

Semiparametric Efficient Estimation of a Conditional Density with Missing or Mismeasured Covariates

By JAMES M. ROBINS† and FUSHING HSIEH and WHITNEY NEWEY
Harvard School of Public Health, Boston, USA *Massachusetts Institute of Technology, Cambridge, USA*

[Received April 1992. Final revision June 1994]

SUMMARY

Pepe and Fleming, and Carroll and Wand have recently proposed estimators in a parametric model for the density of a random variable Y conditional on a vector of covariates (X, V) when data on one of the regressors X is missing for some study subjects. We propose a new class of estimators that remains consistent and asymptotically normal even when the probability that X is missing depends on the observed V and Y , includes an estimator whose asymptotic variance attains the semiparametric variance bound for the model and, when the data are missing completely at random, includes an estimator that is asymptotically equivalent to the inefficient estimators proposed by Pepe and Fleming and by Carroll and Wand. The optimal estimator in our class depends on the unknown probability law generating the data. When the vector V of non-missing regressors has at most two continuous components, we propose an adaptive semiparametric efficient estimator and compare the performance of the proposed semiparametric efficient estimator with the estimators proposed by Pepe and Fleming and Carroll and Wand in a small simulation study. When V has many continuous components, we propose an alternative class of adaptive estimators that should have high efficiency.

Keywords: MEASUREMENT ERROR; MISSING COVARIATES; MISSING DATA; SEMIPARAMETRIC EFFICIENCY; VALIDATION SAMPLE

1. INTRODUCTION

In applied problems it is common for an investigator to specify a parametric model, say $f[Y_i|X_i, V_i; \alpha_0]$, for the density of a possibly multivariate dependent variable Y_i conditional on a set of regressor variables (X_i, V_i) . Either by happenstance or design, data on one of the regressors, say X_i , may be missing for a subset of the study subjects. For example, X_i might represent subject i 's response to a confidential personal question which some subjects may refuse to answer. As a second example, one of the regressors may be mismeasured. Then the mismeasured value Z_i is a component of V_i and is said to be a surrogate for the true value X_i . In this setting an investigator will often measure X_i without error on a small sample of the subjects—the validation sample.

Pepe and Fleming (1991) and Carroll and Wand (1991) have recently proposed estimators for α_0 that extract information from the non-validation sample. Their estimators are asymptotically normal and unbiased whatever the distribution of the missing or mismeasured regressor X_i , conditional on the remaining regressors V_i .

†*Address for correspondence:* Department of Epidemiology and Biostatistics, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA.
E-mail: Robins@HSPH.HARVARD.EDU

The class of estimators proposed in this paper, like the Pepe-Fleming and Carroll-Wand estimators, is consistent and asymptotically normal without specifying the distribution of X_i given V_i . The estimators proposed have several potential advantages over the Pepe-Fleming and Carroll-Wand estimators.

First, Pepe and Fleming (1991) and Carroll and Wand (1991) assumed that missingness does not depend on the outcome Y_i . In contrast, our estimators remain asymptotically normal and unbiased even when missingness depends on Y_i , as would be the case if subjects with certain values of Y_i are rare, and the investigator wished to overrepresent such subjects in the validation sample.

Secondly, their estimators require a preliminary nonparametric estimate of the law of the missing covariate X_i conditional on the vector V_i . If the vector V_i includes more than two continuous covariates, completely nonparametric estimation of the conditional law of X_i given V_i is not generally practical owing to the 'curse of dimensionality' (Huber, 1985). Therefore, their approach is not available (Carroll and Wand, 1991). In contrast, there are estimators in our class which remain asymptotically unbiased and may be computationally simple whatever the dimension of V_i .

Thirdly, our class of estimators contains an 'estimator' whose asymptotic variance attains the semiparametric variance bound for regular estimators of α_0 in the sense of Begun *et al.* (1983). The particular estimator in our class whose asymptotic variance attains the bound depends on the true but unknown probability law generating the data, and, thus, is not available for data analysis. However, when Y_i and V_i are discrete, we show how to construct a semiparametric efficient adaptive estimator based on a nonparametric estimate of the law of X_i given V_i . We also propose candidate semiparametric efficient adaptive estimators when Y_i has one or more continuous components and/or V_i contains at most two continuous components.

The paper is organized as follows. In Section 2, we formalize our model and describe the estimators proposed by Pepe and Fleming (1991) and Carroll and Wand (1991). In Section 3, we derive the semiparametric efficient score and the semiparametric variance bound for our model. In Section 4, we propose a new class of estimators motivated by the form of the efficient score. In Section 5, we construct an adaptive semiparametric efficient estimator when Y_i and V_i are discrete and propose candidate adaptive semiparametric efficient estimators when Y_i and/or V_i contain continuous components. In Section 6, we present the result of a small simulation study comparing the Pepe-Fleming estimator with our semiparametric efficient estimator. In Section 7, we describe adaptive estimators based on optimal combinations of unbiased estimating functions which should have good efficiency properties even when V_i contains many continuous components. We conclude with a discussion.

2. THE MODEL AND PREVIOUSLY PROPOSED ESTIMATORS

We assume that the conditional distribution of Y_i given X_i and V_i is known up to a finite vector of unknown parameters, i.e.

$$f[Y|X, V] = f[Y|X, V; \alpha_0] \quad (1)$$

where $f[Y|X, V; \alpha]$ is a known density with respect to a measure μ that has

two continuous derivatives with respect to α and α_0 is a q -vector of unknown parameters lying in the interior of a fixed compact set. We have suppressed the i -subscript denoting the study subject. Our goal is to estimate α_0 when X is not always observed. Specifically we shall assume that there is an indicator variable Δ such that $\Delta = 1$ if an individual is in the validation sample, and, thus, X is observed, and $\Delta = 0$ if an individual is in the non-validation sample, whereas $W \equiv (Y, V)$ is always observed. In the measurement error example, we regard the surrogate Z as a component of V . We shall assume that $(\Delta_i, Y_i, X_i, V_i)$, $i = 1, \dots, n$, are independent and identically distributed, that X is missing at random given W (Rubin, 1976) and the probability of observing complete data is bounded away from 0, i.e.

$$f[\Delta | W, X] = f[\Delta | W], \quad (2a)$$

$$\pi(W) \equiv \text{pr}[\Delta = 1 | W] > \sigma > 0 \quad \text{for some } \sigma \text{ with probability 1.} \quad (2b)$$

Equations (1) and (2) characterize a semiparametric model indexed by the parameter $\alpha \in R^q$ and an 'infinite dimensional' nuisance parameter g with likelihood $\Pi_i L_i(\alpha, g)$, with

$$L(\alpha, g) \equiv f(V; g_1) f(\Delta | W; g_2) \{f\{Y|X, V; \alpha\} f(X|V; g_3)\}^\Delta \times \left\{ \int f(Y|x, V; \alpha) dF(x|V; g_3) \right\}^{1-\Delta}, \quad (3)$$

where $g = (g_1, g_2, g_3)$ with true value $g_0 = (g_{01}, g_{02}, g_{03})$. The parameters g_1, g_2 and g_3 take values in the set G_1 of marginal densities for V , the set G_2 of conditional densities for Δ given W and the set G_3 of conditional densities for X given V . $F(\cdot | V; g_3)$ is the distribution function corresponding to $f(\cdot | V; g_3)$. If $f(X|V)$ were known, the efficient estimator of α_0 would be the parametric maximum likelihood estimator $\hat{\alpha}_{ML}$ of α_0 that solves

$$0 = \sum_i \frac{\partial \{\ln L_i(\alpha; g_0)\}}{\partial \alpha} \equiv \sum_i S_{\alpha,i}(\alpha) = \sum_i \Delta_i S_{\alpha,i}^c(\alpha) + (1 - \Delta_i) S_{\alpha,i}^{\bar{c}}(\alpha), \quad (4)$$

where $S_{\alpha}^c(\alpha) = s_{\alpha}^c(Y, X, V, \alpha) \equiv \partial \{\ln f(Y|X, V; \alpha)\} / \partial \alpha$ is the score with respect to α for a subject with complete data (i.e. a validation sample member). Also, in equation (4),

$$S_{\alpha}^{\bar{c}}(\alpha) \equiv s_{\alpha}^{\bar{c}}(W, \alpha) \equiv \frac{\int s_{\alpha}^c(Y, V, x, \alpha) f(Y|x, V; \alpha) dF(x|V)}{\int f(Y|x, V; \alpha) dF(x|V)} \quad (5)$$

is the score for a non-validation sample member and $S_{\alpha}(\alpha)$ is the total score with respect to α . Note that $S_{\alpha}^{\bar{c}} = E[S_{\alpha}^c | W]$, where $S_{\alpha}^{\bar{c}} = S_{\alpha}^{\bar{c}}(\alpha_0)$, $S_{\alpha}^c = S_{\alpha}^c(\alpha_0)$ and $S_{\alpha} = S_{\alpha}(\alpha_0)$ are the scores evaluated at the truth.

