

Partially linear models with missing response variables and error-prone covariates

BY HUA LIANG

Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, U.S.A.
hliang@bst.rochester.edu

SUOJIN WANG AND RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
sjwang@stat.tamu.edu carroll@stat.tamu.edu

SUMMARY

We consider partially linear models of the form $Y = X^T\beta + \nu(Z) + \varepsilon$ when the response variable Y is sometimes missing with missingness probability π depending on (X, Z) , and the covariate X is measured with error, where $\nu(z)$ is an unspecified smooth function. The missingness structure is therefore missing not at random, rather than the usual missing at random. We propose a class of semiparametric estimators for the parameter of interest β , as well as for the population mean $E(Y)$. The resulting estimators are shown to be consistent and asymptotically normal under general assumptions. To construct a confidence region for β , we also propose an empirical-likelihood-based statistic, which is shown to have a chi-squared distribution asymptotically. The proposed methods are applied to an AIDS clinical trial dataset. A simulation study is also reported.

Some key words: Confidence region; Empirical likelihood; Estimating equation; Measurement error; Missing data; Missing not at random; Nonparametric regression; Semiparametric estimation.

1. INTRODUCTION

The partially linear model assumes that the response variable Y depends on variable X in a linear way but is related to another independent variable Z in an unspecified manner:

$$Y = X^T\beta + \nu(Z) + \varepsilon, \quad (1)$$

where X is a p -vector covariate, Z is a scalar covariate, the function $\nu(\cdot)$ is unknown, and the model error ε has mean zero conditional on (X, Z) .

There is a substantial literature on kernel-based methods for partially linear models and their generalizations; see for example Engle et al. (1986), Speckman (1988), Robinson (1988), Severini & Staniswalis (1994), Zeger & Diggle (1994), Opsomer & Ruppert (1999) and Härdle et al. (2000), among many others. Liang et al. (1999) considered model (1) with error-prone X . Recently, missing-data issues have been considered, with Liang et al. (2004) considering the case in which X in (1) is missing at random, while in Wang et al. (2004) the response Y is missing at random.

In this paper, we consider the missing response case, but in addition we allow some of the components of X to be measured with error. Our motivation is from AIDS clinical trials, where the response, variable viral load RNA, can be missing. In addition, the covariates, CD4 measurements, are measured with error. Measurement error in predictors causes a bias in the estimated regression coefficient. While Liang et al. (1999) considered the measurement error problem, they did not allow for missing responses.

We assume that we can observe a surrogate W related to X by

$$W = X + U. \quad (2)$$

If $\delta = 1$ indicates that Y is observed and $\delta = 0$ indicates that Y is missing, we assume that the measurement error U is independent of (Y, Z, X, δ) and with $E(U) = 0$, $\text{cov}(U) = \Sigma_{uu}$. We first assume that Σ_{uu} is known, and later extend the results to the general case.

It is important to stress that, under the setting considered in this paper, see (3) below, the missingness of Y is allowed to depend on (X, Z) , but not otherwise on W . Since the true X is not observable, Y is therefore not missing at random. Since we will make no further assumption, such as about the distribution of X or about the missing-data probabilities, what we are dealing with here is conceptually quite different from most studies of missing data in which missing at random or missing completely at random is assumed.

2. ESTIMATION AND MAIN RESULTS

2.1. Known measurement error covariance matrix

As described previously, we assume that, if X were observed, the missing data mechanism would follow the missing at random mechanism in the sense that

$$\text{pr}(\delta = 1|X, Z, Y) = \text{pr}(\delta = 1|X, Z) = \pi(X, Z), \quad (3)$$

for some unknown $\pi(X, Z)$. Since we also assume that the measurement errors U are independent of (Y, Z, X, δ) , $\text{pr}(\delta = 1|X, Z, Y, W) = \pi(X, Z)$. Furthermore, we assume that $\{(Y_i, W_i, Z_i, \delta_i), i = 1, \dots, n\}$ are independent and identically distributed.

We use locally constant smoothers with fixed bandwidths for simplicity in presenting the derivations of the theoretical results: extensions to local polynomial estimation are straightforward, with no change in the limiting distribution of the estimator of β . In what follows, we write $A^{\otimes 2} = AA^T$. Also, define $m_x(z) = E(\delta X|Z = z)/E(\delta|Z = z)$, $m_w(z) = E(\delta W|Z = z)/E(\delta|Z = z)$ and $m_y(z) = E(\delta Y|Z = z)/E(\delta|Z = z)$. Let $\tilde{X}_i = X_i - m_x(Z_i)$, $\tilde{W}_i = W_i - m_w(Z_i)$, $\tilde{Y}_i = Y_i - m_y(Z_i)$, and denote $\text{cov}(\tilde{X}\delta)$ by $\Sigma_{X|Z}$.

Note that $\delta Y = \delta X^T \beta + \delta v(Z) + \delta \varepsilon$. From our assumptions, it follows that $E(\delta Y|Z) = E(\delta X^T|Z)\beta + E(\delta|Z)v(Z)$. If X is observed and $m_x(z)$ and $m_y(z)$ are known, one can obtain a least-squares-type estimator of β as

$$\left[\sum_{i=1}^n \delta_i \{X_i - m_x(Z_i)\}^{\otimes 2} \right]^{-1} \left[\sum_{i=1}^n \delta_i \{X_i - m_x(Z_i)\} \{Y_i - m_y(Z_i)\} \right].$$

The expectation of the i th term in the second summation is $E\{\pi(X, Z)\tilde{X}\tilde{Y}\} = \Sigma_{X|Z}\beta$. It is easily shown that this estimator is consistent and asymptotically normal (Wang et al., 2004).

