Biometrika (2001), **88,** 4, *pp.* 1179–1185

Printed in Great Britain

Semiparametric regression for clustered data

BY XIHONG LIN

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

xlin@sph.umich.edu

AND RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

carroll@stat.tamu.edu

SUMMARY

We consider estimation in a semiparametric partially generalised linear model for clustered data using estimating equations. A marginal model is assumed where the mean of the outcome variable depends on some covariates parametrically and a cluster-level covariate nonparametrically. A profile-kernel method allowing for working correlation matrices is developed. We show that the nonparametric part of the model can be estimated using standard nonparametric methods, including smoothing-parameter estimation, and the parametric part of the model can be estimated in a profile fashion. The asymptotic distributions of the parameter estimators are derived, and the optimal estimators of both the nonparametric and parametric parts are shown to be obtained when the working correlation matrix equals the actual correlation matrix. The asymptotic covariance matrix of the parameter estimator is consistently estimated by the sandwich estimator. We show that the semiparametric efficient score takes on a simple form and our profile-kernel method is semiparametric efficient. The results for the case where the nonparametric part of the model is an observation-level covariate are noted to be dramatically different.

Some key words: Clustered data; Generalised estimation equation; Kernel method; Longitudinal data; Nonparametric regression; Partially linear model; Profile method; Sandwich estimator; Semiparametric efficiency bound; Semiparametric efficient score.

1. INTRODUCTION

The analysis of clustered data, as in longitudinal studies and familial studies, where clusters refer to subjects and families respectively, is challenged by the fact that observations within each cluster are likely to be correlated. We consider in this paper estimation in a semiparametric partially generalised linear model for clustered data using estimating equations. The model assumes that the mean of the outcome variable depends on some covariates parametrically and a cluster-level covariate nonparametrically.

Suppose the data are arranged in a series of n clusters. For the jth observation, $j = 1, \ldots, m_i$, of the ith cluster, $i = 1, \ldots, n$, we have a response Y_{ij}, a $p \times 1$ vector of covariates X_{ij} and a scalar cluster-level covariate Z_i. For example, Z_i is a time-independent covariate, such as a baseline covariate, in longitudinal data; or a family-level covariate, such as social economic index, in familial studies. The mean of Y_{ij}, given all the covariates in the cluster, is μ_{ij} and satisfies

$$g(\mu_{ij}) = X_{ij}^{\mathrm{T}}\beta_0 + \theta(Z_i), \tag{1}$$

where $g(.)$ is a monotonic differentiable link function. The marginal variance is $\mathrm{var}(Y_{ij}) = \phi w_{ij}^{-1} v(\mu_{ij})$, where ϕ is a scale parameter, w_{ij} is a known weight and $v(.)$ is a known variance function.

For independent data, the partially generalised linear model has been the subject of considerable investigation; see Speckman (1988), Hastie & Tibshirani (1990), Severini & Wong (1992) and Severini & Staniswalis (1994), among many others. Severini & Staniswalis (1994) suggest estimating $\theta(.)$ using standard nonparametric regression methods such as kernel regression with standard bandwidths, and estimating β_0 using the profile method. They show that the final estimator of β_0 is asymptotically semiparametrically efficient.

We study the profile-kernel generalised estimating equation method of Severini & Staniswalis (1994) in the semiparametric model (1) for clustered data. We show that most of the pleasant properties of the independent-data case carry over, namely that off-the-shelf nonparametric regression methods can be used, ordinary parametric rates of convergence apply for any working correlation matrix, the most efficient estimator is obtained when the working and true correlation matrices coincide and the methods are semiparametric efficient. The working correlation matrix refers to a user-specified correlation matrix that may or may not be the true correlation matrix.

This result is different from the case where the nonparametric part $\theta(.)$ of model (1) depends on an observation-level covariate Z_{ij}, that is $g(\mu_{ij}) = X_{ij}^{\mathrm{T}}\beta_0 + \theta(Z_{ij})$, where Z_{ij} varies within each cluster, such as a time-varying covariate in longitudinal data. Lin & Carroll (2001) show that, in this case, if we use the profile-kernel method of Severini & Staniswalis (1994) appropriate for the clustered data context, both of the independent data regression results are false: an $n^{1/2}$-consistent estimator of β_0 requires the assumption of working independence or an artificial undersmoothing when the true covariance is used, by which we mean choosing the bandwidth parameter smaller than that given by crossvalidation, and even then it is not semiparametric efficient.