When selection into the validation sample does not depend on Y as well as X , i.e.

$$f[\Delta | W, X] = f[\Delta | V] \quad (6)$$

and $f(X|V)$ is unknown, Carroll and Wand (1991) and Pepe and Fleming (1991) proposed estimating α_0 by $\hat{\alpha}^*$ that solves

$$0 = \sum_i \hat{S}_{\alpha,i}(\alpha) \equiv \sum_i \Delta_i S_{\alpha,i}^c(\alpha) + (1 - \Delta_i) \hat{S}_{\alpha,i}^c(\alpha) \quad (7)$$

where $\hat{S}_{\alpha}^c(\alpha)$ is defined by expression (5) when $dF(x|V)$ is replaced by $d\hat{F}(x|V)$, with $\hat{f}(x|V)$, a nonparametric density estimate of $f(x|V)$ based on the validation sample alone. For example, if V is discrete, Pepe and Fleming propose-

$$\hat{f}(x|V) = \sum_j \Delta_j I(V_j = V) I(X_j = x) / \sum_j \Delta_j I(V_j = V).$$

If V has continuous components, the kernel estimators proposed by Carroll and Wand are employed. If equation (6) holds, and data on non-validation sample members were *not* available, an efficient estimator of α_0 is the complete case estimator $\hat{\alpha}_{\text{val}}$ defined as the solution to $\sum_i \Delta_i S_{\alpha,i}^c(\alpha) = 0$. The estimators $\hat{\alpha}_{\text{val}}$ and $\hat{\alpha}^*$ may be inconsistent if equations (2) hold but equation (6) does not.

3. SEMIPARAMETRIC EFFICIENT SCORE

As in Cuzick (1992), the purpose of this section is to motivate our class of estimators and to provide a lower bound for the asymptotic variance of any regular estimator of α_0 . We begin by reviewing the theory of semiparametric efficiency bounds described in the survey paper of Newey (1990), the monograph of Bickel *et al.* (1993) and Begun *et al.* (1983). We then derive the efficient score and semiparametric variance bound for the semiparametric model defined by the restrictions equations (1) and (2). Again suppose that the data consist of n independent copies of an observed random vector and let $L(\alpha, g)$ be the likelihood for a single subject in a semiparametric model indexed by a parameter $\alpha \in R^q$ of interest and a nuisance parameter g taking values in some infinite dimensional parameter set. Define a regular parametric submodel to be a regular parametric model with parameters (α, η) and subject-specific likelihood contribution $L(\alpha, \eta)$ with true values (α_0, η_0) , where for each η the distribution $L(\alpha, \eta)$ equals a distribution $L(\alpha, g)$ allowed by the semiparametric model. A regular parametric model is a model whose square-root density is mean square differentiable with respect to η and has a non-singular information matrix (Bickel *et al.*, 1993). Define the nuisance tangent set τ to be the mean-squared closure of the linear span of all random vectors bS_η , where S_η is the score for η in some regular parametric submodel (typically, $S_\eta = \partial \{\ln L(\beta_0, \eta_0; Z)\} / \partial \eta$) and b is a constant matrix with q rows. We shall consider τ as a subset of the Hilbert space of $q \times 1$ random vectors H with inner product $E(H_1' H_2) < \infty$. The projection of any vector H on τ exists and is the unique vector $\Pi(H|\tau)$ in τ satisfying $E\{[H - \Pi(H|\tau)]' A\} = 0$ for all A in τ . The semiparametric variance bound for regular estimators of α_0 is the supremum of the Cramer-Rao variance bounds for α_0 over all regular parametric submodels and equals the inverse of the variance of $S_{\text{eff}} \equiv S_\alpha - \Pi(S_\alpha|\tau) \equiv \Pi[S_\alpha|\tau^\perp]$ where S_α is the score for α and τ^\perp is the orthogonal complement of τ . An estimator is regular if its convergence to its limiting distribution is locally uniform. A more precise definition of a regular estimator is given in Bickel *et al.* (1993). S_{eff} is called the efficient score (Begun *et al.*, 1983).

In Appendix A we prove the following proposition.

Proposition 1. In the semiparametric model characterized by equations (1) and (2), $S_{\text{eff}} = U(\phi_{\text{op}})$ where, for any function $\phi(w)$ taking values in R^q ,

$$U(\phi) \equiv U^{(1)} + U^{(2)}(\phi), \quad U^{(1)} \equiv \Delta S_{\alpha}^c - \Delta E[S_{\alpha}^c | X, V, \Delta = 1],$$

$$U^{(2)}(\phi) \equiv -\pi^{-1} \Delta E[(1 - \Delta) \phi(W) | X, V] + (1 - \Delta) \phi(W)$$

and $\phi_{\text{op}}(W)$ is the unique solution to the functional equation

$$\begin{aligned} \phi(W) = E^W[S_{\alpha}^c] - E^W\{E^{XV}[\pi(W)S_{\alpha}^c]/E^{XV}[\pi(W)]\} \\ - E^W\{E^{XV}[\{1 - \pi(W)\} \phi(W)]/E^{XV}[\pi(W)]\} \end{aligned} \quad (8)$$

where $E^{A|B}[C] \equiv E[C|A, B]$. When equation (6) is true, equation (8) simplifies to

$$\phi(W) = E^W[S_{\alpha}^c] - \left\{ \frac{1 - \pi(V)}{\pi(V)} \right\} E^W\{E^{XV}[\phi(W)]\}. \quad (9)$$

After the first draft of this paper had been written, we learned that an expression equivalent to equation (9) had been independently derived by Bickel *et al.* (1993). Further Hasminskii and Ibragimov (1983) proved that $U^{(2)}(\phi_{\text{op}})$ is the efficient score in the model characterized by equations (1), (2) and (6) when data on Y are not obtained for validation sample members ($\Delta = 1$).

The following corollary says that prior knowledge concerning $\pi(W)$ does not increase the efficiency with which α_0 can be estimated.

Lemma 1. $S_{\text{eff}} = U(\phi_{\text{op}})$ is also the efficient score in the three additional semiparametric models (a)-(c) that impose, in addition to equations (1) and (2), the conditions

- (a) $\pi(W)$ is completely known,
- (b) equation (6) is true and
- (c) $\pi(W)$ is known up to an unknown parameter θ , i.e.

$$\pi(W) = \pi(W; \theta_0) \quad (10)$$

where θ_0 is an unknown t -dimensional parameter and $\pi(W; \theta)$ is a known function of W taking values in $(0, 1]$.

We note that, when data on X are missing by happenstance and W is multi-dimensional, an investigator will often choose to specify a parametric model (10) for the unknown missingness process.

