Estimation of Regression Coefficients When Some Regressors Are Not Always Observed

James M. ROBINS, Andrea ROTNITZKY, and Lue Ping ZHAO*

In applied problems it is common to specify a model for the conditional mean of a response given a set of regressors. A subset of the regressors may be missing for some study subjects either by design or happenstance. In this article we propose a new class of semiparametric estimators, based on inverse probability weighted estimating equations, that are consistent for parameter vector α_0 of the conditional mean model when the data are missing at random in the sense of Rubin and the missingness probabilities are either known or can be parametrically modeled. We show that the asymptotic variance of the optimal estimator in our class attains the semiparametric variance bound for the model by first showing that our estimation problem is a special case of the general problem of parameter estimation in an arbitrary semiparametric model in which the data are missing at random and the probability of observing complete data is bounded away from 0, and then deriving a representation for the efficient score, the semiparametric variance bound, and the influence function of any regular, asymptotically linear estimator in this more general estimation problem. Because the optimal estimator depends on the unknown probability law generating the data, we propose locally and globally adaptive semiparametric efficient estimators. We compare estimators in our class with previously proposed estimators. We show that each previous estimator is asymptotically equivalent to some, usually inefficient, estimator in our class. This equivalence is a consequence of a proposition stating that every regular asymptotic linear estimator of α_0 is asymptotically equivalent to some estimator in our class. We compare various estimators in a small simulation study and offer some practical recommendations.

KEY WORDS: Cox proportional hazards model; Linear regression; Logistic regression; Measurement error; Missing covariates; Missing data; Nonlinear regression; Semiparametric efficiency; Survey sampling; Two-stage case-control studies; Validation study.

1. INTRODUCTION

In applied problems it is common to specify a model $g(X_i^*; \alpha)$ for the conditional mean of the response Y_i of a subject i given a set of regressors X_i^*, where α is an unknown parameter vector and $g(X_i^*; \alpha)$ is a known function such as $\alpha' X_i^*$ or $[1 + \exp\{-\alpha' X_i^*\}]^{-1}$. A subset X_i of the regressors $X_i^* \equiv (X_i', V_i')'$ may be missing for some study subjects, either by design or by happenstance. For example, X_i may be very expensive to measure and thus can be obtained only on a subsample of subjects, the validation sample. As a second example, X_i may represent responses to a set of personal questions that some subjects may refuse to answer. If the probability π_i that X_i is completely observed depends only on the vector V_i of other regressors, asymptotically unbiased estimates of the true value α_0 of α may be obtained by a complete case analysis; that is, by the (possibly nonlinear, possibly weighted) least squares regression of Y_i on X_i^* among subjects with complete data. If π_i depends on both Y_i and V_i, then the complete case estimator may be inconsistent.

One goal of this article is to propose a new class of estimators, based on inverse probability weighted estimating equations, that are asymptotically normal and unbiased for α_0 when (a) the data are missing at random in the sense of Rubin (1976), (b) π_i is bounded away from 0, and (c) the π_i are either known (as in a designed study) or can be parametrically modeled as in Rosenbaum (1987). Restriction (a) implies that π_i may depend on subject i's observed data, including Y_i, but not on the missing data. Restriction (b) guarantees the existence of $n^{1/2}$-consistent estimates of α_0.

Restrictions (a)–(c) plus the conditional mean model $g(X_i^*; \alpha)$ constitute a semiparametric model for the data. Define a semiparametric estimator to be one that is guaranteed to be asymptotically normal and unbiased for α_0 under the sole restrictions imposed by the model. A second goal of this article is to show that the optimal estimator in our class has the minimum possible asymptotic variance among all regular semiparametric estimators of α_0. That is, the asymptotic variance of the optimal estimator attains the semiparametric variance bound in the sense of Begun et al. (1983). Regularity is a technical condition that prohibits super-efficient estimators by specifying that the convergence of the estimator to its limiting distribution is locally uniform.

A third goal is to compare estimators in our class with previously proposed estimators, many of which were reviewed by Little (1993). Even when π_i depends only on V_i, the optimal estimator in our class is more efficient than the complete case estimator, because the optimal estimator extracts information available from subjects with incomplete data. In this setting, Dagenais (1973), Gourieroux and Montfort (1981), Beale and Little (1975), Pepe and Fleming (1991), and Carroll and Wand (1991) have previously proposed estimators that extract information from subjects with incomplete data. The Dagenais (1973), Beale and Little (1975), and Gourieroux and Montfort (1981) estimators assumed that $g(X_i^*; \alpha)$ was $\alpha' X_i^*$, whereas the Pepe and Fleming (1991) and Carroll and Wand (1991) estimators allowed for nonlinear regression functions. In Sections 4 and 6 we show that these previous estimators are asymptotically equivalent to inefficient estimators in our class.

* James M. Robins is Professor of Epidemiology and Biostatistics, and Andrea Rotnitzky is Assistant Professor of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Lue Ping Zhao is Associate Professor, Department of Epidemiology, Fred Hutchinson Cancer Center, Seattle, WA 98104. The research was partially supported by National Institutes of Health Grants 2 P30 ES00002, 1-R29-GM48704-01A1, R01AI32475, R01-ES03405, K04-ES00180, and GM-29745. The authors are grateful to Mark Van der Laan, Richard Gill, Whitney Newey, Fushing Hsieh, Daniel Rabinowitz, Donald Blevins and Sander Greenland for their help. Andrea Rotnitzky was additionally supported in part by a Mellon Foundation Faculty Award.

When Y_i is Bernoulli, V_i is discrete, and π_i may depend on both Y_i and V_i, we show that estimators previously proposed by Horvitz and Thompson (1952), Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981), Kalbfleisch and Lawless (1988), Breslow and Cain (1988), Imbens (1992), Flanders and Greenland (1991), and Zhao and Lipsitz (1992) are, with the exception of efficient but computationally challenging estimator of Cosslett (1981), asymptotically equivalent to some inefficient estimator in our class. These equivalences are corollary to Proposition 3.2 of Section 3, which states that every regular asymptotically linear estimator of α_0 is asymptotically equivalent to some estimator in our class, where an estimator $\hat{\alpha}$ is asymptotically linear if $n^{1/2}(\hat{\alpha} - \alpha_0)$ is asymptotically equivalent to an average of independent and identically distributed random variables.

Breslow and Cain (1988), Flanders and Greenland (1991), Weinberg and Wacholder (1993), Zhao and Lipsitz (1992), Carroll, Wang, and Wang (1993), and Carroll, Gail, and Lubin (1993) have considered the estimation of logistic regression models in two-stage case control designs. In Section 6 we show that the optimal estimator in our class attains the semiparametric variance bound in this class of designs even in the presence of the nonrandom differential measurement error considered by Carroll et al. (1993).

The article is organized as follows. In Section 2 we formalize our semiparametric model, propose a class of estimators, study their performance in a small simulation study, and provide some practical recommendations for applications. In Section 3 we show that the optimal estimator in our class attains the semiparametric variance bound.

In Sections 4 and 5 we note that the optimal estimator depends on the true but unknown probability distribution generating the data and thus is not available for data analysis; we thus propose locally and globally adaptive semiparametric efficient estimators.

If we impose additional mild regularity conditions, such as bounds on higher order moments, then even when Y_i has many continuous components and π_i is a complex function of Y_i, prior knowledge of the selection probabilities π_i can be exploited to construct a locally semiparametric efficient "inverse-probability weighted" estimator that should perform well in moderate sized samples (in the sense that, under all data generating processes, i.e. probability laws, allowed by our semiparametric model, the estimator will be approximately normal and centered on α_0); in contrast, due to the "curse of dimensionality," if the π_i are completely unknown, no estimator of α_0 exists with reasonable performance in the moderate sized samples occurring in practice. Yet, when data are missing at random, the likelihood principle implies that inference on α_0 should not depend on whether the π_i are known or completely unknown. With π_i known, our inverse-probability weighted estimators perform well but violate the likelihood principle; in contrast, globally efficient estimators which ignore prior knowledge concerning π_i and maximize a smoothed version of the likelihood function do (approximately) satisfy the likelihood principle but perform poorly in moderate samples. Analogous issues arise in randomized studies in which the probability of the treatment depends, by design, on a known but complex function of pre-treatment predictor variables (Robins, Mark, and Newey 1992; Robins and Morgenstern 1987).

In Section 6 we show that we can improve on the efficiency of inefficient estimators in our class by estimating the selection probabilities π_i even when the π_i are known. In Sections 4 and 6 we consider the relationship between our estimators and previously proposed estimators. In Sections 1–6.2, we assume that the data vectors are independent and identically distributed across subjects; in Section 6.3 we generalize our results to include the nonindependent "fixed fraction" sampling design considered by Breslow and Cain (1988). In Sections 1–6.3, we assume that the π_i are known as in a designed study. In Section 6.4 we allow the data to be missing by happenstance and assume a parametric model for the unknown π_i. In Sections 1–6 we assume that X_i is completely observed or completely unobserved. In Section 7 we generalize our previous results to arbitrary missing data patterns, thereby allowing different components of X_i to be missing on different subjects. In Section 8 we note that the semiparametric problem that we are considering is a special case of the general problem of estimating the parameters of an arbitrary semiparametric model in the presence of data missing at random. In Proposition 8.1, following Robins and Rotnitzky (1992), we provide representations for the efficient score, the semiparametric variance bound, and the influence function of any regular asymptotically linear estimator in this general estimation problem (provided that the probability π_i of observing complete data is bounded away from 0). We then prove our major propositions by specializing Proposition 8.1 to the semiparametric model characterized by the model $g(X_i^*; \alpha)$ for the conditional mean of Y_i given X_i^*. We conclude with a discussion of how our methods can be extended to other semiparametric models, such as the Cox proportional hazards model, with data missing at random.

2. THE MODEL AND A CLASS OF ESTIMATORS

2.1 The Regression Model

We develop our results in the context of the following example modeled on a large ongoing epidemiologic study, the Nurses Health Study (Stampfer et al. 1985). A prospective 5-year follow-up study of 100,000 previously healthy women was undertaken. At start of follow-up, a blood serum sample was obtained from each study subject and frozen for later analysis. A co-investigator wishes to study the effect of the antioxidants serum vitamin A (X_1) and vitamin E (X_2) recorded at start of follow-up on the subsequent development of coronary artery disease. Examples of possible outcome variables Y of interest include (a) Y is a Bernoulli random variable that takes the value 1 if a subject develops a myocardial infarction (i.e., a heart attack) over the 5-year follow-up period and 0 otherwise; (b) $Y = (Y_1, \ldots, Y_5)'$ is a vector of yearly electrocardiogram (EKG) results dichotomized as normal or abnormal, a multivariate discrete outcome; and (c) $Y = (Y_1, \ldots, Y_5)'$ is a vector of yearly measurements of the width in millimeters of a subject's left main stem coronary artery based on a noninvasive dye study, a continuous multivariate outcome. Due to the high cost of laboratory analyses

and to the small amount of stored serum per subject, the other investigators will only permit the stored serum of roughly 2% of the study subjects to be thawed and assayed for the exposures X_1 and X_2 of interest. On all subjects, error-prone estimates, $V_1 = (V_{11}, V_{12})'$, of $X = (X_1, X_2)'$ were obtained based on the results of a dietary questionnaire. The subjects with X recorded are said to be validation sample members. The investigator randomly selects subjects into the validation sample with selection probabilities that may depend on the surrogates V_1, other confounding factors V_2 such as age and race, and the outcome Y. Such selection rules offer efficiency advantages by overrepresenting subjects with extreme or rare values of Y, V_1, and V_2 in the validation sample (Breslow and Cain 1988). We summarize the data in the following notation:

n = total number of study subjects;
Y = outcome variable;
X = true exposures;
V_1 = mismeasured surrogates for X;
V_2 = a vector of confounding variables, such as age and race, with first component the constant 1;
$V = (V_1', V_2')'$;
$X^* = (X', V')'$; and
$\Delta = 1$ if X is fully observed (validation sample) and $\Delta = 0$ otherwise.

Our goal is to estimate the parameters α_0 of the regression model for the conditional mean of Y,

$$Y = g(X^*; \alpha_0) + \varepsilon, \qquad E[\varepsilon | X^*] = 0, \tag{1}$$

where $g(X^*; \alpha)$ is a known smooth function with dimension equal to that of Y and α_0 is a q vector of unknown parameters. Under model (1), when Y is Bernoulli, it is common to take

$$g(X^*; \alpha) = \{1 + \exp(-\alpha' X^*)\}^{-1}, \tag{2}$$

with $\alpha' X^* \equiv \alpha_1' X + \alpha_2' V_2 + \alpha_3' V_1$. It is often supposed that V_1 is a pure surrogate for X in the sense that Y and V_1 are independent given X and V_2 and thus, for example, that $\alpha_3 = 0$ in model (2). But it might occasionally be inappropriate to impose the restriction $\alpha_3 = 0$ in (2) a priori, because V_1 may happen to capture other aspects of the diet correlated with X that need to be adjusted for in the analysis.

2.2 Extraneous Surrogates

In epidemiologic studies it is very common to obtain data on additional extraneous surrogates $V^\dagger$ that are, in themselves, of no interest in the sense that the scientific goal remains the estimation of the conditional mean of Y given X^* rather than the conditional mean of Y given $(X^*, V^\dagger)$. As an example, suppose that at end of follow-up or at time of diagnosis of myocardial infarction (whichever comes first), all subjects are readministered the dietary questionnaire. Let $V^\dagger$ be the vitamin A and E measurements based on the second questionnaire. We would *not* adjust for $V^\dagger$ in our regression model (1), because the error in $V^\dagger$ will likely be nonrandom (i.e., differential) as the cases' ($Y = 1$) concern over their health status may make them more motivated than controls ($Y = 0$) to accurately recall their dietary habits. As a second example, it is standard epidemiologic practice *not* to adjust in model (1) for any covariate $V^\dagger$ (e.g., serum cholesterol) that is measured after start of follow-up (a) if the covariate is an intermediate variable on the causal pathway from vitamin exposure to disease (Y) or (b) more generally, if the covariate is affected by vitamin exposure (even if not an intermediate variable) (Robins 1987; Robins, Blevins, Ritter, and Wulfsohn 1992; Rosenbaum 1984; Weinberg 1992). In the absence of missing data, according to Proposition 3.1, efficient estimation of the parameters of model (1) will not use data on $V^\dagger$. In the presence of missing data, we show in Proposition 4.2 that it is useful to collect data on the extraneous surrogates $V^\dagger$, because when $V^\dagger$ is correlated with X conditional on (Y, V), data on $V^\dagger$ will increase the precision with which we can estimate the parameters of (1). We adopt the notation

$V^\dagger$ = extraneous surrogates; $\quad W^\dagger = (Y', V^{\dagger\prime})'$;

$$W = (W^{\dagger\prime}, V')'; \qquad L = (W', X')'.$$

Throughout we suppose that X is a continuous multivariate exposure, although our results also hold for discrete X. But W may have all continuous, all discrete, or mixed continuous and discrete components. In particular, this implies that X and its surrogate V_1 may be measured on different scales.

2.3 The Missing Data Mechanism

We assume until Section 7 that with probability 1, either X_1 and X_2 are both observed or X_1 and X_2 are both unobserved. We suppose until Section 6.3 that (Δ_i, X_i, W_i), $i = 1, \ldots, n$ are independently and identically distributed random vectors with i indexing study subjects. Because in our study the selection probabilities did not depend on X, we shall assume that

$$\Pr[\Delta = 1 | W, X] = \Pr[\Delta = 1 | W], \tag{3}$$

where the i subscript has been suppressed. Equation (3) implies that X is missing at random in the sense of Rubin (1976). Write $\pi(W) = \Pr[\Delta = 1 | W]$. Because the selection probabilities are under the control of the investigator in our example, we assume until Section 6.4 that

$$\pi(w) \text{ is a known function.} \tag{4}$$

Further, we assume that

$$\pi(W) > \sigma > 0 \text{ with probability 1,} \tag{5}$$

so that probability of being selected into the validation sample is bounded away from 0. A semiparametric model is characterized both by the available data and by restrictions on the joint distribution of that data. For each subject, we shall call

$$L = (W', X')' \tag{6}$$

the full data and

$$(\Delta, L_{\text{obs}}) \tag{7}$$

the observed data, where, following Little and Rubin (1987), L_{obs} is the observed component of L so $L_{\text{obs}} = L$ if $\Delta = 1$ and $L_{\text{obs}} = W$ if $\Delta = 0$. Let the semiparametric model "full"

be characterized by restriction (1) and data (6). Let the semiparametric model "obs" be characterized by (1), (3)–(5), and data (7). We shall propose a class of semiparametric estimators for the parameters α_0 of model "obs." To motivate these estimators, we first review well-known estimators of α_0 in model "full."

2.4 A Class of Estimators

The estimators $\hat{\alpha}^F(h)$ of α_0 in model "full" will be indexed by a $q \times t$ function of X^*, $h(X^*)$, satisfying the local identification condition $E\{h(X^*)\partial g(X^*; \alpha_0)/\partial\alpha'\}$ nonsingular, with t the dimension of Y and q the dimension of α_0. Here and throughout a superscript F will denote estimators and statistics based on the full data (6). Specifically, $\hat{\alpha}^F(h)$ solves $0 = \bar{D}^F(\alpha, h) \equiv n^{-1}\sum_i D_i^F(\alpha, h)$ with

$$D^F(\alpha, h) = h(X^*)\varepsilon(\alpha), \qquad \varepsilon(\alpha) = Y - g(X^*; \alpha). \quad (8)$$

Our semiparametric estimators $\hat{\alpha}(h, \phi)$ of α_0 in model "obs" will depend on $h(X^*)$ and on an arbitrary fixed $q \times 1$ function of w, $\phi(w)$, satisfying $E[\phi(W)'\phi(W)] < \infty$. Let π and ϕ denote the random variables $\pi(W)$ and $\phi(W)$. Then $\hat{\alpha}(h, \phi)$ solves $0 = \bar{D}(\alpha, h, \phi) \equiv n^{-1}\sum_i D_i(\alpha, h, \phi)$, where

$$D(\alpha, h, \phi) = \Delta D^F(\alpha, h)/\pi - A(\phi),$$
$$A(\phi) \equiv (\Delta - \pi)\phi/\pi. \quad (9)$$

Before studying the properties of $\hat{\alpha}(h, \phi)$, it will be convenient to review the well-known asymptotic properties of $\hat{\alpha}^F(h)$. It will be useful to define an asymptotically linear estimator (Newey 1990a). An estimator $\hat{\alpha}$ of α_0 is asymptotically linear with influence function B if

$$n^{1/2}(\hat{\alpha} - \alpha_0) = n^{-1/2}\sum_i B_i + o_p(1),$$

$E(B) = 0$, $E(B'B) < \infty$. If $\hat{\alpha}$ is asymptotically linear, then by the central limit theorem and Slutsky's theorem, $n^{1/2}(\hat{\alpha} - \alpha_0)$ is asymptotically normal with mean 0 and variance $E[BB']$. Asymptotically linear estimators $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ with the same influence function are asymptotically equivalent in the sense that $n^{1/2}(\hat{\alpha}^{(1)} - \hat{\alpha}^{(2)}) = o_p(1)$. Conversely, two asymptotically linear estimators that are asymptotically equivalent must have the same influence function. The following result is well known (see, for example, Manski 1988) under regularity conditions (1)–(9) provided in Appendix B, which we henceforth assume to be true.

Proposition 2.1. Under model "full" with probability approaching 1, there exists a unique solution $\hat{\alpha}^F(h)$ to $\bar{D}^F(\alpha, h) = 0$ such that $\hat{\alpha}^F(h)$ is asymptotically linear with influence function $\{\kappa(h)\}^{-1}D^F(h)$ where $D^F(h) \equiv D^F(\alpha_0, h)$ with $\kappa(h) \equiv E[h(X^*)\partial g(X^*; \alpha_0)/\partial\alpha']$. Further $\kappa(h) = -E[\partial D^F(\alpha_0, h)/\partial\alpha']$.

The corresponding properties of $\hat{\alpha}(h, \phi)$ are as follows.

Proposition 2.2.

a. Under model "obs," with probability approaching 1 there exists a unique solution $\hat{\alpha}(h, \phi)$ to $\bar{D}(\alpha, h, \phi) = 0$ such that $\hat{\alpha}(h, \phi)$ is asymptotically linear with influence function $\{\kappa(h)\}^{-1}D(h, \phi)$ where $D(h, \phi) = D(\alpha_0, h, \phi)$.

b. The asymptotic variance of $n^{1/2}\{\hat{\alpha}(h, \phi) - \alpha_0\}$ can be consistently estimated by $\{\hat{\kappa}(h)\}^{-1}\hat{\Omega}(h, \phi)\{\hat{\kappa}(h)'\}^{-1}$, where $\hat{\kappa}(h) = -n^{-1}\sum_i \partial D_i(\alpha, h, \phi)/\partial\alpha'$, and $\hat{\Omega}(h, \phi) = n^{-1}\sum_i D_i(\alpha, h, \phi)D_i(\alpha, h, \phi)'$ evaluated at $\hat{\alpha}(h, \phi)$.

The fundamental identity used in the proof of Proposition 2.2 in Appendix B is $E[D(h, \phi)] = E[D^F(h)] = 0$ because, by (3) and (5), for any $b(L)$, $E[\Delta b(L)/\pi] = E[E\{\Delta|L\}b(L)/\pi] = E[b(L)]$.

