

Deletion diagnostics for the generalised linear mixed model with independent random effects

B. Ganguli,^{a,*†} S. Sen Roy,^a M. Naskar,^b E. J. Malloy^c and E. A. Eisen^d

The Generalised linear mixed model (GLMM) is widely used for modelling environmental data. However, such data are prone to influential observations, which can distort the estimated exposure–response curve particularly in regions of high exposure. Deletion diagnostics for iterative estimation schemes commonly derive the deleted estimates based on a single iteration of the full system holding certain pivotal quantities such as the information matrix to be constant. In this paper, we present an approximate formula for the deleted estimates and Cook’s distance for the GLMM, which does not assume that the estimates of variance parameters are unaffected by deletion. The procedure allows the user to calculate standardised DFBETAs for mean as well as variance parameters. In certain cases such as when using the GLMM as a device for smoothing, such residuals for the variance parameters are interesting in their own right. In general, the procedure leads to deleted estimates of mean parameters, which are corrected for the effect of deletion on variance components as estimation of the two sets of parameters is interdependent. The probabilistic behaviour of these residuals is investigated and a simulation based procedure suggested for their standardisation. The method is used to identify influential individuals in an occupational cohort exposed to silica. The results show that failure to conduct post model fitting diagnostics for variance components can lead to erroneous conclusions about the fitted curve and unstable confidence intervals. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Cook’s distance; deletion diagnostics; DFBETAs; exposure–response; generalised linear mixed models.

1. Introduction

1.1. Influential observations in environmental epidemiology

The motivation for this paper arises out of the need to simultaneously address the twin issues of modelling nonlinear exposure–response relationships while discounting the effect of influential observations for a cohort study of lung cancer and silica exposure [1].

Heterogeneity in response commonly leads to unusual shapes of the dose response in survival analysis [2]. Concurrently, smoothing methods have become also common for analyzing such data. Smoothing is particularly appropriate for survival analysis of large occupationally exposed cohorts because the exposure–response relationship is often attenuated with a rising relative risk that flattens or declines at the highest exposures [3]. The impact of observations with the highest values of exposure on the shape of the dose response curve is of great interest because a plateau or decline in risk at high levels may appear to undercut evidence for a causal relationship. Ad hoc procedures such as omitting the highest percentiles of the exposure are common but as our application section shows, influential values need not be restricted to the highest exposures. Residual analysis methods are sometimes used but are often inconclusive and not sufficiently well developed for many common models for environmental data. In a previous paper

^aDepartment of Statistics, University of Calcutta, Kolkata, India

^bICAR-Central Inland Fisheries Research Institute, India

^cDepartment of Mathematics and Statistics, American University, Washington DC, WA, U.S.A.

^dDepartment of Environmental Health Sciences, University of California, Berkeley, UC, U.S.A.

*Correspondence to: B. Ganguli, Department of Statistics, University of Calcutta, Kolkata, India.

†E-mail: bgstat@gmail.com

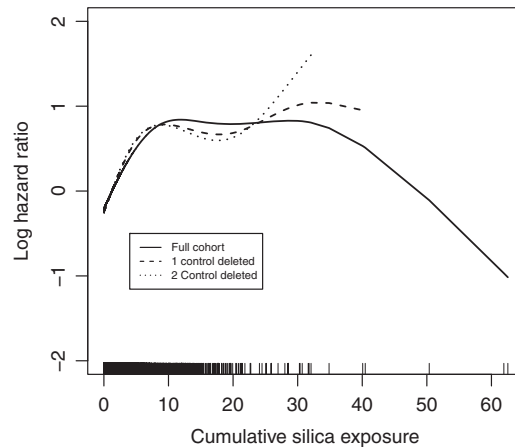


Figure 1. Plot of the estimated log hazard ratio from the motivating silica dataset, including fits from the full dataset and after exclusion of the highest exposures.

Ganguli *et al.* [4], the authors have shown that application of different ad hoc diagnostic procedures can lead to conflicting conclusions about the functional form of the log hazard ratio for silica exposure.

In a previous reanalysis of the motivating data [5], penalized splines were used in a Cox model framework to estimate the exposure–response relationship. As shown in Figure 1, the selected curve rose to a maximum hazard at 10 mg/m³-years, and then after a long plateau, gradually declined. Informal regression diagnostics revealed that the subject with the highest exposure at the end of follow-up (62 mg/m³-years) appeared in 72 of the risk sets defined for the 77 lung cancer cases. The removal of the subject truncated the upper half of the exposure range beyond 40 mg/m³-years, and with it, removed most of the decline in the fitted curve. The post model fitting deletion methods developed in this paper are applied to the same occupational cohort study of silica [1] that provided the basis for the earlier analysis.

1.2. Literature review

Post model fitting deletion diagnostics [6] using criterion-based methods such as Cook’s distance [7] offer a relatively objective means to identify influential data points.

An alternative class of methods for identification of influential points focuses on measurement of local influence [8–13]. As is evident from the name, these methods provide an estimate of local influence in response to small perturbations of the model. They require the user to specify a priori null and alternative hypotheses about the nature of influence. By contrast, Cook’s distance is better suited for exploratory analysis of influence and detection of gross violations.

Much of the literature on criterion-based methods has focused on the classical linear model, an attractive feature of which is that well-known matrix identities can be used to derive explicit expressions for the ‘deleted’ estimates. Outside the realm of the linear model, Fung *et al.* [14] derives Cook’s distance for a class of semiparametric models, while Barlow [15] derives the distance for the proportional hazards model. Preisser and Qaqish [16] considers deletion diagnostics for generalised estimating equations. A particularly simple and elegant framework for deletion diagnostics for *linear* mixed models is provided by Haslett and Hayes [17] and Haslett [18] where the authors show that the best linear unbiased predictor after deletion of an observation is equivalent to the best linear unbiased predictor from the complete subset with the deleted observation replaced by its predictor based on the remaining data points. However, each of these make the assumption that deletion of an individual does not affect the estimate of the variance components. While Haslett and Dillane [19] derives separate deletion diagnostics for variance parameters, the authors recognise that estimation of mean and variance parameters are inter-linked and that one should really derive deleted estimates of regression coefficients, which have been corrected for any changes to estimated variances as a result of deletion. Welsch [20] argues that the estimate of a regression parameter is only meaningful in the context of a standard error estimate, which in turn requires deletion diagnostics for variance parameters. This is particularly true when variance components are used as a device for smoothing [21]. Therneau and Grambsch [22] points out that failure to re-estimate the variance can lead to underestimation for larger influence points. Sen Roy and Guria [23]

shows that a proper combination of the full set and deleted set estimators leads to a smaller estimate of the standard error and hence to more efficient estimation. Ganguli *et al.* [4] reviews the application of existing diagnostic methods for Cox models and finds them inadequate when splines are included as regressors.

Our objective in this paper is to develop an approximation for Cook's distance for a particular class of GLMMs considered by Breslow and Clayton [24] with independent random effects. While our primary focus is on using GLMMs as a device for penalised smoothing, the GLMM with independent random effects has a vast scope including random intercept modelling, correlated errors, clustering, spatial modelling, censored data modelling and combinations of the aforementioned. This has been discussed by Ruppert *et al.* [21]. The proposed method enables the user to study the impact of deletion on point estimates of the fixed and random coefficients as well as on corresponding confidence/prediction intervals. In general, influence detection when the underlying relationship is non-linear is a tricky problem and some of the pertinent issues have been discussed by Manchester and Blanchard [25]. However, the GLMM provides a particularly convenient framework for modelling a non-linear dose response relationship and the interested reader is referred to [21] and [26].

