

Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders

James M. Robins and Steven D. Mark

Harvard School of Public Health, 665 Huntington Avenue,
Boston, Massachusetts 02115, U.S.A.

and

Whitney K. Newey

Department of Economics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, U.S.A.

SUMMARY

In order to estimate the causal effects of one or more exposures or treatments on an outcome of interest, one has to account for the effect of "confounding factors" which both covary with the exposures or treatments and are independent predictors of the outcome. In this paper we present regression methods which, in contrast to standard methods, adjust for the confounding effect of multiple continuous or discrete covariates by modelling the conditional expectation of the exposures or treatments given the confounders. In the special case of a univariate dichotomous exposure or treatment, this conditional expectation is identical to what Rosenbaum and Rubin have called the propensity score. They have also proposed methods to estimate causal effects by modelling the propensity score. Our methods generalize those of Rosenbaum and Rubin in several ways. First, our approach straightforwardly allows for multivariate exposures or treatments, each of which may be continuous, ordinal, or discrete. Second, even in the case of a single dichotomous exposure, our approach does not require subclassification or matching on the propensity score so that the potential for "residual confounding," i.e., bias, due to incomplete matching is avoided. Third, our approach allows a rather general formalization of the idea that it is better to use the "estimated propensity score" than the true propensity score even when the true score is known. The additional power of our approach derives from the fact that we assume the causal effects of the exposures or treatments can be described by the parametric component of a semiparametric regression model. To illustrate our methods, we reanalyze the effect of current cigarette smoking on the level of forced expiratory volume in one second in a cohort of 2,713 adult white males. We compare the results with those obtained using standard methods.

1. Introduction

1.1 *The Problem*

In order to estimate the causal effect of one or more exposures or treatments on an outcome of interest, one has to account for the effect of "confounding factors" which both covary with the exposures or treatments and are independent predictors of the outcome. If few in number, categorical confounding factors are commonly dealt with by stratification. When there are many confounding factors or when some of the factors are continuous, regression methods are used. In this paper we present regression methods which, in contrast to standard methods, adjust for confounding by modelling aspects of the marginal association of the exposures of interest with the confounders rather than by modelling the independent

Key words: Causal inference; Covariance adjustment; Epidemiologic methods; Propensity score; Semiparametric efficiency; Semiparametric regression.

association of the confounders with the outcome. Specifically, we will model the conditional expectation of the exposures given the confounders. These methods of estimation will be particularly useful when prior knowledge regarding the association of the confounders with exposure status is more precise than knowledge regarding their association with the outcome.

For concreteness, we shall attempt to estimate the effect of being a current cigarette smoker on the level of forced expiratory volume in one second (FEV1) in a cohort of 2,713 adult white male former and current cigarette smokers from the initial cross-sectional data collected in the Harvard Six Cities Study (Dockery et al., 1988). We shall estimate this effect while adjusting for the presence of the 22 potential confounding factors listed in Table 1 that include past smoking history, past respiratory symptoms, age, height, and coexistent heart disease. In this example the exposure of interest is dichotomous and we assume that there is no interaction between that exposure and the confounders. That is, we assume that the absolute effect of current smoking on FEV1 does not depend on a subject's age, weight, previous smoking history, etc. In this setting the most common approach to estimating the effect of current smoking on FEV1 would be to postulate a linear regression model

$$Y_i = \beta_1 + \beta S_i + \sum_{k=2}^K \beta_k X_{k,i} + \varepsilon_i, \quad E[\varepsilon_i | S_i, X_i] = 0, \quad (1)$$

where $Y_i, S_i, X_i = (X_{2,i}, \dots, X_{K,i})$ are respectively random variables representing subject i 's FEV1 level, current smoking status ($S_i = 1$ if a current smoker and $S_i = 0$ otherwise), and values on a vector X_i of potential confounding factors. Note that the parameter of interest, β , is distinguished from the "nuisance" parameters $(\beta_1, \dots, \beta_K)$ by the absence of a subscript. For notational simplicity, we shall assume that (Y_i, S_i, X_i) are independent and identically distributed random vectors, although, with minor modifications, our results will hold if the X_i are fixed constants and the (ε_i, S_i) are independent across subjects.

Define $\sigma^2(S, X) = \text{var}[\varepsilon_i | S, X]$. We write $\sigma^2(S, X) = \sigma^2$ if the errors ε_i are homoscedastic. Unless stated otherwise, we shall assume homoscedastic errors, although we do not assume that this fact is known to the data analyst. The ε_i are not assumed to be independent of the (S_i, X_i) .

Suppose we are unwilling to assume that the independent association of the confounders X_i with the outcome Y_i has a known functional form. In that case, we would generalize model (1) to

$$Y_i = \beta S_i + h(X_i) + \varepsilon_i, \quad E[\varepsilon_i | S_i, X_i] = 0, \quad (2)$$

where $h(X_i)$ is an unknown real-valued function of the vector X_i . Model (2) has a

Table 1
Twenty-two potential confounders of the effect of current smoking on FEV1

| | |
|--|-------------------------------------|
| Age | History of emphysema |
| Age-squared | Past history of asthma |
| Height | Current asthma |
| Body mass index | Former cigar smoker |
| Chronic cough | Current cigar smoker level = hi |
| Recurrent bouts of coughing | Current cigar smoker level = medium |
| History of treatment for heart disease | Current cigar smoker level = lo |
| Chronic phlegm production | Former pipe smoker |
| Chronic wheeze | Current pipe smoker level = hi |
| Total years of cigarette smoking | Current pipe smoker level = medium |
| Lifetime pack-years smoked | Current pipe smoker level = lo |

semiparametric regression function with parametric component βS_i and nonparametric component $h(X_i)$. This paper is concerned with the estimation of β from model (2). Robinson (1988) has provided an asymptotically normal and unbiased estimator of β under a large-sample limiting model in which the number of confounding factors remains fixed as the sample size grows. His estimator relies on the fact that, under such a limiting model, the unknown function $h(X_i)$ can be consistently estimated by nonparametric regression techniques. In epidemiologic research, the number of confounding variables can be quite large. In these instances, the more appropriate limiting model would be one in which we allowed the number of confounding factors contained in X_i to increase with the sample size (Huber, 1981).

It is difficult to generalize Robinson's approach based on nonparametric estimation of $h(X_i)$ when the dimension of X_i is large. As a consequence, to obtain consistent estimators of β , we shall consider making additional a priori assumptions beyond those specified by model (2). The standard approach would be to assume that $h(X_i)$ is known a priori except for a finite number of unknown parameters. As an example, the linear regression model (1) assumes that

$$h(X_i) = \beta_1 + \sum_{k=2}^K \beta_k X_{k,i}.$$

In contrast to the standard approach, in this paper we shall suppose that prior information concerning the marginal association of S_i with X_i is sharper than that concerning the form of $h(X_i)$. Thus we shall leave $h(X_i)$ completely unspecified and instead specify parametric models for the marginal association of S_i and X_i . Specifically, we shall consider parametric models for $E(S|X_i) = p(S = 1|X_i)$ such as the logistic regression model

$$p[S = 1|X_i; \alpha] = \frac{\exp(\alpha_1 + \sum_{k=2}^K \alpha_k X_{k,i})}{1 + \exp(\alpha_1 + \sum_{k=2}^K \alpha_k X_{k,i})}, \tag{3}$$

where $\alpha = (\alpha_1, \dots, \alpha_K)$. We shall show that we can obtain asymptotically normal and unbiased estimators of β in model (2) provided our model (3) for $p(S = 1|X_i)$ is correctly specified.

Although correctly specified parametric models for either $h(X_i)$ or $p(S = 1|X_i)$ will provide asymptotically normal and unbiased estimates of β , nonetheless, as discussed in the next paragraph, least squares estimators of β based on models for $h(X_i)$ will always be at least as efficient as any estimator of β based on models for $p(S = 1|X_i)$. This suggests that, for reasons of efficiency, it is always preferable to model $h(X_i)$ rather than $p(S = 1|X_i)$. But if, as we assume in this paper, our prior information concerning $h(X_i)$ is less sharp than that concerning $p(S = 1|X_i)$, we would choose not to model $h(X_i)$ in order to protect against specification bias.

In order to explain why the ordinary least squares estimator of β based on a correctly specified model for $h(X_i)$ is always at least as efficient as any estimator of β based on models for $E[S|X_i]$, we need to review some results from the theory of semiparametric efficiency bounds derived by Chamberlain (1987; and Discussion Paper 1494, Harvard Institute of Economic Research, 1990) and exposted by Newey (1990). For the moment suppose again that, as in equation (1), we were able to correctly specify a parametric model, say, $q(X_i; \theta)$ for $h(X_i)$ depending on a parameter vector θ . In equation (1), $\theta = (\beta_1, \dots, \beta_K)$. Chamberlain (1987) showed that the estimator of β obtained by fitting the model $Y_i = \beta S_i + q(X_i; \theta) + \varepsilon_i$ by unweighted, possibly nonlinear, least squares is the most efficient possible estimator of β that is guaranteed to be asymptotically normal and unbiased under the sole prior restrictions that $E[\varepsilon_i|S_i, X_i] = 0$ and $h(X_i) = q(X_i; \theta)$. [If, as in equation (1), $q(X_i; \theta)$ is linear in θ , we fit using ordinary least squares. Otherwise, we fit using nonlinear least

squares.] Therefore if, as in model (2), we are unwilling to specify a parametric form for $h(X_i)$ and yet want our estimator of β to be asymptotically normal and unbiased whatever be $h(X_i)$, the asymptotic variance of any such estimator clearly cannot be less than the supremum of the asymptotic variances of the least squares estimators of β taken over the set of all possible parametric models for $h(X_i)$. This supremum is called the semiparametric efficiency bound for an estimator of β under model (2) (Bickel et al., 1992) and was shown by Chamberlain (discussion paper cited previously) to equal $n^{-1}\sigma^2/E[\text{var}(S|X)]$, where n is the sample size.

