

Estimation in Partially Linear Models With Missing Covariates

Hua LIANG, Suojin WANG, James M. ROBINS, and Raymond J. CARROLL

The partially linear model $Y = X^T\beta + \nu(Z) + \epsilon$ has been studied extensively when data are completely observed. In this article, we consider the case where the covariate X is sometimes missing, with missingness probability π depending on (Y, Z) . New methods are developed for estimating β and $\nu(\cdot)$. Our methods are shown to outperform asymptotically methods based only on the complete data. Asymptotic efficiency is discussed, and the semiparametric efficient score function is derived. Justification of the use of the nonparametric bootstrap in this context is sketched. The proposed estimators are extended to a working independence analysis of longitudinal/clustering data and applied to analyze an AIDS clinical trial dataset. The results of a simulation experiment are also given to illustrate our approach.

KEY WORDS: Bootstrap; Clustered data; Efficient score; Estimating equation; Local linear regression; Longitudinal data; Missing covariates; Nonparametric regression; Semiparametric estimation.

1. INTRODUCTION

Perhaps the most common model used in analyzing observational studies of the causal effect of a possibly multivariate treatment or exposure $X^T = (X_1, \dots, X_p)$ on a continuous response Y when data are available on one or more continuous pretreatment confounding variables Z is the partial linear model

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

where β is an unknown parameter, $\nu(\cdot)$ is a smooth unknown function of Z , $E(\epsilon|X, Z) = 0$, and the joint distribution of the regressors (X, Z) is left completely unspecified. Robins, Mark, and Newey (1992) proved that this model arises whenever we assume (a) no unmeasured confounders (i.e., ignorability of treatment X within levels of Z) and (b) a constant additive effect of treatment X on the mean of Y . In particular, given assumption (a), this model is guaranteed to be correctly specified under the causal null hypothesis of no effect of treatment X on Y , because the causal null hypothesis implies (1) with $\beta = 0$. Thus, under (a), an asymptotically correct $1 - \alpha$ confidence interval for β in model (1) provides an asymptotic distribution-free α -level test of the causal null hypothesis of no exposure effect. Tests of $\beta = 0$ based on lower-dimensional models that impose parametric functional forms on either $\nu(Z)$ and/or the density of $X|Z$ do not provide asymptotically distribution-free tests of the causal null hypothesis under (a). Even when (a) cannot be assumed to hold, model (1) remains useful and robust, because a large sample test of $\beta = 0$ under model (1) remains an asymptotic distribution-free test of the important associational hypothesis that (a) Y is mean independent of X given Z and that (b) Y is conditionally independent of X given Z .

For these reasons, estimation of β in model (1) has been the subject of considerable study (see Härdle, Liang, and Gao

2000 for a summary). Our contribution in this article is to study model (1) when data on X are not fully observed for some study subjects, whether by design (as in two-stage studies) or by happenstance. The problem of missing exposure variables in regression has been treated in great detail by Robins, Rotnitzky, and Zhao (1994); however, these authors assumed a parametric functional form for $\nu(Z)$. For the aforementioned reasons, it is clearly important to relax, as we do in this article, the assumption that the functional form of $\nu(Z)$ is known. As was done by Robins et al. (1994), we allow the missingness probabilities to depend on both Y and Z , but not on the unobserved value of X . Our results include both the case where the missingness probabilities are known (as in a designed two-stage study) and the case where they are unknown. Our results build on the work of Wang, Wang, Gutierrez, and Carroll (1998), who considered the nonparametric problem (no X) with missing data (see also Cheng 1990, 1994; Cheng and Chu 1996).

The article organized as follows. In Section 2 we define the missing-data mechanism for the problem and define our methods of estimation. In Section 3 we describe our asymptotic results. Not only do we derive the asymptotic distribution of our estimators of β , but we also describe three extensions. First, we compare our methods with methods that use only the complete data with appropriate Horvitz-Thompson (HT) weighting, and show that our methods are asymptotically more efficient. Second, along with deriving analytic standard error estimates, we also justify the use of the nonparametric bootstrap in this context. Finally, we show that our methods can be extended to longitudinal and clustered data when working independence is used as the method of estimation, thus extending the work on nonparametric regression for correlated data using working independence (Zeger and Diggle 1994; Hoover, Rice, Wu, and Yang 1998; Fan and Zhang 2000; Lin and Ying 2001) to the missing-data context.

In Section 4 we study asymptotic efficiency for estimation of β in model (1). Here we derive the semiparametric efficient score function and the semiparametric information bound. The semiparametric efficient score function is a solution to a complex integral equation, but in a special case we are able to derive the score function explicitly and compare the result with our methods. In Section 5 we report the results of a small simulation

Hua Liang is Assistant Member, St. Jude Children's Research Hospital, Memphis, TN 38105 (E-mail: hua.liang@stjude.org). Suojin Wang (E-mail: sjwang@stat.tamu.edu) is Professor and Raymond J. Carroll (E-mail: carroll@stat.tamu.edu) is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. The authors thank the editor, an associate editor, and two referees for their great patience, constructive comments, and useful suggestions. Liang's research was supported in part by the National Institute of Allergy and Infectious Diseases (U01 AI38855) and the American Lebanese Syrian Associated Charities. Wang and Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). Robins' research was supported by the National Institutes of Health (AI-32475).

study, and in Section 6 we present the results of the analysis of an AIDS study. We provide concluding remarks in Section 7, and give proofs in the Appendix. Our asymptotic work uses the general asymptotic theory for semiparametric models developed by Newey (1994) and Robins et al. (1994).

2. THE MODEL AND ESTIMATORS

Let $\delta = 1$ if X is observed and $\delta = 0$ otherwise. Assume that the X 's are missing at random (MAR) in the sense that

$$\pi(Y_i, Z_i) = P(\delta_i = 1 | X_i, Z_i, Y_i) = P(\delta_i = 1 | Z_i, Y_i). \quad (2)$$

In this article, we first assume that the missing-data probability, $\pi(Y, Z)$, is known. Later we show that its estimation with an error of order $o_p(n^{-1/4})$ can be undertaken without affecting the asymptotic properties of our proposed estimate of β . Moreover, we first assume that $(Y_i, X_i, Z_i, \delta_i)$, $i = 1, \dots, n$, are independent and identically distributed (iid). Then we extend the case to the longitudinal/clustered data setting in Section 3.4.

For general parametric models $E(Y|X) = g(X, \theta)$, Robins et al. (1994) proposed the estimating equation

$$\Psi(\cdot, \theta) = \frac{\delta}{\pi} p(X)\{Y - g(X, \theta)\} - \frac{\delta - \pi}{\pi} \phi(Y, \theta) \quad (3)$$

for some user-supplied function $p(\cdot)$, where $\phi(y, \theta)$ is a general function. When there is no Z , the optimal choice of $\phi(\cdot)$ is $\phi(Y) = E\{p(X)(Y - X^T\beta)|Y\}$. If $\beta = 0$ (no X) in (1), then the topic becomes a nonparametric problem, for which Wang et al. (1998) developed HT weighted local linear kernel methods. Our methods can be looked on as combining the parametric procedures of Robins et al. with the nonparametric procedure of Wang et al.

Here is the intuition behind our method. If there were no missing data, then let $\hat{v}(Z, \beta)$ and $\hat{m}(Z)$ be nonparametric regressions of $Y - X^T\beta$ and X on Z . Then under normality and homoscedascity, the semiparametric optimal score function for β is $\{X - \hat{m}(Z)\}\{Y - X^T\beta - \hat{v}(Z, \beta)\}$. Effectively, what we do is apply (3) to this score function to compensate for the missing data, with $p(X)$ being $X - \hat{m}(Z)$ and $g(X, \theta)$ being $X^T\beta + \hat{v}(Z, \beta)$, the result being (5).

To understand our methods, we first define the HT weighted local linear kernel method. Let $K(\cdot)$ be a symmetric density function and let h be a suitable bandwidth. Then for any response $q(Y, X)$, the local estimate at z_0 is the intercept in the regression of $q(Y, X)$ on $(Z - z_0)/h$ with weights $K_h(Z - z_0)\delta/\pi(Y, Z)$, where $K_h(v) = h^{-1}K(v/h)$, that is, the solution α_0 in the equation

$$0 = \sum_{i=1}^n K_h(Z_i - z_0) \frac{\delta_i}{\pi(Y_i, Z_i)} \left(\frac{1}{(Z_i - z_0)/h} \right) \times \left\{ q(Y_i, X_i) - \alpha_0 - \alpha_1 \left(\frac{Z_i - z_0}{h} \right) \right\}. \quad (4)$$

Motivated by (4), we define our estimator as follows. The procedure comprises four stages, and our method of estimation is simple and noniterative.

Step 1. For any β , let $\hat{v}(\cdot, \beta, \pi)$ be a weighted nonparametric regression of $Y - X^T\beta$ on Z using the HT local linear kernel method, that is, (4) with $q(Y, X) = Y - X^T\beta$.

Step 2. Form the corresponding HT local linear kernel regression function $\hat{m}(z, \pi)$ for estimating $m(z) = E(X|z)$ by regressing X on Z , that is, (4) with $q(Y, X) = X$.

Step 3. Let $\hat{\phi}(Y, Z, \beta, \hat{v}, \hat{m})$ be a function of Y and Z that is linear in β , specifically, an estimate of $E[\{X - E(X|Z)\}\{Y - v(Z, \beta) - X^T\beta\}|Y, Z]$ [denote $\phi(Y, Z)$]; see the statement following Step 4.

Step 4. Solve for β in the equation

$$0 = \sum_{i=1}^n \{X_i - \hat{m}(Z_i, \pi)\} \times \{Y_i - \hat{v}(Z_i, \beta, \pi) - X_i^T\beta\} \frac{\delta_i}{\pi(Y_i, Z_i)} - \sum_{i=1}^n \hat{\phi}(Y_i, Z_i, \beta, \hat{v}, \hat{m}) \frac{\delta_i - \pi(Y_i, Z_i)}{\pi(Y_i, Z_i)}. \quad (5)$$

Because (5) is linear in β , it can be solved without iteration. We call the solution $\hat{\beta}_{\text{all}}$.