In § 2 we describe the problem formulation and the profile-kernel generalised estimating equation methods. Section 3 gives our main results, § 4 gives a brief data example and concluding remarks are given in § 5. Technical details are collected into an appendix.

2. SEMIPARAMETRIC MODEL FOR CLUSTERED DATA

In our semiparametric marginal model, we model the effects of X_{ij} $(p \times 1)$ parametrically and the effects of the scalar and cluster-level covariate Z_i nonparametrically, and we treat the within-cluster correlation parameters as nuisance parameters. Let U be a vector of ones. In matrix notation, if we set $\mu_i = (\mu_{i1}, \ldots, u_{im_i})^{\mathrm{T}}$, $Y_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and X_{i} similarly, $i = 1, \ldots, n$, we have $\mu_i = \mu\{X_i\beta_0 + \theta(Z_i)U\}$, where $\mu(.) = g^{-1}(.)$.

The natural local-linear version of the profile-kernel method of Severini & Staniswalis (1994) is defined as follows. Let $K(.)$ be a symmetric kernel function normalised without loss of generality to have variance one, and let h be the bandwidth satisfying $h \sim n^{-1/5}$, which is obtained by choosing h using leaving-one-cluster-out crossvalidation. For any fixed β, we estimate $\theta(z)$ by solving the estimating equations in $\alpha = (\alpha_0, \alpha_1)^{\mathrm{T}}$ given below. If we estimate α at z by $\hat{\alpha}$, the local linear kernel estimator of $\theta(z)$ is $\hat{\theta}(z, \beta) = \hat{\alpha}_0$. For fixed β, the estimating equations for α are

$$\sum_{i=1}^{n} K_{ih}(z) T_i^{\mathrm{T}}(z) \Delta_i(z, \beta, \alpha) V_i^{-1}(z, \beta, \alpha, \delta)\{Y_i - \mu_i(z, \beta, \alpha)\} = 0, \tag{2}$$

where

$$\begin{aligned} K_{ih}(z) &= h^{-1}K\{(Z_i - z)/h\}, \quad \mu_{ij}(z, \beta, \alpha) = \mu\{X_{ij}^{\mathrm{T}}\beta + \alpha_0 + \alpha_1(Z_i - z)/h\}, \\ T_i(z) &= \{U, U(Z_i - z)/h\}, \quad \mu_i(z, \beta, \alpha) = \{\mu_{i1}(z, \beta, \alpha), \ldots, \mu_{im_i}(z, \beta, \alpha)\}^{\mathrm{T}}, \\ \Delta_i(z, \beta, \alpha) &= \mathrm{diag}\{\mu_{ij}^{(1)}(z, \beta, \alpha)\}, \quad S_i(z, \beta, \alpha) = \mathrm{diag}[\phi w_{ij}^{-1} v\{\mu_{ij}(z, \beta, \alpha)\}], \\ V_i(z, \beta, \alpha, \delta) &= S_i^{1/2}(z, \beta, \alpha) R_i(\delta) S_i^{1/2}(z, \beta, \alpha). \end{aligned} \tag{3}$$

Here $\mu^{(1)}(.)$ is the first derivative of $\mu(.)$, S_i contains the marginal variances of the Y_{ij}, and $R_i(\delta)$ is

an invertible working correlation matrix depending on a parameter vector δ, which can be estimated using the method of moments (Liang & Zeger, 1986). Estimation of β proceeds by solving the profile estimating equations

$$\sum_{i=1}^{n} \frac{\partial \mu\{X_i\beta + \hat{\theta}(Z_i, \beta)U\}^{\mathrm{T}}}{\partial \beta} V_i^{-1}(\beta, \delta)[Y_i - \mu\{X_i\beta + \hat{\theta}(Z_i, \beta)U\}] = 0, \tag{4}$$

where $V_i(.)$ takes the same form as (3) except that S_i is evaluated at the $\mu\{X_{ij}^{\mathrm{T}}\beta + \hat{\theta}(Z_i, \beta)\}$.

