Zewotir's Paper

# 1 Demidenko's I Influence

The concept of I Influence is generalized to the non linea regression model.

## 1.1 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) Beckman et al. (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.

## 1.2 Standardized and studentized residuals

To alleviate the problem caused by inconstant variance, the residuals are scaled (i.e. divided) by their standard deviations. This results in a 'standardized residual'. Because true standard deviations are frequently unknown, one can instead divide a residual by the estimated standard deviation to obtain the 'studentized residual.

## 1.3 Residual Analysis for Linear Models, LME models and GLMs

**Keywords:**

- Residuals (*Beginners*),
- Testing the Assumption of Normality (*Beginners*)
- Diagnostic Plots with the `plot` function
- Cook's Distance
- DFFits and DFBeta
- Standardized and Studentized Residuals
- Influence Leverage and Outlierness

## 1.4 Identifying outliers with a LME model object

The process is slightly different than with standard LME model objects, since the ***influence*** function does not work on lme model objects. Given ***mod.lme***, we can use the plot function to identify outliers.

## 1.5 Diagnostics for Random Effects

Empirical best linear unbiased predictors EBLUPS provide the a useful way of diagnosing random effects.

EBLUPs are also known as "shrinkage estimators" because they tend to be smaller than the estimated effects would be if they were computed by treating a random factor as if it was fixed (West et al )

## 1.6 Case Deletion Diagnostics for Mixed Models

**?** notes 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

## 1.7 Methods and Measures

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.

Zewotir and Galpin (2005) lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,
- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

# 2 Demidenko's I Influence

The concept of I Influence is generalized to the non linea regression model.

Zewotir Notepad

> Abstract: Linear mixed models are extremely sensitive to outlying responses and extreme points in the fixed and random effect design spaces. Few diagnostics are available in standard computing packages. We provide routine diagnostic tools, which are computationally inexpensive. The diagnostics are functions of basic building blocks: studentized residuals, error contrast matrix, and the inverse of the response variable covariance matrix. The basic building blocks are computed only once from the complete data analysis and provide information on the influence of the data on different

aspects of the model fit. Numerical examples provide analysts with the complete pictures of the diagnostics.

Key words: Case deletion, influential observations, randomeffects, statistical diagnostics, variance components ratios.

Description: The influence of observations on statistical inference is of importance in statistical data analysis. A practical and well-established approach to influence analysis is based on case deletion. We provide computationally inexpensive diagnostic tools for linear mixed models. The diagnostics are a function of basic building blocks, computed only once from the complete data analysis, and provide information on the influence of the data on different aspects of the model fit.

Residual standard deviation. Roy Subject effects, replicate in subject, Cardiac data PEFR data from BLand Royal Melbourne Hospital. Roy demonstrates that correlation can be described under the model formulation.

$$Y_i = x.\beta + Z.u + epsilon]$$

Laird Ware form (litte et al)

For the purpose of comparison of both approaches, we compute the limits of agreement for two methods described in well known data sets.

$\epsilon$ is an $ntimes1$ vector of error terms Zewotir provides routine diagnostics tools that are computationally inexpensive. $\boldsymbol{u}_i$ is a $q_i \times 1$ vector of random variables from $\mathcal{N}(0\sigma^2.I)$

Christensen Petersen and Johnson studied case deletion diagnostics.

# 3 The Hat Matrix

The projection matrix $H$ (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values $Y$, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^TX)^{-1}X^T \tag{1}$$

$H$ describes the influence each observed value has on each fitted value. The diagonal elements of the $H$ are the 'leverages', which describe the influence each observed value

has on the fitted value for that same observation. The residuals ($R$) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{2}$$

The variances of $Y$ and $R$ can be expressed as:

$$\begin{aligned} \text{var}(Y) &= H\sigma^2 \\ \text{var}(R) &= (I - H)\sigma^2 \end{aligned} \tag{3}$$

Updating techniques allow an economic approach to recalculating the projection matrix, $H$, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

# 4 Zewotir Measures of Influence in LME Models

Zewotir and Galpin (2005) describes a number of approaches to model diagnostics, investigating each of the following;

- Variance components
- Fixed effects parameters
- Prediction of the response variable and of random effects
- likelihood function

## 4.1 Cook's Distance

- For variance components $\gamma$: $CD(\gamma)_i$,
- For fixed effect parameters $\beta$: $CD(\beta)_i$,
- For random effect parameters $\boldsymbol{u}$: $CD(u)_i$,
- For linear functions of $\hat{beta}$: $CD(\psi)_i$