Thus, if we are able to correctly specify a parametric model for $h(X_i)$, the least squares estimator of β always has variance no greater than the efficiency bound $n^{-1}\sigma^2/E[\text{var}(S|X)]$. In contrast, if under model (2), we are unable to specify a parametric model for $h(X_i)$, but instead correctly specify a model for $E[S|X_i]$, no estimator that is asymptotically unbiased for β for all $h(X_i)$ can have variance less than the bound $n^{-1}\sigma^2/E[\text{var}(S|X)]$. This is a consequence of the fact that $\{(S_i, X_i), i \in (1, \dots, n)\}$ is ancillary for β under model (2) (Cox and Hinkley, 1974) and, as discussed by Newey (1990), knowledge concerning the marginal distribution of an ancillary statistic does not affect the semiparametric efficiency bound for the estimation of β .

It needs to be stressed that, even when we can obtain a consistent estimator of β in model (2), it does not follow that the parameter β can be interpreted as the causal effect of current cigarette smoking on FEV1. We now describe conditions under which β does have a causal interpretation.

1.2 A Causal Model

Following Rubin (1978), let $Y_{S=1,i}$ be subject i 's FEV1 had subject i been a current smoker. If subject i is a current smoker in the actual study, then $Y_{S=1,i}$ equals his observed FEV1 Y_i . If subject i is not a current smoker, $Y_{S=1,i}$ is missing. Similarly, $Y_{S=0,i}$ is subject i 's FEV1 if subject i were, possibly contrary to fact, a current nonsmoker. Rubin defined the average causal effect of current smoking among subjects with observed covariates level X_i to be $E[Y_{S=1}|X_i] - E[Y_{S=0}|X_i]$. Now, under our model (2), we know that $E[Y|X_i, S = 1] - E[Y|X_i, S = 0] = \beta$ since $E[Y|X_i, S = 1] = \beta + h(X_i)$ and $E[Y|X_i, S = 0] = h(X_i)$.

Thus a sufficient condition for β to equal the average causal effect of current smoking at each level X_i is that, for each X_i ,

$$E[Y_{S=s}|X_i] = E[Y|X_i, S = s], \quad s \in \{0, 1\}. \quad (4a)$$

Under Rubin's causal model, equation (4a) is equivalent to

$$E[Y_{S=s}|X_i] = E[Y_{S=s}|X_i, S = s]. \quad (4b)$$

We shall assume that equation (4b) holds and thus β has a causal interpretation when X_i is the vector of 22 potential confounding variables described above. The assumption that equation (4b) holds is nonidentifiable in the sense that it is compatible with any joint distribution for the observable random variables (S_i, X_i, Y_i) . When equation (4b) holds, we shall call model (2) a *semiparametric causal regression model*. Equation (4b) says that, conditional on the joint level of the 22 potential independent risk factors X_i , the mean of $Y_{S=s}$ among subjects who actually receive treatment $S = 1$ equals that among subjects who actually receive treatment $S = 0$. We do not assume that equation (4b) holds when X_i is a proper subset of the 22 potential confounding variables.

The mathematical results in this paper are concerned only with the estimation of β in model (2) and do not depend on whether equation (4b) holds. Of course, in general, we are interested in the estimation of β only when we believe it has a causal interpretation.

1.3 Relationship to the Propensity Score

Rosenbaum and Rubin (1983, 1984, 1985) and Rosenbaum (1984, 1987, 1988) have also considered estimating the causal effect of a dichotomous treatment such as S_i on an outcome Y_i by modelling $p(S = 1|X_i)$ when equation (4) holds. These authors call $p[S = 1|X_i]$ the propensity score. In contrast to their approach, our approach straightforwardly allows the treatment or exposure S_i to be continuous or ordinal rather than simply dichotomous. Furthermore, as discussed in the Appendix, our approach allows S_i to be multivariate so that we can, say, estimate the independent effects of current cigarette smoking and past cigarette smoking. In addition, our "regression" approach does not require subclassification or matching on the propensity score $p[S = 1|X_i]$ even when X_i has continuous components so that the potential for "residual" confounding, i.e., bias, due to the fact that one has not precisely matched on $p[S = 1|X_i]$ is avoided. The additional power of our approach derives from the fact that we assume the causal effect of exposure can be described by the parametric component of a semiparametric causal regression model such as model (2).

Rosenbaum (1984, 1988) also considered specifying causal models to avoid the need to match or subclassify on the propensity score. In general, Rosenbaum is concerned with small-sample (exact) rather than large-sample (asymptotic) inference. As a consequence, his causal models tend to be even more restrictive than model (2). Specifically, he assumes a constant treatment effect model—that is, $Y_{S=1,i} = \beta + Y_{S=0,i}$ for all subjects i —although his results would still hold under the weaker assumption that the distributions of $Y_{S=1,i}$ and $Y_{S=0,i}$ differed by a "shift" parameter β . Furthermore, as he points out, his "exact" methods do not allow one to adjust for the confounding effects of continuous covariates.

Finally, as discussed in Section 2, our approach allows a rather general formalization of the idea that it is better to use the "estimated" propensity score than the "true" propensity score even when the true score is known (Rosenbaum, 1987).

2. Estimators Based on Models for the Conditional Expectation of Exposure Given Confounders

2.1 An Infeasible Estimator

In this section, we consider estimators of β under model (2) when we can specify accurate models for $E(S|X_i)$. Note that when S_i is dichotomous, models for $E(S|X_i)$ are models for $p(S = 1|X_i)$. Initially, for pedagogic purposes, we shall assume that we know $E(S|X_i)$ exactly. That is, we assume exact prior knowledge of the expected value of S for every combination of the confounders X_i . Subsequently we make the more tenable assumption that we know $E(S|X_i)$ up to a finite vector of unknown parameters. We allow $\sigma^2(S, X)$ to depend on (S, X) . Henceforth, we adopt the following notational convention: β will refer to the true but unknown value of the coefficient of S_i in model (2); β^\dagger will refer to any hypothesized, possibly incorrect, value for β .

The estimator we shall consider, which we call the E-estimator,

$$\hat{\beta}_E = \frac{\sum_{i=1}^n Y_i [S_i - E(S|X_i)]}{\sum_{i=1}^n S_i [S_i - E(S|X_i)]}, \quad (5)$$

is based on a suggestion by Newey (1990).

It is shown in Theorem A.1 in the Appendix that $\hat{\beta}_E$ has a limiting normal distribution with mean β .

The consistency of $\hat{\beta}_E$ is based on the fact that model (2) implies that

$$E[z_i | X_i, S_i] = E[z_i | X_i], \quad (6)$$

where $z_i = Y_i - S_i\beta$. In the proof of Theorem A.1 in the Appendix, it is shown that equation (6) implies the identity

$$E[U(\beta)] = 0, \quad (7)$$

where, for any β^\dagger , $U(\beta^\dagger) = \sum_{i=1}^n (Y_i - S_i\beta^\dagger)(S_i - E(S|X_i))$. The E-estimator $\hat{\beta}_E$ is the solution β^\dagger to the unbiased estimating equation $U(\beta^\dagger) = 0$.

2.2 A Feasible Estimator

Of course, the estimator $\hat{\beta}_E$ is not feasible since, in practice, $E(S|X_i)$ is unknown. We can overcome this difficulty if we assume a priori that the logistic regression model equation (3) holds. We then estimate $E(S|X_i)$ by logistic regression and subsequently estimate β by

$$\tilde{\beta}_E = \frac{\sum_{i=1}^n Y_i [S_i - \hat{E}(S|X_i)]}{\sum_{i=1}^n S_i [S_i - \hat{E}(S|X_i)]}, \quad (8)$$

where $\hat{E}(S|X_i)$ is the fitted value $\hat{p}_i \equiv p[S = 1 | X_i; \hat{\alpha}]$ of $p[S = 1 | X_i]$, and $\hat{\alpha}$ is the maximum likelihood estimator of α from the logistic regression. Note that we use the symbol $\tilde{\beta}_E$ rather than $\hat{\beta}_E$ to represent the feasible estimator of equation (8).

