ON THE USE OF COOK'S DISTANCE FOR LONGITUDINAL MIXED EFFECTS REGRESSION MODEL

Frans E.S. Tan[1], Mario J.N. Ouwens, Martijn P.F. Berger

Department of Methodology and Statistics

University of Maastricht, The Netherlands

Abstract

Mixed effects models for longitudinal data with fixed as well as random parameters are often used to describe average profiles. Influence measures are usually constructed to detect influential subjects and observations for the fixed regression parameters, treating the subject specific parameters as nuisance parameters. One of these measures is the well-known Cook's Distance. We show that this statistic may fail to detect or may incorrectly detect influential observations due to the random effects variances and covariances. A conditional version of Cook's Distance is proposed that deals with the above mentioned problem.

Keywords: Influential Observations, Cook's Distance, Random effects, conditional Cook's Distance

---

[1] Correspondence to: Frans E.S. Tan, University of Maastricht, Department of Methodology and Statistics, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Phone: +31433882278. E-mail: Frans.Tan@Stat.Unimaas.Nl

An influence measure for longitudinal mixed effects regression models

1. Introduction

In the past decade there is a large number of literature on longitudinal models with random effects (Diggle, Liang and Zeger, 1994, Vonesh and Carter, 1987 among others). In spite of this growing interest, diagnostic methods in a longitudinal context remains a somewhat neglected topic. This article discusses the problem of identifying an observation with a demonstrably large impact on estimated regression parameters. Such observations are called influential observations by Belsley, Kuh, and Welsch (1980). Chatterjee and Hadi (1986) review several influence measures. Each of them stresses different aspects of influence on the calculated values of estimates. Some measures are constructed to detect observations that may influence some regression parameters, or a linear combination of these parameters and/or the estimated variance of these estimated regression parameters. Barrett and Ling (1992), Cook (1968), De Gruttola, Ware and Louis (1987), Rohlf (1975) and Siotani (1959) among others, have suggested extensions to multivariate regression. The existing influence measures for longitudinal data are developed in the context of fixed effects models with structured or unstructured variance-covariance matrices. One of these measures is Cook's Distance. Although it seems natural to use Cook's Distance as an influence measure for mixed effects regression models, some cautions should be made with respect to the application of this measure, Several suggestions have been made in Literature. Christensen, Pearson and Johnson, (1992) proposed a two step method to detect influential observations. In the first step a diagnostic tool is used to evaluate the influence on the estimates of the variance components. They noticed that observations that have a large impact on the estimated variance components will affect the detection of influential observations for the estimated regression parameters. In the second step they used the extension of Cook's Distance, Banerjee (1998) noticed that the effectiveness of Cook's Distance as an influence measure in the longitudinal data setting is limited. Banerjee and Frees (1997) have applied the concept of partial influence to take into account the effects on subject specific parameters and measure the impact of a subject on the population parameters. Recently, Lesaffre and Verbeke (1998) and Ouwens, Tan and Berger (1999) have used the local influence concept to detect influential subjects (second level) and observations (first level), respectively.

In this article we argue that Cook's Distance is rather insensitive to changes of the estimates of the subject specific regression parameters. This may occasionally lead to an observation being incorrectly detected as influential. A conditional version of Cook's Distance is proposed that addresses the above mentioned problem by first conditioning on the subjects in the sample. In section two the data from the London Growth Study (Tanner, Whitehouse, Marubini and Resele, 1976) are described and the problem of detecting influential observations is motivated. In section three the underlying regression model is specified. In section four Cook's Distance is investigated in detail. Examples will be given to demonstrate the shortcomings of Cook's Distance for mixed effects models. A conditional version is proposed in section five and some numerical examples are given that show the similarities and discrepancies between the two representations of Cook's Distance. Some general conclusions will also be discussed in detail. Finally, the London growth study data will be analysed for influential observations in section six.

2. The London Growth Study: Data description and problem formulation

A more elaborate description and analysis of the London Growth Study can be found in Tanner et al. (1976), as part of the so-called Harpenden Growth Study of several hundred boys and girls in a children's home in the country just outside London Between 1948 and 1972. Each child was observed two times a year until the first signs of puberty appeared, followed by measurements every three month until the growth spurt ends, and then each year until twenty and finally each five year. Tanner et al. (1976) have analysed 55 boys and 35 girls whose measurements were made regularly until they reached adult size. Goldstein (1979) selected growth data of twenty pre-adolescent girls between six and ten years old and named this the London Growth Study. One of the topics of the London Growth Study was to describe the mean growth of the girls for each of three groups of mothers: named short mother (< 155 cm). medium mother (155 - 164) and tall mother (>164). Table 1 shows the relationship between height and age of the girls for the three different groups. This table can also be found in Goldstein (1979). Inspection of the data reveals that the second observation of the fifth girl is perhaps a typo. During the analysis, the influence of this observation on the average rate of change should be evaluated. Furthermore, the influence of the other observations, which might not be apparent in this stage, should also be evaluated.

	Age				
Girl	6	7	8	9	10
Short mother					
1	111.0	116.4	121.7	126.3	130.5
2	110.0	115.8	121.5	126.6	131.4
3	113.7	119.7	125.3	130.1	136.0
4	114.0	118.9	124.6	129.0	134.0
5	114.5	112.0	126.4	131.2	135.0
6	112.0	117.3	124.4	129.2	135.2
Mean	112.5	116.7	124.0	129.2	135.2
Medium mother					
7	116.0	122.0	126.6	132.6	137.6
8	117.6	123.2	129.3	134.5	138.9
9	121.0	127.3	134.5	139.9	145.4
10	114.5	119.0	124.0	130.0	135.1
11	117.4	123.2	129.5	134.5	140.0
12	113.7	119.7	125.3	130.1	135.9
13	113.6	119.1	124.8	130.8	136.3
Mean	116.2	121.9	127.9	133.2	138.5
Tall mother					
14	120.4	125.0	132.0	136.6	140.7
15	120.2	128.5	134.6	141.0	146.5
16	118.9	125.6	132.1	139.1	144.0
17	120.7	126.7	133.8	140.7	146.0
18	121.0	128.1	134.3	140.3	144.0
19	115.9	121.3	127.4	135.1	141.1
20	125.1	131.8	141.3	146.8	152.3
Mean	120.3	126.7	133.6	139.9	144.9