The above formula cannot be applied directly when X is measured with error, and $m_x(z)$ and $m_y(z)$ are unknown. However, by our assumptions, $E(\delta W|Z) = E(\delta X|Z)$. We thus

propose an estimator of β that corrects for attenuation:

$$\hat{\beta}_n = \left(\sum_{i=1}^n \delta_i [\{W_i - \hat{m}_w(Z_i)\}^{\otimes 2} - \Sigma_{uu}] \right)^{-1} \left[\sum_{i=1}^n \delta_i \{W_i - \hat{m}_w(Z_i)\} \{Y_i - \hat{m}_y(Z_i)\} \right], \quad (4)$$

where $\hat{m}_w(z)$ and $\hat{m}_y(z)$ are nonparametric regression estimators. Let $K(\cdot)$ be a symmetric density function and h be a suitable bandwidth, and define $K_h(z) = K(z/h)/h$. These estimators take the form

$$\hat{m}_w(z) = \frac{\sum_{i=1}^n K_h(Z_i - z) \delta_i W_i}{\sum_{i=1}^n \delta_i K_h(Z_i - z)}, \quad \hat{m}_y(z) = \frac{\sum_{i=1}^n K_h(Z_i - z) \delta_i Y_i}{\sum_{i=1}^n \delta_i K_h(Z_i - z)}. \quad (5)$$

Remark 1. Alternative estimators are readily constructed, but generally suffer from complications. For example, since $Y - E(Y|Z) = \{X - E(W|Z)\}^T \beta + \varepsilon$, an obvious approach is to estimate $E(W|Z)$ and $E(Y|Z)$ using all the data. The former is easy: any nonparametric regression will do. However, the latter is problematic, because of the missing responses, the possibility that missingness depends on X and the fact that X is unobserved. There does not appear to be an easy way to estimate $E(Y|Z)$ consistently under the current conditions. Note that one of the most important features of the proposed approach is that, through use of the standard measurement error model (2), it can handle the missing not at random case with ease and still provide \sqrt{n} -consistent estimators, as shown in the theorems that follow.

Before presenting our first main result, we note that throughout the paper we make some general assumptions that are listed in the Appendix.

THEOREM 1. *Assume that $\{(Y_i, W_i, Z_i, \delta_i), i = 1, \dots, n\}$ are independent and identically distributed. Under Assumptions A1–A7 in the Appendix, $n^{1/2}(\hat{\beta}_n - \beta)$ is asymptotically normally distributed with mean 0 and covariance matrix $\Sigma_\beta = \Sigma_{X|Z}^{-1} \Gamma \Sigma_{X|Z}^{-1}$, where*

$$\Gamma = E \left[\delta \left\{ (\varepsilon - U^T \beta) \tilde{X} \right\}^{\otimes 2} \right] + E(\delta U U^T \varepsilon^2) + E \left[\delta \left\{ (U U^T - \Sigma_{uu}) \beta \right\}^{\otimes 2} \right].$$

The proof of Theorem 1 and those of Theorems 2 and 3 given below generally use a technique similar to that used by Liang et al. (2004) to prove their Theorems 1 and 2. We omit the details, which can be found in an earlier version of this article, available from the authors.

Remark 2. In typical nonparametric kernel regression, bandwidth selection plays a key role in the performance of nonparametric estimators in terms of their bias and variance. In partially linear models, β is of main interest, and $v(z)$ is a nuisance function. Based on Assumption A2, only the rate of order $n^{-1/5}$ is needed to lead to the same limit distribution for estimating β . In implementing our proposed estimation procedure, we adopt Ruppert et al.'s (1995) method of choosing the bandwidth. Our limited experience indicates that the numerical performance of the resulting estimators of β is stable around the selected bandwidth.

From the proof of Theorem 1, it is seen that Σ_β can be estimated via a standard sandwich method as follows. Let

$$\hat{\Sigma}_{X|Z} = n^{-1} \sum_{i=1}^n \delta_i \{ \{W_i - \hat{m}_w(Z_i)\}^{\otimes 2} - \Sigma_{uu} \},$$

$$\hat{\Gamma} = n^{-1} \sum_{i=1}^n \delta_i \left(\{W_i - \hat{m}_w(Z_i)\} [Y_i - \hat{m}_y(Z_i) - \{W_i - \hat{m}_w(Z_i)\}^T \hat{\beta}_n] + \Sigma_{uu} \hat{\beta}_n \right)^{\otimes 2}$$

and $\hat{\Sigma}_\beta = \hat{\Sigma}_{X|Z}^{-1} \hat{\Gamma} \hat{\Sigma}_{X|Z}^{-1}$. Then it is easily shown that $\hat{\Sigma}_\beta$ is a consistent estimator of Σ_β .

2.2. *Estimated measurement error covariance matrix*

The covariance matrix Σ_{uu} is generally unknown and needs to be estimated. The usual method of doing so (Carroll et al., 1995, Ch.3) is by partial replication, so that we observe $W_{ij} = X_i + U_{ij}$, $j = 1, \dots, m_i$. For notational simplicity, we assume that $m_i \equiv 2$. Extension to more general settings is straightforward; see Liang et al. (1999) for a related discussion. Let \bar{W}_i be the sample mean of the replicates W_{ij} . A consistent, unbiased, method-of-moments estimator for Σ_{uu} is

$$\hat{\Sigma}_{uu} = n^{-1} \sum_{i=1}^n \sum_{j=1}^2 (W_{ij} - \bar{W}_i)(W_{ij} - \bar{W}_i)^T.$$

The corresponding estimator of β is

$$\hat{\beta}_{n,2} = \left(\sum_{i=1}^n \delta_i \left[\{ \bar{W}_i - \hat{m}_{\bar{w}}(Z_i) \}^{\otimes 2} - (1/2) \hat{\Sigma}_{uu} \right] \right)^{-1} \left[\sum_{i=1}^n \delta_i \{ \bar{W}_i - \hat{m}_{\bar{w}}(Z_i) \} \{ Y_i - \hat{m}_y(Z_i) \} \right], \quad (6)$$

where $\hat{m}_{\bar{w}}(z)$ is the locally constant estimator of $m_w(z)$ based on the data $\{(\bar{W}_i, Z_i), i = 1, \dots, n\}$. We now present the following theorem.

