Journal of Multivariate Analysis 100 (2009) 2002–2017

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Influence diagnostics and outlier tests for varying coefficient mixed models

Zaixing Li [a], Wangli Xu [b], Lixing Zhu [b,c,*]

[a] China University of Mining & Technology, Beijing, China
[b] RenMin University of China, Beijing, China
[c] Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 28 July 2007
Available online 23 June 2009

AMS 2000 subject classifications:
primary 62G35
secondary 62G08
62G10

Keywords:
Conditional influence
Cook's distance
"Delete=Replace" identity
Influence diagnostics
Joint influence
Outlier tests

ABSTRACT

In this paper, we consider subset deletion diagnostics for fixed effects (coefficient functions), random effects and one variance component in varying coefficient mixed models (VCMMs). Some simple updated formulas are obtained, and based on which, Cook's distance, joint influence and conditional influence are also investigated. Besides, since mean shift outlier models (MSOMs) are also efficient to detect outliers, we establish an equivalence between deletion models and MSOMs, which is not only suitable for fixed effects but also for random effects, and test statistics for outliers are then constructed. As a byproduct, we obtain the nonparametric "delete = replace" identity. Our influence diagnostics methods are illustrated through a simulated example and a real data set.

1. Introduction

Since not all the observations in a data set play an equal role in determining estimators, tests and other statistics, it is important to detect influential observations in data analysis. To identify anomalous observation(s), various approaches have been proposed in the literature, including deletion diagnostics, MSOMs and local influence analysis.

For linear models (LMs), [1] used case deletion to identify influential observations and defined a distance that is now termed as Cook's distance in the literature, to measure the effect of removing one observation on estimator or fitted value. Besides, [2,3], among others, studied diagnostics for generalized linear models. Smoothing methods which allow for a nonparametric relationship to be estimated at least for some of the predictors were also considered. For instance, [4–6].

Some works have been done for correlated data. For example, [7,8] discussed LMs; [9,10] considered linear mixed models (LMMs). Haslett and Dillane [11] proved a 'delete = replace' identity in LMs and applied it to deletion diagnostics for estimators of variance components. Xiang, Tse and Lee [12] investigated generalized linear mixed models.

However, except that [13] studied influence diagnostics and outlier tests for semiparametric mixed models (SMMs), to the best of our knowledge, there are few references about influence analysis when the nonparametric part appears in correlated data. As an extension of LMMs, VCMMs can be used more reasonably to represent underlying covariate effects by representing these covariate effects by smooth but otherwise arbitrary functions of time, with random effects used to model

* Corresponding author at: RenMin University of China, Beijing, China.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0047-259X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.06.008

the correlation induced by among-subject and within-subject variation. In practice, varying coefficient models (VCMs) other than VCMMs have been studied in the literature such as [14–17]. Zhang [18] first introduced generalized varying coefficient mixed models (GVCMMs) for longitudinal data and considered testing whether the coefficient function is polynomial or not. Except for this, little is found about VCMMs. Its influence analysis is certainly of interest and becomes the target of the paper. The following points are worth to mentioning.

- Compared with [13], where, given the smoothing parameter and the covariance matrix of Gaussian responses, deletion diagnostics for fixed effects was considered, we investigate influence diagnostics for fixed effects, random effects and the deletion updated formula for the variance component in VCMMs. Furthermore, the normality assumption on the distribution of the involved variables, which is commonly assumed in the literature of mixed models, can be avoided.
- For LMs [19], a broader class of parametric models [20], and SMMs [13], the equivalence, of the estimators for fixed effects, between the subset deletion model and the MSOM has been established. In this paper, we first show that the same phenomenon holds not only for the estimators of fixed effects but also for the predictors of random effects. More than this, we establish the equivalence between the estimators of the indicator parameters in MSOMs and the conditional residual predictors in deletion models. Our methods can also be used to other mixed models, such as SMMs, to obtain the equivalence.
- Under the normality assumption on random effects and errors, the intuitive "delete = replace" identity for nonparametric cases is also established from another perspective.

The rest of the paper is organized as follows. In Section 2, we present two kinds of estimating the coefficient functions and of predicting the random effects in VCMMs with general random errors. The updated formulas of subset deletion diagnostics for fixed effects, random effects and the variance component are developed in Section 3. The intuitive "Delete = Replace" identity in nonparametric cases is obtained. Moreover, Cook's distance, joint influence and conditional influence based on these updated formulas are also investigated. In Section 4, we present another diagnostic method—MSOMs and show a close connection to deletion diagnostics. The methods are illustrated through a simulated data set and a real data set in Section 5. All the proofs for the theorems are postponed in the Appendix. Throughout the paper, we shall use letters in bold to denote matrices and vectors.