Table 1

Data of the London Growth Study

3. Model Specification

Suppose $\boldsymbol{y}_i$, $i = 1,..., N$ is a vector of responses of subject i at n_i time points. These responses are described by the longitudinal mixed effects regression model:

$$\boldsymbol{y}_i = X_i\beta + Z_i\boldsymbol{b}_i + \epsilon_i \ , \qquad \textbf{(1)}$$

where the $n_i \times 1$ vectors $\epsilon_1, \epsilon_2,...,\epsilon_N$ are supposed to be measurement errors (or disturbances), which are independently normally distributed with mean zero and covariance matrix $\sigma^2 I_i$, $i = 1,...,N$, I_i is the identity matrix of rank n_i and σ^2 is the common variance of the measurement errors. β is a $p \times 1$ fixed effects regression parameter vector and $\boldsymbol{b}_i$ is a $q \times 1$ vector of random effects regression parameters, which are independently and normally distributed with mean zero and covariance matrix D. Finally, the matrices X_i and Z_i are the $n_i \times p$ and $n_i \times q$ design matrix of rank p and rank q, respectively. For longitudinal data the columns of the Z_i matrices are often functions of the time components. In general, the matrix X_i consists of time varying and time independent covariates. Furthermore, the Z_i matrices are submatrices of the X_i matrices. The variance of $\boldsymbol{y}_i$ is equal to $V_i = Z_i D Z_i' + \sigma^2 I_i$. The block diagonal matrix V is the matrix with V_i 's as block diagonal.

Model (1) may serve at least two different types of objectives. The first type of objectives brings into focus the group results. In this case, the expected value of the responses $\boldsymbol{y}_i$ is equal to the average profile $X_i\beta$ with variance-covariance matrix V_i. An influential observation in this context is supposed to have al large impact on the estimated fixed regression parameters, and the fitted response $\hat{\boldsymbol{y}}_i$ is equal to the estimated average profile $X_i\hat{\beta}$. The London growth study data dealt also with group results. The main objective was to compare mean growth of pre-adolescent girls for each of the three groups. The second type of objectives deals with the situation that the estimation of the subject specific profiles is of main interest. Examples can be found in educational research (Tan, 1994). One of the examples is the analysis of progress-test data. A progress-test intends to measure the progress of the students in a school, and describes and predicts the growth of knowledge of each student. In this case, the expected value of the responses $\boldsymbol{y}_i$ in model (1) is equal to the subject specific profile $X_i\beta + Z_i\boldsymbol{b}_i$ with variance σ^2. An influential observation in this context may have a large impact on the estimated subject specific parameters $\hat{\boldsymbol{b}}_i$ or on the estimated fixed parameters $\hat{\beta}$ or on both.

The regression parameters can be estimated, for example, by using the procedure proposed by Laird and Ware (1982). The variance-covariance matrices D and V of the random regression parameters are estimated by the M.L. method. $\hat{\beta}$ and $\hat{\boldsymbol{b}}_i$ are given by the following formulas:

$$\hat{\beta}=\left(\sum_{i=1}^{N} X_i\hat{V}_i^{-1}X_i\right)^{-1}\sum_{i=1}^{N} X_i\hat{V}_i^{-1}\mathbf{y}_i \qquad (2)$$

and

$$\hat{\boldsymbol{b}}_i=\hat{D}Z_i'\hat{V}_i^{-1}\left(\mathbf{y}_i-X_i\hat{\beta}\right), \qquad (3)$$

respectively. A method to detect influential observations for the fixed regression parameter is Cook's Distance, which will be discussed in the following section.

4. Cook's Distance

Cook's Distance is based on the concept of the influence function introduced by Hampel (1974). At population level an influence function measures the influence on a parameter when one observation is deleted. In a standard linear regression context, where the regression parameters are fixed, the finite sample approximation of the influence function leads to the distance measure developed by Cook (1977):

$$C_j=\frac{(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(j)})'V^{-1}(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(j)})}{p}. \qquad (4)$$

Where $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_{(j)}$ are the fitted values of $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots \mathbf{y}_N')'$, using the sample with and without observation j, respectively, and p is the number of independent design parameters. Note that the variance matrix V of $\mathbf{y}$ is generally unknown and must be estimated from the data. The popularity of Cook's Distance is partly due to the simplicity of its interpretation. C_j can be interpreted as the distance between the two vectors of fitted values when the fitting is done including or excluding the j-th observation. It measures the amount of perturbation of the fitted response $\hat{\mathbf{y}}$ due to an influential observation.

In particular, Cook's Distance identifies observations (level one) with an unusual large impact on the average profile $X\hat{\beta}$. Influential observations that have an unusual large impact on the subject specific parameters $\hat{b}$ cannot always be detected by Cook's Distance and may occasionally lead to an observation being incorrectly detected as influential. The following example may clarify these statements. The samples in the following illustrations contain 20 subjects, for which every subject is measured nine times at equidistant time points. The data are generated from the following underlying model:

$$y_k = (\beta_0 + b_{0k}) + \beta_1 t + \epsilon_k, \qquad \textbf{(5)}$$

where t is the time component, y_k is the response of subject k at time t, β_0 and β_1 are the fixed intercept and slope, respectively, and b_{0k} is the random intercept corresponding with the k-th subject, which is normally distributed with mean zero and variance δ^2. The scattergram of the relationship between the response y_k and the time points in Figure 1 shows a situation in which it is known that exactly one observation has a large impact on the estimated value of $\hat{b}_{0k}$. Without loss of generality we suppose that an observation of subject *8* has a large impact on the estimated value of $\hat{b}_{08}$. The relationship between the response y_8 for subject *8* and the time points is also made apparent in Figure 1. The data for this illustration are generated such that the outlying observation has a large impact on the estimated subject specific parameter $\hat{b}_{08}$ and hence on the estimated fixed intercept $\hat{\beta}_0$ but not on the estimated fixed slope $\hat{\beta}_1$.