2. Methodology

2.1. Model

Our underlying model is a special case of the GLMM of Breslow and Clayton [24], which assumes independence of random effects. For each of n individuals, we observe an outcome y_i together with p covariates \mathbf{x}_i associated with fixed effects. The random effects are components of the r -dimensional vector, \mathbf{b} partitioned as $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_k^T)^T$, with \mathbf{b}_j being of dimension $r_j, j = 1, \dots, k$ and $r = \sum_{j=1}^k r_j$. The corresponding coefficients are denoted by $\mathbf{z}_i = (\mathbf{z}_{i1}^T, \dots, \mathbf{z}_{ik}^T)^T$, where $\mathbf{z}_{ij} = (z_{ij1}, \dots, z_{ijr_j})^T$. Conditioning on the r -dimensional vector of random effects, we have,

$$E(y_i|\mathbf{b}) = \mu_i \quad \text{and} \quad \text{Var}(y_i|\mathbf{b}) = \phi a_i v(\mu_i), \quad i = 1, \dots, n. \quad (1)$$

The a_i 's are known scalars, and $v(\cdot)$ is a known function. The unknown parameter ϕ accounts for overdispersion. Here, μ_i is related to a linear predictor η_i via the relationship,

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\alpha} + \mathbf{z}_i^T \mathbf{b}$$

for some known function $g(\cdot)$. We assume that $\mathbf{b} \sim N_r(\mathbf{0}, \mathbf{D})$ where $\mathbf{D} = ((\sigma_j^2 \mathbf{I}_{r_j}))_{j=1, \dots, k}$. Explicit expressions for parameter estimates are not available and $\boldsymbol{\alpha}$, \mathbf{b} and \mathbf{D} are estimated from an iterative system of equations. Restricted maximum likelihood (REML) estimating equations [27] are used for the variance parameters. Writing $\mathbf{y} = (y_1, \dots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ and $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^T$, the equations are,

$$\begin{bmatrix} \mathbf{X}^T \mathbf{W} \mathbf{X} & \mathbf{X}^T \mathbf{W} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W} \mathbf{X} & \mathbf{Z}^T \mathbf{W} \mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{W} \mathbf{Y} \\ \mathbf{Z}^T \mathbf{W} \mathbf{Y} \end{bmatrix}, \quad (2)$$

and

$$\mathbf{Y}^T \mathbf{Q} \mathbf{Z}_j \mathbf{Z}_j^T \mathbf{Q} \mathbf{Y} = \text{trace}(\mathbf{Q} \mathbf{Z}_j \mathbf{Z}_j^T). \quad (3)$$

Here, $\mathbf{Y} = \boldsymbol{\eta} + (\mathbf{y} - \boldsymbol{\mu}) * g'(\boldsymbol{\mu})$, and assuming a canonical link

$$\begin{aligned} \mathbf{W} &= \text{diag}((\phi a_i g'(\mu_i))), \\ \mathbf{V} &= \mathbf{W}^{-1} + \mathbf{Z} \mathbf{D} \mathbf{Z}^T = \mathbf{W}^{-1} + \sum_{j=1}^k \sigma_j^2 \mathbf{Z}_j \mathbf{Z}_j^T, \\ \text{and } \mathbf{Q} &= \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}, \end{aligned}$$

with $*$ representing component-wise multiplication and \mathbf{Z} partitioned as $[\mathbf{Z}_1, \dots, \mathbf{Z}_k]$ corresponding to the k sets of independent random effects.

Our application is based on the approximation of a Cox model with nonlinear log hazard ratios by a GLMM as illustrated by Therneau and Grambsch [22], and Ganguli and Wand [28]. The use of GLMMs as a device for scatterplot smoothing and penalised regression is well known and has been elaborated upon by Ruppert *et al.* [21], Speed [29], Wahba [30], and Babette and John [31]. The essence of the idea is to use a flexible set of basis functions corresponding to a fixed number of knots as explanatory variables, which can model nonlinearity, and the distributional assumption on the coefficients of these basis functions serves as a device for smoothing. Best linear unbiased prediction can be used to derive the fitted model [32]. The principle of best prediction can be extended in several ways, for example, as a device for penalised log likelihood-based estimation for generalised responses belonging to the one parameter exponential family [33] and as a device for kriging in case of geospatial data [34]. Extensive discussion on technical issues such as knot selection, alternative choices of basis functions, integral evaluation and so on has been provided by Ruppert *et al.* [21]. Therneau and Grambsch [22] further extends this to modelling censored data with nonlinear log hazard ratios by using the counting process representation of Cox models. By introducing an approximation to the baseline hazard function, the authors are able to reduce the model to an approximate Poisson GLMM.

2.2. Deletion estimates

A first set of deleted estimates can be obtained by application of the results of Haslett and Hayes, and Hayes and Haslett [17, 35], viz.:

$$\hat{\alpha}^{(i,0)} = \hat{\alpha}_n - \mathbf{B}_i \tilde{e}_i,$$

$$\hat{\mathbf{b}}^{(i,0)} = \hat{\mathbf{b}}_n - \hat{\tau} \mathbf{Z}^T \mathbf{Q}_i \tilde{e}_i,$$

where

$$\hat{\tau} = \text{diag}(\hat{\tau}_1, \dots, \hat{\tau}_k) = \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_k^2)$$

$$\mathbf{B} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} = [\mathbf{B}_1, \dots, \mathbf{B}_n]$$

$$\mathbf{Q} = [\mathbf{Q}_1, \dots, \mathbf{Q}_n] = ((q_{ij}))_{i,j=1,\dots,n}$$

$$\tilde{e}_i = q_{ii}^{-1} \mathbf{Q}_i^T \mathbf{Y},$$

and $\hat{\alpha}_n$ and $\hat{\mathbf{b}}_n$ represent the estimates from the full dataset. Similarly, estimates of the variance parameters after deletion can be approximated using the results of Haslett and Dillane [19], viz.

$$\mathbf{T} \hat{\tau}^{(i,0)} = \tilde{\mathbf{s}}^{(i)}, \tag{4}$$

where

$$\mathbf{T} = \left(\left(\text{trace} \mathbf{Z}_i \mathbf{Z}_i^T \mathbf{Q} \mathbf{Z}_j \mathbf{Z}_j^T \mathbf{Q} \right) \right)_{i,j},$$

$$\tilde{\mathbf{s}}^{(i)} = \mathbf{Y} \mathbf{P}_i \mathbf{Q} \mathbf{Z} \mathbf{Z}^T \mathbf{Q} \mathbf{P}_i \mathbf{Y} + q_{ii}^{-1} \left(\left(\text{trace} \mathbf{Q} \mathbf{Z}_j \mathbf{Z}_j^T \mathbf{Q} \mathbf{E}_i \right) \right).$$

Here, $\mathbf{P}_i = \mathbf{I}_n - q_{ii}^{-1} \mathbf{E}_i \mathbf{Q}$, and \mathbf{E}_i is the i th unit vector of order n .