The joint solution of (2) and (4) is called the profile-kernel estimator $\{\hat{\theta}(z), \hat{\beta}\}$. The Fisher scoring algorithm can be used to solve (2) and (4). A sandwich-type estimator can be used to estimate the covariance matrix of $\hat{\beta}$ and takes the form

$$\mathrm{cov}(\hat{\beta}) = P^{-1}\left(\sum_{i=1}^{n} Q_i^{\mathrm{T}} V_i^{-1} e_i e_i^{\mathrm{T}} V_i^{-1} Q_i\right) P^{-1}, \tag{5}$$

where

$$Q_i = \Delta_i\{X_i + U\,\partial\hat{\theta}(Z_i, \beta)/\partial\beta\}, \quad e_i = Y_i - \mu\{X_i\beta + \hat{\theta}(Z_i, \beta)U\}, \quad P = \sum_{i=1}^{n} Q_i^{\mathrm{T}} V_i^{-1} Q_i,$$

and Q_i, V_i and $P_{\mathbf{i}}$ are all evaluated at $\{\hat{\beta}, \hat{\theta}(.)\}$. Calculations of $\partial\hat{\theta}(Z_i, \beta)/\partial\beta$ in Q_i can proceed in the same way as that described in Appendix C of Lin & Carroll (2001).

3. Main results

We study in this section the asymptotic properties of the profile-kernel estimator $\{\hat{\theta}(z), \hat{\beta}\}$. Our main results are summarised in the following two theorems. Theorem 1 provides the asymptotic properties of the kernel estimator of the nonparametric function $\hat{\theta}(z)$. Theorem 2 gives the asymptotic properties of the profile estimator $\hat{\beta}$. We assume in both Theorems that the cluster size $m_i = m < \infty$, $n \to \infty$ and $h = O(n^{-1/5})$. Furthermore, we assume the estimated correlation parameter vector $\hat{\delta}$ is $n^{1/2}$-consistent for some δ_0, that is $n^{1/2}(\hat{\delta} - \delta_0) = O_p(1)$ for some δ_0. Note that the method-of-moments estimators of δ satisfy this condition. Suppose the true covariance matrix of Y_i is Σ. The proofs are given in the Appendix.

Theorem 1. *If $\hat{\beta}$ is $n^{1/2}$-consistent, then, for any specified covariance matrix V, the asymptotic bias and variance of the kernel estimator $\hat{\theta}(z) = \hat{\theta}(z, \hat{\beta})$ are*

$$\mathrm{bias}\{\hat{\theta}(z)\} = (h^2/2)\theta^{(2)}(z) + o(h^2)$$

$$\mathrm{var}\{\hat{\theta}(z)\} = \frac{\int K^2(v)\,dv}{nh} \frac{E(U^{\mathrm{T}}\Delta V^{-1}\Sigma V^{-1}\Delta U \mid Z = z)}{\mathscr{A}^2(Z)f(z)} + o\{(nh)^{-1}\},$$

where $f(.)$ is the density function for Z, $\mathscr{A}(z) = E(U^{\mathrm{T}}\Delta V^{-1}\Delta U \mid Z = z)$, and the jth component of $\mu(.)$ in Δ and V is evaluated at $X_{ij}\beta_0 + \theta(Z_i)$. In the mean squared error sense the optimal rate of convergence occurs when $h \sim n^{-1/5}$. The most efficient estimator of $\hat{\theta}(z)$ is obtained when $V = \Sigma$, that is when the limit of the estimated correlation matrix $R(\delta_0)$ is the true correlation matrix for some δ_0. In this case, the variance of $\hat{\theta}(.)$ is $\{nhf(z)\}^{-1}\int K^2(v)\,dv\mathscr{A}^{-1}(z)$ with $V = \Sigma$ in the definition of $\mathscr{A}(z)$.

Theorem 1 shows that, if we use the conventional kernel method and conventional smoothing, for example choosing the bandwidth parameter h using leaving-one-cluster-out crossvalidation, the estimator of the nonparametric component $\hat{\theta}(.)$ has the properties of standard nonparametric estimators and is most efficient when assuming the true correlation matrix.

