Paper 189-29

Mixed Model Influence Diagnostics

Oliver Schabenberger, SAS Institute Inc., Cary, NC

ABSTRACT

Linear models for uncorrelated data have well established measures to gauge the influence of one or more observations on the analysis. For such models, closed-form update expressions allow efficient computations without refitting the model. When similar notions of statistical influence are applied to mixed models, things are more complicated. Removing data points affects fixed effects and covariance parameter estimates. Update formulas for "leave-one-out" estimates typically fail to account for changes in covariance parameters. Moreover, in repeated measures or longitudinal studies, one is often interested in multivariate influence, rather than the impact of isolated points. This paper examines extensions of influence measures in linear mixed models and their implementation in the MIXED procedure.

INTRODUCTION

A statistical model, whether of the fixed-effects or mixed-effects variety, represents how you think your data were generated. Following model specification and estimation, it is of interest to explore the model-data agreement by raising questions such as

- Does the model-data agreement support the model assumptions?
- Should model components be refined, and if so, which components? For example, should regressors be added or removed, and is the covariation of the observations modeled properly?
- Are the results sensitive to model and/or data? Are individual data points or groups of cases particularly influential on the analysis?

In classical linear models, this examination of model-data agreement has traditionally revolved around

- the informal, graphical examination of estimates of model errors to assess the quality of distributional assumptions: *residual analysis*
- overall measures of *goodness-of-fit*
- the quantitative assessment of the inter-relationship of model components; for example, *collinearity diagnostics*
- the qualitative and quantitative assessment of influence of cases on the analysis: *influence analysis*.

The sensitivity of a model is studied through measures that express its stability under perturbations. You are not interested in a model that is either overly stable or overly sensitive. Changes in the data or model components should produce commensurate changes in the model output. The difficulty is to determine when the changes are substantive enough to warrant further investigation, possibly leading to a reformulation of the model or changes in the data (such as dropping outliers). This paper is primarily concerned with stability of linear mixed models to perturbations of the data; that is, with *influence analysis*. Broadly defined, "influence" is understood as the ability of a single or multiple data points, through their presence or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model. The goal of influence analysis is not primarily to mark data points for deletion so that a better model fit can be achieved for the reduced data, although this might be a result of influence analysis. The goal is rather to determine which cases are influential and the manner in which they are important to the analysis. Outliers, for example, may be the most noteworthy data points in an analysis. They can point to a model breakdown and lead to development of a better model.

In recent years, mixed models have become invaluable tools in the analysis of experimental and observational data. In these models, more than one term can be subject to random variation. Mixed model technology enables you to analyze complex experimental data with hierarchical random processes, temporal, longitudinal, and spatial data, to name just a few important applications. The general linear mixed model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is a $(n \times 1)$ vector of observed data, $\mathbf{X}$ is an $(n \times p)$ fixed-effects design or regressor matrix of rank k, $\mathbf{Z}$ is a $(n \times g)$ random-effects design or regressor matrix, $\boldsymbol{\gamma}$ is a $(g \times 1)$ vector of random effects, and $\boldsymbol{\epsilon}$ is an $(n \times 1)$ vector of model errors (also random effects). The distributional assumptions made by the MIXED procedure are as follows: $\boldsymbol{\gamma}$ is normal with mean **0** and variance $\mathbf{G}$; $\boldsymbol{\epsilon}$ is normal with mean **0** and variance $\mathbf{R}$; the random components $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independent. Parameters of this model are the fixed-effects $\boldsymbol{\beta}$ and all unknowns in the variance matrices $\mathbf{G}$ and $\mathbf{R}$. The unknown variance elements are referred to as the *covariance parameters* and collected in the vector $\boldsymbol{\theta}$.

The concept of critiquing the model-data agreement applies in mixed models in the same way as in linear fixed-effects models. In fact, because of the more complex model structure, you can argue that model and data diagnostics are even more important. For example, you are not only concerned with capturing the important variables in the model. You are also concerned with "distributing" them correctly between the fixed and random components of the model. The mixed model structure presents unique and interesting challenges that prompt us to reexamine the traditional ideas of influence and residual analysis.

This paper presents the extension of traditional tools and statistical measures for influence and residual analysis to the linear mixed model and demonstrates their implementation in the MIXED procedure (experimental features in SAS 9.1). The remainder of this paper is organized as follows. The "Background" section briefly discusses some mixed model estimation theory and the challenges to model diagnosis that result from it. The diagnostics implemented in the MIXED procedure are discussed in the "Residual Diagnostics in the MIXED Procedure" section (page 3) and the "Influence Diagnostics in the MIXED Procedure" section (page 5). The syntax options and suboptions you use to request the various diagnostics are briefly sketched in the "Syntax" section (page 9). The presentation concludes with an example.

BACKGROUND

PARAMETER ESTIMATION

In the MIXED procedure, estimation of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ typically proceeds as follows. The restricted log likelihood of the data $\mathbf{Y}$ is formed, given the fixed-effects matrix $\mathbf{X}$. The fixed-effects are profiled from the residual log likelihood and if possible, a residual variance parameter σ^2 is also profiled. Minus twice the resulting function is the objective function for minimization via a ridge-stabilized Newton-Raphson algorithm. This setup has important implications. Because the objective function usually depends nonlinearly on the covariance parameters $\boldsymbol{\theta}$, finding a solution to the (restricted) likelihood problem is an iterative process. Because $\boldsymbol{\beta}$ (and σ^2) have been profiled from the optimization, the iterative optimization involves only the covariance parameters. On convergence, estimates of the fixed effects and predictions of the random effects $\boldsymbol{\gamma}$ are obtained by solving the mixed model equations. These quantities are

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{Y} \\
\widehat{\boldsymbol{\gamma}} &= \mathbf{G}(\widehat{\boldsymbol{\theta}})\mathbf{Z}'\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)
\end{aligned}$$

where $\mathbf{V}(\boldsymbol{\theta}) = \text{var}[\mathbf{Y}] = \mathbf{Z}\mathbf{G}(\boldsymbol{\theta})\mathbf{Z}' + \mathbf{R}(\boldsymbol{\theta})$ is the marginal variance of the data.

If there are no random effects $\mathbf{Z}$ and $\mathbf{R} = \sigma^2\mathbf{I}$, the general linear mixed model reduces to a classical linear model that can be fit with the GLM procedure. In that case, residual and influence measures are well known. For example, you can easily compute studentized residuals, Cook's D, DFBETAs and so forth; refer to the

"Influence Diagnostics" section in Chapter 61, "The REG Procedure" (*SAS/STAT User's Guide*), for details and interpretation. Why, then, does it take such extra effort to apply these ideas in the mixed model?

