Residuals and Outliers in Bayesian Random Effects Models

Robert E. Weiss*
Department of Biostatistics
UCLA School of Public Health
Los Angeles CA 90024-1772 U.S.A.
rob@sunlab.ph.ucla.edu

August 11, 1994

Abstract

Common repeated measures random effects models contain two random components, a random person effect and time-varying errors. An observation can be an outlier due to either an extreme person effect or an extreme time varying error. Outlier statistics are presented that can distinguish between these types of outliers. For each person there is one statistic per observation, plus one statistic per random effect. Methodology is developed to reduce the explosion of statistics to two summary outlier statistics per person; one for the random effects and one for the time varying errors. If either of these screening statistics are large, then individual statistics for each observation or random effect can be inspected. Multivariate, targeted outlier statistics and goodness-of-fit tests are also developed. Distribution theory is given, along with some geometric intuition.

Key Words: Bayesian Data Analysis, Goodness-of-Fit, Hierarchical Models, Observed Errors, Repeated Measures.

1 Introduction.

Residual analysis in various forms is perhaps the only method for internally checking a particular model-data combination for lack of fit. Residual analysis can provide clues to useful and needed model elaboration but is only beginning

*This work was supported by grant #GM50011-01 from NIGM. The author thanks Kathryn Chaloner and Sandy Weisberg for helpful feedback.

to be developed for Bayesian hierarchical models. Some recent work on residual and outlier analysis is Albert and Chib (1993a, 1993b, 1994), Chaloner (1994), and Weiss and Lazaro (1992) and Weiss (1994b). Here I define a residual as an unobserved random error; a residual estimate is a point estimate of the unobserved random error; usually the estimate is suitable for plotting. In contrast, an outlier may be a particular observation whose unobserved error is far from the expected value (an extreme observation), or it may be a case that does not belong to the population under study, regardless of whether it is an extreme observation. In this paper, I use the first definition of outlier and define an outlier statistic as a measure of how outlying a particular unobserved error is. Plots of repeated measures data and residuals often lead to the informal identification of certain cases as outliers (Weiss and Lazaro 1992; Weiss 1994b). It is desirable to have formal methods for identifying outliers as well.

This paper develops a Bayesian approach to residual and outlier statistics for the random effects model, a common model often used to analyze continuous repeated measures data. This paper extends the approach of Chaloner and Brant (1988) and Chaloner (1991, 1994) to the general hierarchical model, with primary application to random effects models (REM) for repeated measures (RM) data. The essential idea is to have residual estimates for plotting and informal identification of outliers and outlier statistics for formal identification and labeling of outliers.

Chaloner and Brant (1988) and Chaloner (1991) developed Bayesian outlier statistics for the linear model and for models with censored data based on the posterior distribution of the unobserved error terms. The residual estimates are posterior means of unknown random errors. For the linear model $Y = X\beta + \epsilon$, where Y is an n-vector of observations, X is an $n \times p$ matrix of predictors, usually with an initial column of ones; β is a p vector of unknown regression coefficients, and $\epsilon \sim N(0, \sigma^2 I)$. Elements of Y, X, and ϵ corresponding to a single case are denoted by y_i, x_i^t, $1 \times p$ and ϵ_i respectively; all are from the i^{th} row of the respective vector or matrix. The posterior probability that ϵ_i is large, $O(i,k)$ for a fixed constant k is defined as

$$O(i,k) = P(|\epsilon_i| > k\sigma)|Y).$$

This is the posterior probability that the unobserved residual is more than k standard deviations from its prior mean of zero. This is called a Posterior Outlier Statistic (POS); in the sequel POS is also used generically to identify a statistic similar to $O(i,k)$. The constant k may be chosen to be a familiar number such as 2 or 3. Alternately it can be chosen so the prior expected number of outliers is a small constant α. For example, for $\alpha = .05$, choose $k = \Phi^{-1}(0.5+.5(0.95)^{1/n})$, where $\Phi(a)$ is the standard normal cumulative distribution function, and n is the sample size (Chaloner and Brant 1988, Chaloner 1994). Albert and Chib (1993a, 1993b, 1994b) have considered the use of these measures in binary data models and in time series models.

Section 2 presents the normal theory repeated measures random effects model and gives a short review of residuals and outlier detection for this model. Section 3 gives general distribution theory and some geometric intuition about residual summaries. Section 4 develops basic outlier statistics for these models and section 5 develops several alternative special purpose outlier statistics not previously seen in either Bayesian or frequentist examples. Section 6 gives two examples, and the paper closes with discussion.

2 The Repeated Measures Random Effects Model

2.1 The Model and Notation

The basic RM REM is

$$\begin{aligned} Y_i &= X_i\alpha + Z_i\beta_i + \epsilon_i \\ \beta_i &\sim N(0, \sigma^2 D), \\ \epsilon_i &\sim N(0, \sigma^2 I), \end{aligned} \tag{1}$$

for $i = 1, ..n$; where $Y_i = (y_{i1}, \ldots, y_{in_i})^t$ is the n_i by 1 vector of repeated measurements on subject i taken at times $t_i = (t_{i1}, \ldots, t_{in_i})^t$; X_i, n_i by p, and Z_i, n_i by q are fixed vectors of covariates; α is a p by 1 parameter vector of fixed effects, β_i is a q by 1 parameter vector of random effects with prior mean 0 and prior covariance matrix $\sigma^2 D$, and σ^2 is the sampling variance. Very commonly, Z_i has columns which are polynomials in t_i, $q < p$, and X_i has a sub matrix of q columns equal to Z_i. The entire vector Y_i will be referred to as a subject, while individual elements y_{ij} will be called an observation. The total number of observations is $N = \sum n_i$. The prior considered here is a flat prior except possibly for D with $p(D, \sigma^2, \alpha) \propto p(D)$ which leads to a proper posterior if $p(D)$ is proper. The particular choice of prior does not really matter for the material presented here. Additional notation will be useful; let x_{ij}^t be the j^{th} row of X_i, z_{ij}^t be the j^{th} row of Z_i.

The marginal distribution of Y_i for model 1 is

$$Y_i | \alpha, \sigma, D \sim N(X_i\alpha, V_i) \tag{2}$$

where

$$V_i = \sigma^2(I + Z_i D Z_i^t)$$

The Gibbs sampler (Gelfand, Hills, Racine-Poon, and Smith 1990, Zeger and Karim 1991) permits straightforward Markov Chain Monte Carlo sampling from the posterior of the parameters (θ) given the data. The discussion about computation assumes that samples $\theta^{(\ell)}, \ell = 1..L$ of some kind are available from the posterior of the parameters.

2.2 Review of Residuals and Outliers in Random Effects Models.

In the remainder of this paper I take a fully Bayes approach, but this section reviews both Bayesian and empirical Bayesian methodology for outlier detection.