Theorem 2. *The following results hold for the profile-kernel estimator $\hat{\beta}$.*

(i) *For any specified covariance matrix V, if $h \sim n^{-1/5}$,*

$$n^{1/2}(\hat{\beta} - \beta_0) \to N\{0, C^{-1}E(\tilde{X}^{\mathrm{T}}\Delta V^{-1}\Sigma V^{-1}\Delta\tilde{X})C^{-1}\},$$

in distribution, where

$$\tilde{X} = X - UB(Z), \quad B(z) = E(U^{\mathrm{T}}\Delta V^{-1}\Delta X \,|\, Z = z)/\mathscr{A}(z), \quad C = E(\tilde{X}^{\mathrm{T}}\Delta V^{-1}\Delta\tilde{X}).$$

The above asymptotic covariance matrix of $\hat{\beta}$ *is consistently estimated by the sandwich covariance estimator* (5). *The optimal estimator of* $\hat{\beta}$ *is obtained when* $V = \Sigma$, *that is when* $R(\delta_0)$ *is the true correlation matrix for some* δ_0. *In this case, the asymptotic covariance matrix for* $\hat{\beta}$ *is* C^{-1} *with* $V = \Sigma$ *in the definition of* C.

(ii) *If* Y_i *follows the quadratic exponential family* (*Diggle et al.*, 1994, *p.* 147), *the profile-kernel estimator* $\hat{\beta}$ *is semiparametric efficient when* $V = \Sigma$, *that is when* $R(\delta_0)$ *is the true correlation matrix for some* δ_0.

Theorem 2 shows that, if we use the conventional profile-kernel method and smooth by, say, leaving-one-cluster-out crossvalidation, the estimator of the parametric component $\hat{\beta}$ is $n^{1/2}$-consistent for any working correlation matrix R and is semiparametric efficient when assuming the true correlation matrix. Note that the asymptotic distribution of $\hat{\beta}$ does not require the quadratic exponential family assumption of the Y_{ij}. This assumption is needed to compute the semiparametric efficiency bound

4. APPLICATION TO INFECTIOUS DISEASE DATA

We applied our methods to a longitudinal infectious disease study involving 275 preschool children who were re-examined every three months for 18 months for the presence of respiratory infection, a binary variable, modelled via logistic regression (Diggle et al., 1994, p. 158). The primary interest was in studying the relationship of the risk of infectious disease to Vitamin A deficiency. Other variables included in the study were age, gender, height, stunting and dummy variables for the visit number. In this example, the clusters are the children. We considered baseline age to be the cluster-level covariate modelled nonparametrically, with the other variables modelled parametrically. This model separates out the cross-sectional age effect and the longitudinal time effect. Preliminary analysis suggested that the correlations of observations within subjects were no more than 0·10, so in our analysis we assumed working independence while accounting for the correlation using the sandwich estimator. The bandwidth was estimated using the empirical-bias bandwidth selector method of Ruppert (1997).

In Fig. 1(a), we display the semiparametric estimate of the effect of baseline age $\theta(.)$, along with its 95% pointwise 'confidence interval'. Figure 1(a) shows that the risk of respiratory infection increases with age for children less than two years old and decreases with age for children older than two. Figure 1(b) compares the fit from the semiparametric model to those of the linear and quadratic baseline age models. Table 1 compares estimates of the parameters using the semiparametric model, as well as when modelling the effect of baseline age as linear or quadratic. Although not displayed here, we also fitted a cubic term to baseline age and it was highly statistically significant. Hence it is desirable to model the baseline age effect nonparametrically.

Our analysis suggests that there is some evidence for the vitamin A effect on the risk of respiratory infection. The effects of sex and time are significant as well. Misspecification of the age effects using the linear and quadratic functions results in considerable bias in the parametric covariate effects, especially the stunting effect.

5. DISCUSSION

Our results stand in stark contrast to what happens if the covariate modelled nonparametrically is observation-level and not, as in our paper, at the cluster-level. For example, when the covariate is observation-level, for example time-varying in longitudinal data, estimation of the nonparametric part using the method of Severini & Staniswalis (1994) generally requires one to assume working independence or to undersmooth artificially, and even then it is not semiparametric efficient (Lin & Carroll, 2001).

