Residuals and Outliers in Repeated Measures Random Effects Models

Robert E. Weiss*

October 17, 1995

Abstract

An approach for developing Bayesian outlier and goodness of fit statistics is presented for the linear model and extended to a hierarchical random effects model for repeated measures data. Diagnostics for univariate outliers, missing covariates, multivariate outliers and global goodness of fit are developed. Distribution theory for the posterior of the residuals is worked out. A local approach is used to show how omitted covariates and fixed and random effects affect residual summaries. Standard plots are interpreted in light of these understandings.

Key Words: Bayesian Data Analysis, Goodness-of-Fit, Hierarchical Models, Longitudinal Data, Outlier, Philosophy of Statistics, Shrinkage.

1 Introduction.

This paper develops a Bayesian approach to residual analysis and extends the approach to the random effects model (REM) used to model repeated

*Robert E. Weiss is Assistant Professor, Department of Biostatistics, Box 177220; UCLA School of Public Health; Los Angeles CA 90095-1772 U.S.A.; email rob@rem.ph.ucla.edu. This work was supported by grant #GM50011 from the NIGM. A portion of this work was completed while the author was on sabbatical at the University of Minnesota, Department of Applied Statistics. The author thanks Charlie Zhang for computing and plotting help; Kathryn Chaloner, Sandy Weisberg, Andrew Gelman for comments and encouragement; and two anonymous referees and the associate editor for substantial helpful comments.

measures (RM) data. Time or other metameter used to index the repeated measures characterizes repeated measures data and the example emphasizes the discovery of missing time effects. Except for time as a covariate, much of the discussion includes the one-way RE analysis of variance (ANOVA) and the random coefficients model as special cases. Distribution theory is fully developed and the presentation substantially extends earlier efforts in Chaloner and Brant (1988), Chaloner (1994), Hill (1965) and Hodges (1994).

The remainder of the section introduces notation for the RM normal REM model and reviews previous work on residual analysis for REMs. Section 2 presents philosophy for Bayesian residual analysis and presents numerous diagnostics in the context of the linear model. Section 3 extends the diagnostics to multivariate data and to the hierarchical random effects model. Section 4.1 develops distribution theory for the random effects model. A local approach is used to understand the effects of missing predictors on the residuals. The diagnostics are illustrated using data from a weight loss study. The paper closes with discussion.

1.1 The Repeated Measures Random Effects Model

1.2 The Model and Notation

The basic RM REM is

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

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 matrices of known 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 $q \leq p$; σ^2 is the sampling variance; $R_i = R_i(\upsilon)$ is a correlation matrix with diagonal elements R_{ijj} and υ is a vector of unknown parameters. The entire vector Y_i will be referred to as a subject, while individual elements y_{ij} will be called observations. The prior is flat except possibly for D with $p(\alpha, \sigma^2, \upsilon, D) \propto p(D, \upsilon)$ which leads to a proper posterior if $p(D, \upsilon)$ is proper. 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 and let $N = \sum n_i$. By far the most common model in current practice has $R_i = I$.

The marginal sampling distribution $Y_i|\alpha, \sigma^2, D, \upsilon$ is

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

where

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

The Gibbs sampler (Gelfand, Hills, Racine-Poon, and Smith 1990; Zeger and Karim 1991; Gilks, Wang, Yvonnet, and Coursaget 1993) permits straightforward Markov chain Monte Carlo sampling from the posterior of the parameters $\theta = (\alpha, \beta_1, \ldots, \beta_n, D, \sigma^2, \upsilon)$ given the data. I assume that samples $\theta^{(\ell)}, \ell = 1..L$ are available from $p(\theta|Y)$.

1.3 Previous work

There is a limited amount of previous research on REM data analysis and outlier detection from an empirical Bayes (EB), Bayesian or other point of view. Some references are Beckman, Nachtsheim and Cook (1987); Christensen, Pearson, and Johnson (1992); Dempster and Ryan (1985); Gilks, Wang, Yvonnet, and Coursaget (1993); Hocking, Green and Bremer (1989); Lange, Little and Taylor (1989); Lange and Ryan (1989); Little (1988); Louis (1988); Moulton (1987); Rutter and Elashoff (1995); Waternaux, Laird and Ware (1989). Hodges (1994) provides a general structure for the development of classical residual analysis in hierarchical models and gives many examples of diagnostic methods in the one way random effects model.

Let hatted quantities $\hat{\alpha}$, and $\hat{\beta}_i$ denote EB estimates of α and β_i and assume $R_i = I$. Louis (1988) defined residuals as $Y_i - X_i\hat{\alpha}$. Waternaux, Laird and Ware (1989) defined the $\hat{\beta}_i$ as residuals and plot $\hat{\beta}_i$ against possible covariates. Weiss and Lazaro (1992) define residuals $\hat{E}_i = Y_i - X_i\hat{\alpha}_i - Z_i\hat{\beta}_i$. They plot the $\hat{E}_i$ in an EB residual plot for repeated measures data called a parallel plot. 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. They show by example that this plot can display outliers, missing terms in the linear predictor $x_{ij}^t\alpha + z_{ij}^t\beta_i$, and lack of fit. In this paper I show that these ad hoc understandings are correct. Furthermore, I show it is also necessary to plot the $\hat{\beta}_i$'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).

2 Bayesian Residual Analysis

The discussion of this section is in terms of the linear regression model $y_i = x_i^t\beta + \sigma\epsilon$ with $\epsilon_i \sim N(0,1)$; the remainder of the paper extends and applies the ideas to analysis of the RM REM. Zellner (1975) initiated Bayesian residual analysis of the linear model.

The goal of classical residual analysis to date has been to analyze residual estimates, usually created by estimating β by $\tilde{\beta}$ and subtracting $\tilde{e}_i = y_i - x_i^t\tilde{\beta}$. This causes a problem, since for each estimator $\tilde{\beta}$, the properties of $\tilde{e}_i$ must be studied (confer Cook, Hawkins and Weisberg 1992, last paragraph). The goal of Bayesian residual analysis is to directly analyze the vector $\epsilon = (\epsilon_i)$

of residuals, through plotting or by use in further diagnostic calculations. First we determine what residual function(s) $g(\epsilon)$ we would inspect were ϵ known. Since ϵ is unknown, we then investigate the posterior distribution $p(g(\epsilon)|Y)$ instead of $g(\epsilon)$. Sometimes it is difficult to work with the entire posterior $p(g(\epsilon)|Y)$. One might then work with summaries of $p(g(\epsilon)|Y)$ such as $\mathrm{E}[g(\epsilon)|Y]$; such posterior expectations might be called residual summaries.

Many interesting residual functions are possible, I give several here; work is ongoing to develop more. Let $w = (w_i)$ be a covariate not in the regression. The first example of an interesting residual function is ϵ_i which is useful for plotting against covariates x_i from the regression or covariates w_i not in the regression. For informal outlier identification one might use ϵ_i or even more simply, $\psi_i = 1\{|\epsilon_i| > k\}$, the indicator function that $|\epsilon_i|$ is large; in particular, is $|\epsilon_i|$ larger than some $k > 0$; I generally take $k = 2$ or $k = 3$. Consider the classical hypothesis test of $H_0 : \epsilon_i \sim N(0,1)$ versus $H_A : \epsilon_i \sim N(\omega, 1)$; if ϵ_i were known, we would reject H_0 at the level $2 * \Phi(-k)$ provided $|\epsilon_i| > k$. We can take the outcome of the test to be the 0–1 indicator function $\psi_i = 1\{|\epsilon_i| > k\}$. To check for a needed covariate, we might plot ϵ against w. A numerical summary of this plot is $\gamma_w = w^t\epsilon/(w^tw)^{1/2}$, which is the classical test statistic for checking if the slope coefficient in the simple linear regression of ϵ on w is equal to zero. Again we can turn it into a $0-1$ indicator function $\psi_w = 1\{|\gamma_w| > k\}$, which corresponds to the classical test of $H_0 : \mathrm{E}[\gamma_w] = 0$ tested at level $2 * \Phi(-k)$.