Step 3 is the only point requiring comment, because it anticipates nonparametric regression with the two "covariates" Y and Z . In particular, we let $\hat{\phi}(y, z, \beta, \hat{v}, \hat{m}) = \hat{E}(X|y, z)y - \hat{E}(X|y, z)\hat{v}(z, \beta, \pi) - \{\hat{E}(XX^T|y, z)\}\beta - \hat{m}(z, \pi)y + \hat{m}(z, \pi)\hat{v}(z, \beta, \pi) + \hat{m}(z, \pi)\{\hat{E}(X|y, z)\}^T\beta$, where $\hat{E}(X|y, z)$ and $\hat{E}(XX^T|y, z)$ are the HT bivariate local linear estimators of $E(X|y, z)$ and $E(XX^T|y, z)$, similar to the definition of $\hat{m}(z, \pi)$. For example, the HT bivariate local linear estimator of the j th element of $E(X|y, z)$ is the solution of α_0 to

$$\arg \min_{\alpha_0, \alpha_{1y}, \alpha_{1z}} \sum_{i=1}^n \{X_{ij} - \alpha_0 - \alpha_{1y}(Y_i - y) - \alpha_{1z}(Z_i - z)\}^2 K_{\lambda_1, \lambda_2}^*(Y_i - y, Z_i - z) \frac{\delta_i}{\pi(Y_i, Z_i)},$$

where $K_{\lambda_1, \lambda_2}^*(\cdot, \cdot)$ is a two-dimensional density function with bandwidths λ_1 and λ_2 and X_{ij} is the j th element of X_i . Similarly, one may define the estimator of each element of $E(XX^T|y, z)$.

Remark 1. There are a host of alternative methods for the regressions in Steps 1–3, including higher degree local polynomial kernel methods, kernel methods with varying bandwidths, smoothing and regression splines, and so on. We chose local linear smoothers with fixed bandwidths because theoretical results can be derived for them. It is obvious that the same results will apply for any kernel-based method, and because splines and kernels are in some sense asymptotically equivalent (Silverman 1984), for splines as well. In addition, proposition 2.1 of Newey (1994) suggests that our asymptotic results will depend not on the form of estimation in Steps 1–3, but only on the steps themselves.

3. MAIN RESULTS

3.1 Asymptotic Results for Our Estimators

$$\hat{\beta}_{\text{all}} \text{ and } \hat{\beta}^*_{\hat{\pi}^*, \text{all}}$$

Our results can be proven under suitable regularity conditions, but in particular we assume that \hat{m} , \hat{v} , and $\hat{\phi}$ converge

uniformly at order $o_p(n^{-1/4})$. This holds for both univariate and bivariate nonparametric regression with properly (not necessarily optimally) chosen bandwidths. For uniformity, one must first assume (as in Severini and Staniswalis 1994 or Carroll, Knickerbocker, and Wang 1995) that (Y, Z) has compact support with marginal and joint densities bounded away from 0. Then, using the techniques of Mack and Silverman (1982) or Marron and Härdle (1986), one can derive the required uniformity.

In what follows, we denote $A \cdot A^T$ by $A^{\otimes 2}$, and for any random variable (vector) ζ , let $\tilde{\zeta} = \zeta - E(\zeta|Z)$ and $\hat{\zeta} = \zeta - \hat{E}(\zeta|Z)$, where $\hat{E}(\zeta|Z)$ is a local linear estimator of $E(\zeta|Z)$. For example, $\tilde{X}_i = X_i - E(X_i|Z_i)$ and $\hat{Y}_i = Y_i - \hat{E}(Y_i|Z_i)$, $\Sigma_{X|Z} = \text{cov}\{X - E(X|Z)\}$.

In addition to the requirements just stated, we need the following conditions, which are assumed to hold throughout the remainder of the article.

Assumption 1. (a) $\Sigma_{X|Z} = E(\tilde{X}\tilde{X}^T)$ is a positive-definite matrix, $E(\epsilon|X, Z) = 0$, and $E(|\epsilon|^3|X, Z) < \infty$.

(b) The bandwidths in Steps 1 and 2 are of order $n^{-1/5}$, and the bandwidths λ_1 and λ_2 for estimating $E(X|Y, Z)$ and $E(XX^T|Y, Z)$ are of order $n^{-1/6}$.

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

(d) The density function of Z , $f_Z(z)$, and the density function of (Y, Z) are bounded away from 0 and have bounded continuous second derivatives.

(e) $E(Y|Z)$, $E(X|Z)$, and $v(\cdot)$ have bounded and continuous second derivatives.

(f) $E(X|Y, Z)$ and $E(XX^T|Y, Z)$ have bounded first derivatives.

(g) The probability function $\pi(y, z) > 0$ on the support of (Y, Z) , and has a bounded continuous second derivative.

In this section we first develop the asymptotics for estimators of β with the missing probability known or unknown. We also address estimation of the covariance of the estimators of β and the benefit of estimating the missingness probability π for the HT estimator even if π is known. We first treat the case of known π and then deal with the case of unknown π .

Theorem 1. Assume that $(Y_i, X_i, Z_i, \delta_i)$, $i = 1, \dots, n$, are iid. Under Assumption 1, $\sqrt{n}(\hat{\beta}_{\text{all}} - \beta)$ is asymptotically normally distributed with mean 0 and covariance matrix $\Sigma_\beta = \Sigma_{X|Z}^{-1} C \Sigma_{X|Z}^{-1}$, where

$$C = E\left(\frac{\epsilon^2}{\pi} \tilde{X}\tilde{X}^T\right) - E\left[\frac{1-\pi}{\pi} \{E(\tilde{X}\epsilon|Y, Z)\}^{\otimes 2}\right].$$

The proof is given in the Appendix. The result corresponds to known results where there are no missing data and for parametric problems (no Z).

Remark 2. At least in theory, only rates of convergence for the bandwidths are necessary; no theory of optimal bandwidths is needed, because all bandwidths with rates of convergence specified in the following paragraph lead to the same limit distribution for estimating β .

The bandwidths in Assumption 3(b) can be constructed in a standard fashion: we used the approach of Ruppert, Sheather, and Wand (1995) to construct, say, h_{opt} . For Step 3, the bandwidths λ_1 and λ_2 are required to converge at usual nonparametric rates. We used $\lambda_1 = \lambda_2 = gn^{1/30}$, where g is the bandwidth as estimated by Ruppert et al. in the regression of X on Z . The multiplication by $n^{1/30}$ is meant to give the correct rate of convergence; because $g = O(n^{-1/5})$, then, say, $\lambda_1 = O(n^{-1/6})$, the optimal bandwidth order for bivariate kernel regression.

In our simulation study and example, we experimented with bandwidths around the selected values, and the results did not change significantly.

We now suppose that π is unknown but can be estimated. The essential conditions are that $\inf_{y,z} \pi(y, z) > 0$ and that $\sup_{y,z} |\hat{\pi}(y, z) - \pi(y, z)| = o_p(n^{-1/4})$. For a properly parameterized model with compact support for (Y, Z) , this follows automatically. For nonparametric models, two-dimensional nonparametric regression is required. As in Remark 2, the rate of convergence $\hat{\pi}(\cdot)$ for bandwidths of order $n^{-1/6}$ is $n^{-1/3} = o(n^{-1/4})$, and the convergence is uniform as described at the start of this section; again smoothness, continuity, and compactness conditions are all required.

When $\pi(Y_i, Z_i)$ is replaced by $\hat{\pi}(Y_i, Z_i)$ in (4) and (5), the estimator is called $\hat{\beta}_{\hat{\pi}, \text{all}}$. We have the following result for $\hat{\beta}_{\hat{\pi}, \text{all}}$.

Theorem 2. Under the conditions described earlier, $\hat{\beta}_{\hat{\pi}, \text{all}}$ has the same normal limit distribution as that of $\hat{\beta}_{\text{all}}$.

The proof of this result is given in the Appendix.

Checking the proof of Theorem 1, we see that Σ_β can be estimated via a standard sandwich method as follows. Let

$$\begin{aligned} \hat{\Sigma}_{X|Z} &= n^{-1} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}(Y_i, Z_i)} \{X_i - \hat{m}(Z_i, \hat{\pi})\}^{\otimes 2}; \\ \hat{C} &= n^{-1} \sum_{i=1}^n \left(\frac{\delta_i}{\hat{\pi}(Y_i, Z_i)} \{X_i - \hat{m}(Z_i, \hat{\pi})\} \hat{\epsilon}_i \right. \\ &\quad - \frac{\delta_i - \hat{\pi}(Y_i, Z_i)}{\hat{\pi}(Y_i, Z_i)} \\ &\quad \times [\hat{E}(X_i|Y_i, Z_i) - \hat{m}(Z_i, \hat{\pi})] \\ &\quad \times \{Y_i - \hat{v}(Z_i, \hat{\beta}_{\hat{\pi}, \text{all}})\} \\ &\quad - \hat{E}(X_i X_i^T|Y_i, Z_i) \hat{\beta}_{\hat{\pi}, \text{all}} \\ &\quad \left. + \hat{m}(Z_i, \hat{\pi}) \hat{E}(X_i|Y_i, Z_i) \hat{\beta}_{\hat{\pi}, \text{all}} \right)^{\otimes 2}. \end{aligned}$$

and $\hat{\Sigma}_\beta = \hat{\Sigma}_{X|Z}^{-1} \hat{C} \hat{\Sigma}_{X|Z}^{-1}$, where $\hat{\epsilon}_i = Y_i - \hat{v}(Z_i, \hat{\beta}_{\hat{\pi}, \text{all}}) - X_i^T \hat{\beta}_{\hat{\pi}, \text{all}}$. Then it is easily shown that $\hat{\Sigma}_\beta$ is a consistent estimator of Σ_β .

3.2 Complete-Data Estimators: Asymptotics and Comparisons With Our Method

It is possible to obtain a limit result when only the complete data are used, that is, when we use only the first part of (5) as our

estimating equation. In this case it may be shown that the corresponding HT estimator, say $\widehat{\beta}_{\text{part}}$, is still asymptotically normal, but the asymptotic variance of $\sqrt{n}(\widehat{\beta}_{\text{part}} - \beta)$ increases to

$$\Sigma_{\text{big}} = \Sigma_{X|Z}^{-1} E \left(\frac{\epsilon^2}{\pi} \widetilde{X} \widetilde{X}^T \right) \Sigma_{X|Z}^{-1}. \tag{6}$$

Therefore, $\widehat{\beta}_{\text{part}}$ is not as efficient as $\widehat{\beta}_{\text{all}}$ unless $E(\widetilde{X}\epsilon|Y, Z) = 0$.