There has been a limited amount of methodology developed for RM REM outlier detection from an empirical Bayes point of view. Dempster and Ryan (1985) and Lange and Ryan (1989), discuss methods for checking the normality of the random effects distribution. Their plots can be used informally for detection of outlying random effects. Waternaux, Laird and Ware (1988), define empirical Bayes estimates of the random effects as residuals and plot them to check for outliers. Lange, Little and Taylor (1989) also plot the random effects, and they and Little (1988) and Louis (1988) present frequentist statistics for checking for outliers. Louis (1992) has called for the extension of Chaloner and Brant's (1988) and Chaloner's (1991) work on residuals to the hierarchical model. Weiss and Lazaro (1992) define residuals

$$\hat{E}_i = Y_i - X_i \hat{\alpha}_i - Z_i \hat{\beta}_i,$$

where the hatted quantities $\hat{\alpha}$, and $\hat{\beta}_i$ are empirical Bayes estimates of the parameters (see Laird and Ware 1982; or Weiss and Lazaro 1992) from maximizing the likelihood resulting from (2). They plot the E_i in an empirical Bayes residual plot for repeated measures data. The plot consists of points $\hat{E}_{ij}$ vs. t_{ij} for all i, using line segments to connect consecutive time points within a case. Points for differing cases are not connected by lines. This plot can show outliers, missing terms in the linear predictor $X_i\alpha + Z_i\beta_i$, and lack of fit (Weiss and Lazaro 1992).

Weiss (1994b) indicates that the residual plots are useful whether the residuals are calculated using a Bayes analysis substituting posterior means for empirical Bayes estimates or by ad hoc or REML (Laird and Ware 1982) estimates.

Weiss (1994b) also advocates plotting estimates $\bar{\beta}_i$ of the random effects and the fixed effects residuals $R_i = Y_i - X_i\bar{\alpha}_i$. The main purpose for plotting the R_i is to check the covariance specification, while outlier detection is one use for the plots of $\bar{\beta}_i$'s and E_i's. For Bayesian accommodation of outliers and robustification of hierarchical models, see Sharples (1990); Wakefield, Smith, Racine-Poon, and Gelfand (1994); Lange, Little, and Taylor (1989); Lange, Carlin and Gelfand (1992a, 1992b) and Seltzer (1993).

3 Distribution Theory.

This section develops distribution theory and some geometric intuition for calculating and understanding residual summaries. Relevant earlier work for the one way random effects model can be found in Hill (1965) and Chaloner (1994).

Model (1) can be written in more compact notation. Let $Y = (Y_1^t, \ldots, Y_n^t)^t$, be the vector of all responses and similarly, let $X = (X_1^t, \ldots, X_n^t)^t$ be the matrix of the coefficients of α; Let $Z = \text{diag}(Z_i)$, be an irregular block diagonal matrix N by nq of the matrices of coefficients of β_i; the blocks may be of differing sizes; let $\beta^t = (\beta_1^t, \ldots, \beta_n^t)$ be the nq vector of random effects; and let $\epsilon^t = (\epsilon_1^t, \ldots, \epsilon_n^t)$. Then the model (1) can be written in the form

$$\begin{aligned} Y &= X\alpha + Z\beta + \epsilon, \\ \beta &\sim N_{nq}(0, \sigma^2 I \otimes D), \\ \epsilon &\sim N_N(0, \sigma^2 I). \end{aligned} \tag{3}$$

I use a prior $p(\alpha, \beta, \sigma^2, D) \propto p(D)$, which will lead to a proper posterior if $p(D)$ is chosen properly. Poor choice of prior leads to an improper posterior, which in turn can lead to convergence problems with Gibbs sampling procedures. The joint posterior can be written as

$$p(\alpha, \beta | \sigma^2, D, Y) p(\sigma^2 | D, Y) p(D|Y).$$

The following matrices are used in the conditional posterior means of α and β and in the associated discussion.

$$\begin{aligned} H_X &= X(X^tX)^{-1}X^t \ , \\ Q_X &= I - H_X \ , \\ H_Z &= Z(Z^tZ + I \otimes D^{-1})^{-1}Z^t \ , \\ Q_Z &= I - H_Z \ , \end{aligned}$$

$$\begin{aligned} P_Z &= Z(Z^tQ_XZ + I \otimes D^{-1})^{-1}Z^tQ_X \text{ , and} \\ P_X &= X(X^tQ_ZX)^{-1}X^tQ_Z \ . \end{aligned}$$

Define the quadratic form $Q(Y) = .5Y^tQ_X(I - P_Z)Y$. Then

$$\left(\begin{array}{c} \alpha \\ \beta \end{array} \right) \Bigg| D, \sigma^2, Y \sim N \left(\left(\begin{array}{c} \bar{\alpha}(D) \\ \bar{\beta}(D) \end{array} \right), \left(\begin{array}{cc} V_{\alpha\alpha} & V_{\alpha\beta} \\ V_{\alpha\beta}^t & V_{\beta\beta} \end{array} \right) \right),$$

$$p(\sigma^2|D, Y) = \frac{Q(Y)^{\frac{N-p-2}{2}}}{(\sigma^2)^{\frac{N-p}{2}}\Gamma((N-p-2)/2)} \exp\{-Q(Y)/(\sigma^2)\} \ ,$$

where $\Gamma(a)$ is the usual gamma function, and

$$\begin{aligned} & p(D|Y) \propto \\ & \quad |X^tX|^{-1/2}|D|^{-n/2}|(Z^tQ_XZ + I \otimes D^{-1})|^{-1/2} \\ & \quad *\Gamma\{(N-p-2)/2\}Q(Y)^{-(N-p-2)/2}p(D) \ . \end{aligned}$$

The location parameters α and β are conditionally normal given D and σ^2 with conditional means that can be written in either of two useful forms

$$\begin{aligned} \bar{\alpha}(D) &= (X^tQ_ZX)^{-1}X^tQ_ZY & (4) \\ &= (X^tX)^{-1}X^t(I - P_Z)Y & (5) \\ \bar{\beta}(D) &= (Z^tZ + I \otimes D^{-1})^{-1}Z^t(I - P_X)Y & (6) \\ &= (Z^tQ_XZ + I \otimes D^{-1})^{-1}Z^tQ_XY & (7) \end{aligned}$$

and conditional variances and covariance

$$\begin{aligned} V_{\alpha\alpha} &= \sigma^2(X^tQ_ZX)^{-1} \ , \\ V_{\beta\beta} &= \sigma^2(Z^tQ_XZ + I \otimes D^{-1})^{-1} \ , \\ V_{\alpha\beta} &= -(X^tX)^{-1}X^tZV_{\beta\beta} \ . \end{aligned}$$

The scale parameter σ^2 is conditionally inverse Γ with $N-p$ degrees of freedom and $Q(Y)$ shape parameter. The posterior of D is not of convenient form for sampling. Inspection of $p(D|Y)$ suggests that using a prior of the form $p(D) \propto |D|^{-c}$ will lead to an improper posterior if the function x^{-c} is not integrable over intervals $(0, a)$, $a > 0$.