Multivariate outlier statistics are also useful. Let W be an n by r full rank matrix of constants. Checking for ϵ outlying in the r dimensional space W is assessed by ϵ^tW. A priori $\gamma_W = \epsilon^tW(W^tW)^{-1}W^t\epsilon$ is distributed as a χ^2_r random variable with r degrees of freedom. A residual function for outlyingness is $\psi_W = 1\{\gamma_W > k\}$; one might also consider $\mathrm{E}[\gamma|Y]$. The cutoff k can be selected as $k = \chi^2_r(1-a)$, the $1-a$ quantile of a χ^2_r, where $a = 2 * \Phi(-2)$ or $a = 2 * \Phi(-3)$ so that the prior probability of an outlier matches with the choice of $k = 2$ or $k = 3$ respectively for univariate outlier detection. Specific choices of W might be i.) several covariates not in the regression; ii.) for $r = 2$, take W to have columns w and $w^2 \equiv (w_i^2)$, to check for a possible quadratic effect of w; iii.) to check for an r dimensional multivariate outlier, take the columns of W to be indicator functions and $W^tW = I$ is an r dimensional identity matrix.

In the lack of particular information, taking $W = I$, the n dimensional identity gives a goodness of fit test with $\gamma_I = \epsilon_i^t\epsilon_i = (Y - X\beta)^t(Y - X\beta)/\sigma^2$ which is a priori distributed χ_n^2; γ_I might be large if the prior and the likelihood support different values of (β, σ^2). This goodness of fit check is at the extreme end of a continuum of possible checks. At the other end of the spectrum, specific prior information reduces the dimension of W and should improve the power of such tests. It seems likely that ψ_I will have low power as compared to ψ_w where w is a known potential predictor and k is chosen from the χ_1^2 distribution.

Since ϵ is unknown, it has a posterior distribution in a Bayesian analysis; we can also consider the posterior distribution of functions of ϵ. When plotting, samples $\epsilon^{(l)}$ from $p(\epsilon|Y)$ may be plotted against W; many samples can be investigated through dynamic graphics. For small samples and for simple displays such as index plots, summaries of the posteriors $p(\epsilon_i|Y)$ can be plotted as in Chaloner and Brant (1988) and Albert and Chib (1995b). The posterior expectation of ψ_i corresponds to the residual summary $O_\epsilon(i, k) = P(|\epsilon_i| > k|Y)$, which was proposed as an outlier statistic by Chaloner and Brant (1988) and extended to additional models by Chaloner (1991, 1994) and Albert and Chib (1993a, 1993b, 1995a and see also 1995b). Bonferroni style choices for k are discussed by Chaloner and Brant (1988). Meng (1994) and Hodges (1994) also mix Bayesian and classical approaches for model checking.

A second goodness of fit statistic is based on the univariate residual functions $\psi_i = 1\{|\epsilon_i| > k\}$. Define $\Psi = \sum_i \psi_i$ and $\pi = 2\Phi(-|k|)$. A priori, Ψ is distributed binomial(n, π). The posterior distribution of Ψ can be investigated to determine model fit. For example, if $\mathrm{E}[\Psi|Y] > n\pi + k_\Psi n\pi(1-\pi)$, then we have evidence that there are too many outliers in the data. If $\mathrm{E}[\Psi|Y] < n\pi - k_\Psi n\pi(1-\pi)$, we have evidence that too few outliers are in the model. If $n\pi$ and $n(1-\pi)$ are large, say greater than 5, we might pick $k_\Psi = 2$ or $k_\Psi = 3$, as with a normal random variable. For $n\pi$ small, we can compute the prior distribution of Ψ which is binomial(n, π) and compare the posterior distribution with the prior. Taking $k = .125661 = \Phi^{-1}(.55)$ and checking $1 - \psi_i$ checks for the number of residuals near 0, specifically the inner $10\% = 2 * (.55 - .5)100\%$ of the normal distribution. A good fitting model should have approximately $.1n$ residuals in this range. An inappropriate prior for σ^2 which strongly shrinks σ^2 towards zero could produce a deficit of residuals near zero.

3 Outlier Identification in Random Effects Models

This section discusses the generalization of the outlier statistics and goodness of fit statistics to the random effects model. Checking for omitted predictors is delayed until after the distribution theory in the next section.

3.1 General Considerations

The residual summaries of the previous section provide methodology for identification of outliers, missing covariates and lack of fit. There are several methods for extending these statistics to multivariate data in general and hierarchical models in particular. 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 $\mu_i = (\mu_{ij})$ and v_{ijj} is the jj element of V_i. 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 EB 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 $Y_i^* = \Gamma_i Y_i$, where $\Gamma_i V_i \Gamma_i^T = \Lambda_i$, Λ_i is diagonal with diagonal elements λ_{ijj} the eigenvalues of V_i so that Y_i^* is the principle component decomposition of Y_i, and then evaluating $P(|Y_{ij}^* - \Gamma_{ij}^t \mu_i| > k\lambda_{ij}^{1/2})$, where Y_{ij}^* is the j^{th} element of Y_i^* and Γ_{ij}^t is the j^{th} row of Γ_i. This particular decomposition will not be followed further in this paper because 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 multiple values of k can be used for $M_1(i,j,k)$ and also for $M_2(i,k)$ as well as elsewhere.

For hierarchical models such as (1), the hierarchical structure suggests several outlier statistics. 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 $M_1(i,j,k)$. 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; for cases with $n_i = 1$, 1 and 3 are identical as are 2 and 4, and 5 and 7.

3.2 Univariate Outlier Statistics

Definitions 7 and 8 are simple and interpretable in the context of a hierarchical random effects model. Each level of the hierarchy produces a set of outlier statistics and by analogy, a set of residuals and residual means suitable for plotting. In general, it is possible to define a residual for any quantity with a known, proper prior.

At the base level, we have the residuals $\epsilon_{ij} = y_{ij} - x_{ij}^t\alpha - z_{ij}^t\beta_i|\sigma^2$ are a priori normal with mean zero and variance $\sigma^2 R_{ijj}$. Corresponding univariate outlier statistics are

$$O_\epsilon(i,j,k) = P(|\epsilon_{ij}| > k\sigma R_{ijj}^{1/2}|Y),$$

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 residual summary for β_{ij} is the posterior probability that β_{ij} is more than k standard deviations from its prior mean of zero

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

where d_{jj} is the j^{th} diagonal element of D. At the third level of the hierarchy, the parameters α have improper priors and no prior mean, so no residual exists for plotting purposes and we do not check them for outlyingness. Diagnostics for nonlinear parameters such as (v, D, σ^2) will be worked out elsewhere. When interest lies specifically in high and low outliers, 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

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

The posterior distributions of the β_i's are taken straight from the Gibbs 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 using results in section 4.1.

3.3 Summary Outlier Statistics

The outlier statistics $O_\epsilon(i,j,k)$, and $O_\beta(i,j,k)$ will produce $n_i + q$ outlier statistics per case which can be a burden in large data sets. 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, as this might miss sets of observations that have large ϵ_i and β_i but are not the most extreme.

A single number summary measure of the $O_\epsilon(i,j,k)$ that is directly useful for screening cases for 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 good screening statistic for the individual $O_\epsilon(i,j,k)$ since

$O_\epsilon(i,j,k) \le O_\epsilon(i,+,k)$; if $O_\epsilon(i,+,k)$ is small or zero, all of the $O_\epsilon(i,j,k)$ are necessarily small or zero. It can be interpreted as the posterior expected number of ϵ outliers for case i. When there is more than one random effect, a parallel construction leads to

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

3.4 Multivariate Outlier Statistics

A disadvantage of univariate diagnostics is that they cannot find multivariate outliers. Multiple $O_\epsilon(i,j,k)$ outliers for one case may suggest that the case is a multivariate outlier in $\mathcal{R}^{n_i}$ but a subject can have no univariate outliers and still be a multivariate outlier. Many readers will be familiar with the example plot of data from a highly correlated bivariate distribution with one case that is not extreme on either coordinate axis, but is unusual in the bivariate distribution.

Consider a situation where $\epsilon_{i1} = 4\sigma$ is a large outlier, but $\epsilon_{ij} = 0$ for $j > 1$. Then the univariate outlier statistic for case i, observation 1 will identify ϵ_{i1} as an outlier, and the summary measure also identifies something suspect for case i. The multivariate outlier statistic only measures outlyingness by $\sum_j \epsilon_{ij}^2$, which will not be unusually large compared to the $1 - \Phi(-3)$ quantile of a χ^2 with the appropriate degrees of freedom if $n_i > 1$. A contrasting situation which we would like to identify occurs when all the ϵ_{ij} are moderately large, say equal to $\pm 2\sigma$. When n_i is greater than 5, a multivariate outlier statistic will identify this as an outlying case, even though no individual univariate $O_\epsilon(i,j,3)$ is nonzero. Thus both univariate and multivariate outlier statistics are useful.