2.5 A Simulation Study and Some Practical Recommendations

To demonstrate various properties of the estimators in our class, we conducted three small simulation experiments. We provide some practical advice for applications guided by the results of these experiments. As in example 2 of Pepe and Fleming (1991), we generated for each of n subjects a normally distributed exposure $X \sim N(0, 1)$; a dichotomous surrogate $V_1 = I[X + \nu > 0]$ with $\nu \sim N(0, 1)$, ν independent of X, and $I[A] = 1$ if A is true and 0 otherwise; a nonrandom intercept $V_2 \equiv 1$; and an outcome Y from the logistic model (2) with $\alpha_3 = 0$, $\alpha_2 = -1$, and $\alpha_1 = 0$, 1, or 2 depending on the experiment. Subjects were randomly selected into the validation sample with probability .10. The sample size n was 2,000. In this section we assume that data on extraneous surrogates $V^\dagger$ were not generated, so $W = (Y, V')'$. Each experiment is based on 1,000 replications. Rows 1–3 of Table 1 provide Monte Carlo averages and estimated asymptotic relative efficiencies (ARE) for three different estimators of α_1 in our class. The estimated ARE of any estimator $\hat{\alpha}_1$ is calculated as the ratio of the Monte Carlo variance of the semiparametric efficient estimator reported in row 3 to that of $\hat{\alpha}_1$. Following Pepe and Fleming (1991), in all our analyses we assumed that it is known that V_1 is a pure surrogate and thus $\alpha_3 = 0$, so only α_1 and α_2 are estimated. Thus α in model (2) was redefined to be $(\alpha_1, \alpha_2)'$, so $h(X^*)$ is of length 2. The estimator $\hat{\alpha}_1(h^F_{\text{eff}}, 0)$ of row 1 uses $h(X^*)$ equal to $h^F_{\text{eff}}(X^*) \equiv (1, X)'$ and $\phi(W) \equiv 0$. The choice of h^F_{eff} was motivated by the fact that (a) in the absence of missing data, the choice $h^F_{\text{eff}}(X^*)$ is efficient because $\hat{\alpha}_1^F(h^F_{\text{eff}})$ is maximum likelihood, and (b) in the presence of missing data, $\hat{\alpha}_1(h^F_{\text{eff}}, 0)$ is maximum likelihood were data on nonvalidation subjects unavailable. When π is constant, $\hat{\alpha}_1(h^F_{\text{eff}}, 0)$ is algebraically equivalent to the estimators of Manski and Lerman (1977), Kalbfleisch and Lawless (1988), and Manski and McFadden (1981) discussed in Section 6.2. The estimator $\hat{\alpha}_1(h^F_{\text{eff}}, \hat{\phi}^{h^F_{\text{eff}}})$ of row 2 uses $\hat{\phi}^h(W) = \hat{E}[D^F(\hat{\alpha}, h)|W]$, with

$$\hat{E}(B|W = w) \equiv \sum_i \Delta_i B_i I(W_i = w) \Big/ \sum_i \Delta_i I(W_i = w) \quad (10)$$

as the sample average of B_i among validation sample members for whom $W_i = w$ and $\hat{\alpha} = \hat{\alpha}(h^F_{\text{eff}}, 0)$. Reading from Table 1, we note that $\hat{\alpha}_1(h^F_{\text{eff}}, \hat{\phi}^{h^F_{\text{eff}}})$ is more efficient than $\hat{\alpha}_1(h^F_{\text{eff}}, 0)$. This reflects the fact that according to Proposi-

Table 1. Results of a Simulation Study

Row	Estimator	Previously proposed by	Monte Carlo average of $\hat{\alpha}_1$, α_1 = 0	α_1 = 1	α_1 = 2	Estimated ARE of $\hat{\alpha}_1$, α_1 = 0	α_1 = 1	α_1 = 2
1	$\hat{\alpha}(h^F_{\text{eff}}, 0)$	Manski and Lerman (1977) Manski and McFadden (1981) Kalbfleisch and Lawless (1988)	−.02	1.01	2.03	.24	.47	.76
2	$\hat{\alpha}(h^F_{\text{eff}}, \hat{\phi}^{h^F_{\text{eff}}})$		−.01	1.02	2.04	.34	.58	.84
3	$\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}^{\hat{h}_{\text{eff}}})$*		.01	1.01	2.01	1.00	1.00	1.00
4	$\hat{\alpha}(h^F_{\text{eff}}, 0)$	Flanders and Greenland (1991)	−.01	1.01	2.03	.34	.58	.84
5	$\hat{\alpha}_{\text{PFCW}}$	Pepe and Fleming (1991) Carroll and Wand (1991)	.01	1.01	2.00	1.00	.97	.74
6	$\hat{\alpha}_{\text{MM}}(q_{\text{MM}})$	Breslow and Cain (1988)	−.01	1.02	2.03	.34	.58	.85
7	$\hat{\alpha}(h^F_{\text{eff}}, 0, \hat{\psi}^{(1)})$		.00			.65		
8	$\hat{\alpha}(h^F_{\text{eff}}, 0, \hat{\psi}^{(2)})$		.00			1.33		
9	$\hat{\alpha}(h^F_{\text{eff}}, 0, \hat{\psi}^{(3)})$		.00			1.10		

NOTE: The estimators in rows 7–9 use data on extraneous surrogates $V^\dagger$.
* Semiparametric efficient estimator in model "obs" when $W = (Y, V')'$.

tions 2.3 and 2.4, for a fixed h, the asymptotic variance of $\hat{\alpha}_1(h, \phi)$ is minimized at $\phi^h(W) = E[D^F(h)|W]$ and that $\hat{\alpha}(h, \phi^h)$ and $\hat{\alpha}(h, \hat{\phi}^h)$ are asymptotically equivalent.

Comparing rows 2 and 3 shows that in the presence of missing data, h^F_{eff} is no longer efficient. Rather, the optimal $h(X^*)$, $h_{\text{eff}}(X^*)$, is characterized in Equations (26)–(29) in Section 5.2 and depends both on the data through the surrogate V_1 and on the probability law generating the data. For example, in Section 5.3 we show that $h_{\text{eff}}(X^*) = (1, .1X + (1 - .1)E(X|V))'$ when $\alpha_{0,1} = 0$. In Section 5.2 we describe how to compute the consistent estimator $\hat{h}_{\text{eff}}(X^*)$ of $h_{\text{eff}}(X^*)$ used in row 3. In Sections 4 and 5 we prove that $\hat{\alpha}_1(h_{\text{eff}}, \phi^{h_{\text{eff}}})$ attains the semiparametric variance bound for our model "obs" and that $\hat{\alpha}_1(\hat{h}_{\text{eff}}, \hat{\phi}^{\hat{h}_{\text{eff}}})$ is asymptotically equivalent to $\hat{\alpha}_1(h_{\text{eff}}, \phi^{h_{\text{eff}}})$ and is thus semiparametric efficient. Rows 4–6 of Table 1, discussed in Section 6, compare the performance of previously proposed inefficient estimators with that of the efficient estimator of row 3. The estimators reported in rows 7–9 illustrate the efficiency advantage of collecting data on extraneous surrogates and are discussed in Section 6.4.

Practical Recommendations. The foregoing simulation results help motivate the following practical advice. The estimator $\hat{\alpha}(h^F_{\text{eff}}, 0)$ is simplest to compute. But if efficiency is of concern, then one can compute $\hat{\alpha}(h^F_{\text{eff}}, \hat{\phi}^{h^F_{\text{eff}}})$ with only slightly greater effort. If even further increases in efficiency are required, then with some additional effort one can use the formulas in Section 5.2 to compute the efficient estimator $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}^{\hat{h}_{\text{eff}}})$. These recommendations assume that, as in our simulation experiment, $W^\dagger = (Y', V^{\dagger\prime})'$ was discrete. But when $W^\dagger$ has continuous components, estimation of the optimal function $h_{\text{eff}}(X^*)$ can be computationally challenging requiring iterative calculations, as discussed in Section 4.2. Thus for nondiscrete $W^\dagger$, although the estimator $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}^{\hat{h}_{\text{eff}}})$ provided in Section 4.2 can (with difficulty) be computed, in practice the more easily computed estimator $\hat{\alpha}(h^F_{\text{eff}}, \hat{\phi}^{h^F_{\text{eff}}})$ will suffice when efficiency is not of major concern. Here h^F_{eff} and $\hat{\phi}^h$ are the generalizations given in Equation (17) and Section 2.7 of h^F_{eff} and $\hat{\phi}^h$ defined earlier. Except when Y is Bernoulli, estimation of the unknown function $\text{var}(\varepsilon|X^*)$ in Equation (17) will be necessary.

In certain instances one might choose to forego weighted estimators. For example, suppose that in model "obs" π is constant, data on $V^\dagger$ were not obtained, $V_2 \equiv 1$, $g(X^*; \alpha_0)$ is linear in X and does not depend on the surrogate V_1, and Y is continuous and univariate. Thus

$$W^\dagger = Y,\ g(X^*; \alpha_0) = \alpha_{0,0} + \alpha'_{0,1}X, \quad \text{and} \quad \pi \text{ is a constant } \rho. \tag{11}$$

Then a reasonable practical approach to estimating α_0 is to compute the parametric maximum likelihood estimator (MLE), $\hat{\alpha}_{\text{MLE}}$, of α_0 in the fully parametric model that imposes the additional assumption that the joint distribution of (ε, X, V) was generated by the "normal-normal measurement error model"

$$V \sim N(\mu, \Sigma),\ X|V \sim N(\gamma' V, \Omega), \quad \varepsilon|X, V \sim N(0, \sigma^2), \tag{12}$$

with $(\mu, \Sigma, \gamma, \Omega, \sigma^2)$ unknown parameters to be estimated. We show in Section 5.1 that $\hat{\alpha}_{\text{MLE}}$ is (a) asymptotically normal and unbiased for α_0 of (11) even if (12) is false and (b) attains the semiparametric variance bound if (12) is true. Result (a) depends critically on the linearity of $g(X^*; \alpha_0)$ in (11).

2.6 Efficiency for Fixed h

Proposition 2.3. For fixed h, the asymptotic variance of $\hat{\alpha}(h, \phi)$ is uniquely minimized at the asymptotic variance of $\hat{\alpha}(h, \phi^h)$ with

$$\phi^h \equiv E[D^F(h)|W]. \tag{13}$$

Proof. It follows from Proposition 2.2 that the ϕ^h minimizing the asymptotic variance of $\hat{\alpha}(h, \phi)$ minimizes the variance of $D(h, \phi)$. Now, because $\Delta D^F(\alpha, h)/\pi = D^F(\alpha, h) + (\Delta - \pi)D^F(\alpha, h)/\pi$, we have

$$D(\alpha, h, \phi) = D^F(\alpha, h) + J(\alpha, h, \phi), \tag{14}$$

where $J(\alpha, h, \phi) \equiv (\Delta - \pi)[D^F(\alpha, h) - \phi]/\pi$. Further, by (3) and (5), for all α, $(\Delta - \pi)[D^F(\alpha, h) - \phi]/\pi$ has mean 0 given (W, X) and thus is uncorrelated with $D^F(\alpha, h)$. Hence $\text{var}[D(\alpha, h, \phi)] = \text{var}[D^F(\alpha, h)] + E[(1 - \pi)\pi^{-1}E\{[D^F(\alpha, h) - \phi]^{\otimes 2} | W\}]$, which is minimized, in the positive definite sense, at $\phi = E[D^F(\alpha, h) | W]$, where $A^{\otimes 2} = AA'$.

The decomposition (14) provides insight into our estimation problem. Because $J(\alpha, h, \phi)$ has mean 0 for all α (not just for α_0) and is uncorrelated with the full data estimating function $D^F(\alpha, h)$, $J(\alpha, h, \phi)$ cannot help in identification or estimation of α_0. That is, as far as the estimation of α_0 is concerned, $J(\alpha, h, \phi)$ is just random noise added to $D^F(\alpha, h)$, representing the penalty we pay for not having observed X_i and thus $D_i^F(\alpha, h)$ for all subjects i. The variance of $J(\alpha, h, \phi)$ is a quantitative measure of this penalty that is minimized at ϕ^h for fixed h.

2.7 Adaptive Estimation of ϕ^h

The estimator $\hat{\alpha}(h, \phi^h)$ is not feasible, because ϕ^h depends on the unknown population quantity $E[D^F(h) | W]$. Hence we shall study the properties of estimators $\hat{\alpha}(h, \hat{\phi}^h)$, where $\hat{\phi}^h$ is an estimate of the unknown function ϕ^h computed as follows. Suppose that we have specified a regression model

$$E(D^F(h) | W) = l(W; \lambda_0), \tag{15}$$

where $l(W; \lambda)$ is a known regression function smooth in a finite dimensional parameter λ. We set $\hat{\phi}^h(W)$ equal to $l(W; \hat{\lambda})$, where $\hat{\lambda}$ is the possibly nonlinear least squares estimator of λ_0 from the regression of $D_i^F(\hat{\alpha}, h)$ on W_i in the validation sample and $\hat{\alpha} = \hat{\alpha}(h, \phi)$ for some preliminary $\phi(W)$. For W discrete, we take $l(W; \lambda_0)$ to be a saturated regression model in W; then $\hat{\phi}^h(W) = \hat{E}[D^F(\hat{\alpha}, h) | W]$, as defined in (10). Under the mild regularity conditions in Appendix B, we prove the following proposition.

Proposition 2.4. Under model "obs," $\hat{\alpha}(h, \hat{\phi}^h)$ is asymptotically equivalent to $\hat{\alpha}(h, \phi^\dagger)$, where $\phi^\dagger(w)$ is the probability limit of $\hat{\phi}^h(w)$. The asymptotic variance of $\hat{\alpha}(h, \hat{\phi}^h)$ can be consistently estimated as in Proposition 2.2, with $\hat{\phi}^h(W)$ in place of $\phi(W)$.

Hence when (15) is correctly specified, $\phi^\dagger(w)$ equals $\phi^h(w)$ and $\hat{\alpha}(h, \hat{\phi}^h)$ is efficient for fixed h. If (15) is misspecified, then $\hat{\alpha}(h, \hat{\phi}^h)$ remains asymptotically normal and unbiased for α_0. Although the asymptotics are the same, one can often improve finite sample performance by using $\tilde{\alpha}(h, \tilde{\phi}^h)$ rather than $\hat{\alpha}(h, \hat{\phi}^h)$, where $\tilde{\alpha}(h, \tilde{\phi}^h)$ is the limit of (i.e., iterates to convergence) the $\hat{\alpha}^{(j)}(h, \hat{\phi}^{h(j)})$ $j = 1, 2, \ldots$, where $\hat{\alpha}^{(1)}(h, \hat{\phi}^{h(1)}) = \hat{\alpha}(h, \hat{\phi}^h)$ and $\hat{\alpha}^{(j+1)}(h, \hat{\phi}^{h(j+1)})$ is defined like $\hat{\alpha}(h, \hat{\phi}^h)$ but with $\hat{\alpha}$ in $D_i^F(\hat{\alpha}, h)$ equal to $\hat{\alpha}^{(j)}(h, \hat{\phi}^{(j)})$. Finally, when W has only one or two continuous components, we could have taken $\hat{\phi}^h(w)$ to be the predicted value at w from a nonparametric (e.g., kernel or series regression) of $D_i^F(\hat{\alpha}, h)$ on W_i in the validation sample. Newey (1993a, 1993b) provided regularity conditions under which the resulting estimator of α_0 is consistent, asymptotically normal.

3. ASYMPTOTIC EFFICIENCY AND ASYMPTOTIC EQUIVALENCE

In this section we show that our class of estimators $\hat{\alpha}(h, \phi)$ contains a member whose asymptotic variance attains the semiparametric variance bound for model "obs" and that any regular asymptotic linear estimator of α_0 has the same influence function as a member of our class. We first formally define the semiparametric variance bound for the class of regular estimators following Begun et al. (1983), Bickel, Klaassen, Ritov, and Wellner (1993), and Newey (1990a). Suppose that the data consist of n independent realizations of a random variable Z. Let $\mathcal{L}(\alpha, \theta; Z)$ be the likelihood function for a single subject in a semiparametric model indexed by a q-dimensional parameter α and a nuisance parameter θ taking values in an infinite-dimensional set. Let (α_0, θ_0) index the distribution generating Z. For example, in model "full" Z is $L = (X', W')'$ and $\mathcal{L}(\alpha, \theta; Z) = \mathcal{L}^F(\alpha, \theta; L)$, where

$$\mathcal{L}^F(\alpha, \theta; L) = f(V; \theta_3) f(X | V; \theta_2) f[\varepsilon(\alpha) | X, V; \theta_1] \times f[V^\dagger | Y, V, X; \theta_4] \tag{16a}$$

and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, restricted only by $\int t\, dF(t | X, V; \theta_1) = 0$ because the error has mean 0. In model "obs" $Z = (\Delta, L_{\text{obs}})$ and $\mathcal{L}(\alpha, \theta; Z) = \mathcal{L}(\alpha, \theta; \Delta, L_{\text{obs}})$, with

$$\mathcal{L}(\alpha, \theta; \Delta, L_{\text{obs}}) = \mathcal{L}^F(\alpha, \theta; X, W)^\Delta \times \left\{\int\int \mathcal{L}^F(\alpha, \theta; x_1, x_2, W)\, dx_1\, dx_2\right\}^{1-\Delta} \pi^\Delta (1 - \pi)^{1-\Delta}. \tag{16b}$$

Define a regular parametric submodel to be a regular fully parametric model with parameters (α, η) and likelihood $\mathcal{L}(\alpha, \eta; Z)$ with true values (α_0, η_0), where the subprefix refers to the fact that for each η, the distribution $\mathcal{L}(\alpha, \eta; Z)$ is a distribution $\mathcal{L}(\alpha, \theta; Z)$ allowed by the semiparametric model. An estimator is regular in a regular parametric model if locally it converges uniformly to its limiting distribution. A more precise definition of a regular estimator has been given by Bickel et al. (1993) and Newey (1990a). An estimator is regular in a semiparametric model if it is regular in every regular parametric submodel. Because LeCam and Hajek proved that in a regular parametric model, the Cramer–Rao variance bound for α_0 is a lower bound for the asymptotic variance of any regular estimator of α_0, it follows that in a semiparametric model, the supremum of the Cramer–Rao bounds for α_0 over all regular parametric submodels is a lower bound for the asymptotic variance of any regular estimator (Begun, Hall, Huang, and Wellner 1983; Bickel et al. 1993; Newey 1990a). The supremum is referred to as the semiparametric variance bound of the semiparametric model. Since, by missing the at random assumption, the term $(1 - \pi)^{1-\Delta}$ is not contained within the integral in (16b), it follows that, for likelihood-based inference concerning α_0, it does not matter whether π is completely known or unknown. Therefore, the semiparametric variance bound for α_0 in model "obs" is the same whether or not π is known. We now motivate our principal result for the missing data case with the following full-data result.

Proposition 3.1. Suppose that $\tilde{\alpha}^F$ is a regular, asymptotically linear (RAL) estimator of α_0 in the semiparametric model "full." Then (a) its influence function lies in the set $\{[\kappa(h)]^{-1}D^F(h)\}$, and (b) there exists a unique h^F_{eff} such that $\{\text{var}[D^F(h^F_{\text{eff}})]\}^{-1}$ equals the semiparametric variance bound. In addition, $\kappa(h^F_{\text{eff}}) = \text{var}[D^F(h^F_{\text{eff}})]$, so the asymptotic variance of $\hat{\alpha}(h^F_{\text{eff}})$ attains the bound.

Proposition 3.2. Suppose that $\hat{\alpha}$ is a RAL estimator of α_0 in the semiparametric model "obs"; then (a) its influence function lies in the set $\{[\kappa(h)]^{-1}D(h, \phi)\}$, and (b) there exists a unique h_{eff} and ϕ_{eff} such that $\{\text{var}[D(h_{\text{eff}}, \phi_{\text{eff}})]\}^{-1}$ equals the semiparametric variance bound for the model. In addition, $\kappa(h_{\text{eff}}) = \text{var}[D(h_{\text{eff}}, \phi_{\text{eff}})]$, so the asymptotic variance of $\hat{\alpha}(\phi_{\text{eff}}, h_{\text{eff}})$ attains the bound.

The proofs of these propositions are deferred to Section 8. Proposition 3.1 is a corollary of results provided by Newey (1990a), Chamberlain (1987), and Newey and Powell (1990). Proposition 3.2 implies that any RAL estimator α_0 in model "obs" must be asymptotically equivalent to an estimator $\hat{\alpha}(h, \phi)$ in our class. $D^F(h^F_{\text{eff}})$ and $D(h_{\text{eff}}, \phi_{\text{eff}})$ are called the efficient scores in their respective semiparametric models. Chamberlain (1987) showed that

$$h^F_{\text{eff}}(X^*) = [\partial g(X^*; \alpha_0)/\partial\alpha]\{\text{var}(\varepsilon \mid X^*)\}^{-1} \quad (17)$$

in model "full." We derive h_{eff} and ϕ_{eff} for model "obs" in the following section. Propositions 3.1 and 3.2 imply that $\text{var}\{D^F[h^F_{\text{eff}}]\} - \text{var}\{D[h_{\text{eff}}, \phi_{\text{eff}}]\}$ is the information about α_0 lost due to missing data.

We next provide an interesting restatement of Proposition 3.2a. Consider any correctly specified model $\pi(\psi)$ for the known missingness process $\pi \equiv \Pr[\Delta = 1 \mid W]$; that is,

$$\pi = \pi(\psi_0), \quad (18)$$

with $\pi(\psi) \equiv \pi(W; \psi)$ a smooth function of W and a finite-dimensional parameter ψ whose range lies in (0, 1]. Let $S_\psi \equiv S_\psi(\psi_0)$ be the score for ψ evaluated at ψ_0. Note that $S_\psi(\psi_0) \equiv \partial \log[\mathcal{L}^{\text{mis}}(\psi_0)]/\partial\psi$ equals $A(\phi_\psi)$, where

$$\mathcal{L}^{\text{mis}}(\psi) \equiv \{\pi(\psi)\}^{\Delta}\{1 - \pi(\psi)\}^{1-\Delta},$$

$$\phi_\psi \equiv \pi\partial \text{ logit } \pi(\psi_0)/\partial\psi. \quad (19)$$

Conversely, given $\phi(W)$ taking values in R^q, $A(\phi)$ is S_ψ in the model logit $\pi(\psi) = \text{logit } \pi + \psi\phi/\pi$, with $\psi_0 = 0$. Furthermore, we noted previously that $A(\phi)$ has mean 0 given (W, X). Conversely, according to Proposition 8.2, any random variable B that is a function of the observed data (7) with mean 0 conditional on the full data $L = (W', X')'$ has the representation $A(\phi)$ for some function ϕ. Hence we obtain the following result.

Restatement of Proposition 3.2a. The influence function of any RAL estimator α_0 in the missing data model "obs" lies in the set whose elements are formed as follows. First, take the influence function of an arbitrary RAL estimator in the model "full," multiply by the indicator variable for full (i.e., complete) data, and divide by the probability of having full data; then add an arbitrary element from the set of scores S_ψ for correctly specified regular parametric models for the *known* missingness process or, equivalently, from the set of functions of the observed data that have mean 0 conditional on the full data.

4. EFFICIENCY CALCULATIONS AND COMPARISONS

In Section 4.2 we derive h_{eff} and ϕ_{eff}, propose some adaptive locally semiparametric efficient estimators, and compare the efficiency of our proposed estimators to that of some previously proposed estimators. To motivate these results, we first study the simpler problem of estimating the "best linear predictor" of Y given X^*.

4.1 Global and Local Efficient Estimation of Best Linear Predictors

Suppose that Y is univariate and α_0 and X^* are vectors of dimension q. Let $\mathbf{i}(X^*)$ be the identity function; that is, $\mathbf{i}(X^*) = X^*$. The asymptotic variance of $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$ generally will not attain the semiparametric variance bound in model "obs," because $h_{\text{eff}}(X^*)$ will not equal $\mathbf{i}(X^*)$. Nevertheless, the asymptotic variance of $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$ attains the efficiency bound for estimation of the best linear predictor of Y given X^*. Formally, let models "full b" and "obs b" be defined like their counterpart models "full" and "obs" with $g(X^*; \alpha_0) = \alpha_0'X^*$, except that $E[\varepsilon \mid X^*] = 0$ is replaced by the weaker condition that $E[X^*\varepsilon] = 0$. Then α_0' equals $E[YX^{*\prime}]\{E[X^*X^{*\prime}]\}^{-1}$. We call $\alpha_0'X^*$ the best linear predictor because α_0 minimizes $E[(Y - \alpha'X^*)(Y - \alpha'X^*)']$. Model "full b" does not restrict the distribution of the data (6) and thus can be viewed as "saturated." In the Appendix we prove the following proposition.

Proposition 4.1. The asymptotic variance of $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$ attains the semiparametric variance bound for α_0 in model "obs b."