Following up on this point, Robins et al. (1994) have shown that even if π is known, after properly estimating π and substituting π by its estimator in $\widehat{\beta}_{\text{part}}$, the resulting estimator, say $\widehat{\beta}_{\widehat{\pi}, \text{part}}$, will generally have a smaller covariance matrix than $\widehat{\beta}_{\text{part}}$. It will also be no less, and sometimes more, variable than $\widehat{\beta}_{\text{all}}$. For our model, a similar phenomenon occurs, as we now show. Assume that $\pi = \pi(\psi)$ follows a parametric model with parameter ψ and define

$$J_{\pi} = E \left\{ \frac{\widetilde{X}}{\pi} \left(\frac{\partial \pi}{\partial \psi} \right)^T \epsilon \right\} \left[E \left\{ \frac{1}{\pi(1-\pi)} \frac{\partial \pi}{\partial \psi} \left(\frac{\partial \pi}{\partial \psi} \right)^T \right\} \right]^{-1} \\ \times E \left\{ \frac{\widetilde{X}}{\pi} \left(\frac{\partial \pi}{\partial \psi} \right)^T \epsilon \right\}^T.$$

We show in Appendix A.3 that

$$\text{cov}^a \{ n^{-1/2} (\widehat{\beta}_{\widehat{\pi}, \text{part}} - \beta) \} \\ = \Sigma_{X|Z}^{-1} \left\{ E \left(\frac{\epsilon^2}{\pi} \widetilde{X} \widetilde{X}^T \right) - J_{\pi} \right\} \Sigma_{X|Z}^{-1}. \tag{7}$$

Here and in the sequel we denote the asymptotic covariance matrix of an estimator by $\text{cov}^a(\cdot)$. One consequence of this result is

$$\text{cov}^a \{ n^{1/2} (\widehat{\beta}_{\text{part}} - \beta) \} \geq \text{cov}^a \{ n^{1/2} (\widehat{\beta}_{\widehat{\pi}, \text{part}} - \beta) \} \\ \geq \text{cov}^a \{ n^{1/2} (\widehat{\beta}_{\text{all}} - \beta) \}, \tag{8}$$

where “ $A \geq B$ ” means that the matrix $A - B$ is semipositive definite and “ $A > B$ ” means that the matrix $A - B$ is positive definite. The first inequality is strict unless $E \left\{ \frac{\widetilde{X}\epsilon}{\pi} \left(\frac{\partial \pi}{\partial \psi} \right)^T \right\} = 0$, and the second inequality is also strict unless $J_{\pi} = E \left[\frac{1-\pi}{\pi} \times \{ E(\widetilde{X}\epsilon|Y, Z) \}^2 \right]$ from proposition 6.1 of Robins et al. (1994). The second part of (8) is generally strict. For example, if we assume that X and ψ are one dimensional, then

$$E \left\{ E(\widetilde{X}\epsilon|Y, Z) \frac{1}{\pi} \frac{\partial \pi}{\partial \psi} \right\}^2 \\ \leq E \left[\frac{1-\pi}{\pi} \{ E(\widetilde{X}\epsilon|Y, Z) \}^2 \right] E \left\{ \frac{1}{\pi(1-\pi)} \left(\frac{\partial \pi}{\partial \psi} \right)^2 \right\}$$

by the Cauchy-Schwarz inequality, and so

$$J_{\pi} \leq E \left[\frac{1-\pi}{\pi} \{ E(\widetilde{X}\epsilon|Y, Z) \}^2 \right].$$

The last inequality is strict unless $\text{var}\{E(\widetilde{X}\epsilon|Y, Z)(\partial\pi/\partial\psi)/\pi\} = 0$ and $E(\widetilde{X}\epsilon|Y, Z) = \frac{c}{1-\pi}(\partial\pi/\partial\psi)$ for some constant c , or $\partial\pi/\partial\psi = 0$.

In particular, let $Y = \beta X + \gamma Z + \epsilon$, where X, Z , and ϵ are all independent of one another and are normal(0, 1). Also, let $\pi(Y, Z)$ in (2) be $H(\psi Y)$, where $H(t) = 1/\{1 + \exp(-t)\}$. Then $\widetilde{X} = X$ and $\partial\pi/\partial\psi = Y(1 - H)H$. It follows that

$E\{\widetilde{X}\epsilon(\partial\pi/\partial\psi)/\pi\} = E\{XY\epsilon(1 - H)\}$, which is not 0 unless $\psi = 0$. This shows the first strict inequality of (8). For the second inequality, we have $E(\widetilde{X}\epsilon|Y, Z) = E(X\epsilon|Y) = \beta Y^2/(\beta^2 + \gamma^2 + 1)^2 - \beta/(\beta^2 + \gamma^2 + 1)$ by direct calculation using the properties of the multivariate normal distribution. This is a quadratic function of Y . In contrast, $c(\partial\pi/\partial\psi)/(1 - \pi) = cY/\{1 + \exp(-\psi Y)\} \neq E(\widetilde{X}\epsilon|Y, Z)$ for any constant c , which implies that the second inequality of (8) is strict.

3.3 The Bootstrap

It is intuitively clear that the bootstrap can be used to construct standard error estimates, because, as can be shown, the estimators of β are regular estimators with linear expansions. Actually proving this is quite complex, however. In Section A.4 we provide a sketch of the argument. Our simulations show that the analytic sandwich-based covariance matrix $\widehat{\Sigma}_{\beta}$ and the bootstrap covariance matrix result in nearly identical coverage probabilities.

3.4 Longitudinal/Clustered Data

Our results can be extended to marginal regression with clustered and longitudinal data in the context that the method used is working independence; that is, the correlation structure is ignored in constructing the estimates, but is used in constructing standard errors of these estimates. With no missing data, there is an extensive literature in nonparametric regression (Zeger and Diggle 1994; Hoover et al. 1998; Fan and Zhang 2000; Lin and Ying 2001) and some results in the partially linear marginal model (Lin and Carroll 2001).

We give explicit results in the case that each cluster has exactly M observations; the more general case is easy but notationally complex. Briefly, there are $i = 1, \dots, n$ clusters, and within a cluster the data are $(Y_{ik}, X_{ik}, Z_{ik}, \delta_{ik})$ for $k = 1, \dots, M$. Models (1) and (2) are assumed to hold marginally, with $(\epsilon_{i1}, \dots, \epsilon_{iM})$ having a working covariance matrix Σ with diagonals Σ_d . The method of estimation is working independence, by which we mean that we treat the data as if they were independent; for example, weight the contributions of each term in the clustered data version of (5) by the inverse of the appropriate diagonal element of Σ_d .

In Section A.5 we derive the limiting distribution of the resulting working independence estimators. The asymptotic covariance matrix is easily estimated by sandwich methods in this case. Although we have no proof, it would appear to us that the bootstrap can be justified along the same lines as the independent data case (see Secs. 3.3 and A.4).

4. ASYMPTOTIC SEMIPARAMETRIC EFFICIENCY

This section contains results on asymptotic semiparametric efficiency in the sense of Robins et al. (1994) in the iid setting given in Section 2.

4.1 General Theory

If one wishes to obtain a semiparametric efficient estimator in the presence of missing data, then it is necessary to derive the efficient score in the model, something we do in Section A.6. Here we state the main results. Let the distribution of ϵ be denoted by F_{ϵ} , and let S_{β} be the derivative of the

log-likelihood with respect to β . Consider functions of the form $H(X, Z, \beta, \nu, \pi, F_\epsilon) = H(X, Z, \cdot)$ with the property that $E\{H(X, Z, \cdot)|Z\} = 0$, where π may be completely unknown. Then all regular estimators are based on estimating functions of the form

$$\begin{aligned} \Psi(Y, X, Z, H, \cdot) &= \frac{\delta}{\pi(Y, Z)} \{Y - X^T \beta - \nu(Z)\} H(X, Z, \cdot) \\ &\quad - \frac{\delta - \pi(Y, Z)}{\pi(Y, Z)} E[\{Y - X^T \beta - \nu(Z)\} H(X, Z, \cdot) | Y, Z]. \end{aligned} \tag{9}$$

In our case, we have chosen $H(X, Z, \cdot) = X - E(X|Z)$. For a given $H(X, Z, \cdot)$, the asymptotic covariance matrix of the resulting regular estimate of β is

$$[E\{H(X, Z, \cdot)X^T\}]^{-1} \text{cov}\{\Psi(Y, X, Z, H, \cdot)\} \times [E\{H(X, Z, \cdot)X^T\}^T]^{-1}.$$

As in Robins et al. (1994), the optimal choice of $H(X, Z, \cdot)$, $H_{\text{opt}}(X, Z, \cdot)$, is the solution to a relatively complex integral equation. We now describe the optimal choice of $H(X, Z, \cdot)$. Let $\mu(X, Z) = E(\epsilon S_\beta | X, Z)$ and $K(X, Z) = E\{\{\epsilon^2/\pi(Y, Z)\} | X, Z\}$. Dropping arguments, define

$$G_{\text{opt}}(X, Z, H_{\text{opt}}) = E\{[\epsilon\{1 - \pi(Y, Z)\}/\pi(Y, Z)] \times E\{\epsilon H_{\text{opt}}(X, Z) | Y, Z\} | X, Z\}.$$

Let $P_1(X, Z) = 1/K(X, Z)$ and $P_2(X, Z, H_{\text{opt}}) = \mu(X, Z) + G_{\text{opt}}(X, Z, H_{\text{opt}})$. Then the optimal choice of $H(X, Z, \cdot)$ solves

$$H_{\text{opt}}(X, Z) = P_1(X, Z) \left(P_2(X, Z, H_{\text{opt}}) - \frac{E\{P_1(X, Z)\{P_2(X, Z, H_{\text{opt}})\} | Z\}}{E\{P_1(X, Z) | Z\}} \right). \tag{10}$$

Even in relatively simple situations, solving the integral equation for $H_{\text{opt}}(\cdot)$, and thus obtaining an efficient estimator, although possible, seems difficult. Because of this difficulty, as a practical matter we suggest that one implement our simpler estimators.

4.2 Comparisons With Our Method

The obvious question is whether our choice of $H(X, Z) = X - E(X|Z)$ causes a major loss of efficiency. Presumably it does in some cases, but study of this issue is made complex by the need to solve for the efficient $H_{\text{opt}}(\cdot)$. However, in certain circumstances our estimators ($\hat{\beta}_{\text{all}}, \hat{\beta}_{\hat{\pi}, \text{all}}$) should be reasonably efficient. For example, they are known to be semiparametric efficient when the errors are homoscedastic and no data are missing (Chamberlain 1992), so that with small to moderate amounts of missing data, they should not be too inefficient.