DISTINCTION FROM LINEAR MODELS

The differences between perturbation and residual analysis in the linear model and the linear mixed model are connected to the important facts that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ depend on the estimates of the covariance parameters, that $\widehat{\boldsymbol{\beta}}$ has the form of an (estimated) generalized least squares (GLS) estimator, and that $\boldsymbol{\gamma}$ is a random vector.

- In a mixed model, you can consider the data in a conditional and an unconditional sense. If you imagine a particular realization of the random effects, then you are considering the conditional distribution $\mathbf{Y}|\boldsymbol{\gamma}$. If you are interested in quantities averaged over all possible values of the random effects, then you are interested in $\mathbf{Y}$; this is called the marginal formulation. In a clinical trial, for example, you may be interested in drug efficacy for a particular patient. If random effects vary by patient, that is a conditional problem. If you are interested in the drug efficacy in the population of all patients, you are using a marginal formulation. Correspondingly, there will be conditional and marginal residuals, for example.
- The estimates of the fixed effects $\boldsymbol{\beta}$ depend on the estimates of the covariance parameters. If you are interested in determining the influence of an observation on the analysis, you must determine whether this is influence on the fixed effects for a given value of the covariance parameters, influence on the covariance parameters, or influence on both.
- Mixed models are often used to analyze repeated measures and longitudinal data. The natural experimental or sampling unit in those studies is the entity that is repeatedly observed, rather than each individual repeated observation. For example, you may be analyzing monthly purchase records by customer. An influential "data point" is then not necessarily a single purchase. You are probably more interested in determining the influential customer. This requires that you can measure the influence of sets of observations on the analysis, not just influence of individual observations.
- The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of $\boldsymbol{\beta}$ and σ^2, which exclude the ith observation, can be computed without re-fitting the model. Such update formulas are available in the mixed model only if you assume that the covariance parameters are not affected by the removal of the observation in question. This is rarely a reasonable assumption.
- The application of well-known concepts in model-data diagnostics to the mixed model can produce results that are at first counter-intuitive, since our understanding is steeped in the ordinary least squares (OLS) framework. As a consequence, we need to revisit these important concepts, ask whether they are "portable" to the mixed model, and gain new appreciation for their changed properties. An important example is the ostensibly simple concept of leverage. The definition of leverage adopted by the MIXED procedure can, in some instances, produce negative values, which are mathematically impossible in OLS. Other measures that have been proposed may be non-negative, but trade other advantages. Another example are properties of residuals. While OLS residuals necessarily sum to zero in any model (with intercept), this not true of the residuals in many mixed models.

RESIDUAL DIAGNOSTICS IN THE MIXED PROCEDURE

A residual is the difference between an observed quantity and its estimated or predicted value. In the mixed model you can distinguish marginal residuals r_m and conditional residuals r_c. A *marginal residual* is the difference between the observed data and the estimated (marginal) mean,

$$r_{mi} = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$$

A *conditional residual* is the difference between the observed data and the predicted value of the observation,

$$r_{ci} = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} - \mathbf{z}_i'\widehat{\boldsymbol{\gamma}}$$

In a model without random effects $\boldsymbol{\gamma}$, the two sets of residuals coincide. The name conditional residual stems from the fact that $\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}$ is the conditional mean of y_i.

Residuals are used to examine model assumptions and to detect outliers and potentially influential data point. The raw residuals r_{mi} and r_{ci} are usually not well suited for these purposes. Even if the true model errors are uncorrelated and have equal variance, the residuals will exhibit correlations and their variances will differ. The interpretation of raw residuals is further made difficult if the variances of the observations differ. A data point with a smaller raw residual may be more troublesome than a data point with a large residual, if the variance of the former observation is less. To account for the unequal variance of the residuals, various studentizations are applied.

A random variable is said to be *standardized* if the difference from its mean is scaled by its standard deviation. The residuals above have mean zero but their variance is unknown, it depends on the true values of $\boldsymbol{\theta}$. Standardization is thus not possible in practice. Instead, you can compute *studentized* residuals by dividing a residual by an estimate of its standard deviation. If that estimate is independent of the ith observation, the process is termed *external studentization*. This is usually accomplished by excluding the ith observation when computing the estimate of its standard error. If the observation contributes to the standard error computation, the residual is said to be *internally studentized*.

A residual can be divided by a standard error that is not necessarily its own. For example, dividing $y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ by the estimated standard deviation of Y_i is not a studentization, because var$[Y_i]$ is not the variance of r_{mi}. The divisor is "proper," however, if the variability of $\widehat{\boldsymbol{\beta}}$ is ignored. Such residuals are called *Pearson-type* residuals. Rather than divide each individual residual by the variance of an observation, you can also consider the vector of residuals and the estimated variance $\mathbf{V}(\widehat{\boldsymbol{\theta}})$. Let $\widehat{\mathbf{C}}$ denote a matrix such that $\widehat{\mathbf{C}}\widehat{\mathbf{C}}' = \mathbf{V}(\widehat{\boldsymbol{\theta}})$, for example, its lower-triangular Cholesky root. Then the *scaled residuals* $\mathbf{r}_c = \widehat{\mathbf{C}}^{-1}\mathbf{r}_m$ have zero mean and are approximately uncorrelated. They are not exactly uncorrelated, because $\widehat{\mathbf{C}}$ is an estimated matrix and $\mathbf{V}$ is not the variance of $\mathbf{r}_m$. Scaled residuals can be useful to diagnose whether the covariance structure of the mixed model has been specified correctly.

The MIXED procedure produces raw residuals with the OUTP= and OUTPM= options of the MODEL statements. Starting with SAS 9.1, the procedure also computes studentized, Pearson, and scaled residuals. These residuals can be written to the OUTP= or OUTPM= data sets or graphed with the ODS Graphics facility (experimental in SAS 9.1).

The following table summarizes the available residuals. Expressions for the variances of $\mathbf{r}_m$ and $\mathbf{r}_c$ can be found in Chapter 46, "The MIXED Procedure" (*SAS/STAT User's Guide*).