Definitions 1, 5 and 6 might be used to check for multivariate outliers. Definition 1 seems less 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 $\xi_{\epsilon i} = \epsilon_i^t(\sigma^2 R_i)^{-1}\epsilon_i$ and $\xi_{\beta i} = \beta_i^t D^{-1}\beta_i$. A priori $\xi_{\epsilon i}$ is distributed $\chi^2_{n_i}$, and similarly, $\xi_{\beta i}$ is a priori distributed χ^2_q. Define

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

where k is chosen to be $\chi^2_{n_i}(1-a)$, and similarly for $\mathrm{MVO}_{\beta_i}(k)$, except k would be chosen based on a χ^2_q. Like $O_\epsilon(i,j,k)$ and $O_\beta(i,j,k)$, these statistics have interpretations as the posterior expected probability of rejecting a null hypothesis. The alternative hypothesis for $\mathrm{MVO}_\epsilon(i,k)$ is $\epsilon_i|\omega_i, \sigma^2, v_i \sim N_{n_i}(\omega_i, \sigma^2 R_i)$ with null hypothesis $H_0 : \omega_i = 0$. Similarly, for $\mathrm{MVO}_\beta(i,k)$, the alternative model is $\beta_i|\omega_i, D \sim N(\omega_i, D)$. The outlier statistics are again the posterior probabilities of rejecting $H_0 : \omega_i = 0$ at the level a, where $k = \chi^2_{n_i}(1-a)$ or $k = \chi^2_q(1-a)$.

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

3.5 Goodness-of-fit

This subsection generalizes the regression model goodness of fit checks to the random effects model. Let $\epsilon = (\epsilon_1^t, \ldots, \epsilon_n^t)^t)$, $\beta = (\beta_1^t, \ldots, \beta_n^t)$, and let $B_j = (\beta_{1j}, \ldots, \beta_{nj})$ be the vector of j^{th} random effects.

As with the regression model we can take $\epsilon^t(\mathrm{diag}(R_i)^{-1})\epsilon/\sigma^2$ as a goodness of fit statistic, where $\mathrm{diag}(R_i)$ is the block diagonal matrix with blocks R_i. A priori it is distributed χ^2_N, where $N = \sum n_i$. Similar goodness of fit statistics can be based on $B_j^t B_j/d_{jj}$ and $\beta^t(I \otimes D^{-1})\beta$. As with the linear model, these statistics will probably have low power as compared to statistics based on prior information about the direction of lack of fit. Still, they should be able to detect disagreement between prior and likelihood.

More interesting are the following statistics based on Ψ. Consider

$$\Psi_\epsilon(a) = \sum_{i=1}^{n} \mathrm{MVO}_{\epsilon i}(k_i(a))$$

and

$$\Psi_\beta(a) = \sum_{i=1}^{n} \mathrm{MVO}_{\beta i}(k),$$

where $k_i(a)$ is chosen to be the $\chi^2_{n_i}(1-a)$ distributions, and $k = \chi^2_q(1-a)$. Then a priori, $\Psi_\epsilon(a)$ and $\Psi_\beta(a)$ are distributed binomial(n, a).

A goodness of fit statistic may not be perfectly useful in a complex model like the RM REM. Unlike a graphic, it does not tell us what part of the model is the root cause for a particular value. For example, $\Psi_\epsilon(a)$ or $\Psi_\beta(k)$ 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 several goodness of fit statistics for ϵ and β and the possibility of the targeted diagnostics in subsection 4.4 make goodness of fit statistics potentially much more useful.

4 The Distribution of the Residuals

Prior to discussing diagnostics for missing predictors it is important to understand how the residuals behave if the model is correct and how they behave under model misspecification. The next subsection gives distribution theory for ϵ and β given a flat prior and $R_i = I$. Subsection 4.2 interprets the results of this section, with emphasis on the posterior means of ϵ and β. Subsection 4.3 takes a local approach to determining how unmodeled covariates affect the residuals and subsection 4.4 gives outlier statistics for detecting omitted covariates.

4.1 Distribution Theory

The goal of this section is to get as close to $p(\epsilon|Y)$ and $p(\beta|Y)$ as possible. The covariance matrix of ϵ_i is assumed to be $\sigma^2 I$. The calculation of the outlier statistics depends on the unknown value of σ, and integrating σ^2 converts normal densities to t, and does not aid understanding. It is not possible to marginalize with respect to D in closed form, so much of the analysis is conditional on D and σ^2.

Relevant earlier distribution theory in the one way REM can be found in Hill (1965), Hodges (1994) and Chaloner (1994). Hodges gives geometric intuition for the one way REM and Duncan and Horn (1972) give relevant geometry and algebra for the Bayesian linear model.

Model (1) can be written in more compact notation. Define $Y = (Y_1^t, \ldots, Y_n^t)^t$; $X = (X_1^t, \ldots, X_n^t)^t$. Let $Z = \mathrm{diag}(Z_i)$, be an irregular block diagonal matrix N by nq; the blocks may be of differing sizes. Then (1) with $R_i = I$ can be written in the form

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

The joint posterior can be written as

$$p(\alpha, \beta|\sigma^2, D, Y)p(\sigma^2|D, Y)p(D|Y). \tag{4}$$

The following matrices are used in the posterior means of α, ϵ and β and in the associated discussion. They come from inspecting the distributions in (4) as opposed to, for example, the methods of Hodges (1994), which took a least squares approach.

$$\begin{aligned} H_X &= X(X^t X)^{-1} X^t \ , \\ Q_X &= I - H_X \ , \end{aligned}$$

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

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

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

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

where a bar indicates the posterior mean as in $\bar{\alpha} = E[\alpha|Y]$, and $\bar{\alpha}(D) = E[\alpha|D,Y]$ and similarly for β and ϵ; $\Gamma(a)$ 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 $\alpha, \beta|D, \sigma^2, Y$ are normal with means that can be written in either of two useful forms

$$
\begin{aligned}
\bar{\alpha}(D) &= (X^tQ_ZX)^{-1}X^tQ_ZY && (5) \\
&= (X^tX)^{-1}X^t(I - P_Z)Y && (6) \\
\bar{\beta}(D) &= (Z^tZ + I \otimes D^{-1})^{-1}Z^t(I - P_X)Y && (7) \\
&= (Z^tQ_XZ + I \otimes D^{-1})^{-1}Z^tQ_XY. && (8)
\end{aligned}
$$

The variances and covariances from $p(\alpha, \beta|\sigma^2, D, Y)$ are

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

The scale parameter σ^2 is conditionally inverse χ^2 with $N-p$ degrees of freedom and $2Q(Y)$ shape parameter. The marginal 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 unobserved residuals ϵ are linear combinations of β and α.

$$\epsilon | D, \sigma^2, Y \sim N(\bar{\epsilon}(D), \sigma^2 M)$$

where

$$\begin{aligned} \bar{\epsilon}(D) &= Y - X\bar{\alpha}(D) - Z\bar{\beta}(D) = Q_Z(I-P_X)Y && (9) \\ &= Q_X(I-P_Z)Y && (10) \end{aligned}$$

and

$$M = X(X^tQ_ZX)^{-1}X^t + Q_XP_Z - H_XZ(Z^tQ_XZ + I \otimes D^{-1})^{-1}Z^tH_X.$$

The covariance of ϵ and β is $Q_XZV_{\beta\beta}$, 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. Integrating out σ^2 above gives a t distribution for $\epsilon|D,Y$, however, this does not help in understanding or interpreting the residuals ϵ or in computing outlier and goodness-of-fit statistics; due to the nature of these calculations the cutoff depends on σ. 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.

The next subsection interprets $\bar{\beta}(D)$ and $\bar{\epsilon}(D)$ and the various posteriors.

4.2 Interpretation

It is not simple to extend ideas of projection matrices, fitted values and residuals from the linear regression model to the RM REM. Because there are two levels in the model, there are two fitted values $X_i\bar{\alpha}$ and $X_i\bar{\alpha} + Z_i\bar{\beta}_i$ for any observation, and there are 3 residuals, ϵ, β, and $Z\beta + \epsilon$. Here I concentrate on the ϵ and β. Since these are unobserved, their posterior distributions are inspected instead; for concreteness the discussion centers on the posterior means $\bar{\epsilon}(D)$ and $\bar{\beta}(D)$. The posterior mean is a posteriori unbiased for the unobserved residuals. As in the previous subsection, this discussion is conditional on D and σ^2.

