Published in final edited form as:

J Am Stat Assoc. 2008 September 1; 103(483): 1270–1280. doi:10.1198/016214508000000625.

Empirical Likelihood-Based Estimation of the Treatment Effect in a Pretest–Posttest Study

Chiung-Yu Huang, Jing Qin, and Dean A. Follmann

Mathematical Statisticians, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892.

Chiung-Yu Huang: huangchi@niaid.nih.gov; Jing Qin: jingqin@niaid.nih.gov; Dean A. Follmann: dfollmann@niaid.nih.gov

Abstract

The pretest–posttest study design is commonly used in medical and social science research to assess the effect of a treatment or an intervention. Recently, interest has been rising in developing inference procedures that improve efficiency while relaxing assumptions used in the pretest–posttest data analysis, especially when the posttest measurement might be missing. In this article we propose a semiparametric estimation procedure based on empirical likelihood (EL) that incorporates the common baseline covariate information to improve efficiency. The proposed method also yields an asymptotically unbiased estimate of the response distribution. Thus functions of the response distribution, such as the median, can be estimated straightforwardly, and the EL method can provide a more appealing estimate of the treatment effect for skewed data. We show that, compared with existing methods, the proposed EL estimator has appealing theoretical properties, especially when the working model for the underlying relationship between the pretest and posttest measurements is misspecified. A series of simulation studies demonstrates that the EL-based estimator outperforms its competitors when the working model is misspecified and the data are missing at random. We illustrate the methods by analyzing data from an AIDS clinical trial (ACTG 175).

Keywords

Auxiliary information; Biased sampling; Causal inference; Observational study; Survey sampling

1. INTRODUCTION

The pretest–posttest randomized trial is a popular study design in medical and social science research, used to assess the effect of a treatment or an intervention. Subjects are assessed before and after a randomly assigned treatment intervention. The treatment effect is evaluated by comparing the change in outcome between groups. The analysis of pretest–posttest data has been studied by many researchers, including Brogan and Kutner (1980), Laird (1983), Stanek (1988), Follmann (1991), Yang and Tsiatis (2001), and Bonate (2003). Methods based on direct comparisons, such as the two-sample *t*-test comparing the posttest

© 2008 American Statistical Association

Publisher's Disclaimer: Full terms and conditions of use: <http://amstat.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

measures and the paired t -test comparing the change between pretest and posttest measures, are commonly used to analyze such data. These simple methods do not fully use the available baseline information, however, and thus are not efficient if the baseline covariates are associated with the posttest measure. To improve efficiency, analysis of covariance (ANCOVA) incorporates the pretest measure and other baseline variables in a regression model (Fleiss 1986). Recently, Leon, Tsiatis, and Davidian (2003) reformulated the problem by adopting the framework of potential outcomes in casual inference, where the person randomized to treatment (control) has a counterfactual outcome under control (treatment). By applying the theory of estimation in semiparametric models developed by Robins, Rotnitzky, and Zhao (1994) for missing data, the authors constructed a class of consistent semiparametric estimators based on the observed data and identified the most efficient estimator in the class.

In this article we consider scenarios in which the posttest measures might be missing and the probability of missing depends on the observed covariate information. In this case the usual statistical methods based on comparing complete cases, such as the t -test, paired t -test, and ANCOVA, yield invalid inferences when subjects with complete observation are not representative of the randomized study population. To correct for this bias, various statistical methods, including the estimators of Horvitz–Thompson (1952) and Hájek (1981), have been proposed. These estimators weight the complete cases by the probability of being observed. Davidian, Tsiatis, and Leon (2005) studied a class of consistent semiparametric estimators and identified the most efficient by applying the results of Robins et al. (1994). But constructing of the most efficient semiparametric estimator relies on the correct specification of the underlying relationship between the posttest measures and other covariate information. In practice, the exact relationship is unknown, and thus the semiparametric estimator may not be optimal. As an alternative, we consider estimators constructed based on the empirical likelihood (EL).

The EL was introduced by Owen (1988, 2001) for constructing confidence intervals for the mean or other functions of the distribution function. Suppose that the observed data $\{y_1, \dots, y_n\}$ are realizations of independent and identically distributed (iid) random variables with a common distribution function $F(y)$. It is well known that the empirical distribution function $F_n(y) = n^{-1} \sum_{i=1}^n I(y_i \leq y)$ maximizes the nonparametric likelihood function $\prod_{i=1}^n p_i$ subject to the constraints $\sum_{i=1}^n p_i = 1$, $p_i \geq 0$, where $p_i = dF(y_i)$ is the jump size of F at value y_i . To construct confidence intervals for the population mean, $\mu = \int y dF(y)$, Owen (1988) proposed maximizing the nonparametric likelihood with an additional constraint, $\sum_{i=1}^n y_i p_i = \mu$. Analogous to the likelihood ratio statistic under a parametric model, Owen showed that the EL ratio,

$$2 \left(\max_{p_i \geq 0} \left\{ \sum_{i=1}^n \log p_i : \sum_{i=1}^n p_i = 1, p_i \geq 0 \right\} - \max_{p_i \geq 0} \left\{ \sum_{i=1}^n \log p_i : \sum_{i=1}^n p_i = 1, \sum_{i=1}^n y_i p_i = \mu, p_i \geq 0 \right\} \right),$$

