, G. W. (1987) Collinearity and least squares regression (with discussion). *Statist. Sci.*, **2**, 68–100.
Theil, H. (1971) *Principles of Econometrics*. New York: Wiley.
Wakefield, J. (1996) The Bayesian analysis of population pharmacokinetic models. *J. Am. Statist. Ass.*, **91**, 62–75.
Wakefield, J. C. and Bennett, J. E. (1996) The Bayesian modelling of covariates for population pharmacokinetic models. *J. Am. Statist. Ass.*, **91**, 917–927.
Weisberg, S. (1983) Comments on a paper by RR Hocking. *Technometrics*, **25**, 240–244.
——— (1985) *Applied Linear Regression*, 2nd edn. New York: Wiley.
Weiss, R. E. (1995) Residuals and outliers in repeated measures random effects models. *Technical Report*. Division of Biostatistics, University of California, Los Angeles.
——— (1996) Bayesian model checking with applications to hierarchical models. *Technical Report*. Division of Biostatistics, University of California, Los Angeles.
Weiss, R. E. and Lazaro, C. G. (1992) Residual plots for repeated measures. *Statist. Med.*, **11**, 115–124.
West, M. and Harrison, W. (1989) *Bayesian Forecasting and Dynamic Models*. New York: Springer.
Williams, D. A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.*, **36**, 181–191.

Discussion on the paper by Hodges

A. C. Atkinson (*London School of Economics and Political Science*)

It is now 20 years since the publication of Cook (1977), the most influential early paper on regression diagnostics. So it is appropriate that the Society has today organized, not one, but two, papers on diagnostics, the second appearing elsewhere (Langford and Lewis, 1998).

As the author makes clear, the algebraic basis of the early work on regression diagnostics is that exact formulae are available for the deletion quantities $\hat{\beta}_{(i)}$ and $s^2_{(i)}$, in standard regression notation—expression (4.10) in the paper. But such simple formulae no longer hold if there is a relationship between the mean and the variance. Diagnostics for generalized linear models (Pregibon, 1981) provide an example where the application of formulae analogous to expression (4.10) leads only to approximations to the effect of deletion.

The stimulating paper that we have heard today reformulates a large class of linear models in a form in which standard diagnostics can be applied, although the relationship between the mean and the variance renders the deletion results again approximate.

A class of model for which some results are available is the dynamic linear or structural time series model. One example is the random walk plus noise model

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t & \epsilon_t &\sim N(0, \sigma^2_\epsilon), \\ \alpha_{t+1} &= \alpha_t + \eta_t & \eta_t &\sim N(0, \sigma^2_\eta), \end{aligned} \tag{1}$$

a simpler form of the model of Section 2.1.2. In today's formulation Γ is a function of σ^2_ϵ and of σ^2_η. Dr Hodges uses an estimate of Γ which is not changed on deletion. A first approximation to $\hat{\Gamma}_{(i)}$ can be found by rewriting Γ in terms of σ^2_ϵ and the signal-to-noise ratio $q_\eta = \sigma^2_\eta/\sigma^2_\epsilon$. On the assumption that q_η remains unchanged, exact deletion formulae can be used for σ^2_ϵ, which is analogous to $s^2_{(i)}$ in the regression model. But it seems that full deletion may lead to zero estimates of variances such as σ^2_η, in line with the comments in the paper about likelihood and with the results of Shephard (1993).

In model (1) there are n observations and n constraint cases. So n observations give rise to $2n$ residuals, just as the auxiliary residuals of Harvey and Koopman (1992) yield a series of residuals for each component of the model.

The interpretation of residuals may indeed be less easy than it is for regression, when we hope that there will be no pattern in a plot of residuals against fitted values. But we have learned not to expect this for generalized linear models. An example of what we should expect is in Fig. 5.17 of Collett (1991). So we may learn to live with Fig. 4. Since expression (2.10) with its large matrices will often not be used for

fitting, methods of calculation of the residuals may be important. Sometimes they may be most easily calculated by the use of the equivalence, for regression models, of deletion and mean shift outliers in Section 4.4. Interventions ('shocks') are used by de Jong and Penzer (1998) to test for outliers in the lower level of model (1). They also use a single scalar intervention in the higher level to test for a jump in the random walk. The calculations, provided that variances are not re-estimated, are little more complicated than those for regression, relying on the algebra of Section 4.1 for added variables.

I have some more detailed comments on Sections 4 and 5. Added variable plots can sometimes fail if important information is contained in points with high leverage, which may give small residuals. It may be helpful to supplement them with index plots of deletion versions of quantities of interest. I was not a referee of the paper, and I was surprised by the choice of constructed variable to test for a transformation. Andrews's variable ignores the Jacobian of the transformation and results in a test with lower power than the other constructed variables. Fig. 8.12 of Atkinson (1985) illustrates this point.

The paper's diagnostics are concerned with single-case deletions. As we know, these can fail owing to masking if several outliers or influential cases are present. One approach is to use very robust procedures. Examples of diagnostics following least median of squares regression are given by Atkinson (1986). However, more insight into the structure of the data may sometimes be obtained by using the forward search for robust estimators (Hadi, 1992; Atkinson, 1994), calculating diagnostic quantities of interest as the search progresses.

Several examples for transformations of univariate and multivariate data are given by Riani and Atkinson (1998). Fig. 9 shows a typical graph, in fact for the modification of the Box and Cox poison data which Andrews (1971) used to illustrate his procedure. Five values of λ were taken (-1, -0.5, 0, 0.5 and 1). For each the data were transformed and an initial outlier-free basic subset was found. The five forward searches used the values of the least squares residuals to decide which cases were to be included in the next subset. Fig. 9 shows the values of the five score statistics, calculated using the constructed variable including the Jacobian. The vertical jumps in the value indicate where the outlier was included—in the last step of the search for $\lambda = -1$, -0.5, 0. It is clear, for example, that $\lambda = -1$ is supported by all the data except the outlier and, on the contrary, that support for the log-transformation depends on the outlier.

Similar insight into the structure of the data can be obtained for the discriminant analysis of multivariate data (Atkinson and Riani, 1997). Fig. 10 is a forward search on Mahalanobis distances for Fisher's iris data, on the assumption of a common covariance matrix for all three groups, so that linear discriminant analysis is appropriate. It is clear that the three groups are not behaving in the same way and that a rather different set of cases are entering group one at the end—a set which can readily be identified (once one has been alerted) by looking at scatterplots of the data.

My question for Dr Hodges is whether the forward search can be extended to hierarchical models. Is

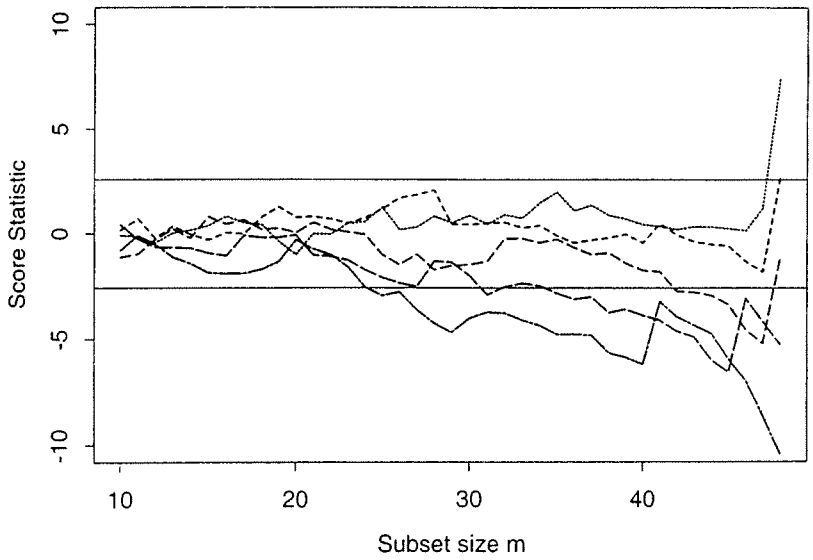

Fig. 9. Modified poison data: score statistic for power transformation as the subset size m increases; individual searches for each λ (— · — ·, $\lambda = 1$; — — —, $\lambda = 0.5$; – – – –, $\lambda = 0$; · · · · ·, $\lambda = -0.5$; ·····, $\lambda = -1$); the upward jumps, resulting from the introduction of the outlier, occur at different values of m for the various searches