The matrices H_X, Q_X and P_X are projection matrices, for example H_XY is the projection of Y onto $\mathcal{S}(X)$, the column space of X and $P_XY = X\bar{\alpha}(D)$ is the projection adjusted for the variance structure of Y_i in equation (2). Also $P_ZY = Z\bar{\beta}(D)$ and P_Z removes the appropriate amount of variation from Y for estimating β and $(I - P_Z)$ removes the appropriate amount of variation in Y so that $X\bar{\alpha}(D) = H_X(I - P_Z)Y$. On the other hand, H_Z, Q_Z, and P_Z are submatrices of projection matrices. This means for example, that $H_ZH_Z \leq H_Z$, where $A \leq B$ for matrices means that $B - A$ is positive semidefinite. It also means that $I - P_Z$ does not remove all of the variation of Y in $\mathcal{S}(Z)$, some variation remains.

This is not necessarily bad; to fix ideas, consider the one way ANOVA model where $X_i = Z_i \equiv 1$ are n_i vectors of ones. The fixed effects ANOVA has $D^{-1} = 0$, and the usual frequentist residual e_i is the same as $\bar{\epsilon}_i(D)$ with $D^{-1} = 0$. Then $1^te_i = 0$ for all i, even though $1^t\epsilon_i$ has a prior variance of $n_i\sigma^2$ and is almost certainly not zero; 1^te_i is not an especially good point estimate of $1^t\epsilon_i$. It has sampling variability equal to zero. The idea that e_i is an estimate of ϵ_i is not usually made in the classical diagnostics literature. In the Bayesian analysis, $1^t\epsilon_i$ has posterior variability $n_i\sigma^2$; its prior distribution is the same as the posterior variance and we do not learn about it from the data. In contrast in the Bayesian analysis of the random effects one way ANOVA, the shrinkage/projection matrix H_Z, with $D^{-1} > 0$, removes the appropriate amount of variation from Y projected onto $\mathcal{S}(Z)$ for estimation of $Z\beta$, but leaves a portion of Y in $\mathcal{S}(Z)$ so that $1^t\bar{\epsilon}(D)_i$ is not zero; in the random effects model we are able to learn about the posterior of $1^t\epsilon_1$ from the data.

Formulae (5) - (10) are a random effects model generalization of the usual regression formulae for coefficient estimates and residuals with two sets of predictors X and Z. The posterior means $\bar{\alpha}(D)$, $\bar{\beta}(D)$ and $\bar{\epsilon}(D)$ can be thought of as estimates achieved in either of two ways. For $\bar{\epsilon}(D)$, one can first remove the variation in Y accounted for by $X\bar{\alpha}(D)$ followed by the variation accounted for by $Z\bar{\beta}(D)$ or the order can be reversed. These are equations (9) or (10) respectively. In both cases, the second removal of residual variation by Q_Z or Q_X ignores the presence of the other predictor in the model, while the first removal of residual variation by $(I - P_X)$ or $(I - P_Z)$ removes the appropriate amount of variation adjusted for the other predictor. From (5), $\bar{\alpha}(D)$ is the estimated regression coefficient from the regression of Y on X adjusted for Z, and from (6), it is the unadjusted regression of $(I - P_Z)Y$ on X. The interpretations of $\bar{\beta}(D)$ falls in between those of $\bar{\epsilon}(D)$ and $\bar{\alpha}(D)$; it is (equation 8) a shrunken regression (a regression with a proper prior) of Y on Z adjusted for regressors X or (equation 7) it is a shrunken regression of $(I - P_X)Y$ on Z.

As the variance ratio D varies, the projection H_Z varies from the projection matrix $Z(Z^tZ)^{-1}Z^t$ when $D^{-1} = 0$ to the 0 matrix when $D = 0$. The matrix D adjudicates between two fixed effects models with predictors X or predictors X and Z. Hodges (1994) has more discussion on this point.

4.3 Effects of Y perturbation on $\bar{\epsilon}(D)$ and $\bar{\beta}(D)$

This section considers the effect on $\bar{\epsilon}(D)$ and $\bar{\beta}(D)$ of local perturbations of Y. Consider the effect of replacing Y by $Y + \delta M$ where δ is small or goes to zero, and where M is a vector of perturbations. The question of primary interest is how the residual summaries $\bar{\epsilon}$ and $\bar{\beta}$ change; if the perturbation affects the residual means, then for large values of δ the perturbation should be discoverable from appropriate inspections of $\bar{\epsilon}$ and $\bar{\beta}$. The analysis is conditional on D, since the unconditional analysis is difficult. Three special cases are of particular interest, i.) $M \in \mathcal{S}(X)$; ii.) $M \in \mathcal{S}(Z)$ but $M \in \mathcal{S}^{\perp}(X)$ that is, M is in the orthogonal complement of X, and $X^t M = 0$. It is convenient to assume that $\mathcal{S}(X)$ is the null space; and iii.) $M \in \mathcal{S}^{\perp}(X, Z)$. These special cases guide our interpretations in practice.

Case i, $M \in \mathcal{S}(X)$. This is the easiest case. The posterior means of $\bar{\epsilon}(D)$ and $\bar{\beta}(D)$ do not change from (10) and (8). Neither $\bar{\sigma}^2(D)$ nor the posterior $p(D|Y)$ change, since both depend on Y only through $Q(Y)$ and $Q(Y)$ will not change.

Case ii, $M \in \mathcal{S}(Z)$, and $\mathcal{S}(X)$ is the null space. Then

$$\left.\frac{\partial \bar{\epsilon}(D)}{\partial \delta}\right|_{\delta=0} = Q_Z M$$

and

$$\left.\frac{\partial \bar{\beta}(D)}{\partial \delta}\right|_{\delta=0} = (Z^t Z + I \otimes D^{-1})^{-1} Z^t M$$

so that $\partial Z\bar{\beta}(D) + \bar{\epsilon}(D)/\partial\delta = M$, and the perturbation is divided up between the two residuals. Understanding where the bulk of the perturbation goes requires an understanding of the eigenstructure of H_Z for the eigenvectors in $\mathcal{S}(Z)$. If $Z_i^t Z_i$ is large compared to D^{-1}, then the eigenvalues are close to 1 and most of the perturbation will end up in $\bar{\beta}(D)$, if D^{-1} is of the same size as $Z_i^t Z_i$, then the perturbation will be split between $\bar{\beta}(D)$ and $\bar{\epsilon}(D)$. If D^{-1} is large compared to $Z_i^t Z_i$, then the perturbation will mostly affect $\bar{\epsilon}$. In the special case of the one-way REM, if the $n_i/(n_i + D^{-1})$ are close to 1, then most of the variation of M in $\mathcal{S}(Z)$ will be in $Z\bar{\beta}$. In the weight loss example of the next section, D^{-1} from the one-way REM is approximately .6, with n_i

nominally 8 and we can expect to see unmodeled fixed effects in $\mathcal{S}(Z)$ in the $\bar{\beta}$ and not in $\bar{\epsilon}(D)$.

Case iii, $M \in \mathcal{S}^{\perp}(X, Z)$. Although $Q_X M = M$, we have $Z^t M = 0$, so that by (8), $\bar{\beta}(D)$ will not change as a function of δ. Also, $H_X M = 0$, so $\bar{\alpha}(D)$ will not change, and all of the perturbation will show up in $\bar{\epsilon}(D)$, and $\partial\bar{\epsilon}(D)/\partial\delta = M$. To find unmodeled fixed effects not in $\mathcal{S}(Z)$ and any unmodeled random effects one should inspect ϵ.

4.4 Diagnostics for missing elements of the linear predictor.

Consider a full rank matrix W, N by r of known time varying covariates; let $R = \text{diag}(R_i)$; I do not assume $R_i = I$ in this subsection. To see if W is useful as either a fixed or random effect we can consider $O_\epsilon(W, k) = P(\epsilon^t R^{-1} W (W^t R^{-1} W)^{-1} W^t R^{-1} \epsilon / \sigma^2 > k | Y)$; a priori the quadratic form is distributed χ^2_r; choose $k = \chi^2_r(1-a)$. If W is an indicator function, or a matrix of r indicator functions, then this reduces to a univariate or multivariate outlier statistic. A common choice for W would be functions of time $t = (t_1^t, \ldots, t_n^t)^t$. In a random intercept model, taking $W = t$ would check for both a missing fixed or random time trend. Taking W to be a vector of ones and zeros with ones whenever $t_{ij} = t_0$ checks for bias at a particular time t_0. Similarly, contrasts in time can be taken to check for quadratic or other nonlinear time effects. For unbalanced data, the columns of W could be a spline basis evaluated at t_{ij} to check for a smooth grand mean trend that differs from the current model.