THEOREM 2. *Under the general conditions of Theorem 1, the estimator $\hat{\beta}_{n,2}$ given in (6) is consistent and asymptotically normal with covariance matrix $\Sigma_{X|Z}^{-1} \Gamma^* \Sigma_{X|Z}^{-1}$, where*

$$\Gamma^* = E \left[\delta \left\{ (\varepsilon - \bar{U}^T \beta) \tilde{X} \right\}^{\otimes 2} \right] + E(\delta \bar{U} \bar{U}^T \varepsilon^2) + E \left[\delta \left\{ (\bar{U} \bar{U}^T - \Sigma_{uu}/2) \beta \right\}^{\otimes 2} \right].$$

By a straightforward but tedious derivation, Theorem 2 can be proved in a manner similar to Theorem 1.

The standard error estimators can also be derived. A consistent estimator of $\Sigma_{X|Z}$ in this case is given by

$$n^{-1} \sum_{i=1}^n \delta_i \left[\{ \bar{W}_i - \hat{m}_{\bar{w}}(Z_i) \}^{\otimes 2} - (1/2) \hat{\Sigma}_{uu} \right].$$

The Γ^* can be estimated as follows. Let

$$R_i = \{ \bar{W}_i - \hat{m}_{\bar{w}}(Z_i) \} \left[Y_i - \hat{m}_y(Z_i) - \{ \bar{W}_i - \hat{m}_{\bar{w}}(Z_i) \}^T \hat{\beta}_{n,2} \right] \\ + (1/2) \left\{ (W_{i1} - W_{i2})^{\otimes 2} - \hat{\Sigma}_{uu} \right\} \hat{\beta}_{n,2}.$$

Then a consistent estimator of Γ^* is the sample covariance matrix of the $R_i \delta_i$'s (Liang et al., 1999).

3. ESTIMATION OF THE MEAN $E(Y)$

It is of interest to estimate the mean $E(Y) = \theta$. Cheng (1994) studied this problem in the purely nonparametric regression case, while Wang et al. (2004) studied the partially linear model with X observed. Here we construct three estimators of θ when X is not observed. The methods are analogous to those of Wang et al. (2004) in the case in which X is observed. We obtain that the three estimators are asymptotically equivalent.

In a manner similar to Cheng (1994), we can construct two estimators of θ as follows:

$$\hat{\theta}_{n,ave} = n^{-1} \sum_{i=1}^n \delta_i Y_i + n^{-1} \sum_{i=1}^n (1 - \delta_i) \{W_i^T \hat{\beta}_n + \hat{v}_n(Z_i)\},$$

$$\hat{\theta}_{n,est} = n^{-1} \sum_{i=1}^n \{W_i^T \hat{\beta}_n + \hat{v}_n(Z_i)\},$$

where $\hat{v}_n(z) = \hat{m}_y(z) - \hat{m}_w^T(z) \hat{\beta}_n$ is a nonparametric regression estimator of $v(z)$ based on the completely observed data of $\{(Z_i, Y_i - W_i^T \hat{\beta}_n), i = 1, \dots, n\}$. One can easily show that $\hat{v}_n(z) - v(z) = o_p(n^{-1/3})$, in a way similar to Liang et al. (1999). This rate satisfies our assumption to establish the asymptotic normality of the estimators of θ .

Let $s_n(z) = \sum_{i=1}^n \delta_i K_h(Z_i - z) / \sum_{i=1}^n K_h(Z_i - z)$, $s(z) = E(\delta | Z = z)$ and $P(Z, \delta) = \delta / s(Z)$. We define a third estimator of θ as

$$\hat{\theta}_{n,wei} = n^{-1} \sum_{i=1}^n \frac{\delta_i}{s_n(Z_i)} Y_i + n^{-1} \sum_{i=1}^n \left\{ 1 - \frac{\delta_i}{s_n(Z_i)} \right\} \{W_i^T \hat{\beta}_n + \hat{v}_n(Z_i)\}.$$

Note that, if we try to substitute $s_n(z)$ by an estimator of $\pi(x, z)$, a problem arises because X is measured with error, so that the exact X is not available for estimating $\pi(X, Z)$. In the following theorem, we establish the asymptotic normality of the three estimators, showing that they are asymptotically equivalent.

THEOREM 3. *In addition to the assumptions of Theorem 1, assume that $nh^4 \rightarrow 0$. Then $n^{1/2}(\hat{\theta}_{n,\cdot} - \theta)$ asymptotically has a normal distribution with mean 0 and variance $E\{P(Z, \delta)\varepsilon + \{1 - P(Z, \delta)\}U^T \beta + E(\tilde{W}^T) \Sigma_{X|Z}^{-1} \delta \{\tilde{W}(\varepsilon - U^T \beta) + \Sigma_{uu} \beta\}\}^2 + E\{X^T \beta + v(Z) - \theta\}^2$, where \cdot indicates 'ave', 'est' or 'wei'.*

4. INFERENCE BASED ON EMPIRICAL LIKELIHOOD

Based on our estimators for the covariance matrix or its bootstrap version, one can give a confidence region for either β or $\theta = E(Y)$. Although we have confirmed that the estimator $\hat{\Sigma}_\beta$ given in §2 is consistent, its finite-sample behaviour may be affected by the need to plug in several estimated terms. Furthermore, the confidence region derived by this procedure is based on a normal approximation, which may be not optimistic in small samples. An alternative method is to use the empirical-likelihood principle; see Owen (2001), Qin (1994), Qin & Lawless (1994) and Chen (1994). In the remainder of this section, we assume that the ε_i are independent, identically distributed and independent of (W_i, Z_i) . We need only to study the empirical-likelihood-based confidence interval for β since the case of θ is similar and simpler.