4. A CLASS OF ESTIMATORS

The form of the efficient score suggests that, given a correctly specified model $\pi(W; \theta)$ for $\pi(W)$, we consider estimators $\hat{\alpha}(\phi, \hat{\theta})$ solving

$$0 = n^{1/2} \bar{U}(\alpha, \phi, \hat{\theta}) = n^{-1/2} \sum_i U_i(\alpha, \phi, \hat{\theta})$$

where $\hat{\theta}$ satisfies

$$\sum_i S_{\theta,i}(\hat{\theta}) = 0,$$

$$S_\theta(\theta) = \partial(\ln[\pi(W; \theta)^\Delta \{1 - \pi(W; \theta)\}^{1-\Delta}]) / \partial\theta,$$

$$U(\alpha, \phi, \theta) = U^{(1)}(\alpha, \theta) + U^{(2)}(\alpha, \phi, \theta),$$

$$U^{(1)}(\alpha, \theta) = \Delta S_\alpha^c(\alpha) - \Delta E_{\alpha, \theta}[S_\alpha^c(\alpha) | X, V, \Delta = 1],$$

$$U^{(2)}(\alpha, \phi, \theta) = \{-\Delta / \pi(W; \theta)\} E_\alpha[\{1 - \pi(W; \theta)\} \phi(W) | X, V] + (1 - \Delta) \phi(W)$$

and, by equations (2), for any H ,

$$\Delta E_{\alpha, \theta}[H | X, V, \Delta = 1] = \frac{\Delta E_\alpha[\pi(W; \theta)H | X, V]}{E_\alpha[\pi(W; \theta) | X, V]}$$

with

$$E_\alpha[h(Y, X, V) | X, V] = \int h(y, X, V) f(y | X, V; \alpha) d\mu(y). \quad (11)$$

We shall suppress the θ -argument when $\theta = \theta_0$. When equation (6) holds and $\theta = \theta_0$, $U^{(1)}(\alpha)$ and $U^{(2)}(\alpha, \phi)$ simplify to

$$U^{(1)}(\alpha) = \Delta S_\alpha^c(\alpha)$$

and

$$U^{(2)}(\alpha, \phi) = -\Delta \left\{ \frac{1 - \pi(V)}{\pi(V)} \right\} E_\alpha^{XV}[\phi(W)] + (1 - \Delta) \phi(W),$$

since $E_\alpha[S_\alpha^c(\alpha) | X, V] = 0$ by the conditional mean zero property of conditional scores; this version of the estimating function $U^{(2)}(\alpha, \phi)$ was originally proposed by Hasminskii and Ibragimov (1983). Our $U^{(2)}(\alpha, \theta)$ is the natural generalization of the Hasminskii and Ibragimov version to the setting in which missingness can depend on Y as well as V . In Appendix B, we prove the following proposition.

Proposition 2. Subject to the regularity conditions provided in Appendix B, under equations (1), (2) and (10)

- (a) with probability approaching 1, the solutions $\hat{\alpha}(\phi, \hat{\theta})$ and $\hat{\alpha}(\phi) \equiv \hat{\alpha}(\phi, \theta_0)$ to $\bar{U}(\alpha, \phi, \theta) = 0$ and $\bar{U}(\alpha, \phi, \theta_0) = 0$ exist and are unique,
- (b) $n^{1/2}\{\hat{\alpha}(\phi, \hat{\theta}) - \alpha_0\}$ and $n^{1/2}\{\hat{\alpha}(\phi) - \alpha_0\}$ are regular asymptotically normal estimators with asymptotic mean 0 and asymptotic variances $I(\phi)^{-1}C(\phi)\{I(\phi)'\}^{-1}$ and $I(\phi)^{-1}\Omega(\phi)\{I(\phi)'\}^{-1}$ respectively, where

$$I(\phi) = E[\partial U(\alpha_0, \phi) / \partial \alpha'] = -E[U(\phi)S_\alpha'], \quad \Omega(\phi) = E[U(\phi)^{\otimes 2}]$$

with $U(\phi) \equiv U(\alpha_0, \phi)$, $C(\phi) = E[\text{Resid}\{U(\phi), S_\theta\}^{\otimes 2}]$, $S_\theta = S_\theta(\theta_0)$, $\text{Resid}(A, B) \equiv A - E(AB')E(BB')^{-1}B$ is the residual from the population least squares regression A on B ,

- (c) $I(\phi)$, $\Omega(\phi)$ and $C(\phi)$ can be consistently estimated by $\hat{I}(\phi)$, $\hat{\Omega}(\phi)$ and $\hat{C}(\phi)$ where, with $\hat{\alpha} \equiv \hat{\alpha}(\phi, \hat{\theta})$,

$$\hat{I}(\phi) \equiv \hat{I}(\hat{\alpha}, \phi, \hat{\theta}) \equiv -n^{-1} \sum_i U_i(\hat{\alpha}, \phi, \hat{\theta}) S_{\alpha, i}(\hat{\alpha})',$$

$\hat{\Omega}(\phi) = n^{-1} \sum_i \hat{U}_i(\phi)^{\otimes 2}$ with $\hat{U}_i(\phi) = U_i(\hat{\alpha}, \phi, \hat{\theta})$, $\hat{C}(\phi) = n^{-1} \sum_i \hat{\text{Resid}}\{\hat{U}_i(\phi), S_{\theta, i}(\hat{\theta})\}^{\otimes 2}$ and $\hat{\text{Resid}}(A_i, B_i)$ is the residual for subject i from the least squares regression of A_i and B_i , $i = 1, \dots, n$, and

- (d) $\text{var}^A[n^{1/2}\{\hat{\alpha}(\phi, \hat{\theta}) - \alpha_0\}] \geq \text{var}(S_{\text{eff}})^{-1}$ with equality if and only if $\phi(W) = \phi_{\text{op}}(W)$ where $A \geq B$ means that $A - B$ is non-negative definite.

A key to the consistency of $\hat{\alpha}(\phi, \hat{\theta})$, under our regularity conditions, is that $U(\alpha, \phi)$ is an unbiased estimating function for α , i.e. $E_{\alpha}[U(\alpha, \phi)] = 0$ for all α . Part (b) implies that, when $\phi(W) \neq \phi_{\text{op}}(W)$, we may increase efficiency by estimating the parameters θ_0 of the missingness model (10) even if θ_0 is known. This efficiency advantage is attributable to the fact that the missingness process is an 'ancillary process' in that the maximum likelihood estimator of α in any parametric submodel with variation independent parameters does not depend on model (10) for missingness. Robins *et al.* (1992), Robins (1992) and Robins and Morgenstern (1987) considered other settings in which estimating the known parameters of such an ancillary process increases the efficiency of an inefficient estimator. Robins and Morgenstern (1987) offered the following heuristic explanation. Estimating the known parameters of an ancillary process effectively conditions on a function of an exact or approximate ancillary statistic, resulting in inferences that are closer to an efficient likelihood-based inference which conditions on all exact and approximate ancillaries.

To calculate $n^{1/2} \bar{U}(\alpha, \phi, \hat{\theta})$ requires that, for each subject, we evaluate integrals of the form of equation (11). If, as with examples studied by Pepe and Fleming (1991) and Carroll and Wand (1991), Y is discrete, then equation (11) is a simple sum. If Y has continuous components this will require a one-dimensional or multidimensional numerical integration. McFadden (1989) and Pakes and Pollard (1989) considered unbiased simulation estimates of equation (11). We emphasize that to compute $\hat{\alpha}(\phi, \hat{\theta})$ for a given function $\phi(W)$ we have no need to estimate the unknown law of X given V . Pepe and Fleming and Carroll and Wand considered only the case in which $\pi(W) = \pi$, i.e. selection into the validation sample is completely at random. In this case, the estimator $\hat{\alpha}^*$ solving equation (7) is asymptotically equivalent to the estimator $\hat{\alpha}(\phi^*)$ in our class where $\phi^*(W) \equiv S_{\alpha}^{\xi}$. To see this, we note that Pepe and Fleming and Carroll and Wand both show in their appendixes that

$$n^{1/2}(\hat{\alpha}^* - \alpha_0) = \{-E[\partial S_{\alpha}(\alpha_0)/\partial \alpha']\}^{-1} n^{1/2} \bar{U}(\alpha_0, \phi^*) + o_p(1).$$

But $E[\partial S_{\alpha}(\alpha_0)/\partial \alpha']^{-1} = I(\phi^*)$. Further, when $\pi(W) = \pi$, the complete case estimator $\hat{\alpha}_{\text{val}}$ is the estimator $\hat{\alpha}(\phi_{\text{val}})$ in our class with ϕ_{val} being the function that is identically 0.

Suppose that $\pi(W)$ is known. Even then, since $\phi_{\text{op}}(W)$ depends on the unknown joint distribution of the data, $\hat{\alpha}(\phi_{\text{op}})$ is not a feasible 'estimator'. Therefore, we might attempt to construct a feasible adaptive semiparametric efficient estimator $\hat{\alpha}(\hat{\phi}_{\text{op}})$ by replacing the unknown $\phi_{\text{op}}(W)$ by a consistent estimate $\hat{\phi}_{\text{op}}(W)$. In the following section we formally carry out this construction in the special case in which Y and V are discrete. We also propose candidate semiparametric efficient estimators outside this special case.