Bickel et al. (1993) independently derived the bound for model "obs b" when $\pi(W)$ is constant. It follows from Proposition 2.4 that if W_i is discrete and we use (10), then $\hat{\alpha}(\mathbf{i}, \hat{\phi}^{\mathbf{i}})$ is semiparametric efficient in model "obs b." If W_i has continuous components, then $\hat{\alpha}(\mathbf{i}, \hat{\phi}^{\mathbf{i}})$ will be locally semiparametric efficient in model "obs b" at restriction (15) with $h(X^*) = \mathbf{i}(X^*)$, where we have used the following definition due to Bickel et al. (1993).

Definition. Given a semiparametric model, say A, and an additional restriction R on the joint distribution of the data not imposed by the model, we say that an estimator $\hat{\alpha}$ is *locally semiparametric efficient in model A at R* if $\hat{\alpha}$ is a semiparametric estimator in model A whose asymptotic variance attains the semiparametric variance bound for model A when R is true. Informally, $\hat{\alpha}$ is the most efficient possible estimator of α_0 when model A and restriction R are both true that is guaranteed to remain asymptotically normal and unbiased for α_0 when A is true but R is false.

It is of interest to compare $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$ to some previously proposed estimator of α_0 in model "obs b." Gourieroux and Montfort (1981) discussed a semiparametric estimator of α_0 in model "obs b" when $\pi(W)$ was a constant ρ. They proposed computing a weighted regression of Y_i on (X_i, V_i), $i = 1, \ldots, n$, except they first estimated the unobserved X_i

for nonvalidation sample subject i by its predicted value $\hat{\gamma}V_i$ and then used an estimated weight $\hat{\omega}$, $\hat{\omega} \leq 1$, to downweight the nonvalidation sample observations. Here $\hat{\gamma}$ is the ordinary least squares (OLS) estimator from the multivariate regression of X on V in the validation sample; that is, $\tilde{\alpha}_{GM} = (\tilde{\alpha}'_{GM,1}, \tilde{\alpha}'_{GM,2})'$ solves

$$0 = n^{-1/2} \sum_i D_{GM,i}(\alpha, \hat{\gamma}) = n^{-1/2} \sum_i \Delta_i X_i^* \varepsilon_i(\alpha) + (1 - \Delta_i)\hat{\omega}(V_i'\hat{\gamma}', V_i')'(Y_i - \alpha_1'\hat{\gamma}V_i - \alpha_2'V_i), \quad (20)$$

with $\alpha' = (\alpha_1', \alpha_2')$. Dagenais (1973) and Beale and Little (1976) proposed estimators that differ from the Gourieroux and Montfort estimator only in the choice of weight $\hat{\omega}$. In Appendix C, we show that $\tilde{\alpha}_{GM}$ has the same influence function as $\hat{\alpha}(\mathbf{i}, \phi_{GM})$ with

$$\phi'_{GM} = \{(\gamma_0 V)'(Y - \alpha'_{0,1}\gamma_0 V - \alpha'_{0,2}V), V'(Y - \alpha'_{0,1}\gamma_0 V - \alpha'_{0,2}V)\}\{(1 - \rho) + \rho\omega^{-1}\}^{-1}, \quad (21a)$$

where $\alpha_0' X^* = \alpha'_{0,1}X + \alpha'_{0,2}V$, $\gamma_0 = E[XV']\{E(VV')\}^{-1}$, and ω is the limit of $\hat{\omega}$. The influence function of $\tilde{\alpha}_{GM}$, in contrast to that of $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$, is always linear in Y_i. Thus one would not expect $\tilde{\alpha}_{GM}$ to be efficient. In fact $\tilde{\alpha}_{GM}$ wil be efficient if and only if $\phi_{GM} = \phi^{\mathbf{i}}$. But $\phi^{\mathbf{i}\prime} \equiv E[X^*\varepsilon | W]'$ equals

$$(\{E(X|V, Y)Y - E(XX'|Y, V)\alpha_{0,1} - E(X|V, Y)\alpha'_{0,2}V\}', V'(Y - \alpha'_{0,1}E(X|V, Y) - \alpha'_{0,2}V)). \quad (21b)$$

Comparing (21a) and (21b), $\phi_{GM} = \phi^{\mathbf{i}}$ if and only if (a) $\omega = 1$, $E[X|V, Y] = \gamma_0 V$, and either $\alpha_{0,1} = 0$ or $\text{var}[X|Y, V] = 0$, or (b) $\omega = 0$ and $\text{var}(\varepsilon) = 0$.

Let $\hat{\alpha}^*_{MLE}$ be the MLE of $\alpha_0 = (\alpha'_{0,1}, \alpha'_{0,2})'$ in the fully parametric multivariate normal model defined by (12) and

$$Y = W^\dagger, E(Y|X^*) = \alpha_{0,1}X + \alpha_{0,2}V, \quad \pi \text{ is a constant } \rho. \quad (22)$$

Results of Gourieroux and Montfort (1981) and Little (1993) imply that $\hat{\alpha}^*_{MLE}$ is a semiparametric estimator in model "obs b." Thus $\hat{\alpha}^*_{MLE}$ is locally semiparametric efficient and asymptotically equivalent to $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$ at the joint restriction (12) and (22). $\hat{\alpha}^*_{MLE}$ will generally be inefficient and thus not asymptotically equivalent to $\hat{\alpha}(\mathbf{i}, \phi^{\mathbf{i}})$ when either (12) or (22) is false.

4.2 Functional Equations for h_{eff} and ϕ_{eff} in Model "obs"

In Section 8 we prove the following proposition.

Proposition 4.2. In model "obs," $h_{eff}(X^*)$ is the unique solution to the functional (integral) equation

$$h(X^*) = \{\partial g(X^*; \alpha_0)/\partial\alpha\}t(X^*) + E[\mathbf{r}\{h(X^*)\varepsilon\}\varepsilon'|X^*]t(X^*) \quad (23)$$

and

$$\phi_{eff} = \phi^{h_{eff}}, \quad (24)$$

where $t(X^*) = \{E[\varepsilon\varepsilon'/\pi|X^*]\}^{-1}$, $\mathbf{r}(B)$ is the operator $(1 - \pi)\pi^{-1}E[B|W]$ and, again, $\phi^{h_{eff}} \equiv E[h_{eff}(X^*)\varepsilon|W]$. We use bold lowercase letters to denote operators.

Equation (23) is not directly useful for data analysis, because (a) the solution $h_{eff}(X^*)$ is a function of the unknown true distribution generating the data, and (b) even were the true distribution known, except in special cases such as those discussed in Section 5, $h_{eff}(X^*)$ will not exist in closed form in the sense that it cannot be explicitly represented as a function of the true distribution. But as shown in Appendix A, for any distribution allowed by model "full" with likelihood $\mathcal{L}^F(\alpha, \theta; L)$ given by (16a), Equation (23) can be solved iteratively by the method of successive approximation. That is, given the mth iterate $h_m(X^*)$, $h_{m+1}(X^*)$ is obtained by evaluating the right side of (23) at $h_m(X^*)$, with expectations computed with respect to $\mathcal{L}^F(\alpha, \theta; L)$. $h_m(X^*)$ will converge to the unique solution $h(X^*; \alpha, \theta)$ of (23) under $\mathcal{L}^F(\alpha, \theta; L)$ as $m \to \infty$ for any initial function $h_0(X^*)$ (Kress 1989). This implies a quite general approach to obtaining locally efficient adaptive estimators in model "obs," as follows:

1. Specify a fully parametric model $\mathcal{L}^F(\alpha, \eta; X, W)$.
2. Estimate η by $\hat{\eta}$ solving $\sum_i \pi_i^{-1}\Delta_i \partial \log \mathcal{L}^F(\hat{\alpha}, \eta; X, W)/\partial\eta = 0$, with $\hat{\alpha} \equiv \hat{\alpha}(h, \phi)$ a preliminary estimate of α_0.
3. Solve (23), by successive approximation if necessary, and evaluate (24) under the law $\mathcal{L}^F(\hat{\alpha}, \hat{\eta}; X, W)$ to obtain $\hat{h}_{eff} \equiv h(\hat{\alpha}, \hat{\eta})$, $\hat{\phi}_{eff} \equiv \phi(\hat{\alpha}, \hat{\eta})$.

Then, by a proof analogous to that of Proposition 2.4 given in Appendix B, $\hat{\alpha}(\hat{h}_{eff}, \hat{\phi}_{eff})$ will be asymptotically equivalent to $\hat{\alpha}(h^\dagger, \phi^\dagger)$, where $h^\dagger(X^*)$ and $\phi^\dagger(W)$ are the probability limits of $\hat{h}_{eff}(X^*)$ and $\hat{\phi}_{eff}(W)$ if $h(X^*; \alpha, \eta)$ and $\phi(W; \alpha, \eta)$ have, for example, bounded second derivatives with respect to (α, η) with probability 1. In particular, if $\mathcal{L}^F(\alpha, \eta; X, W)$ is correctly specified with true values (α_0, η_0), then $(\hat{\alpha}, \hat{\eta})$ will be consistent for (α_0, η_0), $h^\dagger = h_{eff}$, and $\phi^\dagger = \phi_{eff}$. Hence $\hat{\alpha}(\hat{h}_{eff}, \hat{\phi}_{eff})$ is locally semiparametric efficient at the parametric model $\mathcal{L}^F(\alpha, \eta; X, W)$. In fact, because V is ancillary, $h^\dagger = h_{eff}$ and $\phi^\dagger = \phi_{eff}$ even if the model $f(V; \eta_3)$ for the law of V is misspecified. In Appendix D we describe "simulation" estimators for the conditional expectations on the right sides of Equations (23) and (24) that avoid the need for numerical integration. In moderate sized samples, the estimator $\hat{\alpha}(\hat{h}_{eff}, \hat{\phi}_{eff})$ should perform well (in the sense that it will be approximately normal and centered on α_0) provided the dimension of η is not too large.

Equations (23) and (24) allow us to deduce conditions under which data on an extraneous surrogate $V^\dagger$ that is not adjusted for in model (1) will not provide additional information about α_0. Suppose that the selection probability $\pi(W)$ does not depend on the extraneous surrogates $V^\dagger$, so $V^\dagger$ may be ignored without introducing bias. From (23) and (24), it follows that the efficient score $D(h_{eff}, \phi_{eff})$ will depend on $V^\dagger$ unless $E[h_{eff}(X^*)\varepsilon|W]$ does not depend on $V^\dagger$ or π is equal to 1 almost surely. It follows that a sufficient condition for the efficient score not to depend on $V^\dagger$ is that $V^\dagger$ and X are conditionally independent given (Y, V). Furthermore, if W is a perfect surrogate for X in the sense that

$W = c(W)$ with probability 1 for some fixed function $c(\cdot)$, then $h^F_{\text{eff}}(X^*)$ solves (23), $\hat{\alpha}(h^F_{\text{eff}}, \phi^{h^F_{\text{eff}}}) = \hat{\alpha}^F(h^F_{\text{eff}})$, and asymptotically no information is lost due to missing data.

5. SPECIAL CASES WHERE h_{eff} AND ϕ_{eff} EXIST IN CLOSED FORM

5.1 Normal-Normal Measurement Error Model

Consider model "obs" satisfying (11) and suppose in truth that (12) is satisfied for some unknown $\eta = \eta_0$, where $\eta' = (\eta_1', \eta_2', \eta_3')$, $\eta_1 = \sigma^2$, $\eta_2 = (\gamma, \Omega)$, and $\eta_3 = (\mu, \Sigma)$. Then with η replacing θ, let $\mathcal{L}(\alpha, \eta; \Delta, L_{\text{obs}})$ be the missing-data likelihood (16b) corresponding to the fully parametric multivariate normal model defined by the restrictions of (11)–(12) and the model "obs." The parametric efficient score for α in this fully parametric model is $S_{\alpha,\text{eff}} \equiv S_\alpha - E[S_\alpha S_\eta'] \{E[S_\eta S_\eta']\}^{-1} S_\eta$, with $S_\alpha \equiv \partial \log \mathcal{L}(\alpha_0, \eta_0; \Delta, L_{\text{obs}})/\partial\alpha$ and $S_\eta \equiv \partial \log \mathcal{L}(\alpha_0, \eta_0; \Delta, L_{\text{obs}})/\partial\eta$. Then

$$h_{\text{eff}}(X^*) = E[S_{\alpha,\text{eff}} \mid Y = 1, X^*] - E[S_{\alpha,\text{eff}} \mid Y = 0, X^*] \quad (25)$$

is a closed-form expression for $h_{\text{eff}}(X^*)$ in terms of the parameters (α_0, η_0) of (11) and (12). We obtained (25) without explicitly solving (23) by arguing as follows. Let $(\hat{\alpha}_{\text{MLE}}, \hat{\eta}_{\text{MLE}})$ maximize $\prod_i \mathcal{L}(\alpha, \eta; \Delta_i, L_{\text{obs},i})$. In Appendix C, we prove that $\hat{\alpha}_{\text{MLE}}$ is a semiparametric RAL estimator of α_0 in model "obs" of (11), even if (12) is false. Hence because it is also a parametric MLE, $\hat{\alpha}_{\text{MLE}}$ is locally semiparametric efficient in model "obs" at restriction (12) with influence function $E[S_{\alpha,\text{eff}} S_{\alpha,\text{eff}}']^{-1} S_{\alpha,\text{eff}}$. Equation (25) is then the special case corresponding to Y univariate of part (b) of the following corollary to Proposition 3.2a.

Corollary 5.1. If B is the influence function of a RAL estimator in model "obs," then $B = D(h, \phi)$ with (a) $\phi(W) = b_2(W)$, where $b_2(W)$ is determined by B through the unique decomposition $B = \Delta b_1(L) + (1 - \Delta) b_2(W)$, and (b) $h(X^*) = (h_1(X^*), \ldots, h_t(X^*))$, with $h_j(X^*) = E[B \mid Y = e_j, X^*] - E[B \mid Y = 0, X^*]$ and e_j is the vector of the same dimension t as Y, with the jth component equal to 1 and all other components 0. Without loss of generality, we have assumed that e_j for $j = (1, \ldots, t)$ and the 0 vector are in the support of Y.

But when $E(Y \mid X^*) \neq \alpha_0'(1, X')'$ and thus (11) is false, $\hat{\alpha}_{\text{MLE}}'(1, X')'$ is not a consistent estimator of the best linear predictor of Y given $(1, X')'$; that is, of $E[Y(1, X')] E[(1, X')'(1, X')]^{-1}(1, X')'$. Compare with the properties of $\hat{\alpha}^*_{\text{MLE}}$ described in the final paragraph of Section 4.1.

5.2 $W^\dagger$ Discrete

We next show that, as in our simulation experiment, if $W^\dagger = (Y', V^{\dagger\prime})'$ is discrete, then $h_{\text{eff}}(X^*)$ always exists in closed form. We first note that because $\phi_{\text{eff}} = E[h_{\text{eff}}(X^*)\varepsilon \mid W]$, Equation (23) can be rewritten as

$$h_{\text{eff}}(X^*) = \{\partial g(X^*; \alpha_0)/\partial\alpha\} t(X^*) + E[(1 - \pi)\pi^{-1}\phi_{\text{eff}}\varepsilon' \mid X^*] t(X^*). \quad (26)$$

Hence if we obtain a closed-form expression for ϕ_{eff}, (26) becomes a closed-form expression for $h_{\text{eff}}(X^*)$. Now, after right multiplying Equation (26) by ε and then taking conditional expectations given W, we obtain that ϕ_{eff} solves

$$m(W) = \phi - E\{E[(1 - \pi)\pi^{-1}\phi\varepsilon' \mid X^*] t(X^*)\varepsilon \mid W\}, \quad (27)$$

with $m(W) \equiv E[\{\partial g(X^*; \alpha_0)/\partial\alpha\} t(X^*)\varepsilon \mid W]$. In Appendix A we show that ϕ_{eff} is the unique solution to (27). When $W^\dagger$ is discrete, (27) is a finite-dimensional matrix equation and admits a closed-form solution as follows. Denote the lth components of the q vectors $\phi(W)$ and $m(W)$ by $\phi_l(W)$ and $m_l(W)$. Write $\phi_l(W) \equiv \phi_l(W^\dagger, V)$. Then for $W^\dagger$ discrete with S levels, say $w_1^\dagger, \ldots, w_S^\dagger$, we will (in a slight abuse of notation) define $\phi_l(V)$ to be the S vector-valued function of V with sth component $\phi_l(w_s^\dagger, V)$. The S vector $m_l(V)$ is defined analogously in terms of $m_l(W)$. Then, by (27), we obtain the closed-form solution

$$\phi_{\text{eff},l}(V) = \{I_{S\times S} - q(V)\}^{-1} m_l(V), \quad (28)$$

where $I_{S\times S}$ is the $S \times S$ identity matrix; $q(V)$ is the $S \times S$ matrix with k, s entry, $q_{ks}(V)$, equal to

$$E[E\{(1 - \pi)\pi^{-1} I(W^\dagger = w_s^\dagger)[Y - g(X, V; \alpha_0)]' \mid X, V\} \times t(X, V)\{y_k - g(X, V; \alpha_0)\} \mid W^\dagger = w_k^\dagger, V], \quad (29)$$

and y_s and y_k are the realizations of Y corresponding to $W^\dagger = w_s^\dagger$ and $W^\dagger = w_k^\dagger$.

Adaptive Estimation for Discrete $W^\dagger$. When $W^\dagger$ is discrete, adaptive local semiparametric efficient estimation based on a parametric model $\mathcal{L}^F(\alpha, \eta; L)$ proceeds as in Section 4.2, except in Step 3 we now evaluate (28) and then (26) under $\mathcal{L}^F(\hat{\alpha}, \hat{\eta}; L)$ to obtain $\hat{\phi}_{\text{eff}}$ and $\hat{h}_{\text{eff}}$. But when V is also discrete, we modify Steps 1 and 2 in that we leave $f(X \mid V)$ completely unrestricted, so that $\eta_2 \equiv \theta_2$ is "infinite dimensional"; we estimate $f(X \mid V)$ by the nonparametric unsmoothed estimator

$$f(X \mid V; \hat{\theta}_2) \equiv f(X \mid V; \hat{\eta}_2) \equiv \sum_i \Delta_i \pi_i^{-1} I(V_i = V) I(X_i = X) \div \sum_i \Delta_i \pi_i^{-1} I(V_i = V); \quad (30)$$

and we estimate $\eta^* = (\eta_1', \eta_3', \eta_4')'$ by $\hat{\eta}^*$ solving $\sum_i \Delta_i \pi_i^{-1} \partial \log \mathcal{L}^F(\hat{\alpha}, \hat{\eta}_2, \eta^*; L_i)/\partial\eta^* = 0$. A mean value expansion shows that $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$ remains asymptotically equivalent to $\hat{\alpha}(h^\dagger, \phi^\dagger)$; thus $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$ will attain the bound if the parametric models $f[\varepsilon \mid X, V; \eta_1]$ and $f[V^\dagger \mid Y, X^*; \eta_4]$ are correctly specified.

Suppose that, as in our simulation experiment, Y is Bernoulli and data on extraneous covariates $V^\dagger$ were not collected, so $Y = W^\dagger$. Then Equation (1) completely specifies the conditional law of Y given $X^* \equiv (X', V')'$, and the parameters θ_1 and θ_4 are not present in $\mathcal{L}^F(\alpha, \theta; L)$. Thus if V is discrete, then $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$ will be globally semiparametric efficient, provided we estimate $f(X \mid V)$ by (30). In fact, it can be shown that $\hat{\alpha}(\hat{h}_{\text{eff}}, \hat{\phi}_{\text{eff}})$ will be globally efficient even if $\hat{h}_{\text{eff}}$ and $\hat{\phi}_{\text{eff}}$ are obtained by estimating conditional ex-

pectations given W using Equation (10) rather than the law $\mathcal{L}^F(\hat{\alpha}, \hat{\eta}; L)$.

5.3 X Independent of Y Given V

When X is independent of Y given V and $Y = W^\dagger$, it is straightforward to check that (23) is solved by $h_{\text{ind}}(X^*) = h^F_{\text{eff}}(X^*)\sigma^2(V)t(V) + E[h^F_{\text{eff}}(X^*)|V]\{1 - \sigma^2(V)t(V)\}$ and $\phi^{h_{\text{ind}}}(W) = E[h^F_{\text{eff}}(X^*)|V]\varepsilon$, where $t(V) = E[\varepsilon\varepsilon'/\pi|V]^{-1}$, $\sigma^2(V) \equiv E[\varepsilon\varepsilon'|V]$, and $h^F_{\text{eff}}(X^*)$ is given by (17). Thus if $g(X^*; \alpha) = l(\alpha_0 + \alpha_1 X + \alpha_2 V)$ for known $l(\cdot)$, then $\hat{\alpha}(h_{\text{ind}}, \phi^{h_{\text{ind}}})$ will be locally semiparametric efficient at the joint restriction that $\alpha_{0,1} = 0$ and that ε is independent of X given V.

6. ESTIMATION OF THE SELECTION PROBABILITIES, COMPARISONS WITH PREVIOUS ESTIMATORS, AND DATA MISSING BY HAPPENSTANCE

6.1 Estimation of Selection Probabilities

We will show that we can improve on the efficiency of inefficient estimators in our class by estimating the selection probabilities $\pi(W)$ even when they are known. Related phenomena have been noted by Robins and Morgenstern (1987), Rosenbaum (1987), Lancaster (1990), Robins and Rotnitzky (1992), Robins, Mark, and Newey (1992), and Imbens (1992). Suppose that $W' = (W^{\dagger\prime}, V')$ is discrete, and let $\hat{\pi}(w) \equiv \sum_i \Delta_i I(W_i = w)/\sum_i I(W_i = w)$ be the empirical probability of selection into the validation sample given $W = w$. Set $\hat{\pi} = \hat{\pi}(W)$, and let $\hat{D}(\alpha, h, \phi) = \Delta h(X^*)\varepsilon(\alpha)/\hat{\pi} - (\Delta - \hat{\pi})\phi/\hat{\pi}$. A corollary to Proposition 6.1 below is the following.

Corollary 6.1. Under model "obs" with W discrete, with probability approaching 1 there exists a unique solution $\hat{\tilde{\alpha}}(h, \phi)$ to $\hat{\tilde{D}}(\alpha, h, \phi) \equiv n^{-1}\sum_i \hat{D}_i(\alpha, h, \phi) = 0$. Further, $\hat{\tilde{\alpha}}(h, \phi)$ is asymptotically equivalent to $\hat{\alpha}(h, \phi^h)$.

As predicted by Corollary 6.1, reading from rows 2 and 4 of Table 1 we observe that $\hat{\tilde{\alpha}}_1(h^F_{\text{eff}}, 0)$ and $\hat{\alpha}_1(h^F_{\text{eff}}, \phi^{h^F_{\text{eff}}})$ are essentially equally efficient in our simulation experiment. Indeed, Corollary 6.1 implies that $\hat{\tilde{\alpha}}(h, \phi)$ is always more efficient than $\hat{\alpha}(h, \phi)$ unless $\phi = \phi^h$. In fact Corollary 6.1 implies that $\hat{\tilde{\alpha}}(h_{\text{eff}}, 0)$ is a pseudo–complete-case estimator that attains the efficiency bound, where we define a pseudocomplete case estimator of α_0 to be an estimator that uses the nonvalidation sample (i.e., the cases with incomplete data) only to estimate the selection probabilities $\pi(w)$.