Figure 1. Nine repeated measures for twenty subjects including subject 8

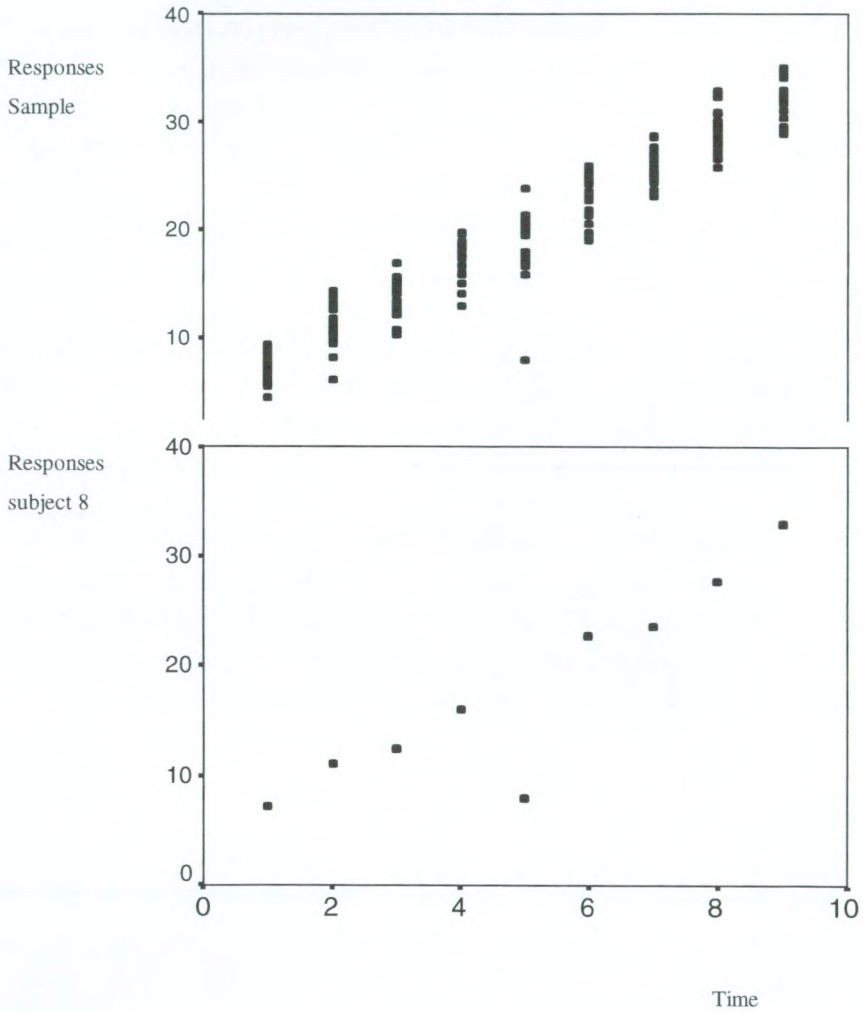


Figure 2 shows the relative changes (due to the deletion of a level one observation) in the estimated fixed intercept $\hat{\beta}_0$, the estimated fixed slope $\hat{\beta}_1$ and $\hat{b}_{0k}$, respectively. The relative change $\hat{b}_{0k}$ is calculated relative to the sum of the standard deviation δ and the true value of the fixed intercept.

Figure 2. Relative changes in $\hat{b}_{0k}$, $\hat{\beta}_0$ and $\hat{\beta}_1$

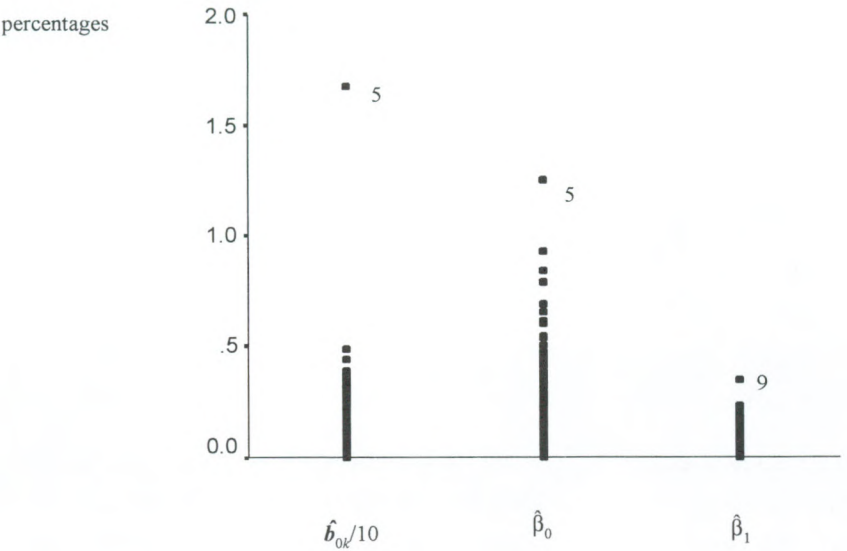
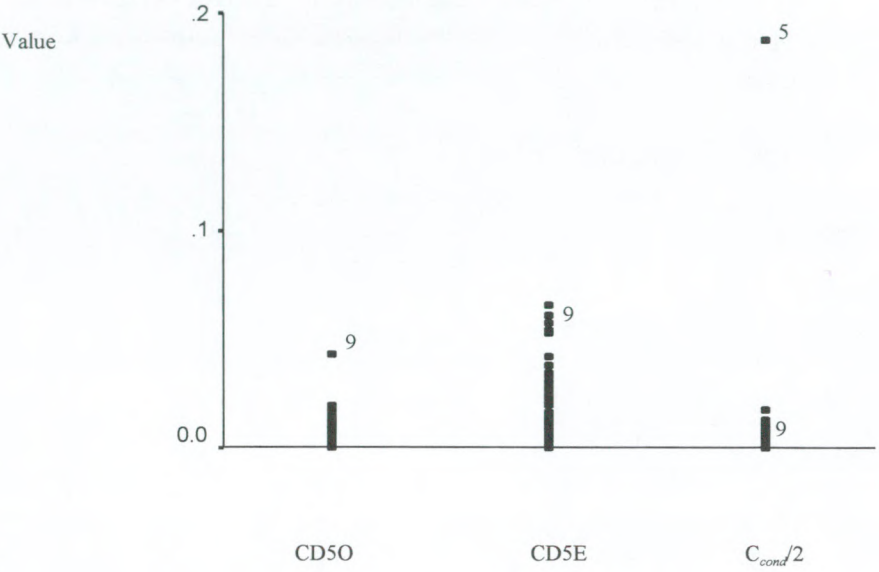