The discussion for the remainder of this section only is conditional on D. The matrices H_X, and Q_X are projection matrices, for example H_XY is the projection of Y onto the column space of X. On the other hand, H_Z, Q_Z, P_Z, and P_X are submatrices of projection matrices, but this paragraph proceeds as if, for example, H_Z is the projection onto the column space of Z: this is inaccurate, but the relationship is made more clear in the next paragraph. Formulae (4) - (7) are the random effects model generalization of the usual regression formulae for coefficient estimates with two sets of predictors X_1 and X_2. In particular, the posterior mean of α (as a function of D) can be thought of in several ways. From (4), $\bar{\alpha}(D)$ are the estimated regression coefficients from the weighted regression of Y on X, with weights the V_i matrices from equation (2) and $\bar{\alpha}(D)$ are the coefficients from the regression of Y on X adjusted for Z. From (5), $\bar{\alpha}(D)$ is also the coefficients from the unweighted regression of $(I - P_Z)Y$ on X. The estimate $\bar{\beta}(D)$ can be thought of as the regression of Y on Z adjusted for X (7), and shrunk towards zero by the prior and it is the shrunken regression of $(I - P_X)Y$ on Z (6). Similarly, the variance matrix $V_{\alpha\alpha}$ is the usual regression variance of α adjusted for Z, and $\bar{\beta}(D)$ has variance $V_{\beta\beta}$ which is from the regression of Y on Z adjusted for X and is smaller because of the information supplied by the prior.

The matrices H_X, Q_X, H_Z, Q_Z, P_Z and P_X are all principle N by N submatrices of projection matrices of order $N + nq$. Define the partitioned matrices $Z^{*t} = (Z^t | I \otimes D^{-1/2})^t$, $Y^{*t} = (Y^t | 0^t)$ and $X^{*t} = (X^t | 0^t)$, where Z^*, X^*, and Y^* all have the same length $N + nq$, and each has the same number of columns as Z, X, and Y, respectively. One wayto think of the $*$ space is that the first N coordinates correspond to observed data, and the last nq coordinates correspond to prior information or prior observations. In the augmented $*$ space, we can consider projection onto the space spanned by the columns of $X*$, and projections onto the column space of Z^*, and projections onto the orthogonal spaces and projections onto the column space adjusting for the other matrix. Define $e^*(a^*, b^*) = (I - b^*(b^{*t}b^*)^{-1}b^{*t})a^*$ as the residuals from the regression of a^* on b^*, where a^* and b^* are appropriately sized matrices and a^* may be a matrix. Then $\bar{\alpha}$ are the fitted coefficients from regressing $e^*(Y^*, Z^*)$ on $e^*(X^*, Z^*)$, and $\bar{\beta}$ are the fitted coefficients from regressing $e^*(Y^*, X^*)$ on $e^*(Z^*, X^*)$. In other words, in the $*$ problem, the coefficient estimates $\bar{\alpha}$ and $\bar{\beta}$ are properly adjusted for each other, in exactly the same fashion that regression coefficients in linear regression are adjusted for the other predictors in the model.

Let $E = E(D) = Y - X\bar{\alpha}(D) - Z\bar{\beta}$ be the conditional vector of residuals. If D were known, then these would be used for plotting and the visual identification of outliers. As it is, the E to be used for plotting are the averages of the $E(D)$ over the distribution of D. In the $*$ problem, define $E^* = Y^* - X^*\bar{\alpha} - Z^*\bar{\beta}$. Then $E^{*t} = (E^t | (I \otimes D^{-1/2})\bar{\beta})$, and the E^* residuals includes not just the $E(D) = Y - X\bar{\alpha}(D) - Z\bar{\beta} = (I - P_X - P_Z)Y$ but also the random effects estimates $\bar{\beta}$ appropriately standardized by the prior standard deviations $(I \otimes D^{-1/2})$. This justifies calling the random effects residuals by Waternaux, Laird and Ware (1989) and missing data in Laird, Lange, and Stram (1987). Here the transformed random effects estimates are both residuals and individual regression coefficients and they are prior observations. The E^* residuals are the projections of the Y^* data onto the orthogonal complement of the X^* and Z^* spaces. They are a priori mean zero and could also be used for plotting. Weiss (1994b) preferred to plot the $\bar{\beta}$'s unstandardized and separately from the E residuals; the untransformed $\bar{\beta}$'s are easier to interpret.

The unobserved residuals ϵ are conditionally distributed

$$\epsilon | D, \sigma^2, Y \sim N(E, \sigma^2 M)$$

where M is

$$M = X(X^t Q_Z X)^{-1} X^t + Q_X P_Z - H_X Z (Z^t Q_X Z + I \otimes D^{-1})^{-1} Z^t H_X.$$

The covariance of ϵ and β is $Q_X Z V_{\beta\beta}$, which will mostly be large only for ϵ's and β's from the same case, since Q_X will be close to the identity, and Z and $V_{\beta\beta}$ will be block diagonal and nearly block diagonal respectively. Details and special cases of the distributions and projection matrices will be worked out in a future paper. Integrating out σ^2 above gives a t distribution for $\epsilon|D$, however, this does not help in computing the outlier statistics of the next section, since due to the nature of the outlier calculation, the cutoff depends on σ. Finally, it is probably not a good idea to condition upon D in assessing the uncertainty in the residuals: D is inherently uncertain, in the sense that the properness of its posterior depends on the chosen prior.

4 Outlier Statistics

The POS provide a method for formal identification of outliers. There are several methods for extending these residuals to multivariate data and to hierarchical models. For general multivariate data $Y_i \sim N(\mu_i, V_i)$, where (2) is a special case, we can consider either of the statistics

$$M_1(i, j, k) = P(|y_{ij} - \mu_{ij}| > k v_{ijj}^{1/2} | Y)$$

or

$$M_2(i, k) = P((Y_i - \mu_i)^t V_i^{-1} (Y_i - \mu_i) > k | Y)$$

where $\mu_i = (\mu_{ij})$. The outlier statistic $M_1(i,j,k)$ treats each observation univariately. The kernel $(Y_i - \mu_i)^t V_i^{-1}(Y_i - \mu_i)$ in the M_2 statistic is a Bayesian analog of the empirical Bayes outlier statistics of Little (1988) and Lange, Little and Taylor (1989); M_2 checks the whole case for joint outlyingness along the elliptical contours of the sampling distribution. One could also consider, similar to Louis (1988), rotating to the principle coordinates $W_i = \Gamma_i Y_i$, where $\Gamma_i V_i \Gamma_i^T = \Lambda_i$, Λ_i is diagonal with diagonal elements λ_{ij} the eigenvalues of V_i so that W_i is the principle component decomposition of Y_i, and then evaluating $P(|W_{ij} - \Gamma_{ij}^t \mu_i| > k\lambda_{ij}^{1/2}$, where W_{ij} is the j^{th} element of W_i and Γ_{ij}^t is the j^{th} row of Γ. This particular decomposition will not be followed further in this paper for two reasons: the problem of dealing with a parameterized covariance matrix, which is solvable; and two, the procedure introduced in the next paragraph provides a readily interpretable set of statistics. The value k is used generically as the cutoff for these and other statistics, however different values of k should be used for both $M_1(i,j,k)$ and $M_2(i,k)$ and elsewhere. It can be useful to entertain several values of k for the same statistic.