Let F be the distribution function which assigns probability p_i at points (W_i, Y_i, Z_i) . Then $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for each i . Our semiparametric empirical-likelihood ratio is defined as follows. Note that $E[\delta\{\tilde{W}(\tilde{Y} - \tilde{W}^T \beta) + \Sigma_{uu} \beta\}] = 0$. The empirical-likelihood

ratio function for β may be defined as

$$\mathcal{R}(\beta) = \max \left\{ \prod_{i=1}^n n p_i \mid \sum_{i=1}^n p_i \delta_i \{ \tilde{W}_i (\tilde{Y}_i - \tilde{W}_i^T \beta) + \Sigma_{uu} \beta \} = 0, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\},$$

if $m_w(z)$ and $m_y(z)$ are known. In our model setting, a modified empirical-likelihood ratio function is defined as

$$\mathcal{R}_n(\beta) = \max \left\{ \prod_{i=1}^n n p_i \mid \sum_{i=1}^n p_i \delta_i \left(\{ W_i - \hat{m}_w(Z_i) \} [Y_i - \hat{m}_y(Z_i) - \{ W_i - \hat{m}_w(Z_i) \}^T \beta] + \Sigma_{uu} \beta \right) = 0, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}. \quad (7)$$

THEOREM 4. *Under Assumptions A1–A7, $-2 \log \{ \mathcal{R}_n(\beta) \}$ converges in distribution to a chi-squared distribution with p degrees of freedom.*

Based on this result, a confidence region for β can be given by $\{ \beta : -2 \log \{ \mathcal{R}_n(\beta) \} \leq c_\alpha \}$, where c_α denotes the α quantile of the chi-squared distribution. When Σ_{uu} is unknown, we need replication data in the usual way. In the special case of $m_i \equiv 2$, as in § 2, we can then replace W_i by \bar{W}_i and $\Sigma_{uu} \beta$ by $(1/2) \hat{\Sigma}_{uu} \beta$. The resulting statistic still has the property given in Theorem 4. A justification of this last assertion can be easily obtained by using the fact that

$$E \left[\delta \left\{ \bar{W}_{ir} (\tilde{Y}_i - \bar{W}_{ir}^T \beta) + \frac{1}{2} \hat{\Sigma}_{uu} \beta \right\} \right] = 0,$$

where $\bar{W}_{ir} = \bar{W}_i - m_w(Z_i)$.

We used software R in our numerical work described below. We developed our code using E1.s, a function written by A. B. Owen, for implementing the proposed empirical-likelihood method.

5. A SIMULATION STUDY

To evaluate the performance of the proposed methods, we conducted a small-scale simulation experiment. We generated $n = 100$ and $n = 500$ observations from model (1), assuming that $Y|X, Z \sim N\{\beta_0 + \beta_1 X + \nu(Z), \sigma^2(X, Z)\}$ and that the probability of Y being observed equals $\text{pr}(\delta = 1|Y, X, Z) = \Phi\{\alpha_0 + \alpha_1 X + \nu_1(Z)\}$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. We also assumed that the measurement error follows $W = X + U$, where $U \sim N(0, 0.04)$. In our simulations, two replications for each X were obtained. We set $\alpha_0 = \beta_0 = 0$, $\beta_1 = 1$, $\alpha_1 = 2$, $X \sim \text{Un}(0, 1)$, $Z \sim \text{Un}(0, 1)$, independent of X , and $\nu(z) = 4\{\exp(-3.25z) - 4\exp(-6.5z) + 3\exp(-9.75z)\}$. We considered the following four cases.

Case 1. Here $\nu_1(z) = 0.75z$ and $\sigma^2(x, z) = 0.25$.

Case 2. Here $\nu_1(z) = \sin(z^2)$ and $\sigma^2(x, z) = 0.25$.

Case 3. Here $\nu_1(z) = 0.75z$ and $\sigma^2(x, z) = 0.1\{\sin^2(2\pi x^3) + 0.5z + 0.3\}$. This case illustrated the effect of heteroscedastic error on the estimators and confidence intervals.

Case 4. Here $\nu_1(z) = 0.75z$ and the error ε follows $0.25^2(\chi_2^2 - 2)$, where χ_2^2 is a chi-squared variable with 2 degrees of freedom. This case illustrated the effect of asymmetric error on the estimators and confidence intervals.

The reliability ratio, defined as $\text{var}(X)/\{\text{var}(X) + \text{var}(U)/2\}$, is 0.806 for all the 4 cases. In our nonparametric estimation procedure, we selected bandwidths as in Remark 2. We used the quartic kernel, $K(u) = 15/16(1 - u^2)^2 I_{(|u| \leq 1)}$. We generated 1000 datasets in each of the four cases. In each case, approximately 35% of the Y 's are missing. To estimate the variance of U , we generated replicate samples of W . We computed the naive and correction-for-attenuation estimates of the parametric components, and their asymptotic and empirical-likelihood-based confidence intervals.

The results are given in Table 1(a). The column 'Estimate' gives the average of 1000 estimated coefficients based on the naive and our proposed methods and the column ' R^2 ' gives the coefficients of determination in the (Y, W, Z) naive and (Y, X, Z) analyses; for the latter, R^2 can be calculated directly when X is observable, as is the case in simulations, or can be estimated by $1 - \text{RSS}/S_{yy}$, with $\text{RSS} = \sum_{i=1}^n \delta_i \{Y_i - W_i \hat{\beta}_{n,2} - \hat{v}(Z_i)\}^2 - \hat{\beta}_{n,2}^2 \sigma_{uu}^2$ and $S_{yy} = \sum_{i=1}^n \delta_i (Y_i - \bar{Y})^2$. Our numerical experience indicates that the values of these two R^2 's are generally similar to each other. We present the estimated values of R^2 here and in our data analysis later. The columns 'midpoint' and 'length' give the average midpoint and length of the confidence intervals. The column 'CI(ME)' gives the confidence intervals using the empirical likelihood and normal approximation methods when the measurement errors are accounted for. The lower and upper values are the averages of 1000 simulated corresponding lower and upper values. The column 'Coverage (ME)' gives the corresponding coverage probabilities of the 1000 datasets.