5. SEMIPARAMETRIC EFFICIENT ESTIMATOR

Suppose that V and Y are discrete with \mathbf{V} and \mathbf{Y} levels respectively. Let $\hat{\alpha}^{(0)}$ be a preliminary $n^{1/2}$ -consistent estimator of α_0 . Iteratively for $j = 1, 2, 3, \dots$, set

$$\hat{\alpha}^{(j)} = \hat{\alpha}^{(j-1)} - \hat{I}(\hat{\alpha}^{(j-1)}, \hat{\phi}^{(j-1)})^{-1} \bar{U}(\hat{\alpha}^{(j-1)}, \hat{\phi}^{(j-1)}) \quad (12)$$

where $\hat{\phi}^{(j-1)}(W)$ is the unique solution to equation (8) when equation (8) is modified so that $S_{\alpha}^c, E^{XV}[\]$ and $E^W[\]$ are replaced by $S_{\hat{\alpha}^{(j-1)}}^c, E_{\hat{\alpha}^{(j-1)}}^{XV}[\]$ and $\hat{E}_{\hat{\alpha}^{(j-1)}}^W[\]$, where, with $w \equiv (y, v)$, for any random variable $H \equiv h(Y, X, V)$,

$$\hat{E}_{\alpha}^w[H] = \frac{\sum_i h(y, X_i, v) f(y|X_i, v; \alpha) \Delta_i \pi(W_i)^{-1} I(V_i = v)}{\sum_i f(y|X_i, v; \alpha) \Delta_i \pi(W_i)^{-1} I(V_i = v)} \tag{13}$$

Proposition 3. Suppose that $n^{1/2}(\hat{\alpha}^{(0)} - \alpha_0)$ is $O_p(1)$, Y and V are discrete and equations (1) and (2) hold. Then for $j = 1, 2, \dots, n^{1/2}(\hat{\alpha}^{(j)} - \alpha_0)$ is a regular, asymptotically normal estimator with mean 0 and asymptotic variance $\text{var}(S_{\text{eff}})^{-1}$ which can be consistently estimated by $[n^{-1} \sum_i U_i(\hat{\alpha}^{(j)}, \hat{\phi}^{(j-1)}) U_i(\hat{\alpha}^{(j)}, \hat{\phi}^{(j-1)})']^{-1}$. Thus, for each $j \geq 1$, $\hat{\alpha}^{(j)}$ is a semiparametric efficient estimator of α_0 .

Proof. When V and Y are discrete with \mathbf{V} and \mathbf{Y} levels, any function of ϕ can be identified with the $q \times \mathbf{VY}$ matrix-valued parameter that has columns consisting of the q -vectors $\phi(w)$. Thus $U(\alpha_0, \phi_{\text{op}})$ is a random function of the unknown q -dimensional parameter α_0 and the $q\mathbf{VY}$ -dimensional parameter ϕ_{op} . Proposition 3 then follows by a mean value expansion of $\bar{U}(\hat{\alpha}^{(j-1)}, \hat{\phi}^{(j-1)})$ around $(\alpha_0, \phi_{\text{op}})$ using

- (a) the continuity of the derivatives of $U(\alpha, \phi)$ with respect to α and ϕ ,
- (b) that, by equations (1) and (2), $E[\partial U(\alpha_0, \phi_{\text{op}})/\partial \phi'] = 0$,
- (c) that $\hat{E}_{\alpha}^w(H)$ is $n^{1/2}$ consistent for $E_{\alpha}^w(H)$ and
- (d) that $E[U(\phi_{\text{op}})^{\otimes 2}] = -E[\partial U(\alpha_0, \phi_{\text{op}})/\partial \alpha']$.

We now provide an explicit computational formula for $\hat{\phi}^{(j)}(W)$ that exploits the fact that, when Y is discrete, equation (8) is a finite dimensional matrix equation. For a fixed value v of the covariate V and a fixed $l \in \{1, \dots, q\}$, the l th component of the function $\hat{\phi}^{(j)}(w) = (\hat{\phi}_1^{(j)}(w), \dots, \hat{\phi}_q^{(j)}(w))'$ can be represented by a \mathbf{Y} -vector, $\hat{\phi}^{\dagger}$ say, where the dependence on j, v and l has been suppressed, i.e. $\hat{\phi}^{\dagger} = (\hat{\phi}_l^{(j)}(y_1, v), \dots, \hat{\phi}_l^{(j)}(y_Y, v))'$. Similarly, for each fixed l and v , the difference of the first two terms on the right-hand side of equation (8) is a \mathbf{Y} -vector, say, \hat{Z} . Let \hat{Z} be Z when α_0, E^{XV} and E^W are replaced by their estimates $\hat{\alpha}^{(j)}, E_{\hat{\alpha}^{(j)}}^{XV}$ and $\hat{E}_{\hat{\alpha}^{(j)}}^W$. Finally, for each fixed v and j , define \hat{m} to be the $\mathbf{Y} \times \mathbf{Y}$ matrix with (y^*, y) entry $\hat{E}_{\hat{\alpha}^{(j)}}^{y^*y}\{h(y, X, v)\}$, where

$$h(y, X, v) = \frac{f(y|X, v; \hat{\alpha}^{(j)}) \{1 - \pi(y, v)\}}{E_{\hat{\alpha}^{(j)}}^{XV}\{\pi(Y, v)\}}$$

and let I be the $\mathbf{Y} \times \mathbf{Y}$ identity matrix. Lemma 2 in Appendix A implies that $(I + \hat{m})$ is invertible with probability 1. Set $\hat{\phi}^{\dagger} = (I + \hat{m})^{-1} \hat{Z}$. Hence $\hat{\phi}^{(j)}$ is easily computed by repeating the above calculation for each of the $q\mathbf{V}$ values of (l, v) .

Suppose now that Y is still discrete but V is a univariate continuous covariate. We can continue to calculate $\hat{\phi}^{\dagger}, \hat{\alpha}^{(j)}$ and $\hat{\phi}^{(j)}$ as above except with

$$\hat{E}_{\alpha}^w[H] = \frac{\sum_i h(y, X_i, v) f(y|X_i, v; \alpha) \Delta_i \pi(W_i)^{-1} K\{(V_i - v)/h\}}{\sum_i f(y|X_i, v; \alpha) \Delta_i \pi(W_i)^{-1} K\{(V_i - v)/h\}} \tag{14}$$

where $K(\)$ is a twice-differentiable kernel function, symmetric around 0 such that

$\int K(x) dx = 1$ and h is a bandwidth. It can be shown that $\hat{\alpha}^{(j)}$ is semiparametric efficient under regularity conditions and with the bandwidth appropriately chosen. When V has two continuous components, a bivariate kernel function may be used in equation (14). When V has more than two continuous components the estimators discussed in Section 7 should be considered.

Suppose next that Y has one or more continuous components and V has two or fewer continuous components. Then, for each value v of V , equation (8) is no longer a finite dimensional matrix equation. Furthermore, the solution $\hat{\phi}^{(j-1)}$ to the modified equation (8) will no longer exist in closed form. However, the equation can be numerically solved by any of the several methods described in chapter 12 of Kress (1989), since the operator acting on $\phi(W)$ in equation (8) is, for fixed V , the identity minus a compact operator.

6. SIMULATION STUDY

To compare the performance of the semiparametric efficient estimators $\hat{\alpha}^{(j)}$ with that of $\hat{\alpha}^*$ and $\hat{\alpha}_{\text{val}}$, we conducted a small simulation study. As in example 2 of Pepe and Fleming (1991), for each of n observations we generated a normally distributed covariate $X \sim N(0, 1)$ and a binary outcome variable Y from the logistic model

$$f[Y = 1 | X, V; \alpha_0] = \frac{\exp(\alpha_{0,0} + \alpha_{0,1}X)}{1 + \exp(\alpha_{0,0} + \alpha_{0,1}X)}. \quad (15)$$

An independent additive normal measurement error ε generated the dichotomous surrogate $V = I[X + \varepsilon > 0]$, $\varepsilon \sim N(0, 1)$. Subjects were randomly selected into the validation sample with probability π . We set the sample size n to 2000, rather than to 200 or 100 as in Pepe and Fleming (1991) because

- (a) validation studies are most commonly performed within rather large studies and
- (b) we wished to compare the performance of $\hat{\alpha}^*$, $\hat{\alpha}_{\text{val}}$ and the $\hat{\alpha}^{(j)}$ at sample sizes at which our asymptotic results apply.

Pepe and Fleming still found residual small sample bias in $\hat{\alpha}^*$ at a sample size of 200.