has a limiting chi-squared distribution and can be used to construct confidence intervals for μ . The EL method has been applied to more general settings by others. For example, Qin and Lawless (1994) made connections to estimating equations and demonstrated that the EL method can yield the most efficient estimator by imposing proper constraints. For survey sampling, Chen and Qin (1993), Chen, Sitter, and Wu (2002), Chen, and Wu (2002), and Wu and Sitter (2001) applied the EL method to incorporate auxiliary covariate information to improve efficiency of estimation. For situations where the sampling scheme may be outcome-dependent, Wang and Zhou (2006) proposed a semiparametric EL procedure to

estimate regression parameters. In situations where the outcome is subject to censoring, Zhou and Liang (2005) considered a semiparametric EL-based estimator for a two-sample comparison. Recently, attention has been focused on applying the EL method to general missing-data problems. Wang and Rao (2002) considered EL-based inference with kernel regression imputation, whereas Wang, Linton, and Härdle (2004) applied their method with imputation under a partially linear regression model and Liang, Wang, and Carroll (2007) extended the procedure to allow for error-prone covariates. To estimate parameters in estimating equations, Wang and Chen (2006) proposed an EL method with multiple imputation for missing data. Finally, Qin and Zhang (2007) proposed maximizing a conditional likelihood based only on complete cases with proper constraints that adjust for the bias due to missingness. They did not consider counterfactual missing data as we do.

The present article applies the EL method to the pretest–posttest setting with missing data. We begin by presenting an EL estimator for a pretest–posttest study without missing data. In Section 2 we show that the influence function of this estimator achieves the semiparametric efficiency bound when the true regression function linking the posttest measures to baseline covariates is used to construct the estimator. We also show that the proposed estimator is more efficient than the estimator studied by Leon et al. (2003) when the conjectured function is not the true regression function. In Section 3 we present an EL estimator for situations where the posttest measure is missing at random (MAR) and additional posttreatment covariate information is available. As before, we show that the proposed semiparametric estimator is most efficient when the true regression function is used and is more efficient than that studied by Davidian et al. (2005) under a misspecified working model. Similar techniques were adopted by Tan (2006) in the setting of causal inference, where the counterfactual outcome was missing due to treatment assignment and the measured covariates were not changed by treatment. But the results of Tan (2006) cannot be applied directly to the setting considered in this article, particularly in the case of dropout or the presence of intermediate covariates, whose values may well predict final outcome. In Section 5 we compare the performance of the EL estimator and its competitors through a series of simulations. For illustration, we apply these methods to a data set from AIDS Clinical Trials Group (ACTG) protocol: 175 (Hammer et al. 1996).

2. PRETEST–POSTTEST STUDIES WITHOUT MISSING DATA

Let n denote the number of patients randomized in a pretest–posttest study. Assume that each patient is randomized to either treatment ($R = 1$) or control ($R = 0$) with known probability δ or $1 - \delta$. We use Z to denote the pretest response, but Z can be generalized to include other baseline covariates as well. For the responses, we adopt the potential outcomes framework. Let Y_1 and Y_0 be the posttest measures that a subject potentially would exhibit if assigned to treatment and control. The bivariate outcome (Y_1, Y_0) is never observed simultaneously for any subject. The observed data are $(R = 1, Z, Y_1)$ if the subject is assigned to the treatment group and $(R = 0, Z, Y_0)$ if the subject is assigned to the control group. Because of randomization, the marginal distributions of the baseline Z in the two groups are identical. Let $\mu_1 = E[Y_1]$ and $\mu_0 = E[Y_0]$. Our main interest lies in estimating the treatment effect defined by $\Delta = \mu_1 - \mu_0$. For simplicity, we consider only the case of estimating μ_1 .

Let i denote subjects, $i = 1, \dots, n$. For convenience, we assume that the first n_1 subjects are assigned to the treatment group and the last $n_0 = n - n_1$ subjects are assigned to the control group. We denote the observed data for the first n_1 subjects by $(R_i = 1, z_i, y_{1i})$, $i = 1, \dots, n_1$, and that for the last n_0 subjects by $(R_i = 0, z_i, y_{0i})$, $i = n_1 + 1, \dots, n$.

For situations in which data are MAR, Robins et al. (1994) characterized the form of the influence function for all regular asymptotic linear (RAL) estimators. Recall that an estimator is asymptotically linear if it is asymptotically equivalent to the average of some function (i.e., influence function) taken at the observations. An asymptotic linear estimator is consistent and asymptotically normal under some general conditions. Moreover, an estimator is regular in a semiparametric model if its convergence is locally uniform. This excludes superefficient estimators that have an asymptotic variance smaller than the Cramer–Rao lower bound. Robins et al. (1994) further identified the most efficient RAL estimator that gives the smallest asymptotic variance in semiparametric models when data are MAR. Applying their theory, Leon et al. (2003; henceforth LTD) showed that all RAL estimators of μ_1 based on observed data in a pretest–posttest study have an asymptotically iid representation with influence function

$$\frac{RY_1}{\delta} - \frac{R - \delta}{\delta} h_1(Z),$$

where h_1 is an arbitrary function such that $\text{var}\{h_1(Z)\} < \infty$. The most efficient RAL estimator that achieves the semiparametric efficiency bound has an influence function with $h_1(Z) = E[Y_1|Z]$. Thus the most efficient estimator for μ_1 among RAL can be expressed as