2. VCMM and estimation

In this section, we present the VCMM and related estimators and predictors. The VCMM under study is as

$$y_{ij} = \sum_{l=1}^{p} x_{ij}^{(l)} \beta_l(t_{ij}) + \mathbf{z}_{ij}{}^{\tau}\mathbf{b}_i + \epsilon_{ij} \quad (j = 1, \ldots, n_i;\ i = 1, \ldots, m) \tag{1}$$

where y_{ij} is the response for the ith subject at time point t_{ij}; $\mathbf{b}_i$ having mean zero and covariance $\sigma_b^2\mathbf{D}_i$, are independent $q_i \times 1$ vectors of random effects with covariates $\mathbf{z}_{ij}$; $\beta_l(\cdot)(l = 1, \ldots, p)$ that are associated with covariates $x_{ij}^{(l)}$, are twice-differentiable smooth arbitrary functions on some finite interval.

For the sake of convenience, we rewrite model (1) in a matrix form. Denote the subject-specific vectors by $\mathbf{X}_i^{(l)} = \mathrm{diag}(x_{i1}^{(l)}, \ldots, x_{in_i}^{(l)})$, $\mathbf{Z}_i = (\mathbf{z}_{i1}^{\tau}, \ldots, \mathbf{z}_{in_i}^{\tau})^{\tau}$, $\mathbf{Y}_i$, and $\boldsymbol{\epsilon}_\mathbf{i}$ are defined similarly. Let $\mathbf{t}^0 = (t_1^0, \ldots, t_r^0)^{\tau}$ be the vector of ordered distinct values of the time points $\{t_{ij} : j = 1, \ldots, n_i; i = 1, \ldots, m\}$ with r being the number of distinct time points, $\boldsymbol{\beta}_l = (\beta_l(t_1^0), \ldots, \beta_l(t_r^0))^{\tau}$ $(l = 1, \ldots, p)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\tau}, \ldots, \boldsymbol{\beta}_p^{\tau})^{\tau}$. Let $\mathbf{N}_l (1 \le j \le p)$ be the incidence matrix mapping $\{t_{ij}\}$ to t^0, we have $(\beta_l(t_{11}), \ldots, \beta_l(t_{mn_m}))^{\tau} = \mathbf{N}_l\boldsymbol{\beta}_l$, for $1 \le l \le p$. Write $\mathbf{N} = \mathrm{diag}(\mathbf{N}_1, \ldots, \mathbf{N}_p)$, and similarly for $\mathbf{Z}$. Let $\mathbf{Y}$, $\mathbf{b}$ and $\boldsymbol{\epsilon}$ denote the vectors obtained from stacking up the m subject-specific vectors of the symbol, for instance, $\mathbf{b} = (\mathbf{b}_1^{\tau}, \ldots, \mathbf{b}_m^{\tau})^{\tau}$. Besides, let $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ and $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{N}$.

Thus, model (1) can be rewritten as

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}. \tag{2}$$

The covariance matrix of $\mathbf{b}$ is equal to $\sigma_b^2\mathbf{D}$ with $\mathbf{D} = \mathrm{diag}(\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_m)$. $\mathbf{D}$ is assumed to be given and the covariance matrix of the random error vector $\boldsymbol{\epsilon}$ is assumed to be $\sigma_\epsilon^2\mathbf{R}$ with a given $\mathbf{R}$ and an unknown σ_ϵ^2. The assumptions are identical to those in Crainiceanu and Ruppert [21]. In this paper, $\sigma_b^2/\sigma_\epsilon^2$ and σ_b^2 or σ_ϵ^2 are assumed to be known. For notational simplicity, $\sigma_b^2/\sigma_\epsilon^2\mathbf{D}$ is still denoted by $\mathbf{D}$. Thus, the covariance matrix of $\mathbf{Y}$ is $\sigma_\epsilon^2\mathbf{V}$ with $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}^{\tau}$. Note that in the form, although it seems a linear model, for every observation, there is an unknown parameter $\beta_{ij} = \beta(t_{ij})$. In other words, it is of course not a parametric model at all. We write it in a linear structure only for convenience of presentation for the later development in the following sections. In the next section, we use a nonparametric estimation procedure to estimate the unknowns in the model.