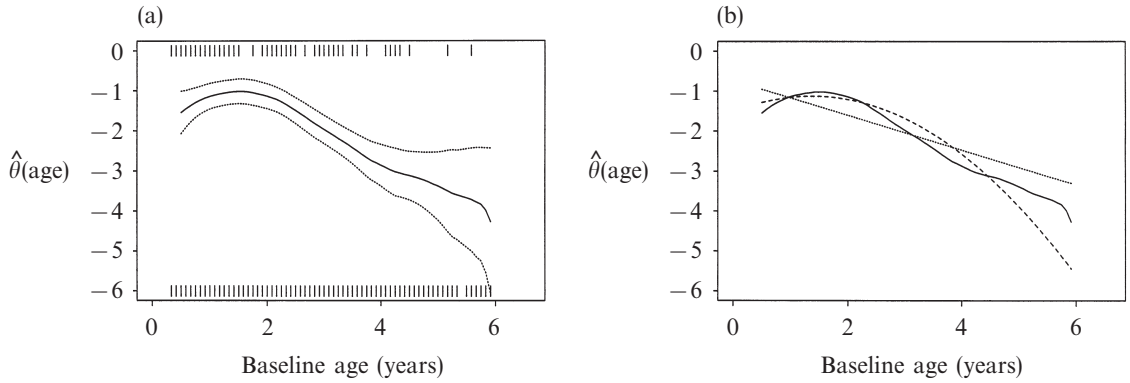

Fig. 1. (a) Estimated kernel estimate $\hat{\theta}$(age) when fitting the semiparametric model to the infectious disease data assuming working independence. Also displayed are 95% pointwise confidence intervals: solid line, $\hat{\theta}$(age); dotted lines, 95% CI. The vertical strokes at 0 and −6 indicate the occurrence of 1 and 0 in the response. (b) Comparison of the kernel estimate $\hat{\theta}$(age) and the estimated linear and quadratic baseline age models: solid line, kernel estimate; dotted line, linear estimate; dashed line, quadratic estimate.

Table 1. *Regression coefficient estimates in the infectious data, assuming working independence and using sandwich estimators of standard errors. Baseline age is the cluster-level covariate modelled nonparametrically in the semiparametric model. The linear and quadratic age models refer to parametric logistic models that assume that the baseline age effect is linear and quadratic, respectively*

	Semiparametric model		Linear age model		Quadratic age model	
	Estimate	SE	Estimate	SE	Estimate	SE
Vitamin A	0·74	0·38	0·87	0·42	0·80	0·41
Sex	−0·63	0·20	−0·39	0·24	−0·53	0·24
Height	−0·03	0·02	−0·04	0·03	−0·04	0·03
Stunting	0·60	0·36	0·27	0·44	0·32	0·44
Visit 2	−1·15	0·31	−1·17	0·39	−1·15	0·39
Visit 3	−0·73	0·30	−0·76	0·37	−0·73	0·37
Visit 4	−1·53	0·32	−1·52	0·41	−1·49	0·41
Visit 5	0·16	0·25	0·06	0·31	0·11	0·30
Visit 6	−0·38	0·27	−0·45	0·33	−0·39	0·33

SE, estimated standard error.

The difference in the results between these two cases can be explained by the different performance of kernel regression of the nonparametric part $\theta(.)$. To obtain the most efficient kernel estimator of $\theta(.)$, one needs to account for the correlation when the covariate Z is cluster-level, but needs to assume independence when the covariate Z is observation-level. In the former case, for each fixed β, one has a single cluster at each Z with correlated replicates of the Y_{ij}. Smoothing across several values of Z brings in all observations of different clusters and hence one needs to account for the correlation. In the latter case, however, since the kernel method is local, the same smoothing brings in different observations from different clusters.

Acknowledgement

We are grateful to the editor, associate editor and referee for helpful suggestions. Lin's research was supported by a grant from the National Cancer Institute. Carroll's research was supported by

a grant from the National Cancer Institute, and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences. Carroll's research took place during a visit to the Centre for Mathematics and its Applications at the Australian National University, with partial support from an Australian Research Council Large Research Grant.

