0.1 Case Deletion Diagnostics

Since the pioneering work of Cook in 1977, deletion measures have been applied to many statistical models for identifying influential observations. Case-deletion diagnostics provide a useful tool for identifying influential observations and outliers.

The key to making deletion diagnostics useable is the development of efficient computational formulas, allowing one to obtain the case deletion diagnostics by making use of basic building blocks, computed only once for the full model.

The computation of case deletion diagnostics in the classical model is made simple by the fact that estimates of β and σ^2, which exclude the ith observation, can be computed without re-fitting the model.

Preisser (1996) describes two type of diagnostics. When the set consists of only one observation, the type is called '*observation-diagnostics*'. For multiple observations, Preisser describes the diagnostics as '*cluster-deletion*' diagnostics.

When applied to LME models, such update formulas are available only if one assumes that the covariance parameters are not affected by the removal of the observation in question. However, this is rarely a reasonable assumption.

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.

0.1.1 Local Influence

Christensen et al. (1992) developed their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted regression problem (conditional on the estimated covariance matrix) for fixed effects.

Christensen, Pearson and Johnson (1992) studied case deletion diagnostics, in particular the equivalent of Cook's distance, for diagnosing influential observations when estimating the fixed effect parameters and variance components.

0.1.2 Deletion Diagnostics

Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models.

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

0.2 Case Deletion Diagnostics for LME models

Haslett and Dillane (2004) remark that linear mixed effects models didn't experience a corresponding growth in the use of deletion diagnostics, adding that McCullough and Searle (2001) makes no mention of diagnostics whatsoever.

Christensen, Pearson and Johnson (1992) describes three propositions that are required for efficient case-deletion in LME models. The first proposition decribes how to efficiently update V when the ith element is deleted.

$$V_{[i]}^{-1} = \Lambda_{[i]} - \frac{\lambda\lambda\prime}{\nu ii} \tag{1}$$

The second of christensen's propostions is the following set of equations, which are

variants of the Sherman Wood bury updating formula.

$$X'_{[i]}V^{-1}_{[i]}X_{[i]} = X'V^{-1}X - \frac{\hat{x}_i\hat{x}'_i}{s_i} \quad (2)$$

$$(X'_{[i]}V^{-1}_{[i]}X_{[i]})^{-1} = (X'V^{-1}X)^{-1} + \frac{(X'V^{-1}X)^{-1}\hat{x}_i\hat{x}'_i(X'V^{-1}X)^{-1}}{s_i - \bar{h}_i} \quad (3)$$

$$X'_{[i]}V^{-1}_{[i]}Y_{[i]} = X\prime V^{-1}Y - \frac{\hat{x}_i\hat{y}'_i}{s_i} \quad (4)$$

Influence on measure component ratios

The general diagnostic tools for variance component ratios are the analogues of the Cook's distance and the Information Ratio.

$$\begin{aligned} CD_U(\gamma) &= (\hat{\gamma}_{(U)} - \hat{\gamma})'[\mathrm{var}(\hat{\gamma})]^{-1}(\hat{\gamma}_{(U)} - \hat{\gamma}) \\ &= -\boldsymbol{g'}_{(U)}(\boldsymbol{Q} - \boldsymbol{G})^{-1}\boldsymbol{Q}(\boldsymbol{Q} - \boldsymbol{G})\boldsymbol{g}_{(U)} \\ &= \boldsymbol{g'}_{(U)}(\boldsymbol{I}_r\mathrm{var}(\hat{\gamma})\boldsymbol{G})^{-2}\mathrm{var}(\hat{\gamma})\boldsymbol{g}_{(U)} \end{aligned}$$

Large values of $CD(\gamma)$ highlight observation groups for closer attentions

$$IR\gamma = \frac{\det(\boldsymbol{Q} - \boldsymbol{G})}{\det(\boldsymbol{Q})}$$

Ideally when all observations have the same influence on the information matrix $IR\gamma$ is approximately one. Deviations from one indicate the group U is influential. Since $\mathrm{var}(\hat{\gamma})$ and $\boldsymbol{I}_r$ are fixed for all observations, $IR\gamma$ is a function of $\boldsymbol{G}$, in turn a function of $\boldsymbol{C}_i$ and c_{ii}.

0.3 Case Deletion Diagnostics for LME models

Schabenberger (2004) examines the use and implementation of influence measures in LME models.