The results are generally in accord with the theory. The impact of the measurement errors on the estimates is substantial. When measurement errors are ignored, the estimators are significantly biased and attenuate to zero. For moderate sample size, the empirical-likelihood-based confidence intervals appear to be superior to those based on the normal approximation. The improvement is greater when the error is nonnormal or its variance is not constant.

At the referee's request, we repeated the procedure above except that X and Z were correlated. We used $X = Z + e$ with $e \sim N(0, 0.06)$, so that the correlation coefficient of X and Z is 0.76, and the reliability ratio is about 0.877. The results are presented in Table 1(b). We can draw the same general conclusion as that from the case where X and Z are independent. The main difference is that the confidence intervals are generally much wider than in the case of independent X and Z , as expected.

6. ANALYSIS OF A DATASET FROM AN AIDS STUDY

In this section, we present an illustrative analysis of the paediatric AIDS clinical trial group PACTG 338 study. One of the purposes of this study is to investigate the effectiveness of antiretroviral medicines, and to see how increasing CD4 cell counts decrease the amount of HIV in the blood, the HIV viral load. We are interested in understanding the pathogenesis of HIV infection and in evaluation of antiretroviral therapies by characterizing the relationship between viral load and CD4 cell counts. Our preliminary investigations suggested that viral load depends linearly on CD4 cell count but nonlinearly on treatment time; see Liang et al. (2004) for a related discussion. We therefore model the relationship between viral load and CD4 cell counts by model (1). Let Y_{ij} be the viral load and let X_{ij} be the CD4 cell count for subject i at treatment time Z_{ij} . The X_{ij} are measured with error (Liang et al., 2003). Here, we treat Y_{ij} as measured without error except for being partially missing. The model we used is

Table 1. *Simulation study. Point estimates for β_1 , with true value 1, together with the coefficients of determination (R^2) and the 95% confidence intervals based on the empirical-likelihood (EL) and normal approximation (Norm) methods. Also displayed are the averages of the lower and upper endpoints of the confidence intervals, their average midpoints, their average lengths and the associated coverage probabilities for the simulated data (a) when X and Z are independent, and (b) when X and Z are strongly correlated*

n	Case	Estimate		R ²	X observed	EL	Confidence interval		Midpoint		Length		Coverage	
		Naive	ME				Naive	X	EL	Norm	EL	Norm	EL	Norm
(a) X and Z are independent														
100	1	0.649	1.027	0.464	0.525	(0.511, 1.624)	(0.438, 1.615)	1.072	1.027	1.114	1.177	93.7	94.3	
	2	0.643	1.021	0.467	0.525	(0.500, 1.627)	(0.438, 1.605)	0.969	1.021	1.126	1.167	94.0	94.7	
	3	0.655	1.042	0.861	0.915	(0.806, 1.318)	(0.740, 1.324)	1.082	1.042	0.522	0.544	95.7	92.7	
	4	0.657	1.048	0.473	0.532	(0.505, 1.676)	(0.364, 1.733)	1.090	1.048	1.171	1.369	96.3	94.3	
500	1	0.665	1.007	0.446	0.502	(0.793, 1.221)	(0.793, 1.221)	0.981	1.007	0.427	0.428	94.7	94.8	
	2	0.658	0.997	0.445	0.501	(0.786, 1.217)	(0.777, 1.218)	1.006	0.997	0.432	0.441	95.8	94.5	
	3	0.660	1.000	0.858	0.917	(0.913, 1.071)	(0.916, 1.083)	1.007	1.000	0.158	0.167	93.7	95.8	
	4	0.663	1.003	0.446	0.502	(0.788, 1.144)	(0.785, 1.221)	0.990	1.003	0.363	0.436	94.8	94.9	
(b) X and Z are strongly correlated														
100	1	0.569	1.082	0.65	0.688	(0.306, 2.016)	(-0.001, 2.165)	1.061	1.082	1.71	2.166	92.7	97.0	
	2	0.589	1.094	0.641	0.677	(0.357, 1.916)	(0.256, 1.931)	0.971	1.094	1.569	1.675	96.3	95.0	
	3	0.568	1.088	0.879	0.922	(0.530, 1.907)	(0.044, 2.132)	1.019	1.088	1.377	2.088	94.6	96.7	
	4	0.564	0.983	0.645	0.674	(0.261, 2.082)	(-0.296, 2.445)	1.172	0.983	1.821	2.741	94.7	97.7	
500	1	0.586	1.013	0.631	0.665	(0.725, 1.227)	(0.720, 1.306)	1.026	1.013	0.522	0.586	94.8	94.2	
	2	0.582	1.005	0.631	0.664	(0.716, 1.221)	(0.715, 1.296)	0.981	1.005	0.544	0.581	95.6	94.5	
	3	0.582	0.991	0.912	0.942	(0.897, 1.062)	(0.892, 1.125)	0.993	0.991	0.168	0.233	98.6	95.3	
	4	0.586	1.011	0.630	0.664	(0.728, 1.228)	(0.727, 1.295)	1.023	1.011	0.510	0.568	96.4	94.6	

$$Y_{ij} = X_{ij}\beta + v(Z_{ij}) + \varepsilon_{ij}, \quad W_{ij} = X_{ij} + U_{ij},$$

where W_{ij} are the observed CD4 cell counts. The first part of this model was applied by Zeger & Diggle (1994) to investigate the relationship between the CD4 cell count, Y_{ij} , and time, Z_{ij} , and other covariates, X_{ij} . If there is no correlation, this longitudinal model reduces to model (1).