Table 1 gives, for $\hat{\alpha} \in \{\hat{\alpha}_{\text{val}}, \hat{\alpha}^* \text{ and } \hat{\alpha}_{\text{eff}}\}$, sample averages and actual coverage rates of 90% univariate confidence intervals for $\alpha_{0,1}$ based on $\hat{\alpha} \pm 1.64[n^{-1} \hat{\text{var}}^A \{n^{1/2}(\hat{\alpha} - \alpha_0)\}]^{1/2}$ based on 200 simulated data sets: $\hat{\alpha}_{\text{eff}}$ is $\hat{\alpha}^{(j)}$ iterated (i.e. j is increased) until the convergence criterion used for $\hat{\alpha}_{\text{val}}$ and $\hat{\alpha}^*$ was met; $\hat{\alpha}^{(0)}$ was set equal to $\hat{\alpha}_{\text{val}}$. The estimated asymptotic relative efficiencies (AREs) are calculated as the square of the ratio of the interquartile ranges for $\hat{\alpha}_{\text{eff}}$ to those of $\hat{\alpha}_{\text{val}}$ and $\hat{\alpha}^*$.

Reading from Table 1, we observe that, as previously noted by Pepe and Fleming (1991), for each π , the ARE of $\hat{\alpha}^*$ is 1.0 when $\alpha_{0,1} = 0$, i.e. when Y and X are independent. The ARE decreases as $\alpha_{0,1}$ increases. In contrast the ARE of $\hat{\alpha}_{\text{val}}$ is a minimum at $\alpha_{0,1} = 0$ and increases as $\alpha_{0,1}$ increases. At $\alpha_{0,1} = 2$ and $\pi = 0.05$, $\hat{\alpha}_{\text{val}}$ is more efficient than $\hat{\alpha}^*$. However, since it would be unusual for $\alpha_{0,1}$ to be 2 when X is distributed as $N(0, 1)$, our simulation results suggest that, in most realistic settings, the efficiency of the Pepe-Fleming and Carroll-Wand estimator is quite good.

TABLE 1
Simulation study of estimators of $\alpha_{0,1}$

π	$\alpha_{0,0}$	$\alpha_{0,1}$	Results for $\hat{\alpha}_{val}$			Results for $\hat{\alpha}^*$			Results for $\hat{\alpha}_{eff}$	
			Monte Carlo average	Actual coverage of 90% intervals	Estimated ARE	Monte Carlo average	Actual coverage of 90% intervals	Estimated ARE	Monte Carlo average	Actual coverage of 90% intervals
0.05	-1	0	-0.010	91.5	0.26	0.0005	93	1.0	0.0004	93
0.05	-1	1	1.01	90.5	0.57	1.01	91.5	0.89	1.01	93
0.05	-1	2	2.04	89.5	0.92	2.02	83	0.74	2.01	88.5
0.1	-1	0	-0.001	89.5	0.21	0.006	87.5	1.0	0.006	88.0
0.1	-1	1	1.04	87.5	0.48	1.01	90	0.89	1.01	89.5
0.1	-1	2	2.03	90	0.82	2.00	88.5	0.82	2.01	90
0.2	-1	0	0.020	92	0.28	0.013	89	1.00	0.012	90.5
0.2	-1	1	1.01	92	0.58	1.00	92.5	0.82	1.00	90.5
0.2	-1	2	2.07	92	0.66	2.02	85	0.75	2.03	93

7. OPTIMAL COMBINATION OF ESTIMATING FUNCTIONS

We can construct estimators with good efficiency properties even when V has many continuous components by optimally combining unbiased estimating functions. These estimators do not require that we estimate the law of X given V . For simplicity, we describe this approach where $\pi(W)$ is known. Let $\psi_1(W), \psi_2(W), \dots$ be a sequence of real-valued functions of W that is complete in mean square, i.e. for any function $\psi(W)$ with finite second moment there are constants $\gamma_j, j \in (1, 2, \dots)$, such that

$$\lim_{k \rightarrow \infty} \left[\int \left\{ \psi(w) - \sum_{j=1}^k \gamma_j \psi_j(w) \right\}^2 dF(w) \right] = 0.$$

When W has continuous components, sequences of polynomials (i.e. power series) are known to be complete. However, in practice, as discussed by Newey (1992), we should first transform any continuous component W_i of W by using a bounded transformation such as $\exp W_i / (1 + \exp W_i)$ to downweight the influence of outlying values of W_i . Another approach is to use splines rather than polynomials. Let $\psi^k(W) = (\psi_1(W), \dots, \psi_k(W))'$, $k > q$. Also let $T^k(\alpha)$ be defined exactly like $U^{(2)}(\alpha, \theta_0)$ except with the k -vector $\psi^k(W)$ replacing $\phi(W)$. Then define $\tilde{\phi}^k(W) = \tilde{\Omega}^{-1} \psi^k(W)$ and $\phi^k(W) = \Omega^{-1} \psi^k(W)$ where $J' = n^{-1} \sum_i T_i^k(\hat{\alpha}) S_{\alpha_i}(\hat{\alpha})'$, $\tilde{\Omega} = n^{-1} \sum_i T_i^k(\hat{\alpha}) T_i^k(\hat{\alpha})'$, $J' = E[T^k(\alpha_0) S_{\alpha'}]$, $\hat{\alpha}$ is a preliminary $n^{1/2}$ -consistent estimator and $\Omega = E[T^k(\alpha_0) T^k(\alpha_0)']$. These functions correspond to the optimal linear combinations of $\psi^k(W)$. Specifically, under the regularity conditions of proposition 2, results of Hansen (1982) on optimal linear combinations of unbiased estimating functions imply that

- $\hat{\alpha}(\tilde{\phi}^k)$ and $\hat{\alpha}(\phi^k)$ are asymptotically equivalent and
- $\hat{\alpha}(\phi^k)$ has asymptotic variance less than or equal to $\hat{\alpha}(\phi)$ for any $\phi(W)$ of the form $c^k \psi^k(W)$ with c^k a $q \times k$ constant matrix.

In practice we would use only a few (e.g. $k = 3$ or $k = 4$) fixed functions to avoid

the poor finite sample performance that would result if $\tilde{\phi}^k(W)$ depended on a large number of estimated linear combination coefficients. Newey (1992) discussed the possibility of selecting the number of functions k by cross-validation. The estimator $\hat{\alpha}(\tilde{\phi}^k)$ will be semiparametric efficient if and only if $\phi_{\text{op}}(W) = \phi^k(W)$. Since the sequence $\psi_1(W), \psi_2(W), \dots$ is complete, it follows that $\phi^k(W)$ should converge to $\phi_{\text{op}}(W)$ in mean square as $k \rightarrow \infty$. This suggests that, if we let k increase with the sample size n at an appropriate rate, $\hat{\alpha}(\tilde{\phi}^k)$ should be semiparametric efficient. Proposition 4 provides the necessary regularity conditions for this result. Although proposition 4 is of theoretical interest, in practice as described above we would choose k to be small, possibly selected by cross-validation.

Consider the one-step version $\tilde{\alpha}$ of the estimator $\hat{\alpha}(\tilde{\phi}^k)$ given by

$$\tilde{\alpha} = \hat{\alpha} + \hat{M}^{-1} \sum_{i=1}^n \hat{U}_i / n$$

with $\hat{M} = n^{-1} \sum_{i=1}^n \hat{U}_i S_{\alpha,i}(\hat{\alpha})'$, $\hat{U}_i = U_i^{(1)}(\hat{\alpha}) + U_i^{(2)}(\hat{\alpha}, \tilde{\phi}^k)$ and $\hat{\alpha}$ a preliminary $n^{1/2}$ -consistent estimator.

Proposition 4. Suppose that

- (a) equations (1) and (2) hold,
- (b) for each k there are $q \times k$ matrices c^k such that $\lim_{k \rightarrow \infty} \{E[\|\phi_{\text{op}}(W) - c^k \psi^k(W)\|^2]\} = 0$ with $\|A\|^2 = A'A$,
- (c) $\pi(W)$ is bounded away from 1,
- (d) the smallest eigenvalue of $E[\psi^k(W)\psi^k(W)']$ is bounded below by ck^{-ck} for a constant $c > 0$ as $k \rightarrow \infty$ and $k \ln k / \ln n \rightarrow 0$,
- (e) $\psi_k(W)$, $k = 1, 2, \dots$ are uniformly bounded,
- (f) $f(y|x, v; \alpha)$ is twice continuously differentiable in α and there is a neighbourhood N of α_0 with both $\int \sup_{\alpha \in N} \|\partial f(y|X, V; \alpha) / \partial \alpha\| d\mu(y)$ and $\int \sup_{\alpha \in N} \|\partial^2 f(y|X, V; \alpha) / \partial \alpha \partial \alpha'\| d\mu(y)$ are bounded and
- (g) S_α satisfies the regularity conditions of appendix B.