Schabenberger (2004) 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 of 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 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.

schabenberger

Schabenberger (2004) notes that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates. Haslett and Dillane (2004) offers an procedure to assess the influences for the variance components within the linear model, complementing the existing methods for the fixed components. The essential problem is that there is no useful updating procedures for $\hat{V}$, or for $\hat{V}^{-1}$. Haslett and Dillane (2004) propose an alternative , and computationally inexpensive approach, making use of the 'delete=replace' identity.

Haslett (1999) considers the effect of 'leave k out' calculations on the parameters β and σ^2, using several key results from Haslett and Hayes (1998) on partioned matrices.

0.3.1 Extension of Diagnostic Methods to LME models

? noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

? develops case deletion diagnostics, in particular the equivalent of Cook's distance, a well-known metric, for diagnosing influential observations when estimating the fixed

effect parameters and variance components. Deletion diagnostics provide a means of assessing the influence of an observation (or groups of observations) on inference on the estimated parameters of LME models. We shall provide a fuller discussion of Cook's distance in due course.

Demidenko (2004) proposes two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

0.3.2 Extending deletion diagnostics to LMEs

Christensen et al. (1992) notes the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect. Christensen et al. (1992) develops these techniques in the context of REML

$$X = \begin{bmatrix} x_i' \\ X(i) \end{bmatrix}, Z = \begin{bmatrix} z_{ij}' \\ Z_{j(i)} \end{bmatrix}, Z = \begin{bmatrix} z_{ij}' \\ Z_{j(i)} \end{bmatrix},$$
$$y = \begin{bmatrix} y_{ij}' \\ y_{j(i)} \end{bmatrix} \text{ and } H = \begin{bmatrix} h_{ii} & h \\ h_{j(i)} & h \end{bmatrix}$$

For notational simplicity, $\boldsymbol{A}_{(i)}$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, $\boldsymbol{a}_i$ denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the (i,j)-th element of $\boldsymbol{A}$.

$\boldsymbol{a}_{(i)}$ denotes a vector $\boldsymbol{a}$ with the i-th element, a_i, removed.

$$\breve{a}_i = \boldsymbol{a}_i - \boldsymbol{A}_{(i)}\boldsymbol{H}_{[i]}\boldsymbol{h}_i \tag{5}$$

0.4 Extension of Diagnostic Methods to LME models

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.

? noted the case deletion diagnostics techniques have not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

? noted the case deletion diagnostics techniques had not been applied to linear mixed effects models and seeks to develop methodologies in that respect. **?** develops these techniques in the context of REML.

Demidenko (2004) extends several regression diagnostic techniques commonly used in linear regression, such as leverage, infinitesimal influence, case deletion diagnostics, Cook's distance, and local influence to the linear mixed-effects model. In each case, the proposed new measure has a direct interpretation in terms of the effects on a parameter of interest, and reduces to the familiar linear regression measure when there are no random effects.

The new measures that are proposed by Demidenko (2004) are explicitly defined functions and do not require re-estimation of the model, especially for cluster deletion diagnostics. The basis for both the cluster deletion diagnostics and Cook's distance is a generalization of Miller's simple update formula for case deletion for linear models. Furthermore Demidenko (2004) shows how Pregibon's infinitesimal case deletion diagnostics is adapted to the linear mixed-effects model.

0.5 Case Deletion Diagnostics for LME models

Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal.

Demidenko (2004) proposes two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

0.6 Extension of technique 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 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

0.7 The CPJ Paper

0.7.1 Case-Deletion results for Variance components

CPJ examines case deletion results for estimates of the variance components, proposing the use of one-step estimates of variance components for examining case influence. The method describes focuses on REML estimation, but can easily be adapted to ML or other methods.

This paper develops their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted regression problem (conditional on the estimated covariance matrix) for fixed effects.

0.7.2 CPJ Notation

$$\boldsymbol{C} = \boldsymbol{H}^{-1} = \begin{bmatrix} c_{ii} & \boldsymbol{c}_i' \\ \boldsymbol{c}_i & \boldsymbol{C}_{[i]} \end{bmatrix}$$

CPJ noted the following identity:

$$\boldsymbol{H}_{[i]}^{-1} = \boldsymbol{C}_{[i]} - \frac{1}{c_{ii}} \boldsymbol{c}_{[i]} \boldsymbol{c}_{[i]}'$$

CPJ use the following as building blocks for case deletion statistics.