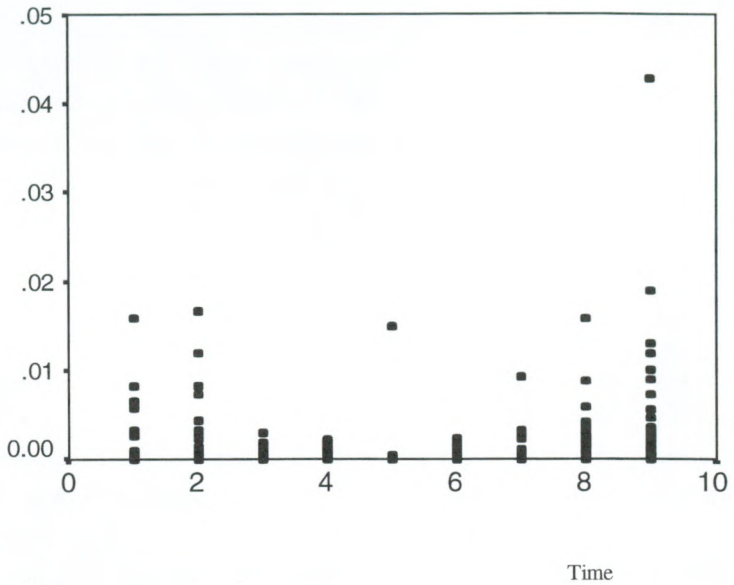
Figure 2 indicates that the fifth observation of subject *8* has the largest impact on $\hat{\beta}_0$, and the ninth observation of subject *8* has the largest impact on the estimated fixed slope $\hat{\beta}_1$. Figure 2 also shows that the value of the $\hat{b}_{08}$ is strongly influenced by the fifth observation of subject *8*. Figure 3 shows that Cook's Distance incorrectly recognizes the ninth observation being most influential. It indicates that the ninth instead of the fifth observation of subject *8* has the largest Cook's Distance (diagram CD50). If the fifth observation is set equal to the average value of subject *8* (diagram CD5E), then the value of Cook's Distance of the ninth observation becomes comparable to the other Cook's Distances. Obviously, the ninth observation of subject *8* is incorrectly recognized to be influential due to the influence of the fifth observation of that subject.

Figure 3. Cook's Distance and Conditional Cook's Distance



The influence of observations on the estimates of β depends on the specific time point. It is well known, that even if none of the observations has a significant impact on the regression parameters, Cook's Distances at the boundaries of the time interval will be larger than at the centre. An obvious way to solve this problem is to compare Cook's Distance per time point across subjects. Figure 4 shows the relationship between Cook's Distances and time points. In general, for each time point there may be one candidate influential observation. A possible procedure to detect an influential observation may be to determine the impact on the estimated regression parameters by sequentially deleting each combination of potentially influential observations. Such a procedure, however, may lead to a large number of possible combinations. Another approach, which may be more efficient, is the use of a conditional Cook's Distance which will be elaborated in the next section.

Figure 4. Cook's Distance over Time



5. Orthogonal decomposition of Cook's Distance and Conditional Cook's Distance

In the previous section the ninth observation of subject *8* was incorrectly detected as being most influential. In this section this incorrect result will be explained and the proposed conditional Cook's Distance will be motivated. Recall that we have started with model (1) as the specification of a mixed effects longitudinal regression model. Furthermore, the Z_i matrices are submatrices of the X_i matrices. Hence model (1) can be rewritten as

$$\begin{aligned} \boldsymbol{y}_i &= (Z_i\ A_i)\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\ &= Z_i\boldsymbol{\beta}_1 + A_i\boldsymbol{\beta}_2 + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i\ , \end{aligned} \tag{6}$$

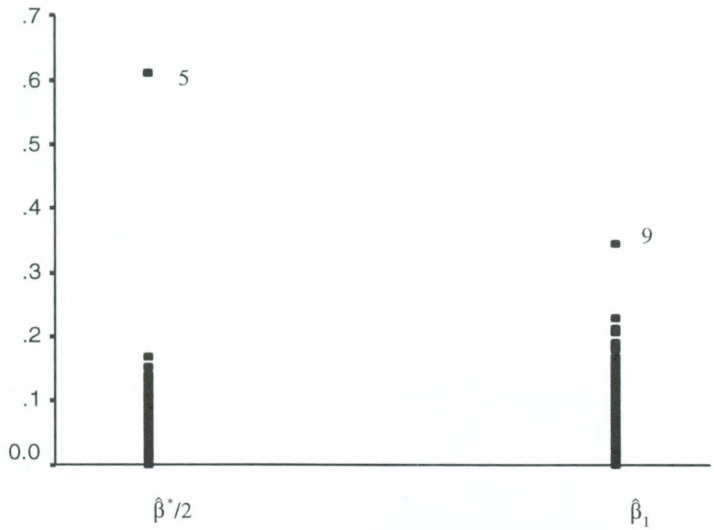
with A_i a n_i x $(p\text{-}q)$ design matrix, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. In general, if deletion of an observation

has an impact on the estimate of $A_i \beta_2$, then the estimate of $Z_i \beta_1$ will also be influenced. The amount of influence depends on both design matrices Z_i and A_i. For example, for model (5), if the axis are located around or far away from the centre point of the scattergram, a change in the estimated slope $\hat{\beta}_1$ due to an influential observation leads to a small or a large change in the estimated intercept $\hat{\beta}_0$, respectively. In order to evaluate the change of a specific fixed parameter estimate not attributed to the change of the other fixed parameter estimates. An orthogonal decomposition of the matrix X_i is constructed, such that