Table 1. Residuals in PROC MIXED

Type of Residual	Marginal	Conditional
Raw	$r_{mi} = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$	$r_{ci} = r_{mi} - \mathbf{z}_i'\widehat{\boldsymbol{\gamma}}$
Studentized	$r_{mi}^{student} = \frac{r_{mi}}{\sqrt{\widehat{\text{var}}[r_{mi}]}}$	$r_{ci}^{student} = \frac{r_{ci}}{\sqrt{\widehat{\text{var}}[r_{ci}]}}$
Pearson	$r_{mi}^{pearson} = \frac{r_{mi}}{\sqrt{\widehat{\text{var}}[Y_i]}}$	$r_{ci}^{pearson} = \frac{r_{ci}}{\sqrt{\widehat{\text{var}}[Y_i\|\boldsymbol{\gamma}]}}$
Scaled	$\widehat{\mathbf{C}}^{-1}\mathbf{r}_m$	

INFLUENCE DIAGNOSTICS IN THE MIXED PROCEDURE

Key to the implementations of influence diagnostics in the MIXED procedure is the attempt to quantify influence, where possible, by drawing on the basic definitions of the various statistics in the classical linear model. On occasion, quantification is not possible. Assume, for example, that a data point is removed and the new estimate of the $\mathbf{G}$ matrix is not positive definite. This may occur if a variance component estimate now falls on the boundary of the parameter space. Thus, it may not be possible to compute certain influence statistics comparing the full-data and reduced-data parameter estimates. However, knowing that a new singularity was encountered is important qualitative information about the data point's influence on the analysis.

The basic procedure for quantifying influence is simple:

1. Fit the model to the data and obtain estimates of all parameters.
2. Remove one or more data points from the analysis and compute updated estimates of model parameters.
3. Based on full- and reduced-data estimates, contrast quantities of interest to determine how the absence of the observations changes the analysis.

We use the subscript (U) to denote quantities obtained without the observations in the set U. For example, $\widehat{\boldsymbol{\beta}}_{(U)}$ denotes the fixed-effects "leave-U-out" estimates. Note that the set U can contain multiple observations.

The influence statistics computed by the MIXED procedure can be coarsely classified as follows.

OVERALL INFLUENCE

An overall influence statistic measures the change in the objective function being minimized. For example, in OLS regression, the residual sums of squares serves that purpose. In linear mixed models fit by maximum likelihood (ML) or restricted maximum likelihood (REML), an overall influence measure is the likelihood distance (Cook and Weisberg 1982, Ch. 5.2). Beckman, Nachtsheim, and Cook (1987) also refer to it as likelihood displacement. The idea is simple. Compute the full data parameter estimates $\widehat{\boldsymbol{\psi}}$ and estimates based on the reduced data, $\widehat{\boldsymbol{\psi}}_{(U)}$. Then the likelihood and restricted likelihood distances are obtained as

$$\begin{aligned} LD_{(U)} &= 2\left\{l(\widehat{\boldsymbol{\psi}}) - l(\widehat{\boldsymbol{\psi}}_{(U)})\right\} \\ RLD_{(U)} &= 2\left\{l_R(\widehat{\boldsymbol{\psi}}) - l_R(\widehat{\boldsymbol{\psi}}_{(U)})\right\} \end{aligned}$$

The likelihood distance gives the amount by which the log-likelihood of the full data changes if one were to evaluate it at the reduced-data estimates. The important point is that $l(\widehat{\boldsymbol{\psi}}_{(U)})$ is not the log-likelihood obtained by fitting the model to the reduced data set. It is obtained by evaluating the likelihood function based on the full data set (containing all n observations) at the reduced-data estimates.

The likelihood distance is a global, summary measure, expressing the joint influence of the observations in the set U on all parameters in $\boldsymbol{\psi}$ that were subject to updating.

If the global measure suggests that the points in U are influential, you should next determine the nature of that influence. In particular, the points can affect

- the estimates of fixed effects
- the estimates of the precision of the fixed effects
- the estimates of the covariance parameters
- the estimates of the precision of the covariance parameters
- fitted and predicted values

It is important to further decompose the initial finding to determine whether data points are actually troublesome. Simply because they are influential "somehow", should not trigger their removal from the analysis or a change in the model. For example, if points primarily affect the precision of the covariance parameters without exerting much influence on the fixed effects, then their presence in the data may not distort hypothesis tests or confidence intervals about $\boldsymbol{\beta}$. They will only do so if your inference depends on an estimate of the precision of the covariance parameter estimates, as is the case for the Satterthwaite and Kenward-Roger degrees of freedom methods and the standard error adjustment associated with the DDFM=KR option.

CHANGE IN PARAMETER ESTIMATES

Because the number of fixed-effects and covariance parameters can be large, the MIXED procedure enables you to compute summary statistics that capture the change in the entire parameter vector. These quadratic forms are based on Cook's D (Cook 1977) and the multivariate DFFITS statistic, (MDFFITS; Belsley, Kuh, and Welsch 1980, p. 32). The primary difference between D and $MDFFITS$ is that the latter uses an externalized estimate of the variance of the parameter estimates, while Cook's D does not. For the fixed effects, the two statistics are

$$\begin{aligned} D(\boldsymbol{\beta}) &= (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(U)})' \widehat{\mathrm{var}}[\widehat{\boldsymbol{\beta}}]^{-} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(U)}) / \mathrm{rank}(\mathbf{X}) \\ \mathrm{MDFFITS}(\boldsymbol{\beta}) &= (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(U)})' \widehat{\mathrm{var}}[\widehat{\boldsymbol{\beta}}_{(U)}]^{-} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(U)}) / \mathrm{rank}(\mathbf{X}) \end{aligned}$$

For both statistics, you are concerned about large values, indicating that the change in the parameter estimate is large relative to the variability of the estimate.

If the covariance parameters are updated during influence analysis, similar statistics can be computed for $\widehat{\boldsymbol{\theta}}$. However, the $D(\boldsymbol{\theta})$ and MDFFITS$(\boldsymbol{\theta})$ statistics do not involve division by a matrix rank.