In the case that the errors are normal and homoscedastic, $\mu(X, Z) = X$, calculations can be done explicitly when missingness depends on Z only [i.e., $\pi(Y, Z) = \pi(Z)$] and when Y is independent of X given Z ($\beta = 0$). In this case, $K(X, Z) = 1/\pi(Z)$ and $\epsilon = Y - \nu(Z)$ is a function of (Y, Z) . Using these facts, it is relatively easy to show that $H_{\text{opt}}(\cdot)$ in (10) is $H_{\text{opt}}(X, Z) = \pi(Z)\{X - E(X|Z)\}$ and that the asymptotic covariance matrix of the semiparametric efficient score and the

asymptotic covariance matrix of our estimator (Theorem 1) are given as

$$\text{cov}_{\text{opt}} = [E\{\pi(Z) \text{cov}(X|Z)\}]^{-1}$$

and

$$\text{cov}(\hat{\beta}_{\hat{\pi}, \text{all}}) = [E\{\text{cov}(X|Z)\}]^{-1} \times E\{\text{cov}(X|Z)/\pi(Z)\} [E\{\text{cov}(X|Z)\}]^{-1}.$$

Under these conditions, it is easily seen by inspection that if the data are missing completely at random, $\pi(Y, Z) \equiv \pi$, then our estimates are semiparametric efficient. Thus our estimator should be approximately semiparametric efficient when β is small, there is little heteroscedasticity, and missingness does not depend too heavily on (Y, Z) .

Further calculations are possible in these cases if $\pi(z)$ is logistic with logits $\alpha_0 + \alpha_1 z$, and if $\text{cov}(X|Z) \equiv \text{cov}(X)$. Let $\theta = E\{\pi(Z)\}$ be the probability of nonmissing data. The asymptotic relative efficiency of the semiparametric efficient score compared with our estimator is $\theta E\{1/\pi(Z)\}$. If $Z = \text{normal}(0, 1)$, and if the "relative risk" of observing X when comparing the 20th and the 80th percentiles of Z ($-.84$ and $.84$) is $R = \exp(2 \times .84 \times \alpha_1)$, this relative risk is the odds of observing X when $Z = .84$ divided by the odds of observing X when $Z = -.84$. Given (θ, R) one can solve for (α_0, α_1) numerically. We have found that if the percentage of missing data θ is no more than 50% and the relative risk is no more than $R = 3$, then our methods have relative efficiency greater than 95% compared to the semiparametric efficient estimator. Even when $R = 5$ and $\theta = .5$, the relative efficiency still is 80%.

Although these results are encouraging, one can presumably construct situations in which the efficiency of our methods is low, for example, when much of the data are missing and the relative risk becomes large. Nonetheless, these simple calculations suggest that in many cases our methods will maintain a reasonable amount of efficiency compared to what could be obtained by semiparametric efficient methods, were routine algorithms for the latter available.

5. A SIMULATION STUDY

In this section we describe simulation results to investigate the finite-sample behavior of $\hat{\beta}_{\hat{\pi}, \text{all}}$ and $\hat{\beta}_{\hat{\pi}, \text{part}}$ given in Section 2.

Our simulations are based on the following model, chosen so that the complete data method will be biased. Assume that $Y|X, Z \sim \text{normal}\{\beta_0 + \beta_1 X + \nu(Z), \sigma^2\}$ and that the probability of X being observed equals $\Pr(\delta = 1|Y, X, Z) = \Phi\{\alpha_0 + \alpha_1 Y + \nu_1(Z)\}$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Then $E(Y|X, Z) = \beta_0 + \beta_1 X + \nu(Z)$. A direct calculation yields

$$E(Y|X, Z, \delta = 1) = \beta_0 + \beta_1 X + \nu(Z) + \frac{\alpha_1 \sigma^2 \varphi(C/\sqrt{d})}{\sqrt{d} \Phi(C/\sqrt{d})},$$

where $\varphi(\cdot)$ is the standard normal density function, $C = \alpha_0 + \nu_1(Z) + \alpha_1\{\beta_0 + \beta_1 X + \nu(Z)\}$, and $d = 1 + \alpha_1^2 \sigma^2$. This means that unless $\alpha_1 = 0$, a complete-data analysis has the bias

$$E(Y|X, Z, \delta = 1) - E(Y|X, Z) = \frac{\alpha_1 \sigma^2 \varphi(C/\sqrt{d})}{\sqrt{d} \Phi(C/\sqrt{d})} \neq 0.$$

We considered three cases where using only the complete data leads to biases:

Table 1. Results of the Simulation Study

Case	Estimation method	Mean of estimates	Mean of analytic SE	Coverage	Mean of bootstrap SE	Coverage	Actual SE
1	Complete data	.432	.095	.914	.097	.932	.089
	Weighted partial data	.461	.105	.932	.098	.928	.097
	All data, π known	.513	.052	.952	.054	.954	.053
	All data, π estimated	.521	.055	.948	.056	.957	.053
2	Complete data	-.132	.114	.925	.125	.943	.098
	Weighted partial data	-.064	.146	.935	.139	.922	.139
	All data, π known	.032	.069	.946	.068	.944	.063
	All data, π estimated	.047	.073	.948	.072	.946	.070
3	Complete data	-.132	.114	.925	.125	.943	.098
	Weighted partial data	.143	.215	.963	.217	.974	.249
	All data, π known	.032	.069	.946	.068	.944	.063
	All data, π estimated	.128	.075	.942	.077	.943	.071

NOTE: "Mean of estimates" is the simulation mean, "Mean of analytic SE" is the mean of the estimated standard errors from the asymptotic formulae, "Mean of bootstrap SE" is the mean of the estimated standard errors from the bootstrap, and "Coverage" is the coverage probability of a nominal 95% confidence interval. The methods are "complete data" = $\hat{\beta}_{\pi, \text{comp}}$; "weighted partial data" = $\hat{\beta}_{\pi, \text{part}}$; "all data, π known" = $\hat{\beta}_{\pi, \text{all}}$; "all data, π estimated" = $\hat{\beta}_{\pi, \text{all}}$; and "Actual SE" = sample standard deviations based on 10,000 independent runs. In case 1, missingness probabilities follow and are fit by a linear probit model. In case 2, missingness probabilities follow and are fit by a semiparametric probit model. In case 3, missingness probabilities follow a semiparametric probit model but are fit by a linear probit model.

Case 1: $X \sim \text{uniform}[0, 1]$ and $Z \sim \text{uniform}[0, 1]$. Also, we assumed that $v_1(z) = \alpha_2 Z$, and used ordinary linear probit regression to estimate the missingness probabilities. Here $\alpha_0 = \beta_0 = v(z) = v_1(z) = 0$, $\sigma = 1$, $\alpha_1 = 2$, and $\beta_1 = 1/2$. This case is the canonical example for our article.

Case 2: $X \sim \text{uniform}[0, 1]$ and $Z \sim \text{uniform}[-1, 1]$. Let $\alpha_0 = \beta_0 = \beta_1 = 0$, $v(z) = 0$, and $\sigma = \alpha_1 = 1$. Here $v_1(z) = \text{sign}(z)z^2$, and we estimated the missingness probabilities using splines via the gamma function in S-PLUS. This case is meant to show what might happen when the missingness probabilities are estimated at a rate faster than $n^{-1/4}$.

Case 3: $X \sim \text{uniform}[0, 1]$ and $Z \sim \text{uniform}[-1, 1]$. Also, we assumed that $v_1(z) = \alpha_2 Z$, and used ordinary linear probit regression to estimate the missingness probabilities. However, in actuality $v_1(z) = \text{sign}(z)z^2$. Let $\alpha_0 = \beta_0 = \beta_1 = 0$, $v(z) = 0$, and $\sigma = \alpha_1 = 1$. This case is meant to be an example of slight model misspecification, since on the range of interest $v_1(z)$ is not too badly nonlinear.

The sample size was $n = 200$. Bandwidths were selected as in Remark 2. We used the quartic kernel, $K(u) = 15/16(1 - u^2)^2 I_{(|u| \leq 1)}$, and $K \star K(u)$ as the bivariate kernel. We generated 10,000 datasets in each of the three cases. For each case, approximately 35% of the X 's were missing. We computed both bootstrap and asymptotic standard errors.

The results, given in Table 1 and Figure 1, are largely in accord with the theory. The asymptotic and bootstrap standard errors are roughly correct, there is little effect to estimating $\pi(\cdot)$ in $\hat{\beta}_{\pi, \text{all}}$, coverage probabilities are near nominal (except for the complete-data estimator), and our method achieves decreases in standard errors of roughly 40%, both in actuality and in estimates.

In summary, the results illustrate our theory and are in agreement with the results of Robins et al. (1994) for purely parametric problems. First, the complete-data estimator is biased. Second, the HT estimator with estimated selection probabilities $\hat{\beta}_{\pi, \text{part}}$ is essentially unbiased, but is more variable than our estimator $\hat{\beta}_{\pi, \text{all}}$. The increases in standard errors, equivalent

to mean squared error efficiencies greater than 2 for $\hat{\beta}_{\pi, \text{all}}$ on average, are surprising but testify to the power of our approach.