Let $M = (m_i)$ be a n by r vector of baseline or time fixed covariates. Consider checking to see if M predicts B_j. One possibility is to use $P(B_j^t M (M^t M)^{-1} M^t B_j / D_{jj} > k | Y)$, where D_{jj} is the j^{th} diagonal element of D and with k chosen from quantiles of the χ^2_r density. If it is decided to enter a single covariate M with $r = 1$ into the model because it is correlated with B_j, the additional predictor vector for person i would be $m_i z_{ij}$.

When W is in $\mathcal{S}(Z)$ and the model has eigenvalues of H_Z not close to zero or one so that residuals in the space $\mathcal{S}(Z)$ are split between $Z\beta$ and ϵ, we can consider the statistic $P((Z\beta + \epsilon)^t(ZDZ^t + R)^{-1}W(W^t(ZDZ^t + R)^{-1}W)^{-1}W^t(ZDZ^t + R)^{-1}(Z\beta + \epsilon)$

5 Weight Loss Data.

Here I illustrate how the diagnostics can be used to aid model development. I illustrate the univariate, summary and multivariate diagnostics, checks for needed time effects and goodness of fit statistics. The uses illustrate the model building process. While an ultimate model is not determined here, the diagnostics clearly indicate the continuing and substantial improvement in the model.

The data set contains up to 8 weekly observations on 38 women enrolled in a diet study. Some measurements were missing for a total of 265 individual observations. A parallel plot of the raw data (not shown) suggests a random intercept is needed. Each person's trend varied around a straight, nearly flat line. Visually extrapolating back to $t = 0$ suggests that each individual intercept is quite different; noise in our visual estimates of intercepts is unlikely to account for the between individual differences in intercepts. It is a random and not a fixed intercept if one would use the weights from this sample as a prior to predict the weight of a new member of the study population. Unless unusual care is taken in drawing the plot, more information is not visible in the original raw data parallel plot. Most calculations are based on a Gibbs sample of size 2001; exceptions are the ϵ diagnostics of section 5.3 and 5.4 which are based on samples of size 1000.

Analysis from the model with a random intercept (model 1) shows the

Model	1	2	3	4
$E[\sigma\|Y]$	4.00	2.81	2.21	1.77

Table 1: Posterior means of σ under the four models.

need for a fixed slope (model 2), and a random slope (model 3); analysis of model 3 shows the need for a general fixed effect structure (model 4), at which point this analysis stops. The diagnostics are used to motivate and validate the increasing complexity of the model. Multiple diagnostics can, of course, point the way to the same needed improvements.

5.1 Plots

First a random intercept model was fit. Figure 1 shows a parallel plot of the posterior means of the ϵ_i's for model 1. A time trend is clearly visible; the residuals for early times are mostly positive, while residuals for later times are mostly negative. A fixed effect for time should clearly be included. Careful inspection of the residual curves, for example by dynamically building this figure one curve at a time suggests that a random slope is also needed; a fanning of the lines at $t = 1$ and $t = 8$ and a narrowing around $t = 3.5$ is barely visible in Figure 1. A random slope does make substantive sense as it suggests that each person loses weight at her own rate. In Figure 1, some outlying cases are identified. These are discussed in the next section. Figures 2 and 3 show parallel plots of the posterior means of the residuals for models 3 and 4. Figure 2 has an '*M*' shape, suggesting unmodeled fixed time effects. In consecutive figures we see a decreasing range of the residuals, indicating smaller residual variation and we also see that the residuals increasingly fill the data rectangle, suggesting increasing uniformity of the residuals across time. The smaller residual variation is also indicated by the posterior means of σ given in table 1 which decrease as the structure in the model increases.

Plots of the posterior means of the random effects are similar for models

3 and 4. Figure 4 shows the scatterplot of the posterior means of the random intercept and slope for model 4. Two cases are identified. There is some evidence of non-normality in that larger negative slopes are increasingly more likely with larger intercepts. Quantile-quantile plots (not shown) of posterior samples of β_i^2/D for models 1 and 2 suggested that there was a single outlier with reasonable probability.

5.2 Case Statistics

Table 2 gives $O_\epsilon(i,j,k)$ and $O_\beta(i,j,k)$ with $k = 3$, for model 3, the random intercept and slope model. Similar tables were examined for the other models also. Models with increasing structure had increasing numbers of observations with non-zero values of $O_\epsilon(i,j,k)$. Model 1 had the fewest identified outlying observations – most are identified in figure 1 and except for observations n_1 and n_2, all identified cases are probable outliers at $k = 3$.

In table 2 the columns headed S_ϵ, M_ϵ, S_β, and M_β are $O_\epsilon(i,+,3)$, $\mathrm{MVO}_\epsilon(i,k)$, $O_\beta(i,j,3)$, $j = 1$ is the intercept, and $j = 2$ for the slope, and $\mathrm{MVO}_\beta(k)$ respectively. A zero indicates $|\epsilon_{ij}^{(l)}|$ or $|\beta_{ij}^{(l)}|$ was never greater than 3. A .00 indicates a value greater than 0 but less than .01 after rounding. A dash indicates a missing y_{ij} value. The constant k has been set to 3 for the summary outlier statistic and to the χ^2 quantile with tail area equal to $2 * \Phi(-3)$ for the multivariate outlier statistics.

Cases 6 and 36 have several outlying ϵ_{ij}'s particularly at week 4; case 9, week 2 is outlying, and possibly case 27 at week 2. Case 6 is identified in Figure 2 at weeks 3, 4, and 7. Case 36 is identified at weeks 4 and 5. For the random effects, case 11 has an unusual random slope, and case 26 has an unusual random intercept. The multivariate and univariate diagnostics $\mathrm{MVO}_\beta(i,3)$ and $O_\epsilon(i,j,3)$ disagree for case 11; for case 9, the ϵ diagnostics disagree. This situation was discussed earlier.

Table 2 can be explored much as a graph might be. As the eye skims the table, outlying cases such as 6 and 7 draw attention to themselves, and with closer inspection it is easy to pick out case 6 week 4, case 9 week 2, and case 36 week 4 as the most probable ϵ outliers. Either column S_ϵ or M_ϵ can be used to screen cases for outliers; for this relatively small data set, screening is not so necessary. Still, it is easy, for example, to go down the single column S_β and find cases 11 and 26. For larger data sets which are becoming common, the screening statistics will be quite useful.

The outlying observations can be identified in Figure 3. They further emphasize the M shape of the residuals across time; outlying observations only occur at the points of the M. The outliers in this figure appear to be due to the quartic time effect still visible in the residuals. This missing quartic effect is formally identified as an outlying residual direction in subsection 5.4.

5.3 Goodness of Fit Checks

Table 3 gives the results of the goodness of fit analysis for the four models. The first two sections are for the β goodness of fit check, checking for an excess of multivariate outliers at $a = .05$ and $a = .01$. The second two sections check for an excess of multivariate ϵ outliers, also at tail areas $a = .05$ and $a = .01$. The row labeled p in the table gives the prior distribution of the number of outliers; this row depends only on a and is the same for the β and ϵ checks. It is fairly clear that there are not an unusual number of multivariate β outliers $\beta^t D^{-1} \beta$ greater than the $\chi^2_q(1 - .05)$ or $\chi^2_q(1 - .01)$ cutoff for any of the models. For models 1 and 2, the posterior probability is essentially 1 that there is at most 1 multivariate β outlier. For models 3 and 4, the posterior distributions of the number of outliers is not too far from the prior distribution, again suggesting nothing unusual.