$$y_i = Z_i \beta_1^* + Z_i^{\perp} \beta_2 + Z_i b_i + \epsilon_i, \qquad (7)$$

where β_1^* is a function of all fixed parameters, and $Z_i^{\perp}$, is the design matrix orthogonal to Z_i. In model (5), a change in the estimate $\hat{\beta}_1^*$ represents a shift, and a change in the estimate $\hat{\beta}_2$ represents a rotation around the center point of the scattergram. Figure 5 shows that deletion of the fifth observation of subject *8* causes the largest shift and deletion of the ninth observation of the same subject causes the largest rotation. Furthermore, a properly chosen influence measure should indicate the fifth observation as being most influential. It can be shown that (Theorem appendix), using Cook's Distance according to equation (4), the contribution to the Cook's Distance score due to a change in $\hat{\beta}_1^*$ is weighted by a monotone function of the inverse of the variance covariance matrix of the random parameters. The contribution due to a change in $\hat{\beta}_2$ is not weighted by a function of this matrix. For example, in the random intercept case the shift parameter estimate is divided by a monotone function of the variance of the random intercept. The effect of the fifth observation on the estimated fixed intercept is larger than on the estimated fixed slope, whereas the effect of the ninth observation on the estimated fixed slope is larger than on the estimated fixed intercept. Consequently, the effect of the fifth observation will be underestimated. It should be noted that a change in the shift parameter estimate is fully ascribed to a change in the subject specific intercept estimates $\hat{b}_{0i}$, i=,...,N. Cook's Distance is obviously less sensitive for changes in $\hat{b}_i$ due to the division by the variances and covariances of the random parameters. We argue that the effect of an observation on an estimated fixed parameter should not be normalized by the between subject variance and covariances.

Figure 5. Relative change in $\hat{\beta}_1$ and $\hat{\beta}^*$



These variances and covariances measure differences between subjects, whereas the effect of deleting an observation reflects the change of an estimated subject parameter irrespective of the between subject variance. This unpleasant feature of Cook's Distance for mixed effects models can be overcome by first conditioning on the subjects in the sample. This conditioning implies that the $\boldsymbol{b}_i$'s are considered to be fixed parameters, and thus we do not have to deal with the variances and covariances of the random parameters. The variance of $\boldsymbol{y}$, conditionally on the subjects in the sample is then equal to the measurement error variance σ^2. The conditional version of Model (1) can be expressed as follows:

$$\boldsymbol{y}=U\boldsymbol{\gamma}+\boldsymbol{\epsilon}, \tag{8}$$

$$U=\begin{bmatrix} X_1 & Z_1 & 0 & 0 & \dots & 0 \\ X_2 & 0 & Z_2 & 0 & \dots & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ X_N & 0 & \dots & \dots & 0 & Z_N \end{bmatrix},$$

Where U is of the form

X_i and Z_i are an n_i x p design matrix of fixed parameters of rank p and an n_i x q design matrix of random effects of rank q, respectively. $\boldsymbol{\gamma}$ is given by

$$\boldsymbol{\gamma}=\left[\beta' \ \boldsymbol{b}_1' \ \boldsymbol{b}_2' \ . \ . \ . \ \boldsymbol{b}_N'\right]', \tag{9}$$

and $\boldsymbol{\epsilon}=(\boldsymbol{\epsilon}_1',\dots,\ \boldsymbol{\epsilon}_N')'$ is the vector of measurement errors.

Conditioning on $\boldsymbol{\gamma}$ and thus on the $\boldsymbol{b}_i$'s and on β implies that the variance matrix of $\boldsymbol{y}$ is equal to $\sigma^2 I$, which is the block-diagonal matrix containing as ith block $\sigma^2 I_i$. Unfortunately, the regression parameters are not identified, because the design matrix U is in general not of full rank. U can only be of full rank if X_i does not contain time independent covariates. Therefore, we propose to use the Laird and Ware estimators as done in the unconditional case. Moreover, by choosing the same set of point estimates, Cook's Distance can be compared with the conditional Cook's Distance. Using the concept of Cook's Distance, the conditional Cook's Distance (conditional on the subjects) is equal to

$$C_{cond_j} = \sum_{i=1}^{N} \frac{((X_i\hat{\beta}+Z_i\hat{\boldsymbol{b}}_i)-(X_i\hat{\beta}_{(j)}+Z_i\hat{\boldsymbol{b}}_{i(j)}))'((X_i\hat{\beta}+Z_i\hat{\boldsymbol{b}}_i)-(X_i\hat{\beta}_{(j)}+Z_i\hat{\boldsymbol{b}}_{i(j)}))}{\sigma^2((N-1)q+p)}, \tag{10}$$

Let $k=(N-1)q+p$, then expression (10) can be decomposed as:

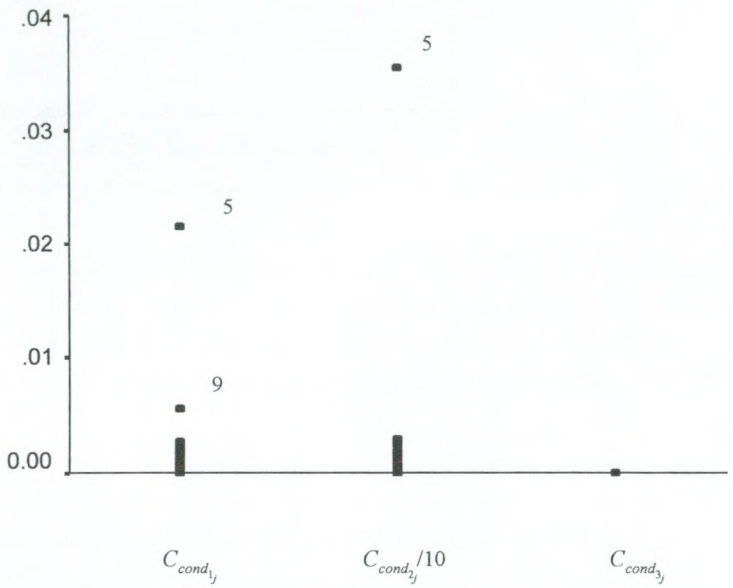
$$C_{cond_j} = C_{cond_{1j}} + C_{cond_{2j}} + C_{cond_{3j}} =$$

$$\frac{(\hat{\beta}-\hat{\beta}_{(j)})'X'X(\hat{\beta}-\hat{\beta}_{(j)})}{k\sigma^2} + \frac{\sum_{i=1}^{N}(\hat{b}_i-\hat{b}_{i(j)})'Z_i'Z_i(\hat{b}_i-\hat{b}_{i(j)})}{k\sigma^2} + \frac{2(\hat{\beta}-\hat{\beta}_{(j)})'\sum_{i=1}^{N}X_i'Z_i(\hat{b}_i-\hat{b}_{i(j)})}{k\sigma^2} . \quad (11)$$