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{R_i Y_{1i}}{\delta} - \frac{R_i - \delta}{\delta} E[Y_1|Z=z_i] \right\} + o_p(n^{-1/2}). \quad (1)$$

The true functional form of $E[Y_1|Z]$ is usually unknown; thus in practice $E[Y_1|Z]$ is replaced by a conjectured function of Z , say $a(Z)$. For now, we assume that $a(Z)$ is prespecified. There is no guarantee that the LTD estimator, with $E[Y_1|Z]$ replaced by $a(Z)$, is most efficient. Another method might yield a more efficient estimator if $E[Y_1|Z]$ is misspecified.

As an alternative, we consider estimators based on the EL method. Our goal is to exploit the “auxiliary” information from the baseline covariates in both groups. To proceed, note that Y_1 is missing in the control group by design and that the likelihood based on observed data for the treatment group and the baseline covariate in the control group is

$$\mathcal{L} = \prod_{i=1}^{n_1} f(z_i, y_{1i}) \prod_{i=n_1+1}^n f(z_i),$$

where $f(z, y_1)$ is the joint density function of (Z, Y_1) and $f(z)$ is the marginal density function of Z , which is common to both groups. Let $p_i = f(z_i, y_{1i})$ for $i = 1, \dots, n_1$ and $r_i = f(z_i)$ for $i = n_1 + 1, \dots, n$. The log-likelihood function is given by

$$\ell = \sum_{i=1}^{n_1} \log p_i + \sum_{i=n_1+1}^n \log r_i. \quad (2)$$

Let $a(z)$ be a guess of $E[Y_1|Z=z]$ with $E\{a(Z)\}^2 < \infty$, and write $a = E\{a(Z)\}$, a guess of μ_1 . Following the EL method, we propose to maximize the log-likelihood function subject to the following constraints:

$$p_i \geq 0, r_i \geq 0, \sum_{i=1}^{n_1} p_i = 1, \sum_{i=n_1+1}^n r_i = 1, \quad (3)$$

and

$$\sum_{i=1}^{n_1} p_i a(z_i) = \sum_{i=n_1+1}^n r_i a(z_i) = a. \quad (4)$$

This latter constraint follows from the fact that the two groups have identical baseline covariate distributions due to randomization and thus $\int \int a(z) f(z, y_1) dy_1 dz = \int a(z) f(z) dz = a$.

For fixed a , we apply the Lagrange multiplier method to maximize (2) with constraints (3) and (4) and obtain the following expressions for the p_i 's and r_i 's:

$$p_i = \frac{1}{n_1} \times \frac{1}{1 + \lambda \{a(z_i) - a\}}, i=1, \dots, n_1, \quad (5)$$

and

$$r_i = \frac{1}{n_0} \times \frac{1}{1 + \tau \{a(z_i) - a\}}, i=n_1+1, \dots, n, \quad (6)$$

where λ and τ are Lagrange multipliers determined by

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{a(z_i) - a}{1 + \lambda \{a(z_i) - a\}} = 0$$

and

$$\frac{1}{n_0} \sum_{i=n_1+1}^n \frac{a(z_i) - a}{1 + \tau \{a(z_i) - a\}} = 0.$$

We then solve these equations to obtain estimates of a , λ , and τ . An EL estimator for μ_1 is given by $\hat{\mu}_{1EL} = \sum_{i=1}^{n_1} \hat{p}_i y_i$, where $\hat{p}_i$ is given by (5) with λ and a are replaced with their estimated values. The large-sample property of $\hat{\mu}_{1EL}$ is presented in Theorem 1, the proof of which is given in the Appendix.

Theorem 1. Suppose that $0 < \delta < 1$ and $E[Y_1^2] < \infty$. Assume that the random treatment assignment R has a binomial distribution with $\Pr(R = 1) = \delta$. Assume that $Z_1, \dots, Z_n$ are iid random variables from a nondegenerate distribution and that the function $a(z)$ satisfies $E\|a(Z)\|^3 < \infty$, where $\|\cdot\|$ is the Euclidean norm. The EL estimator $\hat{\mu}_{1EL}$ has the following asymptotic representation:

$$\hat{\mu}_{1EL} = \frac{1}{n} \sum_{i=1}^n \frac{R_i y_{1i}}{\delta} - E\{Y_1 \psi(Z)\}^T E\{\psi(Z) \psi(Z)^T\}^{-1} \times \left\{ \frac{1}{n} \sum_{i=1}^n \frac{R_i - \delta}{\delta} \psi(Z_i) \right\} + o_p(n^{-1/2}), \quad (7)$$

where $\psi(z_j) = (1, a(z_j))^T$. Moreover, we have $\sqrt{n}(\hat{\mu}_{1EL} - \mu_1) \rightarrow N(0, \sigma_1^2)$, where

$$\sigma_1^2 = \left\{ \frac{1}{\delta} E(Y_1^2) - \mu_1^2 \right\} - \frac{1 - \delta}{\delta} E\{Y_1 \psi(Z)\}^T \times [E\{\psi(Z) \psi(Z)^T\}]^{-1} E\{Y_1 \psi(Z)\}.$$