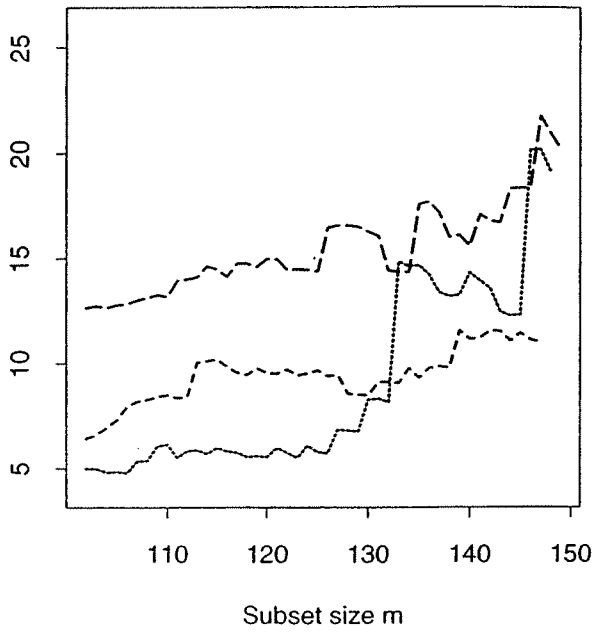

Fig. 10. Iris data: minimum Mahalanobis distances, for the three groups, of units not belonging to the subset: ....., group 1; - - - - -, group 2; – – – –, group 3

refitting at each stage of the search computationally feasible: on what sort of residual should the forward search be based?

Finally I am puzzled by the title of the paper. I can see the algebra, but where explicitly is the geometry?

It gives me great pleasure to propose the vote of thanks.

Jon Wakefield (*Imperial College of Science, Technology and Medicine, London*)

This paper is a welcome and timely addition to the literature on hierarchical models as it addresses an area that has received little attention: the determination of the adequacy of the many layers of assumptions that such models contain. I would first like to emphasize that the determination of model adequacy

(a) should be carried out with respect to the substantive questions of interest and
(b) cannot be carried out in isolation of the context; in particular the interpretation of diagnostics requires subject-matter knowledge.

Suppose that we have, possibly after transformation of the response,

$$y_{ij} = g_1(\theta_i, x_{ij}) + \epsilon_{ij}$$

where the ϵ_{ij} are independent and identically distributed (IID) as $p_1(\cdot|\sigma^2)$, $i = 1, \ldots, N$; $j = 1, \ldots, n_i$. The subscript '1' here reflects the fact that we are modelling a first-stage variable. This paper considers the situation in which g_1 is linear though when the data per individual are not sparse then the following shows that the methods may also be used when g_1 is non-linear. In this case the maximum likelihood estimators $\hat{\theta}_i$ may be obtained and the first stage of the hierarchy may be replaced by $\hat{\theta}_i \sim N\{\theta_i, I(\hat{\theta}_i)^{-1}\}$ where $I(\cdot)$ denotes Fisher's information matrix. This was used for computation in non-linear models by Racine-Poon (1985).

At the second stage we have

$$\theta_i = g_2(\gamma, z_i) + \delta_i$$

where the δ_i are IID as $p_2(\cdot|\Sigma)$. Assessing model adequacy at stages two and higher of the hierarchy is inherently more difficult because the quantities that we are modelling are unobserved. The choice of the functional form g_2 is particularly difficult. In many situations the model will be parameterized such that

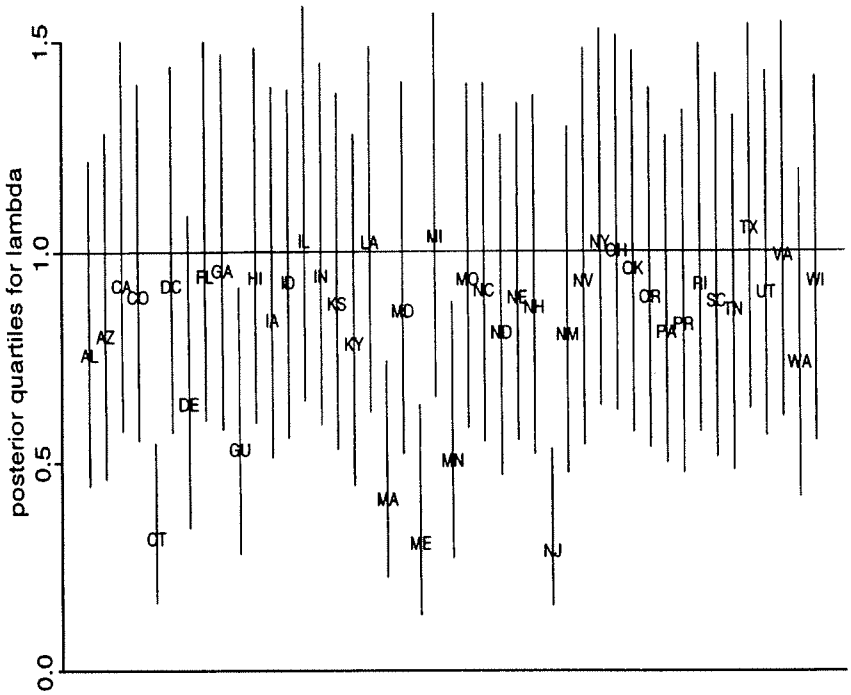

Fig. 11. 50% posterior intervals for outlier detection parameters λ_i, $i = 1, \ldots, 45$: the prior expectation is 1 and the posterior medians are indicated by the state labels

the elements of θ_i take values on the whole real line and a linear model $g_2(\gamma, z_i) = z_i\gamma$ is assumed. More complex models may be assumed if N and n_i are large (the latter ensuring that the δ_i are well estimated), or if the context provides a model based on subject-matter considerations. When a linear model is assumed the choice of which regressors to include for which *element* of θ_i is very difficult, in part because the δ_i are multivariate with correlated elements. The choice will again often be driven by subject-matter and not statistical considerations alone. At this and all other stages model adequacy should be judged with respect to a specific question of interest. Wakefield and Bennett (1996) and Wakefield *et al.* (1998) discuss these issues in the context of population pharmacokinetic–pharmacodynamic modelling.

For the example in the paper I carried out an analysis with a Student t-distribution with low degrees of freedom υ at the second stage. This would be my default choice here for several reasons. Computationally there is negligible difference when using the scale mixture of normals representation where δ_i are assumed to arise from $N(0, \lambda_i^{-1}\Sigma)$ with λ_i distributed as $\text{Ga}(\upsilon/2, \upsilon/2)$. Interesting features are likely to be more prominent since there is less shrinkage and outlying units are flagged and accommodated, making inference more robust. Highlighting outlying units is important as it may help to indicate covariate relationships when individual components of δ_i are examined, and specific units who are outlying may require extra attention. Fig. 11 shows the quartiles of the posterior distributions of the λ_i. We see that many states have the majority of their posterior distributions below the prior expectation of 1, including three of the five New England states, offering a clue that these states are not exchangeable.

The added variables plots of the paper may be constructed to compare each element of θ_i with a particular z_i but if there is correlation in the random effects distribution then the plots for different elements of θ_i will show correlation. I am a little troubled by the use of a single summary estimate of the parameter being used in the construction of the plot, particularly when there is correlation in the random effects distribution. An alternative would be to construct multiple plots using draws from the posterior distribution.

In general the use of the complete posterior distribution of diagnostic quantities should be encouraged as more information is provided. Fig. 12 shows the plot which I would have constructed (Wakefield, 1996) to assess whether the expenses per admission covariate should be used as an explanatory variable. For each state I have taken the quartiles of the posterior distribution and plotted these against the expenses per admission of that state. The smoother may help to detect non-linearities and the use of quartiles accounts for the differing precisions of the premiums. This plot can be compared

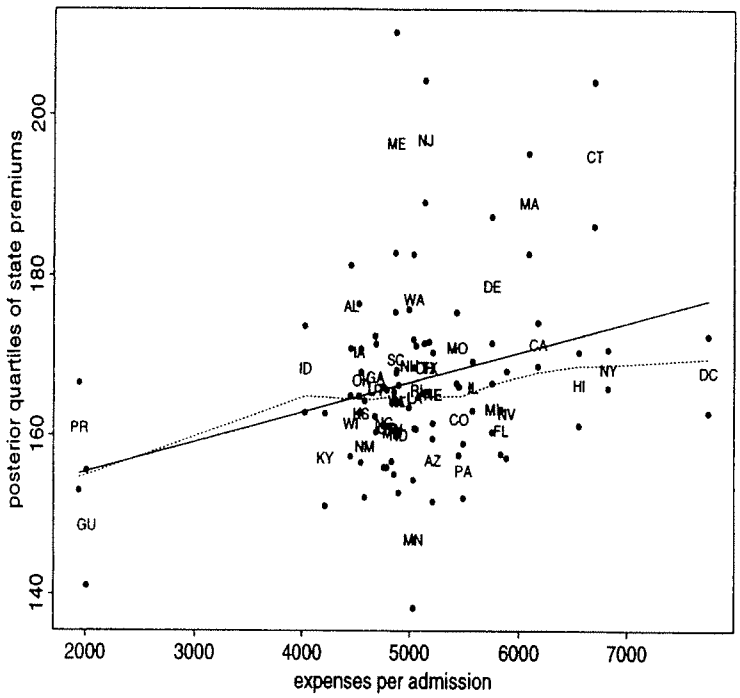

Fig. 12. Posterior summaries of state premiums *versus* expenses per admission covariate: for each state three points are plotted, the lower and upper quartiles and the median (indicated by the state labels) of the posterior distribution for each θ_i; all three points contributed to these fits (———, least squares fit; ·····, smoother)

with Fig. 8 of the paper and is on a more natural scale. I think that more experience using the added variable plots is required; in particular what do the data cases tell us beyond the constraint cases?