The ϵ outliers are a different story. For model 1 and $a = .05$, for example, there is .61 probability that there are 4 outliers, and .31 probability that there are 5 outliers, and a large .05 probability that there are 6 outliers,

case	Week				$O_\epsilon(i,j,3)$				S_ϵ	M_ϵ	O_β		S_β	M_β
#	$i=1$	2	3	4	5	6	7	8			$j=1$	2		
1	0	0	0	.00	0	.00	0	-	.00	0	0	0	0	0
2	0	0	0	0	0	0	-	.00	.00	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	.00	-	0	0	0	.00	0	0	0	0	0
6	0	0	.04	.19	.00	.00	.05	-	.28	.66	0	.00	.00	0
7	.00	.04	0	0	0	.00	.00	.00	.04	.00	0	0	0	0
8	0	0	0	0	0	.00	0	0	.00	0	0	0	0	0
9	0	.56	0	0	0	0	0	0	.56	.00	0	.00	0	0
10	0	0	0	0	-	0	-	0	0	0	0	0	0	0
11	0	.01	0	0	0	0	0	0	.01	.00	0	.13	.13	.03
12	0	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
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	.00	0	-	-	.00	0	0	.00	.00	0
16	.01	0	0	0	.00	0	0	0	.01	.00	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	.00	0	0	0	0	0	0	0	.00	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	-	.00	-	-	.00	0	0	0	0	0
21	0	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
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	.00	0	-	.00	-	.00	.00	0	0	0	0
25	.00	0	0	0	0	0	0	0	.00	0	0	0	0	0
26	0	0	0	0	.05	.03	0	.00	.09	.03	.21	0	.21	.24
27	0	.14	0	.00	.00	0	0	0	.14	.13	0	.01	.01	.00
28	0	0	0	0	0	0	-	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	.00	0	-	.06	0	.07	.03	0	0	0	0
31	.00	0	0	0	0	-	-	-	.00	.00	0	0	0	0
32	0	0	0	0	0	0	-	-	0	0	0	0	0	0
33	.00	0	0	0	0	0	0	0	.00	0	0	0	0	0
34	.00	0	0	0	0	-	-	-	.00	0	0	0	0	0
35	0	0	0	0	.00	-	.01	-	.02	0	0	0	0	0
36	0	0	0	.17	.05	.09	-	-	.31	.52	0	.00	.00	0
37	0	0	0	.00	-	-	-	-	.00	0	0	0	0	0
38	0	0	0	0	-	-	-	-	0	0	0	.00	.00	.00

Table 2: Outlier statistics for the weight loss data set from the model with a random slope and intercept. Columns are 1) case number from 1 to 38; 2-9) $O_\epsilon(i,j,3)$ for $j = 1,\ldots,8$, missing data identified by '-'; 10) S_ϵ is $O_\epsilon(i,+,3)$; 11) M_ϵ is $\mathrm{MVO}_\epsilon(i,k)$; 13-14) $O_\beta(i,j,3)$, $j=1$ is for the intercept, $j=2$ is the slope; 12) S_β is $O_\beta(i,+,3)$, and 12) M_β is $\mathrm{MVO}_\beta(k)$.

model					$p(\Psi_\beta(.05)\|Y)$				
	0	1	2	3	4	5	6	7	8+
p	.1423	.2848	.2773	.1751	.08065	.02887	.00836	.00201	.00050
1	.1449	.8506	.0040	.00050					
2	.1109	.8841	.0040	.0010					
3	.01749	.3243	.3773	.2064	.06447	.00700	.00300		
4	.01549	.3228	.4098	.1989	.04748	.0040	.0010	.00050	
					$p(\Psi_\beta(.01)\|Y)$				
	0	1	2	3	4	5	6	7	8+
p	.6826	.2620	.04896	.00593	.00053	3.60e-05	2.00e-06	9.24e-08	3.743e-09
1	.7281	.2719							
2	.6652	.3348							
3	.3333	.5612	.0960	.00950					
4	.3713	.5257	.09745	.00500	.00050				
					$p(\Psi_\epsilon(.05)\|Y)$				
	0	1	2	3	4	5	6	7	8+
p	.1423	.2848	.2773	.1751	.08065	.02887	.00836	.00201	.00050
1	0	0	0	.019	.610	.309	.054	.007	.001
2	0	0	.001	.009	.093	.585	.255	.050	.007
3	0	0	.078	.285	.284	.209	.096	.034	.013
4	0	.013	.164	.302	.279	.162	.071	.009	
					$p(\Psi_\epsilon(.01)\|Y)$				
	0	1	2	3	4	5	6	7	8+
p	.6826	.2620	.04896	.00593	.00053	3.60e-05	2.00e-06	9.24e-08	3.743e-09
1	0	.003	.036	.639	.305	.017			
2	0	.081	.263	.222	? .329	.099	.006		
3	.018	.052	.650	.234	.041	.005			
4	.019	.364	.477	.124	.014	.002			

Table 3: Goodness of fit statistics. The symbol p indicates the prior distribution of the number of outliers, rows beginning 1, 2, 3, or 4 are posterior distributions of the number of outliers conditional on that model. The β and ϵ posteriors are based on Gibbs samples of size 2001 and 1000 respectively.

which a priori is quite rare. For model 2, the probability is that there are 5 or 6 multivariate ϵ outliers, with probability .034 that there are 7 outliers. Models 3 and 4 fit much better. For model 4, and $a = .01$, there is even 2% probability that there are 0 outliers, which a priori has probability 68% and there is a 36% and 48% probability of only 1 or 2 outliers.

5.4 Outlyingness in a Priori Suspected Directions.

Here we discuss diagnostics to check for particular univariate effects. Linear, quadratic, cubic and quartic effects in time were created, the linear effect W_1 was equal to time minus the grand mean of all 265 times; the quadratic time effect was regressed on the intercept and linear effect and the residuals used as W_2, and so on for the cubic W_3 and quartic effect W_4; each effect is a vector of 265 elements, since there were 265 observations; each effect is orthogonal to all other effects. I then calculated the posterior distributions of $W_m^t\epsilon/(\sigma^2 W_m^t W_m)^{1/2}$, for all four models, where $m = 1, 2, 3, 4$ represents the linear through quartic effects respectively. Table 4 summarizes the results. For model 1, the linear contrast was an enormous outlier, with $O_\epsilon(W_1, 3) = 1$, and a posterior expected mean square $\mathrm{E}[\epsilon^t W(W^t W)^{-1} W^t \epsilon | Y]$ of 108.2. For models 2, 3, and 4, the posterior probabilities are approximately equal to the prior probabilities that the linear effect is an outlier. In contrast, the quartic effect contrast is a strong outlier for models 1, 2, and 3 and is particularly visible in figure 2. The W_2 contrast is a moderate outlier, and only for model 4 is it not an outlier. The cubic effect is apparently not an outlier for any model.

A small simulation study was run for each model to confirm the general interpretation of the results. For each model, 250 random vectors W of length 265 were generated. The components were independent and identically distributed standard normals. The distribution of $W^t\epsilon$ for these 250 vectors was checked. The number in column 6 of table 4 is the fraction of these

250 vectors with $\mathrm{E}[\epsilon^t W(W^t W)^{-1} W^t \epsilon|Y]$ larger than the value in column 3. Similarly, columns 7 and 8 contain the fraction of the vectors with larger values of the outlier statistics $O_\epsilon(W_m, 2)$ and $O_\epsilon(W_m, 3)$, for $m = 1, \ldots, 4$. The simulation study confirms the interpretations of the diagnostics that large values of the mean square, $O_\epsilon(W_m, 2)$, or $O_\epsilon(W_m, 3)$ indicate a direction in which the ϵ residuals are unusual.

6 Discussion.

The primary purpose of residual and outlier analysis is to check model fit. When general lack of fit is found, then plots of residuals might be inspected to indicate needed model improvement. Alternatively, specific checks targeted at particular model components such as missing predictors, missing correlations, misspecified links, or needed transformations can be considered next. Outlier identification per se is an end in itself only after the model is a posteriori quite certain, or when it seems more parsimonious to report a particular model as fitting most of the data with a few exceptions (outliers) than to try to expand the model further.

If it is felt that the posterior uncertainty in the residuals should be accounted for in the residual plots, then a simulation may be run, where samples from the joint posterior of ϵ or β are used in place of the posterior means. The analyst would look at several simulations of the residual plots before deciding if extra fixed or random effects needed to be added to the model. I found looking at parallel plots of residual means useful, while looking at many parallel plots of posterior samples of the residuals seemed confusing. Still, more study is warranted. In contrast, looking at many qq-plots corresponding to residual samples seemed straightforward.