We now present a heuristic argument as to why replacing the known probability π by an estimate $\hat{\pi}$ can improve the efficiency of an inefficient estimator. Suppose that in the example of Section 2.1, there were 100 subjects with $W_i = w$ for a particular realization w and that $\pi(w) = .5$. Also suppose that due to sampling variability, only 40 of the 100 subjects were selected into the validation sample, so that $\hat{\pi}(w) = .4$. Then using the complete case estimating equation $0 = \bar{D}(\alpha, h, 0) = n^{-1}\sum_i \Delta_i h(X_i^*)\varepsilon_i(\alpha)/\pi_i$ is tantamount to assuming that $40/.5 = 80$ subjects had $W_i = w$. On the other hand, using $\hat{\bar{D}}(\alpha, h, 0) = n^{-1}\sum_i \Delta_i h(X_i^*)\varepsilon_i(\alpha)/\hat{\pi}_i$ is tantamount to correctly assuming that $40/.4 = 100$ subjects had $W_i = w$. Therefore, $\hat{\bar{D}}(\alpha, h, 0)$ will be more precise than $\bar{D}(\alpha, h, 0)$ as an estimator of the unobservable "full" data estimating function $\bar{D}^F(\alpha, h) = n^{-1}\sum_i h(X_i^*)\varepsilon_i(\alpha)$. Finally, because $n^{-1}\partial\hat{\bar{D}}(\alpha_0, h, 0)/\partial\alpha$ and $n^{-1}\partial\bar{D}(\alpha_0, h, 0)/\partial\alpha$ both converge to the same limit, $-\kappa(h)$, $\hat{\tilde{\alpha}}(h, 0)$ will be more efficient than $\hat{\alpha}(h, 0)$. The preference for $\hat{\tilde{\alpha}}(h, 0)$ can also be viewed in terms of conditional bias rather than unconditional efficiency. Specifically, $\hat{\alpha}(h, 0)$, in contrast to $\hat{\tilde{\alpha}}(h, 0)$, is asymptotically biased conditional on the approximate ancillary statistic $n^{1/2}(\hat{\pi} - \pi)/\text{var}^A\{n^{1/2}(\hat{\pi} - \pi)\}$.

6.2 Comparison with Previously Proposed Estimators

6.2.1 Missing Completely at Random. A number of alternative estimators of α_0 in model "obs" have been proposed in the special case considered in our simulation experiment, where Y is Bernoulli, $Y = W^\dagger$, and π is a constant ρ. Pepe and Fleming (1991) and Carroll and Wand (1991) proposed estimating α_0 by $\hat{\alpha}_{\text{PFCW}}$ that maximizes $\prod_i \mathcal{L}(\alpha, \hat{\theta}; \Delta_i, L_{\text{obs},i})$ of (16b), except that a nonparametric estimate calculated from the validation sample data is substituted for $f(X|V; \theta_2)$. For V discrete, Pepe and Fleming proposed the estimator (30). If V has continuous components, then the kernel estimators proposed by Carroll and Wand are used. In Appendix C we use Corollary 5.1 to show that, when $g(\cdot; \alpha)$ is logistic, $\hat{\alpha}_{\text{PFCW}}$ is asymptotically equivalent to $\hat{\alpha}(h_{\text{PFCW}}, \phi_{\text{PFCW}})$, with $\phi_{\text{PFCW}} = E[X^*\varepsilon|W]$, $h_{\text{PFCW}}(X^*) = \rho X^* + (1 - \rho)\{E[X^*\varepsilon|Y = 1, V] - E[X^*\varepsilon|Y = 0, V]\}$. When Y is independent of X given V, h_{PFCW} and ϕ_{PFCW} equal h_{ind} and $\phi^{h_{\text{ind}}}$ of Section 5.3, and thus $\hat{\alpha}_{\text{PFCW}}$ will be efficient (Pepe and Fleming 1991). Comparing rows 3 and 5 of Table 1 reveals that when $\alpha_{0,1} \neq 0$, $\hat{\alpha}_{\text{PFCW}}$ is inefficient. To understand why $\hat{\alpha}_{\text{PFCW}}$ is inefficient, suppose that X were also discrete, so that θ_2 in the model $f(X|V; \theta_2)$ is a finite vector of parameters. Then the MLE of α is obtained by maximizing $\prod_i \mathcal{L}(\alpha, \hat{\theta}_2(\alpha); \Delta_i, L_{\text{obs},i})$, where $\hat{\theta}_2(\alpha)$ is the restricted MLE of θ_2 given α, obtained by maximizing (16b) over θ_2 for fixed α. $\hat{\theta}_2(\alpha)$ will depend on data from both the validation and nonvalidation sample members and is efficient for θ_2 when $\alpha = \alpha_0$ is known. In contrast, when $\alpha_{0,1} \neq 0$, Equation (30) is inefficient for θ_2. As a result, $\hat{\alpha}_{\text{PFCW}}$ is inefficient for α_0.

6.3 Missing at Random and Two-Stage Case-Control Designs

Until the final two paragraphs of this subsection, we restrict attention to the special case considered in Section 5.2 and our simulation experiment, in which $Y = W^\dagger$, Y is Bernoulli, and V is discrete. We allow selection to depend on $W = (Y, V')'$. In this setting, Manski and Lerman (1977) and Kalbfleisch and Lawless (1988) generalized an idea of Horvitz and Thompson (1952) and proposed the complete case estimator $\hat{\alpha}_{\text{ML}}(l_{\text{ML}})$ solving $\sum_i \Delta_i D_{\text{ML},i}(l_{\text{ML}}, \alpha) = 0$, where for any $l(X^*, \alpha)$, $D_{\text{ML}}(l, \alpha) = l(X^*, \alpha)\varepsilon(\alpha)/\pi$ and $l_{\text{ML}}(X^*, \alpha) \equiv \{\partial g(X^*; \alpha)/\partial\alpha\}[g(X^*; \alpha)\{1 - g(X^*; \alpha)\}]^{-1}$. Flanders and Greenland (1991) proposed the pseudo–complete-case estimator $\hat{\tilde{\alpha}}_{\text{ML}}(l_{\text{ML}})$ which replaces π by the empirical selection probability $\hat{\pi}$ in the definition of $D_{\text{ML}}(l, \alpha)$. It is straightforward to show that $\hat{\alpha}_{\text{ML}}(l)$ and $\hat{\tilde{\alpha}}_{\text{ML}}(l)$ are asymptotically equivalent to $\hat{\alpha}(h, 0)$ and $\hat{\tilde{\alpha}}(h, 0)$, with $h(X^*)$

$= l(X^*, \alpha_0)$. This becomes an exact algebraic equivalence when, as in our simulation study (row 4), $g(X^*; \alpha)$ is logistic since then $l_{ML}(X^*, \alpha) = X^*$. It follows that the Flanders–Greenland estimator $\hat{\alpha}_{ML}(l_{ML})$ can be made semiparametric efficient by replacing $l_{ML}(X^*, \alpha)$ by $\hat{h}_{eff}(X^*)$ of Section 5.2. In the same setting, Manski and McFadden proposed the complete-case estimator $\hat{\alpha}_{MM}(q_{MM})$, where for any $Q(\alpha) = q(Y, X^*, \alpha)$, $\hat{\alpha}_{MM}(q)$ solves $0 = \sum_i \Delta_i D_{MM,i}(q, \alpha)$, with $D_{MM}(q, \alpha) = \{Q(\alpha) - E_\alpha[\pi Q(\alpha) | X^*]/E_\alpha[\pi | X^*]\}$ and $q_{MM}(Y, X^*, \alpha) = l_{ML}(X^*, \alpha)\varepsilon(\alpha)$. Breslow and Cain (1988) proposed the pseudo–complete-case estimator $\hat{\hat{\alpha}}_{MM}(q_{MM})$ that replaced π by $\hat{\pi}$. As is evident from Table 1, $\hat{\hat{\alpha}}_{MM}(q_{MM})$ in row 6 is more efficient than $\hat{\alpha}_{MM}(q_{MM})$ in row 1, but neither is semiparametric efficient. In Appendix C we prove that $\hat{\alpha}_{MM}(q)$ and $\hat{\hat{\alpha}}_{MM}(q)$ are asymptotically equivalent to $\hat{\alpha}(h, 0)$ and $\hat{\hat{\alpha}}(h, 0)$, where $h(X^*)$ satisfies $q(Y, X^*, \alpha) = \pi^{-1}h(X^*)\varepsilon(\alpha)$ or, equivalently,

$$h(X^*) = \{\pi(1, V)q(1, X^*, \alpha_0) - \pi(0, V)q(0, X^*, \alpha_0)\} - \{\pi(1, V) - \pi(0, V)\}E[Q(\alpha_0) | \Delta = 1, X^*]. \quad (31)$$

It follows that the Breslow–Cain estimator could have been made semiparametric efficient by replacing $q_{MM}(Y, X^*, \alpha)$ with $\pi^{-1}\hat{h}_{eff}(X^*)\varepsilon(\alpha)$. Further, when, as in our simulation experiments, π is constant, the Breslow–Cain and Flanders–Greenland estimators are asymptotically equivalent. Zhao and Lipsitz (1992) discuss estimators in the class $\hat{\hat{\alpha}}_{ML}(l)$ and $\hat{\hat{\alpha}}_{MM}(q)$.

As mentioned in Section 3, the efficiency bound in model "obs" is unchanged when the assumption (4) that π is known is not imposed. Indeed, $\hat{\hat{\alpha}}(\hat{\hat{h}}_{eff}, 0)$ attains the bound and yet does not depend on π, where $\hat{\hat{h}}_{eff}$ is $\hat{h}_{eff}$ of Section 5.2 with $\hat{\pi}$ replacing π. In contrast, as we now argue, the asymptotic variance of the optimal complete case estimator will differ depending on whether π is or is not known. To be precise, let model "obs-complete" be the model "obs" with all nonvalidation sample data discarded and π possibly unknown. Imbens and Lancaster (1991) referred to model "obs-complete" as the "Bernoulli sampling" model and showed that with π *unknown*, (a) this model is equivalent to (i.e., has the same likelihood function as) the stratified sampling discrete choice model previously considered by Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981), and others, and (b) $\hat{\alpha}_{Cosslett}$ proposed by Cosslett (1981, sec. 2.14) and $\hat{\alpha}_{Imbens}$ proposed by Imbens (1992, sec. 3.2) are semiparametric efficient in this model. In Appendix A we prove that Manski–McFadden estimator $\hat{\alpha}_{MM}(q_{MM})$ is semiparametric efficient in the model "obs-complete" when π is *known*.

Although no complete-case estimator will be semiparametric efficient in model "obs" for all components of α_0, in a logistic model saturated with respect to a discrete V there exist complete-case estimators that are efficient for the slope parameters for X. Specifically, suppose that V has T levels, $(1, 2, \ldots, T)$, and $g(X^*; \alpha) = 1/[1 + \exp\{-\sum_{t=1}^{T} \alpha_t^{(1)} X_1 I(V = t) + \alpha_t^{(2)} X_2 I(V = t) + \alpha_t^{(3)} I(V = t)\}]$, with $\alpha' = (\alpha^{(1)'}, \alpha^{(2)'}, \alpha^{(3)'})$ and $\alpha^{(j)} = (\alpha_1^{(j)}, \ldots, \alpha_T^{(j)})'$ and $\alpha^{(1)}$ and $\alpha^{(2)}$ are the slope parameters. Suppose that for the moment we strengthen both model "obs" and "obs-complete" by imposing the additional assumption that the marginal distribution of W, $f(W)$, is known. Then the efficiency bound in the two strengthened models will be identical, because the nonvalidation sample data can be useful only for estimating $f(W)$ but $f(W)$ is now known. But Cosslett (1981, sec. 2.20) showed that $\hat{\alpha}^{(j)}_{Cosslett}$, $\hat{\alpha}^{(j)}_{MM}(q_{MM})$, and $\hat{\alpha}^{(j)}_{val}$, $j = 1, 2$, are asymptotically equivalent and semiparametric efficient with π either known or unknown in the strengthened model "obs-complete." Here $\hat{\alpha}_{val}$ maximizes $\prod_i \Delta_i \mathcal{L}^F(\alpha, \theta; L_i)$. Hence these same complete-case estimators must be efficient in the less restrictive model "obs." Weinberg and Wacholder (1992) obtained this result for X discrete. Finally, Cosslett (1981, sec. 2.21) proposed an estimator based on a nonparametric maximum likelihood estimate of the law of X given V that uses nonvalidation data and proved that this estimator is semiparametric efficient in model "obs." Cosslett (1981) and Imbens (1992) noted that this estimator may be difficult to compute, although Weinberg and Wacholder (1992) recently proposed using the EM algorithm to simplify the computation.

Two-Stage Case-Control Studies. We now extend our results to include two-stage case-control designs with $g(X^*; \alpha)$ logistic, Y Bernoulli. Data on extraneous surrogates $V^\dagger$ may again be available. In the first stage, we select n subjects. Specifically, for $i = 1, \ldots, n$, with probability γ_0, subject i is randomly selected from the cases $(Y = 1)$, and with probability $(1 - \gamma_0)$, subject i is randomly selected from the controls $(Y = 0)$, and W_i is recorded. In the second stage, first-stage subjects are independently selected with probability $\pi(W_i)$ to have X_i recorded. The design described in Section 2.1 is the special case of this two-stage design with $\gamma_0 = \Pr(Y = 1)$. Let model "obs†" be model "obs," except that we now allow γ_0 and $\Pr(Y = 1)$ to differ and γ_0 is unknown, and write the logistic model (2) as $g(X^*; \alpha) = [1 + \exp\{-[a + b'X^*_{-1}]\}]^{-1}$, where $\alpha_0 = (a_0, b_0')'$ and X^*_{-1} is X^* less the component that is the constant 1. Write $\hat{\alpha}(h, \phi)' = (\hat{a}(h, \phi), \hat{b}(h, \phi)')$. In Appendix C we prove the following lemma.

Lemma (6.1). In model "obs†" with $g(X^*; \alpha)$ given by (2), (a1) any RAL estimator of b_0 is asymptotically equivalent to some $\hat{b}(h, \phi)$; (a2) the estimator given in Proposition 2.2b is consistent for $\text{var}^A(n^{+1/2}[\hat{b}(h, \phi) - b_0])$; (b) $\hat{b}(\hat{h}_{eff}, \hat{\phi}_{eff})$ as defined in Section 4.2 for $W^\dagger$ nondiscrete and in Section 5.2 for $W^\dagger$ discrete is locally semiparametric efficient for b (even were γ_0 known); (c) if fixed fractions γ^* and $1 - \gamma^*$ of the n subjects are randomly sampled from the cases $(Y = 1)$ and controls $(Y = 0)$, then parts (a) and (b) remain true.

When V is discrete and the selection probabilities $\pi(W)$ do not depend on the differentially misclassified measurements $V^\dagger$, the b component of the Flanders–Greenland (1991) and Breslow–Cain estimators (1988) are consistent for b_0 but inefficient. Carroll et al. (1993) and Carroll, Wang, and Wang (1993) proposed estimators of b_0 in settings in which the selection probabilities may depend on $V^\dagger$; however, in contrast to our estimators $\hat{b}(h, \phi)$, for consistency the estimators proposed by these authors require a correctly specified parametric model for the law of $V^\dagger$ given (Y, X^*).

6.4 Arbitrary W

We now return to the general case of model "obs"; in particular, W may now again have continuous components. We suppose that we have a correctly specified parametric selection model for π; that is, Equation (18) holds. Let $\hat{\psi}$ maximize $\prod_i \mathcal{L}_i^{\text{mis}}(\psi)$ so that $0 = \sum_i S_{\psi,i}(\hat{\psi})$. Define $D(\alpha, h, \phi, \psi)$ like $D(\alpha, h, \phi)$, but with the true selection probabilities replaced by those specified by the parameter ψ in the selection model. Let $\hat{\alpha}(h, \phi, \psi)$ solve $\sum_i D_i(\alpha, h, \phi, \psi) = 0$ so $\hat{\alpha}(h, \phi, \psi_0) = \hat{\alpha}(h, \phi)$. Define $A^{\otimes 2} = AA'$. In Appendix B we prove the following.

Proposition 6.1. Given a correctly specified model for the missingness process with score $S_\psi = A(\phi_\psi)$ in model "obs," (a) with probability approaching 1, $\hat{\alpha}(h, \phi, \hat{\psi})$ exists and is unique; (b) $\hat{\alpha}(h, \phi, \hat{\psi})$ is a regular asymptotically linear estimator of α_0 with influence function $\{\kappa(h)\}^{-1}\text{Resid}\{D(h, \phi), S_\psi\} = \{\kappa(h)\}^{-1}D(h, \phi_*)$, where $\text{Resid}(A, B) = A - E(AB')\{\text{var}(B)\}^{-1}B$ is the residual from the population least squares regression of A on B, and $\phi_* \equiv \phi + E[D(h, \phi)S_\psi'][\text{var}(S_\psi)\}^{-1}\phi_\psi$; (c) $\text{var}^A[n^{1/2}\{\hat{\alpha}(h, \phi) - \alpha_0\}] \geq \text{var}^A[n^{1/2}\{\hat{\alpha}(h, \phi, \hat{\psi}) - \alpha_0\}] \geq \text{var}^A[n^{1/2}\{\hat{\alpha}(h, \phi^h) - \alpha_0\}]$, where the first inequality is strict unless $E[D(h, \phi)S_\psi'] = 0$, and the second inequality is strict unless $\phi_* = \phi^h$ (i.e., unless $b\phi_\psi = \phi^h - \phi$ for some matrix b); (d) given J nested correctly specified models, $j = 1, \ldots, J$, for the missingness process ordered by the increasing dimension of the parameter vectors $\psi^{(j)}$, the asymptotic variance of $\hat{\alpha}(h, \phi, \hat{\psi}^{(j)})$ is nonincreasing with j; and (e) the asymptotic variance of $n^{-1/2}\{\hat{\alpha}(h, \phi, \hat{\psi}) - \alpha_0\}$ can be consistently estimated by $\{\hat{\kappa}(h, \hat{\psi})\}^{-1}\hat{\Omega}(h, \phi, \hat{\psi})\{\hat{\kappa}(h, \hat{\psi})\}^{-1}$; $\hat{\kappa}(h, \hat{\psi}) \equiv n^{-1}\sum_i \partial D_i(\alpha, h, \phi, \hat{\psi})/\partial\alpha'$; $\hat{\Omega}(h, \phi, \hat{\psi}) = n^{-1}\sum_i \hat{\text{Resid}}_i\{D(h, \phi), S_\psi\}^{\otimes 2}$, where $\hat{\text{Resid}}_i\{D(h, \phi), S_\psi\}$ is the residual for subject i from the least squares regression of the $D_i(\hat{\alpha}, h, \phi, \hat{\psi})$ on $S_{\psi,i}(\hat{\psi})$ for $i = 1, \ldots, n$.

To illustrate the efficiency advantage of collecting data on extraneous surrogates, in our simulation experiment we generated extraneous surrogates $V^\dagger = (V_1^\dagger, V_2^\dagger)'$ according to $V_j^\dagger \mid Y, X^* \sim N(X - .12 + .24Y, \sigma_j^2)$, with $\sigma_1^2 = \frac{1}{2}$, $\sigma_2^2 = \frac{1}{16}$, $\text{cov}(V_1^\dagger, V_2^\dagger \mid Y, X^*) = 0$. $\hat{\psi}^{(1)}$, used in row 7 of Table 1, was obtained by fitting the model

$$\text{logit}\,\pi(\psi^{(1)}) = \sum_{j=0}^{j=1}\sum_{m=0}^{m=1}\psi_{jm}I\{(Y, V) = (j, m)\} + \psi_1 V_1^\dagger + \psi_2 V_1^\dagger Y. \quad (32)$$

Note that the true values of ψ_1 and ψ_2 are 0, because $\pi = .1$ is constant. $\psi^{(2)}$ of row 8 added $\psi_3 V_2^\dagger + \psi_4 V_2^\dagger Y$ to model (32). Comparison of the efficiencies of row 4 with row 7 and of row 7 with row 8 illustrates Proposition 6.1d, because row 4 is based on fitting the model (32) with ψ_1 and ψ_2 set to 0 a priori. Note that although the row 3 estimator is semiparametric efficient when $W' = (Y, V)$, it is less efficient than the row 8 estimator based on $W' = (Y, V, V_1^\dagger, V_2^\dagger)$, which illustrates the efficiency advantage of collecting data on extraneous surrogates. The estimators in rows 4 and 7–9 can be computed by a canned logistic regression program that allows for individual weights (i.e., $\Delta_i\pi_i(\hat{\psi})^{-1}$), although the standard errors output by the program are based on the inverse of the (weighted) information matrix and thus will be invalid.

Although our proof of Proposition 6.1 assumes $\pi(\hat{\psi})$ to be $n^{1/2}$-consistent for π, results of Newey (1993a) suggest that it is necessary and sufficient for $\pi(\hat{\psi})$ to be greater than $n^{1/4}$ consistent for π, which limits the number of free parameters in model $\pi(\psi)$. The effect of moderate overparameterization of $\pi(\psi)$ is illustrated by the fact that the estimated asymptotic relative efficiency (ARE) (i.e., Monte Carlo efficiency) in row 9 is less than that in row 8, even though $\pi(\psi^{(3)})$ added the terms $\psi_5 V_1^{\dagger 2} + \psi_6 V_2^{\dagger 2} + \psi_7 V_1^{\dagger 2} Y + \psi_8 V_2^{\dagger 2} Y$ to model $\pi(\psi^{(2)})$ of row 8. Thus the finite sample efficiencies recorded in the table conflict with the theoretical asymptotic efficiencies of Proposition 6.1d. In our experience, moderate overparameterization of $\pi(\psi)$ produces significant finite sample bias in our estimated variance of $\hat{\alpha}$ but little bias in $\hat{\alpha}$ itself, suggesting that in this setting, inference (e.g., confidence intervals) should be based on bootstrap estimates of the variability of $\hat{\alpha}$ rather than on the variance estimator of Proposition 6.1e.

Nonindependent Sampling. Proposition 6.1 can be extended to settings in which the assumption that the $(Y_i, X_i, \Delta_i, V_i)$, $i = (1, \ldots, n)$, are independent random vectors is inappropriate. As an example, suppose that $W = (W^{*\prime}, W^{**\prime})'$, W^* is discrete with S levels $w_1^*, \ldots, w_S^*$ and, following Breslow and Cain (1988), the investigator uses a "fixed fraction" sampling design in which he or she selects a fraction ρ_s without replacement from the N_s subjects in level w_s^* of W^* into the validation sample for $s = 1, \ldots, S$. Because one cannot sample a fractional person, we let $\rho_s \equiv \rho_s(N_s)$ depend on N_s; however, we assume constants $\pi(w_s^*) > 0$ such that $\rho_s(N_s) - \pi(w_s^*)$ is $O_p(N_s^{-1})$. For example, given constants $\pi(w_s^*)$, we could set $\rho_s(N_s)N_s$ to be the smallest integer greater than or equal to $\pi(w_s^*)N_s$. We continue to assume that the L_i, $i = (1, \ldots, n)$, are independent. On the other hand, the Δ_i are not independent conditional on $\{W_i; i = 1, \ldots, n\}$, because the sampling design ensures that $N_s^{-1}\sum_i \Delta_i I(W_i^* = w_s^*) = \rho_s$. In Appendix B, we prove the following lemma.