For hierarchical models, such as (1) the hierarchical structure suggests several outlier statistics be considered. For model (1), one could compute the posterior probability that

1. Y_i is far away from its marginal mean $X_i\alpha$,
2. Y_i is far away from $X_i\alpha + Z_i\beta_i$,
3. y_{ij} is far away from $x_{ij}^t\alpha$,
4. y_{ij} is far away from $x_{ij}^t\alpha + z_{ij}^t\beta_i$,
5. ϵ_i is far away from 0,
6. β_i is far away from 0,
7. ϵ_{ij} is far away from 0, and
8. β_{ij} is far away from 0.

These 8 possible computations will be called *definitions* for lack of a better term. Some simplification is immediate: 2 and 5 are identical, and 4 and 7; the different mathematics suggest different ways of thinking about the same thing. The first definition is the same as $M_2(i,k)$, and the third is $M_1(i,k)$. Four elements of the above list are univariate (3, 4, 7, 8) and four are multivariate.

When there is only one random effect, 6 and 8 are identical and for cases with $n_i = 1$, 1 and 3 are identical and 2 and 4 are identical.

Definitions 7 and 8 are simple and interpretable in the context of a random effects model. Each level of the hierarchy produces a set of outlier statistics and by analogy, a set of residuals suitable for plotting.

At the base level, we have the $\epsilon_i = Y_i - X_i\alpha - Z_i\beta_i$ as random errors which are a priori normal with mean zero, given the parameters. The posterior mean $\bar{\epsilon}_i$ of the ϵ_i is a summary of the posterior that is suitable for plotting and informal identification of outliers. The corresponding Posterior Outlier Statistics are

$$O_\epsilon(i, j, k) = P(|\epsilon_{ij}| > k\sigma|Y),$$

the posterior probability that ϵ_{ij} lies more than k prior standard deviations from its prior mean. At the second level of the hierarchy, we have the β_i, a priori mean zero with prior variance D given the parameters. The posterior mean $\bar{\beta}_i$ is suitable for plotting and informal identification of outliers. The POS for β_{ij} is the posterior probability that β_{ij} is more than k standard deviations from its prior mean of zero is

$$O_\beta(i, j, k) = P(|\beta_{ij}| > k\sigma d_{jj}^{1/2}|Y),$$

where d_{jj} is the j^{th} diagonal element of D. At the third level of the hierarchy, the parameters (α, σ^2, D) have improper priors and no prior conditional mean, so no residual exists for plotting purposes and no POS. When interest lies specifically in high and low outliers as might happen in the areas of chemometrics, quality control, or human research, the absolute values in the definitions of $O_\epsilon(i, j, k)$ and $O_\beta(i, j, k)$ can be replaced by a $+$ or $-$ sign as appropriate.

Posteriors of the residuals ϵ_i can be formed from samples of the posterior as

$$\epsilon_i^{(l)} = Y_i - X_i\alpha^{(l)} - Z_i\beta_z^{(l)} .$$

The posterior distributions of the β_i's are taken straight from the posterior samples. More accurate computations can be produced by integrating over some or all of α, β_i, and σ^2 and then calculating the posterior density as a resulting mixture of closed form computations. These calculations are given in section 3.

The POS's $O_\epsilon(i, j, k)$ and $O_\beta(i, j, k)$ make for easy inference and followup action. Cases with large outlier statistics can be identified in plots of $\bar{\epsilon}_{ij}$ and $\bar{\beta}_{ij}$. While there will be confounding between β_i and ϵ_i whenever $n_i \leq q$, even for n_i only somewhat greater than q, as in the examples in section 6, the POS's distinguish quite satisfactorily between ϵ and β outliers. When $O_\beta(i,j,k)$ is large, it suggests checking that the particular individual i is actually a member of the population of interest. For example, one might inspect baseline predictors for accuracy or unusualness. When $O_\epsilon(i,j,k)$ is large, it suggests first checking if the observation y_{ij} was recorded in error, or if either x_{ij} or z_{ij} are recorded in error or if the conditions during measurement were in some way aberrant. If several $O_\epsilon(i,j,k)$ are large, it additionally suggests checking whether there is something unique about individual i and again whether individual i is a member of the population of interest. When many $O_\epsilon(i,j,k)$ or $O_\beta(i,j,k)$ are large, it suggests that there may be a problem with the model; this idea suggests a goodness of fit check may be possible and is discussed further in the next section.

5 Summary, Multivariate, Specialty and Goodness of fit Outlier Statistics

The recommended outlier statistics $O_\epsilon(i,j,k)$, and $O_\beta(i,j,k)$ will produce n_i+q outlier statistics per case in a general data set, which can be a moderate burden on the data analyst. One sensible way to reduce the burden is to set the value of k high enough to ensure that only a few cases are identified as extreme. Sometimes this is undesirable, since this might miss sets of observations that have large ϵ_i and β_i but are not the most extreme. The next two subsections consider statistics to screen cases so that the individual POS's $O_\epsilon(i,j,k)$, and $O_\beta(i,j,k)$ need only be considered for a few cases. Subsection 5.1 considers the multivariate statistics resulting from outlier definitions 5 and 6. Subsection 5.2 suggests a direct summary of $O_\epsilon(i,j,k)$, and $O_\beta(i,j,k)$. Subsections 5.3 and 5.4 discuss targeted outlier statistics and general purpose goodness of fit statistics.

5.1 Multivariate Outlier Statistics

Definitions 1, 5 and 6 might be used to reduce the burden of n_i+q statistics per case. Definition 1 seems less than useful, because it ignores information available in the hierarchical model. Definitions 5 and 6 are analogs of definition $M_2(i,k)$ that treat the different levels separately. Define $\xi_{\epsilon i} = \epsilon_i^t(\sigma^2 I)^{-1}\epsilon_i$ and $\xi_{\beta i} = \beta_i^t D^{-1}\beta_i$. A priori given the parameters, $\xi_{\epsilon i}$ is a χ^2 random variable with n_i degrees of freedom, and similarly, $\xi_{\beta i}$ is a priori distributed as a $\chi^2(q)$. It makes sense to compare them a posteriori to quantiles of the chi-square distribution with n_i and q degrees of freedom. Define

$$\text{MVO}_{\epsilon i}(k) = P(\xi_{\epsilon i} > k|Y)$$

where k is chosen to be $\chi^2(1-a, n_i)$, the $100(1-a)$th percentile of the χ^2 distribution with n_i degrees of freedom, and similarly for $\text{MVO}_{\beta i}(k)$, except k would be chosen based on a chi^2 with q degrees of freedom. In parallel with the Chaloner and Brant (1988) suggestion, one might choose $a = n^{-1}.05$ so that a priori the expected probability of finding an ϵ outlier is .05 over all, and similarly for $\text{MVO}_{\beta i}(k)$.

The posterior means of the ξ_i statistics for either ϵ's or β's could be plotted in QQ-plots as in Lange, Little, and Taylor (1989), Dempster and Ryan (1985) and Lange and Ryan (1989). If the n_i are different, the posterior means could be back transformed using the appropriate χ^2 distribution and then retransformed to a convenient distribution for plotting.