Unfortunately, the results of Haslett and Hayes [17], and Hayes and Haslett [35] cannot be applied to subsequent iterations to arrive at the deleted estimates as the i th iteration equation post deletion is not directly related to the i th iteration equation of the full model. Instead, we consider the systems (2) and (3) at convergence and denote by $\hat{\alpha}_n$, $\hat{\mathbf{b}}_n$ and $\hat{\mathbf{D}}_n$ the estimators computed from the full system based on n observations. At the next iteration of the deleted system, the quantities \mathbf{Y} , $\boldsymbol{\mu}$, $\boldsymbol{\eta}$, \mathbf{V} , \mathbf{W} and \mathbf{Q} are updated using the first step deleted estimates $\hat{\alpha}^{(i,0)}$, $\hat{\mathbf{b}}^{(i,0)}$ and $\hat{\tau}^{(i,0)}$ to obtain the next set of iterative equations,

$$\begin{bmatrix} \mathbf{X}^{(i)T} \mathbf{W}^{(i,0)} \mathbf{X}^{(i)} & \mathbf{X}^{(i)T} \mathbf{W}^{(i,0)} \mathbf{Z}^{(i)} \\ \mathbf{Z}^{(i)T} \mathbf{W}^{(i,0)} \mathbf{X}^{(i)} & \mathbf{Z}^{(i)T} \mathbf{W}^{(i,0)} \mathbf{Z}^{(i)} + \mathbf{D}^{(i,0)-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(i)T} \mathbf{W}^{(i,0)} \mathbf{Y}^{(i)} \\ \mathbf{Z}^{(i)T} \mathbf{W}^{(i,0)} \mathbf{Y}^{(i)} \end{bmatrix}, \tag{5}$$

where $\mathbf{X}^{(i)}$ and $\mathbf{Z}^{(i)}$ are the design and variance matrices constructed after deletion of the i th individual and $\mathbf{D}^{(i,0)}$ and $\mathbf{W}^{(i,0)}$ are variance matrices constructed based on $\hat{\alpha}^{(i,0)}$, $\hat{\mathbf{b}}^{(i,0)}$ and $\hat{\tau}^{(i,0)}$. As the one-step deleted

estimates $\hat{\alpha}^{(i,0)}$, $\hat{\mathbf{b}}^{(i,0)}$ and $\hat{\tau}^{(i,0)}$ usually constitute small corrections and are often under estimates [22], we make the assumption that the solution to the full system corresponding to (5) may be approximated by $\hat{\alpha}^{(i,0)}$, $\hat{\mathbf{b}}^{(i,0)}$ and $\hat{\tau}^{(i,0)}$. Then by a reapplication of the results of Haslett and Hayes [17], and Hayes and Haslett [35], improved deleted estimates are given by

$$\begin{aligned}\hat{\alpha}^{(i)} &= \hat{\alpha}_n - \mathbf{B}_i \tilde{e}_i - \mathbf{B}_i^* \tilde{e}_i^*, \\ \hat{\mathbf{b}}^{(i)} &= \hat{\mathbf{b}}_n - \hat{\tau} \mathbf{Z}^T \mathbf{Q}_i \tilde{e}_i - \hat{\tau}^* \mathbf{Z}^T \mathbf{Q}_i^* \tilde{e}_i^*\end{aligned}$$

where the superscript * denotes that these quantities were calculated based on $\hat{\alpha}^{(i,0)}$, $\hat{\mathbf{b}}^{(i,0)}$ and $\hat{\tau}^{(i,0)}$. Using a Taylor series expansion, we can linearise $\hat{\alpha}^{(i)}$ as

$$\begin{aligned}\hat{\alpha}^{(i)} &= \hat{\alpha}_n - \mathbf{B}_i \tilde{e}_i - \mathbf{B}_i^* \tilde{e}_i^* \\ &\simeq \hat{\alpha}_n - \mathbf{B}_i \tilde{e}_i - \left\{ \mathbf{B}_i \tilde{e}_i + \begin{pmatrix} \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \alpha^T} & \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \mathbf{b}^T} & \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \tau^T} \end{pmatrix} \begin{pmatrix} \hat{\alpha}^{(i,0)} - \hat{\alpha}_n \\ \hat{\mathbf{b}}^{(i,0)} - \hat{\mathbf{b}}_n \\ \hat{\tau}^{(i,0)} - \hat{\tau}_n \end{pmatrix} \right\}.\end{aligned}$$

Hence,

$$\hat{\alpha}^{(i)} - \hat{\alpha}_n = -2\tilde{e}_i \mathbf{B}_i - \begin{pmatrix} \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \alpha^T} & \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \mathbf{b}^T} & \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \tau^T} \end{pmatrix} \begin{pmatrix} \hat{\alpha}^{(i,0)} - \hat{\alpha}_n \\ \hat{\mathbf{b}}^{(i,0)} - \hat{\mathbf{b}}_n \\ \hat{\tau}^{(i,0)} - \hat{\tau}_n \end{pmatrix} \quad (6)$$

$$= f_{\alpha}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i), \quad (7)$$

where, as shown in Appendix A1,

$$\begin{aligned}\frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \alpha^T} &= -\tilde{e}_i \mathbf{B} \frac{\partial \mathbf{V}}{\partial \alpha^T} (\mathbf{Q}_i \otimes \mathbf{I}_p) + \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T \frac{\partial \mathbf{V}}{\partial \alpha^T} \{(\tilde{e}_i \mathbf{Q}_i - \mathbf{QY}) \otimes \mathbf{I}_p\}, \\ \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \mathbf{b}^T} &= -\tilde{e}_i \mathbf{B} \frac{\partial \mathbf{V}}{\partial \mathbf{b}^T} (\mathbf{Q}_i \otimes \mathbf{I}_q) + \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T \frac{\partial \mathbf{V}}{\partial \mathbf{b}^T} \{(\tilde{e}_i \mathbf{Q}_i - \mathbf{QY}) \otimes \mathbf{I}_q\}, \\ \frac{\partial \tilde{e}_i \mathbf{B}_i}{\partial \tau^T} &= -\tilde{e}_i \mathbf{B} \frac{\partial \mathbf{V}}{\partial \tau^T} (\mathbf{Q}_i \otimes \mathbf{I}_k) + \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T \frac{\partial \mathbf{V}}{\partial \tau^T} \{(\tilde{e}_i \mathbf{Q}_i - \mathbf{QY}) \otimes \mathbf{I}_k\},\end{aligned}$$

and $f_{\alpha}(\cdot)$ is defined in (6) and (7). Note that this enables us to approximate $\hat{\alpha}^{(i)}$ in terms of quantities obtained from the full model fit. Similarly, we can obtain expressions for the DFBETAs,

$$\hat{\mathbf{b}}^{(i)} - \hat{\mathbf{b}}_n = f_{\mathbf{b}}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i), \quad (8)$$

$$\hat{\tau}^{(i)} - \hat{\tau}_n = f_{\tau}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i) \quad (9)$$

as well as similar expressions for DFFIT residuals in terms of the output of the full model fit.

2.3. Standardisation and critical values

The next step is to derive a suitable standardisation factor for the differences in (6) and (8). The usual approach in the literature is to standardise these by a positive definite matrix, which can be chosen to reflect specific concerns [36]. A standard procedure following Cook [7] is to use the variances of the corresponding parameters from the full model for standardisation and the percentage points of the \mathbf{F} distribution as a critical value. In this case, the corresponding ‘*p value*’ gives the level of the smallest confidence ellipsoid based on the full data that contains the estimate.

However, there has been some discussion in the literature regarding the choice of the norming matrix to be used for standardisation. Beckman and Cook [6] argues that it should not depend on *i*, whereas the norm suggested by Belsley *et al.* [37] does so. Dempster and Gasko [36] provides an interesting geometric perspective on various alternative standardisation criteria and their connection with residual analyses.