An alternative to computing $P(|\epsilon_{ij}| > \sigma|Y)$ is to compute the posterior distribution of the one sided p-value, $p(1 - \Phi(|\epsilon_{ij}|\sigma^{-1})|Y)$. Meng (1994) and

effect	model	mean square	$O_\epsilon(2)$	$O_\epsilon(3)$	Simulation Results MS	$O_\epsilon(2)$	$O_\epsilon(3)$
linear							
	1	108.2	1	1	0	0	0
	2	.95	.038	.002	.300	.128	.020
	3	1.00	.049	.002	.332	.156	.056
	4	1.05	.057	.004	.320	.152	.020
quadratic							
	1	3.82	.362	0	.040	.044	.080
	2	5.71	.987	0	.004	.040	.192
	3	9.54	.999	.582	0	0	0
	4	.96	.044	.002	.344	.500	.072
cubic							
	1	.64	0	0	.472	.400	.080
	2	1.43	0	0	.192	.384	.040
	3	2.21	.010	0	.096	.352	.116
	4	.91	.031	.002	.356	.184	.012
quartic							
	1	14.8	1	1	0	0	0
	2	28.73	1	1	0	0	0
	3	47.79	1	1	0	0	0
	4	.99	.050	.005	.344	.556	.052

Table 4: Checks for needed polynomial effects. The first column indicates an effect linear, quadratic, cubic or quartic in time. The second column indicates model 1, 2, 3, or 4. The mean square is $E[(W^t\epsilon)^2/(W^tW\sigma^2)|Y]$, which should be approximately 1 if there is no missing predictor. The next two columns are $O_\epsilon(2) = O_\epsilon(W,2) = P(|W^t\epsilon| > 2(W^tW)^{1/2}\sigma|Y)$ and $O_\epsilon(3) = O_\epsilon(W,3) = P(|W^t\epsilon| > 3(W^tW)^{1/2}\sigma|Y)$. The last three columns are the fraction of times in 250 trials that random normal, independent predictors produced a statistic (mean square, $O_\epsilon(2)$, $O_\epsilon(3)$) with a larger value than the tabled value.

Gelman, Meng and Stern (1995) introduce contrasting Bayesian approach to testing that inspects the posterior distribution of the p-value. See also Vounatsou and Smith (1995) for analysis of residuals from discrete data.

References

Albert, J. H. and Chib S. (1993a), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.

Albert, J. H. and Chib S. (1993b), "Bayes Inference Via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts," *Journal of Business and Economic Statistics*, 11, 1-15.

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

Albert, J. H. and Chib S. (to appear), "Bayesian Residual Analysis for Binary Response Regression Models," *Biometrika.*

Beckman, R. J., Nachtsheim, C. J. and Cook, R. D. (1987), "Diagnostics for Mixed-Model Analysis of Variance," *Technometrics*, 29, 413-426.

Chaloner, K. (1991), "Bayesian Residual Analysis in the Presence of Censoring," *Biometrika*, 78, 637-644.

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.

Chaloner, K. and Brant R. (1988), "A Bayesian Approach to Outlier Detection and Residual Analysis," *Biometrika*, 75, 651-660.

Christensen, R., Pearson, L. M. and Johnson, W. (1992), "Case-Deletion Diagnostics for Mixed Models" *Technometrics*, 34, 38-45.

Cook, R. D., Hawkins, D. M. and Weisberg, S. (1992), "Comparison of Model Misspecification Diagnostics Using Residuals From Least Mean of Squares and Least Median of Squares" *Journal of the American Statistical Association*, 87, 419-424.

Dempster, A. P. and Ryan, L. M. (1985), "Weighted Normal Plots," *Journal of the American Statistical Association*, 80, 845-850.

Duncan, D. B. and Horn, S. D. (1972), "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," *Journal of the American Statistical Association*, 67, 815-821.

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," *Journal of the American Statistical Association*, 85, 972-985.

Gelman, A., Meng, X. L. and Stern, H. (in press), "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies (with discussion)," *Statistica Sinica.*

Gilks, W. R., Wang, C. C., Yvonnet, B., and Coursaget, P. (1993), "Random-Effects Models for Longitudinal Data Using Gibbs Sampling," *Biometrics*, 49, 441-453.

Hill, B. M. (1965), "Inference About Variance Components in the One-Way Model," *Journal of the American Statistical Association*, 60, 806-825.

Hocking, R. R., Green, J. W., and Bremer, R. H. (1989), "Variance-Component Estimation with Model-Based Diagnostics," *Technometrics*, 31, 227-239.

Hodges, J. (1994), "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics," University of Minnesota, Division of Biostatistics, Technical Report 94-009.

Lange, N. and Ryan, L. (1989), "Assessing Normality in Random Effects Models," *The Annals of Statistics*, 17, 624-642.

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," *Journal of the American Statistical Association*, 87, 615-625.

Lange, N., Carlin, B. P. and Gelfand, A. E. (1992b), Rejoinder, *Journal of the American Statistical Association*, 87, 631-632.

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.

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

Louis, T. A. (1988), "General Methods for Analyzing Repeated Measures," *Statistics in Medicine*, 7, 29-45.

Meng, X.L. (1994), "Posterior Predictive P-values" *Annals of Statistics*, 22, 1142-1160.

Moulton, B. R. (1987), "Diagnostics for Group Effects in Regression Analysis," *Journal of Business and Economic Statistics*, 5, 275-282.

Rutter, C. M. and Elashoff, R. M. (1995), "Analysis of Longitudinal Data: Random Coefficient Regression Modeling," *Statistics in Medicine*, to appear.

Seltzer, M. H. (1993), "Sensitivity Analysis for Fixed Effects in the Hierarchical Model: A Gibbs Sampling Approach," *Journal of Educational Statistics*, 18, 207-235.

Sharples, L. D. (1990), "Identification and Accommodation of Outliers in General Hierarchical Models," *Biometrika*, 77, 445-453.

Vounatsou, P. and Smith, A. F. M. (1995), "Bayesian Analysis of Ring-Recovery Data Via Markov Chain Monte Carlo Simulation," *Biometrics*, 51, 687-708.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. E. (1994), "Bayesian Analysis of Linear and Nonlinear Population Models by Using the Gibbs Sampler," *Applied Statistics*, 43, 201-221.

Waternaux, C., Laird, N. M. and Ware J. H. (1989), "Methods for Analysis of Longitudinal Data: Blood-Lead Concentrations and Cognitive Development," *Journal of the American Statistical Association*, 84, 33-41.

Weiss, R. E. and Lazaro, C. G. (1992), "Residual Plots for Repeated Measures," *Statistics in Medicine*, 11, 115-124.

Zeger, S. L. and Karim, M. R. (1991), "Generalized Linear Models with Random Effects," *Journal of the American Statistical Association*, 86, 79-86.

Zellner, A. (1975), "Bayesian Analysis of Regression Error Terms," *Journal of the American Statistical Association*, 70, 138-144.

Figure Captions

Figure 1. Parallel plot of the ϵ residual means for model 1. Case numbers for some outlying observations are identified. A decreasing trend across time is visible.

Figure 2. Parallel plot of the ϵ residual means for model 3. Case numbers for some outlying observations are identified in each plot. An '*M*' shape is visible across time.

Figure 3. Parallel plot of the ϵ residual means for model 4. Case numbers for some outlying observations are identified in each plot. No obvious time trend is visible.

Figure 4. Scatterplot of posterior means of the random effects from model 4 with cases 11 and 26 identified. There is some suggestion that heavier women, that is, those with larger individual intercepts, may have larger slopes than women with smaller and negative intercepts.

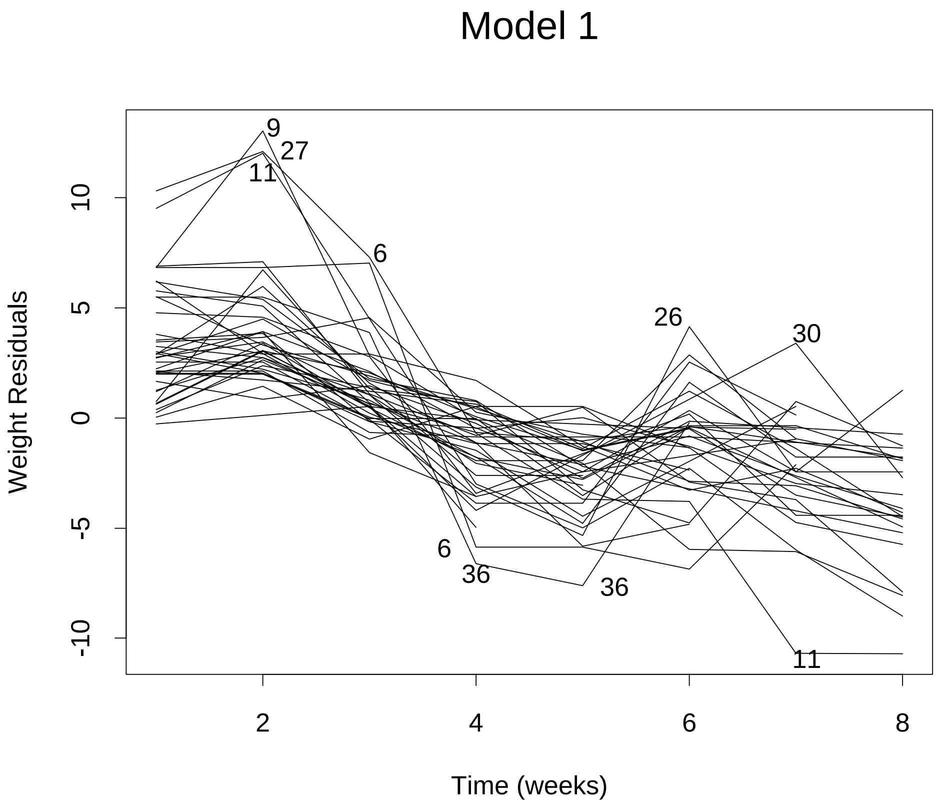
Model 1
Weight Residuals
10
5
0
-5
-10
9
27
11
6
26
30
6
36
36
11
2
4
6
8
Time (weeks)

