

Chapter 1

Model Diagnostics

1.1 Introduction

In classical linear models model diagnostics have been become a required part of any statistical analysis, and the methods are commonly available in statistical packages and standard textbooks on applied regression. However it has been noted by several papers that model diagnostics do not often accompany LME model analyses. Model diagnostic techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations.

1.1.1 Model Data Agreement

? describes the examination of model-data agreement as comprising several elements; residual analysis, goodness of fit, collinearity diagnostics and influence analysis.

1.2 Introduction

In classical linear models model diagnostics have been become a required part of any statistical analysis, and the methods are commonly available in statistical packages and standard textbooks on applied regression. However it has been noted by several papers

that model diagnostics do not often accompany LME model analyses. Model diagnostic techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations.

1.2.1 Model Data Agreement

? describes the examination of model-data agreement as comprising several elements; residual analysis, goodness of fit, collinearity diagnostics and influence analysis.

What is Influence

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 *schabenberger*.

1.2.2 Model Data Agreement

Schabenberger(20XX) describes the examination of model-data agreement as comprising several elements;

- residual analysis,
- goodness of fit,
- collinearity diagnostics
- influence analysis.

1.2.3 Influence Diagnostics: Basic Idea and Statistics

The general idea of quantifying the influence of one or more observations relies on computing parameter estimates based on all data points, removing the cases in question from the data, refitting the model, and computing statistics based on the change between full-data and reduced-data estimation.

1.2.4 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].

1.2.5 What is Influence

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 (schabenberger).

1.2.6 What is Influence

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 (Schabenberger, 20XX).

1.2.7 Quantifying Influence

The basic procedure for quantifying influence is simple as follows:

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

? introduces powerful tools for local-influence assessment and examining perturbations in the assumptions of a model. In particular the effect of local perturbations of parameters or observations are examined.

1.2.8 Influence Analysis for LME Models

The linear mixed effects model is a useful methodology for fitting a wide range of models. However, linear mixed effects models are known to be sensitive to outliers. ? advises that identification of outliers is necessary before conclusions may be drawn from the fitted model.

Standard statistical packages concentrate on calculating and testing parameter estimates without considering the diagnostics of the model. The assessment of the effects of perturbations in data, on the outcome of the analysis, is known as statistical influence analysis. Influence analysis examines the robustness of the model. Influence analysis methodologies have been used extensively in classical linear models, and provided the basis for methodologies for use with LME models. Computationally inexpensive diagnostics tools have been developed to examine the issue of influence (?). Studentized residuals, error contrast matrices and the inverse of the response variance covariance matrix are regular components of these tools.

1.2.9 Influence Statistics for LME models

Influence statistics can be coarsely grouped by the aspect of estimation that is their primary target:

- overall measures compare changes in objective functions: (restricted) likelihood distance (Cook and Weisberg 1982, Ch. 5.2)
- influence on parameter estimates: Cook's (Cook 1977, 1979), MDFFITS (Belsley, Kuh, and Welsch 1980, p. 32)
- influence on precision of estimates: CovRatio and CovTrace
- influence on fitted and predicted values: PRESS residual, PRESS statistic (Allen 1974), DFFITS (Belsley, Kuh, and Welsch 1980, p. 15)
- outlier properties: internally and externally studentized residuals, leverage

1.3 Extension of techniques to LME Models

Model diagnostic techniques, well established for classical models, have since been adapted for use with linear mixed effects models. Diagnostic techniques for LME models are inevitably more difficult to implement, due to the increased complexity.

Beckman, Nachtsheim and Cook (1987) ? applied the local influence method of Cook (1986) to the analysis of the linear mixed model.

While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known.

If the global measure suggests that the points in U are influential, the nature of that influence should be determined. In particular, the points in U can affect the following

- 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.

1.3.1 What is Influence

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 (?).

1.3.2 Quantifying Influence

The basic procedure for quantifying influence is simple as follows:

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

? introduces powerful tools for local-influence assessment and examining perturbations in the assumptions of a model. In particular the effect of local perturbations of parameters or observations are examined.

1.4 Residual diagnostics

For classical linear models, residual diagnostics are typically implemented as a plot of the observed residuals and the predicted values. A visual inspection for the presence of trends inform the analyst on the validity of distributional assumptions, and to detect outliers and influential observations.

Extension of techniques to LME Models

Model diagnostic techniques, well established for classical models, have since been adapted for use with linear mixed effects models. Diagnostic techniques for LME models are inevitably more difficult to implement, due to the increased complexity.

Beckman, Nachtsheim and Cook (1987) ? applied the local influence method of Cook (1986) to the analysis of the linear mixed model.

While the concept of influence analysis is straightforward, implementation in mixed models is more complex. Update formulae for fixed effects models are available only when the covariance parameters are assumed to be known.

If the global measure suggests that the points in U are influential, the nature of that influence should be determined. In particular, the points in U can affect the following

- 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.