In the special case where $a(z)$ is the true regression function $E(Y|z)$, we have

$$E\{Y_1\psi(Z)^T\}[E\{\psi(Z)\psi(Z)^T\}]^{-1} = (0, 1).$$

It follows that

$$\hat{\mu}_{1EL} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{R_i Y_{1i}}{\delta} - \frac{R_i - \delta}{\delta} a(z_i) \right\} + o_p(n^{-1/2}).$$

Thus if $a(z) = E(Y_1|z)$, then the EL estimator and the LTD estimator have the same influence function, and both achieve the semiparametric efficiency lower bound. But if $a(z)$ is not the true regression function, then the two estimators are not asymptotically equivalent. Consider the following class of influence functions:

$$\left\{ \frac{RY_1}{\delta} - \rho_1 \frac{R - \delta}{\delta} - \rho_2 \frac{R - \delta}{\delta} a(Z); \rho_1 \in R, \rho_2 \in R \right\}.$$

The LTD estimator is a member of this class with $\rho_1 = 0$ and $\rho_2 = 1$. The optimal choice of (ρ_1, ρ_2) that minimizes the variance is given by

$$(\rho_1, \rho_2)^T = \text{cov} \left\{ \frac{RY_1}{\delta}, \frac{R - \delta}{\delta} \psi(Z) \right\} \times \text{var} \left\{ \frac{R - \delta}{\delta} \psi(Z) \right\}^{-1} = E\{Y_1\psi(Z)^T\} E\{\psi(Z)\psi(Z)^T\}^{-1},$$

which is the influence function of the EL estimator. Thus the proposed EL estimator is more efficient than the LTD estimator when the prespecified $a(z)$ is not $E(Y_1|z)$.

3. PRETEST–POSTTEST STUDY WITH MISSING DATA

In this section we consider pretest–posttest studies in which the posttest measure can be missing but all other variables are observed. We assume that additional covariate information X is collected after intervention but before follow-up. Such covariate information is commonly available in longitudinal studies in which X might be an intermediate measure of the response. Similar in the previous section, here we use counterfactual X_1 and X_0 to denote the covariate a subject would exhibit during the intervention period if assigned to treatment and control.

Let D_1 and D_0 denote the indicator of a nonmissing outcome if a subject is assigned to treatment and control; thus we only see one of the pair. For convenience, we assume that Y_1 is observed for the first n_{11} treatment subjects and missing for the next $n_{10} = n_1 - n_{11}$ treatment subjects. The control counts n_{01} and n_{00} are analogous. Thus in the treatment group we observe $(R_i = 1, D_{1i} = 1, z_i, x_{1i}, y_{1i})$ for $i = 1, \dots, n_{11}$ and $(R_i = 1, D_{1i} = 0, z_i, x_{1i})$ for $i = n_{11} + 1, \dots, n_1$. Similarly, in the control group we observe $(R_i = 0, D_{0i} = 1, z_i, x_{0i}, y_{0i})$ for $i = n_1 + 1, \dots, n_1 + n_{11}$ and $(R_i = 0, D_{0i} = 0, z_i, x_{0i})$ for $i = n_1 + n_{11} + 1, \dots, n$. The data structure of a pretest–posttest study is given in Table 1. We assume that among those randomized to treatment, Y_1 is MAR, with the probability of being observed given by

$$P(D_1 = 1 | R = 1, Z, X_1, Y_1) = P(D_1 = 1 | R = 1, Z, X_1) = w_1(Z, X_1),$$

where $w_1(z, x_1) = \varepsilon > 0$ is a known function. A similar definition is extended for subjects in the control group; thus missingness can be different for the two groups.

Davidian et al. (2005; henceforth DTL), showed that the RAL estimators of μ_1 in a pretest–posttest study with posttest measures MAR have an influence function of the form

$$\frac{RD_1 Y_1}{\delta w_1(Z, X_1)} - \frac{R - \delta}{\delta} h_1(Z) - \frac{R\{D_1 - w_1(Z, X_1)\}}{\delta w_1(Z, X_1)} g_1(Z, X_1),$$

where h_1 and g_1 are arbitrary functions such that $\text{var}\{h_1(Z)\} < \infty$ and $\text{var}\{g_1(Z, X_1)\} < \infty$. The popular Horvitz–Thompson (HT) estimator (1952) has the aforementioned influence function with $h_1(Z) = g_1(Z, X_1) = 0$. By applying the theory of Robins et al. (1994), DTL (2005) showed that the semiparametric estimator given by

$$\hat{\mu}_{1DTL} = \frac{1}{n} \sum_{i=1}^n \frac{R_i D_{1i} Y_{1i}}{\delta w_1(z_i, x_{1i})} - \frac{1}{n} \sum_{i=1}^n \frac{R_i - \delta}{\delta} h_1(z_i) - \frac{1}{n} \sum_{i=1}^n \frac{R_i \{D_{1i} - w_1(z_i, x_{1i})\}}{\delta w_1(z_i, x_{1i})} g_1(z_i, x_{1i})$$