The PACTG 338 study consists of 297 children, who were clinically stable and who had not had prior treatment with protease inhibitors. They were subjected to a regimen containing a 2- or 3-drug protease inhibitor containing regimen, i.e., ritonavir plus 1 or 2 nucleoside analogues, or to a dual nucleoside analogue regimen. There were 2287 observations, among which 404 (17.6%) viral load RNA values were missing. The ranges of viral load (\log_{10}) and CD4 cell counts are (2.6, 6.21) and (51, 3284); the mean and median of CD4 cell counts are 824.55 and 746; the mean and median of HIV RNA levels are 3.518 and 3.346; and the specimens were obtained on weeks 0, 4, 8, 12, 24, 36, 48, 60, 72 and 84. See Nachman et al. (2000) for a detailed explanation of the study. The CD4 cell counts are used to follow response to HIV medications, as a measure of adherence to treatment and most importantly to guide decisions regarding opportunistic infection prophylaxis. Some patients may fail to go to clinical trial centres for an HIV viral load measurement when they feel that their immunity is strong enough or too weak. Therefore, the assumption that the missing RNA levels depend on true CD4 cell counts and not on measured counts and treatment time appears to be at least reasonable.

We ignored the correlation structure when computing the estimates, using the so-called working independence assumption. As pointed out in equation (2) of Lin & Carroll (2001), working independence has some model-robustness advantages over estimation methods that account for correlation, with a corresponding loss of efficiency. To reduce the marked skewness of CD4 cell counts, and make treatment times equally spaced, we take log-transformations of both variables. We used the same kernel function as in the simulation study in § 5, and obtained a bandwidth of $h = 0.124$ in the same manner as described there. We assumed that the measurement errors U_{ij} were independent and normally distributed with mean zero and variance σ_u^2 . In the absence of validation or replication data, as in Lin & Carroll (2000), we conducted a sensitivity analysis by taking σ_u^2 to be one quarter and one half of the variance of W .

We applied the methods proposed in §§2 and 4, assuming $\sigma_u^2 = 0$, which naively ignores measurement error, $\sigma_u^2 = 0.068$ and $\sigma_u^2 = 0.135$. For β , we give estimated values, along with the normal approximation, bootstrap and empirical-likelihood confidence intervals in the full-data part of Table 2. The bootstrap intervals were based on 200 replications. The R^2 's shown in the table are calculated as was described in §5 for the coefficient of determination in the (Y, X, Z) regression analysis if X were observed. The estimated values of β corresponding to $\sigma_u^2 = 0.068$ and $\sigma_u^2 = 0.135$ increased by 31.2% and 48.8% in absolute value, respectively, compared to the naive estimate, and the confidence intervals were widened accordingly. As expected, when the possibility of measurement errors was taken into account, we found a somewhat stronger negative association between viral load and CD4 cell counts. Whether or not there is correlation between within-subjects observations, the bootstrap method generally produces correct confidence intervals, while the other two approaches can lead to incorrect, typically too optimistic, confidence intervals. However, in this special case, any correlation effect appears to be minimal since all three methods produced similar confidence intervals with the empirical-likelihood intervals being

Table 2. AIDS study. Estimates of β , with the 95% confidence intervals (CI) based on the normal approximation (Norm), empirical likelihood (EL) and bootstrap methods. Also given are the estimated R^2 's of the coefficient of determination for the (Y, X, Z) regression analysis if X were observed.

	$\sigma_u^2 = 0$	$\sigma_u^2 = 0.068$	$\sigma_u^2 = 0.135$
Full data			
Estimate of β	-0.125	-0.164	-0.186
R^2	0.331	0.422	0.436
CI(Norm)	(-0.144, -0.106)	(-0.207, -0.121)	(-0.244, -0.128)
CI(EL)	(-0.132, -0.109)	(-0.183, -0.128)	(-0.221, -0.130)
CI(Boot)	(-0.143, -0.107)	(-0.205, -0.123)	(-0.245, -0.127)
Independent data			
Estimate of β	-0.129	-0.159	-0.212
R^2	0.138	0.162	0.214
CI(Norm)	(-0.190, -0.068)	(-0.236, -0.082)	(-0.320, -0.104)
CI(EL)	(-0.181, -0.068)	(-0.226, -0.076)	(-0.310, -0.114)

slightly shorter, as in the simulation studies, while the normal and bootstrap intervals are virtually identical.

Since our approximate confidence intervals assumed working independence, which may well not be the case in the current data analysis, we also demonstrate how they perform for data that actually do satisfy the assumption of having independent observations. For this purpose, we took one observation from each child at random, and repeated the exercise 1000 times. The results are presented in the independent-data part of Table 2. All the entries were computed in the same way as for the full data except that they are the averages of 1000 estimates. Again, the two methods produced similar results with the empirical-likelihood intervals being slightly shorter.

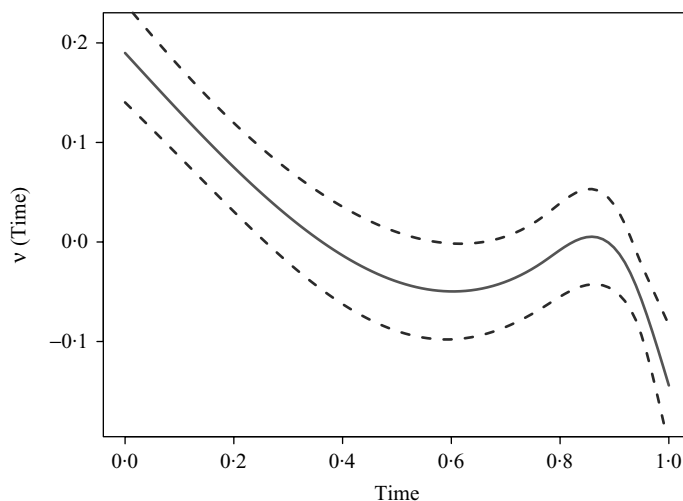


Fig. 1. AIDS study. The solid curve represents the estimated values of $v(z)$ based on the complete observations, and the dotted lines indicate the confidence intervals.