As shown in Theorem A.1 in the Appendix, it follows from Pierce (1982) and Newey (1990) that when the logistic model of equation (3) is true, $\tilde{\beta}_E$ is asymptotically normal and unbiased and its asymptotic covariance matrix can be consistently estimated by

$$\text{var}_{\text{est}}^{\wedge}(\tilde{\beta}_E) = \text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E) - \hat{Q}[\text{var}_{\text{est}}^{\wedge}(\hat{\alpha})]\hat{Q}^T \quad (9)$$

where

$$\text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E) = \frac{\sum_{i=1}^n \tilde{z}_i^2 (S_i - \hat{p}_i)^2}{[\sum_{i=1}^n S_i (S_i - \hat{p}_i)]^2}, \quad (10)$$

$\tilde{z}_i \equiv Y_i - \tilde{\beta}_E S_i$, \hat{Q}^T is the K -vector with j th component

$$\hat{Q}_j = \frac{-\sum_{i=1}^n \tilde{z}_i \hat{p}_i (1 - \hat{p}_i) X_{j,i}}{\sum_{i=1}^n S_i (S_i - \hat{p}_i)}$$

(where we define $X_{j,i} = 1$ when $j = 1$), and $\text{var}_{\text{est}}^{\wedge}(\hat{\alpha})$ is the estimated covariance matrix (i.e., the inverse of the observed information matrix) from the fit of the logistic model equation (3). The observed information matrix has (j, k) entry $-\sum_{i=1}^n \hat{p}_i (1 - \hat{p}_i) X_{j,i} X_{k,i}$. The estimator $\text{var}_{\text{est}}^{\wedge}(\tilde{\beta}_E)$ is not guaranteed to be positive-definite. A positive-definite consistent variance estimator is obtained by replacing $\hat{p}_i (1 - \hat{p}_i)$ by $(S_i - \hat{p}_i)^2$ both in the numerator of \hat{Q}_j and in the observed information matrix.

Even though $\hat{\beta}_E$ is infeasible when $p[S = 1 | X_i]$ is unknown, $\text{var}_{\text{est}}^{\wedge}(\tilde{\beta}_E)$ is still a feasible consistent estimator of its asymptotic variance. Therefore it follows from equation (9) that one generates a more precise estimate of β by *estimating* the propensity score $E(S|X_i)$ than by using the true population value of the propensity score even were the latter known. That is, $\text{var}_{\text{est}}^{\wedge}(\tilde{\beta}_E)$ is always less than or equal to $\text{var}_{\text{est}}^{\wedge}(\hat{\beta}_E)$. As discussed in the Appendix, this result depends on the fact that $\hat{\alpha}$ is an efficient estimator of α . The preference for $\tilde{\beta}_E$ compared to $\hat{\beta}_E$ when the parameter α of model (3) is known can also be viewed in terms of conditional bias. Specifically, it can be shown that, conditional on the ancillary statistic $[\text{var}_{\text{est}}^{\wedge}(\hat{\alpha})]^{-1/2}(\hat{\alpha} - \alpha)$, $\hat{\beta}_E$ becomes asymptotically biased while $\tilde{\beta}_E$ remains asymptotically unbiased (Robins and Morgenstern, 1987; Rosenbaum, 1987; Efron and Hinkley, 1978).

In Table 2 we present four different estimates $\tilde{\beta}_E$ of β based on specifying four different

Table 2
 Estimates $\hat{\beta}_E$ under four different specifications for $p[S = 1 | X_i]$

| Analysis | Covariates $X_{k,i}$ included in logistic model of equation (3) for $p[S = 1 X_i]$ | $\hat{\beta}_E$ | $\text{var}_{\text{est}}^{\Delta}(\hat{\beta}_E) \times 10^{-4}$ | $\text{var}_{\text{est}}^{\Lambda}(\hat{\beta}_E) \times 10^{-4}$ |
|----------|---|-----------------|--|---|
| (1) | Constant term only | -.0580 | 9.49 | 159.0 |
| (2) | Constant, chronic cough (Yes, No) | .0429 | 9.36 | 167.0 |
| (3) | Constant, pack-years of smoking | .0520 | 7.45 | 157.9 |
| (4) | Constant, 22 covariates in Table 1 | -.1133 | 8.82 | 332.6 |

logistic regression models for $p[S = 1 | X_i]$. In the first analysis in Table 2, we assume no confounding. That is, we fit only a constant term α_1 in equation (3). In the second analysis X_i in equation (3) is the single binary covariate—history of chronic cough. In the third analysis X_i is the single continuous covariate—lifetime number of pack-years. In the fourth analysis X_i in equation (3) is the 22-vector of potential confounders. The striking efficiency advantage attributable to estimating the propensity score $p[S = 1 | X_i]$ can be obtained by comparing $\text{var}_{\text{est}}(\hat{\beta}_E)$ to $\text{var}_{\text{est}}(\hat{\beta}_E)$ in Table 2.

Under the assumptions that (a) the coefficient β in equation (2) has a causal interpretation [i.e., equation (4b) holds] when X_i is the 22-vector of confounders and (b) the model for $p[S = 1 | X_i]$ used in analysis (4) is true, analysis (4) provides a consistent estimator of this causal β . Therefore, we estimate that current smoking causes a decrease of .1133 liter in FEV1. A 95% confidence interval for β is $-.113 \pm (1.96)(.00088)^{1/2} = (-.170, -.056)$.

Under assumptions (a) and (b), we now provide sufficient conditions for the simpler analyses (1)–(3) also to provide consistent estimators of the “causal” β associated with model (2) with X_i the 22-vector of covariates.

We shall restrict attention to analysis (3) since the conditions for analyses (1) and (2) are similar. Let X_{k^*} be the covariate “lifetime number of pack-years” used in analysis (3). $\hat{\beta}_E$ from analysis (3) will be consistent for the causal β if either of the following is true:

Sufficient condition (1): With X_i the 22-vector of covariates, $\alpha_k = 0$ for the 21 covariates X_k in the logistic model (3) other than X_{k^*} (i.e., lifetime pack-years is the only predictor of current smoking among the 22 potential confounding factors).

Sufficient condition (2): The unknown function $h(X_i) = h(X_{2,i}, \dots, X_{K,i})$, $K = 22$, is actually only a function of $X_{k^*,i}$ (i.e., lifetime pack-years is the only independent risk factor among the 22 potential confounding factors) and $p[S = 1 | X_{k^*,i}]$ follows a linear logistic model.

In general it would be unlikely that an investigator would be willing to assume that either of the above sufficient conditions held, and thus would tend to rely on analysis (4).

Suppose equation (4b) holds and consider the test of the null hypothesis $\beta = 0$ that rejects if $\hat{\beta}_E \pm 1.96[\text{var}_{\text{est}}(\hat{\beta}_E)]^{1/2}$ fails to include 0. Then except for the assumption that the model (3) for $p[S = 1 | X_i]$ is correctly specified, this test is an “otherwise asymptotically distribution-free” .05 α -level test of the sharp null hypothesis of no causal effect of exposure, i.e., of the hypothesis $Y_{S=1,i} = Y_{S=0,i} = Y_i$ for all subjects i .

Rosenbaum (1984, §4.2) proposes a test of this null hypothesis that will be “otherwise asymptotically distribution-free” under the condition that $(Y_{S=1,i}, Y_{S=0,i})$ and S_i are conditionally independent given X_i .

3. Relationship of E-Estimators to Ordinary Least Squares

The ordinary least squares (OLS) estimator of β in equation (1) can be written

$$\hat{\beta}_{OLS} = \frac{\sum Y_i(S_i - \hat{P}(S|X_i))}{\sum [S_i - \hat{P}(S|X_i)]^2}, \tag{11}$$

where summation signs without indexes will refer to sums over individuals and where $\hat{P}(S|X_i)$ is the fitted value from the OLS regression of S on X_i and the constant one. Now the right-hand side of equation (11) can be written as

$$\hat{\beta}_{OLS} = \frac{\sum Y_i(S_i - \hat{P}(S|X_i))}{\sum S_i(S_i - \hat{P}(S|X_i))}$$

using the fact that, for OLS, the empirical correlation of the fitted values and the residuals is zero. Now suppose we had modelled $E(S|X_i) = p[S = 1 | X_i]$ by the linear probability model $p[S = 1 | X_i; \alpha] = \alpha_1 + \sum_{k=2}^K \alpha_k X_{k,i}$ rather than by a logistic model, and we fit the linear probability model by least squares. Then $\hat{E}(S|X_i) = \hat{P}(S|X_i)$. Therefore, from its definition $\hat{\beta}_E = \hat{\beta}_{OLS}$. In the previous section we showed that $\hat{\beta}_E$ is consistent if our model for $E(S|X_i)$ is true. It follows, as pointed out by Newey (1990), that if, in truth, $E(S|X_i) = \alpha_1 + \sum_{k=2}^K \alpha_k X_{k,i}$ [i.e., $E(S|X_i)$ is linear in X_i], then $\hat{\beta}_{OLS}$ is consistent for β even if $h(X_i)$ is nonlinear and thus equation (1) is false. Nonetheless if $h(X_i)$ is nonlinear, the estimate of the variance of $\hat{\beta}_{OLS}$ provided by standard software packages is inconsistent, and equation (9) must be used. If $E(S|X_i)$ is not linear in X_i , the ordinary least squares estimate of β would, in general, be inconsistent if the unknown function $h(X_i)$ is, in truth, nonlinear in X_i .

Table 3 shows $\hat{\beta}_{OLS}$ from the fit of equation (1) for the four choices of X_i as in Table 2. Note that $\hat{\beta}_{OLS} = \hat{\beta}_E$ in analysis (2). This reflects the fact that when X_i is a single dichotomous covariate, $E(S|X_i)$ is simultaneously linear and linear logistic, and $\hat{E}(S|X_i) = \hat{P}(S|X_i)$. For similar reasons $\hat{\beta}_{OLS} = \hat{\beta}_E$ in analysis (1). $\hat{\beta}_E$ from analysis (3) using the continuous variable "pack-years" is not identical to the OLS estimate since $\hat{E}(S|X_i) \neq \hat{P}(S|X_i)$. The fact that $\hat{\beta}_E$ and $\hat{\beta}_{OLS}$ are close can be explained by the near linearity of $\hat{E}(S|X_i)$ in our data, which can be checked by plotting $\hat{E}(S|X_i)$ versus X_i .