Lemma 6.2. In a model characterized by (1), data (7), and a "fixed fraction" sampling design, the estimators $\hat{\alpha}(h, \phi, \hat{\psi})$ remain asymptotically normal and unbiased for α_0 with asymptotic variance that can still be consistently estimated as in Proposition 6.1e, provided that the model $\pi(\psi)$ used to estimate π has nested within it the saturated model in W^*,

$$\text{logit}\,\pi(W; \psi^*) = \sum_{s=1}^{S}\psi_s^* I(W^* = w_s^*), \quad \psi^* = (\psi_1^*, \ldots, \psi_S^*)'. \quad (33)$$

6.5 Data Missing by Happenstance

In many epidemiologic studies data are missing by happenstance rather than design, and thus the missingness mechanism $\pi(W)$ is not known. For example, a subset of subjects may simply refuse to have blood drawn for Vitamin

A and E measurements. In this setting suppose that we were willing to continue to assume that the restrictions of model "obs" held except for the assumption that the missingness mechanism is completely known, but that we correctly specify a parametric model for the missing data mechanism with score $S_\psi(\psi)$. Denote the resulting semiparametric model by "obs*". Then in Appendixes A and B we prove the following.

Proposition 6.2. In model "obs*" (a) the conclusions of Proposition 6.1 remain true, (b) the influence function of any RAL estimator of α_0 lies in the set $\{\text{Resid}[\kappa(h)^{-1}D(h, \phi), S_\psi]\}$, and (c) the efficient score remains $D(h_{\text{eff}}, \phi_{\text{eff}})$.

Part (b) of Proposition 6.2 can be restated as follows: When the missingness process is known only up to an unknown parameter ψ_0, the influence of any RAL estimator lies in the set whose elements consist of the residual from the population regression of an arbitrary "influence function of an RAL estimator when the missingness process is completely known" on the "score S_ψ for the missingness process." As an alternative to replacing π by $\pi(\hat{\psi})$ in defining the estimator $\hat{\alpha}(h, \phi, \hat{\psi})$, we could have estimated π by a nonparametric regression estimator $\hat{\pi}$ to protect against misspecification bias. Newey (1993a) considered nonparametric regression estimates based on series and provided regularity conditions under which the resulting estimator of α_0 would be asymptotically equivalent to $\hat{\alpha}(h, \phi^h)$. Of course, if the dimension of W is large, such asymptotic results are of little practical relevance due to the "curse of dimensionality," and some (not necessarily parametric) model for $\pi(W)$ will be required.

7. OTHER MISSING DATA PATTERNS

We have assumed that any subjects with incomplete data have exactly the same set of covariates missing; that is, X_1 and X_2 were either both observed or both unobserved. In this section this assumption is relaxed.

7.1 Monotone Missing-Data Patterns

Suppose that it were more expensive to assay Vitamin E (X_2) than Vitamin A (X_1). Then the following sequential design might be used in the study of Section 2.1. An initial sample of subjects is randomly selected to have X_1 assayed. A subsample of this initial sample is then randomly selected to have X_2 assayed. The probability of selection into the initial sample may depend on W. The probability of selection into the subsample may depend on both W and X_1. $\hat{\alpha}(h, \phi)$, as previously defined, may now be inconsistent since X may no longer be independent of Δ given W, so $E[D(\alpha, h, \phi)] \neq E[D^F(h)] = 0$. Further, if selection does not depend on X_1, then $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$, although consistent, may now be inefficient, because it ignores data on X_1 when X_1 is observed but X_2 is missing. Note that we can generalize this sequential design to a vector $X = (X_1, \ldots, X_M)$ of $M \geq 2$ exposures. Indeed, it will be both more general and more convenient to consider a sequential design that would allow the components of W as well as X to be missing. To do so, write $W' \equiv (W_1, \ldots, W_{M^\dagger})$ and $L' \equiv (L_1, \ldots, L_K) = (W', X')$, $K = M^\dagger + M$ with each component L_k and W_k of L and W univariate. Thus $L_1 = W_1$, $L_{M^\dagger} = W_{M^\dagger}$, $L_{M^\dagger+1} = X_1$, and $L_K = X_M$. Set $L_0 \equiv 0$. Let $R_k = 1$ if L_k is observed and $R_k = 0$ otherwise, with $R_0 \equiv 1$. Let $R = (R_1, \ldots, R_K)'$ be the K vector of missing data indicators with realization r. Let $L_{(r)}$ be the vector of observed components of L when $R = r$ and let $\mathbf{1}_j, j = (0, \ldots, K)$, be the K vector in which the first j components are 1 and the last $K - j$ components are 0. We shall say there is a monotone missing-data pattern if, as in our sequential design, the observed data for each subject i is

$$\{R, L_{(R)}\}, \tag{34a}$$

with

$$R = \mathbf{1}_j \quad \text{for some} \quad j \in \{0, \ldots, K\} \text{ with probability 1.} \tag{34b}$$

Equation (34b) implies that $R_K = 1$ if and only if $I(R = \mathbf{1}) = 1$, where $\mathbf{1} \equiv \mathbf{1}_K$ is the K vector of 1s and $I(R = \mathbf{1})$ is the indicator for complete data on L. Data (7) was the special case of monotone missingness, in which with probability 1, either $R = \mathbf{1}_{M^\dagger}$ or $R = \mathbf{1}$ with $\Delta = I(R = \mathbf{1})$ and $L_{\text{obs}} = L_{(R)}$. The data in our sequential design were *missing at random* in the sense of Rubin (1976). That is, the probability of observing $R = r$ depends only on the observed components $L_{(r)}$ of L,

$$\pi(r) \equiv \Pr[R = r|L] = \Pr[R = r|L_{(r)}]. \tag{35}$$

When (35) is true, we write $\pi(r)$ as $\pi(r, L_{(r)})$ when we wish to make explicit the dependence on $L_{(r)}$. Suppose that

$$\pi(\mathbf{1}) > \sigma > 0 \quad \text{with probability 1,} \tag{36}$$

so each subject has a positive probability of having complete data, and that

$$\pi(r) \text{ is known.} \tag{37}$$

Redefine the semiparametric model "obs" to be the model characterized by (1), (35)–(37), and the observed data (34a). To obtain consistent estimates of α_0 in this model, let $\bar{L}_k \equiv (L_0, L_1, \ldots, L_{k-1})'$ be shorthand for $L_{(r)}$ with $r = \mathbf{1}_{(k-1)}$. Hence $\bar{L}_{K+1} = L$. Put $\bar{L}_0 = 1$. Define $\pi_k = \Pr[R_k = 1 | R_{k-1} = 1, \bar{L}_k]$, $\bar{\pi}_k = \prod_{m=1}^{k} \pi_m$, $\pi_0 = \bar{\pi}_0 = 1$. When, as we assume in this subsection, (34b) holds and thus missingness is monotone, it can be verified that $\pi(\mathbf{1}) = \bar{\pi}_K$ and that (35) is equivalent to $\Pr[R_k = 1 | R_{k-1} = 1, L] = \Pr[R_k = 1 | \bar{L}_k, R_{k-1} = 1] \equiv \pi_k$. Now redefine

$$D(\alpha, h, \phi) = I(R = \mathbf{1})D^F(\alpha, h)/\pi(\mathbf{1}) - A(\phi) \tag{38}$$

and

$$A(\phi) \equiv \sum_{k=1}^{K} (R_k - \pi_k R_{k-1})\bar{\pi}_k^{-1}\phi_k(\bar{L}_k), \tag{39}$$

with $\phi \equiv (\phi_1, \ldots, \phi_K)$ and $\phi_k \equiv \phi_k(\bar{L}_k)$ taking values in R^q. Furthermore, redefine $\mathcal{L}^{\text{mis}}(\psi_0)$ and ϕ_ψ of equation (19) as $\mathcal{L}^{\text{mis}}(\psi_0) = \prod_{k=1}^{K} \{\pi_k(\psi_0)^{R_k}\{1 - \pi_k(\psi_0)\}^{1-R_k}\}^{R_{k-1}}$, $\phi_{\psi,k} = \bar{\pi}_k \partial \text{ logit } \pi_k(\psi_0)/\partial\psi$, where $\pi_k(\psi)$ is a correctly specified model for π_k; that is,

$$\pi_k = \pi_k(\psi_0), \qquad k = (1, \ldots, K), \tag{40}$$

and $\pi_k(\psi) = \pi(\bar{L}_k, \psi)$ is a known function taking values in $(0, 1]$. In Section 8 and Appendixes A and B, we prove that

with these redefinitions, Propositions 2.2 and 3.2 and the restatement of Proposition 3.2 remain true. Furthermore, Propositions 2.3, 4.2, 6.1, and 6.2 remain true when we redefine $\phi^h \equiv (\phi_1^h, \ldots, \phi_K^h)$ and $\mathbf{r}(B)$ by

$$\phi_k^h(\bar{L}_k) = E[D^F(h)|\bar{L}_k] \tag{41}$$

and

$$\mathbf{r}(B) \equiv \sum_{k=1}^{K} (1 - \pi_k)\bar{\pi}_k^{-1} E[B|\bar{L}_k] \tag{42}$$

and replace π by $\pi(\mathbf{1})$ in the definition of $t(X^*)$ following equation (24). Set $\zeta(B) = R_K B/\bar{\pi}_K$. In view of the identity $E(B|\bar{L}_k) = E[\zeta(B)|\bar{L}_k, R_{k-1} = 1]\bar{\pi}_{k-1}$ given in Proposition 8.2b, an adaptive estimator $\hat{\alpha}(h, \hat{\phi}^h)$ can be constructed as in Section 2.7, with $\hat{\phi}_k^h$ obtained by multiplying $\bar{\pi}_{k-1}$ by the estimated predicted value from the regression of $\zeta[D^F(\hat{\alpha}, h)]$ on functions of $\bar{L}_k$ among subjects with $R_{k-1} = 1$.

7.2 Missing at Random with Arbitrary Missing-Data Patterns

We next consider estimating the parameter α_0 in model "obs" with arbitrary nonmonotone missing-data patterns. As an example, suppose that in the study of Section 2.1 an initial sample of approximately 4% was selected, in which one-half of the subjects had their Vitamin A (X_1) assayed and one-half had their Vitamin E (X_2) assayed. Suppose, further, that approximately 50% of subjects in whom Vitamin A had been assayed were selected to have their Vitamin E assayed as well (with the selection probability depending on their observed Vitamin A levels) and vice versa. Then the missing data pattern will be nonmonotone and yet satisfy (35)–(37). Redefine

$$A(\phi) = \{\pi(\mathbf{1})\}^{-1} I(R = \mathbf{1}) \times \left\{ \sum_{r \neq \mathbf{1}} \pi(r)\phi_r(L_{(r)}) \right\} - \sum_{r \neq \mathbf{1}} I(R = r)\phi_r(L_{(r)}), \tag{43}$$

where the sum is over all possible realizations of R other than $\mathbf{1}$, $\phi = \{\phi_r; r \neq \mathbf{1}\}$, and $\phi_r = \phi_r(L_{(r)})$ is a function of $L_{(r)}$ taking values in R^q. That is, $A(\phi) = \pi(\mathbf{1})^{-1} I(R = \mathbf{1}) E[\phi(R, L_{(R)})|L] - \phi(R, L_{(R)})$. Similarly, redefine $\mathcal{L}^{\text{mis}}(\psi_0) = \pi(R, \psi_0)$ and $\phi_{\psi,r} = \partial \log\{\pi(r, \psi_0)\}/\partial\psi$, where $\pi(r, \psi)$ is a correctly specified parametric model for $\pi(r)$ satisfying (35); that is,

$$\pi(r) = \pi(r; \psi_0) \tag{44}$$

and $\pi(r; \psi) = \pi(r, L_{(r)}, \psi)$ is a function of ψ taking values in (0, 1] satisfying $\sum_r \pi(r, \psi) = 1$ for each ψ with probability 1. In Section 8 and Appendixes A and B, we prove that with these redefinitions, (a) Propositions 2.2 and 3.2, and the restatement of Proposition 3.2a are true, and (b) Propositions 2.3, 4.1, 6.1, and 6.2 are true (1) when ϕ^h has elements $\phi_r^h \equiv E(B^{*h}|L_{(r)})$, where B^{*h} is defined to be the unique q vector-valued function of L satisfying

$$\mathbf{m}(B^{*h}) = h(X^*)\varepsilon, \tag{45}$$

with $\mathbf{m}(B^*) \equiv \sum_r \pi(r)E[B^*|L_{(r)}]$ and (2)

$$\mathbf{r}[h(X^*)\varepsilon] \equiv \{\pi(\mathbf{1})\}^{-1} \sum_{r \neq \mathbf{1}} \pi(r)E[B^{*h}|L_{(r)}]. \tag{46}$$

Note that with nonmonotone missing data, ϕ^h (the optimal choice of ϕ for fixed h) no longer exists in closed form. Also, Corollary 5.1 remains true, except with the condition $\phi_r(L_{(r)}) = b(r, L_{(r)})$ for $r \neq \mathbf{1}$ replacing condition (a) where $B = \sum_r I(R = r)b(r, L_{(r)})$. Note that even if the law of L had been known, because with arbitrary missing data patterns the solution $h_{\text{eff}}(X^*)$ to (23) depends on $B^{*h_{\text{eff}}}$ through (46) and $B^{*h_{\text{eff}}}$ depends on $h_{\text{eff}}(X^*)$ through (45), more work is required to obtain h_{eff} and ϕ_{eff}. In Appendix A, we show that $B^{*h_{\text{eff}}}$ solves

$$B^* = \{\partial g(X^*; \alpha_0)/\partial\alpha\}\{\text{var}(\varepsilon|X^*)\}^{-1}\varepsilon - \{\mathbf{m}(B^*) - B^*\} + E[\mathbf{m}(B^*) - B^*|X^*]\,\text{var}(\varepsilon|X^*)^{-1}\varepsilon. \tag{47}$$

We prove in Appendix A that the solution $B_{\text{eff}}^* \equiv B^{*h_{\text{eff}}}$ is unique and can be obtained by successive approximation. Then ϕ_{eff} has elements $\phi_{r,\text{eff}} = E[B_{\text{eff}}^*|L_{(r)}]$, $r \neq \mathbf{1}$. Further, by (45), $h_{\text{eff}} = (h_{\text{eff},1}, \ldots, h_{\text{eff},t})$, with $h_{\text{eff},j}(X^*) = E[\mathbf{m}(B_{\text{eff}}^*)|Y = e_j, X^*] - E[\mathbf{m}(B_{\text{eff}}^*)|Y = 0, X^*]$, with e_j as in Corollary 5.1. As in Section 4.2 we can obtain locally semiparametric efficient estimates by computing expectations under $\mathcal{L}^F(\hat{\alpha}, \hat{\eta}; L)$.

When the data are missing by happenstance and there are arbitrary missing-data patterns, it is necessary to specify and fit models $\pi(r, L_{(r)}, \psi)$ for $\pi(r, L_{(r)})$. Robins, Greenland, and Rotnitzky (1994) have provided appropriate models and fitting algorithms.

Finally suppose that only data on X are missing but there are arbitrary missing data patterns in X and (35) is false, so the data are not missing at random in the sense of Rubin (1976). Then $\hat{\alpha}(h, \phi)$, using $D(\alpha, h, \phi)$ and $A(\phi)$ as given by (38) and (43), can be inconsistent for α_0. But suppose that (3)–(5) remain true. Then, even with arbitrary missing-data patterns in X, $\hat{\alpha}(h, \phi)$ with $D(\alpha, h, \phi)$ and $A(\phi)$ based on (9) is a semiparametric estimator in the model defined by restrictions (1) and (3)–(5). Furthermore, a proof analogous to that of theorem 3 of Robins and Rotnitzky (1995) implies that $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$, with h_{eff} and ϕ_{eff} as initially defined in Section 4.2, will attain the semiparametric variance bound for this model.

8. ESTIMATION IN ARBITRARY SEMIPARAMETRIC MODELS WITH MISSING DATA

In Proposition 8.1 we provide representations for the efficient score and the influence function of any RAL estimator in an arbitrary semiparametric model with data missing at random. We prove Propositions 2.3, 3.1, 3.2, 4.2, and 6.2 by specializing to the case in which the "full data" semiparametric model is characterized by the conditional mean restriction (1).

8.1 An Arbitrary Semiparametric Model

Consider the generic semiparametric model with likelihood $\mathcal{L}(\alpha, \theta; Z)$ of Section 3. We define the nuisance tangent

space Λ to be the mean squared closure of the set of all random vectors bS_η, where S_η is the score for η in some regular parametric submodel (usually, $S_\eta = \partial \log \mathcal{L}(\alpha_0, \eta_0; Z)/\partial\eta$) and b is a conformable constant matrix with q rows (i.e., $\Lambda = \{A \in R^q : E[\|A\|^2] < \infty)$, and there exists $b_j S_{\eta,j}$ with $\lim_{j\to\infty} E[\|A - b_j S_{\eta,j}\|^2] = 0\}$, where each b_j is a matrix of constants and $\|A\|^2 = A'A$. We shall consider Λ as a subset of the Hilbert space of $q \times 1$ random vectors H with inner product $E[H_1'H_2]$ and $E[H'H] < \infty$. In our examples Λ is a linear subspace. The projection $\Pi(H|\Lambda)$ of any vector H on a closed linear space such as Λ exists and is the unique vector $A \in \Lambda$ minimizing $E[(H - A)'(H - A)]$. Π is the projection operator. $\Pi(H|\Lambda)$ is also the unique element of Λ satisfying $E[\{H - \Pi(H|\Lambda)\}'A] = 0$ for all $A \in \Lambda$. The semiparametric variance bound equals the inverse of the variance of $S_{\text{eff}} \equiv \Pi[S_\alpha|\Lambda^\perp]$, where S_α is the score for α (usually, $S_\alpha = \partial \log \mathcal{L}(\alpha_0, \theta_0; Z)/\partial\alpha$) and "$\perp$" denotes an orthogonal complement (Begun et al. 1983; Bickel et al. 1993). S_{eff} is called the efficient score.

Definition. For any set $\mathcal{F}$ of random variables, let $\mathcal{F}_0$ be the subset with mean 0. Let $\Lambda_0^\perp \equiv (\Lambda^\perp)_0$.

Lemma 8.1.

a. In any semiparametric model, the influence function of any RAL estimator of α_0 is in $\Lambda_{0*}^\perp \equiv \{E[AS_\alpha']^{-1}A;\ A \in \Lambda_0^\perp\} \equiv \{A \in \Lambda_0^\perp;\ E[AS_\alpha'] = I_{q\times q}\}$.
b. Further, $A \in \Lambda_0^\perp$ implies $E(AS_{\text{eff}}') = E(AS_\alpha')$. Theorem (2.2) of Newey (1990a) implies Lemma 8.1.

Even if $\Lambda_{0*}^\perp$ contains nonzero elements, Lemma 8.1 does not guarantee that any RAL estimator exists: however, in sufficiently smooth models, including all those studied in this paper, $\Lambda_{0*}^\perp$ can be identified with the space of influence functions of RAL estimators (Bickel and Ritov 1990; Newey 1990a).

We now specialize the foregoing general results to "full data" and "missing data" models. Again let $L' = (L_1, \ldots, L_K)$ be a multivariate random variable with each L_k univariate, let

$$\mathcal{L}^F(\alpha, \theta; L) \tag{48}$$

be the likelihood for a single subject when L is fully observed in a semiparametric model "full" indexed by $\alpha \in R^q$ and an infinite-dimensional nuisance parameter θ, and let S_α^F, Λ^F, S_{eff}^F, and $\Lambda_{0*}^{F,\perp}$ be the score for α, the nuisance tangent space, the efficient score, and the space of influence functions in model "full." Let the semiparametric model "obs" with likelihood $\mathcal{L}(\alpha, \theta; R, L_{(R)})$ now be characterized by the observed data (34a), restriction (48) on the law of L, the missing-at-random assumption (35), restriction (36), and the additional restriction that

$$\pi(r, L_{(r)}) \in \{\pi(r, L_{(r)}; \gamma): \gamma \in \boldsymbol{\gamma}\}, \tag{49}$$

where, for each γ, $\pi(r, L_{(r)}; \gamma)$ is a density for $\Pr[R = r|L]$ satisfying (35) and $\boldsymbol{\gamma}$ may be infinite dimensional. Henceforth let S_α, Λ, S_{eff}, and $\Lambda_{0*}^\perp$ be the score for α, the nuisance tangent space, the efficient score, and the space of influence functions in the missing data model "obs." Define the tangent space for the model (49) for the missing process to be $\Lambda^{(3)} = \{A^{(3)} \in R^q$; there exists $b_j S_{\psi,j}$ with $\lim_{j\to\infty} E[\|A^{(3)} - b_j S_{\psi,j}\|^2] = 0\}$, where S_ψ is the score at the truth for a regular parametric submodel $\pi(r, L_{(r)}; \psi)$ of the model $\pi(r, L_{(r)}; \gamma)$ and the b_j are constant matrices. We shall assume that $\Lambda^{(3)}$ is linear. Define $\Lambda^{(2)} \equiv \{A^{(2)} = a^{(2)}(R, L_{(R)}) \in R^q; E[A^{(2)}|L] = 0\}$ to be the space of functions of the observed data with mean 0 given the full data L. In Appendix A we prove the following.

Lemma 8.2.

a. $\Lambda^{(3)} \subseteq \Lambda^{(2)}$.
b. If the model (49) is completely nonparametric in the sense that it is unrestricted except for the condition $\sum_r \pi(r, L_{(r)}; \gamma) = 1$, then $\Lambda^{(2)} = \Lambda^{(3)}$.

Lemma 8.3. If (36) is true, then with $A(\phi)$ as in (43), (a) $E[A(\phi)|L] = 0$ and (b) given any $a^{(2)}(R, L_{(R)}) \in \Lambda^{(2)}$, $a^{(2)}(R, L_{(R)}) = A(\phi)$, with $-\phi_r(L_{(r)}) = a^{(2)}(r, L_{(r)})$ for $r \neq \mathbf{1}$. Hence $\Lambda^{(2)} = \{A(\phi)\}$.

Henceforth let $B = b(R, L_{(R)})$ and $D = d(R, L_{(R)})$ represent generic functions of $(R, L_{(R)})$ and let $B^* = b^*(L)$, and $D^* = d^*(L)$ represent generic functions of the full data L. Define the operators $\mathbf{g}$, $\mathbf{m}$, $\mathbf{u}$, and $\mathbf{v}$ by $\mathbf{g}(B^*) \equiv \sum_r I(R = r)E(B^*|L_{(r)}, R = r)$, $\mathbf{m}(B^*) \equiv E[\mathbf{g}(B^*)|L] = \sum_r \pi(r)E(B^*|L_{(r)}, R = r)$, $\mathbf{u}(B^*) = \{\pi(\mathbf{1})\}^{-1}I(R = \mathbf{1})B^*$, and $\mathbf{v}(B^*, A^{(2)}) = \mathbf{u}(B^*) + A^{(2)} - \Pi[\mathbf{u}(B^*) + A^{(2)}|\Lambda^{(3)}]$. When the data are missing at random (i.e., Eq. (35) is true), $E(B^*|L_{(r)}, R = r) = E(B^*|L_{(r)})$ and $\mathbf{g}(B^*)$ and $\mathbf{m}(B^*)$ simplify. Our fundamental result is the following rather abstract proposition, a verbal description of which is provided in Remark 1.