The curve of the estimated nonparametric function of treatment time and the corresponding confidence bands for the case of $\sigma_u^2 = 0$ are shown in Fig. 1. The results for the other cases of $\sigma_u^2 = 0.068$ and $\sigma_u^2 = 0.135$ are similar and are therefore not shown. The confidence bands were obtained by 200 bootstrap replications, in which patients were resampled. The plot indicates that the viral load RNA levels rapidly decrease after initial antiviral treatment, then become flat and even rebound a little bit and finally decrease rapidly.

7. DISCUSSION

The point estimation methods and normal limit distributions are readily extended to longitudinal and repeated measures contexts, if one uses working independence, i.e. ignores the correlation structure when computing the estimator but uses it in computing asymptotic covariance matrices. However, as was suggested in our data analysis, how to employ the empirical-likelihood procedure for correlated data appears to be a difficult issue, and is currently under our investigation.

The proposed estimators are based on the observed data, but exclude the observed covariates (W, Z) when Y is missing except that all the W 's are used to estimate Σ_{uu} when its estimation is desired. Although we have not derived the efficiency bound for the estimator of β , we conjecture that little gain, if any, can be obtained if we include those observations (W, Z) associated with missing Y 's; see Bickel et al. (1993, p. 146) for a related result.

ACKNOWLEDGEMENT

The authors thank the editor and a referee for their helpful suggestions and constructive comments. Liang's research was partially supported by three grants from the National Institute of Allergy and Infectious Diseases. Wang and Carroll's research was partially 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. The work of Raymond Carroll occurred during a visit to the Centre of Excellence for Mathematics and Statistics of Complex Systems at the Australian National University, whose support is gratefully acknowledged.

APPENDIX

Technical details

Assumption A1. The matrix $E\left\{\pi(X, Z)\tilde{X}\tilde{X}^T\right\}$ is positive-definite, $E(\varepsilon|X, Z) = 0$, and $E(|\varepsilon|^3|X, Z) < \infty$.

Assumption A2. The bandwidths used in estimating $m_w(z)$ and $m_y(z)$ are of order $n^{-1/5}$.

Assumption A3. The function $K(\cdot)$ is a bounded symmetric density function with compact support and satisfies $\int K(u)du = 1$, $\int K(u)udu = 0$ and $\int u^2K(u)du = 1$.

Assumption A4. The density function of Z , $f_Z(z)$, is bounded away from zero and has bounded continuous second derivatives.

Assumption A5. The functions $m_y(z)$ and $m_w(z)$ have bounded and continuous second derivatives.

Assumption A6. The probability function $\pi(x, z)$ is bounded away from zero on the support of (X, Z) , and has bounded continuous second partial derivatives.

Assumption A7. The random variable U satisfies $E(\|U\|^3) < \infty$.

We first point out the following facts, which can easily be shown by Assumptions A2–A5:

$$\hat{m}_w(z) - m_w(z) = o_p(n^{-1/4}), \quad \hat{m}_y(z) - m_y(z) = o_p(n^{-1/4}). \quad (\text{A1})$$

In the rest of the Appendix, we prove Theorem 4. We first present a lemma, whose proof can be found in Liang et al. (1999).

LEMMA A1. *Assume that random variables a_i and b_i satisfy $Ea_i = 0$ and $\|b_i\| = o_p(n^{-1/4})$. Then*

$$\sum_{i=1}^n a_i b_i \xi_i = o_p(n^{1/2}),$$

where ξ_i are independent variables with zero conditional mean and finite variance.

Proof of Theorem 4. Let

$$\Omega_i = (\{W_i - \hat{m}_w(Z_i)\}[Y_i - \hat{m}_y(Z_i) - \{W_i - \hat{m}_w(Z_i)\}^T \beta] + \Sigma_{uu} \beta) \delta_i.$$

A standard simplification as in Owen (2001, p. 61) yields that

$$p_i = \frac{1}{n(1 + a^T \Omega_i)}, \quad (\text{A2})$$

for $i = 1, \dots, n$, where $a = (a_1, \dots, a_p)^T$ is the solution of the equation

$$n^{-1} \sum_{i=1}^n \frac{\Omega_i}{1 + a^T \Omega_i} = 0. \quad (\text{A3})$$

Mimicking the proof Theorem 3.2 of Owen (2001), we have

$$\|a\| = O_p(n^{-1/2}). \quad (\text{A4})$$

On the other hand, based on the assumptions, Theorem 1 and the strong law of large numbers, we have

$$\max_{1 \leq i \leq n} \|\Omega_i\| = o_p(n^{1/2}). \quad (\text{A5})$$

Note that

$$n^{-1} \sum_{i=1}^n \frac{\Omega_i}{1 + a^T \Omega_i} = n^{-1} \sum_{i=1}^n \Omega_i (1 - a^T \Omega_i) + n^{-1} \sum_{i=1}^n \frac{(a^T \Omega_i)^2 \Omega_i}{1 + a^T \Omega_i}.$$

The second term is $o_p(n^{-1/2})$ since $|a^T \Omega_i| = o_p(1)$ and

$$\sum_{i=1}^n (a^T \Omega_i)^2 \Omega_i \leq \|a\| \max_{1 \leq i \leq n} |a^T \Omega_i| \sum_{i=1}^n \|\Omega_i\|^2 = O_p(n^{-1/2}) o_p(1) O_p(n) = o_p(n^{1/2}).$$

It then follows from (A3) that

$$a = \left(\sum_{i=1}^n \Omega_i \Omega_i^T \right)^{-1} \sum_{i=1}^n \Omega_i + o_p(n^{-1/2}). \quad (\text{A6})$$

A similar argument using $\sum_{i=1}^n p_i = 1$ yields that

$$0 = n^{-1} \sum_{i=1}^n \frac{a^T \Omega_i}{1 + a^T \Omega_i} = n^{-1} \sum_{i=1}^n a^T \Omega_i - n^{-1} \sum_{i=1}^n (a^T \Omega_i)^2 + o_p(n^{-1}).$$

Therefore, we have

$$\sum_{i=1}^n a^T \Omega_i = \sum_{i=1}^n (a^T \Omega_i)^2 + o_p(1). \quad (\text{A7})$$