We now discuss a modification of the estimator $\hat{\beta}_E$ that has an even closer connection to OLS than does $\hat{\beta}_E$. Define

$$\hat{\beta}_{Em} = \frac{\sum Y_i(S_i - \hat{E}(S|X_i))}{\sum [S_i - \hat{E}(S|X_i)]^2}. \tag{12}$$

When $E(S|X_i)$ is nonlinear (e.g., logistic), $\hat{\beta}_{Em}$ will not in general equal $\hat{\beta}_E$. Nonetheless,

Table 3
Estimates $\hat{\beta}_{OLS}$ under four different specifications for covariates included in $\beta_1 + \sum \beta_k X_{k,i}$ in model equation (1)

| Analysis | Covariates $X_{k,i}$ included in equation (1) | $\hat{\beta}_{OLS}$ | $\text{var}_{est}(\hat{\beta}_{OLS})^a \times 10^{-4}$ |
|----------|---|---------------------|--|
| (1) | Constant term only | -.0580 | 9.50 |
| (2) | Constant, chronic cough (Yes, No) | .0429 | 9.39 |
| (3) | Constant, pack-years of smoking | .0492 | 7.64 |
| (4) | Constant, 22 covariates in Table 1 | -.1199 | 8.68 |

^a Using White's (1980) heteroscedastic consistent variance estimator.

$\hat{\beta}_{Em}$ and $\hat{\beta}_E$ have the same asymptotic distribution. One obtains $\hat{\beta}_{Em}$ by regressing Y_i versus “the residuals” $S_i - \hat{E}(S|X_i)$ using OLS regression with no intercept.

4. Two-Stage E-Estimators

Throughout this section we assume that the logistic model equation (3) is correctly specified with X_i the vector of 22 covariates. Then $\hat{\beta}_E$ is a consistent estimator of β in equation (2) without making any assumptions about the form of $h(X_i)$. Suppose now that we have an a priori guess as to the shape of $h(X_i)$. For concreteness, suppose we believed that $h(X_i)$ was linear or at least nearly linear in X_i , i.e., $h(X_i) = \beta_1 + \sum_{k=2}^K \beta_k X_{k,i}$. We can now consider how to develop an estimator, say $\hat{\beta}^*$, that may be much more efficient than $\hat{\beta}_E$ if our guess concerning the shape of $h(X_i)$ is correct or nearly correct, and will remain consistent, asymptotically normal no matter how wrong our guess may be. To construct $\hat{\beta}^*$, we proceed in two steps. First we compute $\hat{E}(S|X_i)$ and $\hat{\beta}_E$ as before. We then regress that $\hat{z}_i = Y_i - \hat{\beta}_E S_i$ on X_i . We then define $\hat{\beta}^*$ to be the solution β^* to the estimating equation

$$0 = U^*(\beta^*) \equiv \sum \left(Y_i - \beta^* S_i - \hat{\beta}_1 - \sum_{k=2}^K \hat{\beta}_k X_{k,i} \right) (S_i - \hat{E}(S|X_i)),$$

where $(\hat{\beta}_1, \dots, \hat{\beta}_K)$ are the OLS estimates from the regression of \hat{z}_i on X_i . Therefore,

$$\hat{\beta}^* = \frac{\sum (Y_i - \hat{\beta}_1 - \sum_{k=2}^K \hat{\beta}_k X_{k,i}) [S_i - \hat{E}(S|X_i)]}{\sum S_i (S_i - \hat{E}(S|X_i))}.$$

In Theorems A.1 and A.3 in the Appendix we show that $\hat{\beta}^*$ is asymptotically normal and unbiased even if the proposed linear model for $h(X_i)$ is incorrect.

A consistent estimator of $\text{var}^A(\hat{\beta}^*)$ is

$$\frac{\sum \hat{u}_i^2 (S_i - \hat{p}_i)^2}{[\sum S_i (S_i - \hat{p}_i)]^2} - (\hat{Q}^*) \text{var}_{\text{est}}(\hat{\alpha}) (\hat{Q}^*)^T, \tag{13}$$

where $\hat{u}_i = Y_i - \hat{\beta}^* S_i - \sum_{k=2}^K \hat{\beta}_k X_{k,i} - \hat{\beta}_1$ and $(\hat{Q}^*)^T$ has components

$$\hat{Q}_j^* = \frac{-\sum \hat{u}_i \hat{p}_i (1 - \hat{p}_i) X_{j,i}}{\sum S_i (S_i - \hat{p}_i)}.$$

In our example $\hat{\beta}^*$ is $-.117$ with $\text{var}^A(\hat{\beta}^*) = 8.79 \times 10^{-4}$ when X_i is the 22-vector of covariates.

In the final paragraph of the Appendix we show that, if the linear model postulated for $h(X_i)$ were correct, then (1) \hat{Q}^* converges to zero in probability so the correction term could be ignored; and (2) if $\sigma^2(S, X) = \sigma^2$, then $\text{var}^A(\hat{\beta}^*) = n^{-1} \sigma^2 / E[\text{var}(S|X)]$.

When $h(X_i)$ is linear, $\hat{\beta}_{OLS}$ will be consistent asymptotically normal and $\text{var}^A(\hat{\beta}_{OLS})$ will be less than or equal to $\text{var}^A(\hat{\beta}^*)$, with equality when $E(S|X_i)$ is linear in X_i . Of course, if neither $h(X_i)$ nor $E(S|X_i)$ is linear, $\hat{\beta}^*$ but not $\hat{\beta}_{OLS}$ remains consistent [provided the nonlinear model for $E(S|X_i)$ is correctly specified]. When $\sigma^2(S, X) = \sigma^2$ and $h(X_i)$ is, in truth, linear, $\hat{\beta}^*$ has the smallest asymptotic variance among all estimators that remain asymptotically unbiased even were $h(X_i)$ nonlinear (Chamberlain, discussion paper cited previously). That is, it attains the semiparametric efficiency bound for model (2).

5. Discussion

Suppose again that β in model (2) is causal [i.e., equation (4b) holds] when X_i is the 22-vector of covariates. Then the validity of our E-estimators of the causal effect of current smoking on FEV1 requires that the semiparametric regression model (2) and logistic

regression (3) be correctly specified. Specification of (3) can be checked using the techniques described by Landwehr, Pregibon, and Shoemaker (1984). The no-interaction assumption of model (2) can be checked by nesting (2) in the more general semiparametric regression model of the Appendix that includes interactions between current smoking S_i and the covariates in X_i , and then testing whether the interaction coefficients are nonzero.

We note that, rather than simply modelling $p[S = 1 | X_i]$ by the linear no-interaction logistic model equation (3), we could continue to add to equation (3) additional terms such as powers of the $X_{k,i}$ (e.g., $X_{k,i}^2, X_{k,i}^3$) for continuous covariates and all orders of interaction between the various covariates and their powers (e.g., $X_{3,i}^2 \cdot X_{4,i} \cdot X_{5,i}^3$). This will greatly increase the number of free coefficients in our model for $p[S = 1 | X_i]$. As we add these additional terms, we derive two benefits. First, we decrease any asymptotic bias in $\hat{\beta}_E$ (or $\hat{\beta}^*$) due to possible misspecification of the linear no-interaction model for $p[S = 1 | X_i]$. Second, when the linear no-interaction logistic model is *correctly specified* and thus the additional terms are not necessary to make $\hat{\beta}_E$ unbiased, the asymptotic variance of $\hat{\beta}_E$ (or $\hat{\beta}^*$) is nonincreasing and will usually decrease as the number of free parameters in the model for $p[S = 1 | X_i]$ increases [see Pierce (1982) and Corollary A.1 of the Appendix]. Thus, rather than having the usual tradeoff between efficiency and bias, we find that increasing the number of free parameters can lead to improvements in both bias and efficiency. This apparent "free lunch" must be tempered by two facts. First, no matter how many terms we add, $\text{var}^{\wedge}(\hat{\beta}_E)$ and $\text{var}^{\wedge}(\hat{\beta}^*)$ will always exceed $n^{-1}\sigma^2/E[\text{var}(S|X)]$ (with homoscedastic errors) (Chamberlain, 1987). Second, the results we have derived require that the estimates of the free parameters in the model for $p[S = 1 | X_i]$ are $n^{1/2}$ -consistent. [Newey (1990) suggests that $n^{1/4}$ -consistency is sufficient.] This limits the number of free parameters we may have in our model for $p[S = 1 | X_i]$ as a function of sample size. For example, we could not allow the number of free parameters to equal the total sample size. Cross-validation techniques for model selection should be useful in choosing a proper ratio of sample size to parameters. Moderate and small-sample simulation studies are needed as a guide to practice.

We note that when the linear no-interaction logistic model (3) is *misspecified*, the asymptotic variance of the (now potentially biased) estimator $\hat{\beta}_E$ based on a misspecified model for $p[S = 1 | X_i]$ can be less than the asymptotic variance of the estimator $\hat{\beta}_E$ based on a more richly parameterized, correctly specified model in which the misspecified model is nested. This phenomenon is evident in a comparison of analyses (3) and (4) in Table 2. The estimated variance of $\hat{\beta}_E$ in analysis (3) is less than that in analysis (4), because covariates other than "pack-years of smoking" are also important predictors of current smoking.