Then

$$n^{1/2}(\tilde{\alpha} - \alpha_0) \xrightarrow{d} N(0, E[S_{\text{eff}} S_{\text{eff}}']^{-1}).$$

The minimum eigenvalue condition in (d) is similar to a condition used in Newey (1988, 1992) to show efficiency of similar estimators in other models. For example, it will be satisfied if the elements of $\psi^k(W)$ consist of distinct multivariate powers of W and the distribution of W has an absolutely continuous component with density bounded away from 0 on an open set.

Note that proposition 4 describes an approach to constructing a semiparametric efficient estimator that does not require that we explicitly estimate the law of X given V .

8. FINAL REMARKS

We have assumed that the missing covariate X is univariate. In fact, our results continue to hold for multivariate X provided that, for each subject, X is either completely observed or completely unobserved. Robins *et al.* (1994) and Robins and Rotnitzky (1992) have suggested how the methods of this paper can be extended to allow different components of X to be missing on different individuals.

When Y is Bernoulli, Breslow and Cain (1988), Breslow and Zhao (1989), Flanders and Greenland (1991), Kalbfleisch and Lawless (1988), Zhao and Lipsitz (1992), Manski and McFadden (1981) and Manski and Lerman (1977) have also proposed inefficient estimators for the model considered in this paper. Furthermore, Cosslett (1981) proposed an efficient but computationally challenging estimator. Robins *et al.* (1994) showed how each of the inefficient estimators can be modified so that they become semiparametric efficient.

Robins *et al.* (1994) considered estimation in a parametric model for the conditional mean of Y given (X, V) with data missing at random; when Y is Bernoulli, this model coincides with the semiparametric model studied in this paper. Finally, Robins and Rotnitzky (1992), Robins *et al.* (1994) and Robins (1995) discuss estimation in arbitrary semiparametric models with data missing or coarsened at random and the probability of observing complete data bounded away from 0. The model studied in this paper is an example of such a model.

ACKNOWLEDGEMENTS

This research was partially supported by National Institutes of Health grants 2 P30 ES00002, R01AI32475, R01-ES03405 and K04-ES00180 and the National Science Foundation.

APPENDIX A

Proof of proposition 1. We first calculate the nuisance tangent set τ and then S_{eff} . By a purely formal differentiation of the likelihood (3) with respect to g_1 , g_2 and g_3 , we expect the nuisance tangent set τ to be the linear space given by the closure of

$$\{a(Y, X, V, \Delta) = a_1(V) + a_2(\Delta, W) + a_3(Y, X, V, \Delta)\}, \quad (16)$$

where each of the functions $a_1(\cdot)$, $a_2(\cdot)$ and $a_3(\cdot, \cdot, \cdot, \cdot)$ take values in R^q and represent either conditional or unconditional scores (Bickel *et al.*, 1993). As conditional and unconditional scores, they satisfy the restrictions $E[a_1(V)] = 0$, $E[a_2(\Delta, W)|W] = 0$ and $a_3(Y, X, V, \Delta) = \Delta a_3^c(X, V) + (1 - \Delta)E^W[a_3^c(X, V)]$ for some $a_3^c(X, V)$ with $E[a_3^c(X, V)|V] = 0$. We can calculate that $E[A_1 A_2] = E[A_1 A_3] = E[A_2 A_3] = E[S_\alpha A_1] = E[S_\alpha A_2] = 0$ where, for example, $A_2 = a_2(\Delta, W)$. Thus S_{eff} is the residual from the projection of S_α on τ_3 . Furthermore τ_1 , τ_2 and τ_3 are mutually orthogonal where $\tau_1 = \{A_1\}$, $\tau_2 = \{A_2\}$ and $\tau_3 = \{A_3\}$. Let $S_\alpha - S_{\text{eff}} = \hat{A}_3 = \Delta \hat{A}_3^c + (1 - \Delta)E^W[\hat{A}_3^c]$ so that

$$S_{\text{eff}} = \Delta(S_\alpha^c - \hat{A}_3^c) + (1 - \Delta)E^W(S_\alpha^c - \hat{A}_3^c). \quad (17)$$

The key step in our proof is to define

$$Q_{\text{eff}} \equiv \pi(W)\{S_\alpha^c - \hat{A}_3^c\} + \{1 - \pi(W)\}E^W(S_\alpha^c - \hat{A}_3^c). \quad (18)$$

Our strategy is to prove the following three lemmas which together imply proposition 1. Redefine $\phi_{\text{op}}(W) \equiv E^W(Q_{\text{eff}})$.

Lemma 2. $\phi_{\text{op}}(W)$ solves equation (8).

Lemma 3. Equation (8) has a unique solution.

Lemma 4. $S_{\text{eff}} = U(\phi_{\text{op}})$.

We shall first prove the following preliminary lemma.

Lemma 5. Equations (1) and (2) imply

$$E^{XV}[Q_{\text{eff}}] = 0 \quad \text{almost surely.} \quad (19)$$

Proof. Equation (19) will hold if we can show that

$$E[Q'_{\text{eff}}A_1] = E[Q'_{\text{eff}}A_3^c] = 0 \quad (20)$$

for all A_1 and A_3^c . To see why, note that $E^{XV}[Q_{\text{eff}}] = (E^{XV}[Q_{\text{eff}}] - E^V[Q_{\text{eff}}]) + E^V[Q_{\text{eff}}] = A_3^c + A_1^*$, say, since $E[Q_{\text{eff}}] = 0$. Thus, equation (20) would imply $0 = E[Q'_{\text{eff}}E^{XV}(Q_{\text{eff}})] = E[E^{XV}(Q_{\text{eff}})'E^{XV}(Q_{\text{eff}})]$ from which equation (19) would follow. To establish equation (20) note that

$$S_{\text{eff}} = \Delta \pi(W)^{-1}Q_{\text{eff}} - \{\Delta - \pi(W)\} \pi(W)^{-1} \phi_{\text{op}}(W), \quad (21)$$

since, from the definition of Q_{eff} ,

$$\phi_{\text{op}}(W) = E^W[S_{\alpha}^c - \dot{A}_3^c]. \quad (22)$$

Now equations (21) and (2) imply that $S_{\text{eff}} - \Delta \pi(W)^{-1}Q_{\text{eff}}$ is in τ_2 , so that, for all A_3 , $0 = E[(S_{\text{eff}} - \Delta \pi(W)^{-1}Q_{\text{eff}})'A_3]$ by the orthogonality of τ_2 and τ_3 . But $E[S'_{\text{eff}}A_3] = 0$ since by definition S_{eff} is orthogonal to τ_3 . Hence, $0 = E[\Delta \pi(W)^{-1}Q'_{\text{eff}}A_3] = E[\Delta \pi(W)^{-1}Q'_{\text{eff}}A_3^c] = E[Q'_{\text{eff}}A_3^c]$, where the last equality is by equation (2). By an identical argument $E[Q'_{\text{eff}}A_1] = 0$.

Remark. Lemma 5 is a special case of lemma (A.6) in Robins *et al.* (1994), since $Q_{\text{eff}} = E[S_{\text{eff}}|L]$, and $S_{\text{eff}} \in \tau^\perp$ where $L \equiv (W, X)$.

Proof of lemma 2. Use equation (19) to set the conditional expectation of the right-hand side of equation (18) with respect to (X, V) equal to 0 and then solve for \dot{A}_3^c to obtain

$$\dot{A}_3^c = E^{XV}[\pi(W)]^{-1}(E^{XV}[\pi(W)S_{\alpha}^c] + E^{XV}[\{1 - \pi(W)\} \phi_{\text{op}}(W)]). \quad (23)$$

Now take the conditional expectation of both sides of equation (23) with respect to W , and then use equation (22) to replace $E^W[\dot{A}_3^c]$ by $E^W(S_{\alpha}^c) - \phi_{\text{op}}(W)$. On rearranging, we obtain equation (8).