Proposition 8.1. In model "obs," with $A(\phi)$ as in (43):

a1. If $B \in \Lambda_0^\perp$, then the decomposition $B = \mathbf{u}\{E(B|L)\} + [B - \mathbf{u}\{E(B|L)\}]$ satisfies $E(B|L) \in \Lambda_0^{F,\perp}$ and $B - \mathbf{u}\{E(B|L)\} = A(\phi) \in \Lambda^{(2)}$, with $\phi_r(L_{(r)}) = b(r, L_{(r)})$ for $r \neq \mathbf{1}$.
a2. Further, if $B \in \Lambda_{0*}^\perp$, then $E(B|L) \in \Lambda_{0*}^{F,\perp}$.
b. $\Lambda_0^\perp = \{\mathbf{v}(B^*, A^{(2)});\ B^* \in \Lambda_0^{F,\perp}, A^{(2)} \in \Lambda^{(2)}\}$.
c1. If $B^* \in \Lambda_0^{F,\perp}$, $A^{(2)} \in \Lambda^{(2)}$, then $E[\mathbf{v}(B^*, A^{(2)})S_{\text{eff}}'] = E[B^*S_{\text{eff}}^{F\prime}]$.
c2. $\Lambda_{0*}^\perp = \{\mathbf{v}(B^*, A^{(2)}),\ B^* \in \Lambda_{0*}^{F,\perp},\ A^{(2)} \in \Lambda^{(2)}\} = \{[E(B^*S_{\text{eff}}^F)]^{-1}\mathbf{v}(B^*, A^{(2)});\ B^* \in \Lambda_0^{F,\perp},\ A^{(2)} \in \Lambda^{(2)}\}$.
d. $\Pi[\mathbf{u}(D^*)|\Lambda^{(2)}] = \mathbf{u}(D^*) - \mathbf{g}\{\mathbf{m}^{-1}(D^*)\} = A(\phi^*)$, where $\mathbf{m}^{-1}(D^*)$ is the unique B^* solving $\mathbf{m}(B^*) = D^*$ and $\phi_r^*(L_{(r)}) = E[\mathbf{m}^{-1}(D^*)|L_{(r)}]$ for $r \neq \mathbf{1}$.
e1. $S_{\text{eff}} = \mathbf{u}(D_{\text{eff}}^*) - \Pi[\mathbf{u}(D_{\text{eff}}^*)|\Lambda^{(2)}] = \mathbf{g}[\mathbf{m}^{-1}(D_{\text{eff}}^*)]$, with D_{eff}^* the unique $D^* \in \Lambda^{F,\perp}$ solving $\Pi[\mathbf{m}^{-1}(D^*)|\Lambda^{F,\perp}] = S_{\text{eff}}^F$.
e2. $B_{\text{eff}}^* \equiv \mathbf{m}^{-1}(D_{\text{eff}}^*)$ is the unique B^* satisfying $B^* = S_{\text{eff}}^F - \Pi[\mathbf{m}(B^*) - B^*|\Lambda^F]$ which can be solved by successive approximation.

Remark 1. When $\pi(r)$ is known and $\Lambda^{(3)}$ is thus empty, part c2 of Proposition 8.1 is equivalent to the restatement of Proposition 3.2a, except that to be precise, we should have referred to the set of "mean square limits of scores" rather than simply to the set of "scores." When model (49) is a parametric model with score S_ψ, part c2 of Proposition 8.1

is equivalent to the restatement of Proposition 6.2b, because in that case for any B, $B - \Pi(B|\Lambda^{(3)})$ is the residual from the population least squares regression of B on S_ψ. Proposition 8.1e1 and Eq. (36) implies both that $\text{var}(S_{\text{eff}})$ is invertible and that S_{eff} does not depend on $\Lambda^{(3)}$ and thus on the model (49) for the missingness process. It follows that, even if models (48) for the full data and (49) for the missing at random mechanism are completely unrestricted, the law of L and the response probabilities $\pi(r, L_{(r)})$ are locally identified. Proposition 8.1e2 is used to prove that the solutions to (23) and (47) are unique. The remaining parts of Proposition 8.1 are needed for the proofs of Propositions 2.3, 4.1, and 4.2. It can be seen from the proof provided in Appendix A, that when $\pi(r) \equiv \Pr[R = r|L]$ is completely known (i.e., $\Lambda^{(3)}$ is empty), Proposition 8.1 remains true even without the missing at random assumption (35). Van der Laan (1993, pp. 29–30) proved that $B^* = \mathbf{m}^{-1}(D^*)$ can be obtained by solving $B^* = D^* + [B^* - \mathbf{m}(B^*)]$ by successive approximation. The following proposition specializes parts of Proposition 8.1 to the case of monotone missingness.

Proposition 8.2. In model "obs," if missingness is monotone [i.e., Eq. (34b) is true], then, with $A(\phi)$ given by (39):

a1. $\mathbf{g}(D^*) = R_K D^* + \sum_{k=1}^{K} (R_{k-1} - R_k) E(D^*|\bar{L}_k)$.
a2. $\mathbf{m}(D^*) = \bar{\pi}_K D^* + \sum_{k=1}^{K} (1 - \pi_k)\bar{\pi}_{k-1} E(D^*|\bar{L}_k)$.
b. $E(D^*|\bar{L}_k) = E(D^*|\bar{L}_k, R_{k-1} = 1) = E(R_K D^*/\bar{\pi}_K|\bar{L}_k, R_{k-1} = 1)\bar{\pi}_{k-1}$.
c. $\Lambda^{(2)} = \{A(\phi)\}$.
d. $\Pi[\mathbf{u}(D^*)|\Lambda^{(2)}] = A(\phi^*)$ with $\phi_k^* = E(D^*|\bar{L}_k)$.
e. $\mathbf{m}^{-1}(D^*) = D^*/\bar{\pi}_K - \sum_{k=1}^{K} (1 - \pi_k)\bar{\pi}_k^{-1} E(D^*|\bar{L}_k)$.

Propositions 8.1 and 8.2 are proved in Appendix A. $\mathbf{m}^{-1}(D^*)$ and $\Pi[\mathbf{u}(D^*)|\Lambda^{(2)}]$ do not in general exist in closed form with arbitrary patterns of missingness. But according to Propositions 8.2d and 8.2e, they do exist in closed form when the missing pattern is monotone. The rightmost expression in part b is a useful representation of $E[D^*|\bar{L}_k]$ as an explicit function of the observables $(R, L_{(R)})$.

8.2 Specialization to the Conditional Mean Model (1)

The following proposition concerning the full data model (1) is needed for the proofs of Propositions 2.3, 3.1, 3.2, 4.1, and 4.2 and is proved in Appendix A.

Proposition 8.3. In the model "full" characterized by restriction (1) with likelihood $\mathcal{L}^F(\alpha, \theta; L)$ of (16a), (a) $\Pi(D^*|\Lambda_0^{F,\perp}) = E[D^*\varepsilon'|X^*]\{\text{var}(\varepsilon|X^*)\}^{-1}\varepsilon$, (b) $\Lambda_0^{F,\perp} = \{D^F(h)\}$, (c) $S_{\text{eff}}^F = D^F(h_{\text{eff}}^F)$ with h_{eff}^F as in equation (17), (d) $E[D^F(h)S_{\text{eff}}^F] = E[h(X^*)\{\partial g(X^*; \alpha_0)/\partial\alpha\}] = \kappa(h)$, and (e) $\Lambda_{0^*}^{F,\perp} = \{[\kappa(h)]^{-1}D^F(h)\}$.

Proposition 3.1 is an immediate corollary of Lemma 8.1 and Proposition 8.3.

Proof of Proposition 3.2. Consider first an arbitrary missing data pattern. Then, with $A(\phi)$ given by (43) and $D(\alpha, h, \phi)$ given by (38), Proposition 3.2a is Proposition 8.1c2 with $\Lambda_{0^*}^{F,\perp}$ as in Proposition 8.3e. Proposition 3.2b follows from (a) $S_{\text{eff}} \in \Lambda_0^{\perp}$ so $S_{\text{eff}} = D(h_{\text{eff}}, \phi_{\text{eff}})$ for some h_{eff} and ϕ_{eff} by Proposition 8.3b and (b) $\kappa(h_{\text{eff}}) = E[D^F(h_{\text{eff}})S_{\text{eff}}^{F\prime}] = E[D(h_{\text{eff}}, \phi_{\text{eff}})S_{\text{eff}}'] = \text{var}(S_{\text{eff}})$ by Proposition 8.3d and Proposition 8.1c1. If missingness is monotone, then, by Proposition 8.2c, Proposition 3.2 is true with $A(\phi)$ as defined in (39).

Proof of Propositions 2.3, 4.1, and 4.2. To prove Proposition 2.3, for arbitrary missing-data patterns, note that by Proposition 2.2, the asymptotic variance of $\hat{\alpha}(h, \phi)$ is $\text{var}\{[\kappa(h)]^{-1}D(h, \phi)\}$. Thus with h fixed, we need to minimize $\text{var}\{D(h, \phi)\} \equiv \text{var}[\mathbf{u}\{D^F(h)\} - A(\phi)]$ over $A(\phi) \in \Lambda^{(2)}$. By the definition of a projection, the minimizer will be $\Pi[\mathbf{u}\{D^F(h)\}|\Lambda^{(2)}]$, which by Proposition 8.1d equals $A(\phi^h)$ with $\phi_r^h \equiv E[B^{*h}|L_{(r)}]$ because, by its definition (45), $B^{*h} \equiv \mathbf{m}^{-1}\{D^F(h)\}$, proving Proposition 2.3. Proposition 4.1 then follows from the fact that in the saturated model "full b," the OLS estimator $\hat{\alpha}^F(\mathbf{i})$ is known to be RAL, and $\Lambda_{0^*}^{F,\perp}$ consists of a single element (Bickel et al. 1993; Newey 1990a). Turning now to Proposition 4.2, Propositions 8.1e and 8.3b imply that $h_{\text{eff}}(X^*)$ solves $S_{\text{eff}}^F = \Pi[\mathbf{m}^{-1}(D^*)|\Lambda^{F,\perp}]$, with $D^* \equiv h(X^*)\varepsilon$. Hence Propositions 8.3a and 8.3c and definition (45) of B^{*h} imply that $h_{\text{eff}}(X^*)$ solves $\partial g(X^*; \alpha_0)/\partial\alpha = E[B^{*h}\varepsilon'|X^*]$. But definitions (45) and (46) imply that $B^{*h} = \{\pi(\mathbf{1})\}^{-1}h(X^*)\varepsilon - \mathbf{r}[h(X^*)\varepsilon]$. Thus $h_{\text{eff}}(X^*)$ solves $\partial g(X^*; \alpha_0)/\partial\alpha = h(X^*)\{t(X^*)\}^{-1} - E[\mathbf{r}\{h(X^*)\varepsilon\}\varepsilon'|X^*]$, which gives (23). Equation (24) follows from Proposition 2.3. Uniqueness is proved in Appendix A.

To prove Propositions 2.3, 4.1, and 4.2 when the missing data pattern is monotone, note by Proposition 8.2d, ϕ^h is as in (41) with $A(\phi)$ now as in (39). Further, after substituting the right side of the identity in Proposition 8.2e for $\mathbf{m}^{-1}(D^*)$ in the identity $S_{\text{eff}}^F = \Pi[\mathbf{m}^{-1}(D^*)|\Lambda^{F,\perp}]$, we again obtain (23), but now with $\mathbf{r}(h(X^*)\varepsilon)$ as in (42).

Proof of Propositions 6.2b and 6.2c. Proposition 6.2b follows from Proposition 3.2 and the discussion in Remark 1. Proposition 6.2c follows from Proposition 8.1e1.

8.3 Additional Considerations

In any semiparametric model in which data are missing at random, the probability of observing full (complete) data is bounded away from 0, and the response probabilities are known or can be modelled, Proposition 8.1 can be used to construct a class of estimators that contains both an estimator whose asymptotic variance attains the semiparametric variance bound and an estimator asymptotically equivalent to any given RAL estimator. As an example, consider the Cox proportional hazards model,

$$\lambda_T(t|X^*) = \lambda_0(t)\exp(\alpha_0' X^*), \tag{50}$$

where, in the example of Section 2.1, T is time to myocardial infarction, $\lambda_T(t|X^*)$ is the hazard of T at time t given X^*, and V no longer includes the constant 1. We assume cen-

soring C, defined as the minimum of time to death without myocardial infarction and end of follow-up, is independent of T given X^*. Redefine $W = (T^* \equiv \min(T, C), Y \equiv I(T = T^*), V^{\dagger\prime}, V')'$. Now let model "full" be defined by data $L = (W', X')'$ and restriction (50). Let model "obs" now be characterized by (50), (3)–(5), and data (7); that is, (Δ, L_{obs}). Pugh, Robins, Lipsitz, and Harrington (1992) proved that, similar to Proposition 2.2, the influence function of $\hat{\alpha}(h, \phi)$ solving $\sum_i D_i(\alpha, h, \phi) = 0$ equals $\kappa(h)^{-1}\{\Delta B^F(h)/\pi - A(\phi)\}$ where $\kappa(h)^{-1}B^F(h)$ is the influence function of $\hat{\alpha}^F(h)$ solving $\sum_i D_i^F(\alpha, h) = 0$. Here $A(\phi)$ is as in (9) and $h \equiv h(t, X^*)$, $D_i^F(\alpha, h) = I(T_i < C_i)\{h(T_i^*, X_i^*) - \sum_j e^{\alpha' X_j^*} I(T_j^* \geq T_i^*) h(T_i^*, X_j^*)/\sum_j e^{\alpha' X_j^*} I(T_j^* \geq T_i^*)\}$, $D_i(\alpha, h, \phi) = \Delta_i \pi_i^{-1} I(T_i < C_i)\{h(T_i^*, X_i^*) - [\sum_j e^{\alpha' X_j^*} I(T_j^* \geq T_i^*) h(T_i^*, X_j^*) \Delta_j \pi_j^{-1}/\sum_j e^{\alpha' X_j^*} I(T_j^* \geq T_i^*) \Delta_j \pi_j^{-1}]\} - A_i(\phi)$, $\kappa(h) = E\{B^F(h) B^F(h_{\text{eff}}^F)\}'$, $h_{\text{eff}}^F(t, X^*) = X^*$, $B^F(h) = \int_0^\infty dM(u)\{h(u, X^*) - E[h(u, X^*) | T = u]\}$, $dM(u) = dN(u) - \lambda_0(u)\exp(\alpha_0' X^*) I(T^* \geq u)\, du$, and $N(u) = I[T^* \leq u, T^* = T]$. Note that $\hat{\alpha}^F(h_{\text{eff}}^F)$ is the usual Cox partial likelihood estimator. Ritov and Wellner (1988) proved that any RAL estimator of α_0 in model "full" is asymptotically equivalent to $\hat{\alpha}^F(h)$ for some h; thus, by Proposition 8.1c2 (i.e., by the restatement of Proposition 3.2i), any RAL estimator of α_0 in model "obs" is asymptotically equivalent to some estimator $\hat{\alpha}(h, \phi)$. Further, by a proof analogous to our proofs of Propositions 2.3, 4.1, and 4.2, it can be shown that the asymptotic variance of $\hat{\alpha}(h_{\text{eff}}, \phi_{\text{eff}})$ attains the semiparametric variance bound for model "obs," where $\phi_{\text{eff}} = E[B^F(h_{\text{eff}}) | W]$ and $h_{\text{eff}}(u, X^*)$ solves $h(u, X^*) = E[h(u, X^*) | T = u] + \nu(u, X^*)\{X^* + E[\pi^{-1}B^F(h) - (1 - \pi)\pi^{-1}E\{B^F(h) | W\} | X^*, T^* \geq u] + E[\{\pi^{-1}(u, 1, V^\dagger, V) - 1\}E\{B^F(h) | T = u, V^\dagger, V\} | X^*]\}$ with $\nu(u, X^*) = E[\pi^{-1}(u, 1, V^\dagger, V) | X^*]$. The case cohort design (Prentice, 1986) is a special case of model "obs" with $\pi(u, 1, V^\dagger, V) = 1$ and $\pi(u, 0, V^\dagger, V) = \rho$, with ρ the subcohort sampling fraction. The nested case-control design of Thomas (1977) is also a special case of model "obs" modified so as to allow for nonindependent sampling as in Section (6.3). Proposition 8.1a can be used to show that, even when data on the extraneous surrogates $V^\dagger$ are not obtained, the estimators of Prentice (1986) and Lin and Ying (1993) for the case-cohort design and the partial likelihood estimator of Thomas (1977) for the nested case-control design are generally inefficient. Similar results have been obtained by Robins, Hsieh, and Newey (1995) and Rotnitzky and Robins (1993) for parametric models for the conditional density of Y given X^* with missing covariates or outcome data, by Robins and Rotnitzky and Zhao (1995) for parametric models for the conditional mean of Y given X^* with missing data on Y, and by Robins and Rotnitzky (1992) and Robins (1993a,b) in the accelerated failure time model and in the Cox proportional hazards model in the presence of dependent censoring attributable to time-dependent covariates that simultaneously predict failure and censoring. Analogues of Propositions 8.1 and 8.2 hold when the data are coarsened at random in the sense of Heitjan and Rubin (1991) allowing extension of our results to models with missingness (e.g., censoring) in continuous time (Robins and Rotnitzky 1992; Robins 1993a,b).

APPENDIX A: SEMIPARAMETRIC EFFICIENCY

In Appendix A, we prove Lemmas 8.2–8.3, Propositions 8.1–8.3, that Equations (23), (27) and (47) have unique solutions, that Equations (23) and (47) can be solved by successive approximation, and that the Manski-McFadden estimator of Section 6 is semiparametric efficient in model "obs-complete" with $\pi(W)$ known.

Proof of Lemma 8.2

$\Lambda^{(3)} \subseteq \Lambda^{(2)}$, because (a) any score S_ψ is in $\Lambda^{(2)}$ by the conditional mean 0 property of scores, and (b) $\Lambda^{(2)}$ is closed because it is the inverse image of the closed set of $\{0\}$ under the continuous mapping $E\{\cdot | L\}$. To prove part b, we note that for any bounded $A^{(2)}$ in $\Lambda^{(2)}$, the submodel $\pi(R, L_{(R)})(1 + \psi' A^{(2)})$ defined on a sufficiently small open ball around $\psi_0 = 0$ is regular with score $A^{(2)}$ by lemma C.4 of Newey (1990b). But any function in $\Lambda^{(2)}$ can be approximated in mean square by bounded functions.

Proof of Lemma 8.3

Part (a) is an easy calculation. For part (b), note that $0 = E[a^{(2)}(R, L_{(R)}) | L] = \sum_r \pi(r) a^{(2)}(r, L_{(r)})$ implies $a^{(2)}(\mathbf{1}, L) = \{\pi(\mathbf{1})\}^{-1} \sum_{r \neq \mathbf{1}} \pi(r) a^{(2)}(r, L_{(r)})$.

The proof of Proposition 8.1 requires a series of lemmas concerning model "obs" of Section 8.

Lemma A.1. In model "obs," Λ is the closure of $\{\mathbf{g}(A^F) + A^{(3)}; A^{(3)} \in \Lambda^{(3)}, A^F \in \Lambda^F\}$ and $S_\alpha = \mathbf{g}(S_\alpha^F)$. Lemma A.1 states that the score for α in model "obs" is the conditional expectation of the score for α in model "full" given the observed data.

Proof. Lemma A.1 follows from proposition A5.5 of Bickel et al. (1993) concerning scores in missing-data models, Lemma C.4 of Newey (1990b) and the fact that, by the missing-at-random assumption (35), $\mathbf{g}(A^{(3)}) = A^{(3)}$.

Lemma A.2. If (36) is true and $B = b(R, L_{(R)})$, then (a) $B - \mathbf{u}\{E(B|L)\} = A(\phi)$, with $\phi_r(L_{(r)}) = -b(r, L_{(r)})$, and (b) $B - \mathbf{u}\{E(B|L)\} \in \Lambda^{(2)}$.

Proof. Part a is a straightforward calculation. Part b follows from part a either by Lemma 8.3 or directly from $E[\mathbf{u}\{E(B|L)\} | L] = E(B|L)$.

Lemma A.3. For all B^* and all $A^{(2)} \in \Lambda^{(2)}$, $E[\mathbf{g}(B^*) A^{(2)\prime}] = 0$.

Proof. $E[\mathbf{g}(B^*) A^{(2)\prime}] = E[B^* A^{(2)}] = E[B^* E\{A^{(2)} | L\}'] = E[B^* 0'] = 0$.

Lemma A.4. In model "obs," $\Lambda^{(1)}$ and $\Lambda^{(3)}$ are mutually orthogonal closed linear spaces, where $\Lambda^{(1)} \equiv \{\mathbf{g}(A^F); A^F \in \Lambda^F\}$. Further, $\Lambda = \Lambda^{(1)} \oplus \Lambda^{(3)}$, where $\oplus$ denotes the direct sum of two spaces.

Proof. $\Lambda^{(3)} \perp \Lambda^{(1)}$ by Lemmas A.3 and 8.2. $\Lambda^{(3)}$ is closed by definition and linear by assumption. Linearity of $\Lambda^{(1)}$ follows from the assumed linearity of Λ^F. To prove that $\Lambda^{(1)}$ is closed, we note that it is the image of the closed set Λ^F under the linear operator $\mathbf{g}(\cdot)$. Hence it suffices to show that $\mathbf{g}(\cdot)$ has a continuous inverse. But $\mathbf{g}(\cdot)$ has a continuous inverse by part a of Proposition A1.7 of Bickel et al. (1993), because $E[\|\mathbf{g}(A^F)\|^2] \geq E[I(R = \mathbf{1}) A^{F\prime} A^F] = E[\pi(\mathbf{1}) A^{F\prime} A^F] \geq \sigma E[\|A^F\|^2]$, where the final inequality is by (36). Hence $\Lambda = \Lambda^{(1)} \oplus \Lambda^{(3)}$ by Lemma A.1.

Lemma A.5. In model "obs," if $B^* \in \Lambda_0^{F,\perp}$ and $A^{(2)}$ is in $\Lambda^{(2)}$, then $\mathbf{v}(B^*, A^{(2)}) \in \Lambda_0^\perp$.