The first term $C_{cond_{1j}} = \frac{(\hat{\beta}-\hat{\beta}_{(j)})'X'X(\hat{\beta}-\hat{\beta}_{(j)})}{k\sigma^2}$, can be interpreted as a distance measure for the estimated fixed effects parameters. This term is equal to the Cook's Distance, but without normalizing random variance and covariances (see corollary appendix). The second term $C_{cond_{2j}} = \sum_{i=1}^{N}\frac{(\hat{b}_i-\hat{b}_{i(j)})'Z_i'Z_i(\hat{b}_i-\hat{b}_{i(j)})}{k\sigma^2}$ can be interpreted as a distance measure for the change in the estimated subject specific regression parameters. A major advantage of the decomposition (11) is that the total amount of influence as given by formula 10 can be divided into an amount of influence on the overall parameters and an amount of influence on the subject specific parameters.

To illustrate the merits of the conditional Cook's Distance the random intercept example is used again, where the fifth observation of subject *8* was influential. The conditional Cook's Distances are plotted in figure 3. Figure 3 indicates that the fifth observation has a large conditional Cook's Distance, and thus the conditional Cook's Distance correctly detects the fifth observation as influential. Figure 6 shows the values of the decomposition terms of formula 11 for the illustration mentioned above. Note that the fifth observation has a relatively large impact on $\hat{\beta}_1^*=\hat{\beta}_0+5\hat{\beta}_1$ as well as on $\hat{b}_{0k}$. Unlike Cook's Distance, the conditional Cook's Distance does not normalize the effect of the fifth observation by a function of the random intercept variance. Figure 6 makes apparent that this variance is substantial, because now the fifth observation of subject *8* instead of the ninth has the largest influence value. The second term shows that the fifth observation has the largest effect on the estimated subject parameters. The first term of the decomposition also emphasizes that this fifth observation is an influential observation for the estimated overall intercept as well. As demonstrated in this figure, the last term of the decomposition are extremely small .

Figure 6. Values of the three terms of the decomposition



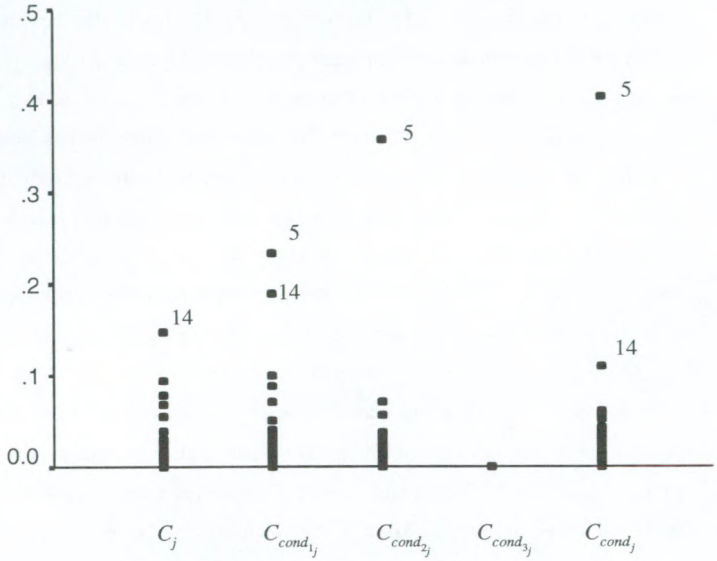
6. The London Growth Study: An analysis for influential observations

The model that we use for the London growth study is a random intercept longitudinal regression model with time versus group interaction. Note that Goldstein (1979) has used a fixed effects model with time versus group interaction. The inter-subject variation for each of the three groups is small. The discrepancy between Cook's Distance and its conditional version is mainly due to this variation. Therefore, the first and second group are combined to obtain a larger inter-subject variation. The model is specified as follows.

$$y_i = (\beta_0 + b_{0i}) + \beta_1 G + \beta_2 t + \beta_3 G * t + \epsilon_i, \qquad (11)$$

where G is the indicator for tall mothers, β_j, $j = 0,...,3$ are fixed regression parameters, ϵ_i is the error term, b_{0i} is the intercept of subject i, and y_i is the length of girl i at time point (age) t. Figure 7 shows the Cook's Distances, the conditional Cook's Distances and the values of the three terms of decomposition (11). Cook's Distance C_j indicates that the fifth observation of the first girl in the second group (girl number 14) is most influential. The conditional Cook's Distance C_{cond_j}, however, indicates that the second observation of the fifth girl of the first group (girl number 5) is most influential. The same observation is indicated as most influential by the first and the second terms $C_{cond_{1j}}$ and $C_{cond_{2j}}$, respectively. The fifth observation of first girl of the second group is also indicated by the first term as influential. A closer inspection of the data in Table 1 shows, that the growth curve of the height of girl is flatter than the other curves in her group. Table 2 shows that deleting observation five changes upwards the fixed regression slope estimate by means of $\hat{\beta}_3$. Cook's Distance correctly detects this change. The average growth of the first group remains unchanged. Furthermore, the second observation of the fifth girl of the first group is lower than the first observation of the same person. Deleting this observation changes upwards the estimated intercept $\hat{\beta}_0$ as well as the estimated slope $\hat{\beta}_1$ of the average growth. The average growth of the second group remains unchanged. Cook's Distance does not detect this change. The conditional Cook's Distance however, points out that the deletion of this observation has the largest impact on a fixed parameter- and subject parameter estimate as well. This illustration demonstrates that Cook's Distance is less sensitive for influential observations on the fixed parameters associated to the random effects and that the second term of the conditional Cook's Distance is sensitive for observations that are influential for the random effects.