The nice thing about the $O_\epsilon(i,j,k)$ and $O_\beta(i,j,k)$ outlier diagnostics introduced earlier is that they directly suggest useful actions for a data analyst to take. A disadvantage of the univariate diagnostics is that they cannot find multivariate outliers. Many readers will be familiar with the plot of data from a correlated bivariate distribution with a single observation that is not extreme on any univariate measure, but is unusual in the bivariate distribution. In particular, in the equicovariance model, an observation could be high on early time values and low on later times - suggesting a missing slope effect. Diagnostics based on $\xi_{\epsilon,i}$ can find multivariate outliers but will not perfectly agree with diagnostics based on $O_\epsilon(i,j,k)$ since they ask different questions.

5.2 Summary Outlier Statistics

A single number summary measure of the $O_\epsilon(i,j,k)$ that is directly useful for identifying cases with large $O_\epsilon(i,j,k)$ is

$$O_\epsilon(i,+,k) = \sum_{j=1}^{n_i} O_\epsilon(i,j,k).$$

This statistic might be disdained because of its lack of sophistication. On the other hand, it is a perfect screening statistic for the individual $O_\epsilon(i,j,k)$ since if it is small, none of the $O_\epsilon(i,j,k)$ can be large. It can be interpreted as the posterior expected number of ϵ outliers for case i. When there is more than one random effect, an exactly parallel construction leads to

$$O_\beta(i,+,k) = \sum_{j=1}^{q} O_\beta(i,j,k).$$

5.3 Targeted Outlier Statistics

Special purpose outlier statistics can be developed to target specific model problems. Two examples are given, although more can certainly be considered. A statistic that checks if $\psi_{i1} = (\epsilon_{i2}, \epsilon_{i4})$ is outlying is

$$O_{\psi,1}(i,k) = P(\psi_{i1}^t \psi_{i1} > k\sigma^2 | Y).$$

A priori, $O_\psi(i,k)$ is χ^2 with 2 degrees of freedom, and checks to see if $(\epsilon_{i2}, \epsilon_{i4})$ is a bivariate outlier. The constant k would be based on some quantile of a $\chi^2(2)$. Alternatively, one can consider if $\psi_2 = \epsilon_{i2} - \epsilon_{i4}$ is an outlier. That is, is the difference between the two residuals unusually large? The relevant outlier statistic becomes:

$$O_{\psi,1}(i,k) = P(\psi_{i2}^2 > 2k\sigma^2 | Y),$$

and k is taken from an appropriate quantile of a χ^2 with 1 df. The 2 on the right hand side is because $\epsilon_{2i} + \epsilon_{4i}$ are a priori independent and identically distributed (iid).

5.4 Goodness of Fit

We can calculate the a priori expected value and variance of the number of outliers in a model. Since the POS are posterior expected number of outliers, this suggests that a goodness of fit statistic is possible. Consider

$$\Omega_\epsilon(k) = \sum_{i=1}^{n} \mathrm{MVO}_{\epsilon i}(k)$$

and

$$\Omega_\beta(k) = \sum_{i=1}^{n} \mathrm{MVO}_{\beta_i}(k).$$

Assume that k is chosen to be the a probability tail quantile of the appropriate χ^2 distributions. Then a priori, $\Omega_\epsilon(k)$ and $\Omega_\beta(k)$ has a priori expectation na, and standard deviation equal to $\mathrm{SD}_\Omega = (na(1-a) + V_1 + V_2)^{1/2}$, where V_1 is a correction for posterior uncertainty, since the exact values of the ϵ_i are unknown, and V_2 is a correction for possible computation based uncertainty such as a finite sample size in a Gibbs sampling (Gelfand and Smith 1990) algorithm. The variance V_2 can usually be made arbitrarily small in Gibbs sampling. It is unclear how to compute V_2, since the individual outlier statistics $MVO_\epsilon(k)$ can be highly correlated a posteriori, and because the first term in SD_Ω is a sampling variance while the second term is due to the fact that we have posterior uncertainty. In large samples, assuming n_i large also, then V_2 will be small.

A goodness of fit statistic may not be perfectly useful in the RM REM. Unlike a graphic, it does not tell us what part of the model is the root cause for it's size. In the current examples, $\Omega_\epsilon(k)$ or $\Omega_\beta(k)$ might be large because there are a) missing fixed effects; b) missing random effects; c) non-normal ϵ or β errors; d) outliers; or e) missing predictors. Even deleting an extra useful predictor could lead to adequate goodness of fit. The presence of two goodness of fit statistics, $\Omega_\epsilon(k)$ and $\Omega_\beta(k)$ and the possibility of more targeted goodness of fit statistics in the previous section make them potentially more useful. In the weight loss example in section 6, $\Omega_\epsilon(k)$ is large while $\Omega_\beta(k)$ is small, suggesting that the plot of the ϵ_i residuals should be inspected, and not the plot of the β residuals.

5.5 The General Hierarchical Model

Lindley and Smith (1972) consider a general hierarchical model where $y|\theta_1, C_1 \sim N(A_1\theta_1, C_1)$ at the first stage; $\theta_1|\theta_2, C_2 \sim N(A_2\theta_2, C_2)$ at the second stage, and; at the third stage, $\theta_2|\theta_3, C_3 \sim N(A_3\theta_3, C_3)$. Then for each stage where the residuals are conditionally zero, there will be an outlier statistic for each element θ_{ij} of θ_i of the form $P(|\theta_{ij} - A_{ij}^t\theta_{(i+1)}| > kC_{ijj}^{1/2})$, where A_{ij}^t is the jth row of A_i, and C_{ijj} is the jth diagonal element of C_j. The ideas of this section can be used both to reduce the number of outlier statistics and to develop special purpose statistics.

6 Examples

6.1 Pediatric Pain Example

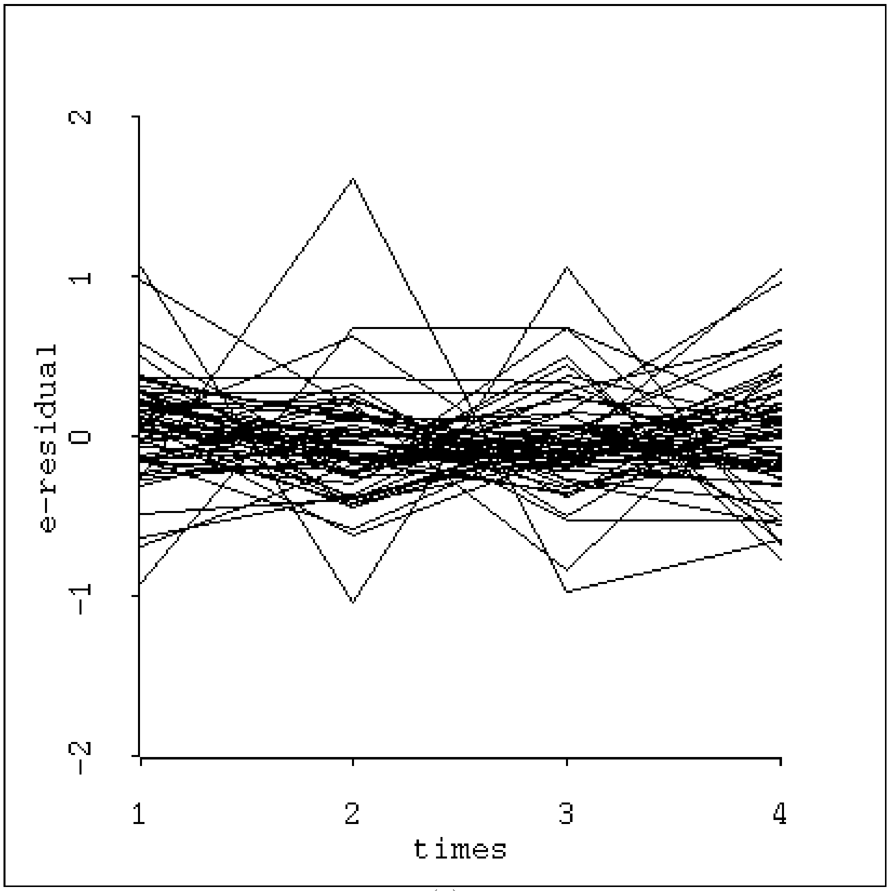
The model for this data is (1), with $q = 1$, $p = 8$, $n = 64$ and usually four observations per child. The observations are the log of the time that a child can keep his or her arm immersed voluntarily in very cold water. The X_i matrix for each child is a column of ones, a column of 0's or 1's depending on whether