6. DATA ANALYSIS OF AN AIDS CLINIC TRIAL GROUP STUDY

In this section, we present an analysis of an AIDS clinical trial group (ACTG 315) study. The purpose of this study was to investigate the relationship between virologic and immuno-

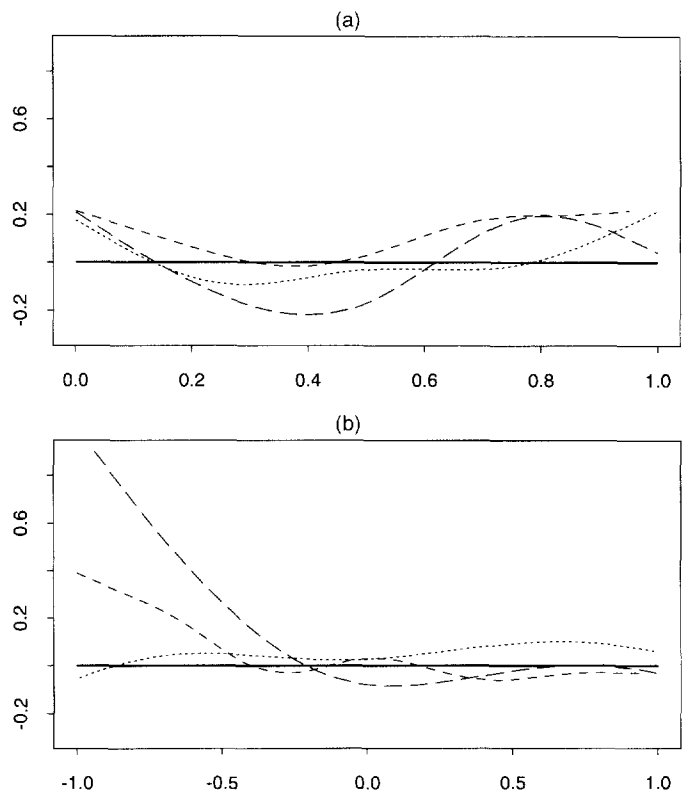


Figure 1. Estimates of Function $v(z)$ for Cases (a) and (b). The solid curve is for the true function, the dotted curve is based on the estimates from $\hat{\beta}_{\pi, \text{all}}$, the short dashed line is based on the estimates from $\hat{\beta}_{\pi, \text{part}}$, and the long dashed line is based on the complete data without weighting.

logic responses in AIDS clinical trials. In general, it is believed that the virologic response RNA (measured by viral load) and immunologic response (measured by CD4+ cell counts) are negatively correlated during treatment. Our preliminary investigations suggested that viral load depends linearly on CD4+ cell count but nonlinearly on treatment time.

In this study, both viral load and CD4+ cell counts were scheduled to be measured after initiation of antiviral therapy. There are a total of 514 observations, with 13.8% of CD4+ cell counts missing. Most of the missing values of the covariate CD4+ cell counts occurred because the covariate and the viral load were measured at different times. In other words, the missingness does not depend on the values being missing, and in this sense is MAR (Little and Rubin 1987). As Wu (2002) stated, "the MAR assumption should be reasonable for this study."

One way to model situations in which viral load depends linearly on CD4+ cell counts but nonlinearly on time is through the partially linear model (1), with Y representing viral load, X CD4+ cell count, and Z time. A similar type of model was considered by Zeger and Diggle (1994), with Y representing CD4+ cell counts, Z time, and X other covariates.

This dataset comprises 53 patients, with the number of observations per patient ranging from 3 to 11, with a median of 10 and a mean of 9.7. We have applied the estimators presented in the previous sections to analyze the dataset, assuming that the missingness mechanism follows a logistic model. The method of estimation was working independence without weighting; that is, in Section 3.4, Σ_d was assumed to be $\sigma^2 \mathbf{I}$.

We used the same kernels as in the simulation study (Sec. 5), finding bandwidths $h = .15$ and $\lambda_1 = \lambda_2 = .205$. The times were standardized to the unit interval. Confidence intervals were obtained by 200 bootstrap replications, where in the bootstrap patients were resampled (see Secs. 3.3 and 3.4 for discussion).

A simple logistic regression analysis suggests that the missingness of CD4+ cell counts depends on $Y =$ viral load, as well as $Z =$ treatment time. Because only 13% of the CD4+ cell counts were missing, it is not too surprising in retrospect that the three estimates and their bootstrap confidence intervals were similar:

$$\hat{\beta}_{\hat{\pi},all} = -.1072 (-.1445, -.0698),$$

$$\hat{\beta}_{\hat{\pi},part} = -.1016 (-.1396, -.0635),$$

and

$$\hat{\beta}_{comp} = -.1003 (-.14, -.0606).$$

The curves of the three estimated nonparametric functions of treatment time and the corresponding confidence bands are shown in Figure 2. They are also similar, and indicate that the viral load RNA levels rapidly decrease after initial antiviral treatment. Then the viral load RNA levels become flat and even rebound a little bit. The only real potential difference among the three analyses is that in Figure 2, the estimated threshold point is somewhat earlier for the complete-case method than for the other two methods.

For comparison, we also fitted a linear regression model on the data and obtained the estimated mean function as $\text{RNA} = 1.28 - .534 \times \text{Time} - .11 \times \text{CD4+}$. Note that the estimated

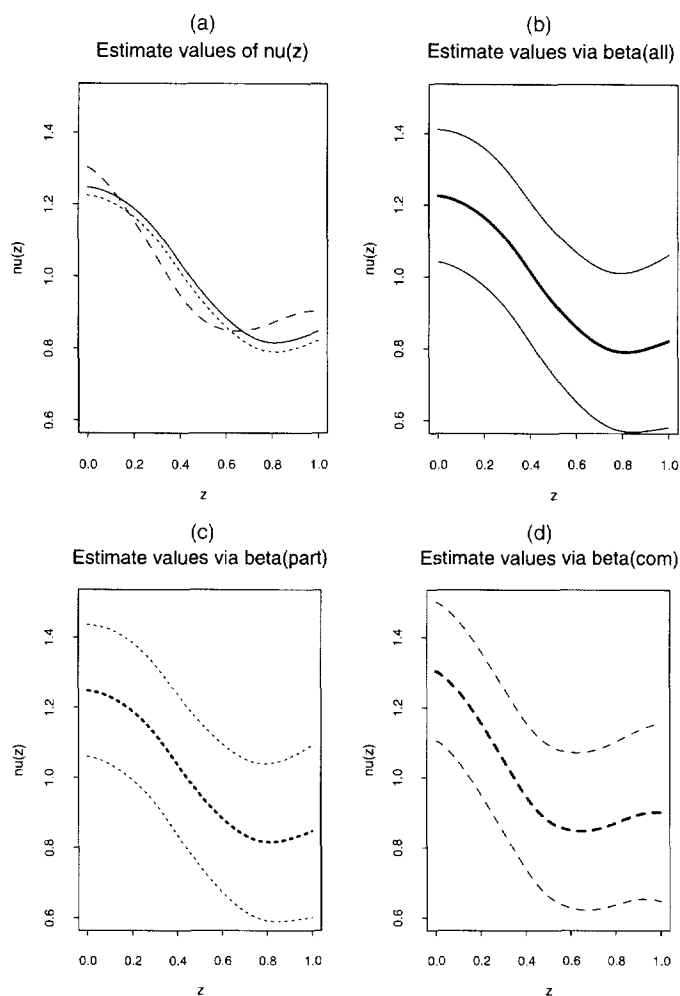


Figure 2. Estimates and the Corresponding Confidence Intervals of Function $\nu(z)$ for the ACTG 315 Dataset. The solid curves are based on the estimates from $\hat{\beta}_{\hat{\pi},all}$, the dotted lines are based on the estimates from $\hat{\beta}_{\hat{\pi},part}$, and the dashed lines are based on the complete data without weighting. (a) presents three estimates; (b), (c), and (d) show each estimate and the corresponding confidence band.

coefficient of CD4+ is in the three foregoing confidence intervals, and that the negative coefficient of Time indicates that viral load decreases with treatment time. But this linear regression does not reflect the changing trend of the viral load over the treatment time.

7. DISCUSSION

We have introduced four estimators, $\hat{\beta}_{part}$, $\hat{\beta}_{\hat{\pi},part}$, $\hat{\beta}_{all}$, and $\hat{\beta}_{\hat{\pi},all}$. The first two are based on the complete cases only, whereas the latter two use all of the data; the designation $\hat{\pi}$ means that the selection probabilities are estimated. Our analyses indicate that $\hat{\beta}_{all}$ and $\hat{\beta}_{\hat{\pi},all}$ are asymptotically equivalent, whereas $\hat{\beta}_{\hat{\pi},part}$ is generally more efficient than $\hat{\beta}_{part}$ but less efficient than $\hat{\beta}_{all}$. In effect, many of the lessons learned from the semiparametric missing-data literature carry over to the partially linear model. Our estimator $\hat{\beta}_{\hat{\pi},all}$ is easy to understand, is simple to implement, and has good behavior in finite-sample examples. The corresponding estimator for the nonparametric component $\nu(z)$ also appears to perform well.

In Section 4 we derived the score function for asymptotic semiparametric efficient estimation of β . This involves solving

an integral equation, and we have not implemented it. However, as mentioned in Section 4, our estimator is approximately semiparametric efficient when β is small and there is little heteroscedasticity.

Recall that $\widehat{\beta}_{\widehat{\pi}, \text{part}}$ is the estimator using the complete data only but with estimated selection probabilities. To obtain the covariance matrix of $\widehat{\beta}_{\widehat{\pi}, \text{part}}$ explicitly and compare it with that of $\widehat{\beta}_{\text{part}}$, we have assumed that π can be modeled parametrically. Note that nonparametric estimation of π with a rate higher than $n^{-1/4}$ does not effect the asymptotic distribution of the estimator $\widehat{\beta}_{\widehat{\pi}, \text{all}}$. An interesting question is whether we have similar conclusions for $\widehat{\beta}_{\widehat{\pi}, \text{part}}$ when π is estimated nonparametrically with a rate higher than $n^{-1/4}$. This is a topic for further study.

A referee has mentioned that it would be an interesting problem to estimate $E(Y)$. A simple estimate is the sample mean of the Y 's, but Cheng (1994) has shown in the purely nonparametric context with missing Y 's that more efficient estimation of $E(Y)$ is possible using the nonparametric function estimate. We conjecture that the same is true in the partially linear model with missing X 's, but we have not carried through with the calculations.

Our example (and, consequently, the article) concerns the case where X is missing but Z is not. We have not studied the case where in model (1), Z , not X , is missing. It should be possible to extend our results to this case, however. Steps 1–4 would change only in the sense that $\pi(Y, Z)$ would be replaced by $\pi(Y, X)$ and that conditioning on (Y, Z) in Step 3 would be replaced by conditioning on (Y, X) . This would actually make some parts of Step 3 easier, because terms such as $E(XX^T|Y, Z)$ would now be replaced by $E(XX^T|Y, X) = XX^T$. We conjecture that the obvious analog of Theorem 1 holds in this case.