Influence Diagnostics Basic Idea and Statistics

The general idea of quantifying the influence of one or more observations relies on computing parameter estimates based on all data points, removing the cases in question from the data, refitting the model, and computing statistics based on the change between full-data and reduced-data estimation.

Influence statistics can be coarsely grouped by the aspect of estimation that is their primary target: overall measures compare changes in objective functions: (restricted) likelihood distance (Cook and Weisberg 1982, Ch. 5.2) influence on parameter estimates: Cooks (Cook 1977, 1979), MDFFITS (Belsley, Kuh, and Welsch 1980, p. 32) influence on precision of estimates: CovRatio and CovTrace influence on fitted and predicted values: PRESS residual, PRESS statistic (Allen 1974), DFFITS (Belsley, Kuh, and Welsch 1980, p. 15) outlier properties: internally and externally studentized residuals, leverage For linear models for uncorrelated data, it is not necessary to refit the model after removing a data point in order to measure the impact of an observation on the model. The change in fixed effect estimates, residuals, residual sums of squares, and the variance-covariance matrix of the fixed effects can be computed based on the fit to the full data alone. By contrast, in mixed models several important complications arise. Data points can affect not only the fixed effects but also the covariance parameter estimates on which the fixed-effects estimates depend.

Furthermore, closed-form expressions for computing the change in important model quantities might not be available. This section provides background material for the various influence diagnostics available with the MIXED procedure. See the section Mixed Models Theory for relevant expressions and definitions. The parameter vector denotes all unknown parameters in the and matrix. The observations whose influence is being ascertained are represented by the set and referred to simply as "the observations in ." The estimate of a parameter vector, such as , obtained from all observations except those in the set is denoted . In case of a matrix , the notation represents the matrix with the rows in removed; these rows are collected in . If is symmetric, then notation implies removal of rows and columns. The vector comprises the responses of the data points

being removed, and is the variance-covariance matrix of the remaining observations. When , lowercase notation emphasizes that single points are removed, such as .

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.

1.5 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 observation 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.

1.5.1 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 procedures 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.

1.5.2 Influential Observations : DFBeta and DFBetas

Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set. dfbeta refers to how much a parameter estimate changes if the observation in question is dropped from the data set. Note that with k covariates, there will be k+1 dfbetas (the intercept, β_0 , and 1 β for each covariate). Cook's distance is presumably more important to you if you are doing predictive modeling, whereas dfbeta is more important in explanatory modeling.

1.5.3 Leverage

leverage is a term used in connection with regression analysis and, in particular, in analyses aimed at identifying those observations that are far away from corresponding average predictor values. Leverage points do not necessarily have a large effect on the outcome of fitting regression models.

Leverage points are those observations, if any, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.[1]

Modern computer packages for statistical analysis include, as part of their facilities for regression analysis, various quantitative measures for identifying influential observations: among these measures is partial leverage, a measure of how a variable contributes to the leverage of a datum.

1.5.4 Residual

Residual (or error) represents unexplained (or residual) variation after fitting a regression model. It is the difference (or left over) between the observed value of the variable and the value suggested by the regression model.

The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the residual (e). Each data point has one residual.

Residual = Observed value - Predicted value

$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero. That is, $\sum e = 0$ and $\bar{e} = 0$.

1.5.5 Introduction

In statistics and optimization, statistical errors and residuals are two closely related and easily confused measures of the deviation of an observed value of an element of

a statistical sample from its "theoretical value". The error (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true function value, while the residual of an observed value is the difference between the observed value and the estimated function value.

The distinction is most important in regression analysis, where it leads to the concept of studentized residuals.

1.5.6 Residual

A residual (or fitting error), on the other hand, is an observable estimate of the unobservable statistical error. Consider the previous example with men's heights and suppose we have a random sample of n people. The sample mean could serve as a good estimator of the population mean. Then we have:

The difference between the height of each man in the sample and the unobservable population mean is a statistical error, whereas The difference between the height of each man in the sample and the observable sample mean is a residual. Note that the sum of the residuals within a random sample is necessarily zero, and thus the residuals are necessarily not independent. The statistical errors on the other hand are independent, and their sum within the random sample is almost surely not zero.

Other uses of the word "error" in statistics:

The use of the term "error" as discussed in the sections above is in the sense of a deviation of a value from a hypothetical unobserved value. At least two other uses also occur in statistics, both referring to observable prediction errors:

- Mean square error or mean squared error (abbreviated MSE) and root mean square error (RMSE) refer to the amount by which the values predicted by an estimator differ from the quantities being estimated (typically outside the sample from which the model was estimated).
- Sum of squared errors, typically abbreviated SSE or SSe, refers to the residual sum of squares (the sum of squared residuals) of a regression; this is the sum of

the squares of the deviations of the actual values from the predicted values, within the sample used for estimation. Likewise, the sum of absolute errors (SAE) refers to the sum of the absolute values of the residuals, which is minimized in the least absolute deviations approach to regression.