Our approach is based on the result,

$$\begin{aligned} f_{\alpha}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i) - f_{\alpha}\left(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, q_{ii}^{-\frac{1}{2}}Z\right) &\xrightarrow{P} 0 \\ f_{\mathbf{b}}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i) - f_{\mathbf{b}}\left(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, q_{ii}^{-\frac{1}{2}}Z\right) &\xrightarrow{P} 0 \\ f_{\tau}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i) - f_{\tau}\left(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, q_{ii}^{-\frac{1}{2}}Z\right) &\xrightarrow{P} 0, \end{aligned} \tag{10}$$

where Z is a standard normal deviate. This result has been proved in Appendix (A2). The implication of this is that we can generate K standard normal deviates $\{Z_k, k = 1, \dots, K\}$ and estimate the covariance matrix of say, $f_{\alpha}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i)$, by the empirical covariance of $f_{\alpha}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, q_{ii}^{-\frac{1}{2}}Z_k), k = 1, \dots, K$, which can then be used to standardise the residuals $f_{\alpha}(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\tau}_n, \tilde{e}_i)$. Similar approaches can be used for the random effects \mathbf{b} and the variance components τ as well as for the DFFITs $\hat{\eta}_i - \hat{\eta}_i^{(i)}$. Simulation based critical values can also be derived for each of the quantities based on the results in (10).

3. Application

3.1. Application to the silica exposure dataset

In this section, we apply our methods to the motivating dataset. The connection between Cox models and Poisson regression is well established [38, 39]. Here, we use the equivalence discussed by Therneau and Grambsch [22], and Ganguli and Wand [28] who approximate a Cox proportional hazards model with a non linear log hazard ratio by a Poisson mixed effects model. We fit a model of the form,

$$\lambda(t|x) = \lambda_0(t) \exp\{\beta_0 \text{HISP} + f(x(t))\}, \tag{11}$$

where t is the age at which a subject dies from lung cancer, HISP is an indicator for whether or not a subject is Hispanic and $x(t)$ is cumulative silica exposure at time t . The function $f(\cdot)$ is unknown but assumed to be smooth.

Using I as the indicator function and δ as the censoring indicator (1, if uncensored and 0, if censored) and writing $N(t) = I(T \leq t, \delta = 1)$, it follows from Ganguli and Wand [28] that (11) can be approximated as,

$$E(N(t)|\mathbf{b}) = \exp\left\{\beta_0 \text{HISP} + \beta_1 x(t) + \sum_{k=1}^K b_k |x(t) - \kappa_k|^3\right\} \hat{\Lambda}, \tag{12}$$

where the κ_k 's are a set of K quantiles of x and the components of $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_K) \sim \text{i.i.d.} N(0, \sigma_b^2)$ and $\hat{\Lambda}$, is the estimator of the cumulative hazard function of the standard Cox model with log linear hazard.

Following Ganguli and Wand [28], and Therneau and Grambsch [22], this can be approximated by a GLMM with a log-linear link function,

$$\log\{E(N(t)|\mathbf{b})\} = \log(\hat{\Lambda}) + \beta_0 \text{HISP} + \beta_1 x(t) + \sum_{k=1}^K b_k |x(t) - \kappa_k|^3,$$

where $\hat{\Lambda}$, the estimated cumulative hazard from a standard Cox model assuming a linear log hazard function for cumulative silica hazard, serves as an approximate offset.

Figure 2 shows the DFFITs in panel (a) and DFBETAs for the variance component σ_b^2 in panel (b). Panels (c) and (d) show the estimates of the log hazard function after excluding subjects identified as influential in panel (a) and after excluding subjects identified as influential based on both panels (a) and (b), respectively. Subjects were considered to be influential if their standardised DFBETA or DFFIT values exceeded the simulation based benchmark described in Section 2.3. We can make the following observations:

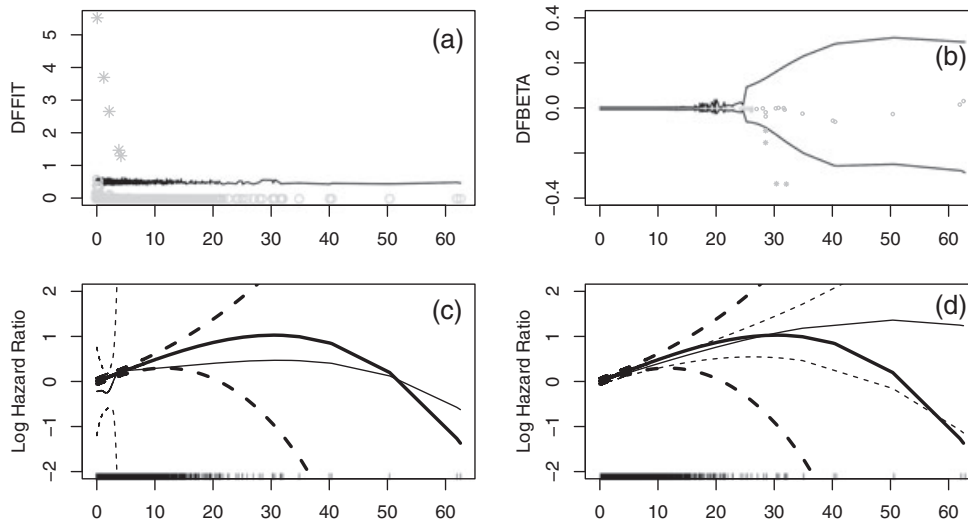


Figure 2. Influence diagnostics for the silica data. The horizontal axis represents cumulative silica exposure ($mg/m^3 - yr$). Panel (a) shows DFFITs and panel (b) shows DFBETAs for the variance component. Panels (c) and (d) give the estimated log hazard function after excluding observations identified as influential only in panel (a) and in both panels (a) and (b), respectively. Thick solid and dotted lines correspond to fitted values and 95% confidence intervals from the complete dataset. Thin solid and dotted lines correspond to estimates after deletion of influential subjects.

- The log hazard continues to dip beyond an exposure of $50 mg/m^3$ -year in the first case but approaches a plateau in the second.
- The two figures (Figure 2 (c)–(d)) show a remarkable contrast. In the former case (Figure 2(c)), deletion does not affect the shape of the dose response curve significantly although it does attenuate the response by about 50% at $30 mg/m^3$ -years. However, in the latter case (Figure 2(d)), deletion appears to have a more dramatic effect and leads to a different shape of the dose response curve. The results suggest that the curve based on the full dataset might be picking up spurious structure at the upper range of exposures.
- The 95% pointwise confidence intervals are indicated using bold lines for the full dataset and dotted lines corresponding to post-deletion. The results make a strong case for conducting post model fitting deletion diagnostics for variance components as the confidence intervals obtained by dropping observations with influential DFFIT values only is very wide. Note that individuals who are influential for variance component estimation are still included for estimation in panel (c) and it appears that the inclusion of these observations leads to an unstable estimate of the variance components and hence to an unreliable confidence interval. By contrast, the confidence intervals obtained in panel (d) are tighter.

Our learning from this application is firstly that influential observations need not occur at the highest extremes of exposure. Second, we can distinguish between observations, which are influential for the shape of the estimated dose response curve and those which impact its value but not its shape. The individual with the highest exposure is influential for the estimated shape of the log hazard ratio. The optimum degrees of freedom estimate decreases from approximately three to two as a result of dropping the individual with the highest exposure. Finally, we note that failure to conduct post model fitting diagnostics for variance components can lead to erroneous and unstable confidence interval estimation.