It would appear possible, at least in principal, to extend model (1) to quasi-likelihood models with mean $\mu\{X^T\beta + v(Z)\}$ and variance function $\sigma^2 V\{X^T\beta + v(Z)\}$. Major work on this model when no data are missing was done by Severini and Staniswalis (1994); the notation of Lin and Carroll (2001) is closer to ours. Our approach can be viewed as taking the following steps: (a) computing in a closed form their profile likelihood score $\widetilde{X}(\widetilde{Y} - \widetilde{X}^T\beta)$ to model (1) theoretically, (b) basing estimation on the analog to the Robins et al. (1994) estimating (3), and (c) estimating the required nonparametric functions in the result. In quasi-likelihood models there is no closed form in (a), and this would appear to be the major difficulty in extending our approach. There is, however, a profile quasi-likelihood estimating equation defined in an iterative fashion (Severini and Staniswalis 1994; Lin and Carroll 2001). Using this, we would expect that a result very much like Theorem 1 would hold. Following these two references, we would expect to replace ϵ by $\{Y - \mu(\cdot)\}/V(\cdot)$ and \widetilde{X} by $\widetilde{X}^*\mu^{(1)}(\cdot)/V(\cdot)$, where $\mu^{(1)}$ is the derivative of μ , and $\widetilde{X}^* = X - B(Z)/A(Z)$, with $A(Z) = E\{[\mu^{(1)}(\cdot)]^2/[\sigma^2(\cdot)V(\cdot)]|Z\}$ and $B(Z) = E\{X[\mu^{(1)}(\cdot)]/[\sigma^2(\cdot)V(\cdot)]|Z\}$, with “ \cdot ” being $X^T\beta + v(Z)$. In principle, the method proposed in this article can also be extended to partial nonlinear regression models. One main difficulty with this approach is that the estimating equations for β will no longer be linear in β , which creates potential problems when solving for $\widehat{\beta}$. A detailed investigation of these issues would be interesting, but is beyond the scope of this article.

APPENDIX: PROOFS

A.1 Sketch Proof of Theorems 1 and 2

The result here follows immediately from the general results of Newey (1994). On a formal mathematical basis, all of the uniform convergence results require that (Y, Z) have compact support, that their joint density is positive on that support, and that $\pi(\cdot) > 0$. We use these assumptions without comment in what follows.

Let $m_1(Z) = E(X|Z)$, $m_2(Z) = E(Y|Z)$, $m_3(Y, Z) = E(X|Y, Z)$, and $m_4(Y, Z) = E(XX^T|Y, Z)$. Then $v(z)$ in model (1) is $m_2(z) - m_1^T(z)\beta$. Define

$$\begin{aligned} \Psi(m_1, m_2, m_3, m_4, \pi, \beta, Y, X, Z, \delta) &= \{X - m_1(Z)\}[Y - m_2(Z) - \{X - m_1(Z)\}^T\beta] \frac{\delta}{\pi} \\ &\quad - \{Y - m_2(Z)\}\{m_3(Y, Z) - m_1(Z)\} \\ &\quad - \{m_4(Y, Z) - m_3(Y, Z)m_1^T(Z) - m_1(Z)m_3^T(Y, Z) \\ &\quad + m_1(Z)m_1^T(Z)\}\beta \frac{\delta - \pi}{\pi}. \end{aligned} \tag{A.1}$$

Our estimators $\widehat{\beta}_{\text{all}}$ and $\widehat{\beta}_{\widehat{\pi}, \text{all}}$ solve the estimating equations

$$0 = \sum_{i=1}^n \Psi(\widehat{m}_1, \widehat{m}_2, \widehat{m}_3, \widehat{m}_4, \pi, \beta, Y_i, X_i, Z_i, \delta_i)$$

and

$$0 = \sum_{i=1}^n \Psi(\widehat{m}_1, \widehat{m}_2, \widehat{m}_3, \widehat{m}_4, \widehat{\pi}, \beta, Y_i, X_i, Z_i, \delta_i).$$

Let

$$\begin{aligned} D(m_1^* - m_1, m_2^* - m_2, m_3^* - m_3, m_4^* - m_4, \pi^* - \pi, \beta, Y, X, Z, \delta) &= \sum_{j=1}^4 \frac{\partial \Psi}{\partial m_j} (m_j^* - m_j) + \frac{\partial \Psi}{\partial \pi} (\pi^* - \pi), \end{aligned}$$

where the partial derivatives are the Frechet partial derivatives. It is easy to obtain that

$$\begin{aligned} \frac{\partial \Psi}{\partial \pi} &= -\{X - m_1(Z)\}[Y - m_2(Z) - \{X - m_1(Z)\}^T\beta] \frac{\delta}{\pi^2} \\ &\quad + \{Y - m_2(Z)\}\{m_3(Y, Z) - m_1(Z)\} \\ &\quad - \{m_4(Y, Z) - m_3(Y, Z)m_1^T(Z) \\ &\quad - m_1(Z)m_3^T(Y, Z) + m_1(Z)m_1^T(Z)\}\beta \frac{\delta}{\pi^2} \end{aligned}$$

and

$$\frac{\partial \Psi}{\partial m_2} = \{m_3(Y, Z) - m_1(Z)\} \frac{\delta - \pi}{\pi} - \{X - m_1(Z)\} \frac{\delta}{\pi}.$$

Other Frechet partial derivatives, $\frac{\partial \Psi}{\partial m_1}$, $\frac{\partial \Psi}{\partial m_3}$, and $\frac{\partial \Psi}{\partial m_4}$, are more cumbersome to express because there are matrices and up to four-dimensional arrays involved in the expressions, but these expressions are in essence the same as those given earlier. It follows from direct calculation that $E(\frac{\partial \Psi}{\partial \pi}) = 0$ and $E(\frac{\partial \Psi}{\partial m_j}) = 0$ for $j = 1, \dots, 4$.

In addition,

$$\begin{aligned} &\|\Psi(m_1^*, m_2^*, m_3^*, m_4^*, \pi^*, \beta, Y, X, Z, \delta) \\ &\quad - \Psi(m_1, m_2, m_3, m_4, \pi, \beta, Y, X, Z, \delta) \\ &\quad - D(m_1^* - m_1, m_2^* - m_2, m_3^* - m_3, m_4^* - m_4, \\ &\quad \pi^* - \pi, \beta, Y, X, Z, \delta)\| \\ &= O(\|m_1^* - m_1\|^2 + \|m_2^* - m_2\|^2 \\ &\quad + \|m_3^* - m_3\|^2 + \|m_4^* - m_4\|^2 + \|\pi^* - \pi\|^2). \end{aligned} \tag{A.2}$$

where $\|h\|$ denotes a norm for the function h , such as the Sobolev norm, that is a supremum norm for a function and its derivatives. Equation (A.2) is Newey's assumption 5.1(i). We have made the assumptions that $\widehat{m}_j(\cdot) - m_j(\cdot) = o_p(n^{-1/4})$ for $j = 1, 2, 3, 4$ and that $\widehat{\pi}(\cdot) - \pi(\cdot) = o_p(n^{-1/4})$. This is assumption 5.1(ii) of Newey (1994). Again, Newey's assumption 5.2 holds by the expression of $D(\cdot, \beta, Y, X, Z, \delta)$. In addition, it follows from the foregoing statements that for any $(m_1^*, m_2^*, m_3^*, m_4^*, \pi^*)$,

$$E\{D(m_1^* - m_1, m_2^* - m_2, m_3^* - m_3, m_4^* - m_4, \pi^* - \pi, \beta, Y, X, Z, \delta)\} = 0,$$

thus verifying Newey's assumption 5.3; his $\alpha(z) = 0$, according to his discussion just before his lemma 5.1. By that lemma, it follows that both $\widehat{\beta}_{\pi^*, \text{all}}$ and $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ have the same limit distribution as the solution to the equation

$$0 = \sum_{i=1}^n \Psi(m_1, m_2, m_3, m_4, \pi, \beta, Y_i, X_i, Z_i, \delta_i). \quad (\text{A.3})$$

It is an easy calculation to show that the solution to (A.3) has the same limit distribution as described in the statement of Theorem 1. This completes the proof.

A.2 Proof of (6)

We use the same line of argument as in Section A.1, although direct calculations are also possible. Here the estimating function is

$$\Psi_1(m_1, m_2, \pi, \beta, Y, X, Z, \delta) = \{X - m_1(Z)\} [Y - m_2(Z) - \{X - m_1(Z)\}^T \beta] \frac{\delta}{\pi}. \quad (\text{A.4})$$

In our claim of (6), $\pi(\cdot)$ is known, so that in applying the results of Newey, we need only consider estimation of m_1 and m_2 . The same argument as in Section A.1 now applies directly to (A.4). This completes the proof.

A.3 Proof of (7)

Let $\pi_\psi = (\partial\pi/\partial\psi)$. The joint estimating functions are (A.4) and

$$\Psi_2(\pi, Y, X, Z, \delta) = \frac{\pi_\psi(Y, Z, \psi)}{\pi(Y, Z, \psi)\{1 - \pi(Y, Z, \psi)\}} \{\delta - \pi(Y, Z, \psi)\}. \quad (\text{A.5})$$

As in Section A.2, the asymptotic distributions of $\widehat{\beta}_{\widehat{\pi}^*, \text{part}}$ and $\widehat{\psi}$ are the same as if $m_1(\cdot)$ and $m_2(\cdot)$ were known. The solutions of (A.4) and (A.5) thus have the standard form of parametric estimating equations. Detailed but routine calculations yield (7).

A.4 Brief Justification of the Bootstrap for Independent Data

Here we provide a sketch of an argument showing that the bootstrap can be used to construct consistent estimates of standard errors for independent data. We believe that the same type of argument would apply for clustered data with working independence, but that the argument for this more general case would be much more involved.

Define the bootstrap estimator $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ and two other versions, $\widehat{\beta}_{\pi^*, \text{all}}$ and $\widehat{\beta}_{\pi^*, \text{part}}$, to be the solution to the estimating equations

$$0 = \sum_{i=1}^n \Psi(\widehat{m}_1^*, \widehat{m}_2^*, \widehat{m}_3^*, \widehat{m}_4^*, \widehat{\pi}^*, \beta, Y_i^*, X_i^*, Z_i^*, \delta_i^*),$$

$$0 = \sum_{i=1}^n \Psi(\widehat{m}_1, \widehat{m}_2, \widehat{m}_3, \widehat{m}_4, \widehat{\pi}, \beta, Y_i^*, X_i^*, Z_i^*, \delta_i^*),$$

and

$$0 = \sum_{i=1}^n \Psi(m_1, m_2, m_3, m_4, \pi, \beta, Y_i^*, X_i^*, Z_i^*, \delta_i^*),$$

where Ψ is given in (A.1) and $*$ indicates resampled data or the corresponding estimates. Note that for each i , all functions, $\widehat{m}_1^*, \widehat{m}_2^*, m_1$, and so on, are evaluated at bootstrap-resampled observation (Y_i^*, Z_i^*) .