### 4.1.1 Random Effects

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

### 4.1.2 linear functions

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

## 4.2 Information Ratio

# 5 Zewotir Measures of Influence in LME Models

Zewotir and Galpin (2005) describes a number of approaches to model diagnostics, investigating each of the following;

- Variance components
- Fixed effects parameters
- Prediction of the response variable and of random effects
- likelihood function

## 5.1 Cook's Distance

- For variance components $\gamma$: $CD(\gamma)_i$,
- For fixed effect parameters $\beta$: $CD(\beta)_i$,
- For random effect parameters $\boldsymbol{u}$: $CD(u)_i$,
- For linear functions of $\hat{beta}$: $CD(\psi)_i$

### 5.1.1 Random Effects

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

### 5.1.2 linear functions

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

# 6 Zewotir Measures of Influence in LME Models

Zewotir and Galpin (2005) describes a number of approaches to model diagnostics, investigating each of the following;

- Variance components
- Fixed effects parameters
- Prediction of the response variable and of random effects
- likelihood function

## 6.1 Cook's Distance

- For variance components $\gamma$: $CD(\gamma)_i$,
- For fixed effect parameters $\beta$: $CD(\beta)_i$,
- For random effect parameters $\boldsymbol{u}$: $CD(u)_i$,
- For linear functions of $\hat{beta}$: $CD(\psi)_i$

### 6.1.1 Random Effects

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

### 6.1.2 linear functions

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

## 6.2 Information Ratio

# 7 Zewotir Measures of Influence in LME Models

Zewotir and Galpin (2005) describes a number of approaches to model diagnostics, investigating each of the following;

- Variance components
- Fixed effects parameters
- Prediction of the response variable and of random effects
- likelihood function

# 8 Measures 2

## 8.1 Cook's Distance

- For variance components $\gamma$

Diagnostic tool for variance components

$$C_{\theta i} = (\hat{(\theta)}_{[i]} - \hat{(\theta)})^T \text{cov}(\hat{(\theta)})^{-1} (\hat{(\theta)}_{[i]} - \hat{(\theta)})$$

## 8.2 Variance Ratio

- For fixed effect parameters $\beta$.

## 8.3 Cook-Weisberg statistic

- For fixed effect parameters $\beta$.

## 8.4 Andrews-Pregibon statistic

- For fixed effect parameters $\beta$.

The Andrews-Pregibon statistic $AP_i$ is a measure of influence based on the volume of the confidence ellipsoid. The larger this statistic is for observation $i$, the stronger the influence that observation will have on the model fit.

# 9 Efficient Updating Theorem

Zewotir and Galpin (2005) describes the basic theorem of efficient updating.

- $$m_i = \frac{1}{c_{ii}}$$

### 9.0.1 linear functions

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

# 10 Efficient Updating Theorem

Zewotir and Galpin (2005) describes the basic theorem of efficient updating.

- $$m_i = \frac{1}{c_{ii}}$$

### 10.0.1 Random Effects

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

### 10.0.2 linear functions

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

## 10.1 Information Ratio

# 11 Efficient Updating Theorem

Zewotir and Galpin (2005) describes the basic theorem of efficient updating.

- $$m_i = \frac{1}{c_{ii}}$$

### 11.0.1 Random Effects

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

### 11.0.2 Random Effects

A large value for $CD(u)_i$ indicates that the $i-$th observation is influential in predicting random effects.

### 11.0.3 linear functions

$CD(\psi)_i$ does not have to be calculated unless $CD(\beta)_i$ is large.

## 11.1 Information Ratio

# 12 Computation and Notation

with $\boldsymbol{V}$ unknown, a standard practice for estimating $\boldsymbol{X\beta}$ is the estime the variance components $\sigma_j^2$, compute an estimate for $\boldsymbol{V}$ and then compute the projector matrix $A$, $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$.

**?** remarks that $\boldsymbol{D}$ is a block diagonal with the $i-$th block being $u\boldsymbol{I}$

# 13 Haslett's Analysis

For fixed effect linear models with correlated error structure Haslett (1999) showed that the effects on the fixed effects estimate of deleting each observation in turn could be cheaply computed from the fixed effects model predicted residuals.

A general theory is presented for residuals from the general linear model with correlated errors. It is demonstrated that there are two fundamental types of residual associated with this model, referred to here as the marginal and the conditional residual.

These measure respectively the distance to the global aspects of the model as represented by the expected value and the local aspects as represented by the conditional expected value.

These residuals may be multivariate.

Haslett and Hayes (1998) developes some important dualities which have simple implications for diagnostics.

## 13.1 Methods and Measures

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.