Proof. Because $\Lambda = \Lambda^{(1)} \oplus \Lambda^{(3)}$ and $E[\mathbf{v}(B^*, A^{(2)})] = 0$, it is sufficient to show that $\mathbf{v}(B^*, A^{(2)})$ is contained in $\Lambda^{(1),\perp}$ and $\Lambda^{(3),\perp}$. By its definition, $\mathbf{v}(B^*, A^{(2)}) \in \Lambda^{(3),\perp}$. That $\mathbf{v}(B^*, A^{(2)})$

$\in \Lambda^{(1),\perp}$ follows from $E[\mathbf{v}(B^*, A^{(2)})\mathbf{g}(A^F)'] = E[\mathbf{u}(B^*)\mathbf{g}(A^F)'] = E[B^*A^{F\prime}]$, where the first equality is by Lemma A.3 and the second equality is by $E[\mathbf{u}(B^*)\mathbf{g}(A^F)'] = E[\{\pi(\mathbf{1})\}^{-1}I\ (R = \mathbf{1})B^*A^{F\prime}]$.

Lemma A.6. In model "obs," $B \in \Lambda_0^{(1),\perp}$ if and only if $E(B|L) \in \Lambda_0^{F,\perp}$.

Proof. The lemma follows from $E[B\mathbf{g}(A^F)'] = E[\mathbf{u}\{E(B|L)\}\mathbf{g}(A^F)'] = E[E(B|L)A^{F\prime}]$, where the first equality is by Lemmas A.2 and A.3.

Proof of Proposition 8.1

Because $B \in \Lambda_0^{\perp} \Rightarrow B \in \Lambda_0^{(1),\perp}$, part a1 follows from Lemmas A.6 and A.2. Part a2 follows from $B \in \Lambda_{0*}^{\perp}$ implies $I_{q\times q} = E[BS_\alpha'] = E[\mathbf{u}\{E(B|L)\}\mathbf{g}(S_\alpha^F)'] = E[E(B|L)S_\alpha^{F\prime}]$, where the second equality is by $S_\alpha = \mathbf{g}(S_\alpha^F)$, and Lemmas A.2–A.3. Part b follows immediately from part a1 and Lemma A.5. Turn next to part c1. Because $B^* \in \Lambda_0^{F,\perp}$ implies that $\mathbf{v}(B^*, A^{(2)}) \in \Lambda_0^{\perp}$ by Lemma A.5, in view of Lemma 8.1b it suffices to note $E[\mathbf{v}(B^*, A^{(2)})S_\alpha'] = E[B^*S_\alpha^{F\prime}]$, because $E[\mathbf{v}(B^*, A^{(2)})S_\alpha'] = E[\mathbf{u}(B^*)\mathbf{g}(S_\alpha^F)]$ by $S_\alpha = \mathbf{g}(S_\alpha^F)$ and Lemma A.3. Part c2 is a consequence of parts b and c1 and Lemma 8.1b. Turn now to part d. Van der Laan (1993, pp. 29–30) proved that $\mathbf{m}^{-1}$ exists as a bounded linear operator. A modified version of his proof follows. [The proof in Robins and Rotnitzky (1992) unnecessarily and incorrectly argued that the operator $\pi(\mathbf{1})^{-1}\mathbf{m}(B^*) - B^*$ was always compact.] Let $\mathbf{i}$ be the identity operator. An operator $\mathbf{a}$ is a contraction if $\sup_{\|B^*\|=1} E[\|\mathbf{a}(B^*)\|^2] < c < 1$. Now $\mathbf{m}^{-1}$ is a bounded linear operator if $\mathbf{i} - \mathbf{m}$ is a contraction, since $\mathbf{m} = \mathbf{i} - (\mathbf{i} - \mathbf{m})$ and the identity minus a contraction has a bounded inverse (Kress, 1989, p. 16). But $\mathbf{m}$ and thus $\mathbf{i} - \mathbf{m}$ are self-adjoint, because $E[\mathbf{m}(B^*)'D^*] = E[B^{*\prime}\mathbf{m}(D^*)]$. By $\mathbf{i} - \mathbf{m}$ self-adjoint, $\sup_{\|B^*\|=1} E[\|[\mathbf{i} - \mathbf{m}](B^*)\|^2] = \sup_{\|B^*\|=1} E[B^{*\prime}[\mathbf{i} - \mathbf{m}](B^*)]$ (Kress 1989, thm. 15.9). But $\sup_{\|B^*\|=1} E[B^{*\prime}[\mathbf{i} - \mathbf{m}](B^*)] < 1 - \sigma$, since $E[B^{*\prime}\mathbf{m}(B^*)] = E[\mathbf{g}(B^*)'\mathbf{g}(B^*)] \geq E[\pi(\mathbf{1})\|B^*\|^2] \geq \sigma E[\|B^*\|^2]$, where the last inequality uses (36). It remains to show $\Pi[\mathbf{u}(D^*)|\Lambda^{(2),\perp}] = \mathbf{g}(B^*)$ with $B^* \equiv \mathbf{m}^{-1}(D^*)$. Now $\mathbf{g}(B^*) \in \Lambda^{(2),\perp}$ by Lemma A.3. Further, because $D^* = \mathbf{m}(B^*) = E[\mathbf{g}(B^*)|L]$, $\mathbf{u}(D^*) - \mathbf{g}(B^*) = A(\phi^*) \in \Lambda^{(2)}$ by Lemma A.2, completing the proof. To prove part e1, note that $S_{\text{eff}} \equiv S_\alpha - \Pi(S_\alpha|\Lambda) = S_\alpha - \Pi(S_\alpha|\Lambda^{(1)})$, because $S_\alpha = \mathbf{g}(S_\alpha^F) \perp \Lambda^{(3)}$ by $\Lambda^{(3)} \subseteq \Lambda^{(2)}$ and Lemma A.3. So with $A_\alpha^F \in \Lambda^F$ solving $\Pi(S_\alpha|\Lambda^{(1)}) = \mathbf{g}(A_\alpha^F)$ and $B_{\text{eff}}^* \equiv S_\alpha^F - A_\alpha^F$, $S_{\text{eff}} = \mathbf{g}(B_{\text{eff}}^*) \in \Lambda_0^{\perp}$ and $\Pi[B_{\text{eff}}^*|\Lambda^{F,\perp}] = \Pi[S_\alpha^F|\Lambda^{F,\perp}] \equiv S_{\text{eff}}^F$. Hence by part a1, $S_{\text{eff}} = \mathbf{u}\{\mathbf{m}(B_{\text{eff}}^*)\} + [\mathbf{g}(B_{\text{eff}}^*) - \mathbf{u}\{\mathbf{m}(B_{\text{eff}}^*)\}]$, with $D_{\text{eff}} \equiv \mathbf{m}(B_{\text{eff}}^*) \in \Lambda_0^{F,\perp}$ and $\mathbf{g}(B_{\text{eff}}^*) - \mathbf{u}\{\mathbf{m}(B_{\text{eff}}^*)\} = -\Pi[\mathbf{u}(D_{\text{eff}})|\Lambda^{(2)}]$ by part d, which proves part e1 except for uniqueness. To prove part e1, first note $B_{\text{eff}}^* \equiv S_\alpha^F - A_\alpha^F$ solves the equation in part e2 because $\mathbf{g}(B_{\text{eff}}^*) \in \Lambda_0^{\perp}$ implies $\mathbf{m}(B_{\text{eff}}^*) \in \Lambda^{F,\perp}$ by Lemma A.6. To prove uniqueness, let B_0^* be the difference between two solutions B^* to the equation in part (e.2), so $0 = \mathbf{m}(B_0^*) + \Pi[B_0^* - \mathbf{m}(B_0^*)|\Lambda^{F,\perp}]$. Thus $\mathbf{m}(B_0^*) \in \Lambda^{F,\perp}$. Hence $0 = \Pi[B_0^*|\Lambda^{F,\perp}]$, so $B_0^* \in \Lambda^F$. Thus $0 = E[\mathbf{m}(B_0^*)B_0^{*\prime}] \geq \sigma E[B_0^*B_0^{*\prime}]$ by (36). Hence $B_0^* = 0$ with probability 1. This implies uniqueness in part e1 as well. That the equation in (e.2) can be solved by successive approximation is proved below.

Proof of Proposition 8.2

Parts a1 and a2 are easy calculations. To prove part b, note that (35) and (36) implies that $f[L|\bar{L}_k, R = \mathbf{1}_{k-1}] = f[L|\bar{L}_k]$, which implies $E[D^*|\bar{L}_k] = E[D^*|\bar{L}_k, R = \mathbf{1}_{k-1}] = E[D^*|\bar{L}_k, R_{k-1} = 1]$. Also, $E[D^*|\bar{L}_k] = E[R_KD^*/\bar{\pi}_K|\bar{L}_k] = E[R_KD^*/\bar{\pi}_K|\bar{L}_k,\ R_{k-1} = 1]\Pr[R_{k-1} = 1|\bar{L}_k]$. But $\Pr[R_{k-1} = 1|\bar{L}_k] = \bar{\pi}_{k-1}$.

To prove part c, one can calculate that $A(\phi)$, defined as in (39), equals $A(\phi^\dagger)$ as defined in (43) if

$$\phi_r^\dagger = \sum_{k=0}^{j}(1 - \pi_k)\bar{\pi}_k^{-1}\phi_k(\bar{L}_k) - \bar{\pi}_j^{-1}\phi_{j+1}(\bar{L}_{j+1}) \quad \text{with} \quad r = \mathbf{1}_j,\ 0 \leq j < K. \tag{A.1}$$

(A.1) can be reexpressed as the following recursive formula for the $\phi_k(\bar{L}_k)$ in terms of $\phi^\dagger$:

$$\phi_{j+1}(\bar{L}_{j+1}) = \{-\phi_r^\dagger + \sum_{k=1}^{j}(1 - \pi_k)\bar{\pi}_k^{-1}\phi_k(\bar{L}_k)\}\bar{\pi}_j, \quad \text{with} \quad r = \mathbf{1}_j,\ j = (0, \ldots, K), \quad \text{where, by convention,} \sum_{k=1}^{0} \equiv 0. \tag{A.2}$$

Part c now follows from (A.1), (A.2), and Lemma 8.3. To prove part d, note that by part c, $A(\phi^*) \in \Lambda^{(2)}$. Also, one can write $\mathbf{u}(D^*) = D^* + \sum_{k=1}^{K}(R_k - \pi_kR_{k-1})\bar{\pi}_k^{-1}D^*$, so that $\mathbf{u}(D^*) - A(\phi^*) = D^* + \sum_{k=1}^{K}(R_k - \pi_kR_{k-1})\bar{\pi}_k^{-1}\{D^* - E[D^*|\bar{L}_k, R_{k-1} = 1]\}$. Using this representation, one can use iterated conditional expectations to show that $E[\{\mathbf{u}(D^*) - A(\phi^*)\}A(\phi)'] = 0$ for all $A(\phi) \in \Lambda^{(2)}$, proving part d. Part e follows at once from part d, Proposition 8.1d and the identity $E[\bar{\pi}_K^{-1}I(R_K = 1)\mathbf{g}\{\mathbf{m}^{-1}(D^*)\}|L] = \mathbf{m}^{-1}(D^*)$.

Proof of Proposition 8.3

To prove part a, following arguments essentially identical to those in lemma A.5 of Newey and Powell (1990), we have $\Lambda^F = \{A_1 + A_2 + A_3: A_1 = a_1(X^*), A_2 = a_2(X^*, \varepsilon)$, and $A_3 = a_3(L)$, with $E(A_1) = 0$, $E(A_2|X^*) = 0$, $E(\varepsilon A_2'|X^*) = 0$, and $E(A_3|\varepsilon, X^*) = 0\}$. The restriction $E(\varepsilon A_2'|X^*) = 0$ is a consequence of $E[\varepsilon|X^*] = 0$. Further, it is easy to show that $\Lambda_1^F \equiv \{A_1\}$, $\Lambda_2^F \equiv \{A_2\}$, and $\Lambda_3^F \equiv \{A_3\}$ are mutually orthogonal. Consider any D^* with $E(D^*) = 0$. Then $\Pi(D^*|\Lambda^F) = \Pi(D^*|\Lambda_1^F) + \Pi(D^*|\Lambda_2^F) + \Pi(D^*|\Lambda_3^F)$. Further, it is easy to verify that $\Pi(D^*|\Lambda_3^F) = D^* - E(D^*|X^*, \varepsilon)$, $\Pi(D^*|\Lambda_2^F) = E(D^*|X^*, \varepsilon) - E(D^*|X^*) - E\ (D^*\varepsilon'|X^*)E(\varepsilon\varepsilon'|X^*)^{-1}\varepsilon$, and $\Pi(D^*|\Lambda_1^F) = E(D^*|X^*)$. So finally, $\Pi(D^*|\Lambda^F) = D^* - E(D^*\varepsilon'|X^*)E(\varepsilon\varepsilon'|X^*)^{-1}\varepsilon$, proving part a. Part b follows directly from a; c follows from a and the identity $E(S_\beta^F\varepsilon'|X^*) = \partial g(X^*; \alpha_0)/\partial\alpha$ or from Chamberlain (1987). Part d is a straightforward calculation, and e follows from b.

Uniqueness of the Solutions to Equations (23), (27), and (47)

Since Equation (47) is a special case of the equation in Proposition 8.1e2, it has a unique solution. Further, because $h_{\text{eff}}(X^*)$, solving (23) implies $\mathbf{m}^{-1}(h_{\text{eff}}(X^*)\varepsilon)$ solves (47), $h_{\text{eff}}(X^*)$ is unique by $\mathbf{m}^{-1}(\cdot)$ injective. To prove that (27) has a unique solution, let ϕ_j, $j = 1, 2$ be two solutions to (27). Let $h_j(X^*)$ be defined by (26), with ϕ_j substituted for ϕ_{eff}. Then $E[h_j(\mathbf{X}^*)\varepsilon|W] = \phi_j$ by ϕ_j satisfying (27), and thus $h_j(X^*)$ satisfies (23). By uniqueness of the solution to (23), $h_1(X^*) = h_2(X^*)$. Hence by $E[h_j(X^*)\varepsilon|W] = \phi_j$, we conclude that $\phi_1 = \phi_2$.

Proof That the Equation in Proposition (8.1e2), Equations (23) and (47) Can Be Solved by Successive Approximation

Let $\mathbf{a}$ be a bounded linear operator from a Hilbert space to itself. The solution B to $B = D - \mathbf{a}(B)$ can be obtained by successive approximation if $\mathbf{a}$ is a contraction (Kress 1989, p. 17). The operator $\Pi[\mathbf{m} - \mathbf{i}|\Lambda^F]$ in Proposition 8.1e2 is a contraction, since $\mathbf{m} - \mathbf{i}$ is a contraction (Van der Laan 1993) and $\Pi[\cdot|\Lambda^F]$ is a projection operator. Hence Equation (47), as a special case of Proposition 8.1e2, can also be solved by successive approximation. Finally, the operator acting on $h(X^*)$ on the right side of (23) can be shown to be a contraction by applying the Cauchy-Schwartz inequality twice.

Proof That $\hat{\alpha}_{MM}(q_{MM})$ is Semiparametric Efficient in Model "obs-complete" with $\pi(W)$ Known

In this model the likelihood is $\mathcal{L}(\alpha, \theta; Y, X, V) = f[Y, X, V \mid \Delta = 1; \alpha, \theta]$. Redefine $S_\alpha^F \equiv \partial \log f(Y \mid X^*; \alpha_0)/\partial\alpha = h_{\text{eff}}^F(X^*)\varepsilon$ and $\Lambda^F = \{A = a(X, V); E(A \mid \Delta = 1) = 0\}$. Let S_α, S_{eff}, and Λ be the score for α, efficient score, and nuisance tangent space in model "obs-complete" with $\pi(W)$ known. Then it is straightforward to show $S_\alpha = S_\alpha^F - E[S_\alpha^F \mid \Delta = 1]$ and $\Lambda = \{A - E(A \mid \Delta = 1); A \in \Lambda^F\}$. Hence it is sufficient to show $D_{MM}(q_{MM})$ is S_{eff}. To show $D_{MM}(q_{MM}) \in \Lambda_0^\perp$, note that $D_{MM}(q_{MM}) = S_\alpha^F - E[S_\alpha^F \mid X, V, \Delta = 1]$, so $E[D_{MM}(q_{MM})\{A - E(A \mid \Delta = 1)\}' \mid \Delta = 1] = 0$ for all $A \in \Lambda^F$. Finally, $S_\alpha - D_{MM}(q_{MM}) = E[S_\alpha^F \mid X, V, \Delta = 1] - E[S_\alpha^F \mid \Delta = 1] \in \Lambda$. Hence $D_{MM}(q_{MM}) = S_{\text{eff}}$.

APPENDIX B: ASYMPTOTIC NORMALITY

In this Appendix, we prove Lemma 6.2 and Propositions 2.2, 2.4, 6.1, and 6.2a.

We first prove Proposition 6.1 and then Proposition 2.4. Proposition 2.2 is a special case of 6.1, and the proof of Proposition 6.2a is identical to that of Proposition 6.1. We consider the case of arbitrary missing-data patterns so that $S_\psi(\psi)$ is based on a model $\pi(r, \psi)$ satisfying (44), $D(\alpha, h, \phi)$ is given by (38), and $A(\phi)$ is given by (43). Further specialization to the case of monotone missing data is straightforward.

Let $H(\gamma)' = (D^F(h)', S_\psi(\psi)')$, $\gamma' = (\alpha', \psi')$, and $\gamma = \alpha \times \psi$, where α and ψ are the parameter spaces of α and ψ. We prove our propositions under the following nine regularity conditions:

1. γ lies in the interior of a compact set $\boldsymbol{\gamma}$.
2. except in the proof of Lemma 6.2 (L_i, R_i), $i = (1, \ldots, n)$ are independently and identically distributed.
3. $\pi(\mathbf{1}, \psi) > c > 0$ for all $\psi \in \boldsymbol{\psi}$ for some c.
4. $E[H(\gamma)] \neq 0$ if $\gamma \neq \gamma_0$.
5. $\operatorname{var}[H(\gamma_0)]$ is finite and positive definite.
6. $E[\partial H(\gamma_0)/\partial\gamma']$ exists and is invertible.
7. a neighborhood N of γ_0 such that $E[\sup_{\gamma \in N}\|H(\gamma)\|]$, $E[\sup_{\gamma \in N}\|\partial H(\gamma)/\partial\gamma'\|]$, 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}.
8. $f(L, R; \gamma)$ is a regular parametric model with score $S_\gamma(\gamma) = \partial \log f(L, R; \gamma)/\partial\gamma$, where $f(L, R; \gamma)$ is a density that differs from the true density $f(L, R) = f(L, R; \gamma_0)$ only in that γ replaces γ_0.
9. For all γ^* in a neighborhood N of γ_0, $E_{\gamma^*}[H(\gamma^*)]$ and $E_{\gamma^*}[\sup_{\gamma \in N}\|H(\gamma)'H(\gamma)\|]$ is bounded where E_{γ^*} refers to expectation with respect to the density $f(L, R; \gamma^*)$.

Proof of Proposition 6.1

First, note that $E[D(\alpha_0, h, \phi, \psi_0)] = 0$ by (1), (35), (36), and (44). Let $H^*(\gamma)' = (D(\alpha, h, \phi, \psi)', S_\psi(\psi)')$. Then, by $E[D(\alpha_0, h, \phi, \psi_0)] = 0$ and $\pi(\mathbf{1}, \psi)$ bounded below by c, regularity conditions 1–9 hold, with $H^*(\gamma)$ replacing $H(\gamma)$. Theorems 2.6 and 3.4 of Newey and McFadden (1993) or corollary 1 of Manski (1988, chap. 8) imply that if $H^*(\gamma)$ satisfies 1–7 then, with probability approaching 1, there exists a unique solution $\hat{\gamma}$ to $\sum_i H_i^*(\gamma) = 0$ such that $\hat{\gamma}$ is asymptotically linear with influence function $-E[\partial H^*(\gamma_0)/\partial\gamma']^{-1}H^*(\gamma_0)$. By definition of $H^*(\gamma)$, $\hat{\gamma}' = (\hat{\alpha}(h, \phi, \hat{\psi})', \hat{\psi}')$. Hence $\hat{\alpha}(h, \phi, \hat{\psi})$ and $\hat{\psi}$ are asymptotically linear with influence functions $-E[\partial D(\alpha_0, h, \phi)/\partial\alpha']^{-1}\{D(h, \phi) - E[\partial D(\alpha_0, h, \phi, \psi_0)/\partial\psi']E[\partial S_\psi(\psi_0)/\partial\psi']^{-1}S_\psi\}$ and $-E[\partial S_\psi(\psi_0)/\partial\psi']S_\psi$, with $S_\psi = S_\psi(\psi_0)$. Now, under regularity conditions 6, 8, and 9, Lemma c.3 of Newey (1990b) implies that $-E[H^*(\gamma_0)/\partial\gamma'] = E[H^*(\gamma_0)S_\gamma'(\gamma_0)]$, which by theorem 2.2 of Newey (1990a) implies that $\hat{\gamma}$ is regular and that $-E[\partial D(\alpha_0, h, \phi, \psi_0)/\partial\psi'] = E[D(h, \phi)S_\psi']$, $-E[\partial D(\alpha_0, h, \phi, \psi_0)/\partial\alpha'] = E[D(h, \phi)S_\alpha']$, and $-E[\partial S_\psi(\psi_0)/\partial\psi'] = E[S_\psi S_\psi']$. Substituting these identities into the expression for the influence function of $\hat{\alpha}(h, \phi, \hat{\psi})$ and noting that by Proposition 8.1c1, Proposition 8.3d, and Lemma 8.1b, $\kappa(h) = E[D(h, \phi)S_\alpha']$ completes the proof of Propositions 6.1a and 6.1b. The second inequality in Proposition 6.1c follows from Proposition 2.3; the first inequality is a special case of Proposition 6.1d. To prove Proposition 6.1d, we note that $S_{\psi^{(j)}}$ is the first $p^{(j)}$ components of the $p^{(j+1)}$ vector $S_{\psi^{(j+1)}}$, where $S_{\psi^{(j)}}$ is the score for model j and $p^{(j)}$ is the dimension of $\psi^{(j)}$. By standard least squares theory, the variance of the residual from a population regression does not increase as the number of regressors increases, proving Proposition 6.1d. Finally, Proposition 6.1e follows under our regularity conditions, from Theorem 4.5 in Newey and McFadden (1993).

Proof of Proposition 2.4

Redefine $\gamma' \equiv (\alpha_1', \alpha_2', \lambda')$, where α_1 and α_2 are of the dimension of α_0, $H^*(\gamma)' \equiv (D\{\alpha_1, h, l(W; \lambda)\}', D(\alpha_2, h, \phi)', G(\alpha_2, \lambda)')$ with $\phi = \phi(W)$ a given function, $G(\alpha, \lambda) \equiv \Delta[D^F(\alpha, h) - l(W; \lambda)]\{\partial l(W; \lambda)/\partial\lambda\}$. We prove Proposition 2.4 under the assumption that regularity conditions 1–9 hold, with $H^*(\gamma)$ replacing $H(\gamma)$. Define $\gamma_0' = (\alpha_0', \alpha_0', \lambda^{\dagger\prime})$, where $\lambda^\dagger$ satisfies $E[G(\alpha_0, \lambda^\dagger)] = 0$, and note that $E[H^*(\gamma_0)] = 0$ even if (15) is misspecified. Thus by theorem 3.4 of Newey and McFadden (1993) or corollary 1 of Manski (1988, chap. 8), with probability approaching 1 the solution $\hat{\gamma}$ to $\sum_i H_i^*(\gamma) = 0$ is asymptotically linear with influence function $\{-E[\partial H^*(\gamma_0)/\partial\gamma]\}^{-1}H^*(\gamma_0)$. But $\hat{\gamma}'$ is precisely $(\hat{\alpha}(h, \hat{\phi}^h)', \hat{\alpha}(h, \phi)', \hat{\lambda}')$, defined in Section 2.7. Hence $\hat{\alpha}(h, \hat{\phi}^h)$ is asymptotically linear with influence function $-\{E[\partial D^F(\alpha_0, h)/\partial\alpha']\}^{-1}D(\alpha_0, h, \phi^\dagger)$, with $\phi^\dagger(w) = l(w; \lambda^\dagger)$, because $0 = E[\partial D(\alpha_1, h, l(W; \lambda))/\partial\lambda']|_{\gamma=\gamma_0} = E[\partial D(\alpha_1, h, l(W; \lambda))/\partial\alpha_2']|_{\gamma=\gamma_0}$ and $E[\partial D(\alpha_1, h, l(W; \lambda))/\partial\alpha_1']|_{\gamma=\gamma_0} = E[\partial D^F(\alpha_0, h)/\partial\alpha']$. The consistency of the variance estimator follows from theorem 4.5 of Newey and McFadden (1993).