First, we claim that $\widehat{m}_j^*(\cdot) - \widehat{m}_j(\cdot) = o_p(n^{-1/4})$ for $j = 1, 2, 3, 4$, and $\widehat{\pi}^*(\cdot) - \widehat{\pi}(\cdot) = o_p(n^{-1/4})$. We will come back to this claim at the end of the justification of the main bootstrap result.

Following the same arguments as in the proof of Theorem 1 by verifying the assumptions of Newey (1994) for his lemma 5.1, we obtain that $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ and $\widehat{\beta}_{\pi^*, \text{all}}$ have the same limit distribution. Using Newey's lemma 5.1 again, it follows that $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ and $\widehat{\beta}_{\pi^*, \text{part}}$ have the same limit distribution. The foregoing results imply that $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ and $\widehat{\beta}_{\pi^*, \text{all}}$ have the same limit distribution.

Now let $\Phi(\beta, Y, X, Z, \delta) = \Psi(m_1, m_2, m_3, m_4, \pi, \beta, Y, X, Z, \delta)$ and let $\widetilde{\beta}_{\text{all}}$ be the solution to

$$0 = \sum_{i=1}^n \Phi(\beta, Y_i, X_i, Z_i, \delta_i).$$

Then $\widetilde{\beta}_{\text{all}}$ solves

$$0 = \sum_{i=1}^n \Phi(\beta, Y_i^*, X_i^*, Z_i^*, \delta_i^*).$$

A standard bootstrap argument shows that $\widetilde{\beta}_{\text{all}}^* - \widetilde{\beta}_{\text{all}}$ and $\widetilde{\beta}_{\text{all}} - \beta$ have the same limit distribution.

On the other hand, in the proof of Theorem 1 we showed that $\widetilde{\beta}_{\text{all}}$ and $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ are asymptotically equivalent to the first order. Therefore, $\widetilde{\beta}_{\text{all}}^* - \widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ and $\widehat{\beta}_{\widehat{\pi}^*, \text{all}} - \beta$ have the same limit distribution. Combining this with the foregoing result about the relationship between $\widehat{\beta}_{\widehat{\pi}^*, \text{all}}$ and $\widehat{\beta}_{\pi^*, \text{all}}$, we obtain that $\widetilde{\beta}_{\text{all}}^* - \widehat{\beta}_{\pi^*, \text{all}}$ and $\widehat{\beta}_{\pi^*, \text{all}} - \beta$ have the same limit distribution, completing the proof.

We now go back to check the claim that we made at the beginning of this section. Because of space limitations, we give an outline for $\widehat{m}_1^*(\cdot) - \widehat{m}_1(\cdot)$ only, with X univariate.

Let

$$L_n(\alpha) = n^{-1} \sum_{i=1}^n K_h(Z_i - z_0) \frac{\delta_i}{\pi(Y_i, Z_i)} \left(\frac{1}{(Z_i - z_0)/h} \right) \times \left\{ X_i - \alpha_0 - \alpha_1 \left(\frac{Z_i - z_0}{h} \right) \right\},$$

where $\alpha = (\alpha_0, \alpha_1)^T$ and π is first assumed to be known; see (4) with $q(Y, X) = X$. Let

$$L_n^*(\alpha) = n^{-1} \sum_{i=1}^n K_h(Z_i^* - z_0) \frac{\delta_i^*}{\pi(Y_i^*, Z_i^*)} \left(\frac{1}{(Z_i^* - z_0)/h} \right) \times \left\{ X_i^* - \alpha_0 - \alpha_1 \left(\frac{Z_i^* - z_0}{h} \right) \right\}$$

be the bootstrap version of $L_n(\alpha)$. Then, by the standard bootstrap argument (Davison and Hinkley 1997, p. 22), we have

$$E^*\{L_n^*(\alpha) - L_n(\alpha)\} = 0 \quad \text{almost surely,}$$

and

$$\begin{aligned} \text{cov}^*\{L_n^*(\alpha)\} &= n^{-1} \text{cov}^* \left[K_h(Z_1^* - z_0) \frac{\delta_1^*}{\pi(Y_1^*, Z_1^*)} \left(\frac{1}{(Z_1^* - z_0)/h} \right) \right. \\ &\quad \left. \times \left\{ X_1^* - \alpha_0 - \alpha_1 \left(\frac{Z_1^* - z_0}{h} \right) \right\} \right] \end{aligned}$$

$$\begin{aligned}
 &= n^{-1} E^* \left[K_h^2(Z_1^* - z_0) \frac{\delta_1^*}{\pi^2(Y_1^*, Z_1^*)} \right. \\
 &\quad \times \left. \left\{ X_1^* - \alpha_0 - \alpha_1 \left(\frac{Z_1^* - z_0}{h} \right) \right\}^2 \left(\frac{1}{(Z_1^* - z_0)/h} \right)^{\otimes 2} \right] \\
 &\quad + O_p(n^{-1}) \\
 &= n^{-2} \sum_{i=1}^n K_h^2(Z_i - z_0) \frac{\delta_i}{\pi^2(Y_i, Z_i)} \\
 &\quad \times \left\{ X_i - \alpha_0 - \alpha_1 \left(\frac{Z_i - z_0}{h} \right) \right\}^2 \left(\frac{1}{(Z_i - z_0)/h} \right)^{\otimes 2} \\
 &\quad + O_p(n^{-1}) \\
 &= (nh)^{-1} A(z_0) + o_p\{(nh)^{-1}\} = O_p\{(nh)^{-1}\},
 \end{aligned}$$

where E^* and cov^* are the bootstrap expectation and covariance conditional on observed data,

$$\begin{aligned}
 A(z_0) = \int \int K^2(t) \pi^{-1}(y, z_0) \{ m_4(y, z_0) - 2(\alpha_0 + \alpha_1 t) m_3(y, z_0) \\
 + (\alpha_0 + \alpha_1 t)^2 \} \left(\frac{1}{t} \right)^{\otimes 2} f(y, z_0) dt dy,
 \end{aligned}$$

and $f(y, z)$ is the joint density of (Y, Z) . Therefore,

$$L_n^*(\alpha) - L_n(\alpha) = O_p\{(nh)^{-1/2}\} = o_p(n^{-1/4}) \tag{A.6}$$

for a proper choice of h , such as h having a rate of $n^{-1/5}$.

If π is to be estimated with observed data and with resampled data in the bootstrap case, then result (A.6) still holds for the corresponding difference $L_n^*(\alpha) - L_n(\alpha)$. This can be verified by approximating the bootstrap mean and variance of the difference by using a Taylor series expansion. The details are straightforward but lengthy, involving the calculation of the variance of the second-order term in the expansion, which is seen to be negligible.

Finally, recall that $\widehat{m}_1(z_0)$ and $\widehat{m}_1^*(z_0)$ are defined through the estimating equations $L_n(\alpha) = 0$ and $L_n^*(\alpha) = 0$. From (A.6) and using the method of scoring, it is readily seen that $\widehat{m}_1^*(z_0) - \widehat{m}_1(z_0) = o_p(n^{-1/4})$, as was claimed.

A.5 Asymptotic Covariance Matrix in the Longitudinal Case

To describe the asymptotic covariance matrix in the longitudinal case, we need some additional definitions. To maintain notational simplicity, we adapt the notation in the iid case as much as possible and also focus on the special case of equal replications for each subject i , $i = 1, \dots, n$, that is, the number of replications $\equiv M$, say. The results are readily extended to the more general case, although formulas become somewhat more complex. For each i , define $Y_i = (Y_{i1}, \dots, Y_{iM})^T$, $\mathbf{X}_i = (X_{i1}, \dots, X_{iM})^T$, $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iM})^T$, $\delta_i = (\delta_{i1}, \dots, \delta_{iM})^T$, $\pi_i = (\pi_{i1}, \dots, \pi_{iM})^T$, $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{iM})^T$, $\mathbf{F}_i = \text{diag}(\delta_{i1}/\pi_{i1}, \dots, \delta_{iM}/\pi_{iM})^T$, and $\mathbf{G}_i = \text{diag}(\delta_{i1}/\pi_{i1} - 1, \dots, \delta_{iM}/\pi_{iM} - 1)$. Furthermore, let Σ_d be the $M \times M$ diagonal matrix with the diagonal elements of $\text{cov}(Y|\mathbf{X}, \mathbf{Z})$ or, indeed, any working covariance matrix. Then using the same notation as in (A.1), the estimating function for each subject under the working independence assumption is

$$\begin{aligned}
 &\Psi(m_1, m_2, m_3, m_4, \pi, \beta, Y, \mathbf{X}, \mathbf{Z}, \delta) \\
 &= \{\mathbf{X} - m_1(\mathbf{Z})\}^T \Sigma_d^{-1} \mathbf{F} [Y - m_2(\mathbf{Z}) - \{\mathbf{X} - m_1(\mathbf{Z})\} \beta] \\
 &\quad - \{(m_3(Y, \mathbf{Z}) - m_1(\mathbf{Z}))^T \Sigma_d^{-1} \mathbf{G} (Y - m_2(\mathbf{Z})) \\
 &\quad - [m_4(Y, \mathbf{Z}, \delta) - m_3^T(Y, \mathbf{Z}) \Sigma_d^{-1} \mathbf{G} m_1(\mathbf{Z}) \\
 &\quad - m_1^T(\mathbf{Z}) \Sigma_d^{-1} \mathbf{G} \{m_3(Y, \mathbf{Z}) - m_1(\mathbf{Z})\}] \beta\}.
 \end{aligned}$$

where $m_1(\mathbf{Z}_i) = \{m_1(Z_{i1}), \dots, m_1(Z_{iM})\}^T$ with $m_1(Z_{ik}) = E(X_{ik}|Z_{ik})$ for $k = 1, \dots, M$; $m_2(\mathbf{Z}_i)$ and $m_3(Y_i, \mathbf{Z}_i)$ are defined similarly; and $m_4(Y_i, \mathbf{Z}_i, \delta_i) = E^\dagger(\mathbf{X}_i^T \Sigma_d^{-1} \mathbf{G}_i \mathbf{X}_i | Y_i, \mathbf{Z}_i, \delta_i)$, with E^\dagger the conditional expectation of any function of (X_{ik}, δ_{ik}) only on $(Y_{ik}, Z_{ik}, \delta_{ik})$. Here we have assumed that all the regressions are marginal; that is, $m_1(z) = E(X_{ik}|Z_{ik} = z)$ does not depend on k , and the same is true for m_2, m_3 , and m_4 .