CHANGE IN PRECISION OF ESTIMATES

The effect on the precision of estimates is separate from the effect on the point estimates. Data points that have a small Cook's D, for example, can still greatly affect hypothesis tests and confidence intervals, if their influence on the precision of the estimates is large. Following Christensen, Pearson, and Johnson (1992), the MIXED procedure computes functions of the trace and determinants of the variance matrices based on the full-data and the reduced-data estimates:

$$\begin{aligned} \mathrm{COVTRACE}(\boldsymbol{\beta}) &= \left| \mathrm{trace}(\widehat{\mathrm{var}}[\widehat{\boldsymbol{\beta}}]^{-} \ \widehat{\mathrm{var}}[\widehat{\boldsymbol{\beta}}_{(U)}]) - \mathrm{rank}(\mathbf{X}) \right| \\ \mathrm{COVRATIO}(\boldsymbol{\beta}) &= \frac{\det_{ns}(\widehat{\mathrm{var}}[\widehat{\boldsymbol{\beta}}_{(U)}])}{\det_{ns}(\widehat{\mathrm{var}}[\widehat{\boldsymbol{\beta}}])} \end{aligned}$$

where $\det_{ns}(\mathbf{M})$ denotes the determinant of the nonsingular part of matrix $\mathbf{M}$.

The benchmarks of "no influence" are zero for the covariance trace and one for the covariance ratio. If the influence analysis updates the covariance parameters, the MIXED procedure computes similar statistics for $\boldsymbol{\theta}$:

$$\begin{aligned} \text{COVTRACE}(\boldsymbol{\theta}) &= \left|\text{trace}(\widehat{\text{var}}[\widehat{\boldsymbol{\theta}}]^{-}\ \widehat{\text{var}}[\widehat{\boldsymbol{\theta}}_{(U)}]) - q\right| \\ \text{COVRATIO}(\boldsymbol{\theta}) &= \frac{\det_{ns}(\widehat{\text{var}}[\widehat{\boldsymbol{\theta}}_{(U)}])}{\det_{ns}(\widehat{\text{var}}[\widehat{\boldsymbol{\theta}}])} \end{aligned}$$

where q denotes the rank of $\text{var}[\widehat{\boldsymbol{\theta}}]$. The variance matrix that is used in the computation of COVTRACE and COVRATIO for covariance parameters is obtained from the inverse Hessian matrix. You can request a listing of this matrix with the ASYCOV option of the PROC MIXED statement.

EFFECT ON FITTED AND PREDICTED VALUES

The MIXED procedure computes the following statistics to measure influence on fitted and predicted values. The PRESS residual (Allen 1974) is the difference between the observed value and the predicted (marginal) mean, where the predicted value is obtained without the observations in question. Formally,

$$\widehat{e}_{i(U)} = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{(U)}$$

If you compute the influence of individual observations, the MIXED procedure reports these PRESS residuals. When removing sets of observations, PROC MIXED computes the PRESS statistic. This statistic is the sum of the squared PRESS residuals in a deletion set,

$$\text{PRESS}_{(U)} = \sum_{i \in U} \widehat{e}_{i(U)}$$

The effect of observations on fitted values can be measured by the DFFITS statistic of Belsley, Kuh, and Welsch (1980, p. 15). A DFFIT measures the change in predicted values due to removal of a single data point. If this change is standardized by the externally estimated standard error of the predicted value in the full data, you obtain the DFFITS statistic:

$$\text{DFFITS}_i = (\widehat{y}_i - \widehat{y}_{i(u)})/\text{ese}(\widehat{y}_i)$$

OUTLIER CHARACTERISTICS

Internally and externally studentized residuals are excellent statistics for detecting unusual observations. In OLS models with normal errors, externally studentized residuals are t distributed and a yardstick of ± 2 is common in their interpretation. This distributional result does not apply in general for externally studentized residuals in mixed models, particularly when the external studentization involves only an update of the residual variance. But because the studentized mixed model residuals have approximately unit variance, the ± 2 yardstick remains useful.

Another statistic often used to determine the influence potential of data points is the leverage. In the classical linear model, leverages are the diagonal elements h_{ii} of the so-called "Hat" matrix,

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$$

$\mathbf{H}$ is a projector, that is, it is symmetric and idempotent. Its properties are thus easily established. If $\mathbf{X}$ contains an intercept, then $1/n \leq h_{ii} \leq 1$ and $\text{trace}(\mathbf{H}) = \text{rank}(\mathbf{X})$. If you think of $\mathbf{H}$ as the gradient of the fitted values with respect to the data,

$$\mathbf{H} = \frac{\partial \widehat{\mathbf{Y}}}{\partial \mathbf{Y}}$$

then the corresponding expression in the linear mixed model is

$$\mathbf{H}_1 = \mathbf{X}\left(\mathbf{X}'\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}$$

The subscript $_1$ is used to signal that this is not the only possible "leverage" matrix. While it serves the definition in terms of a gradient, it is not a true projector. Although idempotent, $\mathbf{H}_1$ is not symmetric; it is an oblique projector (Christensen 1991).

A true projector can be constructed using the Cholesky decomposition $\mathbf{C}\mathbf{C}' = \mathbf{V}(\widehat{\boldsymbol{\theta}})$, namely

$$\mathbf{H}_2 = \mathbf{C}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{C}'^{-1}$$

Both mixed model leverages have in common that trace$(\mathbf{H}_1) =$ trace$(\mathbf{H}_2) =$ rank$(\mathbf{X})$. Their lower and upper bounds differ, however. For example, the diagonal elements of $\mathbf{H}_1$ can be negative. It appears strange at first to consider a negative "leverage." You can observe negative values in models where data points are highly correlated. Surrounding observations have contributed so much toward the estimate of the mean of y_i, that the observation itself receives a negative weight (screening effect). The projector $\mathbf{H}_2$ is preferable on these grounds, but it has an important shortcoming. Let $\mathbf{X}^* = \mathbf{C}^{-1}\mathbf{X}$. Then it is easy to see that $\mathbf{H}_2$ is the OLS hat matrix

$$\mathbf{H}_2 = \mathbf{X}^*\left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-}\mathbf{X}^{*\prime}$$

If you think of leverage values to express how unusual an observation is in the regressor space, then $\mathbf{H}_2$ does not support this interpretation. It measures how unusual an observation is in the transformed space of the regressors x^*. Also, even if $\mathbf{X}$ contains an intercept, $\mathbf{X}^*$ does not, and thus $1/n$ is not the lower bound for the diagonal elements of $\mathbf{H}_2$. Other leverages have been proposed for correlated error models. For example, Martin (1992) considers the complementary value $\mathbf{V}(\boldsymbol{\theta})^{-1}(\mathbf{I} - \mathbf{H}_1)$ as the generalization of leverage to dependent data. However, the elements of this matrix depend on the scaling of the data.

ITERATIVE VS. NONITERATIVE INFLUENCE ANALYSIS

While the basic idea of influence analysis is straightforward, the implementation in mixed models can be tricky. For example, update formulas for the fixed effects are available only when the covariance parameters are assumed to be known. At most the profiled residual variance can be updated without refitting the model.