3.2. Performance on simulated datasets

In this section, we demonstrate the performance of the method on three different applications all of which may be fitted using GLMMs.

3.2.1. Clustered data. We simulate clustered count data using the model:

$$y_{ij} \sim \text{Poisson}(\mu_{ij}); \quad i = 1, 2, \dots, k; j = 1, 2, \dots, n_i,$$

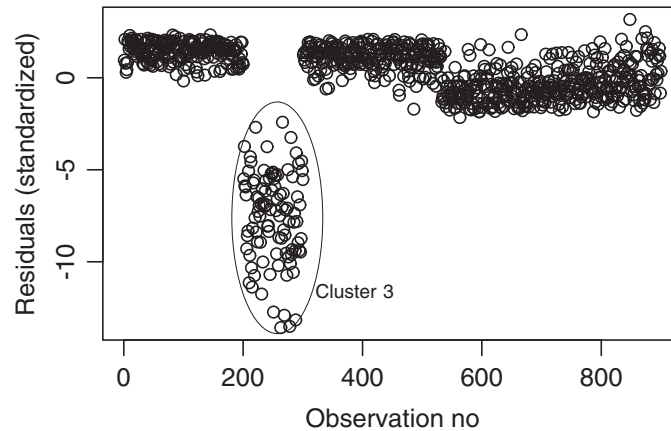


Figure 3. Plot of standardized *DFBETAs* of fixed effects for clustered Poisson counts with all observation in cluster 3 as influential.

with k being the number of clusters. All observations from the i th cluster share a common random intercept, b_i specified by $\log(\mu_{ij}) = b_i + \beta x_{ij}$, where x_{ij} is a subject specific covariate, and β is the corresponding log rate ratio. For the purpose of simulation, we set $\beta = 0.5$, $k = 9$ and $n_i = 100$. The $b_i, i = 1, 2, \dots, 9$, are generated from a $N(0, 0.5)$ distribution, and x_{ij} 's are obtained from a sequence of $(k \times n_i) = 900$ values generated from a $U[0, 1]$ distribution. We introduce an influential subject (here cluster) by replacing all observations in cluster 3 by counts corresponding to $\mu_{ij} = 20$.

Figure 3 shows the corresponding standardized residuals, which clearly identify observations belonging to cluster 3 as influential in terms of high residual values. It is noteworthy that inclusion of the outlying cluster forces the regression line to have a negative slope, whereas the true slope coefficient is positive. Note that the estimates of the variance component are 3.544 and 0.301 corresponding to inclusion and exclusion of the outlying cluster, respectively.

3.2.2. Binary generalized additive model. We generate data from a binary model with a nonlinear dose response, which is modelled using a generalized additive model specified by:

$$y \sim \text{Binary}(p) \quad \text{where} \quad \text{logit}(p) = \beta_0 + f(x). \tag{13}$$

We set $\beta_0 = 0.5$ and $f(x) = 1.5x - 2.5x^2$. We first generate the $x_i, i = 1, 2, \dots, n = 1500$ from the uniform $[0, 1]$ distribution and then generate the binary $y_i, i = 1, 2, \dots, n = 1500$ using (13). Thereafter, we introduce influential values by sorting the x values and replacing all the 0 responses between observations 1480–1495 by 1's.

Following Ruppert *et al.* [21], the model in (13) can be represented as a GLMM as follows:

$$\text{logit}(p|\mathbf{b}) = \beta_0 + \sum_{k=1}^K b_k |x - \kappa_k|^3. \tag{14}$$

Here, the κ_k 's are a set of K quantiles of x and the components of $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_K) \sim \text{i.i.d.}N(0, \sigma_b^2)$. This model can be fitted using the `glmmPQL` routine in R. Note that K is set at 15 following the guidelines provided by Ruppert *et al.* [21].

A plot of the estimated log-odds ratio along with point-wise 95% confidence intervals (Figure 4) suggests spurious structure of the estimated function when computed from the full dataset, whereas the functional form based on the deleted estimates indicates a smoother fit. Note that the REML based estimates of the degrees of freedom based on all and non-influential observations only were 4.5 and 2.8, respectively.

3.2.3. Geostatistical data. We consider here a simulated example from the following geostatistical model:

$$y_i \sim \text{Poisson}(\mu), \quad \text{where} \quad \log(\mu) = s(\mathbf{x}), \tag{15}$$

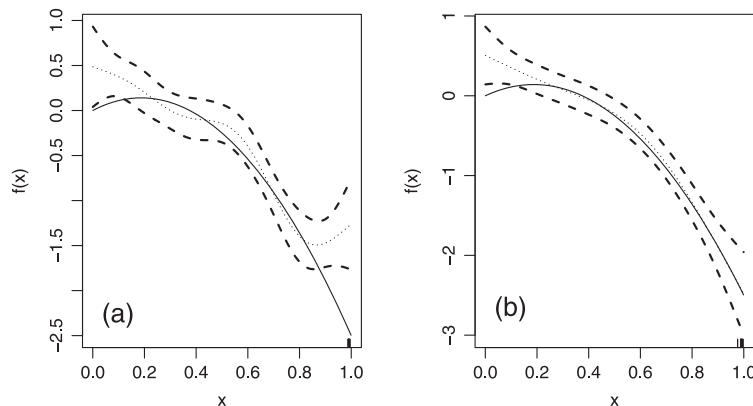


Figure 4. True (continuous line), estimated (dotted lines) and 95% point-wise confidence interval (thick dashed line) of log odds $f(x)$ for a simulated binary generalized additive model (GAM). (a) is obtained by including all observations and (b) by excluding observations flagged as influential.

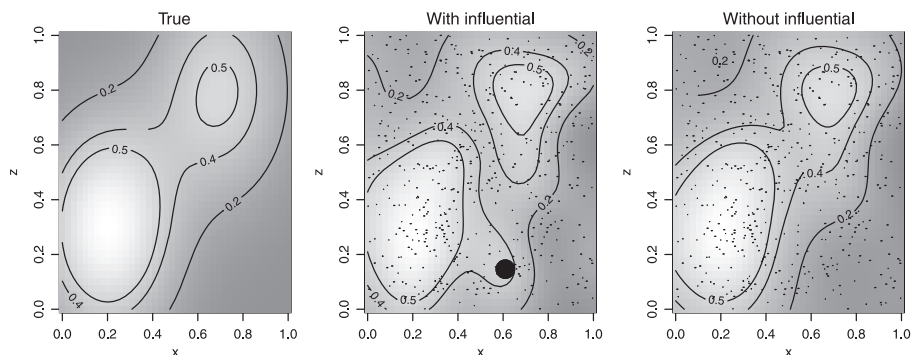


Figure 5. Contour plots showing a comparison of the true and estimated functional forms with and without inclusion of the influential point. DFBETAs for the variance component are overlaid on the estimated surface and the size of the solid circle corresponds to the magnitude of influence.

with $s(\cdot)$ being a bivariate function of $\mathbf{x} = (x_1, x_2)^T$, which we set as

$$s(x_1, x_2) = \frac{0.6}{2\pi\sigma_1\sigma_2} \times e^{-\frac{(x_1-0.2)^2}{\sigma_1^2} - \frac{(x_2-0.3)^2}{\sigma_2^2}} + \frac{0.4}{2\pi\sigma_1\sigma_2} \times e^{-\frac{(x_1-0.7)^2}{\sigma_1^2} - \frac{(x_2-0.8)^2}{\sigma_2^2}},$$

where $\sigma_1 = 0.3$, and $\sigma_2 = 0.4$. Five hundred sets of \mathbf{x} were independently generated from the uniform $[0, 1]$ distribution and based on these, y 's were generated using (15). We replaced the 500th observation with a relatively high value to induce an influential observation. Following Kammann and Wand [34], a low rank model formulation is given by,

$$\log E(y_i|\mathbf{b}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sum_{k=1}^K b_k |\mathbf{x}_i - \boldsymbol{\kappa}_k|^3. \tag{16}$$

Here, the $\boldsymbol{\kappa}_k$'s are a set of K knots chosen by applying a space-filling algorithm to the \mathbf{x}_i 's and the components of $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_K) \sim \text{i.i.d.}N(0, \sigma_b^2)$.