Using some standard algebra, with $\widetilde{\mathbf{X}} = \mathbf{X} - m_1(\mathbf{Z})$, the covariance of the limit distribution can be easily obtained as

$$\{E(\widetilde{\mathbf{X}}^T \Sigma_d^{-1} \widetilde{\mathbf{X}})\}^{-1} C^* \{E(\widetilde{\mathbf{X}}^T \Sigma_d^{-1} \widetilde{\mathbf{X}})\}^{-1},$$

$$C^* = E\{\widetilde{\mathbf{X}}^T \Sigma_d^{-1} \mathbf{F} \epsilon - E(\widetilde{\mathbf{X}}^T \Sigma_d^{-1} \mathbf{G} \epsilon | \mathbf{Y}, \mathbf{Z}, \delta)\}^{\otimes 2}.$$

When $m_j(\cdot)$ ($j = 1, 2, 3, 4$) and $\pi(\cdot)$ are estimated with an error of $o_p(n^{-1/4})$, the resulting estimators have the same limiting distribution. It is readily seen that our estimation steps for these nuisance functions in the working independence setting produce estimation errors of $o_p(n^{-1/4})$. Note that the foregoing asymptotic covariance can be estimated via resampling methods such as the bootstrap. However, if the independence of the observations within subjects holds and $\Sigma_d = \sigma^2 \mathbf{I}$, then the covariance formula reduces to that in Theorem 1, as expected.

A.6 Calculation of the Efficient Score

In this section we show that the efficient score is (9), with $H_{\text{opt}}(\cdot)$ given by (10). We draw on section 8 of Robins et al. (1994), with the exception that their $\Lambda_0^{F, \perp}$ is denoted here by $\Lambda^{\perp, F}$, the linear span of the space of influence function in model (1) when there are no missing data. Bickel, Klaassen, Ritov, and Wellner (1993) and Robins (1994) showed that

$$\Lambda^{\perp, F} = \{h(X, Z)\epsilon \text{ such that } E\{h(X, Z)|Z\} = 0\}.$$

Make the definitions $L = (Y, X, Z)$ = complete data and $D = D(L)$ and $B = B(L)$ generic functions of the complete data; these were called B^* and D^* by Robins et al. (1994). Make the further definitions

$$W = 1/E(\epsilon^2|X, Z)$$

and

$$\prod\{D|\Lambda^{\perp, F}\} = \epsilon W \{E(D\epsilon|X, Z) - E(WD\epsilon|Z)/E(W|Z)\}.$$

Note that $\prod(\cdot)$ is the projection operator into $\Lambda^{\perp, F}$.

Recall that $\pi = \pi(Y, Z)$ is the probability that X is observed. Following Robins et al. (1994), define δ to be the indicator that X is observed and "obs" to be the observed data, so that "obs" = (Y, Z, δ) if $\delta = 0$ and "obs" = (Y, Z, X, δ) if $\delta = 1$. Make the definitions

$$g(D) = E(D|\text{obs})$$

and

$$m(D) = E\{E(D|\text{obs})|L\} = (1 - \pi)E(D|Y, Z) + \pi D.$$

Then $g(D) = \delta D + (1 - \delta)E(D|Y, Z)$. By proposition 8.1.d of Robins et al. (1994), it is easily verified that the inverse of $m(\cdot)$ is

$$m^{-1}(D) = \frac{D}{\pi} - \left(\frac{1}{\pi} - 1\right)E(D|Y, Z).$$

Let S_β be the derivative of the log-likelihood with respect to β : $S_\beta = X$ if ϵ is homoscedastic and normally distributed. Define $\mu = \mu(X, Z) = E(\epsilon S_\beta | X, Z)$. The efficient score in the full data space is

$$S_{\text{eff}}^F = \prod\{S_\beta|\Lambda^{\perp, F}\} = \epsilon W \{\mu - E(W\mu|Z)/E(W|Z)\}.$$

From proposition 8.1.e1 of Robins et al. (1994), we need to find D_{opt} in $\Lambda^{\perp, F}$ such that $S_{\text{eff}}^F = \prod\{m^{-1}(D_{\text{opt}})|\Lambda^{\perp, F}\}$, in which case the efficient score function is $g\{m^{-1}(D_{\text{opt}})\}$. This means that D_{opt} satisfies

$$\begin{aligned} \mu - E(W\mu|Z)/E(W|Z) \\ = E\{m^{-1}(D_{\text{opt}})\epsilon|X, Z\} \\ - E\{WE\{m^{-1}(D_{\text{opt}})\epsilon|X, Z\}|Z\}/E(W|Z). \end{aligned} \quad (\text{A.7})$$

Because the optimal choice of D , say D_{opt} , is in $\Lambda^{\perp, F}$, it is necessarily of the form $D_{\text{opt}} = \epsilon H_{\text{opt}}(X, Z)$. We now compute $H_{\text{opt}} = H_{\text{opt}}(\cdot)$.

Recall that $K = K(X, Z) = E(\epsilon^2/\pi|X, Z)$ and

$$G_{\text{opt}} = G(X, Z, H_{\text{opt}}) = E\{\epsilon(1/\pi - 1)E(\epsilon H_{\text{opt}}|Y, Z)|X, Z\}.$$

Further, define

$$A_{\text{opt}} = \{E(WH_{\text{opt}}K|Z) - E(WG_{\text{opt}}|Z)\}/E(W|Z).$$

It is easily seen that

$$E\{m^{-1}(D_{\text{opt}})\epsilon|X, Z\} = H_{\text{opt}}K - G_{\text{opt}}. \quad (\text{A.8})$$

Now plug (A.8) back into (A.7) to find that

$$\begin{aligned} \mu - E(W\mu|Z)/E(W|Z) = H_{\text{opt}}K - G_{\text{opt}} \\ - \{E(WH_{\text{opt}}K|Z) - E(WG_{\text{opt}}|Z)\}/E(W|Z). \end{aligned}$$

The last term is A_{opt} , so that we divide through by K to find that

$$\mu/K - E(W\mu|Z)/\{KE(W|Z)\} = H_{\text{opt}} - G_{\text{opt}}/K - A_{\text{opt}}/K. \quad (\text{A.9})$$

Recall, however, that in the definition of $\Lambda^{\perp, F}$, $E(H_{\text{opt}}|Z) = 0$. Using this fact and taking expectation of (A.9) conditional on Z , we find that

$$\begin{aligned} -A_{\text{opt}} = [E(\mu/K|Z) - E(W\mu|Z)E\{(1/K)|Z\}/E(W|Z) \\ + E(G_{\text{opt}}/K|Z)]/E\{(1/K)|Z\}. \end{aligned} \quad (\text{A.10})$$

Now put (A.10) back into (A.9) to get the integral equation

$$\begin{aligned} H_{\text{opt}} = (1/K)[\mu + G_{\text{opt}} - E(\mu/K|Z)/E\{(1/K)|Z\} \\ - E(G_{\text{opt}}/K|Z)/E\{(1/K)|Z\}]. \end{aligned} \quad (\text{A.11})$$

From proposition 8.1.e1 of Robins et al. (1994), the efficient score function is $g\{m^{-1}(D_{\text{opt}})\}$, and this is easily seen to equal

$$g\{m^{-1}(D_{\text{opt}})\} = \frac{\delta}{\pi}\epsilon H_{\text{opt}} - \frac{\delta - \pi}{\pi}E(\epsilon H_{\text{opt}}|Y, Z). \quad (\text{A.12})$$

[Received November 2001. Revised October 2003.]

REFERENCES

- Bickel, P. J., Klaassen, C. J., Ritov, Y., and Wellner, J. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
- Carroll, R. J., Knickerbocker, R. K., and Wang, C. Y. (1995), "Dimension Reduction in Semiparametric Measurement Error Models," *The Annals of Statistics*, 23, 161-181.
- Chamberlain, G. (1992), "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567-596.
- Cheng, P. E. (1990), "Applications of Kernel Regression Estimation: A Survey," *Communications in Statistics, Part A—Theory and Methods*, 19, 4103-4134.
- (1994), "Nonparametric Estimation of Mean Functionals With Data Missing at Random," *Journal of the American Statistical Association*, 89, 81-87.
- Cheng, P. E., and Chu, C. K. (1996), "Kernel Estimation of Distribution Functions and Quantiles With Missing Data," *Statistica Sinica*, 6, 63-78.
- Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge, U.K.: Cambridge University Press.
- Fan, J., and Zhang, J. T. (2000), "Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of the Royal Statistical Society, Ser. B*, 62, 303-322.
- Härdle, W., Liang, H., and Gao, J. (2000), *Partially Linear Models*. Heidelberg: Springer Physica-Verlag.
- Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, Y. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809-822.
- Lin, D. Y., and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data," *Journal of the American Statistical Association*, 96, 103-126.
- Lin, X. H., and Carroll, R. J. (2001), "Semiparametric Regression for Clustered Data Using Generalized Estimating Equations," *Journal of the American Statistical Association*, 96, 1045-1056.
- Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*. New York: Wiley.
- Mack, Y., and Silverman, B. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 60, 405-415.
- Marron, J. S., and Härdle, W. (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91-113.
- Newey, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349-1382.
- Robins, J. M. (1994), "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics*, 23, 2379-2412.
- Robins, J. M., Mark, S. D., and Newey, W. K. (1992), "Estimating Exposure Effects by Modelling the Expectation of Exposure conditional on Confounders," *Biometrics*, 48, 479-495.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846-866.
- Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257-1270.
- Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501-511.
- Silverman, B. W. (1984), "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898-916.
- Wang, C. Y., Wang, S., Gutierrez, R. G., and Carroll, R. J. (1998), "Local Linear Regression for Generalized Linear Models With Missing Data," *The Annals of Statistics*, 26, 1028-1050.
- Wu, L. (2002), "A Joint Model for Nonlinear Mixed-Effects Models With Censoring and Covariates Measured With Error, With Application to AIDS Studies," *Journal of the American Statistical Association*, 97, 955-964.
- Zeger, S. L., and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689-699.