The results described in the preceding three paragraphs help to clarify both when E-estimation will and will not be preferable to standard covariance adjustment by least squares. Consider first the case in which the sample size is quite large and the dimension of X_i is small, so that richly parameterized models for either $h(X_i)$ or $p[S = 1 | X_i]$ can be used. Then, as discussed above and in technical detail by Newey (1990), as one adds power and interaction terms to the model (3) for $p[S = 1 | X_i]$, any bias in $\hat{\beta}^*$ and $\hat{\beta}_E$ would tend to zero and the asymptotic variance of $\hat{\beta}^*$, and even $\hat{\beta}_E$, will approach the semiparametric efficiency bound of $n^{-1}\sigma^2/E[\text{var}(S|X)]$. Similarly, in this setting, if we expanded the linear regression model (1) by adding additional terms such as powers of $X_{k,i}$ and interactions between the $X_{k,i}$ and their powers, the bias of $\hat{\beta}_{OLS}$ from the least squares fit of (1) would tend to zero, and the variance of $\hat{\beta}_{OLS}$ would approach the efficiency bound $n^{-1}\sigma^2/E[\text{var}(S|X)]$. Thus, in this setting, the use of highly parameterized models for $h(X_i)$ fit by least squares or the use of highly parameterized models for $p[S = 1 | X_i]$ fit by E-estimation leads to estimators of β with similar properties.

Now, consider the case in which the dimension of X_i is large and/or the sample size is moderate. One is then restricted to choosing parsimonious parametric models for $h(X_i)$ and/or $p[S = 1 | X_i]$. Further, since the ratio of the sample size to the dimension of X_i is small, the power to discriminate between correct and incorrectly specified models for $h(X_i)$ and/or $p[S = 1 | X_i]$ will be poor. If, as is often the case in an etiologic study, our primary interest is in obtaining valid inferences concerning β (e.g., confidence intervals that cover at their nominal rate), it is essential to try to obtain asymptotically unbiased estimators of β . Since, in general, unbiased estimation of β requires that the model used in the analysis be correct, we would prefer E-estimation over least squares estimation if we believed that our ability to specify nearly correct parsimonious models for $p[S = 1 | X_i]$ exceeded our ability to specify such models for $h(X_i)$. This would be the case when the investigator thinks, based on substantive considerations, that his or her knowledge of the shape of the regression surface $p[S = 1 | X_i]$ is sharper than knowledge of the shape of the function $h(X_i)$. In the special case, represented by our example in Section 3, in which the fitted regression surface $\hat{p}[S = 1 | X_i]$ is nearly linear in the X_i , E-estimation and standard covariance adjustment by least squares will provide similar estimates irrespective of whether $h(X_i)$ is or is not linear.

We next consider whether it might be possible to develop robust E-estimators. Even if the linear model (1) were true, the efficiency of $\hat{\beta}_{OLS}$ would be poor if the errors ϵ_i have heavy-tailed distributions (Huber, 1981). If we are willing to assume that, in addition to (1), the errors were independent of the (S_i, X_i) , efficient robust estimation based on M, L, or R estimators is possible (Huber, 1981). If ϵ_i is independent of (S_i, X_i) but model (1) were not true, robust E-estimation of model (2) could be based on solving an unbiased estimating equation of the form $\sum_i m(Y_i - \beta^T S_i, X_i)(S_i - E[S | X_i]) = 0$, where the function $m(Y_i - \beta^T S_i, X_i)$ would be chosen to downweight observations for which $Y_i - \beta^T S_i$ differs greatly from its expected value given X_i . (Such observations will be associated with large values of the residuals.) How to choose the function $m(Y_i - \beta^T S_i, X_i)$ in this setting is outside the scope of this paper.

If $\sigma^2(S, X)$ depends on X alone or on S and X , it is possible to develop "weighted" E-estimators that will be more efficient than the E-estimators $\hat{\beta}_E$ or $\hat{\beta}^*$ (Chamberlain, 1987).

Suppose next that the outcome of interest is a dichotomous disease variable. Then Y_i will be a Bernoulli random variable. In that case, one might no longer wish to specify the semiparametric model (2), i.e.,

$$E[Y_i | X_i, S_i] = h(X_i) + \beta S_i,$$

since the model does not naturally obey the restriction that probabilities must lie in the interval $[0, 1]$. Therefore one might specify a semiparametric logistic model

$$E[Y_i | X_i, S_i] = \frac{\exp[h(X_i) + \beta S_i]}{1 + \exp[h(X_i) + \beta S_i]} \tag{14}$$

Unfortunately, the approach developed in this paper will not allow us to consistently estimate the β of equation (14) even though Bickel et al. (1992) and Chamberlain (discussion paper cited previously) show that, in principle, there should exist an $n^{1/2}$ -consistent estimator of β based on data (X_i, S_i, Y_i) [at least when the dimension of X_i is fixed as the sample size increases]. Our approach fails because it is fundamentally based on the fact that, for model (2), β is identified from the "pseudo-data" (S_i, V_i, Y_i) , where $V_i = E(S | X_i)$. We call V_i "pseudo-data." For example, if V_i were known, our estimator $\hat{\beta}_E$ does not require data on X_i . It can be shown that β in equation (14) is not identified from pseudo-data (S_i, V_i, Y_i) due to the "noncollapsibility" of the logistic parameter β when we collapse from the "raw

data" X_i to V_i . Indeed, Gail, Wieand, and Piantadosi (1984) essentially prove this nonidentifiability result in the special case for which $V_i = \frac{1}{2}$ for all subjects. In fact, suppose X_i were dichotomous and thus e^β was the *common* exposure (S)–disease (Y) odds ratio in the two 2×2 tables indexed by the levels of X . In this special case, the nonidentifiability of β when V_i is a fixed constant for all subjects i is simply a restatement of the following well-known fact. Even when S and X are (marginally) independent, the common odds ratio e^β is not identified from data (S_i, Y_i) since the marginal exposure–disease odds ratio (ignoring X) may differ from e^β and the magnitude of the difference depends on the distribution of X (Gail et al., 1984). However, in contrast to our nonidentifiability results for the β of model (14), if equation (4b) holds, the average causal effect of S on disease Y , i.e., $E[Y_{S=1}] - E[Y_{S=0}]$, is identified from (S_i, V_i, Y_i) (Rosenbaum and Rubin, 1983).

Suppose next that Y_i has a Poisson or overdispersed Poisson distribution. We might then wish to specify semiparametric log-linear models, e.g.,

$$E[Y_i | X_i, S_i] = \exp[h(X_i) + \beta S_i]. \quad (15)$$

For log-linear models, a simple modification of our approach can be used to consistently estimate β from pseudo-data (S_i, V_i, Y_i) . Specifically, since, under model (15), $E[U(\beta)] = 0$, where

$$U(\beta^\dagger) \equiv \sum_{i=1}^n Y_i e^{-\beta^\dagger S_i} (S_i - E[S | X_i]), \quad (16)$$

the solution $\hat{\beta}_E$ to $U(\beta^\dagger) = 0$ will be consistent, asymptotically normal. A feasible consistent estimator $\hat{\beta}_E$ can be obtained from data (S_i, X_i, Y_i) by specifying a (correct) model for $E[S | X_i]$.

The methods of E-estimation can be extended to estimate the causal effect of a time-varying treatment. Specifically, Robins (1989a, 1992a, 1992b, 1992c, 1992d) and Robins et al. (1992) use an extension of E-estimation, which they call G-estimation, to estimate, from observational data, the causal effect of a time-varying treatment both on a survival time outcome and on the evolution of the mean of a continuous outcome variable measured repeatedly over time in the presence of time-dependent confounding factors. Robins (1989a, 1992b, 1992d) uses G-estimation to correct for noncompliance in randomized trials studying the effect of a time-varying treatment both on survival time outcomes and on the evolution of the mean of a continuous outcome variable when noncompliance depends on time-dependent prognostic factors. G-estimation is of particular importance in estimating the causal effect of a time-varying treatment in the presence of time-varying prognostic factors because standard covariance adjustment based on time-dependent Cox proportional hazard models for survival time outcomes or generalized estimating equations (Liang and Zeger, 1986) for repeated measures outcomes cannot consistently estimate the treatment effect (Robins, 1986, 1989a, 1989b, 1992a, 1992b, 1992c).

ACKNOWLEDGEMENTS

This work was supported in part by National Institutes of Health Grants 2 P30 ES00002, R01-ES03405, K04-ES00180, ES01108, and CA09001. We would like to thank Doug Dockery, Frank Speizer, Benjamin Ferris, and other contributors to the Harvard Six Cities Study for their generous sharing of time and data.