Results are plotted in Figure 5. The estimated d.f. with and without influential observations were 23.3 and 21.75, respectively. The influential point was correctly identified, and the estimated functional form after dropping the influential point was similar to the true functional form.

4. Discussion

In this paper, we have developed Cook's distance for fixed and random effects parameters as well as for variance component estimates for the generalised linear mixed model with independent random effects.

Given the wide applicability of the model, these diagnostics have far reaching scope. The simulations and application presented make a strong case for conducting post model fitting deletion diagnostics. First, influential observations need not occur at the highest extremes of exposure. Second, we can distinguish between influential observations, which only modestly influence the values of the log hazard over local regions, and those which influence the shape of the exposure-response curve. Third, observations influential for the variance component estimate occurred at the highest levels of exposure and the deletion of these outliers had more influence on the shape of the curve. In the silica application, when the highest exposed subject was removed, the optimum degrees of freedom dropped from three to two. Using the full dataset leads to overfitting in regions of high cumulative exposure. This is common for exposure-response modelling and possible reasons include the saturation of biological processes, misclassification of exposure at high levels and the healthy worker effect [3, 5, 40]. For the silica dataset, the main problem is the sparsity of data and few cases in the regions of high cumulative exposure.

Empirical observations from the authors suggest that the performance of the method is driven by a combination of the degree of nonlinearity in the exposure-response relationship and the local density of data points. The method performs well in presence of moderate nonlinearity even in regions of sparse data but will tend to spuriously identify influential observations in regions of high nonlinearity. We have also informally compared the performance of the method on simulated datasets with Poisson and binary outcomes and observed that for the same degrees of nonlinearity and data density, influential values are identified more accurately for Poisson outcomes. Further extensions of the method should accommodate deletion of a subset of observations. Note that in our application to the silica data, which has time-varying exposures, we have considered detection of influential exposures and considered an individual as influential if at least one of his observations has been classified as influential. We have assumed throughout the paper that the random effects are independent. While such an assumption is typically made for cases such as our motivating example, where the variance components model serves as a device for smoothing, non-spherical variance structures are commonly used when random effects are used to model subject specific effects or overdispersion. As additional covariance parameters are also estimated by REML and the results of Haslett and Hayes, and Hayes and Haslett [17, 35] continue to be valid, our method can, in theory, be extended to accommodate non-spherical variance structures although at a cost of added computational complexity. Additional research that could be spurred by this paper include the development of models, which include explanations for the occurrence of influential observations such as population heterogeneity, but it is clear that such research requires considerable substantive knowledge.

Appendix

A.1. Derivation of the expression for the derivatives in (6):

We have

$$\frac{\delta \tilde{\mathbf{e}}_i \mathbf{B}_i}{\delta \boldsymbol{\alpha}^T} = \frac{\delta \mathbf{B}_i}{\delta \boldsymbol{\alpha}^T} \tilde{\mathbf{e}}_i \otimes \mathbf{I}_p + \mathbf{B}_i \otimes \mathbf{I}_p \frac{\delta \tilde{\mathbf{e}}_i}{\delta \boldsymbol{\alpha}^T}.$$

To calculate the derivatives, first note that writing $\mathbf{C} = \begin{pmatrix} \mathbf{0}_{p \times p} & \mathbf{X}^T \\ \mathbf{X} & \mathbf{V} \end{pmatrix}$,

$$\mathbf{B} = (\mathbf{I}_p \ \mathbf{0}_{p \times n}) \begin{pmatrix} \mathbf{0}_{p \times p} & \mathbf{X}^T \\ \mathbf{X} & \mathbf{V} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0}_{p \times n} \\ \mathbf{I}_n \end{pmatrix} = (\mathbf{I}_p \ \mathbf{0}_{p \times n}) \mathbf{C}^{-1} \begin{pmatrix} \mathbf{0}_{p \times n} \\ \mathbf{I}_n \end{pmatrix},$$

so that

$$\begin{aligned} \frac{\delta \mathbf{B}}{\delta \boldsymbol{\alpha}^T} &= \left\{ (\mathbf{I}_p \ \mathbf{0}_{p \times n}) \otimes \mathbf{I}_p \right\} \frac{\delta \mathbf{C}^{-1}}{\delta \boldsymbol{\alpha}^T} \left\{ (\mathbf{0}_{n \times p} \ \mathbf{I}_n)^T \otimes \mathbf{I}_p \right\} \\ &= - \left\{ (\mathbf{I}_p \ \mathbf{0}_{p \times n}) \otimes \mathbf{I}_p \right\} \mathbf{C}^{-1} \frac{\delta \mathbf{C}}{\delta \boldsymbol{\alpha}^T} \mathbf{C}^{-1} \left\{ (\mathbf{0}_{n \times p} \ \mathbf{I}_n)^T \otimes \mathbf{I}_p \right\} \end{aligned}$$

Noting that

$$\frac{\delta \mathbf{C}}{\delta \boldsymbol{\alpha}^T} = \begin{pmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \frac{\delta \mathbf{V}}{\delta \boldsymbol{\alpha}^T} \end{pmatrix}$$

and that

$$\frac{\delta \mathbf{V}}{\delta \boldsymbol{\alpha}^T} = \begin{bmatrix} \phi a_1 \frac{g''(\mu_1)}{g'(\mu_1)} \mathbf{x}_1^T & \mathbf{0}^T & \dots & \mathbf{0}^T \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \dots & \phi a_n \frac{g''(\mu_n)}{g'(\mu_n)} \mathbf{x}_n^T \end{bmatrix},$$

we obtain, after some simplification,

$$\frac{\delta \mathbf{B}_i}{\delta \boldsymbol{\alpha}^T} = -\mathbf{B} \frac{\delta \mathbf{V}}{\delta \boldsymbol{\alpha}^T} (\mathbf{Q}_i \otimes \mathbf{I}_p).$$

Similarly,

$$\begin{aligned} \frac{\delta \mathbf{Q}_i}{\delta \boldsymbol{\alpha}^T} &= -\mathbf{Q} \frac{\delta \mathbf{V}}{\delta \boldsymbol{\alpha}^T} (\mathbf{Q}_i \otimes \mathbf{I}_p), \\ \text{and } \frac{\delta q_{ii}}{\delta \boldsymbol{\alpha}^T} &= -\mathbf{Q}_i^T \frac{\delta \mathbf{V}}{\delta \boldsymbol{\alpha}^T} (\mathbf{Q}_i \otimes \mathbf{I}_p). \end{aligned}$$

Also, using the fact that

$$\frac{\delta \tilde{e}_i}{\delta \boldsymbol{\alpha}^T} = q_{ii}^{-1} \mathbf{Q}_i^T \frac{\delta \mathbf{V}}{\delta \boldsymbol{\alpha}^T} \{(\tilde{e}_i \mathbf{Q}_i - \mathbf{Q} \mathbf{Y}) \otimes \mathbf{I}_p\},$$

we obtain the expression for $\frac{\delta \tilde{e}_i \mathbf{B}_i}{\delta \boldsymbol{\alpha}^T}$. Expressions for $\frac{\delta \tilde{e}_i \mathbf{B}_i}{\delta \mathbf{b}^T}$ and $\frac{\delta \tilde{e}_i \mathbf{B}_i}{\delta \boldsymbol{\tau}^T}$ can be similarly found.