Continuing the theme of the use of the complete posterior distribution, Fig. 13 shows normal scores plots constructed from nine draws from the posterior distribution from a second analysis that I carried out using a normal second-stage distribution (to allow comparison with Fig. 6 of the paper). These plots account for the variability in the posterior distributions of each θ_i. Simulating these diagnostics from the 'true' model offers one opportunity for understanding the properties of such plots.

I found some evidence of different within-state variability for the data of the paper. Fig. 14 shows 95% intervals for the standard deviations when different variances were allowed for each. If it were believed that there were different within-state variances then one possibility would be to assume a hierarchy for these also, or to have exchangeability within covariate-defined groups of states. The final decision should be based, if possible, on other data (from previous years for all a subset of the states) and on the aim of the study. For example if one wanted to predict the *observed* premiums within a state then either a common τ^2 or a hierarchical structure is required.

When covariates are included in the model then, as Fig. 15 shows, the behaviour of the Gibbs sampler is very poor because there is little information in the data to isolate the components of within- and between-state variability. In fact Fig. 16 shows that there are two competing explanations for the observed variability in the data. It would be interesting to look at non-Bayesian estimates of τ^2 here since the likelihood is concentrated close to the boundary. Unluckily in the paper the chain was of insufficient length for this behaviour to be recognized and an underestimate of τ^2 resulted in Table 1. The two sets of time series in Fig. 15 correspond to 'good' and 'bad' parameterizations (Gelfand *et al.*, 1995, 1996; Roberts and Sahu, 1997). Interestingly in this case, in line with theoretical results of the aforementioned references, the 'uncentred' parameterization (γ, δ_i) is preferred because the within-state variance σ^2 is larger than the between-state variance τ^2. A strength of the Bayesian approach is the possibility of isolating the source of variability by including prior distributions based on other information.

In conclusion this paper offers some interesting new diagnostic tools but the aim of the analysis and the background context remain of paramount importance. As I hope is obvious from my comments, I enjoyed the paper and I have great pleasure in seconding the vote of thanks.

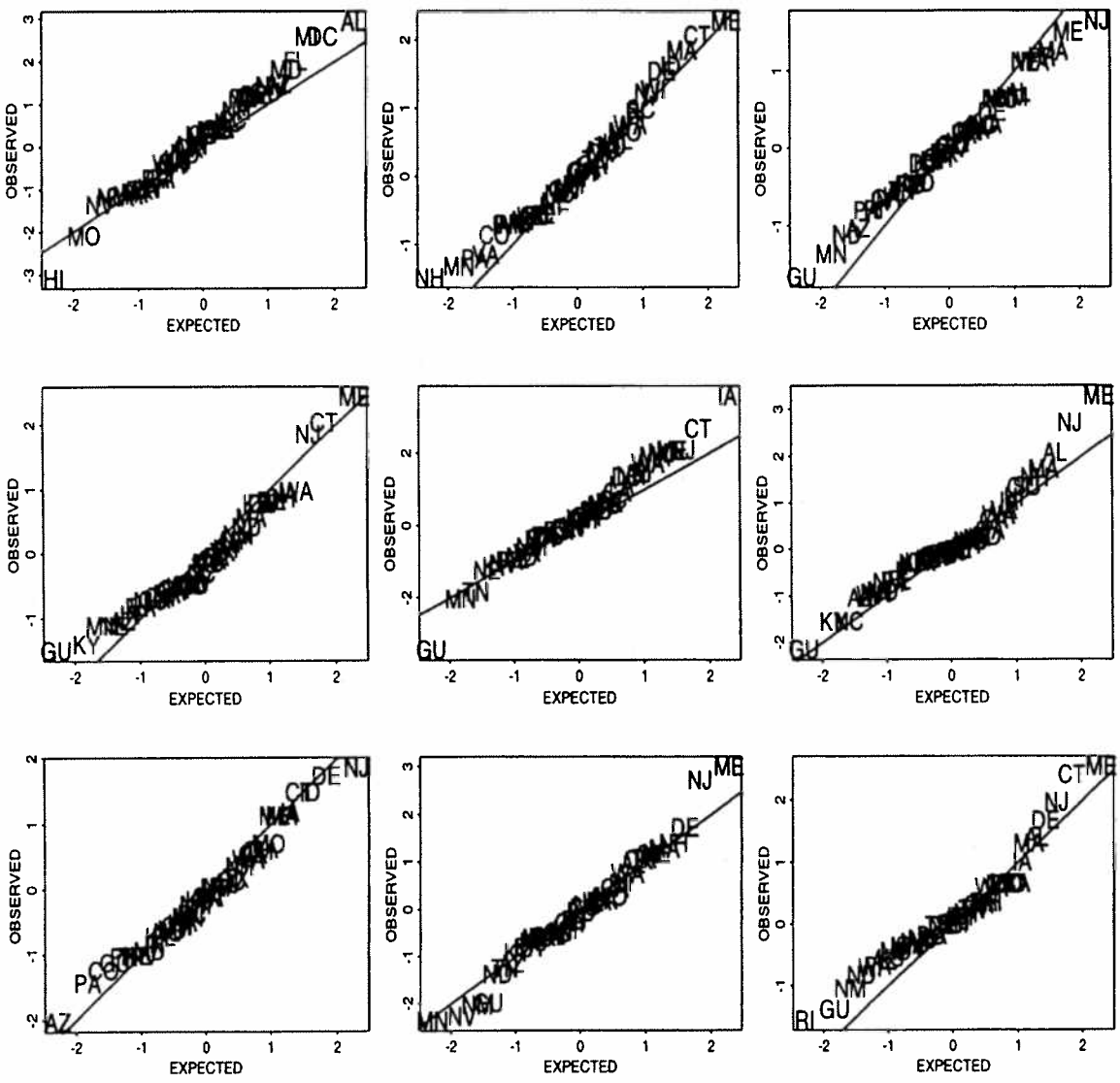

Fig. 13. Normal scores plots of the second-stage state premiums, for nine draws from the posterior distribution

The vote of thanks was passed by acclamation.

J. A. Nelder (*Imperial College of Science, Technology and Medicine, London*)
In December 1995 Professor Youngjo Lee and I gave a paper to the Society on hierarchical generalized linear models (HGLMs) (Lee and Nelder, 1996). In this we defined an extension to GLMs to include random effects, not necessarily normal, and a criterion for fitting them, the h-likelihood. We showed that, given the dispersion components, fixed and random effects could be estimated together using iterative weighted least squares applied to an extended data matrix with extra rows, one for each random effect. On p. 632 of Lee and Nelder (1996), below equation (4.5) you will find a remark (coming from Professor Lee) that 'In conjugate HGLMs the equations (4.4) have the following nice interpretation'. The interpretation that follows is, for the normal–normal HGLM, identical with the reformulation of the speaker. It is, though, more general because it applies to all conjugate HGLMs, such as the Poisson–gamma, binomial–beta and gamma–inverse gamma. I recommend this approach to the author.