A measure of total influence requires updates of all model parameters, and the only way that this can be achieved in general is by removing the observations in question and refitting the model. Because this "brute-force" method involves iterative reestimation of the covariance parameters, it is termed *iterative influence analysis*. Reliance on closed-form update formulas for the fixed effects without updating the (un-profiled) covariance parameters is termed a *noniterative* influence analysis.

An iterative analysis seems like a costly, computationally intensive enterprise. If you compute iterative influence diagnostics for all n observations, then a total of $n + 1$ mixed models are fit iteratively. This does not imply, of course, that the procedure's execution time increases n-fold. Keep in mind that

- iterative reestimation always starts at the converged full-data estimates. If a data point is not influential, then its removal will have little effect on the objective function and parameter estimates. Within one or two iterations, the process should arrive at the reduced-data estimates.
- if complete reestimation does require many iterations, then this is important information in itself. The likelihood surface has probably changed drastically, and the reduced-data estimates are moving away from the full-data estimates.

- without reestimation of all parameters, the full import of the observations in the deletion set U can not be ascertained. Also, influence statistics for θ are not available unless the covariance parameters are updated. A one-step update of the covariance parameters may be a good compromise to prevent long run times and to allow the covariance parameter estimates to react to the data perturbation.

The MIXED procedure is capable of both modes for computing influence statistics. By default, the analysis is non-iterative. Fixed effects are automatically updated. Only a profiled residual variance is updated among the covariance parameters, if the influence statistic requires an external estimate. Influence diagnostics for the covariance parameters are only available in iterative mode.

SYNTAX

This section briefly sketches the PROC MIXED syntax with which you can request the experimental residual and influence diagnostics in SAS 9.1. The experimental status of these features means that both their appearance and their syntax are subject to change in a future release. Please refer to Chapter 46, "The MIXED Procedure" (*SAS/STAT User's Guide*), for greater detail on the options and suboptions associated with residual and influence diagnostics in the MIXED procedure. The section below presents excerpts from that chapter.

Influence diagnostics are requested with the INFLUENCE option of the MODEL statement in PROC MIXED. The suboptions of the INFLUENCE option control the type of influence analysis, for example, iterative or noniterative, point or set deletion, etc. Raw conditional and marginal residuals are computed automatically by the MIXED procedure when the OUTP= or OUTPM= options of the MODEL statement are specified. Starting in SAS 9.1, if you also specify the RESIDUAL option of the MODEL statement, (internally) studentized and Pearson residuals are added to the OUTP= and/or OUTPM= data sets. Externally studentized residuals can be obtained as part of the influence diagnostics.

The results of influence and residual analysis can be displayed with the ODS Graphics facility, an experimental extension to the Output Delivery System, that produces quality graphics as automatically as tables. Please refer to Chapter 15, "Statistical Graphics Using ODS" (*SAS/STAT User's Guide*), in the SAS 9.1 documentation for detailed information about the ODS Graphics facility and to the section "ODS Graphics" (Chapter 46, *SAS/STAT User's Guide*) in the MIXED SAS 9.1 documentation for details specific to graphics implementation in PROC MIXED.

INFLUENCE<(<EFFECT=*effect***>**
<ESTIMATES|EST>
<ITER=*number***>**
<KEEP=*number***>**
<SELECT=*value-list***>**
<SIZE=*number***>)>**

The INFLUENCE option of the MODEL statement in the MIXED procedure computes influence diagnostics by noniterative or iterative methods. The noniterative diagnostics rely on recomputation formulas under the assumption that covariance parameters or their ratios remain fixed. With the possible exception of a profiled residual variance, no covariance parameters are updated. This is the default behavior because of its computational efficiency.

Without further suboptions, PROC MIXED computes single-case deletion diagnostics and influence statistics for each observation in the data set by updating estimates for the fixed effects parameter estimates, and also the residual variance, if it is profiled. The EFFECT=, SELECT=, ITER=, SIZE=, and KEEP= suboptions provide additional flexibility in the computation and reporting of influence statistics.

Table 2. Suboptions of INFLUENCE Option

Description	Suboption
Compute influence diagnostics for individual observations	default
Measure influence of sets of observations chosen according to a classification variable or effect	EFFECT=
Remove pairs of observations and report the results sorted by degree of influence	SIZE=2
Remove triples, quadruples of observations,...	SIZE=
Allow selection of individual observations, observations sharing specific levels of effects, and construction of tuples from specified subsets of observations	SELECT=
Update fixed effects and covariance parameters by refitting the mixed model, adding up to n iterations	ITER=$n > 0$
Compute influence diagnostics for the covariance parameters	ITER=$n > 0$
Update only fixed effects and the residual variance, if it is profiled	ITER=0
Add the reduced-data estimates to the data set created with ODS OUTPUT	ESTIMATES

EFFECT=*effect* specifies an effect according to which observations are grouped. Observations sharing the same level of the *effect* are removed from the analysis as a group. The *effect* must contain only class variables, but need not be contained in the model.

ESTIMATES|EST specifies that the updated parameter estimates should be written to the ODS output data set. The values are not displayed in the "Influence" table, but if you use ODS OUTPUT to create a data set from the listing, the estimates are added to the data set. If ITER=0, only the fixed effects estimates are saved. In iterative influence analyses, fixed effects and covariance parameters are stored.

ITER=*n* controls the maximum number of additional iterations PROC MIXED performs to update the fixed effects and covariance parameter estimates following data point removal. When $n > 0$, the influence analysis is iterative. The default is $n = 0$. If $n > 0$ and METHOD=REML (default) or METHOD=ML, the procedure updates fixed effects *and* variance-covariance parameters after removing the selected observations with additional Newton-Raphson iterations, starting from the converged estimates for the entire data. The process stops for each observation or set of observations if the convergence criterion is satisfied or the number of further iterations exceeds *n*.

KEEP=*n* determines how many observations are retained for display and in the output data set or how many tuples if you specify SIZE=. The output is sorted by an influence statistic as discussed for the SIZE= suboption.