Figure 7. Cook's Distances, values of the three terms of the decomposition and Conditional Cook's Distances



	Sample	Sample without 2nd observation of girl number 5	Sample without 5th observation of girl number14	Sample without both observations
$\hat{\beta}_0$	81.437 (1.218)	82.050 (1.083)	81.437 (1.191)	82.050 (1.050)
$\hat{\beta}_1$	1.686 (2.059)	1.073 (1.828)	0.939 (2.030)	0.331 (1.782)
$\hat{\beta}_2$	5.506 (0.102)	5.445 (0.076)	5.506 (0.098)	5.445 (0.070)
$\hat{\beta}_3$	0.742 (0.173)	0.804 (0.128)	0.849 (0.170)	0.910 (0.120)

Table 2

Estimates and Variances of the Elements of β

7. Discussion

In this article we argue that influential observations that have a large impact on the subject specific parameters cannot always be detected by Cook's Distance due to large between subject variation. Since some of the fixed regression parameters are more sensitive to changes of subject specific parameters, the problem of influential observations should be approached locally. A conditional version of Cook's Distance is proposed that deals with this problem. The conditional measure can be decomposed in a part that measures the influence in the estimated fixed parameters, a part that measures the influence in the estimated subject specific parameters, and a part that can be neglected.

In general, the error terms in model (1) are correlated due to within individual serial correlations. However, in specific applications the effect of serial correlation may be dominated by the combination of random effects and measurement error (Diggle et al. 1994, page 88). In this article we restrict ourselves to uncorrelated error terms. Note that, if an observation has a large influence on an estimated subject parameter, then deletion of this observation will affect the variances and covariances of the random parameters and hence V. The second term, $C_{cond_{2j}}$, of the decomposition (11) replaces the variance component analysis of Christensen et al. (1992).

In practice, subject oriented as well as observation oriented diagnostics should be applied. The proposed method is not meant to be used at subject level. Analysis at subject level alone is not sufficient. Instead, one could think of a combination of the method proposed by Banerjee and Frees (1998) and the conditional Cook's Distance. Based on this two step approach it appears that, at subject level, subject *9* of the London Growth Study is an influential subject.

In conclusion, we argue that for mixed effects longitudinal model the unconditional Cook's Distance is more suitable than Cook's Distance to detect influential (level one) observations. An influential observation is primarily influential for the estimated regression parameters of a specific subject. The amount of influence of an observation should not depend on the amount of variation between subjects. Needless to say that , for fixed effects models both measures are the same. The error terms are supposed to be uncorrelated. Hence, for fixed effects models we are dealing with the common OLS Cook's Distance. The advantage of conditional Cook's Distance arises when random parameters are present.

Literature

Banerjee, M. (1998). Cook's Distance in linear longitudinal models, *Commun. Statist-Theory Meth.,* 27(12),2973-2983

Banerjee, M. And Frees, E.W. (1997). Influence diagnostics for linear longitudinal models, *Journal of the American Statistical Association,* vol. 92, 439, 999-1005

Barrett, B.E. and Ling, R.F. (1992). General Classes of Influence Measures for Multivariate Regression, *Journal of the American Statistical Association, 87,* 184-191

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

Chatterjee, S. and A.S. Hadi (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression, *Statistical Science, 1,* 379-416

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

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics,* 19, 15-18

Cook, R.D. (1986). Assessment of local influence (with discussion), *Journal of the Royal Statistical Society Sector B, 48,* 133-169.

De Gruttola, V., Ware, J.H., and Louis, T.A. (1987). Influence Analysis of Generalized Least Squares Estimators, *Journal of the American Statistical Association, 82,* 911-917.

Diggle, P.J., Liang, K.Y., Zeger, S.L. (1994). *Analysis of longitudinal data,* Oxford: Clarenden Press.

Goldstein, H. (1979). *The design and analysis of longitudinal studies,* London: academic press

Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association, 69,* 383-393.

Lesaffre, E. And Verbeke, G. (1998). Local influence in linear models, *Biometrics,* 54, 570-582.

Ouwens, J.N.M., Tan, F.E.S., and Berger, M.P.F. (1999). *Local influence for repeated measures generalized linear mixed models.* In H. Friendl, A. Berghold, G. Kauermann (eds.), proceedings of the 14th International Workshop on Statistical Modelling, 308-316, Graz, Austria.

Rohlf, F.J. (1975). Generalization of the Gap Test for the Detection of Multivariate Outliers, *Biometrics., 31,* 93-101.

Siotani, M. (1959), The Extreme Value of the Generalized Distances of the Individual Points in the Multivariate Normal Sample, *Annals of the Institute of Statistical Mathematics, 10,* 183-208.

Tan, E.S. (1994), *A stochastic growth model for the longitudinal measurement of ability.* Ph.D Thesis, Maastricht University. Maastricht: Datawyse Maastricht.

Tanner, J.M., Whitehouse, R.H., Marubini, E., and Resele, L.F. (1976). The adolescent growth spurt of boys and girls of the Harpenden Growth Study, *Annals of Human Biology, 3,* 109-126.

Vonesh, E.F., and Carter, R.L., (1987), Efficient inference for random-coefficient growth curve models with unbalanced data, *Biometrika 43,* 617-628.

Appendix

In this appendix the relationship between Cook's Distance and the first term of the Conditional Cook's Distance is derived. Given model (1), Z_i is a submatrix of X_i. Consequently the X_i-space can be decomposed in the Z_i -space and the subspace of X_i orthogonal to the Z_i -space. It follows that for every i the change in the estimates of $\boldsymbol{y}_i$ is equal to

$$X_i(\hat{\beta}-\hat{\beta}_{(j)})=Z_i\boldsymbol{u}_i+Z_i^{\perp}\boldsymbol{v}_i \tag{A1}$$

for some vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, and $Z_i^{\perp}$ a matrix of full rank for which the columns are vectors in the X_i-space and are orthogonal to the columns of Z_i .