is most efficient if $h_1(z) = E[Y_1|z]$, $g_1(z, x_1) = E[Y_1|z, x_1]$, and $w_1(Z, X_1)$ is correctly specified. Moreover, the DTL estimator has an appealing “double-robustness” property that $\hat{\mu}_{1DTL}$ is a consistent estimator if either (a) the regression functions $h_1 = E[Y_1|z]$ and $g_1 = E[Y_1|z, x_1]$ are both correctly specified and the missingness mechanism w_1 is misspecified, or (b) h_1 and g_1 are not correctly specified but w_1 is correctly specified. In practice, the true regression model is unknown, and another method might yield a more efficient estimator if the working regression models are misspecified.

For ease of notation, we use $f(z)$, $f(z, x_1)$, and $f(z, x_1, y_1)$ to denote the marginal distribution of Z and the joint distributions of (Z, X_1) and (Z, X_1, Y_1) . Note that in the control group, the counterfactual (D_1, X_1, Y_1) is missing but Z is observed. Because of randomization, the two treatment groups have identical baseline covariate distributions. As before, the efficiency of estimation of μ_1 should be improved by properly incorporating baseline covariate information from the control group. Let $p_i = f(z_i, x_{1i}, y_{1i})$, $q_i = f(z_i, x_{1i})$, and $r_i = f(z_i)$. The log-likelihood constructed based on observed data from treatment and the baseline covariate from control is given by

$$\ell = \sum_{i=1}^n \{R_i D_{1i} \log p_i + R_i (1 - D_{1i}) \log q_i + (1 - R_i) \log r_i\} + \sum_{i=1}^n [R_i D_{1i} \log \{w_1(z_i, x_{1i})\} + R_i (1 - D_{1i}) \log \{1 - w_1(z_i, x_{1i})\}].$$

Let $u(z)$ be a function of Z and let $v_1(z, x_1)$ be a function of (Z, X_1) with $E[u(Z)^2] < \infty$ and $E[v_1(Z, X_1)^2] < \infty$. Ideally, $u(z)$ will be a good guess for $E[Y_1|Z = z]$ and $v_1(z, x_1)$ will be a good guess for $E[Y_1|Z = z, X_1 = x_1]$. In practice, $u(z)$ or $v_1(z, x_1)$ can be vector functions, but we do not consider such a generality here. We propose to maximize the log-likelihood function subject to the following constraints:

$$p_i \geq 0, q_i \geq 0, r_i \geq 0, \sum_{i=1}^n R_i D_{1i} p_i = \sum_{i=1}^n R_i (1 - D_{1i}) q_i = \sum_{i=1}^n (1 - R_i) r_i = 1,$$

$$\sum_{i=1}^n R_i D_{1i} p_i u(z_i) = \sum_{i=1}^n R_i (1 - D_{1i}) q_i u(z_i) = \sum_{i=1}^n (1 - R_i) r_i u(z_i) = u, \quad (8)$$

$$\sum_{i=1}^n R_i D_{1i} p_i v_1(z_i, x_{1i}) = \sum_{i=1}^n R_i (1 - D_{1i}) q_i v_1(z_i, x_{1i}) = v, \quad (9)$$

and

$$\sum_{i=1}^n R_i D_{1i} p_i w_1(z_i, x_{1i}) = \sum_{i=1}^n R_i (1 - D_{1i}) q_i w_1(z_i, x_{1i}) = w, \quad (10)$$

The first set of constraints ensure that the p_i 's, q_i 's, and r_i 's give proper density functions. The constraints (8) follow from marginalization, $f(z, x_1) = \int f(z, x_1, y_1) dy_1$ and $f(z) = \int f(z, x_1) dx_1$; thus $\int \int \int u(z) f(z, x_1, y_1) dy_1 dx_1 dz = \int \int u(z) f(z, x_1) dx_1 dz = \int u(z) f(z) dz$. Similar reasoning provides (9) and (10). The constraints in (10) are given to correct the selection bias in the observed data.

For fixed (u, v, w) , we apply the Lagrange multiplier method and maximize over the p_i 's, q_i 's, and r_i 's to obtain

$$p_i = \frac{1}{n_{11}} \times \frac{R_i D_{1i}}{1 + \lambda_1 \{\mathbf{a}(z_i, x_{1i}) - \mathbf{a}\}}, \quad q_i = \frac{1}{n_{10}} \times \frac{R_i (1 - D_{1i})}{1 + \lambda_2 \{\mathbf{a}(z_i, x_{1i}) - \mathbf{a}\}},$$

and

$$r_i = \frac{1}{n_0} \times \frac{1 - R_i}{1 + \lambda_3 \{u(z_i) - u\}},$$

where $\mathbf{a}(z_i, x_{1i}) = (u(z_i), v_1(z_i, x_{1i}), w_1(z_i, x_{1i}))^T$, $\mathbf{a} = (u, v, w)^T$, and $\boldsymbol{\lambda}_1 = (\lambda_{11}, \lambda_{12}, \lambda_{13})$, $\boldsymbol{\lambda}_2 = (\lambda_{21}, \lambda_{22}, \lambda_{23})$, and λ_3 are Lagrange multipliers satisfying