SELECT = *value-list* specifies which observations or effect levels are chosen for influence calculations. If SELECT= is not specified, diagnostics are computed for all possible sets, that is

- all observations, if EFFECT= or SIZE= are not given
- all levels of the specified effect, if EFFECT= is specified
- all tuples of size k formed from the observations in *value-list*, if SIZE=k is specified

SIZE= *n* instructs PROC MIXED to remove groups of observations formed as tuples of size *n*. For example, SIZE=2 specifies all $n \times (n-1)/2$ unique pairs of observations. The number of tuples for SIZE=k is $n!/(k!(n-k)!)$ and grows quickly with n and k. Using the SIZE= option

can result in considerable computing time. The MIXED procedure displays by default only the 50 tuples with the greatest influence. You can use the KEEP= option to override this default and to retain a different number of tuples in the listing or ODS output data set.

RESIDUAL

requests that Pearson-type and (internally) studentized residuals be added to the OUTP= and OUTPM= data sets. The RESIDUAL option adds the variables PearsonResid and StudentResid to the OUTP= and OUTPM= data sets.

The option has no effect unless the OUTP= or OUTPM= option are specified or you request statistical graphics with the experimental ODS GRAPHICS statement.

EXAMPLE: INFLUENCE ANALYSIS FOR REPEATED MEASURES

This example uses data from Pothoff and Roy (1964) on repeated growth measurements of 11 girls and 16 boys. The measurements were taken at ages 8, 10, 12, and 14. The following data step creates the data set Pr.

```
data pr;
   input Person Gender $ y1 y2 y3 y4;
   y=y1; Age=8;  output;
   y=y2; Age=10; output;
   y=y3; Age=12; output;
   y=y4; Age=14; output;
   drop y1-y4;
   datalines;
 1   F   21.0   20.0   21.5   23.0
 2   F   21.0   21.5   24.0   25.5
 3   F   20.5   24.0   24.5   26.0
 4   F   23.5   24.5   25.0   26.5
 5   F   21.5   23.0   22.5   23.5
 6   F   20.0   21.0   21.0   22.5
 7   F   21.5   22.5   23.0   25.0
 8   F   23.0   23.0   23.5   24.0
 9   F   20.0   21.0   22.0   21.5
10   F   16.5   19.0   19.0   19.5
11   F   24.5   25.0   28.0   28.0
12   M   26.0   25.0   29.0   31.0
13   M   21.5   22.5   23.0   26.5
14   M   23.0   22.5   24.0   27.5
15   M   25.5   27.5   26.5   27.0
16   M   20.0   23.5   22.5   26.0
17   M   24.5   25.5   27.0   28.5
18   M   22.0   22.0   24.5   26.5
19   M   24.0   21.5   24.5   25.5
20   M   23.0   20.5   31.0   26.0
21   M   27.5   28.0   31.0   31.5
22   M   23.0   23.0   23.5   25.0
23   M   21.5   23.5   24.0   28.0
24   M   17.0   24.5   26.0   29.5
25   M   22.5   25.5   25.5   26.0
26   M   23.0   24.5   26.0   30.0
27   M   22.0   21.5   23.5   25.0
;
```

These data are also analyzed in the MIXED documentation. Here, we perform an iterative influence analysis using the ODS Graphics facility to display the results. The model we consider is one without random

effects, where the correlations among the repeated measurements for each child follow a first-order autoregressive structure. The fixed-effects portion of the model consists of gender, age effect, and their interaction. Because of the clustered data structure, it is of interest to study the influence of clusters (children) on the analysis, rather than the influence of individual observations. A cluster comprises the repeated measurements for each child (Person).

The following statements request the iterative influence analysis and display tabular and graphical output.

```
ods html;
ods graphics off;

proc mixed data=pr;
  class Person Gender;
  model y = Gender Age Gender*Age /
            influence(iter=5 effect=Person est);
  repeated / type=ar(1) subject=Person;
run;

ods graphics off;
ods html close;
```

The graphical displays are requested by specifying the experimental ODS GRAPHICS statement. The ITER=5 suboption of the INFLUENCE option requests the iterative analysis. For each deletion set, covariance parameters are updated up to five times. The deletion sets are defined through the EFFECT= suboption. All observations that have the same level as the Person classification variable comprise a set. This will remove the four observations of each child in turn. The EST option requests graphical displays of the reduced-data parameter estimates for the fixed effects and the covariance parameters.

The “Influence Diagnostics” table shows the results in tabular form. Clearly, child 20 stands out in terms of the summary statistic restricted likelihood distance (RLD). Note that the data for this boy show a marked increase at age 12, followed by a decrease of the growth measurement.

The influence of these four observations on the fixed effects is not that large, however. In fact, this deletion set has one of the smallest Cook’s D statistics. The import of the observations for this child on the variance of the fixed effects is also relatively small. These four observations exert influence primarily on the estimates of the covariance parameters and their precision, as can be seen from the Cook’s D, MDFFITS, and COVRATIO statistic for the covariance parameters.

The Mixed Procedure

Influence Diagnostics for Levels of Person

Person	Number of Observations in Level	Iterations	PRESS Statistic	Cook's D	MDFFITS	COVRATIO
1	4	1	9.6412	0.01119	0.00992	1.3415
2	4	1	3.7181	0.01154	0.01016	1.3815
3	4	1	10.8479	0.02907	0.02591	1.3174
4	4	2	24.4013	0.04667	0.04347	1.1961
5	4	1	1.6900	0.00334	0.00294	1.4051
6	4	1	11.7185	0.01981	0.01776	1.3158
7	4	1	1.2187	0.00307	0.00269	1.4069
8	4	1	5.0250	0.01807	0.01592	1.3675
9	4	1	13.1216	0.03196	0.02875	1.2864
10	4	2	85.2778	0.16899	0.18880	0.7277
11	4	2	69.6436	0.12270	0.12869	0.8583
12	4	2	38.8782	0.05476	0.05435	0.9723
13	4	1	14.8373	0.01148	0.01060	1.2248
14	4	1	6.8024	0.00082	0.00074	1.2851
15	4	1	20.5090	0.03478	0.03244	1.1510
16	4	1	23.5720	0.02813	0.02629	1.1361
17	4	1	8.1930	0.01197	0.01103	1.2538
18	4	1	9.1204	0.00677	0.00619	1.2680
19	4	2	15.6824	0.02364	0.02219	1.1953
20	4	4	42.9162	0.00831	0.00825	0.8175
21	4	2	91.5142	0.12647	0.14980	0.6399
22	4	1	14.2947	0.02470	0.02276	1.2119
23	4	1	6.4046	0.00940	0.00859	1.2817
24	4	2	42.5667	0.14725	0.15055	0.8947
25	4	1	3.6903	0.00450	0.00409	1.2990
26	4	1	7.9353	0.01924	0.01766	1.2430
27	4	2	21.6103	0.02450	0.02298	1.1631