Proof of lemma 3. Let $\phi^{(1)}(W)$ and $\phi^{(2)}(W)$ be two solutions of equation (8). We wish to show that $\phi^*(W) = \phi^{(1)}(W) - \phi^{(2)}(W) = 0$ almost surely. Now $\phi^*(W)$ satisfies

$$0 = \phi^*(W) + E^W(E^{XV}[\{1 - \pi(W)\} \phi^*(W)]/E^{XV}[\pi(W)]) \equiv \phi^*(W) + T^*(W).$$

If $1 - \pi(W) = 0$ almost surely, obviously $\phi^*(W) = 0$. Suppose that $\pi(W) \neq 1$ with non-zero probability. Then

$$\begin{aligned} 0 &= E[\{1 - \pi(W)\} \phi^*(W)' \{\phi^*(W) + T^*(W)\}] = E[\{1 - \pi(W)\} \phi^*(W)' \phi^*(W)] \\ &\quad + E(\{1 - \pi(W)\} \phi^*(W)' E^{XV}[\{1 - \pi(W)\} \phi^*(W)]/E^{XV}[\pi(W)]) \\ &= E[\{1 - \pi(W)\} \phi^*(W)' \phi^*(W)] + E(E^{XV}[\{1 - \pi(W)\} \phi^*(W)]' \\ &\quad \times E^{XV}[\{1 - \pi(W)\} \phi^*(W)]/E^{XV}[\pi(W)]) > 0 \end{aligned}$$

unless $\phi^*(W) = 0$ almost surely by $\pi(W) > \sigma$.

Proof of lemma 4. First use equations (18) and (22) to write

$$Q_{\text{eff}} = \pi(W)(S_{\alpha}^c - \dot{A}_3^c) + \{1 - \pi(W)\} \phi_{\text{op}}(W). \quad (24)$$

Then substitute the right-hand side of equation (23) for \dot{A}_3^c in equation (24) and substitute the result for Q_{eff} in expression (21). Algebraic simplification then shows that the right-hand side of the substituted version of equation (21) equals $U(\phi_{\text{op}})$, proving the lemma.

Proof of lemma 1. Since the projection of S_{α} on τ_2 is 0, the semiparametric variance bounds for all three semiparametric models of lemma 1 are as in proposition 1, since the models only differ in their restrictions on the nuisance scores in τ_2 .

APPENDIX B

Let $H(\gamma)' = (U(\alpha, \phi, \theta)', S_\theta(\theta)')$, $\gamma' = (\alpha', \theta')$ and $\gamma = \alpha \times \theta$ where α and θ are the parameter spaces of α and θ and the dependence of $H(\gamma)$ and ϕ has been suppressed. Let $L(\alpha, g_1, \theta, g_3)$ be the likelihood (3) with $\pi(W; \theta)^\Delta \{1 - \pi(W; \theta)\}^{1-\Delta}$ replacing $f(\Delta | W; g_2)$. We shall prove proposition 2 under the following regularity conditions:

- (i) γ lies in the interior of a compact set γ ;
- (ii) $\pi(W; \theta) > c > 0$ for all $\theta \in \theta$ for some c ;
- (iii) $E[H(\gamma)] \neq 0$ if $\gamma \neq \gamma_0$;
- (iv) $\text{var}\{H(\gamma_0)\}$ is finite and positive definite;
- (v) $E[\partial H(\gamma_0)/\partial \gamma']$ exists and is invertible;
- (vi) $E[\sup_{\gamma \in N} |H(\gamma)|]$, $E[\sup_{\gamma \in N} |\partial H(\gamma)/\partial \gamma'|]$ and $E[\sup_{\gamma \in N} |H(\gamma)H(\gamma)'|]$ are all finite where $|A| \equiv (\sum_{ij} A_{ij}^2)^{1/2}$ for any matrix A with elements A_{ij} where N is some open neighbourhood of γ_0 ;
- (vii) $L(\gamma) \equiv L(\alpha, g_{10}, \theta, g_{20})$ is a regular parametric model with score $S_\gamma(\gamma) = \partial \{\ln L(\gamma)\} / \partial \gamma$;
- (viii) for all $\gamma^* \in N$ $E_{\gamma^*}[H(\gamma)]$ and $E_{\gamma^*}[\sup_{\gamma \in N} |H(\gamma)' H(\gamma)|]$ are bounded where E_{γ^*} refers to expectation with respect to the submodel $L(\gamma^*)$.

We prove proposition 2 for the estimator $\hat{\alpha}(\phi, \hat{\theta})$ since the proof for $\hat{\alpha}(\phi, \theta_0)$ is a special case. First note that, by equations (1), (2) and (10), $E[U(\alpha_0, \phi, \theta_0)] = E[S_\theta(\theta_0)] = 0$. Theorems (2.6) and (3.4) of Newey and McFadden (1994) or corollary 1 of chapter 8 of Manski (1988) then imply that, under regularity conditions (i)-(vi), with probability approaching 1, there is a unique solution $\hat{\gamma}$ to $0 = \sum_i H_i(\gamma)$ and that

$$n^{1/2}(\hat{\gamma} - \gamma_0) = \{-E[\partial H(\gamma_0)/\partial \gamma']\}^{-1} H_i(\gamma_0) + o_p(1).$$

But, by definition of $H(\gamma)$, $\hat{\gamma}' = (\hat{\alpha}(\phi, \hat{\theta})', \hat{\theta}')$. Hence part (a) is proved and

$$n^{1/2}\{\hat{\alpha}(\phi, \hat{\theta}) - \alpha_0\} = -I(\phi) \left(n^{-1/2} \sum_i U_i(\phi) - E[\partial U(\alpha_0, \phi, \theta_0)/\partial \theta'] \{E[\partial S_\theta(\theta_0)/\partial \theta]\}^{-1} S_{\theta,i} \right) + o_p(1).$$

Now, under regularity conditions (v), (vii) and (viii), lemma (c.3) in Newey (1990a) implies the (generalized) information equalities

$$E[\partial U(\alpha_0, \phi, \theta_0)/\partial \alpha'] = -E[U(\phi)S'_\alpha] E[\partial U(\alpha_0, \phi, \theta_0)/\partial \theta'] = -E[U(\phi)S'_\theta],$$

$$E[\partial S_\theta(\theta_0)/\partial \theta] = -E[S_\theta S'_\theta] \text{ which imply, by theorem (2.2) in Newey (1990b), that } \hat{\alpha}(\phi, \hat{\theta})$$

is regular and

$$n^{1/2}\{\hat{\alpha}(\phi, \hat{\theta}) - \alpha_0\} = -I(\phi) n^{-1/2} \sum_i \text{Resid}_i \{U(\phi), S_\theta\} + o_p(1)$$

which proves part (b). Part (c) follows under our regularity conditions from theorem (4.5) in Newey and McFadden (1994). Part (d) follows from the uniqueness of the solution $\phi_{op}(W)$ to equation (8) and the fact that $\text{var}(S_{\text{eff}})^{-1}$ is the semiparametric variance bound.

Remark. If regularity condition (iii) is false, we can only conclude that there is a consistent root $\hat{\alpha}(\phi, \hat{\theta})$ to $0 = \bar{U}(\alpha, \phi, \hat{\theta})$ which satisfies part (b) of proposition 2. There may also be inconsistent roots.

Proof of proposition 4. By $T^k(\alpha_0)$ orthogonal to the tangent set and by $T^k(\alpha_0)$ orthogonal to $U^{(1)}(\alpha_0)$ it follows that $E[T^k(\alpha_0)S'_\alpha] = E[T^k(\alpha_0)S'_{\text{eff}}] = E[T^k(\alpha_0)U^{(2)}(\alpha_0, \phi_{op})']$. Hence $U^{(2)}(\alpha_0, \phi^k)$ is the mean-square projection of $U^{(2)}(\alpha_0, \phi_{op})$ on $T^k(\alpha_0)$. Also, by

conditions (a) and (c) $U^{(2)}(\alpha_0, \phi)$ is mean square continuous as a function of ϕ , so, by the spanning condition (b), $E[\|U^{(2)}(\alpha_0, \phi_{op}) - c^k T^k(\alpha_0)\|^2] \rightarrow 0$ as $k \rightarrow \infty$, and hence $E[\|S_{eff} - U^{(1)}(\alpha_0) - U^{(2)}(\alpha_0, \phi^k)\|^2] \rightarrow 0$. It follows that

$$\sum_{i=1}^n S_{eff,i}/n^{1/2} - \sum_{i=1}^n \{U_i^{(1)}(\alpha_0) + U_i^{(2)}(\alpha_0, \phi^k)\}/n^{1/2} \xrightarrow{p} 0.$$