id	$O_\beta(i)$	$O_\epsilon(i,1)$	$O_\epsilon(i,2)$	$O_\epsilon(i,3)$	$O_\epsilon(i,4)$
	$\bar{\beta}_i$	$\bar{\epsilon}_{i1}$	$\bar{\epsilon}_{i2}$	$\bar{\epsilon}_{i3}$	$\bar{\epsilon}_{i4}$
10	2	0	0	0	0
	(-.99)	(.09)	(-.04)	(-.17)	(-.31)
20	1108	0	0	0	0
	(1.94)	(.38)	(.38)	(.34)	(-.26)
55	488	0	0	0	0
	(1.71)	(.28)	(.28)	(.28)	(-.11)
59	29	146	2	2	0
	(1.32)	(-.91)	(.69)	(.69)	(.10)
30	0	406	366	406	17
	(.93)	(1.06)	(-1.04)	(1.06)	(-.67)
31	0	0	1918	186	15
	(-.61)	(-.25)	(1.61)	(-.97)	(-.65)
61	0	4	0	0	424
	(.27)	(-.68)	-.16)	(-.08)	(1.04)

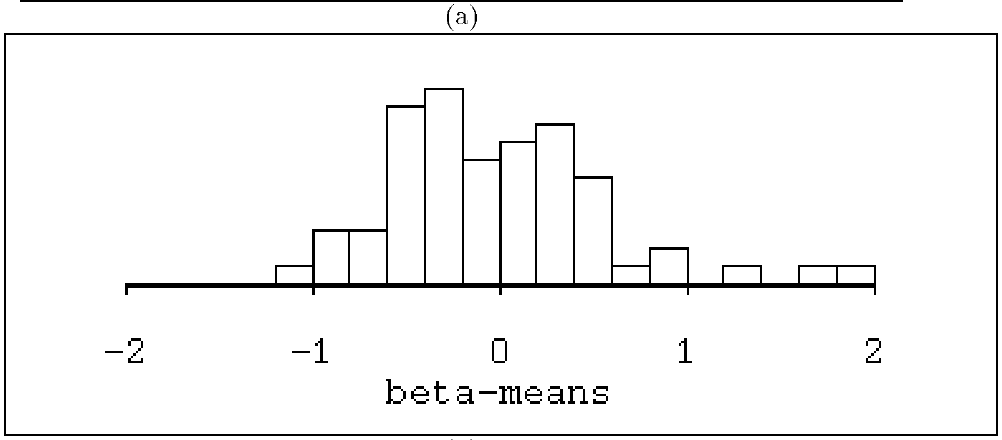
Table 1: Outlier statistics for the pediatric pain data set. Key: id = Case number; $O_\beta(i)$, is the probability that β_i is an outlier with $k = 3$; $O_\epsilon(i,1), O_\epsilon(i,2), O_\epsilon(i,3), O_\epsilon(i,4)$ are the probabilities that the ϵ_{ij}, for $j = 1, ..., 4$ in the i^{th} case are outliers with $k = 3$; the last argument was deleted. All probabilities have been multiplied by 2001, the Gibbs sample size. The second line for each case are the estimated residual means; $\bar{\beta}$ = posterior mean of random effect; $\bar{\epsilon}_{ij}$ are the posterior means of the ϵ_i's.

the child is an *attender* or *distracter*, which is the childs general style of coping with the pain of the cold water. The last 6 columns are all zero except for a single one at time 4 indicating which of three treatment groups the child was randomized to. The three treatments are three forms of counseling given prior to the last measurement. The treatment can be instruction to *attend*, *distract*, or a *null* counseling.

Figure 1(a) shows a parallel plot of the posterior means of the residuals $\bar{\epsilon}_i$ (Weiss and Lazaro 1992; Weiss 1994b), and 1(b) gives a histogram of $\bar{\beta}_i$. Both figures show a certain number of outlying observations. Table 1 gives the POS and the posterior mean residuals for the observations identified by either $O_\beta(i,k)$ and $O_\epsilon(i,j,k)$ as large, for $k = 3$. These outlier diagnostics can distinguish clearly between random effects outliers and sampling error outliers. The statistic $O_\beta(i,3)$ is in column 2, the unneeded third argument has been dropped, and the $O_\epsilon(i,j,3)$ are in the last four columns. Most statistics are 0, and are not shown. A few other cases have $O_\epsilon(i,j,3)$ greater than zero,



(a)



(b)

Figure 1: Residual plots for pain data. (a) Parallel plot of e-residuals. (b) Histogram of posterior means of random effects.

model	tail area	$\Omega_\epsilon(k)$	exp	se	$\Omega_\beta(k)$	exp	se
	.05	4.43	1.9	1.34	2.0	1.90	1.34
random intercept	.01	3.28	.38	.61	.78	.38	.61
	.00270	2.49	.10	.32	.30	.103	.32
	.00270	1.39	.10	.32	.28	.10	.32
random slope and intercept							