$$\sum_{i=1}^n \frac{R_i D_{1i} \{\mathbf{a}(z_i, x_{1i}) - \mathbf{a}\}}{1 + \lambda_1 \{\mathbf{a}(z_i, x_{1i}) - \mathbf{a}\}} = \mathbf{0}, \quad \sum_{i=1}^n \frac{R_i (1 - D_{1i}) \{\mathbf{a}(z_i, x_{1i}) - \mathbf{a}\}}{1 + \lambda_2 \{\mathbf{a}(z_i, x_{1i}) - \mathbf{a}\}} = \mathbf{0},$$

and

$$\sum_{i=1}^n \frac{(1 - R_i) \{u(z_i) - u\}}{1 + \lambda_3 \{u(z_i) - u\}} = 0.$$

Let $\hat{p}_i$, $\hat{q}_i$, and $\hat{r}_i$ denote the solution to these equations. Because $\hat{p}_i$ is the jump size of the distribution function at (z_i, x_{1i}, y_{1i}) , μ_1 can be estimated by

$$\hat{\mu}_{1EL} = \sum_{i=1}^n R_i D_{1i} \hat{p}_i y_{1i}.$$

Moreover, functions of the response distribution, such as percentiles or truncated means, can be estimated directly by using $\hat{p}_i$'s. For example, the l th percentile of the response measure can be estimated by $\inf\{m: \sum_{i=1}^n R_i D_i \hat{p}_i I(y_{1i} \leq m) \geq l/100\}$, and thus the median response is estimated by

$$\inf\left\{m: \sum_{i=1}^n R_i D_i \hat{p}_i I(y_{1i} \leq m) \geq .5\right\}.$$

We begin our study of the large-sample properties of $\hat{\mu}_{1EL}$ by assuming that the probability of seeing a complete case, $w_1(z, x_1)$, is known. Let $\Psi(z, x_1) = (\Psi_1(z, x_1)^T, \Psi_2(z)^T)^T$, where $\Psi_1(z, x_1) = (1, u(z), v_1(z, x_1), w_1(z, x_1))^T$ and $\Psi_2(z) = (1, u(z))^T$. We state the asymptotic normality of $\hat{\mu}_{1EL}$ in Theorem 2. The proof of this theorem is given in a supplemental technical report, available at http://www.amstat.org/publications/jasa/supplemental_materials.

Theorem 2. Suppose that $0 < \delta < 1$ and $E[Y_1^2] < \infty$, and that there exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ so that $\epsilon_1 \leq w_1(z, x_1) \leq 1 - \epsilon_2$ for all (z, x_1) . Moreover, $0 < \Pr(R = 1, D_1 = 1) < 1$. Assume that the function $\Psi(z, x_1)$ satisfies $E\|1 - w_1(Z, X_1)\} w_1(Z, X_1)^{-1} \Psi(Z, X_1)\|^3 < \infty$, where $\|\cdot\|$ is the Euclidean norm. Then the EL estimator, $\hat{\mu}_{1EL}$, has the following asymptotic representation:

$$\hat{\mu}_{1EL} = \frac{1}{n} \sum_{i=1}^n \frac{R_i D_{1i} y_{1i}}{\delta w_1(z_i, x_{1i})} + \mathbf{C} \mathbf{A}^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{B}_i + o_p(n^{-1/2}), \quad (11)$$

where

$$\mathbf{B}_i = \begin{pmatrix} \frac{R_i \{D_{1i} - w_1(z_i, x_{1i})\}}{\delta w_1(z_i, x_{1i}) \{1 - w_1(z_i, x_{1i})\}} \Psi_1(z_i, x_{1i}) \\ \left\{ \frac{R_i D_{1i}}{\delta w_1(z_i, x_{1i})} - \frac{1 - R_i}{1 - \delta} \right\} \Psi_2(z_i) \end{pmatrix},$$

$\mathbf{A} = \text{var}(\mathbf{B}_i)$, and $\mathbf{C}$ is the covariance between $R_i D_{1i} y_{1i} / \{\delta w_1(z_i, x_{1i})\}$ and $\mathbf{B}_i$. Thus as $n \rightarrow \infty$, $\sqrt{n}(\hat{\mu}_{1EL} - \mu_1)$ converges weakly to a normal distribution with mean 0 and variance-covariance matrix $\sigma_1^2 = \sigma^2 - \mathbf{C} \mathbf{A}^{-1} \mathbf{C}^T$, where σ^2 is the variance of the HT estimator,

$$\hat{\mu}_{1HT} = n^{-1} \sum_{i=1}^n R_i D_{1i} y_{1i} / \{\delta w_1(z_i, x_{1i})\}.$$