Figure 1

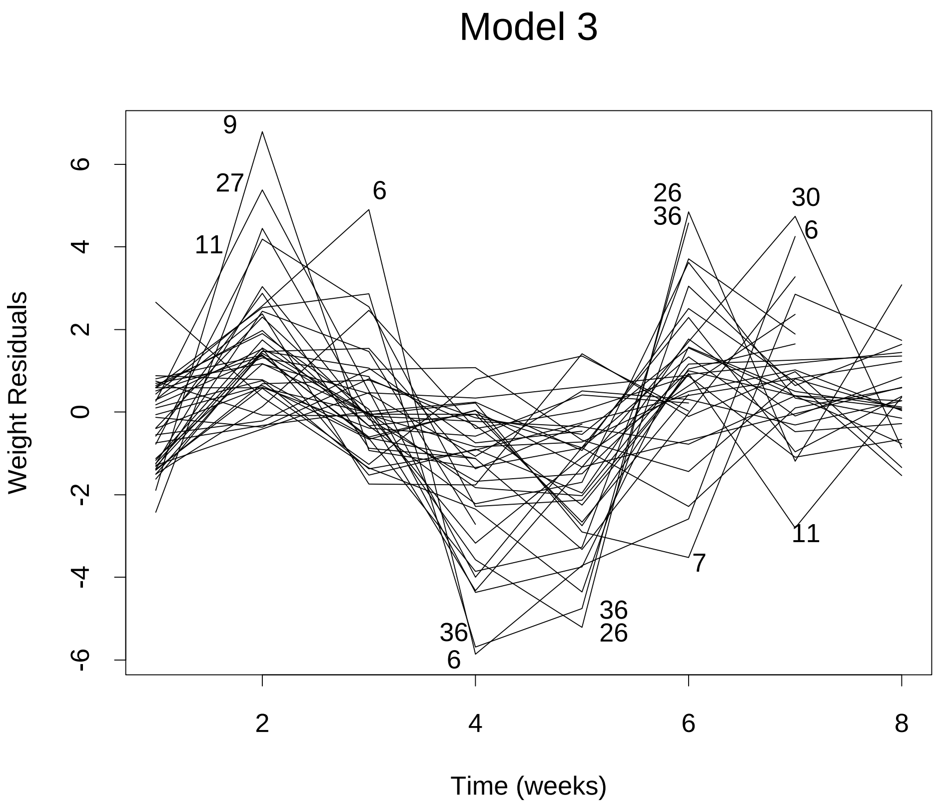
Model 3
Weight Residuals
Time (weeks)
9
27
11
6
26
36
30
6
11
7
36
26
36
6
-6
-4
-2
0
2
4
6
2
4
6
8

Figure 2

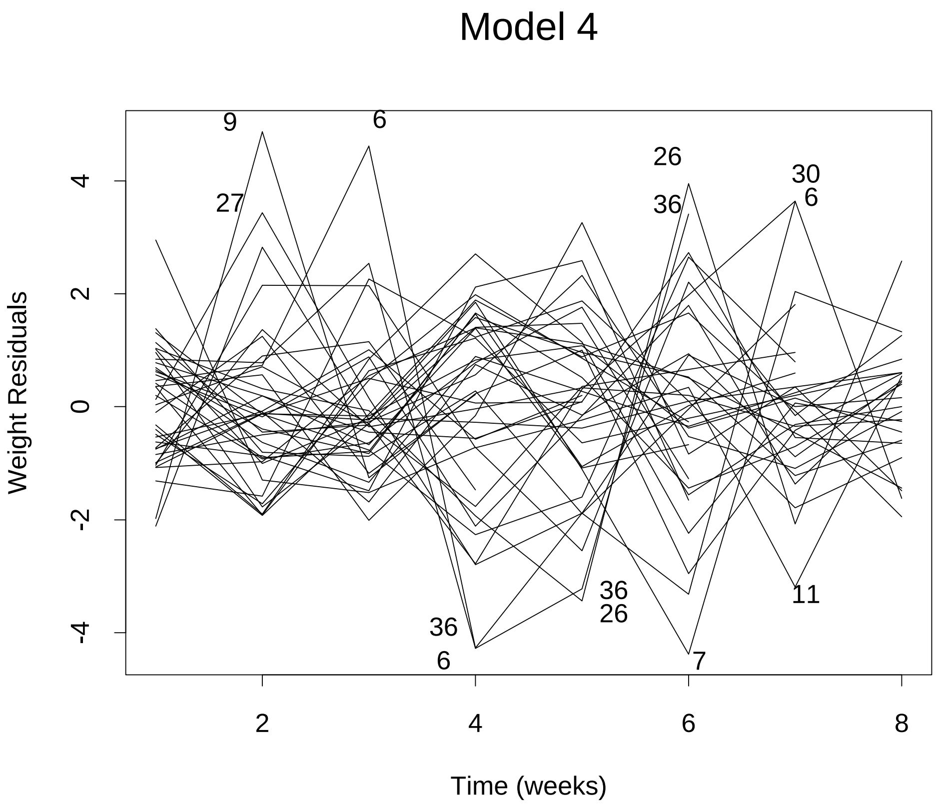
Model 4
Weight Residuals
Time (weeks)
-4
-2
0
2
4
2
4
6
8
9
6
27
26
36
30
6
36
26
36
6
7
11

Figure 3

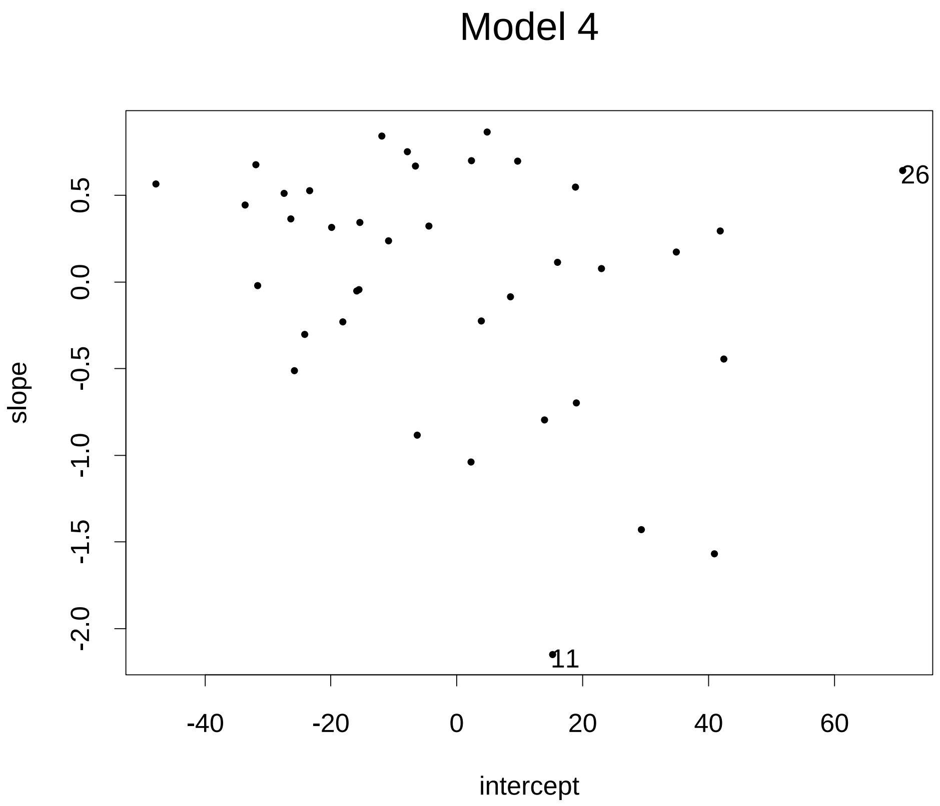

Figure 4