I recently wrote in a script a remark to the effect that at least once a year I hear a speaker enunciate a dichotomy between Bayesian and frequentist modes of inference; at the head of Section 3.2.2 is my 1997 example. I know that this dichotomy is false because I subscribe to a likelihood mode of inference. In the current context, use of the h-likelihood avoids the integration required to form marginal likelihoods,

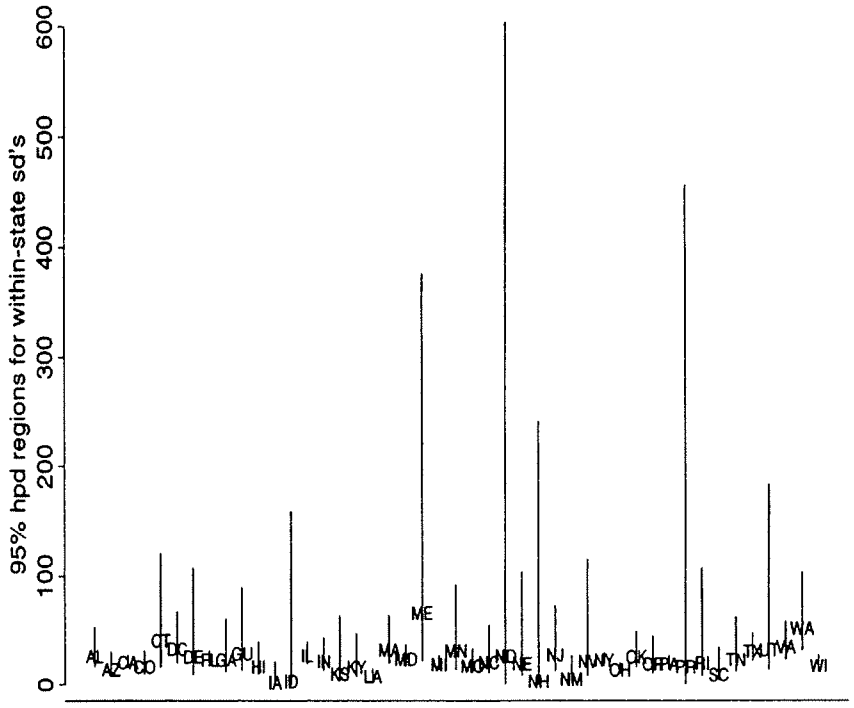

Fig. 14. Posterior distributions of the within-state standard deviations: state labels are at the posterior medians

and the basic fitting does not require Markov chain Monte Carlo methods; furthermore, use of the profile h-likelihood allows likelihood intervals to be determined that are not upset by parameter estimates lying on the edge of the parameter space, or by likelihood surfaces that are not quadratic in the neighbourhood of the mode.

The author discusses the complication introduced into the plot of residuals *versus* fitted values by the presence of random effects in the latter. We noticed this when developing model checking procedures for HGLMs, and Professor Lee has found what we believe to be the solution. This is to use fitted values based on the fixed effects only; the correlation disappears, and the plots are useful. We have not yet looked at extending partial residual or added variable plots to HGLMs, and for these the methods in the paper will make a good starting point.

Nicholas T. Longford (*De Montfort University, Leicester*)

The value of this paper is in describing a general approach (*process*) to diagnostics, rather than offering yet another *procedure* that would need fixes ('further research') for non-standard settings. My image of a Bayesian ideal for model diagnostics is to gather prior information about the kinds of model violations that are more, or less, likely to occur. In some cases, we may have prior information that certain units are more likely to be outliers of a specific or an unspecified kind. Obvious examples in the analysis of the health maintenance organization data are the jurisdictions of Guam and Puerto Rico, and the plans with the smallest enrolment, which may have a radically different composition of the clientele from the other plans. How could such information be incorporated in the process of model checking?

David Draper (*University of Bath*)

It is a pleasure to welcome Jim Hodges to this side of the pond and to these proceedings, and to thank him for an interesting and useful paper. I was a little surprised on reading it, however, to see no direct mention of predictive diagnostics. For instance, in hierarchical models one can

(a) successively omit observations (one at a time, say) at all levels of the hierarachy,
(b) construct predictive distributions for the omitted outcomes based on the non-omitted data,
(c) locate the omitted outcomes in their respective predictive distributions in some appropriate way, e.g. (Draper, 1996) z-scores or ratios of predictive densities at the observed outcomes to their respective maxima of predictive densities, and
(d) aggregate the summaries in (c) in some appropriate way that is sensitive to the model hierarchies.

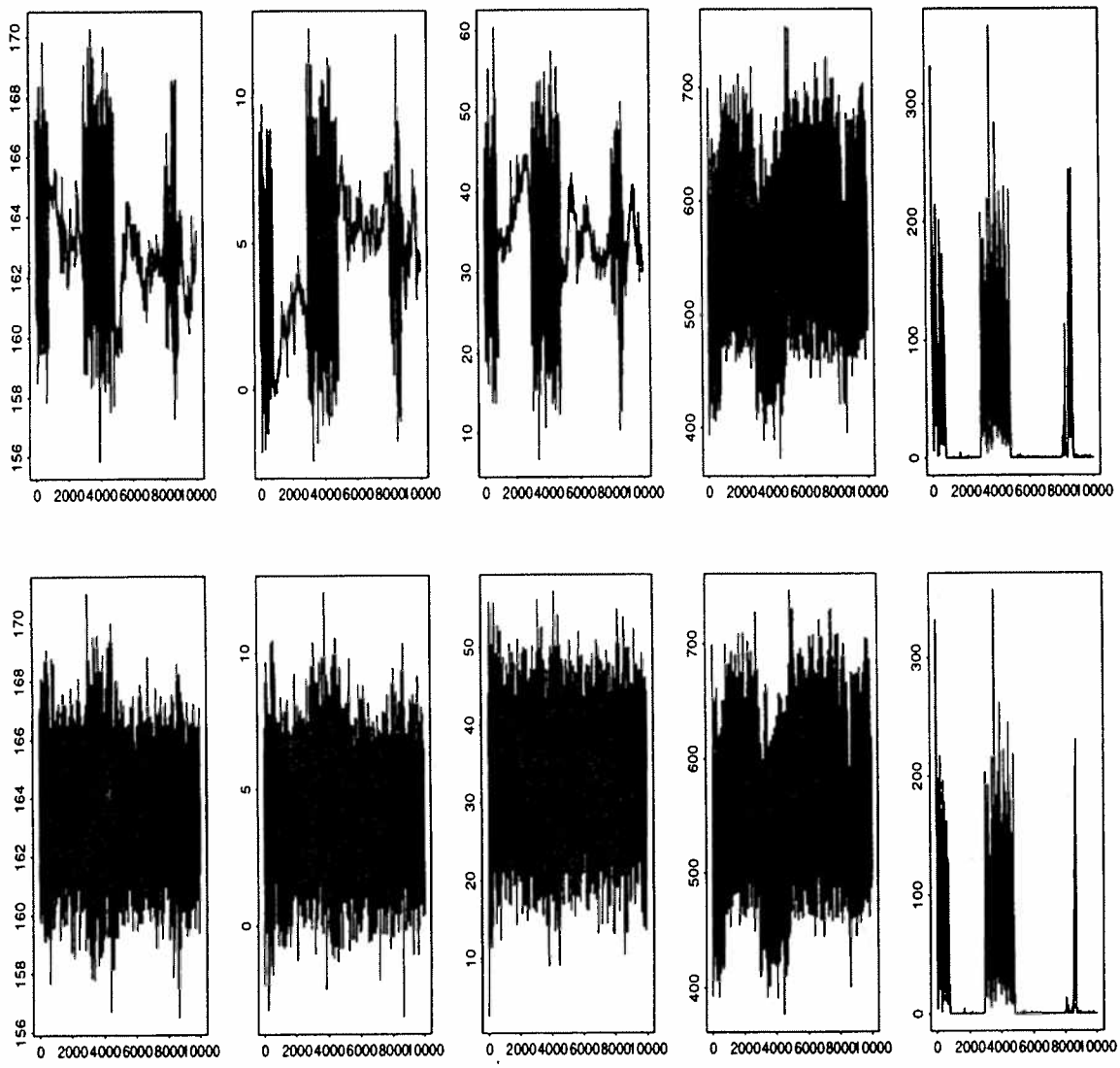

Fig. 15. Markov chain Monte Carlo time series of parameters from an analysis including two second-stage covariates: from left to right the columns represent γ_0, γ_1, γ_2, σ^2 and τ^2; the top row corresponds to the 'poor' parameterization (γ, θ_i) and the bottom row to the 'good' parameterization (γ, δ_i) where $\gamma = (\gamma_0, \gamma_1, \gamma_2)$

The goals of such analyses are both to identify outliers at all levels and to create omnibus graphical fit measures at each level. All this sounds like it should be closely related to the paper's case influence diagnostics, in particular the signed single-parameter analogue to Cook's distance, although the summaries are different, so I would appreciate comments on the following. What are the precise relationships between the author's diagnostics and the above predictive measures; is it likely that certain kinds of extra-model structure are more readily identified with one or other of these approaches; if so, how do we know in practice when to use one and when the other?