RÉSUMÉ

Pour estimer l'influence d'un ou plusieurs facteurs sur une variable d'intérêt, il faut prendre en compte les effets des covariables qui d'une part varient avec les dits facteurs, et d'autre part aident à

prédire la variable d'intérêt, indépendamment de ces facteurs. Dans cet article, nous présentons des méthodes de régression qui, à la différence des méthodes usuelles, ajustent l'effet confondant de plusieurs covariables (continues ou discrètes) par modélisation de l'espérance conditionnelle des différents facteurs en fonction des covariables. Dans le cas particulier d'un seul facteur à deux niveaux, cette espérance conditionnelle est identique à ce que Rosenbaum et Rubin ont appelé le score de propension. Ces auteurs, d'ailleurs, ont aussi proposé des méthodes d'estimation passant par la modélisation de ce score de propension. Nos méthodes généralisent celles de Rosenbaum et Rubin de plusieurs manières. Tout d'abord, notre approche s'étend d'emblée à tous les cas de figure possibles pour les facteurs, chacun d'entre eux pouvant être continu, ordinal ou discret. Ensuite, même dans le cas d'un seul facteur à deux niveaux, notre approche ne nécessite pas de classification ou d'appariement d'après le score de propension, de telle sorte que le risque de "confusion résiduelle" (c'est-à-dire de biais) lié à ces méthodes est évité. Enfin, notre approche permet de conforter l'idée qu'il vaut mieux utiliser le score de propension estimé que le vrai score de propension, même lorsque ce vrai score est connu. Le surcroît de puissance de notre approche provient du fait que nous supposons que l'influence des facteurs peut être décrite par la composante paramétrique d'un modèle de régression semi-paramétrique. A titre d'illustration, nous réanalysons, sur une cohorte de 2,713 adultes blancs de sexe masculin, l'effet du tabac sur la valeur du volume expiratoire maximal seconde, et nous comparons les résultats obtenus avec ceux des méthodes classiques.

REFERENCES

- Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Annals of Statistics* **11**, 432–452.
- Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1992). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore, Maryland: Johns Hopkins University Press.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305–334.
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Dockery, D. W., Speizer, F. E., Ferris, B. G., Ware, J. H., Louis, T. A., and Spiro, A. (1988). Cumulative and reversible effects of lifetime smoking on simple tests of lung function in adults. *American Review of Respiratory Diseases* **137**, 286–292.
- Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 657–687.
- Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431–444.
- Huber, P. (1981). *Robust Statistics*. New York: Wiley.
- Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* **79**, 61–71.
- Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear model. *Biometrika* **73**, 13–22.
- Manski, C. F. (1988). *Analog Estimation Methods in Econometrics*. New York: Chapman and Hall.
- Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.
- Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics* **10**, 475–478.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.
- Robins, J. M. (1989a). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman, and A. Mulley (eds), 113–159. Washington, DC: NCHSR, U.S. Public Health Service.
- Robins, J. M. (1989b). The control of confounding by intermediate variables. *Statistics in Medicine* **8**, 679–701.
- Robins, J. M. (1992a). Correcting for noncompliance in randomized trials using structural nested mean models. *Communications in Statistics*, in press.
- Robins, J. M. (1992b). Estimating the causal effect of a time-varying treatment on survival using a new class of failure time models. *Communications in Statistics*, in press.
- Robins, J. M. (1992c). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, in press.
- Robins, J. M. (1992d). Analytic methods for HIV treatment and cofactor effects. In *Methodological Issues of AIDS Behavioral Research*, D. G. Ostrow and R. Kessler (eds). New York: Plenum.

- Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia (PCP) on the survival of AIDS patients. *Epidemiology*, in press.
- Robins, J. M. and Morgenstern, H. (1987). The foundations of confounding in epidemiology. *Computers and Mathematics with Applications* **14**, 869–916.
- Robinson, P. (1988). Root- N -consistent semiparametric regression. *Econometrica* **56**, 931–954.
- Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* **79**, 565–574.
- Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 387–394.
- Rosenbaum, P. R. (1988). Permutation tests for matched pairs with adjustments for covariates. *Applied Statistics* **37**, 401–411.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.
- Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 33–38.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6**, 34–58.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Received December 1989; revised October 1990 and January 1991; accepted February 1991.

APPENDIX

In this Appendix, we prove the results stated in the text. We assume that

$$Y_i = f(S_i, X_i, \beta) + h(X_i) + \varepsilon_i, \quad E[\varepsilon_i | S_i, X_i] = 0, \quad (\text{A.0})$$

where $f(S_i, X_i, \beta)$ is a linear function of a V -dimensional parameter vector β that takes the value zero when $S_i = 0$. [Extension of our results to nonlinear functions of β is straightforward.] Model (2) in the text is the special case in which $f(S_i, X_i, \beta) = \beta S_i$, for univariate β and dichotomous S_i . (A.0) generalizes (2) by allowing for multivariate exposures, each component of which may be categorical, ordinal, or continuous. For example, we might suppose $S_i = (S_{i1}, \dots, S_{iM})$ and $f(S_i, X_i, \beta) = \sum_{m=1}^M \beta_m S_{mi} + \beta_{M+1} S_i X_i$ with $V = M + 1$. If equation (4b) holds when s is any value of S_i , then $f(S_i, X_i, \beta)$ is the average effect of joint exposure level S_i compared to the baseline level $S_i = 0$ among subjects with covariate level X_i . If $f(S_i, X_i, \beta)$ depends on X_i , we say there is an exposure-covariate interaction.

Define $f_\beta(S_i, X_i)$ to be the V -vector of partial derivatives of $f(S_i, X_i, \beta)$ with respect to the components of β and let

$$E[f_\beta(S_i, X_i) | X_i] = r(X_i; \alpha), \quad (\text{A.1})$$

where $r(\cdot; \cdot)$ is a known function and α is an unknown parameter. Define $R(S_i, X_i; \alpha) \equiv f_\beta(S_i, X_i) - r(X_i; \alpha)$. Note that $E[R(S_i, X_i; \alpha) | X_i] = 0$. If, as in the text, $f_\beta(S_i, X_i) = S_i$ is a Bernoulli random variable, (A.1) is a fully parametric model for S_i given X_i ; otherwise, (A.1) is a semiparametric model for the density $f_\beta(S_i | X_i)$, since the distribution of $R(S_i, X_i; \alpha)$ is completely unrestricted except for having mean zero given X_i .

Now for any nonrandom function $g(x)$, define

$$\begin{aligned} n^{-1/2} U(\beta^\dagger, g, \hat{\alpha}) &= n^{-1/2} \sum_i [Y_i - f(S_i, X_i, \beta^\dagger) - g(X_i)] R(S_i, X_i; \hat{\alpha}) \\ &\equiv n^{-1/2} \sum_i U_i(\beta^\dagger, g, \hat{\alpha}), \end{aligned} \quad (\text{A.2})$$

where $\hat{\alpha}$ is asymptotically equivalent to an $n^{1/2}$ -consistent solution to $0 = \sum_i M_i(\alpha^\dagger) \equiv \sum_i m(S_i, X_i, \alpha^\dagger)$ for some $M_i(\alpha^\dagger)$ satisfying $E[M_i(\alpha)] = 0$. That is, when $M_i(\alpha)$ is continuously differentiable, $n^{1/2}(\hat{\alpha} - \alpha) = -\{E[\partial M_i(\alpha)/\partial \alpha']\}^{-1} n^{-1/2} \sum_i M_i(\alpha) + o_p(1)$, and we say that $-\{E[\partial M_i(\alpha)/\partial \alpha']\}^{-1} M_i(\alpha)$ is the influence function of $\hat{\alpha}$. Chamberlain (1987) proves that $\hat{\alpha}$ is semiparametric efficient for α

under the sole restriction (A.1) on the conditional distribution of S_i given X_i only if $M_i(\alpha)$ equals

$$M_i^{eff}(\alpha) \equiv \frac{\partial r(X_i; \alpha)}{\partial \alpha} \{\text{var}[R(S_i, X_i; \alpha) | X_i]\}^{-1} R(S_i, X_i; \alpha).$$

That is, $\hat{\alpha}$ is semiparametric efficient only if it is asymptotically equivalent to the *optimal* weighted (possibly nonlinear) least squares estimator of α . Henceforth, we shall say that $\hat{\alpha}$ is semiparametric efficient under (A.1) if $\hat{\alpha}$ has influence function $-\{E[\partial M_i^{eff}(\alpha)/\partial \alpha']\}^{-1} M_i^{eff}(\alpha)$. If $f_{\beta}(S_i, X_i) = S_i$ is Bernoulli, semiparametric efficiency under (A.1) is just ordinary parametric efficiency. Our main result is given as Theorem A.1.