A.2. To prove (10):

We make the following three assumptions:

- (1). $n^{-1} \begin{bmatrix} \mathbf{X}^T \mathbf{W} \mathbf{X} & \mathbf{X}^T \mathbf{W} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W} \mathbf{X} & \mathbf{Z}^T \mathbf{W} \mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \rightarrow \boldsymbol{\Sigma}$, which is positive definite.
- (2). \mathbf{W} and $\mathbf{Z} \mathbf{D} \mathbf{Z}^T$ have finite elements with $\mathbf{W} > \mathbf{0}$ and/or the elements of \mathbf{Z} and \mathbf{D} finite.
- (3). $a_i = \phi \frac{g''(\mu_i)}{g'(\mu_i)} \leq c, i = 1, \dots, n$ for some constant c .

Note that $f_\alpha(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\boldsymbol{\tau}}_n, \tilde{e}_i) - f_\alpha(\hat{\alpha}_n, \hat{\mathbf{b}}_n, \hat{\boldsymbol{\tau}}_n, q_{ii}^{-\frac{1}{2}} Z)$ can be expressed as

$$-2 \left(\tilde{e}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{B}_i + \begin{pmatrix} \mathbf{A}_1 - \mathbf{C}_1 & \mathbf{A}_2 - \mathbf{C}_2 & \mathbf{A}_3 - \mathbf{C}_3 \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{B}_3 \end{pmatrix} + \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{C}_3 \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 - \mathbf{D}_1 \\ \mathbf{B}_2 - \mathbf{D}_2 \\ \mathbf{B}_3 - \mathbf{D}_3 \end{pmatrix}. \quad (\text{A.1})$$

Consider the first term in (A.1). By the normal theory model applied to the constructed variable \mathbf{Y} ,

$$\tilde{e}_i \stackrel{\text{asy}}{\sim} N(\mathbf{1}^* \mathbf{Q}_i^T \mathbf{X} \boldsymbol{\alpha}, q_{ii}^{-2} \mathbf{Q}_i^T \mathbf{V} \mathbf{Q}_i) = N(0, q_{ii}^{-1}) \quad , \text{ so that } \tilde{e}_i - q_{ii}^{-\frac{1}{2}} Z \xrightarrow{P} 0.$$

Also,

$$\mathbf{B}_i = \frac{1}{\sqrt{n}} \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X}^T \mathbf{V}^i \tilde{e}_i \rightarrow \mathbf{0} \quad \text{as } n \rightarrow \infty.$$

Here, \mathbf{V}^i is the i th column of \mathbf{V}^{-1} . Now,

$$\begin{aligned} \mathbf{A}_1 - \mathbf{C}_1 &= \\ & - \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{B} \frac{\delta \mathbf{V}}{\delta \alpha^T} (\mathbf{Q} \otimes \mathbf{I}_p) + \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T \frac{\delta \mathbf{V}}{\delta \alpha^T} \left\{ \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{Q}_i \otimes \mathbf{I}_p \right\} \\ & = - \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{B} \mathbf{A} \mathbf{Q}^* \mathbf{X} + \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{A} \mathbf{Q}^* \mathbf{X} - \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i \mathbf{A} q_{ii} \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{E}_i \mathbf{X} \\ & \leq \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \left\{ - \left(\frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{n} \right) + \left(\frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{n} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^i \mathbf{1} \mathbf{X} \right) - \left(\frac{\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}}{n} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^i \mathbf{1} \mathbf{X} \right) \right\} \\ & = o(1)O(1) = o(1), \end{aligned}$$

and $\mathbf{B}_1 = \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \frac{1}{\sqrt{(n)}} \mathbf{X}^T \mathbf{V}^i \frac{1}{\sqrt{n}} \tilde{\epsilon}_1 = O(1)O(1)o(1)$ by assumptions 1–3, so that

$$\mathbf{B}_1(\mathbf{A}_1 - \mathbf{C}_1) = o(1).$$

Similarly, $\mathbf{B}_2(\mathbf{A}_2 - \mathbf{C}_2) = o(1)$. Note that

$$\begin{aligned} \mathbf{A}_3 - \mathbf{C}_3 &= \\ & = \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) \mathbf{B} \mathbf{Z} \mathbf{D}^* (\mathbf{Z}^T \mathbf{Q}_i \otimes \mathbf{I}_k) + \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) (\mathbf{Z}^T \mathbf{Q}_i \otimes \mathbf{I}_k) \\ & \quad - \mathbf{B}_i q_{ii}^{-1} \mathbf{Q}_i^T q_{ii} \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right) (\mathbf{Z}^T \otimes \mathbf{I}_k) \\ & = o(1)O(1)O(1) + O(1)o(1) - O(1)o(1) = o(1), \end{aligned}$$

and that $\mathbf{B}_3 = ((\text{trace} \frac{1}{n^2} (\mathbf{Z}_i^T \mathbf{Q} \mathbf{Z}_j) \frac{1}{n^2} (\mathbf{Z}_j \mathbf{Q} \mathbf{Z}_i)))^{-1} = O(1)o(1) = o(1)$, so that $\mathbf{B}_3(\mathbf{A}_3 - \mathbf{C}_3) = o(1)$. Similarly,

$$\begin{aligned} \mathbf{C}_1 &= \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X}^T \mathbf{V}^i \left(\tilde{\epsilon}_i - q_{ii}^{-\frac{1}{2}} Z \right), \\ \text{and} \\ (\mathbf{B}_1 - \mathbf{D}_1) &= \left(\frac{1}{\sqrt{n}} q_{ii}^{-\frac{1}{2}} Z \right) \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{A} \mathbf{Q}_i^T \mathbf{X} \right) \\ & \quad + \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \left\{ \frac{1}{n} \mathbf{X}^T \mathbf{V}^i \mathbf{Q}_i^T \left(q_{ii}^{-\frac{1}{2}} Z / n \right) \mathbf{A} \mathbf{Q}_i^T \mathbf{X} \right\} \\ & \quad - \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \left\{ \frac{1}{n} \mathbf{X}^T \mathbf{V}^i q_{ii}^{-1} \mathbf{Q}_i^T \mathbf{A} q_{ii} \left(q_{ii}^{-\frac{1}{2}} \frac{Z}{\sqrt{(n)}} \right) \mathbf{E}_i \mathbf{X} \right\}, \end{aligned}$$

so that

$$\mathbf{C}_1(\mathbf{B}_1 - \mathbf{D}_1) = [O(1)]O(1)O(1)o(1) = o(1)$$

Similarly, we can show that $\mathbf{C}_2(\mathbf{B}_2 - \mathbf{D}_2) = o(1)$.