Theorem

Suppose that $\{\boldsymbol{s}_{i1},...,\boldsymbol{s}_{ik_i}\}$ is an orthonormal basis for the kernel of Z_iDZ_i' and $\{\boldsymbol{t}_{i1},...,\boldsymbol{t}_{i\,lm_i}\}$ is an orthonormal basis of eigenvectors of the image of Z_iDZ_i', with eigenvalues λ_{im} . Let $<,>$ be the inproduct. p times Cook's Distance is then equal to

$$pC_j=\sum_{i=1}^{N}\left(\sum_{m=1}^{k_i}\frac{<\boldsymbol{s}_{im},Z_i^{\perp}\boldsymbol{v}_i>^2}{\sigma^2}+\sum_{m=1}^{lm_i}\frac{<\boldsymbol{t}_{im},Z_i\boldsymbol{u}_i>^2}{\sigma^2+\lambda_{im}}\right) \tag{A2}$$

Proof:

$$pC_j=(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}_{(j)})'V^{-1}(\hat{\boldsymbol{y}}-\hat{\boldsymbol{y}}_{(j)})=\sum_{i=1}^{N}(\hat{\beta}-\hat{\beta}_{(j)})'X_i'V_i^{-1}X_i(\hat{\beta}-\hat{\beta}_{(j)}) \tag{A3}$$

Using the union of $\{\boldsymbol{s}_{i1},...,\boldsymbol{s}_{ik_i}\}$ and $\{\boldsymbol{t}_{i1},...,\boldsymbol{t}_{i\,lm_i}\}$ as a basis for the X_i -space $X_i(\hat{\beta}-\hat{\beta}_{(j)})$ can be written as

$$\sum_{m=1}^{k_i}<\boldsymbol{s}_{im},X_i(\hat{\beta}-\hat{\beta}_{(j)})>\boldsymbol{s}_{im}+\sum_{i=1}^{lm_i}<\boldsymbol{t}_{im},X_i(\hat{\beta}-\hat{\beta}_{(j)})>\boldsymbol{t}_{im}. \tag{A4}$$

The $\boldsymbol{s}_{im}$'s and $\boldsymbol{t}_{im}$'s are eigenvectors of the matrix V^{-1} with eigenvalues σ^2 and σ^2

$+ \lambda_{im}$, respectively. Consequently, the Cook's Distance can be written as

$$\sum_{i=1}^{N}\left(\sum_{m=1}^{k_i}\frac{<\boldsymbol{s}_{im},X_i(\hat{\beta}-\hat{\beta}_{(j)})>^2}{\sigma^2}+\sum_{m=1}^{Im_i}\frac{<\boldsymbol{t}_{im},X_i(\hat{\beta}-\hat{\beta}_{(j)})>^2}{\sigma^2+\lambda_{im}}\right). \qquad \text{(A5)}$$

It will be proven below that the $\boldsymbol{s}_{im}$'s are orthogonal to the columns of the Z_i matrix and that the $\boldsymbol{t}_{im}$'s are linear combinations of the columns of Z_i. Expression (A2) can then be derived from expression (A5) using expression (A1) and the statements in the previous sentence, which completes the proof.

$\boldsymbol{s}_{im}$ is orthogonal to the columns of Z_i, for every m, $m = 1,\ldots,k_i$.

Proof:

Take an arbitrary $i \leq N$ and $m \leq k_i$. Then $0=Z_iDZ_i'\boldsymbol{s}_{im}=\boldsymbol{s}_{im}'Z_iDZ_i'\boldsymbol{s}_{im}$. Since D is of full rank it follows that $Z_i'\,\boldsymbol{s}_{im} = 0$. Consequently $\boldsymbol{s}_{im}$ is orthogonal to the columns of Z_i. Because i and m were arbitrary, this holds for all $i \leq N$ and $m \leq k_i$.

The t_{im}'s are linear combinations of the columns of Z_i

Proof:

Take an arbitrary $i \leq N$ and $m \leq Im_i$. Then $\boldsymbol{t}_{im}$ can be written as $\boldsymbol{t}_{im} = Z_i\boldsymbol{b} + \boldsymbol{d}$, for some $\boldsymbol{b}$ and $\boldsymbol{d}$, where $\boldsymbol{d}$ is orthogonal to the columns of Z_i. Furthermore $\boldsymbol{t}_{im}$ is an eigenvector of Z_iDZ_i'. Note that λ_{im} is larger than zero. The result can be derived as follows:

$$\boldsymbol{d}'\boldsymbol{d}=\boldsymbol{d}'(Z_i\boldsymbol{b}+\boldsymbol{d})=\boldsymbol{d}'(\frac{Z_iDZ_i'}{\lambda_{im}}(Z_i\boldsymbol{b}+\boldsymbol{d}))=(\boldsymbol{d}'Z_i)\frac{DZ_i'}{\lambda_{im}}(Z_i\boldsymbol{b}+\boldsymbol{d}))=0. \qquad \textbf{(A6)}$$

This implies that $\boldsymbol{d}$ is zero and thus that $\boldsymbol{t}_{im}$ is a linear combination of the columns of Z_i. Because i and m were arbitrary, this holds for all i, $i \leq N$ and $m \leq Im_i$.

□

Corollary

Let k be the denominator of the Conditional Cook's Distance $(p + (N\text{-}1)q)$. k times the first term of the Conditional Cook's Distance is then equal to

$$kC_{cond_{1_j}} = \sum_{i=1}^{N} \left(\sum_{m=1}^{k_i} \frac{\langle s_{im}, Z_i^{\perp} v_i \rangle^2}{\sigma^2} + \sum_{m=1}^{lm_i} \frac{\langle t_{im}, Z_i u_i \rangle^2}{\sigma^2} \right), \qquad \textbf{(A7)}$$

and the relationship between Cook's Distance and the first term of the Conditional Cook's Distance is given by

$$kC_{cond_{1_j}} = pC_j + \sum_{i=1}^{N} \sum_{m=1}^{lm_i} \frac{\lambda_{im}}{\lambda_{im}+\sigma^2} \frac{\langle t_{mi}, Z_i u_i \rangle^2}{\sigma^2}. \qquad \text{(A8)}$$

Ontvangen: 3 december 1998
Geaccepteerd: 10 november 1999