Frank Critchley (*Birmingham University*)

In warmly welcoming Dr Hodges's paper, I would like to sketch its relationship to a wider approach being pursued with colleagues at Birmingham. The author emphasizes the role played in the linear model by Euclidean geometry. Our approach is similarly inspired by considerations of geometry, viewing it as the mathematically natural way to deliver statistically desirable invariance. In another Minnesotan connection, we also use XLISPSTAT to exploit more of the potential of multivariate dynamic graphics.

Statistical science often proceeds by adopting a *working* problem formulation

$$\mathrm{PF}^0 = (Q, \text{data, model, inference method})$$

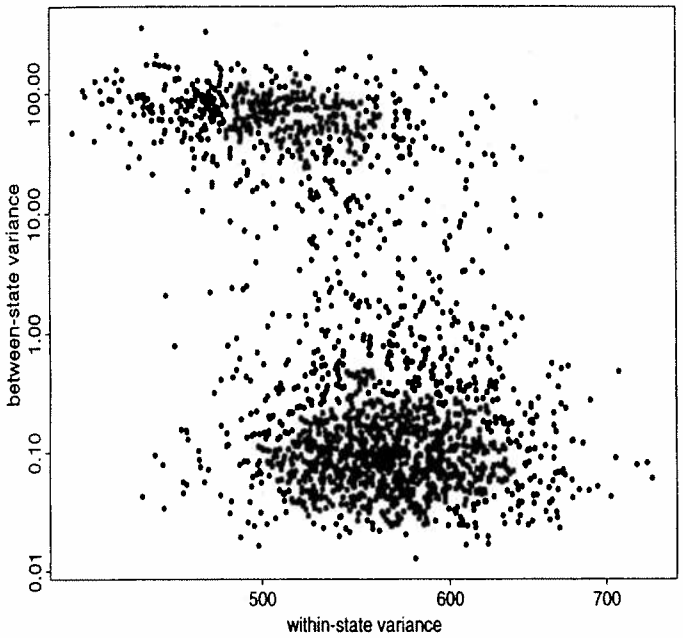

Fig. 16. Samples from the bivariate posterior distribution $p\{\log(\sigma^2), \log(\tau^2)|y\}$: the logarithmic scale clearly shows the two modes which correspond to different explanations for the observed variability in the data

to produce an answer A to the question of interest Q. In this broad conception, the role of diagnostics is to explore 'interesting' alternative problem formulations PF and their effect — if any — on A. In a global extension of Cook (1986), the log-likelihood of the model in PF^0 is identified as the null case $\omega_l = \omega_l^0$ of a set of perturbed log-likelihoods $\{l(\theta; \omega_l): \theta \in \Theta, \omega_l \in \Omega_l\}$. If the inference method calls for a prior $\pi(\theta)$ or utility function $U(\theta; a)$, these are similarly identified as the null cases $\omega_\pi = \omega_\pi^0$ or $\omega_U = \omega_U^0$ of additional perturbation parameters. With ω^0 denoting the null case of the complete perturbation parameter ω, a change $\omega^0 \to \omega$ brings a change $PF^0 \to PF$ in the problem formulation. Its effect on A is monitored by tracking the induced change $\tau(\omega^0) \to \tau(\omega)$ in a suitable target function. In particular contexts this may, for example, be Cook's likelihood displacement function, the Kullback–Leibler divergence between posterior distributions under ω^0 and ω or (the expected utility of) the optimal decision procedure.

Invariance to smooth one-to-one changes $\omega \to \omega^*$ in the perturbation parameter is highly desirable in as much as ω is merely a label. In this rather general set-up it is not immediately clear what it means to go 'straight' from ω^0 to ω, nor how large this perturbation is. A geometrically natural way to answer these questions invariantly is to put a metric tensor on perturbation space Ω. The key question now is whether or not the chosen metric is flat. There is a standard way to check this. The good news is that the answer is in the affirmative for a statistically natural choice of metric and for a wide range of problems. The premium on flatness is that then (and only then) there is a preferred parameterization in which Ω *is* Euclidean, which brings us back to the present paper.

Finally, I have two specific comments. In as much as the higher level parameters reflect a hidden layer, neural networks combined with (Bayesian) graphical models appear to offer a more flexible form of modelling them than, say, equation (2.9). Again, there was no treatment of the familiar variance perturbation scheme. This would be particularly interesting here since, as the author notes, Γ determines the underlying geometry.

A. C. Davison (*Swiss Federal Institute of Technology, Lausanne*)

There seems to be some disagreement about Bayesian model checking, as a comparison of Carlin and Louis (1996) and Gelman *et al.* (1995) shows. One approach, described by Box (1980), starts from the position that all the information about the problem from the data, the current sampling model $f(y|\theta)$ and prior density $\pi(\theta)$ is contained in the joint density

$$f(y, \theta) = f(y)\,\pi(\theta|y).$$

The second term on the right-hand side is to be used for inference about θ, conditional on correctness of the model, while the first contains information on model fit, and may be given a repeated sampling interpretation. This rests on accepting that the argument from the likelihood principle against such an interpretation — and its adjuncts such as significance tests — falls when the model itself is in question. If this is accepted it is useful to decompose $f(y)$ further.

Suppose that a statistic $s = s(y)$ is minimal sufficient for θ, and that a is a distribution constant function of s. Then

$$f(y) = f(y|s)\, f(a) \int f(s|a;\, \theta)\, \pi(\theta)\, \mathrm{d}\theta,$$

whose first two terms can be used to check the sampling model alone; the third provides information on the agreement between it and the prior. Consider the regression–scale model in which

$$y_{ij} = x_{ij}^{\mathrm{T}} \beta_i + \sigma_i \epsilon_{ij}, \qquad j = 1, \ldots, n_i,\ i = 1, \ldots, k,$$

where the ϵ_{ij} are a random sample of variables from a known distribution, and let $\hat{\beta}_i$ and $\hat{\sigma}_i$ denote maximum likelihood estimates of β_i and σ_i. Then the quantities $a_{ij} = (y_{ij} - x_{ij}^{\mathrm{T}} \hat{\beta}_i)/\hat{\sigma}_i$ are distribution constant and can be used for model checking, independent of the prior. The a_{ij} are essentially just the usual scaled residuals. Since the leverages are non-random, calculation and plotting of derived quantities such as standardized residuals and measures of influence are justified in a Bayesian framework and may be given their usual interpretation. This gives a Bayesian argument for using standard methods of model checking at the lowest level of a hierarchical model, regardless of the superstructure above it.

As an example of this, the models used in the paper assume a common value of σ_i for every state, and it is natural to question this. A standard test for unequal variances based on the residuals gives a significance level of 0.005, but as this changes to 0.086 when Washington, with its large outlier, is dropped, the assumption of common variance seems to be supported by the data. This is independent of any prior belief about the values of the variances.

In a model with normal errors, $\hat{\beta}_i$ and $\hat{\sigma}_i$ are sufficient, and the next level of the hierarchy would presumably be checked by comparing these statistics with their posterior means. In more general models this would not be possible, because the conditional distributions $f(s|a;\, \theta)$ would come into play, though presumably some progress could be made from accurate approximations to conditional distributions and densities; see for instance Barndorff-Nielsen and Cox (1994).

Thomas Leonard (*University of Edinburgh*)

I would like to congratulate Dr Hodges on a splendid paper in an important area. The choice of inferential procedure for the variance components is of course of essential importance (e.g. Sun *et al.* (1996)), particularly when inference regarding the first-stage parameters is of primary concern, and frequency properties are also of interest. I have always liked Jim's philosophy of mixing Bayesian and Fisherian ideas, with applied objectives in mind.

In situations where there is a little prior information regarding the second-stage variance components, I find it convenient to use uniform distributions, rather than the log-uniform distributions which typically lead to improper posteriors. (See Strawderman (1971), Leonard (1976) and Leonard and Hsu (1992).) In the special case discussed in Section 5.1, I would employ uniformly flat distributions for both σ^2 and τ^2, as well as for the linear parameters in the model, and then hope for good frequency properties.