It is easy to see that the results in Theorem 2 still hold when $u(z)$ and $v_1(z, x_1)$ are vector functions. Let $u(z) = h_1(z)$ and $v_1(z, x_1) = \{1 - w_1(z, x_1)\} \times (h_1(z), g_1(z, x_1))^T$, and consider the following class of influence functions:

$$\left\{ \frac{R D_1 Y_1}{\delta w_1(Z, X_1)} + \rho_1 \frac{R \{D_1 - w_1(Z, X_1)\}}{\delta w_1(Z, X_1) \{1 - w_1(Z, X_1)\}} \Psi_1(Z, X_1) + \rho_2 \left(\frac{R D_1}{\delta w_1(Z, X_1)} - \frac{1 - R}{1 - \delta} \right) \Psi_2(Z); \rho_1 \in R^5, \rho_2 \in R^2 \right\}.$$

The DTL estimator is a member of the class with $\mathbf{p}_1 = (0, 0, 1 - \delta, 1, 0)^T$ and $\mathbf{p}_2 = (0, -1 + \delta)$. Straightforward algebra shows that the EL estimator has the same influence function as the DTL estimator if $h_1(Z) = E[Y_1|Z]$ and $g_1(Z, X_1) = E[Y_1|Z, X_1]$ (or consistent estimates of them) are used in the estimation procedure. Thus the EL estimator achieves the semiparametric efficiency bound in this case. On the other hand, when the conjectured functions h_1 and g_1 are misspecified, the EL estimator is asymptotically more efficient than

the DTL estimator. This follows from the fact that $\mathbf{A}$ is the variance–covariance matrix of $\mathbf{B}_j$ and $\mathbf{C}$ is the covariance between $\mathbf{B}_j$ and $R_i D_{1j} y_{1j} / \{\delta w_1(z_i, x_{1i})\}$. Thus the influence function of the EL estimator is asymptotically the most efficient one in the class.

Following (11), the EL estimator is asymptotically equivalent to the residual projection of the HT estimator onto the space spanned by $\mathbf{B}_j$. Note that $\mathbf{B}_j$ is E-ancillary in the sense of Small and McLeisch (1989); that is, it has zero expectation regardless of the choice of $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$. Intuitively, the residual of the DTL estimator projection onto the space spanned by $\mathbf{B}_j$ would preserve the double-robustness property, as well as improve efficiency over the DTL estimator. To exploit this, we consider the following estimator:

$$\hat{\mu}_{1ELDR} = \hat{\mu}_{1DTL} - \mathbf{C}_n \mathbf{A}_n^{-1} \times \frac{1}{n} \sum_{i=1}^n \mathbf{B}_i,$$

where $\mathbf{A}_n$ is the empirical variance estimate of $n^{-1} \sum_{i=1}^n \mathbf{B}_i$ and $\mathbf{C}_n = (\mathbf{C}_{1n}, \mathbf{C}_{2n})$ is given by

$$\mathbf{C}_{1n} = \frac{1}{n} \sum_{i=1}^n \frac{R_i D_i}{\delta^2 w_1(z_i, x_{1i})^2} \{y_i - g_1(z_i, x_{1i})\} \boldsymbol{\psi}_1(z_i, x_{1i})$$

and

$$\mathbf{C}_{2n} = \frac{1}{n} \sum_{i=1}^n \left[\frac{R_i D_i}{\delta^2 w_1(z_i, x_{1i})^2} \{y_i - g_1(z_i, x_{1i})\} + \frac{R_i}{\delta^2} \{g_1(z_i, x_{1i}) - h_1(z_i)\} \right] \boldsymbol{\psi}_2(z_i).$$

Straightforward calculation shows that if w_1 is correctly specified, then $\mathbf{C}_n$ converges in probability to the covariance between $\hat{\mu}_{1DTL}$ and $\mathbf{B}_j$. Moreover, if g_1 and h_1 are correctly specified, then $\mathbf{C}_n \rightarrow \mathbf{0}$ in probability. It follows that $\hat{\mu}_{1ELDR}$ is a consistent estimator for μ_1 if either (a) the conjectured functions h_1 and g_1 both are correctly specified and w_1 is misspecified or (b) h_1 and g_1 are not correctly specified but w_1 is correctly specified. Thus $\hat{\mu}_{1ELDR}$ has the desired double-robustness property and is asymptotically more efficient than the DTL estimator when the working model is misspecified.

In practice, the missingness probability $1 - w_1(z, x_1)$ is usually estimated by assuming a logistic regression model $\log[w_1(z, x_1, \boldsymbol{\theta}) / \{1 - w_1(z, x_1, \boldsymbol{\theta})\}] = \boldsymbol{\eta}(z, x_1)^T \boldsymbol{\theta}$, where $\boldsymbol{\eta}(z, x_1) = (1, z, x_1)^T$. Using the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ in $w_1(z, x_1, \hat{\boldsymbol{\theta}})$ in the foregoing estimation procedure and denoting the corresponding estimate as $\hat{\mu}_{1EL}$ (or $\hat{\mu}_{1ELDR}$), we state the asymptotic properties of the corresponding EL estimator in Theorem 3 by assuming the following conditions:

