

schabenberger examines the use and implementation of influence measures in LME models.

Influence is understood to be the ability of a single or multiple data points, through their presences or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model (*schabenberger*).

Outliers are the most noteworthy data points in an analysis, and an objective of influence analysis is how influential they are, and the manner in which they are influential.

schabenberger describes a simple procedure for quantifying influence. Firstly a model should be fitted to the data, and estimates of the parameters should be obtained. The second step is that either single or multiple data points, specifically outliers, should be omitted from the analysis, with the original parameter estimates being updated. This is known as ‘leave one out’ or ‘leave k out’ analysis. The final step of the procedure is comparing the sets of estimates computed from the entire and reduced data sets to determine whether the absence of observations changed the analysis.

A residual is the difference between an observed quantity and its estimated or predicted value. In LME models, there are two types of residuals, marginal residuals and conditional residuals. A marginal residual is the difference between the observed data and the estimated marginal mean. A conditional residual is the difference between the observed data and the predicted value of the observation. In a model without random effects, both sets of residuals coincide.

schabenberger notes that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates.

Abstract

This paper reviews the use of diagnostic measures for LME models in SAS. This text has been widely cited by texts that don't deal with SAS implementations.

0.1 Introduction (Page 1)

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?

Model Validation Framework

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.

Stating the LME Model

The general linear mixed model is

$$Y = X\beta + Zu + \varepsilon$$

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

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 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 points 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, $b(U)$ denotes the fixed-effects *leave-U-out* estimates. Note that the set U can contain multiple observations.

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 β . 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.

0.2 Iterative and non-iterative influence analysis

? highlights some of the issue regarding implementing mixed model diagnostics.

A measure of total influence requires updates of all model parameters.

however, this doesn't increase the procedures execution time by the same degree.

0.2.1 Iterative Influence Analysis

For linear models, the implementation of influence analysis is straightforward. However, for LME models, the process is more complex. Update formulas for the fixed effects are available only when the covariance parameters are assumed to be known. A measure of total influence requires updates of all model parameters. This can only be achieved in general is by omitting observations, then refitting the model.

? describes the choice between iterative influence analysis and non-iterative influence analysis.

0.2.2 Summary of Paper

Standard residual and influence diagnostics for linear models can be extended to LME models. The dependence of the fixed effects solutions on the covariance parameters has important ramifications on the perturbation analysis. Calculating the studentized residuals-And influence statistics whereas each software procedure can calculate both conditional and marginal raw residuals, only SAS Proc Mixed is currently the only program that provide studentized residuals Which are preferred for model diagnostics. The conditional Raw residuals are not well suited to detecting outliers as are the studentized conditional residuals. (schabenbege r)

LME are flexible tools for the analysis of clustered and repeated measurement data. LME extend the capabilities of standard linear models by allowing unbalanced and missing data, as long as the missing data are MAR. Structured covariance matrices for both the random effects G and the residuals R . missing at Random.

A conditional residual is the difference between the observed value and the predicted value of a dependent variable- Influence diagnostics are formal techniques that allow the identification of observations that heavily influence estimates of parameters. To alleviate the problems with the interpretation of conditional residuals that may have unequal variances, we consider scaling. Residuals obtained in this manner are called studentized residuals.

0.2.3 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.

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.

Schabenberger: 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.