This becomes an important issue in Poisson and multinomial situations (e.g. George *et al.* (1994) and Leonard and Hsu (1994)). For example, hierarchical models for Poisson means, using first-stage gamma assumptions, generally require proper distributions for the first-stage parameters (e.g. Leonard and Novick (1986)). However, an application of Dr Hodges's more flexible multivariate normal hierarchies to the logarithms of the Poisson means would introduce the feasibility of the flat second-stage distributions discussed above. Other link functions to Hodges's second-stage models are appropriate in many other situations where the first-stage distribution is not normal. In the normal case with unknown covariance matrices, these hierarchical models can also be applied to the matrices logarithms of the covariance matrix and this permits the modelling of the type of noisy data which might reject the more standard models, described in Jim's paper, via his interesting analyses of residuals.

The following contributions were received in writing after the meeting.

Ronald Christensen (*University of New Mexico, Albuquerque*)
Computing devices can be wonderful things. The most useful that I know is called generalized linear models. I have much the same feeling about this work. It seems like an extremely useful computing device, but it does not strike me as theory for any branch of statistics. In particular, I have a difficult time taking equation (2.7) seriously as a linear model, because Θ is a function of $\mathbf{E}$. I think that the point of the work is that computing devices developed for linear models can also be applied to hierarchical Bayes models by writing equation (2.7). The danger, of which the author is well aware, is that, although equation (2.7) makes very fruitful suggestions about methods of data analysis, since it is only a computing device, each suggestion must be evaluated individually.

Consider a quadratic form $(Y^* - X^*\beta)'W^{-1}(Y^* - X^*\beta)$. For our purposes, this will appear in the exponent of a normal density. In finding frequentist sampling distributions or Bayesian posterior distributions, it is necessary to change the quadratic form into one that depends more directly on β. Write the appropriate oblique projection operator as $A^* \equiv X^*(X^{*\prime}W^{-1}X^*)^{-1}X^{*\prime}W^{-1}$. Then

$$(Y^* - X^*\beta)'W^{-1}(Y^* - X^*\beta) = (Y^* - A^*Y^*)'W^{-1}(Y^* - A^*Y^*) + (A^*Y^* - X^*\beta)'W^{-1}(A^*Y^* - X^*\beta)$$
$$= \mathrm{SSE}^* + (\beta^* - \beta)'(X^{*\prime}W^{-1}X^*)(\beta^* - \beta)$$

where $\beta^* \equiv (X^{*\prime}W^{-1}X^*)^{-1}X^{*\prime}W^{-1}Y^*$ and the first term on the right-hand side does not depend on β. This computational device is fundamentally geometrical. It is essentially the Pythagorean theorem.

As an example, consider the linear model $Y = X\beta + e$, $e \sim N(0, \sigma^2 I)$. Setting $Y^* = Y$, $X^* = X$ and $W = \sigma^2 I$ leads to the sampling distribution of $\hat{\beta}$. Now consider a Bayesian model $Y|\beta \sim N(X\beta, V)$ with a prior for β defined by $\tilde{X}\beta \sim N(\tilde{Y}, \tilde{V})$ (see Bedrick *et al.* (1996) and Christensen (1996), section 2.9). The quadratic form for the joint distribution is

$$(Y - X\beta)'V^{-1}(Y - X\beta) + (\tilde{Y} - \tilde{X}\beta)'\tilde{V}^{-1}(\tilde{Y} - \tilde{X}\beta) = \left(\begin{pmatrix} Y \\ \tilde{Y} \end{pmatrix} - \begin{pmatrix} X \\ \tilde{X} \end{pmatrix}\beta\right)' \begin{pmatrix} V^{-1} & 0 \\ 0 & \tilde{V}^{-1} \end{pmatrix} \left(\begin{pmatrix} Y \\ \tilde{Y} \end{pmatrix} - \begin{pmatrix} X \\ \tilde{X} \end{pmatrix}\beta\right).$$

The computing device equates this to a quadratic form that does not depend on β plus another that does. The posterior for β depends only on the second form. The idea of treating equation (2.7) as a linear model is quite simply a suggestion to use this computing device. Moreover, thinking about equation (2.7) as a linear model *suggests* many things. For example, case deletion methods can clearly be applied. When applied to actual data cases, the meaning is clear. We obtain the posterior without that data case. However, when applied to the rows of $\tilde{Y} - \tilde{X}\beta$, the meaning is less clear. (The actual meaning is the replacement of a proper prior with an improper prior; see Bedrick *et al.* (1996).) Ultimately, the rationale for any such procedure must be based on Bayesian distribution theory.

D. R. Cox (*Nuffield College, Oxford*) **and P. J. Solomon** (*University of Adelaide*)
Dr Hodges's interesting paper fills some of the gaps in assessing model adequacy for hierarchical and related models. A rather different approach examines specific kinds of explicitly formulated departures from the standard model. One such is via a transformation of the response variable (Solomon, 1985) and the references given by Dr Hodges in Section 4.2. Another (Solomon and Cox, 1992) is via a fairly general second-order non-linear model allowing the separation of three features of potential interest: non-normality of the variation within groups, correlation between the variance within a group with the group mean and non-normality of the distribution of the group population means.

Furthermore, at the lowest level of variability, where the repeated measures may be considered as many small random samples, the methods of Cox and Solomon (1986, 1988) are appropriate for detecting unusual structure or non-normality. Cox (1998) discusses issues of adequacy of fit for variance component models, including suggestions for handling outliers and unbalanced data.

Wing K. Fung and H. Gu (*University of Hong Kong*)
By expressing the higher levels of the hierarchical models in terms of artificial 'cases' to the data set, the author suggests formulating the models in the form of ordinary linear models. The rich class of diagnostics in linear regression models can then be generalized to hierarchical models. The reformulation is very general, and it covers models such as random effect and mixed effect models. The paper makes a useful contribution to diagnostics in such models. In regression, two popular statistical

diagnostics are the residual and the leverage measure. The first diagnostic is fully discussed in Section 4.5, but the second is hardly mentioned in the paper. It is well known that the leverage measure also plays an important diagnostic role in many different kinds of linear models such as the mixed effect linear model (Christensen *et al.*, 1992) and the generalized linear model (Preisser and Qaqish, 1996). Can the author comment on the possible construction of the measure in hierarchical models?

The author also discusses the Box–Cox transformation for the outcome variable. Atkinson (1985) has described some alternative parametric transformations. Recently, Cook and Weisberg (1994) and He and Shen (1997) proposed a very general class of transformations which can be estimated by using a variety of smoothers and the B-spline functions. Their methods often work well and give very similar results in practice (Shi and Fung, 1998). Does the author have any comments about applying these transformations in hierarchical models?

Andrew Gelman (*Columbia University, New York*) **and Phillip N. Price** (*Lawrence Berkeley National Laboratory*)

This paper makes the interesting and important contribution of viewing the pseudodata expression of the hierarchical linear model as a unifying tool in model checking. (The analytical and computational use of the formulation for Bayesian inference is well known; see, for example Dempster *et al.* (1991) and Gelman *et al.* (1995), chapter 13.) More generally, the author notes that hierarchical regression models are more difficult to understand than we might imagine, especially if predictors appear at more than one level in the hierarchy. We would like to echo that point with a statistical anecdote of our own (see Price *et al.* (1996)).

We fit a hierarchical linear model to the logarithms of home radon measurements in Minnesota, with random effects for the counties of measurement. The model thus had variance components at the within- and between-county levels, which we examined to obtain an idea of the model's precision. We then added an individual level predictor — an indicator for whether each house had a basement. Adding the predictor caused the estimated (posterior median) within-county variance to decrease (as expected; houses with basements tend to have higher radon levels), but the estimated between-county variance substantially *increased*, which was completely unexpected. What was happening? After some thought, we realized that the counties with more basements happened to have higher random effects (in the second model). In the first model, much of the variation in county radon levels was cancelled by an opposite variation in the proportion of basements. The increased between-county variance in the second model indicates true variation among counties that happened to be masked by the first model.

This sort of pattern, caused by correlation between individual level variables and random effects, does not occur in non-hierarchical regression. We suspect that the tools developed in the paper under discussion will be useful in understanding this sort of data structure, and we look forward to future work in this area.

Finally, we disagree with the claim in Section 3.2.2 that 'Bayesian diagnostics . . . place an extra layer of mathematics between the analyst and the data'. Posterior predictive checks — comparisons of observed data with their predictive distribution under an assumed model — can be performed graphically and, in fact, can be simpler to interpret than classical methods such as Studentized residual plots (see Rubin (1984) and Gelman *et al.* (1996)).

Jian-Xin Pan (*Rothamsted Experimental Station, Harpenden*)

In what follows I concentrate on three topics that have been neglected by the author.