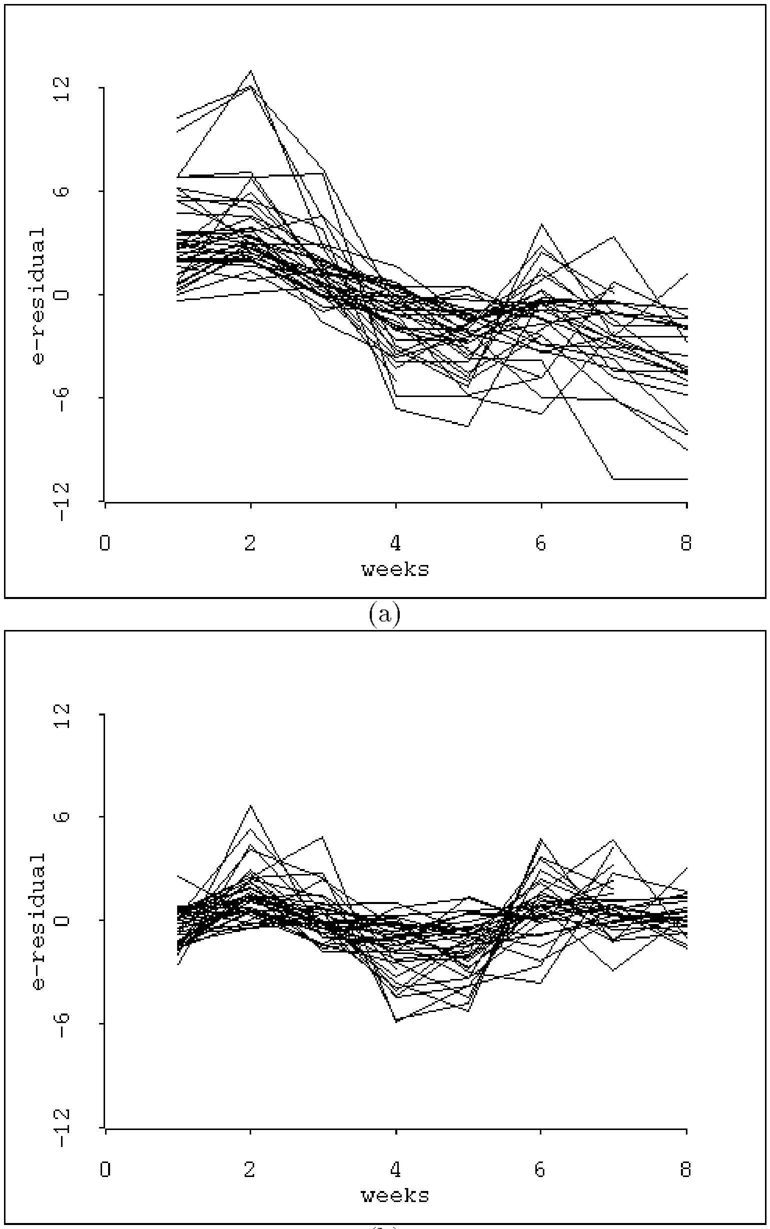
Table 2: Goodness of fit statistics for weight loss data. Tail area is the probability in the tail which k corresponds to; exp is the prior expected value for the statistic, and se is the prior standard error.

but all nonzero $O_\beta(i,3)$'s have been shown. The outlying β cases are generally different from the outlying ϵ residuals. The probability that β_{20} is more than 3 standard deviations away from 0 is .55, a clear indication that this case is an outlier. In contrast the prior probability of any outlying random effect is $64 * 2 * \Phi(-3) = .17$, 64 times twice the normal tail area beyond $+/-3$. Case 30 is an attender taught to attend that has the largest residuals in Figure 1(a); from table 1, the probability that β_{30} is an outlier is 0, but the probability that $\epsilon_{30,1}$, $\epsilon_{30,2}$, and $\epsilon_{30,3}$ are outliers for $k = 3$ is about .2 for all three observations.

6.2 Weight Loss Data

The weight loss data contains observations on 38 women enrolled in a diet study. Measurements were taken weekly, this data set was formed prior to the end of the study, and some data is missing. A plot of the raw data (Weiss 1994b) suggests a random intercept is needed. Table 2 lists goodness of fit statistics for this model for several values of k. No lack of fit is indicated for the random effects, but there are apparently excess outliers in the ϵ residuals; the goodness of fit statistics are substantially larger than expected for tail areas of .01 and .0027. Inspecting Figure 2(a) indeed suggests that a random slope is needed to properly model these data. The residual plot for the model with a random slope and intercept is in part (b) of this figure. Again some structure is apparent, indicating the need for further modeling. The goodness of fit statistic is smaller than the previous model, but still large, indicating excess very large residuals.

For the model with a random intercept and slope, Table 3 gives the outlier statistics for individual ϵ_i and β_i. Case 8 at week 2 corresponds to the largest outlier visible in figure 2(b), with a posterior probability of being greater than 3 prior standard deviations estimated at .56. Cases 5 and 35 are identified as



(b)

Figure 2: Residual plots for weight data. (a) Parallel plot of e-residuals, model with random intercept. (b) Parallel plot of e-residuals from model with random intercept and slope.

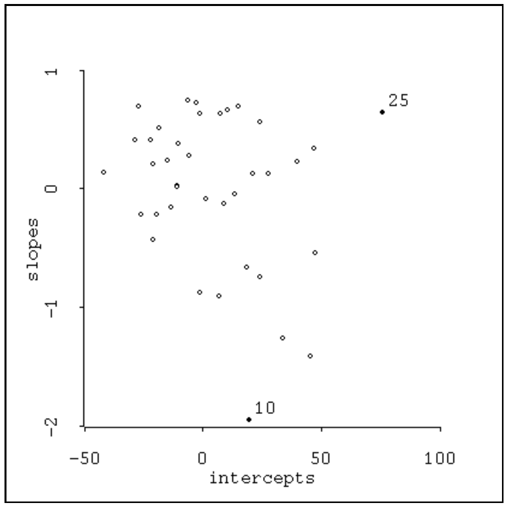


Figure 3: Weight loss data: posterior means of random effects.

having some probability of being univariate outliers, but the multivariate outlier statistic indicates that they are quite probably big outliers. There are not too many β outliers. Cases 10 and 25 are formally identified as univariate outliers. Figure 3 also shows in an informal manner that these cases are univariate outliers. No multivariate outliers are identified, either formally using the POS statistics or with the plot. Figure 3 also shows that heavier study participants tend to have greater weight loss.

7 Discussion.

A competing Bayesian outlier diagnostic is the conditional predictive ordinate (CPO)

$$\text{CPO}_i = \int f(y_i|x_i, \beta_i, \sigma) p(\beta, \sigma_i|Y_{(i)})$$

(Geisser 1980; Geisser 1993 Chap 4; Pettit and Smith 1985; Pettit 1990; Peña and Guttman 1993; Weiss 1994a). For an example of its use in hierarchical models, see Sharples (1990). Peña and Guttman (1993) prefer the CPO to the POS $O(i, k)$ because it better approximates a statistic from a complex mixture model. In contrast, POS is designed to identify cases with large residuals ϵ_i. It would be a mistake to assume that it is not useful because it does not approximate a statistic which it was not designed to imitate. POS has a simple definition and is intuitively appealing. It is bounded between zero and one, and thus is restricted to an interpretable scale. In contrast, CPO_i occurs often as a limiting component of more complicated mixture models, and while it may have a wider range of uses, it is harder to explain and, because it has a range which varies with the problem and even the particular observation, it can be harder to interpret.