Proof of Lemma 6.2

It is straightforward to show that the probability distribution generated by the "fixed sampling plan" is equal to the conditional distribution generated by an independent sample plan with $\pi(W) = \pi(W^*)$ conditional on $U = 0$, where $U = (U_1, \ldots, U_S)'$, $U_s = n^{1/2}(N_s^{\text{val}}/N_s - \rho_s(N_s))\}$, and $N_s^{\text{val}} = \sum_i \Delta_i I(W_i^* = w_s^*)$. Therefore, we can restrict attention to conditional properties of the independent sampling design given $U = 0$. Because $U_s = n^{1/2}\{N_s^{\text{val}}/N_s - \pi(W_s^*)\} - n^{1/2}\{\rho_s(N_s) - \pi(W_s^*)\}$, $\rho_s(N_s) - \pi(W_s^*) = O_p(N_s^{-1})$ by assumption, and $\operatorname{logit}\{N_s^{\text{val}}/N_s\}$ is the MLE of ψ_s^* under model (33), conditioning on $U = 0$ is asymptotically equivalent to conditioning on $n^{-1/2}\sum_i S_{\psi^*,i} = 0$, where S_{ψ^*} is the score for model (33). Thus it suffices to show that the asymptotic conditional distribution of $n^{1/2}\{\hat{\alpha}(h, \phi, \hat{\psi}) - \alpha_0\}$, given $n^{-1/2}\sum_i S_{\psi^*,i}$, equals its unconditional asymptotic distribution. By Proposition 6.1a and the conditional properties of the multivariate normal distribution, the limiting conditional distribution of $n^{1/2}\{\hat{\alpha}(h, \phi, \hat{\psi}) - \alpha_0\}$ is equivalent to that of $\{\kappa(h)\}^{-1}n^{-1/2}\sum_i \text{Resid}_i\{\text{Resid}[D(h, \phi), S_\psi], S_\psi^*\}$. But because, by assumption, model (33) with score S_ψ^* is nested within model $\pi(\psi)$ with score S_ψ, we have $S_\psi^* = bS_\psi$ for some matrix b, so that $\text{Resid}\{\text{Resid}[D(h, \phi), S_\psi], S_\psi^*\} = \text{Resid}[D(h, \phi), S_\psi]$. The lemma then follows by Theorem 6.1a and 6.1e.

APPENDIX C: ASYMPTOTIC EQUIVALENCE WITH PREVIOUS ESTIMATORS

In this Appendix, we prove Lemma 6.1, that $\hat{\alpha}_{\text{MLE}}$ of Section 5.1 is a semiparametric estimator in model "obs" and derive the influence functions of the previously proposed estimators of Sections 4 and 6.

Asymptotic Equivalence of $\hat{\alpha}_{PFCW}$ and $\hat{\alpha}(h_{PFCW}, \phi_{PFCW})$

In their appendixes, Pepe and Fleming (1991) and Carroll and Wand (1991) showed that $\hat{\alpha}_{PFCW}$ has influence function $\kappa^{-1}B_{PFCW}$, with $B_{PFCW} = \Delta\{X^*\varepsilon - (1-\rho)\rho^{-1}E[\phi_{PFCW}|X^*]\} + (1-\Delta)\phi_{PFCW}$. $\kappa = E[B_{PFCW}S'_\alpha]$ by Lemma 8.1. Applying Corollary 5.1, we obtain $\kappa^{-1}B_{PFCW} = \kappa^{-1}D(h, \phi)$, with $h(X^*) = h_{PFCW}(X^*)$ and $\phi(W) = \phi_{PFCW}(W)$. But, by Proposition 2.2, $\kappa^{-1}D(h, \phi)$ is the influence function of $\hat{\alpha}(h_{PFCW}, \phi_{PFCW})$.

Asymptotic Equivalence of $\hat{\alpha}_{MM}(q)$ and $\hat{\alpha}(h, 0)$

Clearly, $\hat{\alpha}_{MM}(q)$ has influence function $\kappa^{-1}\Delta D_{MM}(q, \alpha_0)$, where $\kappa = E[\Delta D_{MM}(q, \alpha_0)S'_\alpha]$; so, by Proposition 2.2, it suffices to show that $D(h, 0) = \Delta D_{MM}(q, \alpha_0)$. Now, given $h(X^*)$, $D(h, 0) = \Delta D_{MM}(q, \alpha_0)$, with $q(Y, X^*, \alpha) \equiv \pi^{-1}h(X^*)\varepsilon$ because $E[\pi Q(\alpha_0)|X^*] = 0$. Conversely, given $q(Y, X^*, \alpha)$, we apply Corollary 5.1 to obtain $h(X^*)$ as given by (31) and $\phi(W) \equiv 0$. Proof of the asymptotic equivalence of $\hat{\tilde{\alpha}}_{MM}(q)$ and $\hat{\tilde{\alpha}}(h, 0)$ follows from Corollary 5.1 after using a Taylor expansion to show that $\hat{\tilde{\alpha}}_{MM}(q)$ has influence function $\kappa^{-1}\{\Delta D_{MM}(q, \alpha_0)\} - (\Delta - \pi)\{E[D_{MM}(q, \alpha_0)|W]\}$.

Proof of the Asymptotic Equivalence of $\hat{\alpha}_{GM}$ and $\hat{\alpha}(\mathbf{i}, \phi_{GM})$

A Taylor expansion around γ_0 and ω gives $n^{-1/2}D_{GM,i}(\alpha_0, \hat{\gamma}) = n^{-1/2}\sum_i D_{GM,i}(\alpha_0, \gamma_0) - (1-\rho)\omega(\gamma_0, 1)'\alpha_{0,1}E(VV')n^{1/2}(\hat{\gamma} - \gamma_0) + o_p(1) = n^{-1/2}\sum_i H_{GM,i} + o_p(1)$, with $H_{GM} = D_{GM}(\alpha_0, \gamma_0) - \Delta\omega(1-\rho)\rho^{-1}(\gamma_0, 1)'\alpha_{0,1}V(X - \gamma_0 V)$, where we have used $n^{1/2}(\hat{\gamma} - \gamma_0) = n^{-1/2}\rho^{-1}\sum_i \Delta_i\{E(VV')\}^{-1}V_i(X_i - \gamma_0 V_i) + o_p(1)$. Hence, by a Taylor expansion around α_0 and Lemma 8.1, $\hat{\alpha}_{GM}$ will have influence function $\kappa^{-1}H_{GM}$, $\kappa = E[H_{GM}S'_\alpha]$. Because Corollary 5.1 is also true for model "obs b," we have $\kappa^{-1}H_{GM} = \kappa^{-1}D(h, \phi)$, where $h(X^*) = E(H_{GM}|Y=1, X^*) - E(H_{GM}|Y=0, X^*) = bX^*$ with $b = (b_1, b_2)$, $b_1 = (\rho, 0)'$, $b_2 = [(1-\rho)\gamma_0, (1-\rho)\omega + \rho]'$, and $\phi = (\gamma_0, 1)'V[Y - \alpha_{0,1}\gamma_0 V - \alpha_{0,2}V]$. Hence, by Proposition 2.2 and Lemma 8.1, $\hat{\alpha}(h, \phi)$ and $\hat{\tilde{\alpha}}_{GM}$ have the same influence functions. But because b is constant, $\hat{\alpha}(h, \phi) = \hat{\alpha}(\mathbf{i}, b^{-1}\phi)$ and $b^{-1}\phi = \phi_{GM}$.

Proof that $\hat{\alpha}_{MLE}$ of Section 5.1 is a Semiparametric Estimator in Model "obs" When (11) is True

Let $\eta^\dagger = (\gamma^\dagger, \Omega^\dagger, \sigma^{2\dagger}, \mu^\dagger, \Sigma^\dagger)$, where $\gamma^\dagger \equiv E[XV']\{E[VV']\}^{-1}$, $\Omega^\dagger = E[(X - \gamma^\dagger V)^{\otimes 2}]$, $\sigma^{2\dagger} = E[(Y - \alpha_{0,0} + \alpha_{0,1}X)^2]$, $\mu^\dagger = E(V)$, and $\Sigma^\dagger = E[(V - \mu^\dagger)(V - \mu^\dagger)']$, where expectations are with respect to the true distribution of the data. Straightforward but tedious calculation gives $E[\partial \log \mathcal{L}(\alpha_0, \eta^\dagger; \Delta, L_{obs})/\partial(\alpha, \eta)] = 0$ even if (12) is false, where $\mathcal{L}(\alpha, \eta; \Delta, L_{obs})$ is the likelihood assuming that (12) were also true. Hence, under our regularity conditions, $(\hat{\alpha}_{MLE}, \hat{\eta}_{MLE})$ will be RAL estimators of $(\alpha_0, \eta^\dagger)$ in model "obs" when (11) is true even if (12) is false.

Proof of Lemma 6.1

To prove Lemma 6.1, recall that $\alpha \equiv (a, b')'$ and define the odds ratio function $\text{or}(X^*_{-1}) = f(X^*_{-1}|Y=1)/f(X^*_{-1}=0|Y=1)/\{f(X^*_{-1}|Y=0)/f(X^*_{-1}=0|Y=0)\}$, where we assume that 0 is in the support of X^*_{-1}. Prentice and Pyke (1979) pointed out that, with the intercept a unrestricted, model "full" based on the logistic model (2) is equivalent to a semiparametric model indexed by b and an infinite-dimensional parameter ρ with the likelihood contribution for a single subject, $\mathcal{L}^F(b, \rho; L) = (\rho_1)^Y(1-\rho_1)^{1-Y}f(X^*_{-1}|Y; b, \rho_2)f[V^\dagger|X^*_{-1}, Y; \rho_3]$, characterized by the sole restriction that $f(X^*_{-1}|Y; b, \rho_2)$ satisfies $\log\{\text{or}(X^*_{-1})\} = b'X^*_{-1}$. Thus model "obs" is equivalent to the semiparametric model with likelihood $\mathcal{L}(b, \rho; \Delta, L_{obs})$ given by (16) modified by replacing $\mathcal{L}^F(\alpha, \theta; L)$ by $\mathcal{L}^F(b, \rho; L)$. $\mathcal{L}^F(b, \rho; L)$ and $\mathcal{L}(b, \rho; \Delta, L_{obs})$ are simply reparameterizations of $\mathcal{L}^F(\alpha, \theta; L)$ and $\mathcal{L}(\alpha, \theta; \Delta, L_{obs})$ of (16a) and (16b). Now the semiparametric model "obs†" also has likelihood function $\mathcal{L}(b, \rho_1, \rho_2, \rho_3; \Delta, L_{obs})$. Thus the asymptotic properties of the semiparametric estimator $\hat{b}(h, \phi)$ and its variance estimator, as well as the form of the efficient score for b and of the orthogonal complement $\Lambda_0^\perp$ to the nuisance tangent space for b, do not depend on whether true value of ρ_1 is $\rho_{10} \equiv \Pr(Y=1)$ or γ_0. It follows that parts a and b of Lemma 6.1 hold for model "obs†." Part b for γ_0 known and part c follow from $\{Y_i; i = 1, \ldots, n\}$ ancillary for b in the parameterization $\mathcal{L}(b, \rho; \Delta, L_{obs})$.

APPENDIX D: COMPUTATIONALLY CONVENIENT SIMULATION ESTIMATORS

We provide, for any $B = b(W, X)$, easily computed simulation estimates $\hat{E}_{\alpha,\eta}(B|W)$ and $\hat{E}_{\alpha,\eta}(B|X^*)$ of $E_{\alpha,\eta}(B|W)$ and $E_{\alpha,\eta}(B|X^*)$, where $E_{\alpha,\eta}[\cdot|\cdot]$ denotes expectations with respect to a law $\mathcal{L}^F(\alpha, \eta; L)$ allowed by the model "full" characterized by (1). Our estimators $\hat{E}_{\alpha,\eta}[\cdot|\cdot]$ are based on the identities $E[B|X^* = x^*] \equiv E[b(W^\dagger, x^*)q(W^\dagger, x^*)]/E[q(W^\dagger, x^*)]$, with $q(W^\dagger, x^*) \equiv f(V^\dagger|Y, x^*)f(Y|x^*)/f(V^\dagger, Y)$, and $E(B|W = w) = E[b(X, w)p(X, w)]/E[p(X, w)]$, where $p(X, w) = f(v^\dagger|X, y, v)f(y|X, v)I(V = v)$, for V discrete and, if V has continuous components, $p(X, w) = f(v^\dagger|X, y, v)f(y|X, v)f(X|v)/f(X)$. These identities motivate generating a large number, say n^*, of independent draws L_j from $\mathcal{L}^F(\alpha, \eta; L)$ and defining (a) $\hat{E}_{\alpha,\eta}[B|X^* = x^*] \equiv \sum_{j=1}^{n^*} b(W_j^\dagger, x^*)\hat{q}(W_j^\dagger, x^*)/\sum_{j=1}^{n^*}\hat{q}(W_j^\dagger, x^*)$, where $\hat{q}(W_j^\dagger, x^*) \equiv f[V_j^\dagger|Y_j, x^*; \eta_4]f[Y_j|x^*; \alpha, \eta_1]/\hat{f}(W_j^\dagger; \alpha, \eta_1, \eta_4)$, $f[Y_j|x^*; \alpha, \eta_1] \equiv f_\varepsilon[Y_j - g(x^*; \alpha)|x^*; \alpha, \eta_1]$, $\hat{f}(W_j^\dagger; \alpha, \eta_1, \eta_4) \equiv \sum_{k=1}^{n^*} f[V_j^\dagger|Y_j, X_k^*; \eta_4]f[Y_j|X_k^*; \alpha, \eta_1]/n^*$, and (b) $\hat{E}_{\alpha,\eta}(B|W = w) \equiv \sum_{j=1}^{n^*} b(X_j, w)\hat{p}(X_j, w)/\sum_{j=1}^{n^*}\hat{p}(X_j, w)$, where $\hat{p}(X_j, w) \equiv f(v^\dagger|X_j, y, v; \eta_4)f(y|X_j, v; \alpha, \eta_1)I(V_j = v)$ for V discrete and $\hat{p}(X_j, w) \equiv f(v^\dagger|X_j, y, v; \eta_4)f(y|X_j, v; \alpha, \eta_1)f(X_j|v; \eta_2)/\hat{f}(X_j; \eta_2)$ for V with continuous components, with $\hat{f}(X_j; \eta_2) \equiv \sum_{k=1}^{n^*} f(X_j|V_k; \eta_2)/n^*$. By the central limit theorem, $\hat{E}_{\hat{\alpha},\hat{\eta}}[B|W = w]$ and $\hat{E}_{\hat{\alpha},\hat{\eta}}[B|X^* = x^*]$ converge to $E_{\hat{\alpha},\hat{\eta}}(B|W = w)$ and $E_{\hat{\alpha},\hat{\eta}}[B|X^* = x^*]$ at a $\sqrt{n^*}$ rate. These simulation estimators were motivated by a similar estimator proposed by Daniel Rabinowitz in an unpublished manuscript.

[Received December 1990. Revised September 1993.]

REFERENCES

Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Ser. B, 37, 129–145.

Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983), "Information and Asymptotic Efficiency in Parametric-nonparametric Models," *The Annals of Statistics*, 11, 432–452.

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in 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.

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993), "Case-Control Studies With Errors in Covariates," *Journal of the American Statistical Association*, 88, 185–199.

Carroll, R. J., and Wand, M. P. (1991), "Semi-Parametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, Ser. B, 53, 573–587.

Carroll, R. J., Wang, S., and Wang, C. Y. (1993), "Asymptotics for Prospective Analysis of Stratified Case-Control Studies," submitted to *Journal of the American Statistical Association*.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–324.

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, Cambridge, MA: MIT Press, pp. 51–111.

Dagenais, M. G. (1973), "The Use of Incomplete Observations and Multiple Regression Analysis: A Generalized Least Squares Approach," *Journal of Econometrics*, 1, 317–328.

Flanders, W. D., and Greenland, S. (1991), "Analytic Methods for Two-Stage Case-Control Studies and Other Stratified Designs," *Statistics in Medicine*, 10, 739–747.

Gourieroux, C., and Montfort, A. (1981), "On the Problem of Missing Data in Linear Models," *Review of Econometric Study*, xlviii, 579–586.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, 19, 2244–2253.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Huber, P. (1985), "Projection Pursuit," *The Annals of Statistics*, 13, 435–474.

Imbens, G. W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models With Choice-Based Sampling," *Econometrica*, 60, 1187–1214.

Imbens, G. W., and Lancaster, T. (1991), "Efficient Estimation and Stratified Sampling," Discussion Paper 1545, Harvard Institute of Economic Research.

Kalbfleisch, J. D., and Lawless, J. F. (1988), "Likelihood Analysis of Multi-State Models for Disease Incidence and Mortality," *Statistics in Medicine*, 7, 149–160.

Kress, R. (1989), *Linear Integral Equations*, Berlin: Springer-Verlag.

Lancaster, T. (1990), "A Paradox in Choice-Base Sampling," working paper, Brown University.

Lin, D. Y., and Ying, Z. (1993), "Cox Regression With Incomplete Covariate Measurements," *Journal of the American Statistical Association*, 88, 1341–1349.

Little, R. J. A. (1993), "Regression With Missing X's: A Review. *Journal of the American Statistical Association*, 87, 1227–1237.

Little, R. J. A., and Rubin, D. (1987), *Statistical Analysis With Missing Data*, New York: John Wiley.

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, Cambridge, MA: MIT Press, pp. 2–50.

Newey, W. K. (1990a), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1990b), "Efficient Estimation of Tobit Models Under Conditional Symmetry," in *Semiparametric and Nonparametric Methods in Econometrics and Statistics*, eds. W. Barnett, J. Powell, and G. Tauchen, Cambridge: Cambridge University Press, pp. 291–336.

——— (1993a), "The Asymptotic Variance of Semiparametric Estimators," submitted to *Econometrica*.

——— (1993b), "Series Estimation of Regression Functionals," MIT mimeo.

Newey, W. K., and McFadden, D. (1993), "Estimation in Large Samples," in *Handbook of Econometrics*, Vol. 4, eds. D. McFadden and R. Engler, Amsterdam: North-Holland.

Newey, W. K., and Powell, J. (1990), "Efficient Estimation of Linear and Type I Censored Regression Models Under Conditional Quantile Restrictions," *Econometric Theory*, 6, 295–317.

Pepe, M. S., and Fleming, T. R. (1991), "A Nonparametric Method for Dealing With Mismeasured Covariate Data," *Journal of the American Statistical Association*, 86, 108–113.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.

Prentice, R. L. (1986), "A Case-cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

Pugh, M., Robins, J. M., Lipsitz, S., and Harrington, D. (1992), "Inference in the Cox Proportional Hazards Model With Missing Covariates," Technical Report, Harvard School of Public Health, Dept. of Biostatistics.

Ritov, Y., and Wellner, J. A. (1988), "Censoring, Martingales, and the Cox Model," in *Contemporary Mathematics: Statistical Inference for Stochastic Processes* (Vol. 80), ed. N. U. Prabhu, Providence, RI: American Mathematical Society, 191–220.

Robins, J. M. (1987), "A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies With Sustained Exposure Periods," *Journal of Chronic Diseases*, 40, Supplement 2, 139s–161s.

Robins, J. M., and Morgenstern, H. (1987), "The Foundations of Confounding in Epidemiology," *Computers and Mathematics with Applications*, 14, 869–916.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992), "G-Estimation of the Effect of Prophylaxis Therapy for *Pneumocystis carinii* Pneumonia on the Survival of AIDS Patients," *Epidemiology*, 3, 319–336.

Robins, J. M., Hsieh, F., and Newey, W. (1995), "Semiparametric Efficient Estimation of a Conditional Density With Missing or Mismeasured Covariates," submitted to the *Journal of the Royal Statistical Society*, Sen. B.

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 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, Boston: Birkhäuser, pp. 297–331.

Robins, J. M. (1993a), "Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers," *Proceedings of the Biopharmaceutical Section, American Statistical Association*, pp. 24–33.

——— (1993b), "Analytic Methods for HIV Treatment and Cofactor Effects," in *Methodological Issues of AIDS Behavioral Research*, eds. D. G. Ostrow and R. Kessler, New York: Plenum Press, pp. 213–287.

Robins, J. M., Greenland, S., and Rotnitzky, A. (1992), "Parametric Models for Nonmonotone Missing Data Processes," Technical Report, Harvard School of Public Health, Dept. of Epidemiology.

Robins, J. M., and Rotnitzky, A. (1994), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*.

Rosenbaum, P. R. (1984), "The Consequences of Adjustment for a Concommitant Variable That Has Been Adversely Affected by Treatment," *Journal of the Royal Statistical Society*, Ser. A, 147, 656–666.

——— (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

Rotnitzky, A., and Robins, J. M. (1993), "Efficient Semiparametric Estimation With Missing Outcomes and Surrogate Data," technical report, Harvard School of Public Health, Dept. of Epidemiology.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Stampfer, M. J., Willett, W. C., Colditz, G. A., Rosner, B., Speizer, F. E., and Hennekens, C. H. (1985), "A Prospective Study of Postmenopausal Estrogen Therapy and Coronary Heart Disease," *New England Journal of Medicine*, 313, 1044–1049.

Thomas, D. C. (1977), "In Appendix to Liddell, F. D. K., McDonald, J. C., and Thomas, D. C.," *Journal of the Royal Statistical Society*, Ser. B, 140, 469–490.

Van der Laan, M. J. (1993), *Efficient and Inefficient Estimation in Semiparametric Models*, Doctoral Dissertation, University of Utrecht, The Netherlands.

Weinberg, C. R. (1993), "Towards a Clearer Definition of Confounding," *American Journal of Epidemiology*, 137, 1–3.

Weinberg, C. R., and Wacholder, S. (1993), "Prospective Analysis of Case-Control Data Under General Multiplicative-Intercept Risk Models," submitted to *Biometrika*.

Zhao, L. P., and Lipsitz, S. (1992), "Design and Analysis of Two-Stage Studies," *Statistics in Medicine*, 11, 769–782.