Masking and swamping effects of outliers

In ordinary regression diagnostics, masking and swamping effects are common (see, for example, Barnett and Lewis (1984)), in which one observation may conceal the importance of another. In the hierarchical models, these effects become more complex because the masking and swamping effects may apply across the hierarchies of the models. In the health maintenance organizations data analysed in the paper, for example, the outlying nature of a state may be masked and swamped by another outlying state or by a small number of plans within the state. For the hierarchical models, how can we effectively deal with masking and swamping effects either within a specific hierarchy or across the hierarchies?

Case influence on covariance Γ

The covariance matrix of the hierarchical models (2.12) is of block diagonal form: $\Gamma = \mathrm{diag}(\Gamma_1, \Gamma_2, \Gamma_3)$ where Γ_i $(i = 1, 2, 3)$ are covariance matrices of the data, the constraint and the prior cases respectively. When Γ is unknown, the author suggested using Gibbs resampling methods to estimate Γ. Is it possible

to calculate the maximum likelihood estimate or the restricted maximum likelihood estimate of Γ (e.g. Patterson and Thompson (1971)) directly? Furthermore, how should we assess a case influence on the covariance Γ when it is deleted from the models? It seems to me that the approximation of case influence for $\hat{\Theta}$ given by expression (4.10) is too rough since the estimate $\hat{\Theta}_{(-r)}$ actually depends on $\hat{\Gamma}_{(-r)}$, the estimate of Γ without the rth case, which has been ignored by the author.

Specific structure of Γ_1

For longitudinal data, one of the most commonly used covariance structures of Γ_1 is the *first-order autoregressive model* $\Gamma_1 = \sigma^2(\rho^{|i-j|})$, where $\sigma^2 > 0$ and $\rho > 0$ (see, for example, Diggle *et al.* (1994), pages 56–58). When the hierarchical models have this specific covariance structure which is embedded in the uppermost block of Γ, how can we make statistical inferences on the parameters Θ, σ^2 and ρ and do statistical diagnostics on those hierarchical models? These problems are very challenging and need to be studied further.

Daniel J. Sargent (*Mayo Clinic, Rochester*)

I would like to elaborate on a comment made by the author in Section 6, where he states that the reformulation of Section 2 has uses to speed computing. By expressing hierarchical linear models as in equation (2.15), multivariate Gibbs sampling based on the full conditional of Θ is straightforward. However, the matrix calculations based on this full conditional may be computationally expensive in large data sets. An alternative is to use the full conditional for Θ as the candidate distribution for a multivariate Metropolis–Hastings algorithm, with infrequent updating of the candidate distribution. Sargent *et al.* (1998) explore this and other issues in the setting of models with both linear and non-linear mean structures, using the reformulation of Section 2 to suggest appropriate candidate distributions for Metropolis–Hastings algorithms. In the non-linear setting constraint and prior cases typically exist, but the data cases must be manufactured. Substantial gains in computing speed are observed in both linear and non-linear models.

Joe Whittaker (*Lancaster University*)

Hodges uses a neat trick of reformulating the underlying additive model to derive added variable plots for hierarchical models. The key to the paper is the rewriting of the model as $\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E}$ at equation (2.7) where the elements of $\mathbf{Y}$ are all known. Projecting this onto the space orthogonal to $\mathbf{X}$ allows expressions for $\mathbf{E}$, from which he derives the diagnostics. I would note that although the title suggests that the trick is only appropriate to hierarchical models it also works more generally with, for instance, additive structures, such as cross-classifications. Also, although equation (2.7) has superficially the form of a linear model, it is neither the fixed effect model, because $\mathrm{var}(\mathbf{Y}) \neq \mathrm{var}(\mathbf{E})$, nor is it the random effect model, because $\mathrm{cov}(\Theta, \mathbf{E}) \neq 0$. It is just an accounting identity.

The **author** replied later, in writing, as follows.

Introductory apologies

Perhaps I should have said (though by now it is obvious) that I am not an expert on diagnostics. I use them and I am especially interested in a particular class of models. I am thus grateful to learn, among other things, that Atkinson's added variable is preferable to Andrews's. I am also not a Markov chain Monte Carlo (MCMC) expert; after further computing, I am convinced that my MCMC results in Section 5.2 would have looked like Wakefield's, if I had run it longer: *mea culpa.*

Christensen and Whittaker are right: the reformulation cannot be interpreted as a linear model in the usual sense. I like both of their terms, 'computing device' and 'accounting identity'.

Atkinson asks, 'Where's the geometry?'. He is right; other work arising from the reformulation is more explicitly geometric than the present paper. None-the-less, I appeal to the truism (due to Hilbert, I think) that algebra is geometry rendered in equations, and geometry is algebra rendered in pictures; in this sense, the current paper is full of geometry.

I certainly would have cited Lee and Nelder (1996) if I had known of it, and I apologize for insufficient diligence. Gelman and Price's example is intriguing; I would like to have tried my methods on their data set, but I have not had the time. Finally, to Critchley: I would bet that this approach is relevant to neural nets, but my comprehensive ignorance of them precludes comment.

Wakefield and Davison reanalyses

Wakefield's Fig. 11 is quite interesting—why is New Jersey (NJ) so outstanding? (Is it because it is moderately extreme and has a relatively large sample size?) Fig. 12 is also interesting but I would expect

a clearer signal if the effect of the other regressors could be partialed out as in the usual added variable plot. Wakefield's Fig. 13 is intriguing because it gives at best tepid confirmation of the long upper tail in my Fig. 6—I wonder whether this upper tail is an artefact of using $\hat{\Gamma}$ to construct my Fig. 6. In Wakefield's Fig. 14, Washington (WA) is unexceptional, whereas in Davison's analysis it is key! Washington only has seven plans; it seems unlikely that the outlier is swamped by the other six plans. Although I like Fig. 14, its message is obscured somewhat because of the long upper tails on several states with few plans (ID, ME, PR, NH and UT). For this plot, perhaps it would be better to have the vertical axis on the log-scale (see Wakefield's Fig. 16), or to treat the within-state precisions or variances as draws from some distribution, i.e. to add a hierarchical layer. Finally, Wakefield's Fig. 16 shows a bimodal distribution like the distribution shown analytically by O'Hagan (1985) in a simpler problem. This should be sobering to those favouring maximization point estimates (Nelder and Pan) or flat priors (Leonard). Probably any point estimate of τ^2 is misleading here: *mea culpa* again.

'These are a few of my favourite things'

Several discussants asked about particular diagnostics. Any diagnostic requiring refitting for each point will not adapt straightforwardly (Nelder, Draper and perhaps Pan) because of the need to re-run the MCMC fitter. Fitting by maximization methods, which avoids this problem, has inherent disadvantages, as noted above. Another possible way to avoid this problem would be to use t-errors instead of normal errors, as suggested by Wakefield, thereby obtaining robust estimates of unknown parameters in Γ_1 and Γ_2. One could then fix Γ at these point estimates and proceed as if the hierarchical model were a single-level model with known error covariance. (But, in cases where bimodality is present, this could be misleading also.) Similarly, the usual measure of leverage should adapt straightforwardly (Fung and Gu), but it would depend on Γ, and so it might be preferable to use such a robust point estimate of Γ.

Fitting a model with t-errors may address other issues as well. For example, this seems preferable to using a perturbation scheme for variances (Critchley, if I understand his comment). It also seems to address masking and/or swamping (Pan) better than any imaginable outlier test.

Wakefield's use of t-errors also suggests a way to introduce prior information into diagnostics (Longford). For example, if we think that Guam and Puerto Rico are more probably outliers than the other states, then the prior for their λ_i could capture this by using a smaller ν than the priors for the other states. Similarly, if Wakefield's Fig. 14 were constructed by modelling the within-state variances hierarchically, we could permit higher variances in the draws of σ_i^2 for states i that are thought likely to have different variances.

Regarding Draper's suggestion, it should be possible to work directly with $\mathbf{Y}$ and $\mathbf{X}$, in contrast with the difficulties described in Section 4.5. It would probably be advisable to keep a separate predictive summary for each level of the hierarchy. The single-parameter Cook distance and Draper's predictive measures address distinct concerns—Cook asks how much parameter estimates change with case deletion, whereas the predictive diagnostic identifies discrepant cases—and thus their roles seem clearly differentiated.

Regarding case influence on Γ (Pan), I would begin with simple case deletion diagnostics adapted from linear regressions and see what happens. Contrary to Pan's assertion, Section 4.4 suggests that using the full data estimate of Γ is neither uniformly benign nor disastrous. At least I feel that I have some understanding of when this diagnostic works, in contrast with a solution based on the largely unexamined maximum likelihood estimate or restricted maximum likelihood.