Theorem A.1 Under regularity conditions given in Corollary 1, Chapter 8 of Manski (1988), there exists a solution $\hat{\beta}_E(g) \equiv \hat{\beta}_E(g, \hat{\alpha})$ to $n^{-1/2}U(\hat{\beta}^\dagger, g, \hat{\alpha}) = 0$ such that $n^{1/2}(\hat{\beta}_E(g) - \beta)$ is asymptotically normal with mean 0 and variance that can be consistently estimated by

$$\hat{I}^{-1} \hat{\Delta}(g) (\hat{I}')^{-1}, \tag{A.3a}$$

where

$$\hat{I}' \equiv n^{-1} \sum_i \partial U_i(\beta, g, \hat{\alpha})' / \partial \beta = n^{-1} \sum_i f_{\beta}(S_i, X_i) R(S_i, X_i; \hat{\alpha})',$$

$$\hat{\Delta}(g) = n^{-1} \sum_i K_i(\hat{\beta}, g, \hat{\alpha}) K_i'(\hat{\beta}, g, \hat{\alpha}),$$

$$K_i(\hat{\beta}, g, \hat{\alpha}) = U_i(\hat{\beta}, g, \hat{\alpha}) - \hat{B}(g) \hat{C}^{-1} M_i(\hat{\alpha}),$$

$$\hat{\beta} \equiv \hat{\beta}_E(g),$$

$$\hat{B}(g) \equiv n^{-1} \sum_i \partial U_i(\hat{\beta}, g, \hat{\alpha}) / \partial \alpha' = n^{-1} \sum_i [Y_i - f(S_i, X_i, \hat{\beta}) - g(X_i)] \frac{\partial R(S_i, X_i; \hat{\alpha})}{\partial \alpha'},$$

$$\hat{C} = n^{-1} \sum_i \partial M_i(\hat{\alpha}) / \partial \alpha'.$$

If $\hat{\alpha}$ is semiparametric efficient under (A.1), the asymptotic variance of $n^{1/2}(\hat{\beta}_E(g) - \beta)$ can be consistently estimated by

$$\hat{I}^{-1} \hat{\Sigma}(g) (\hat{I}')^{-1} - \hat{Q}(g) \hat{\Omega} \hat{Q}'(g), \tag{A.3b}$$

where $\hat{Q}(g) \equiv \hat{I}^{-1} \hat{B}(g)$, $\hat{\Sigma}(g) = n^{-1} \sum_i U_i(\hat{\beta}_E(g), g, \hat{\alpha}) U_i'(\hat{\beta}_E(g), g, \hat{\alpha})'$, and $\hat{\Omega}$ is a consistent estimator of $\text{var}^\wedge[n^{1/2}(\hat{\alpha} - \alpha_0)]$.

Except when $f_{\beta}(S_i, X_i)$ equals a dichotomous S_i (as in the text), $\text{var}[R(S_i, X_i; \alpha) | X_i]$ may be an unknown function of α and X_i . Hence, if one chooses to estimate α by the unweighted (possibly nonlinear) least squares regression of $f_{\beta}(S_i, X_i)$ on X_i , it is necessary to use formula (A.3a) rather than (A.3b), since $\hat{\alpha}$ will then be efficient only if the (unknown) variance of $R(S_i, X_i; \alpha)$ does not depend on X_i .

However, if one has a correctly specified model $\text{var}[R(S_i, X_i; \alpha) | X_i] = \psi(X_i; \theta)$, where $\psi(X_i; \theta)$ is a known function and θ is an unknown parameter, then it is well known that the estimate $\hat{\alpha}$ that solves $0 = \sum_i \{\partial r(X_i; \alpha) / \partial \alpha\} \{\psi(X_i; \hat{\theta})\}^{-1} R(S_i, X_i; \alpha)$ has influence function $-\{E[\partial M_i^{eff}(\alpha) / \partial \alpha']\}^{-1} M_i^{eff}(\alpha)$ and (A.3b) can be used. Here $\hat{\theta}$ is the (possibly nonlinear) multivariate least squares regression estimate of θ obtained by regressing $R(S_i, X_i; \hat{\alpha}) R'(S_i, X_i; \hat{\alpha})$ on X_i , where $\hat{\alpha}$ is obtained from a preliminary unweighted least squares regression of $f_{\beta}(S_i, X_i)$ on X_i .

Application of Theorem A.1 Consider equation (9) in the text. In that setting, $g(X_i) \equiv 0$; $f(S_i, X_i, \beta) = \beta S_i$; $f_{\beta}(S_i, X_i) = S_i$; $R(S_i, X_i; \hat{\alpha}) = S_i - \hat{p}_i$, where $\hat{p}_i = e^{\alpha' X_i} / (1 + e^{\alpha' X_i})$;

$$\frac{\partial R(S_i, X_i; \hat{\alpha})}{\partial \alpha'} = \frac{\partial}{\partial (\alpha')} [e^{\alpha' X_i} / (1 + e^{\alpha' X_i})] |_{\alpha' = \hat{\alpha}} = \hat{p}_i (1 - \hat{p}_i) X_i';$$

$Y_i - f(S_i, X_i, \hat{\beta}_E(g)) - g(X_i) = \tilde{z}_i$; $U_i(\hat{\beta}_E(g), g, \hat{\alpha}) = \tilde{z}_i (S_i - \hat{p}_i)$; $\hat{I} = n^{-1} \sum_i S_i (S_i - \hat{p}_i)$; $\hat{\Sigma}(g) = n^{-1} \sum_i \tilde{z}_i^2 (S_i - \hat{p}_i)^2$; $\hat{Q}(g) = \sum_i \tilde{z}_i \hat{p}_i (1 - \hat{p}_i) X_i' / \sum_i S_i (S_i - \hat{p}_i)$. Substituting into equation (A.3b), we obtain equation (9).

The reader can check that substituting in (A.3b) also gives equation (13) if we set $g(X_i) = \hat{\beta}_1 + \sum_{k=2}^K \hat{\beta}_k X_{k,i}$ above. [As we shall see in Theorem (A.3) below, the fact that $g(X_i)$ is based on estimates $\hat{\beta}_k$ does not affect the asymptotic variance (A.3b).]

Proof of Theorem A.1 For pedagogic purposes we first sketch a proof. We then show how Corollary 1 of Manski's Chapter 8 can be used to formally prove the theorem. Since

$$E[U_i(\beta, g, \alpha)] = 0 \quad (\text{A.4})$$

by (A.0), we have that, under the regularity conditions discussed below, a Taylor expansion and the weak law of large numbers (WLLN) gives

$$0 = n^{-1/2}U(\hat{\beta}_E(g), g, \hat{\alpha}) = n^{-1/2}U(g) + I[n^{1/2}(\hat{\beta}_E(g) - \beta)] + B(g)[n^{1/2}(\hat{\alpha} - \alpha)] + o_p(1),$$

where $n^{-1/2}U(g) \equiv n^{-1/2}\sum U_i(g) \equiv n^{-1/2}U(\beta, g, \alpha)$, $I \equiv E[\partial U_i(\beta, g, \alpha)/\partial \beta']$, which does not depend on g .

$$B(g) \equiv E\left(\frac{\partial U_i(\beta, g, \alpha)}{\partial \alpha'}\right).$$

Thus

$$n^{1/2}(\hat{\beta}_E(g) - \beta) = -I^{-1}[B(g)n^{1/2}(\hat{\alpha} - \alpha) + n^{-1/2}U(g)] + o_p(1). \quad (\text{A.5})$$

By assumption, $n^{1/2}(\hat{\alpha} - \alpha) = -C^{-1}n^{-1/2}\sum_i M_i + o_p(1)$, where $C = E[\partial M_i(\alpha)/\partial \alpha]$, $M_i \equiv M_i(\alpha)$. Hence $n^{1/2}(\hat{\beta}_E(g) - \beta) = -I^{-1}n^{-1/2}\sum_i [U_i(g) - B(g)CM_i] + o_p(1)$. Thus $n^{1/2}(\hat{\beta}_E(g) - \beta)$ is asymptotically normal with mean zero and variance $I^{-1}\Delta(g)I^{-1}$, where $\Delta(g) = \text{var}[U_i(g) - B(g)C^{-1}M_i]$ since $n^{1/2}(\hat{\beta}_E(g) - \beta)$ is a sum of independent mean-zero random variables plus a term of $o_p(1)$. Formula (A.3a) follows by the WLLN.

We next establish (A.3b) for a semiparametric efficient $\hat{\alpha}$ under (A.1) using arguments similar to those in Pierce (1982) and Newey (1990). Let $L_i(\alpha^\dagger, \eta^\dagger) \equiv f(S_i|X_i; \alpha^\dagger, \eta^\dagger)$ be any (regular) parametric submodel with true values α, η for the density of S_i given X_i consistent with the restriction (A.1). Let $S_{a,i} = \partial \ln L_i(\alpha, \eta)/\partial \alpha^\dagger$. Let

$$\tau = \{a(S_i, X_i); a(S_i, X_i) = \partial \ln L_i(\alpha, \eta)/\partial \eta^\dagger \text{ for some parametric submodel}\}.$$

Note $\tau = \{a(S_i, X_i); E[a(S_i, X_i)|X_i] = 0 \text{ and } E[R(S_i, X_i; \alpha)a(S_i, X_i)'|X_i] = 0\}$ since the scores $a(S_i, X_i)$ are restricted only by having a conditional mean of zero and by being conditionally uncorrelated with $R(S_i, X_i; \alpha)$. It follows from Chamberlain (1987), Begun et al. (1983), and Newey (1990) that (a) $S_{a,i} - M_i^{\text{eff}} \in \tau$ and $E[M_i^{\text{eff}}a(S_i, X_i)'] = 0$ for all $a(S_i, X_i) \in \tau$ and (b) $\text{var}^A[n^{1/2}(\hat{\alpha} - \alpha)] = \{E[M_i^{\text{eff}}(M_i^{\text{eff}})']\}^{-1}$. M_i^{eff} is called the efficient score in the semiparametric model (A.1) for the law of S_i given X_i .