All of the discussed outlier statistics have implicitly assumed that larger values indicated more outlying. It is possible that extreme values of a residual such as ϵ_i do not necessarily correspond to more outlying. A simple example occurs in the location model where the y_i are iid with multi-modal sampling density $f(y_i - /mu)$ and mean μ. Outlier identification regions other than $|\epsilon_i| > k\sigma$ could be defined. In situations like this, the outlier statistic could be altered to $P(f(y_i - \mu) < c|Y)$. This brings the POS statistic closer to CPO in conception: the POS is a posterior probability summarizing the distribution of $f(y_i - \mu_i)$, while CPO is the inverse posterior harmonic mean of $f(y_i - \mu_i)$ (confer Weiss 1994a). This connection between the POS and CPO extends to other models, including the regression location model (with known scale) but if the scale parameter σ is unknown then the POS becomes $P(f(y_i - \mu) < c\sigma|Y)$. The definition of CPO does not change, which is actually a flaw of CPO. Suppose a regression model with nonconstant variance known up to an unknown scale parameter. An observation with a variance known to be very large will almost automatically be an outlier while an observation known to have a small variance a priori will not be an outlier, assuming, of course, that the model is correct.

Other summaries of $f(y_i|\theta)$ can serve usefully as outlier statistics. For example, Aitchison and Dunsmore (1975) consider $E[f(y_i|\theta)|Y]$ as a measure of discordancy. Weiss (1994a) shows that influence measures are also posterior summaries of $f(y_i|\theta)$.

References

[1] Albert, J. H. and Chib S. (1993a). Bayesian analysis of binary and polychotomous response data. *JASA*, 88, 669-679.

[2] Albert, J. H. and Chib S. (1993b). Paper in *Journal of Business and Economic Statistics.*

[3] Albert, J. H. and Chib S. (to appear). Bayesian Modeling of Binary Repeated Measures Data with Application to Crossover Trials. To appear in *Bayesian Biostatistics*, (eds) D. A. Berry and D. K. Stangl, New York: Marcel-Dekker.

[4] Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika*, 78, 637-644.

[5] Chaloner, K. (1994). Residual analysis and Outliers in Bayesian Hierarchical Models. In *Aspects of Uncertainty*, (eds) P. R. Freeman and A. F. M. Smith. New York: John Wiley & Sons, 149-157.

[6] Chaloner, K. and Brant R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, 75, 651-660.

[7] Dempster, A. P. and Ryan, L. M. (1985). Weighted normal plots. *JASA*, 80, 845-850.

[8] Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *JASA*, 85, 972-985.

[9] Geisser, S. (1980). Comments on "Sampling and Bayes' inference in scientific modeling and robustness" *JRSS-A*, 143, 416- 417.

[10] Geisser, S. (1993). *Predictive Inference: An Introduction*. London: Chapman and Hall.

[11] Hill, B. (1965). *Journal of the American Statistical Association*, 60, .

[12] Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963-974.

[13] Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, 82, 97-105.

[14] Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17, 624-642.

[15] Lange, N., Carlin, B. P. and Gelfand, A. E. (1992a). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *JASA*, 87, 615-625.

[16] Lange, N., Carlin, B. P. and Gelfand, A. E. (1992b). Rejoinder. *JASA*, 87, 631-632.

[17] Lange, K., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881-896.

[18] Little, R. J. A. (1988). Commentary. *Statistics in Medicine*, 7, 347-355.

[19] Louis, T. A. (1988). General methods for analyzing repeated measures *Statistics in Medicine*, 7, 29-45.

[20] Louis, T. A. (1992). Comment on Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *JASA*, 87, 626-628.

[21] Peña, D. and Guttman, I. (1993). Comparing probabilistic methods for outlier detection in linear models. *Biometrika*, 80, 603-610.

[22] Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *JRSS-B*, 52, 175-184.

[23] Pettit, L. I. and Smith, A. F. M. (1985). Outliers and Influential Observations in Linear Models. In *Bayesian Statistics 2*, Eds. J. M. Bernardo, M. DeGroot, D. Lindley, and A. F. M. Smith, pp. 473-94 Amsterdam: North Holland.

[24] Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics*, 18, 207-235.

[25] Sharples, L. D. (1990). Identification and accommodation of outliers in general hierarchical models. *Biometrika*, 77, 445-453.

[26] Wakefield, J., Smith, A. F. M., Racine-Poon, A., and Gelfand A. (1994). Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. it JRSS-A, 43, 201-221.

[27] Waternaux, C., Laird, N. M. and Ware J. H. (1989). Methods for analysis of longitudinal data: Blood-lead concentrations and cognitive development. *JASA*, 84, 33-41.

[28] Weiss, R. E. and Lazaro, C. G. (1992). Residual plots for repeated measures. *Statistics in Medicine*, 11, 115-124.

[29] Weiss, R. E. (1994a). Bayesian sensitivity analysis. Submitted for publication.

[30] Weiss, R. E. (1994b). Exploratory data graphics for repeated measures data. Technical report, submitted for publication.

[31] Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects. *JASA*, 86, 79-86.

case #	Week $O_\epsilon(i,j,3)$ 1	2	3	4	5	6	7	8	sum(ϵ)	MVO(ϵ)	O_β i	O_β s	sum(β)	MVO(β)
0	0	0	0	1	0	3	0	-	4	0	0	0	0	0
1	0	0	0	0	0	0	-	1	1	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	3	-	0	0	0	3	0	0	0	0	0
5	0	0	77	382	1	1	102	-	563	1315	0	1	1	0
6	1	80	0	0	0	2	1	1	85	5	0	0	0	0
7	0	0	0	0	0	4	0	0	4	0	0	0	0	0
8	0	1122	0	0	0	0	0	0	1122	7	0	3	0	0
9	0	0	0	0	-	0	-	0	0	0	0	0	0	0
10	0	29	0	0	0	0	0	0	29	2	0	258	258	69
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	-	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	1	0	-	-	1	0	0	2	2	0
15	11	0	0	0	8	0	0	0	19	3	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	1	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	-	1	-	-	1	0	0	0	0	0
20	0	0	0	0	0	0	-	0	0	0	0	0	0	0
21	0	0	0	0	0	0	-	-	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	1	0	-	9	-	10	1	0	0	0	0
24	1	0	0	0	0	0	0	0	1	0	0	0	0	0
25	0	0	0	0	106	62	0	9	177	65	424	0	424	476
26	0	273	0	10	1	0	0	0	284	270	0	15	15	8
27	0	0	0	0	0	0	-	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	10	0	-	126	0	136	70	0	0	0	0
30	2	0	0	0	0	-	-	-	2	10	0	0	0	0
31	0	0	0	0	0	0	-	-	0	0	0	0	0	0
32	1	0	0	0	0	0	0	0	1	0	0	0	0	0
33	1	0	0	0	0	-	-	-	1	0	0	0	0	0
34	0	0	0	0	3	-	30	-	33	0	0	0	0	0
35	0	0	0	331	101	179	-	-	611	1039	0	1	1	0
36	0	0	0	4	-	-	-	-	4	0	0	0	0	0
37	0	0	0	0	-	-	-	-	0	0	0	1	1	1

Table 3: Outlier statistics for the diet data set. Columns are 1) case number (from 0 to 37); 2) $2001 * O_\epsilon(i,j,3)$ for $j = 1, \ldots, 8$, missing data identified by '-'; sum(ϵ) is $O_\epsilon(i,+,3)$; MVO(ϵ) is $\mathrm{MVO}_\epsilon(i,k)$; sum($\beta$) is $O_\beta(i,+,3)$, and MVO(β) is $\Omega_\beta(k)$