The methods given here should handle Pan's special structure straightforwardly, but I have not tried them so I cannot say more.

Finally, Fung and Gu ask about more flexible transformation diagnostics. It should be straightforward to adapt Cook and Weisberg (1994) but, because the fitted values are shrunk, its graphic will suffer from the same problems as the usual residual plot (Section 4.5). To use this graphic, it might be better to follow Hilden-Minton's advice for residuals, bearing in mind the dangers noted in Section 4.5. He and Shen's (1997) method is more difficult to adapt. It is probably a bad idea to use the spline basis for all of $\mathbf{Y}$, but it is not clear how to use it on $\mathbf{y}$ only. Perhaps it would work better to apply this method to the random effects formulation (3.1). It may even be possible to ignore the random effects in using this diagnostic, although I do not know how the resulting correlated error term will affect the canonical correlation.

Philosophy of diagnostics

I agree that some Bayesian computations can be useful for diagnostics (Wakefield), but for most of

those which I have seen I stick to my comments in Section 3.2.2 (*contra* Gelman and Price). Davison's approach seems quite appealing in that it allows you to have an idiotic prior but not to damn the sampling model because of it.

Contrary impressions notwithstanding, I am not a believer in the Bayes–frequentist dichotomy (Nelder). But you cannot grind every axe in every paper! I ground this axe a little in Hodges (1996).

Substantive focus

As a final word, I would like to echo Wakefield and Critchley on the importance of using the substantive problem integrally in any diagnostic exercise. I argued in Hodges (1996) that we should not focus on the effect of variant model specifications on intermediate quantities like parameter estimates, but rather on their effect on the answer to the substantive question under study. By my own standard, then, the present paper advocates further sin, but, as Wakefield suggests, such sinful tools can be used in the service of virtue, if one maintains an appropriate attitude.

References in the discussion

Andrews, D. F. (1971) A note on the selection of data transformations. *Biometrika*, **58**, 249–254.

Atkinson, A. C. (1985) *Plots, Transformations, and Regression*. Oxford: Clarendon.

———(1986) Masking unmasked. *Biometrika*, **73**, 533–541.

———(1994) Fast very robust methods for the detection of multiple outliers. *J. Am. Statist. Ass.*, **89**, 1329–1339.

Atkinson, A. C. and Riani, M. (1997) Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis. *Environmetrics*, **8**, 583–602.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics*. London: Chapman and Hall.

Barnett, V. and Lewis, T. (1984) *Outliers in Statistical Data*. New York: Wiley.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996) A new perspective on priors for generalized linear models. *J. Am. Statist. Ass.*, **91**, 1450–1460.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Carlin, B. P. and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Christensen, R. (1996) *Plane Answers to Complex Questions: the Theory of Linear Models*, 2nd edn. New York: Springer.

Christensen, R., Pearson, L. M. and Johnson, W. O. (1992) Case-deletion diagnostics for mixed models. *Technometrics*, **34**, 38–45.

Collett, D. (1991) *Modelling Binary Data*. London: Chapman and Hall.

Cook, R. D. (1977) Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.

———(1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.

Cook, R. D. and Weisberg, S. (1994) Transforming a response variable for linearity. *Biometrika*, **81**, 731–737.

Cox, D. R. (1998) Components of variance: a miscellany. *Statist. Meth. Med. Res.*, to be published.

Cox, D. R. and Solomon, P. J. (1986) Analysis of variability in large numbers of small samples. *Biometrika*, **73**, 543–554.

———(1988) On testing for serial correlation in large numbers of small samples. *Biometrika*, **75**, 145–148.

Dempster, A. P., Rubin, D. B. and Tsutakawa, R. K. (1981) Estimation in covariance components models. *J. Am. Statist. Ass.*, **76**, 341–353.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Draper, D. (1996) Discussion on Posterior predictive assessment of model fitness via realized discrepancies (by A. Gelman, X.-L. Meng and H. S. Stern). *Statist. Sin.*, **6**, 733–807.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parameterizations for normal linear mixed models. *Biometrika*, **82**, 479–488.

———(1996) Efficient parameterizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 165–180. Oxford: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A., Meng, X.-L. and Stern, H. S. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sin.*, **6**, 733–807.

George, E. I., Makov, U. E. and Smith, A. F. M. (1994) Hierarchical analysis for exponential families via Monte Carlo simulation. In *Aspects of Uncertainty* (eds P. R. Freeman and A. F. M. Smith), pp. 181–199. New York: Wiley.

Hadi, A. S. (1992) Identifying multiple outliers in multivariate data. *J. R. Statist. Soc.* B, **54**, 761–771.

Harvey, A. C. and Koopman, S. J. (1992) Diagnostic checking of unobserved components time series models. *J. Bus. Econ. Statist.*, **10**, 377–389.

He, X. and Shen, L. (1997) Linear regression after spline transformation. *Biometrika*, **84**, 474–481.

Hodges, J. S. (1996) Statistical practice as argumentation: a sketch of a theory of applied statistics. In *Modeling and Prediction Honoring Seymour Geisser* (eds J. C. Lee, A. Zellner and W. O. Johnson), pp. 19–45. New York: Springer.

de Jong, P. and Penzer, J. (1998) Diagnosing shocks in time series. Submitted to *J. Am. Statist. Ass.*

Langford, I. H. and Lewis, T. (1998) Outliers in multilevel data. *J. R. Statist. Soc.* A, **161**, 121–160.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.

Leonard, T. (1976) Some alternative approaches to multiparameter estimation. *Biometrika*, **63**, 69–75.

Leonard, T. and Hsu, J. S. J. (1992) Bayesian inference for a covariance matrix. *Ann. Statist.*, **20**, 1669–1696.

——— (1994) The Bayesian analysis of categorical data—a selective review. In *Aspects of Uncertainty* (eds P. R. Freeman and A. F. M. Smith), pp. 283–310. New York: Wiley.

Leonard, T. and Novick, M. R. (1986) Bayesian full rank marginalization for two-way contingency tables. *J. Educ. Statist.*, **11**, 35–56.

O'Hagan, A. (1985) Shoulders in hierarchical models. In *Bayesian Statistics 2* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 697–710. Amsterdam: North-Holland.

Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.

Preisser, J. S. and Qaqish, B. F. (1996) Deletion diagnostics for generalised estimating equations. *Biometrika*, **83**, 551–562.

Price, P. N., Nero, A. V. and Gelman, A. (1996) Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Hlth Phys.*, **71**, 922–936.

Racine-Poon, A. (1985) A Bayesian approach to nonlinear random effects models. *Biometrics*, **41**, 1015–1024.

Riani, M. and Atkinson, A. C. (1998) A unified approach to multivariate transformations and multiple outliers. To be published.

Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B, **59**, 291–317.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Sargent, D. J., Hodges, J. S. and Carlin, B. P. (1997) Structured Markov chain Monte Carlo. *Research Report 98-002*. Division of Biostatistics, University of Minnesota, Minneapolis.

Shephard, N. (1993) Maximum likelihood estimation of regression models with stochastic trend components. *J. Am. Statist. Ass.*, **88**, 590–595.

Shi, P. and Fung, W. K. (1998) A note on transforming a response variable for linearity. *Biometrika*, to be published.

Solomon, P. J. (1985) Transformations for components of variance and covariance. *Biometrika*, **72**, 233–239.

Solomon, P. J. and Cox, D. R. (1992) Nonlinear component of variance models. *Biometrika*, **79**, 1–11.

Strawderman, W. E. (1971) Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.*, **42**, 385–388.

Sun, L., Hsu, J. S. J., Guttman, I. and Leonard, T. (1996) Bayesian methods for variance components. *J. Am. Statist. Ass.*, **91**, 743–752.

Wakefield, J. (1996) The Bayesian analysis of population pharmacokinetic models. *J. Am. Statist. Ass.*, **91**, 62–75.

Wakefield, J. C., Aarons, L. and Racine-Poon, A. (1998) The Bayesian approach to population pharmacokinetic/pharmacodynamic models. In *Case Studies in Bayesian Statistics* (eds B. P. Carlin, A. L. Carriquiry, C. Gatsonis, A. Gelman, R. E. Kass, I. Verdinelli and M. West). New York: Springer.

Wakefield, J. C. and Bennett, J. E. (1996) The Bayesian modelling of covariates for population pharmacokinetic models. *J. Am. Statist. Ass.*, **91**, 917–927.