Now by differentiating the identity $E_{\beta, \alpha^\dagger, \eta^\dagger}[U_i(\beta, g, \alpha^\dagger)] = 0$ with respect to α^\dagger using the chain rule and evaluating at the true values (α, η) , we obtain $B(g) = -E[U_i(g)S_{a,i}]$, where $E_{\beta, \alpha^\dagger, \eta^\dagger}$ refers to expectation with respect to a density that differs from the truth only in that the law of S_i given X_i is $f(S_i|X_i; \alpha^\dagger, \eta^\dagger)$. Similarly differentiating this identity with respect to η^\dagger , we obtain $E[U_i(g)a'(S_i, X_i)] = 0$ for all $a(S_i, X_i) \in \tau$. Thus, by (a) in the last paragraph, we conclude $B(g) = -E[U_i(g)(M_i^{\text{eff}})']$. Similarly, the identity $E_{\beta, \alpha^\dagger, \eta^\dagger}[M_i(\alpha^\dagger)] = 0$ implies $C = -E[M_i(M_i^{\text{eff}})']$. Hence $K_i(g) \equiv U_i(g) - B(g)C^{-1}M_i = U_i(g) - E[U_i(g)(M_i^{\text{eff}})']\{E[M_i(M_i^{\text{eff}})']\}^{-1}M_i$. In the special case in which $M_i = M_i^{\text{eff}}$, $K_i(g)$ is the residual from the (population) least squares regression of $U_i(g)$ on M_i^{eff} , and a standard calculation gives $\text{var}[K_i(g)] = \text{var}[U_i(g)] - B(g)C^{-1}B'(g)$. (A.3.b) then follows by (b) in the last paragraph and the WLLN.

Theorem A.1 is formally proved by noting that it is an immediate consequence of Corollary 1 in Manski's Chapter 8 and the above variance calculations when we set Manski's function $g(z, b)$ equal to $(U_i(\beta^\dagger, g, \alpha^\dagger)', M_i(\alpha^\dagger)')$ and Manski's function $r(x)$ equal to $x'x$, where x is a vector.

Corollary A.1 If $\hat{\alpha}^{(j)}$ is semiparametric efficient under the j th of J nested correctly specified models $E[f_\beta(S_i, X_i)|X_i] = r(X_i; \alpha^{(j)})$, ($j = 1, \dots, J$), with the dimension of $\alpha^{(j)}$ increasing with j , then the asymptotic variance of $\hat{\beta}_E^{(j)}(g) \equiv \hat{\beta}_E(g, \hat{\alpha}^{(j)})$ is nonincreasing with j .

Proof Correct specification implies that, for $j > j^*$, M_i^{eff, j^*} is the first j^* components of $M_i^{\text{eff}, j}$, the efficient score for the j th model. But, by standard least squares theory, the variance of the residual $K_i^{(j)}(g)$ based on the j th model must be less than or equal to that based on model j^* .

The following theorem will be used in proving the claims made in the paragraph following equation (13).

Theorem A.2

- (a) $\text{var}^A[n^{1/2}(\hat{\beta}_E(g) - \beta)] \geq \text{var}^A[n^{1/2}(\hat{\beta}_E(h) - \beta)]$.
 (b) $\text{var}^A[n^{1/2}(\hat{\beta}_E(h) - \beta)] = \text{var}^A[n^{1/2}(\hat{\beta}_E(h) - \beta)]$, where $\hat{\beta}_E(h) \equiv \hat{\beta}_E(h, \alpha)$ and $\hat{\beta}_E(h) \equiv \hat{\beta}_E(h, \hat{\alpha})$.

Proof of (a) (a) is an immediate consequence of the following two lemmas.

Lemma A.1 The function g minimizing $\text{var}^\wedge[n^{1/2}(\hat{\beta}_E(g) - \beta)]$ also minimizes $\text{var}^\wedge[n^{-1/2}U(\beta, g, \hat{\alpha})]$.

Proof By a Taylor expansion, we have

$$0 = n^{-1/2}U(\hat{\beta}_E(g), g, \hat{\alpha}) = n^{-1/2}U(\beta, g, \hat{\alpha}) + n^{-1} \frac{\partial U(\beta, g, \hat{\alpha})}{\partial \beta'} [n^{1/2}(\hat{\beta}_E(g) - \beta)] + o_p(1). \quad (\text{A.6})$$

A further Taylor expansion of $n^{-1}\partial U(\beta, g, \hat{\alpha})/\partial \beta'$ around α_0 and the WLLN proves $n^{-1}\partial U(\beta, g, \hat{\alpha})/\partial \beta' = I + o_p(1)$, proving the lemma.

Lemma A.2 The function h minimizes $\text{var}^\wedge[n^{-1/2}U(\beta, g, \hat{\alpha})]$.

Proof $n^{-1/2}U(\beta, g, \hat{\alpha}) = n^{-1/2}\sum_i \varepsilon_i R(S_i, X_i; \hat{\alpha}) + n^{-1/2}\sum_i [h(X_i) - g(X_i)]R(S_i, X_i; \hat{\alpha}) \equiv A_1 + A_2(g)$, say where we have used (A.0) to substitute $\varepsilon_i + h(X_i)$ for $Y_i - f(S_i, X_i, \beta)$. If we can show $\text{cov}^\wedge(A_1, A_2(g)) = 0$, then $\text{var}^\wedge[n^{-1/2}U(\beta, g, \hat{\alpha})] = \text{var}^\wedge(A_1) + \text{var}^\wedge[A_2(g)]$, which is minimized at $g = h$ since $\text{var}^\wedge[A_2(h)] = 0$. Now A_1 and $A_2(g)$ have zero covariance since (a) $E[A_1 | (S, X)] = 0$ and (b) $A_2(g)$ is fixed given $(S, X) \equiv \{(S_i, X_i); i = 1, \dots, n\}$. (a) and (b) follow from the fact that $E[\varepsilon_i | (S, X)] = 0$ and $\hat{\alpha}$ depends on the data only through (S, X) .

Proof of (b) (b) follows from the fact that $B(h) = 0$ by (A.0).

In general, we do not know $h(X_i)$. Therefore, as in Section 4, we shall hypothesize a model $h(X_i) = g(X_i; \theta)$ where $g(\cdot, \cdot)$ is a known function and θ is a vector of parameters to be estimated. We estimate θ by (possibly nonlinear) least squares regression of $Y_i - f(S_i, X_i, \hat{\beta}_E)$ on X_i , where $\hat{\beta}_E$ is $\hat{\beta}_E(g)$ for $g(X_i) \equiv 0$. Let $\hat{\theta}$ be the (possibly nonlinear) least squares estimator of θ . It is clear that, since $\hat{\beta}_E$ is an $n^{1/2}$ -consistent estimator of β , if the model for $h(X_i)$ were correctly specified, $n^{1/2}(\hat{\theta} - \theta)$ would have a nondegenerate limiting distribution with mean 0. If the model for $h(X_i)$ were misspecified, there still exists θ^* such that $n^{1/2}(\hat{\theta} - \theta^*)$ has a nondegenerate limiting distribution with mean 0. The following theorem shows that we can then use $\hat{\theta}$ to construct an adaptive estimator of β that (1) has the same limiting distribution as $\hat{\beta}_E(h)$ if our model $h(X_i)$ is correctly specified and (2) remains consistent, asymptotically normal even if our model is misspecified.

Theorem A.3 If $n^{1/2}(\hat{\theta} - \theta^*)$ has a nondegenerate limiting distribution with mean 0, then $\hat{\beta}_E[g(X_i, \hat{\theta})]$ has the same limiting distribution as $\hat{\beta}_E[g(X_i, \theta^*)]$. In particular, it will be consistent and asymptotically normal whether or not the hypothesized model for $h(X_i)$ is correct, and it will have the same limiting distribution as $\hat{\beta}_E(h)$ if the model for $h(X_i)$ is correct.

Proof For notational convenience, assume that θ is one-dimensional. It will be sufficient to show that

$$n^{-1/2}U(\beta^\dagger, \hat{\alpha}, g(\hat{\theta})) = n^{-1/2}U(\beta^\dagger, \hat{\alpha}, g(\theta^*)) + o_p(1) \quad (\text{A.7})$$

for $|\beta^\dagger - \beta| = O(n^{-1/2})$. By a Taylor series expansion

$$n^{-1/2}U(\beta^\dagger, \hat{\alpha}, g(\hat{\theta})) = n^{-1/2}U(\beta^\dagger, \hat{\alpha}, g(\theta^*)) + n^{1/2}(\hat{\theta} - \theta^*)[n^{-1}U[\beta^\dagger, \hat{\alpha}, g'(\theta^*)]] \quad (\text{A.8})$$

$$+ n^{1/2}(\hat{\theta} - \theta^*)^2[n^{-1}U[\beta^\dagger, \hat{\alpha}, g''(\hat{\theta}^*)]] \quad (\text{A.9})$$

for some $\hat{\theta}^*$ between $\hat{\theta}$ and θ^* . Now, if $\beta^\dagger = \beta$, by Theorem A.1 and Pierce (1982), $[n^{-1}U[\beta^\dagger, \hat{\alpha}, g'(\theta^*)]]$ converges to 0 in probability since it has mean 0 to $o_p(n^{-1/2})$ with variance converging to 0 as $n \rightarrow \infty$. Further, under regularity conditions, this remains true if $|\beta^\dagger - \beta| = O(n^{-1/2})$. It then follows from Slutsky's theorem that expression (A.8) converges in law to 0 and thus in probability to 0. Further, since $n^{-1}U[\beta^\dagger, \hat{\alpha}, g''(\hat{\theta}^*)]$ is at most $O_p(1)$ and $n^{1/2}(\hat{\theta} - \theta^*)^2$ is $O_p(n^{-1/2})$, it follows that expression (A.9) is $O_p(n^{-1/2})$. Thus equation (A.7) is true.

Theorem (A.3) and part (b) of Theorem (A.2) imply proposition (1) in the paragraph following equation (13). Proposition (2) is an easy calculation.