Influence Diagnostics for Levels of Person

Person	COVTRACE	Cook's D Cov Parms	MDFFITS Cov Parms	COVRATIO CovParms	COVTRACE CovParms	RMSE without deleted level	Restricted Likelihood Distance
1	0.3098	0.03565	0.0326	1.1489	0.1448	2.31316	0.078
2	0.3415	0.05100	0.0456	1.1493	0.1469	2.32185	0.093
3	0.2903	0.01812	0.0165	1.1232	0.1211	2.30818	0.132
4	0.1888	0.05982	0.0534	1.1237	0.1234	2.27905	0.248
5	0.3600	0.06207	0.0548	1.1579	0.1560	2.32689	0.072
6	0.2891	0.05000	0.0463	1.1510	0.1458	2.30704	0.127
7	0.3613	0.06890	0.0611	1.1609	0.1583	2.32712	0.077
8	0.3305	0.05455	0.0495	1.1499	0.1464	2.31862	0.122
9	0.2649	0.03419	0.0319	1.1269	0.1232	2.30052	0.160
10	0.2899	0.61657	0.6538	0.9747	0.0896	2.14317	1.492
11	0.1389	0.32291	0.3103	1.0449	0.1108	2.18660	0.899
12	0.0252	0.04632	0.0488	1.0277	0.0361	2.23808	0.275
13	0.2102	0.02479	0.0231	1.1387	0.1345	2.30444	0.070
14	0.2613	0.03085	0.0272	1.1352	0.1351	2.31932	0.032
15	0.1454	0.00823	0.0077	1.1126	0.1098	2.28639	0.145
16	0.1319	0.00310	0.0030	1.0687	0.0677	2.28347	0.114
17	0.2350	0.05914	0.0545	1.1559	0.1505	2.31089	0.103
18	0.2468	0.04130	0.0377	1.1495	0.1455	2.31460	0.066
19	0.1854	0.03370	0.0347	1.0665	0.0702	2.29951	0.129
20	0.0354	7.42313	19.6236	0.5307	0.1129	2.27626	14.546
21	0.4057	0.82320	0.8390	1.0053	0.1676	2.12644	1.463
22	0.1991	0.02903	0.0271	1.1312	0.1273	2.30118	0.126
23	0.2584	0.02976	0.0263	1.1284	0.1285	2.31849	0.064
24	0.1059	0.48613	0.5319	0.8505	0.1552	2.21905	1.170
25	0.2727	0.03991	0.0351	1.1378	0.1376	2.32227	0.055
26	0.2258	0.01989	0.0181	1.1154	0.1140	2.30915	0.095
27	0.1564	0.02558	0.0237	1.1264	0.1232	2.28908	0.125

It is important to note that for every deletion set the procedure converged in less than the maximum number of specified iterations (ITER=5). The most iterations were required after removal of child 20. The covariance parameters had to move far away from the full-data estimates, which required more iterations than for the other subjects.

The overall influence diagnostic (RLD) and diagnostics for the fixed effects are displayed graphically in Figure 1. The large restricted likelihood distance of child 20 can be clearly seen. The children with the largest effect on the fixed effects estimates are 10, 11, 21, and 24 (Cook's D). They also have fairly large values of the PRESS statistic. If these children were not in the data, the resulting model would estimate rather poorly the mean growth of children that share their growth features. Interestingly, all five of these children have COVRATIO values of less than one. Deleting them from the analysis increases the estimated precision of the fixed effects estimates.

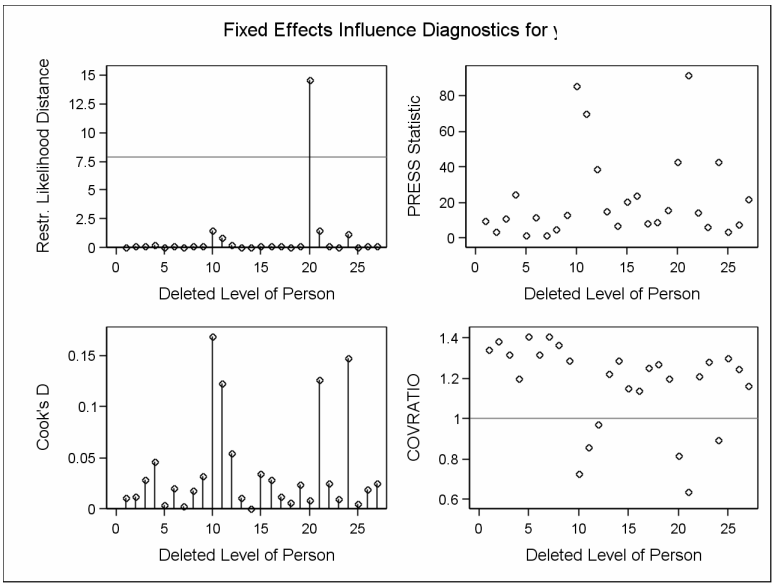

Figure 1. Overall and Fixed Effects Diagnostics

Influence diagnostics regarding the covariance parameters are shown in Figure 2. Again, the influence of child 20 far exceeds that of other subjects in these data. This is expected since its restricted likelihood distance is substantial, while its impact on the fixed effects was rather moderate.

The large value for Cook's D CovParms shows this subset's impact on the covariance parameter estimates. The COVRATIO again shows dramatically that in the absence of this child's observations the covariance parameters can be estimated much more precisely.

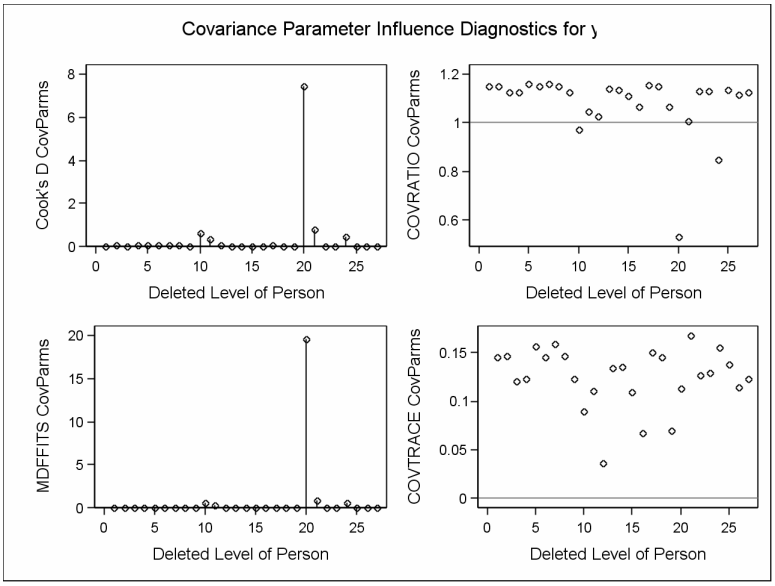

Figure 2. Covariance Parameter Diagnostics

Finally, a very detailed picture can be gained by examining how the individual parameter estimates react to the removal of the 27 sets in turn. The fixed effect estimates for the intercept and age effect are not altered by the removal of any of the first eleven children (Figure 3). This is a result of the singular parameterization of fixed classification effects in the MIXED procedure. The first eleven children are all girls.