APPENDIX

Proofs of Theorems 1 *and* 2

We first show the results assuming the working correlation parameters δ are known. Define

$$\mathscr{A}(z) = E(U^{\mathrm{T}}\Delta V^{-1}\Delta U \,|\, Z = z), \quad \varepsilon_i = Y_i - \mu\{X_i\beta_0 + \theta(Z_i)U\}.$$

It is readily shown that the solution to (2) with β replaced by β_0 has the asymptotic expansion

$$\hat{\theta}(z, \beta_0) - \theta(z) = \left(\frac{h^2}{2}\right)\theta^{(2)}(z) + \{\mathscr{A}(z)f(z)\}^{-1}n^{-1}\sum_{i=1}^{n} K_h(Z_i - z)U^{\mathrm{T}}\Delta_i V_i^{-1}\varepsilon_i + o_p\{(nh)^{-1/2} + h^2\}. \tag{A1}$$

The asymptotic bias and variance of $\hat{\theta}(z)$ given in Theorem 1 can hence be easily derived from (A1). The variance of $\hat{\theta}(z)$ is clearly minimised by choosing $V = \Sigma$.

Using (A1), we have $\partial\hat{\theta}(z, \beta_0)/\partial\beta^{\mathrm{T}} = -B(z) + o_p(1)$, where

$$B(z) = E(U^{\mathrm{T}}\Delta V^{-1}\Delta X \,|\, Z = z)/\mathscr{A}(z).$$

Let $\tilde{X} = X - UB(Z)$, $C_n = n^{-1}\sum_i \tilde{X}_i^{\mathrm{T}}\Delta_i V_i^{-1}\Delta_i\tilde{X}_i$ and $C = E(\tilde{X}^{\mathrm{T}}\Delta V^{-1}\Delta\tilde{X})$. If we use (A1) and $\hat{\theta}(z, \beta_0) - \theta(z) = O_p(n^{-2/5})$ with $h = n^{-1/5}$, a Taylor expansion of (4) shows that

$$C_n n^{1/2}(\hat{\beta} - \beta_0) = D_{1n} + D_{2n} + o_p(1), \tag{A2}$$

where

$$\begin{aligned}
D_{1n} &= n^{-1/2}\sum_{i=1}^{n}\tilde{X}_i^{\mathrm{T}}\Delta_i V_i^{-1}\varepsilon_i,\\
D_{2n} &= n^{-1/2}\sum_{i=1}^{n}\tilde{X}_i^{\mathrm{T}}\Delta_i V_i^{-1}[\mu\{X_i\beta_0 + \theta(Z_i)U\} - \mu\{X_i\beta_0 + \hat{\theta}(Z_i, \beta_0)U\}]\\
&= -D_{21n} - D_{22n} + o_p(1),\\
D_{21n} &= \left(\frac{h^2}{2}\right)n^{-1/2}\sum_{i=1}^{n}\tilde{X}_i^{\mathrm{T}}\Delta_i V_i^{-1}\Delta_i U\theta^{(2)}(Z_i),\\
D_{22n} &= n^{-1/2}\sum_{i=1}^{n}\left[n^{-1}\sum_{i'=1}^{n}\tilde{X}_{i'}^{\mathrm{T}}\Delta_{i'}V_{i'}^{-1}\Delta_{i'}U\{\mathscr{A}(Z_{i'})f(Z_{i'})\}^{-1}K_h(Z_{i'} - Z_i)\right]U^{\mathrm{T}}\Delta_i V_i^{-1}\varepsilon_i.
\end{aligned} \tag{A3}$$

Since $E(\tilde{X}^{\mathrm{T}}\Delta V^{-1}\Delta U \,|\, Z) = 0$, we have $D_{21n} = o_p(1)$. Since the term inside the square bracket of D_{22n} is asymptotically equivalent to $E[\tilde{X}^{\mathrm{T}}\Delta V^{-1}\Delta U\{\mathscr{A}(Z)f(Z)\}^{-1}] + h^2 = h^2$, we have $D_{22n} = o_p(1)$. Hence $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to $N\{0, C^{-1}E(\tilde{X}^{\mathrm{T}}\Delta V^{-1}\Sigma V^{-1}\Delta\tilde{X})C^{-1}\}$. It can be easily shown that the asymptotic covariance matrix of $\hat{\beta}$ is consistently estimated by the sandwich estimator (5). The optimal choice of the working covariance matrix V is clearly the actual covariance Σ, in which case the asymptotic covariance of $\hat{\beta}$ is C^{-1} with $V = \Sigma$ in the definition of C.