Consider $\mathcal{R}_n(\beta)$. Using a Taylor expansion of $\log(1+x)$ in x , we have

$$\begin{aligned} -\log\{\mathcal{R}_n(\beta)\} &= \sum_{i=1}^n \log(1 + a^T \Omega_i) \\ &= \sum_{i=1}^n \{a^T \Omega_i - (1/2)(a^T \Omega_i)^2\} + Q_n. \end{aligned}$$

The remainder term Q_n is bounded by

$$\|a\|^2 \max_{1 \leq i \leq n} |a^T \Omega_i| \sum_{i=1}^n \|\Omega_i\|^2 = O_p(n^{-1}) o_p(1) O_p(n) = o_p(1).$$

Using (A7) and (A6), we have

$$\begin{aligned} -2 \log\{\mathcal{R}_n(\beta)\} &= \sum_{i=1}^n a^T \Omega_i \Omega_i^T a + o_p(1) \\ &= \left(n^{-1/2} \sum_{i=1}^n \Omega_i \right)^T \left(n^{-1} \sum_{i=1}^n \Omega_i \Omega_i^T \right)^{-1} \left(n^{-1/2} \sum_{i=1}^n \Omega_i \right) + o_p(1). \end{aligned}$$

Write $\tilde{\Omega}_i = (\{W_i - m_w(Z_i)\}[Y_i - m_y(Z_i) - \{W_i - m_w(Z_i)\}^T \beta] + \Sigma_{uu} \beta) \delta_i$. Then $\tilde{\Omega}_i - \Omega_i$ can be expressed as

$$\begin{aligned} &\tilde{W}_i [\{\hat{m}_y(Z_i) - m_y(Z_i)\} - \{\hat{m}_w(Z_i) - m_w(Z_i)\}^T \beta] \delta_i \\ &\quad - \{\hat{m}_w(Z_i) - m_w(Z_i)\} \{\{\hat{m}_y(Z_i) - m_y(Z_i)\} - \{\hat{m}_w(Z_i) - m_w(Z_i)\}^T \beta\} \delta_i \\ &\quad + \{\hat{m}_w(Z_i) - m_w(Z_i)\} (\tilde{Y}_i - \tilde{W}_i^T \beta) \delta_i. \end{aligned}$$

It follows from (A1) that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \{\hat{m}_w(Z_i) - m_w(Z_i)\} \{\{\hat{m}_y(Z_i) - m_y(Z_i)\} + \{\hat{m}_w(Z_i) - m_w(Z_i)\}^T \beta\} \delta_i = o_p(1).$$

On the other hand, Lemma A1 implies that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{W}_i [\{\hat{m}_y(Z_i) - m_y(Z_i)\} + \{\hat{m}_w(Z_i) - m_w(Z_i)\}^T \beta] \delta_i &= o_p(1), \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\hat{m}_w(Z_i) - m_w(Z_i)\} (\tilde{Y}_i - \tilde{W}_i^T \beta) \delta_i &= o_p(1). \end{aligned}$$

These results imply that $n^{-1/2} \sum_{i=1}^n \Omega_i$ and $n^{-1/2} \sum_{i=1}^n \tilde{\Omega}_i$ asymptotically have the same normal distribution, and $n^{-1} \sum_{i=1}^n \Omega_i \Omega_i^T$ and $n^{-1} \sum_{i=1}^n \tilde{\Omega}_i \tilde{\Omega}_i^T$ have the same limiting value. The proof is thus complete.

REFERENCES

- BICKEL, P. J., KLAASSEN, C. J., RITOV, Y. & WELLNER, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
- CARROLL, R. J., RUPPERT, D. & STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. New York: Chapman and Hall.
- CHEN, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *J. Mult. Anal.* **49**, 24–40.
- CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Am. Statist. Assoc.* **89**, 81–7.
- ENGLER, R. F., GRANGER, C. W. J., RICE, J. & WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* **81**, 310–20.
- HÄRDLE, W., LIANG, H. & GAO, J. (2000). *Partially Linear Models*. Heidelberg: Springer Physica-Verlag.
- LIANG, H., HÄRDLE, W. & CARROLL, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* **27**, 1519–35.
- LIANG, H., WANG, S., ROBINS, J. M. & CARROLL, R. J. (2004). Estimation in partially linear models with missing covariates. *J. Am. Statist. Assoc.* **99**, 357–67.
- LIANG, H., WU, H. L. & CARROLL, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effect varying-coefficient semiparametric models with measurement error. *Biostatistics* **4**, 297–312.
- LIN, X. & CARROLL, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.* **95**, 520–34.
- LIN, X. & CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.* **96**, 1045–56.
- NACHMAN, S. A., STANLEY, K., YOGEV, R. et al. (2000). Nucleoside analogs plus zidovudine in stable antiretroviral therapy experienced HIV infected children: a randomized controlled trial. *J. Am. Med. Assoc.* **283**, 492–8.
- OPSOMER, J. D. & RUPPERT, D. (1999). A root- n consistent backfitting estimator for semiparametric additive modelling. *J. Comp. Graph. Statist.* **8**, 715–32.
- OWEN, A. B. (2000). *Empirical Likelihood*. London: Chapman and Hall/CRC.
- QIN, J. (1994). Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Ann. Inst. Statist. Math.* **46**, 117–26.
- QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.
- ROBINSON, P. M. (1988). Root- n -consistent semiparametric regression. *Econometrica* **56**, 931–54.
- RUPPERT, D., SHEATHER, S. J. & WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–70.
- SEVERINI, T. A. & STANISWALIS, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* **89**, 501–11.
- SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc. B* **50**, 413–36.
- WANG, Q. H., LINTON, O. & HÄRDLE, W. (2004). Semiparametric regression analysis with missing response at random. *J. Am. Statist. Assoc.* **99**, 334–45.
- ZEGER, S. L. & DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

[Received October 2005. Revised June 2006]