Since there are only two covariance parameter estimates, the AR(1) correlation parameter and the residual variance, Figure 4 contains only two panels. If you recall that the trend over time for child 20 was not monotonic, it is understandable that its absence from the data will increase the AR(1) correlation parameter. If the non-monotonic set of observations is present, the correlation estimate will be suppressed considerably, regardless of which other child is being removed from the analysis. Note that there are other sets of observations, besides child 20, that exert influence on the residual variance.

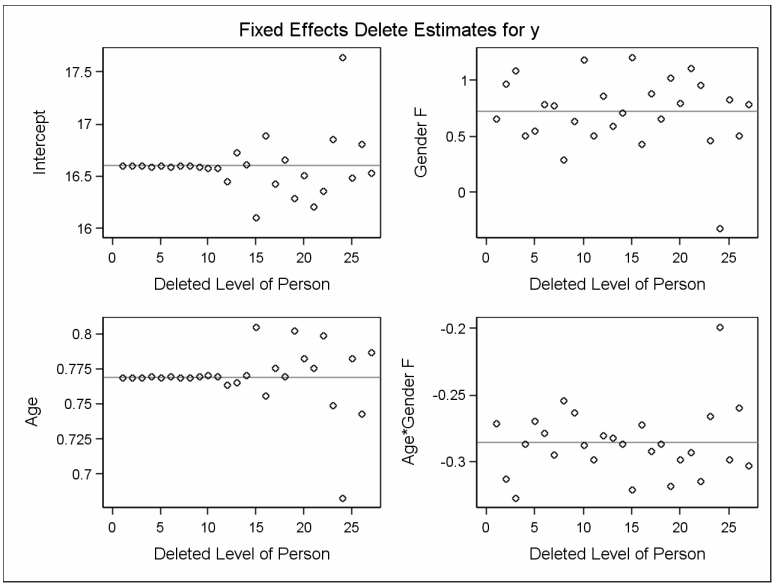

Figure 3. Fixed Effects Delete Estimates

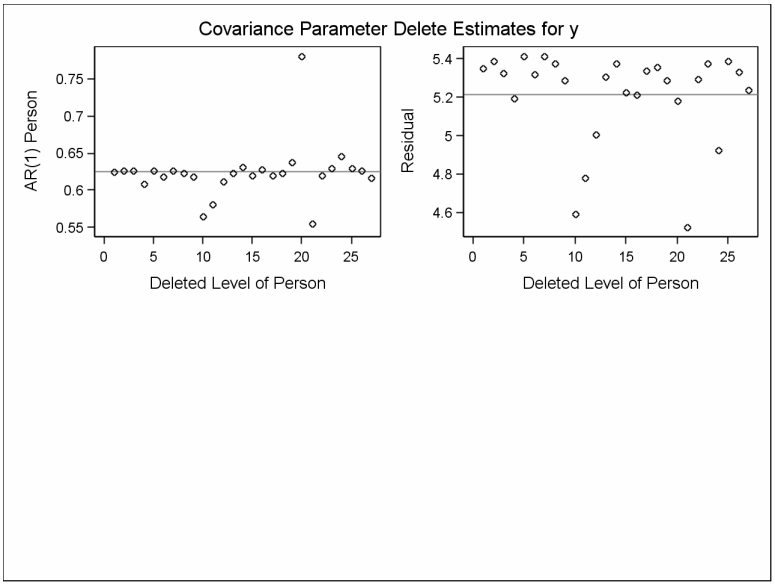

Figure 4. Covariance Parameter Delete Estimates

SUMMARY AND CONCLUSIONS

Standard residual and influence diagnostics for linear models can be extended to linear mixed models. The dependence of fixed-effects solutions on the covariance parameter estimates has important ramifications in perturbation analysis. To gauge the full impact of a set of observations on the analysis, covariance parameters need to be updated, which requires refitting of the model. The experimental INFLUENCE option of the MODEL statement in the MIXED procedure (SAS 9.1) enables you to perform iterative and noniterative influence analysis for individual observations and sets of observations.

The conditional (subject-specific) and marginal (population-averaged) formulations in the linear mixed model enable you to consider conditional residuals that use the estimated BLUPs of the random effects, and marginal residuals which are deviations from the overall mean. Residuals using the BLUPs are useful to diagnose whether the random effects components in the model are specified correctly, marginal residuals are useful to diagnose the fixed-effects components. Both types of residuals are available in SAS 9.1 as an experimental option of the MODEL statement in the MIXED procedure.

It is important to note that influence analyses are performed under the assumption that the chosen model is correct. Changing the model structure can alter the conclusions. Many other variance models have been fit to the data presented in the repeated measures example. You need to see the conclusions about which model component is affected in light of the model being fit. For example, modeling these data with a random intercept and random slope for each child or an unstructured covariance matrix will affect your conclusions about which children are influential on the analysis and how this influence manifests itself.

REFERENCES

Allen, D.M. (1974), "The Relationship between Variable Selection and Data Augmentation and a Method of Prediction," *Technometrics*, 16, 125–127.

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

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics; Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.

Christensen, R. (1991), *Linear Models for Multivariate, Time Series, and Spatial Data*, New York: Springer-Verlag.

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

Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R.D. (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.

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

Martin, R.J. (1992), "Leverage, Influence, and Residuals in Regression Models When Observations are Correlated," *Communications in Statistics, Theory & Methods*, 21, 1183 –1212.

Pothoff, R.F. and Roy, S.N. (1964), "A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems," *Biometrika*, 51, 313–326.

SAS Institute Inc. (2004), *SAS/STAT 9.1 User's Guide*, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION Oliver Schabenberger, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513
Email: oliver.schabenberger@sas.com