We next show that, when Y_i follows a quadratic exponential family, the asymptotic covariance matrix of $\hat{\beta}$ equals the inverse of the semiparametric information bound and is hence semiparametric efficient. Suppressing the cluster index i, suppose that the outcome vector Y of the ith cluster follows a quadratic exponential family with the loglikelihood function

$$l(Y; \gamma_1, \gamma_2) = \gamma_1^{\mathrm{T}}Y + \gamma_2^{\mathrm{T}}W + c_1(\gamma_1, \gamma_2) + c_2(Y),$$

where $c_1(.)$ and $c_2(.)$ are some specific functions, W is a vector containing elements $Y_j Y_k$, for $j, k = 1, \ldots, m$, for binary data $j \neq k$, and γ_2 represents the covariance parameters of Y. Suppose that γ_2 is known. There is a one-to-one transformation between γ_1 and $\mu = E(Y)$, where $\mu = \mu\{X\beta + \theta(Z)U\}$. If we denote $\mathrm{cov}(Y)$ by Σ, the score functions of β and $\theta(z)$ are

$$\dot{l}_\beta = X^{\mathrm{T}}\Delta\Sigma^{-1}[Y - \mu\{X\beta + \theta(Z)U\}], \quad \dot{l}_\theta = U^{\mathrm{T}}\Delta\Sigma^{-1}[Y - \mu\{X\beta + \theta(Z)U\}].$$

If we follow Lin & Carroll (2001), the efficient score can be shown to be $\dot{l}_\beta - \dot{l}_\theta \xi_*(Z)$, where $\xi_*(.)$ satisfies $E[\{\dot{l}_\beta - \dot{l}_\theta \xi_*(Z)\}\dot{l}_\theta \xi(Z)] = 0$ for all $\xi(Z)$. Hence $\xi_*(Z)$ satisfies

$$E[\{X - \xi_*(Z)U\}^{\mathrm{T}}\Delta\Sigma^{-1}\Delta U \,|\, Z] = 0,$$

that is

$$\xi_* = \frac{E(U^{\mathrm{T}}\Delta\Sigma^{-1}\Delta X \,|\, Z)}{E(U^{\mathrm{T}}\Delta\Sigma^{-1}\Delta U \,|\, Z)} = B(Z), \quad X - U\xi_*(Z) = X - UB(Z) = \tilde{X}.$$

The semiparametric efficient score is hence $\tilde{X}^{\mathrm{T}}\Delta\Sigma^{-1}(Y - \mu)$, which is asymptotically the same as the profile-kernel score (4) with $V = \Sigma$. The information bound is $E(\tilde{X}\Delta\Sigma^{-1}\Delta\tilde{X})$, whose inverse is the same as the covariance of the profile-kernel estimator $\hat{\beta}$ with $V = \Sigma$.

We next sketch an argument showing that, if δ is estimated by $\hat{\delta}$ with the property that, for some δ_0, $n^{1/2}(\hat{\delta} - \delta_0) = O_p(1)$, the asymptotic distribution of $\hat{\beta}$ is not affected. Specifically, in (2)–(4), $\hat{\delta}$ is added to the arguments for $\Delta_i(.)$ and $V_i(.)$, while in the latter equation $\hat{\theta}(z, \beta)$ is replaced by $\hat{\theta}(z, \beta, \hat{\delta})$. Using (A1), we first note that $\partial\hat{\theta}(z, \beta_0, \delta_0)/\partial\delta = o_p(1)$. Then (A2) still holds with $\hat{\theta}(z, \beta_0)$ in D_{2n} replaced by $\hat{\theta}(z, \beta_0, \hat{\delta})$. This adds the extra term $-D_{23n}$ to (A3), where

$$D_{23n} = n^{-1}\sum_{i=1}^{n} \tilde{X}_i^{\mathrm{T}}\Delta_i V_i^{-1}\Delta_i U\{\partial\hat{\theta}(Z_i, \beta_0, \delta_0)/\partial\delta^{\mathrm{T}}\}n^{1/2}(\hat{\delta} - \delta_0).$$

Since $\partial\hat{\theta}(Z_i, \beta_0, \delta_0)/\partial\delta = o_p(1)$, we have that $D_{23n} = o_p(1)$. The rest of the proof is the same.

References

Diggle, P. J., Liang, K. Y. & Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, X. & Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.* To appear.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Assoc.* **92**, 1049–62.

Severini, T. A. & Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* **89**, 501–11.

Severini, T. A. & Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768–802.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B **50**, 413–36.

[*Received July* 2000. *Revised March* 2001]