Zewotir and Galpin (2005) lists several established methods of analyzing influence in LME models. These methods include

- Cook's distance for LME models,
- likelihood distance,
- the variance (information) ration,

- the Cook-Weisberg statistic,
- the Andrews-Prebigon statistic.

# 14 Zewotir: Computation and Notation

with $\boldsymbol{V}$ unknown, a standard practice for estimating $\boldsymbol{X\beta}$ is the estime the variance components $\sigma_j^2$, compute an estimate for $\boldsymbol{V}$ and then compute the projector matrix $A$, $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$.

Zewotir remarks that $\boldsymbol{D}$ is a block diagonal with the $i-$th block being $u\boldsymbol{I}$

# 15 Section 3

$$\boldsymbol{X} = \begin{bmatrix} x' \\ \boldsymbol{X} \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} z' \\ \boldsymbol{Z} \end{bmatrix}$$

$\boldsymbol{A}_{(i)}$ denote an $n \times m$ matrix $\boldsymbol{A}$ with the $i-$th row removed.

CPJ used certain statistics as the basic building blocks of case deletion diagnostics.

**Theorem 2 :** Basic Theorem of efficient updating (Zewotir)

# 16 Measures of Influence

Cook's Distance

$$CD_i(\gamma) = \boldsymbol{g}'_{(i)}(\boldsymbol{I} + var(\hat{\gamma})\boldsymbol{G} + \boldsymbol{g}$$

Large values of $CD$ highlights observations for special attention.

Information Ratio

$$IR(\gamma) = det(\boldsymbol{I}_r + var(\hat{\gamma})\boldsymbol{G}$$

- $det(A)$ denotes the determinant of the square matrix $\boldsymbol{A}$.
-

4.2. Influence on fixed effects parameter estimates 4.2.1 Analogue of Cooks Distance The Cooks distance can be extended to measure influence on the fixed effects in the mixed models. Large values of $CD_i(\beta)$ indicates points for further consideration 4.2.2. Analogue of the variance ratio

The variance ratio measures the change of the determinant of the variance of the fixed effects parameter estimates when the i-th case is deleted.

4.2.3 Analogue of the Cook-Weisberg statistic

This statistic is used to measure the change of the confidence ellipsoid value of $\beta$.

$$\boldsymbol{y} \sim N(boldsymbolX\beta, \sigma_e^2 psilonboldsymbolH)$$

The $100(1-\alpha)\%$ confidence ellipsoid for $\beta$ is....

Cook and Weisberg proposed the logarithm of the ratio $E_{(i)}$ to E as a measure of influence.

4.2.4 Analogue of the Andrews Pregibons statistic This is another measure based on the volume of the confidence ellipsoid. $\mathrm{AP}_i$

4.3 Influence on random effects prediction. Analogue of Cooks Distance A large $CD_i(u)$ indicates that the i-th observation is influential in predicting random effects.

4.4 Influence on the likelihood function Likelihood Distance ($LD_i$)

4.5 influence on the linear functions of the fixed effect parameters. All the diagnostics are a function of the following basic building blocks 1) Studentized residuals 2) Error contrast matrix 3) The inverse of the response variable covariance matrix. The basic building blocks are computed once from the complete data set. Zewotir assumes that D is block diagonal with the i-th block being $\gamma.I$. Applications in other notepad Studentized Residuals

$$e_i^s = \frac{e_i}{s_{(i)}^2(1-h_i)}$$

where $e_i^s$ - studentized residual $s_{(i)}^2$ - standard deviation where $i$th obs is deleted $h_i$ - leverage statistic Belsley et al (1980) recommend the use of studentized Residuals to determine whether there is an outlier. Data set 1: Beckmans' aerosol data ( Zewotir) high efficiency particulate air filter. toxic dust, radionuclides, and mists/ Both Beckman and CPJ rank the 14th observation as the most influential. Data set 2: Metal oxide analysis data ( Zewotir) Process and measurement variation on the properties of lots of metal oxides. type 1 and 2 chemist 1 and 2 Sample 1 and 2 Maximum likelihood variance component ratios. Full data and with cases removed.

http://www.tandfonline.com/doi/abs/10.1080/03610910600716795 http://www.ingentaconnect.com/
http://www.sciencedirect.com/science/article/pii/S0047259X09001213

Margina Means ( population averaged) $E[Y_{ij}]$

Conditional means ( Cluster specific) $E[Y_{ij}|\gamma_i]$

# References


Beckman, R., C. Nachtsheim, and R. Cook (1987). Diagnostics for mixed-model analysis of variance. *Technometrics 29*(4), 413–426.

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.

Zewotir, T. and J. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of Data Science 3*, 153–177.