Finally, note that,

$$\begin{aligned} \frac{\mathbf{C}_3}{\sqrt{n}} &= - \left(\frac{1}{\sqrt{n}} q_{ii}^{-\frac{1}{2}} Z \right) \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Z} \right) \mathbf{D}^* (\mathbf{Z}^T \mathbf{Q}_i^T \otimes \mathbf{I}_k) \\ & \quad + \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{X} \mathbf{V}^i q_{ii}^{-1} \mathbf{Q}_i^T \left(\frac{1}{\sqrt{n}} q_{ii}^{-\frac{1}{2}} Z \right) (\mathbf{Z}^T \mathbf{Q}_i \otimes \mathbf{I}_k) \\ & \quad - \left(\frac{1}{n} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}^T \mathbf{V}^i \mathbf{Q}_i^T \left(\frac{1}{\sqrt{n}} q_{ii}^{-\frac{1}{2}} Z \right) (\mathbf{Z}^T \otimes \mathbf{I}_k), \end{aligned}$$

so that $\mathbf{C}_3 (\mathbf{B}_3 - \mathbf{D}_3) = O(\sqrt{n})O(n^{-1}) = o(n^{-\frac{1}{2}}) = o(1)$. This completes the proof. In case \mathbf{V} is replaced by a consistent estimator $\hat{\mathbf{V}}$, we require an additional set of conditions:

$$\begin{aligned} \frac{1}{n} \mathbf{X}^T (\hat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}) \mathbf{X} &\xrightarrow{P} \mathbf{0} \\ \frac{1}{n} \mathbf{X}^T (\hat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}) \mathbf{Z} &\xrightarrow{P} \mathbf{0} \\ \frac{1}{n} \mathbf{X}^T (\hat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}) \tilde{\epsilon}_i &\xrightarrow{P} \mathbf{0} \\ \frac{1}{n} \mathbf{Z}^T (\hat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}) \tilde{\epsilon}_i &\xrightarrow{P} \mathbf{0} \end{aligned}$$

In case the application is to clustered data, we would require the aforementioned set of conditions to hold for each cluster.

Acknowledgements

This research was supported by grant R01 CA081345 from the National Institute of Health. The authors would also like to thank Prof. Checkoway for permission to use the dataset on silica exposure.

References

1. Checkoway H, Heyer NJ, Seixas NS, Welp EA, Demers PA, Hughes JM, Weill H. Dose-response association of silica with nonMalignant respiratory disease and lung cancer mortality in the diatomaceous earth industry. *Americal Journal of Epidemiology* 1997; **145**:680–688.
2. Hernan MA. The hazards of hazard ratios. *Epidemiology* 2010; **21**:13–15.
3. Stayner L, Steenland K, Dosemeci M, Hertz-Picciotto I. Attenuation of exposure–response curves in occupation cohort studies at high exposure levels. *Scandinavian Journal of Work, Enironment & Health* 2003; **29**:317–324.
4. Ganguli B, Naskar M, Malloy EJ, Eisen EA. Determination of the functional form of the relationship of covariates to the log hazard ratio in a Cox model. *Journal of Applied Statistics* 2015; **42**:1091–1105.
5. Eisen EA, Agalliu I, Thurston SW, Coul BA, Checkoway H. Smoothing in occupational cohort studies: an illustration based on penalized splines. *Occupational Environmental Medicine* 2004; **61**:854–860.
6. Beckman RJ, Cook RD. Outlier.....s. *Technometrics* 1983; **25**:119–149.
7. Cook RD. Influential observations in linear regression. *Journal of the American Statistical Association* 1979; **74**:169–174.
8. Pregibon D. Logistic regression diagnostics. *The Annals of Statistics* 1981; **9**:705–724.
9. Ouwens MJ, Tan F, Berger M. Local influence to detect influential data structures for generalized linear mixed models. *Biometrics* 2001; **57**:1166–1172.
10. Cain KC, Lange NT. Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* 1984; **40**:493–499.
11. Lesaffre E, Verbeke G. Local influence in linear mixed models. *Biometrics* 1998; **54**:570–582.
12. Zhu H, Lee S. Local influence for generalized linear mixed models. *The Canadian Journal of Statistics* 2003; **31**:293–309.
13. Fung W, Gu H, Xiang L, Yau KKW. Assessing local influence in principal component analysis with application to haematology study data. *Statistics in Medicine* 2007; **26**:2730–2744.
14. Fung WK, Zhu HY, Wei BC, He X. Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society Series B* 2002; **64**:565–579.
15. Barlow WE. Global measures of local influence for proportional hazards regression models. *Biometrics* 1997; **53**:1157–162.
16. Preisser JS, Qaqish BF. Deletion diagnostics for generalized estimating equations. *Biometrika* 1996; **83**:551–562.
17. Haslett J, Hayes K. Residuals for the linear model with general covariance structure. *Journal of the Royal Statistics Society Series B* 1998; **60**:201–215.
18. Haslett J. Simple derivation of deletion diagnostic results for the general linear model with correlated errors. *Journal of the Royal Statistical Society Series B* 1999; **61**:603–609.
19. Haslett J, Dillane D. Application of delete = replace to deletion diagnostics for variance component estimation in the linear mixed model. *Journal of the Royal Statistical Society Series B* 2004; **66**:131–143.
20. Welsch RE. Influential observations, high leverage points and outliers in linear regression: Comment. *Statistical Science* 1986; **1**:403–405.
21. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: New York, 2003.
22. Therneau TM, Grambsch PM. *Modeling survival data extending the Cox model*. Springer-Verlag: New York, 2000.
23. Sen Roy S, Guria S. Estimation of regression parameters in the presence of outliers in the response. *Statistics* 2009; **43**:531–539.
24. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
25. Manchester L, Blanchard W. When is a curve an outlier? An account of a tricky problem. *The Canadian Journal of Statistics* 1996; **24**:455–466.

26. Grambsch PM, Terneau TM, Fleming TR. Diagnostics plot to reveal functional form for covariates in multiplicative intensity models. *Biometrics* 1995; **51**:1469–1482.
27. Patterson HD, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrics* 1971; **58**: 545–554.
28. Ganguli B, Wand MP. Additive models for geo-referenced failure time data. *Statistics in Medicine* 2006; **25**:2469–2482.
29. Speed T. [That BLUP is a Good Thing: The estimation of random effects]: Comment. *Statistical Science* 1991; **6**(1):42–44.
30. Wahba G. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society Series B* 1978; **40**:364–372.
31. Babette AB, John AR. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 1998; **93**:961–976.
32. Robinson GK. That blup is a good thing. *The estimation of random effects. Statistical Science* 1991; **6**:15–32.
33. Green PJ. Penalized likelihood for general semi-parametric regression models. *International Statistical Review* 1987; **55**:245–259.
34. Kammann E, Wand M. Geoadditive models. *Applied Statistics* 2003; **52**:1–18.
35. Hayes K, Haslett J. Simplifying general least squares. *The American Statistician* 1999; **53**:376–381.
36. Dempster A, Gasko-Green M. New tools for residual analysis. *The Annals of Statistics* 1981; **9**:945–959.
37. Belsley DA, Kuh E, Welsch RE. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons: New York, 1980.
38. Whitehead J. Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society Series C* 1980; **29**:268–275.
39. Ma R, Krewski D, Burnett RT. Random effects Cox models: a Poisson modeling approach. *Biometrika* 2003; **90**:157–169.
40. Steeland K, Ryan S, Mitch K, Jenifer J, Henry DK. Risk estimation with epidemiologic data when response attenuates at high exposure levels. *Environmental Health perspective* 2011; **11**:854–860.