- $\breve{x}_i$
- $\breve{z}_i j$
- $p_i i$
- $\breve{z}_i$
- $\breve{y}_i$
- m_i

All of these terms are a function of a row (or column) of $\boldsymbol{H}$ and $\boldsymbol{H}_{[i]}^{-1}$

0.8 Matrix Notation for Case Delection

0.8.1 Case deletion notation

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i,j)-$th element of $\boldsymbol{A}$.

0.9 CPJ's Three Propositions

Proposition 1

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \nu^{ii} & \lambda_i' \\ \lambda_i & \Lambda_{[i]} \end{bmatrix}$$

$$\boldsymbol{V}_{[i]}^{-1} = \boldsymbol{\Lambda}_{[i]} - \frac{\lambda_i \lambda_i'}{\lambda_i}$$

0.9.1 Proposition 2

(i) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{X}_{[i]} = \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}$

(ii) $= (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y})^{-1}$

(iii) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{Y}_{[i]} = \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}$

0.9.2 Proposition 3

This proposition is similar to the formula for the one-step Newtown Raphson estimate of the logistic regression coefficients given by pregibon (1981) and discussed in Cook Weisberg.

0.10 The CPJ Paper

0.10.1 Case-Deletion results for Variance components

Christensen et al. (1992) examines case deletion results for estimates of the variance components, proposing the use of one-step estimates of variance components for examining case influence. The method describes focuses on REML estimation, but can easily be adapted to ML or other methods.

? examines case deletion results for estimates of the variance components, proposing the use of one-step estimates of variance components for examining case influence. The method describes focuses on REML estimation, but can easily be adapted to ML or other methods.

This paper develops their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted regression problem (conditional on the estimated covariance matrix) for fixed effects.

Christensen et al. (1992) developed their global influences for the deletion of single observations in two steps: a one-step estimate for the REML (or ML) estimate of the variance components, and an ordinary case-deletion diagnostic for a weighted resgression problem (conditional on the estimated covariance matrix) for fixed effects. Lesaffre's approach accords with that proposed by Christensen et al when applied in a repeated measurement context, with a large sample size.

0.10.2 Preisser 2008

Preisser & Qaqish (1996) introduced one-step deletion diagnostics for generalized estimating equations. In this note, we derive a different expression for DBETAm, and show that it is equivalent to the formula of Preisser & Qaqish (1996). We show that

significant computational savings are possible through application of the ShermanMorrisonWoodbury formula (Sherman & Morrison, 1950; Henderson & Searle, 1981) for the inverse of a matrix component in the diagnostic formula.

0.10.3 Partitioning Matrices

Without loss of generality, matrices can be partitioned as if the $i-$th omitted observation is the first row; i.e. $i = 1$.

0.11 Case deletion notation

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i, j)-$th element of $\boldsymbol{A}$.

0.11.1 Partitioning Matrices

Without loss of generality, matrices can be partitioned as if the $i-$th omitted observation is the first row; i.e. $i = 1$.

0.11.2 Effects on fitted and predicted values

$$\hat{e}_{i(U)} = y_i - x\hat{\beta}_{(U)} \tag{6}$$

0.11.3 Matrix Notation for Case Deletion

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i, j)-$th element of $\boldsymbol{A}$.

Bibliography

Christensen, R., L. M. Pearson, and W. Johnson (1992). Case-deletion diagnostics for mixed models. *Technometrics 34*(1), 38–45.

Demidenko, E. (2004). *Mixed Models: Theory And Application*. Dartmouth College: Wiley Interscience.

Haslett, J. (1999). A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society (Series B) 61*, 603–609.

Haslett, J. and D. Dillane (2004). Application of 'delete = replace' to deletion diagnostics for variance component estimation. *Journal of the Royal Statistical Society (Series B) 66*, 131–143.

Haslett, J. and K. Hayes (1998). Residuals for the linear model with general covariance structure. *Journal of the Royal Statistical Society (Series B) 60*, 201–215.

McCullough, C. and S. Searle (2001). *Generalized , Linear and Mixed Models*. Wiley Interscience.

Preisser, J. S. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika 83*(3), 551–5562.

Schabenberger, O. (2004). Mixed model influence diagnostics. 18929.