By part (e) and the dominated convergence theorem, $\int \pi(w)f(y|X, V; \alpha) d\mu(y)$ and $\int \psi_j(w)f(y|X, V; \alpha) d\mu(y)$ are twice continuously differentiable, with derivatives bounded uniformly in j . Also, $\int \pi(w)f(y|X, V; \alpha) d\mu(y)$ is bounded away from 0, so that each element of $T^k(\alpha)$ is twice continuously differentiable with derivatives bounded uniformly in k . It then follows that for any $\hat{\alpha} = \alpha_0 + O_p(1/n^{1/2})$, $|\hat{J} - J| \leq ck |\hat{\alpha} - \alpha_0| = O_p(k/n^{1/2})$ for any matrix norm $|\cdot|$ where \hat{J} is defined in Section 7 and $J \equiv \Sigma_i T_i^k(\alpha_0) S'_{\alpha,i}$. Also, by Chebyshev's inequality, $|\hat{J} - J| = O_p(k/n^{1/2})$ with J as in Section 7, so by the triangle inequality, $|\hat{J} - J| = O_p(k/n^{1/2})$. It also follows similarly that $|\hat{\Omega} - \Omega| = O_p(k^2/n^{1/2})$. Furthermore, by standard matrix results, $\lambda_{\min}(\hat{\Omega}) \geq \lambda_{\min}(\Omega) - |\hat{\Omega} - \Omega|$, for the minimum eigenvalue $\lambda_{\min}(\cdot)$. Also, it can be shown that equation (2b) and the extremal characterization of λ_{\min} imply that, for any constant vector b with $|b| = 1$, $E[\|b' U^{(2)}(\alpha_0, k)\|^2] \geq E[\|b' \phi^k(W)\|^2] = b' E[\phi^k(W) \phi^k(W)'] b \geq \lambda_{\min}(E[\phi^k(W) \phi^k(W)'])$ so that $\lambda_{\min}(\Omega) \geq ck^{-ck}$ by part (d). Therefore, $\lambda_{\min}(\hat{\Omega}) \geq ck^{-ck}$ with probability approaching 1. This result implies that $|\hat{\Omega}^{-1} \hat{J} - \Omega^{-1} J| = O_p(k^{ck}/n^{1/2}) = o_p(1)$. By $U^{(2)}(\alpha, \phi)$ linear in ϕ and $|\Sigma_{i=1}^n T_i^k(\alpha_0)/n^{1/2}| = O_p(k)$,

$$\sum_{i=1}^n \{U_i^{(2)}(\alpha_0, \tilde{\phi}^k) - U_i^{(2)}(\alpha_0, \phi^k)\}/n^{1/2} = -(\hat{J}' \hat{\Omega}^{-1} - J' \Omega^{-1}) \sum_{i=1}^n T_i^k(\alpha_0)/n^{1/2} \xrightarrow{p} 0.$$

Thus

$$\sum_{i=1}^n S_{eff,i}/n^{1/2} - \sum_{i=1}^n \{U_i^{(1)}(\alpha_0) + U_i^{(2)}(\alpha_0, \tilde{\phi}^k)\}/n^{1/2} \xrightarrow{p} 0$$

by the triangle inequality, so by Slutsky's theorem

$$\sum_{i=1}^n \{U_i^{(1)}(\alpha_0) + U_i^{(2)}(\alpha_0, \tilde{\phi}^k)\}/n^{1/2} \xrightarrow{d} N(0, E[S_{eff} S'_{eff}]).$$

Furthermore, differentiating the identity $E_{\alpha}[T^k(\alpha)] = 0$ with respect to α (as allowed by parts (c) and (e)) gives $E[\partial T^k(\alpha_0)/\partial \alpha] = -E[T^k(\alpha_0) S'_{\alpha}]$. Thus, by similar arguments, it can also be shown that M and $-n^{-1} \Sigma_{i=1}^n \partial \{U_i^{(1)}(\hat{\alpha}) + U_i^{(2)}(\hat{\alpha}, \tilde{\phi}^k)\}/\partial \alpha$ both converge in probability to $E[S_{eff} S'_{eff}]$.

Finally proposition 4 follows by the usual mean value expansion argument, with $\Sigma_{i=1}^n \tilde{U}_i/n^{1/2}$ expanded in α around α_0 .

REFERENCES

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983) Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.*, 11, 432-452.
 Bickel, P., Klassen, C., Ritov, Y. and Wellner, J. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
 Breslow, N. E. and Cain, K. C. (1988) Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.
 Breslow, N. E. and Zhao, L. P. (1989) Logistic regression for stratified case-control studies. *Biometrics*, 44, 891-899.

- Carroll, R. J. and Wand, M. P. (1991) Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc. B*, **53**, 573-585.
- Cosslett, S. R. (1981) Efficient estimation of discrete choice models. In *Structural Analysis of Discrete Data with Econometric Applications* (eds C. F. Manski and D. McFadden), pp. 51-111. Cambridge: Massachusetts Institute of Technology Press.
- Cuzick, J. (1992) Semiparametric additive regression. *J. R. Statist. Soc. B*, **54**, 831-843.
- Flanders, W. D. and Greenland, S. (1991) Analytic methods for two stage case-control studies and other stratified designs. *Statist. Med.*, **10**, 739-747.
- Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029-1054.
- Hasminskii, R. Z. and Ibragimov, I. A. (1983) On asymptotic efficiency in the presence of an infinite-dimensional nuisance parameter. In *Proc. Joint USSR-Japan Symp.*, pp. 195-229. Berlin: Springer.
- Huber, P. (1985) Projection pursuit. *Ann. Statist.*, **13**, 435-474.
- Kalbfleisch, J. D. and Lawless, J. F. (1988) Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.*, **7**, 149-160.
- Kress, R. (1989) *Linear Integral Equations*. Berlin: Springer.
- Manski, C. F. (1988) *Analog Estimation Methods in Econometrics*. New York: Chapman and Hall.
- Manski, C. F. and Lerman, S. (1977) The estimation of choice probabilities from choice-based samples. *Econometrica*, **45**, 1977-1988.
- Manski, C. F. and McFadden, D. (1981) Alternative estimators and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data with Econometric Applications* (eds C. F. Manski and D. McFadden), pp. 2-50. Cambridge: Massachusetts Institute of Technology Press.
- McFadden, D. (1989) A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, **57**, 239-265.
- Newey, W. K. (1988) Adaptive estimation of regression models via moment restrictions. *J. Econometr.*, **38**, 301-339.
- (1990a) Efficient estimation of Tobit models under conditional symmetry. In *Semiparametric and Non-parametric Methods in Econometrics and Statistics* (eds W. Barnett, J. Powell and G. Tauchen), pp. 291-336. Cambridge: Cambridge University Press.
- (1990b) Semiparametric efficiency bounds. *J. Appl. Econometr.*, **5**, 99-135.
- (1992) Efficient estimation of semiparametric models via moment restrictions. *Working Paper*. Department of Economics, Massachusetts Institute of Technology, Cambridge.
- Newey, W. K. and McFadden, D. (1994) Estimation in large samples. In *Handbook of Econometrics* (eds D. McFadden and R. Engler), vol. 4. Amsterdam: North-Holland. To be published.
- Pakes, A. and Pollard, D. (1989) Simulation in the asymptotics of optimization estimators. *Econometrica*, **57**, 1027-1057.
- Pepe, M. S. and Fleming, T. R. (1991) A nonparametric method for dealing with mismeasured covariate data. *J. Am. Statist. Ass.*, **86**, 108-113.
- Robins, J. M. (1995) Locally efficient median regression with random censoring and surrogate markers. In *Proc. 1994 Int. Res. Conf. Lifetime Data Models in Reliability and Survival Analysis* (eds N. P. Jewell, A. C. Kimber, M.-L. T. Lee and G. A. Whitmore). Boston: Kluwer. To be published.
- 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. and Morgenstern, H. (1987) The foundations of confounding in epidemiology. *Comput. Math. Applic.*, **14**, 869-916.
- Robins, J. M. and Rotnitzky, A. (1992) Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology—Methodological Issues* (eds N. Jewell, K. Dietz and V. Farewell), pp. 297-331. Boston: Birkhäuser.
- Robins, J. M., Rotnitzky, A., Zhao, L.-P. and Lipsitz, S. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846-866.
- Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581-592.
- Zhao, L. P. and Lipsitz, S. (1992) Design and analysis of two-stage studies. *Statist. Med.*, **11**, 769-782.