- A1** There exist $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $1 - \varepsilon_1 \leq w(z, x_1, \boldsymbol{\theta}) \leq \varepsilon_2$ for all (z, x_1) and for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. Moreover, $0 < P(R = 1, D_1 = 1) < 1$.
- A2** For all (z, x_1) , $w_1(z, x_1, \boldsymbol{\theta})$ admits all third partial derivatives $\partial^3 w_1(z, x_1, \boldsymbol{\theta}) / \partial \theta_j \partial \theta_m$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$; moreover, $|\partial^3 w_1(z, x_1, \boldsymbol{\theta}) / \partial \theta_k \partial \theta_l \partial \theta_m|$ is bounded by an integrable function for all $\boldsymbol{\theta}$'s in this neighborhood.
- A3** $E\{[1 - w_1(Z, X_1, \boldsymbol{\theta})] w_1(Z, X_1, \boldsymbol{\theta})^{-1} \|\boldsymbol{\psi}_1(Z, X_1)\|^3\} < \infty$ for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.

Details of the proof are given in the aforementioned technical report.

Theorem 3. Suppose that $0 < \delta < 1$ and $E[Y_1^2] < \infty$. Under conditions A1–A3, $\tilde{\mu}_{1EL}$ has the following asymptotic representation:

$$\tilde{\mu}_{1EL} = \frac{1}{n} \sum_{i=1}^n \frac{R_i D_{1i} Y_{1i}}{\delta w_1(z_i, x_{1i}, \theta_0)} + \mathbf{C}^* (\mathbf{A}^*)^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{B}_i^* + o_p(n^{-1/2}),$$

where θ_0 is the true parameter value, $\mathbf{B}_i^* = (R_i \{D_{1i} - w_1(z_i, x_{1i}, \theta_0)\} \eta(z_i, x_{1i})^T, \mathbf{B}_i^T)^T$, and $\mathbf{C}^* = (E[Y_1 \{1 - w_1(Z, X_1)\} \times \boldsymbol{\eta}(Z, X_1)^T], \mathbf{C})$, and

$$\mathbf{A}^* = \begin{pmatrix} \mathbf{A}_{11}^* & \mathbf{0} \\ \mathbf{A}_{21}^* & \mathbf{0} \end{pmatrix}$$

with $\mathbf{A}_{11}^* = \delta E[w_1(Z, X_1, \theta_0) \{1 - w_1(Z, X_1, \theta_0)\} \eta(Z, X_1) \times \eta(Z, X_1)^T]$ and $\mathbf{A}_{21}^* = E[\psi_1^*(Z, X_1) \eta(Z, X_1)^T + \{1 - w_1(Z, X_1, \theta_0)\} \psi_2^*(Z, X_1) \eta(Z, X_1)^T]$. Thus in distribution,

$$\sqrt{n}(\tilde{\mu}_{1EL} - \mu_1) \rightarrow N(0, \sigma^{*2}),$$

where $\sigma^{*2} = \sigma^2 - 2\mathbf{C}^* \mathbf{A}^{*-1} (\mathbf{C}^*)^T + \mathbf{C}^* \mathbf{A}^{*-1} \mathbf{B}^* (\mathbf{A}^{*-1})^T \mathbf{C}^{*T}$ and $\mathbf{B}^*$ is the variance–covariance matrix of $\mathbf{B}_i^*$.

In practice, the working models for the means usually are estimated from the observed data. Suppose that the functions $u(z)$ and $v_1(z, x_1)$ can be parameterized as $u(z, \beta_1)$ and $v_1(z, x_1, \beta_2)$; thus $\boldsymbol{\psi}$ is parameterized as $\boldsymbol{\psi}(z, x_1, \beta, \theta)$ with $\beta = (\beta_1, \beta_2)$. Suppose that β is estimated by $\hat{\beta}$ and $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$ for some β_0 . Assume that $\boldsymbol{\psi}(z, x_1, \beta, \theta)$ admits all second partial derivatives $\boldsymbol{\psi}(z, x_1, \beta, \theta) / \beta_k \beta_l$ for all (β, θ) in a neighborhood of (β_0, θ_0) . Further assume that $E\|\boldsymbol{\psi}(z, x_1, \beta, \theta) / \beta\|^2 < \infty$ and the second derivative $\boldsymbol{\psi}(z, x_1, \beta, \theta) / \beta_k \beta_l$ is bounded by an integrable function in the same neighborhood of (β_0, θ_0) . Then the EL estimator with an estimated working model has the same asymptotic representation as in Theorem 3, with $\boldsymbol{\psi}(z, x_1, \theta_0)$ replaced by $\boldsymbol{\psi}(z, x_1, \beta_0, \theta_0)$.

Although we give only the results for estimating μ_1 in this section, the procedure can be applied in an analogous manner to estimate μ_0 . Let $\hat{\mu}_{0EL}$ be the EL estimator for μ_0 . The treatment effect is estimated by $\hat{\Delta} = \hat{\mu}_{1EL} - \hat{\mu}_{0EL}$. The large sample properties of $\hat{\Delta}$ follow directly from the asymptotic representations of $\hat{\mu}_{1EL}$ and $\hat{\mu}_{0EL}$. The derivation is tedious, because $\hat{\mu}_{1EL}$ and $\hat{\mu}_{0EL}$ are correlated through the common baseline covariates. We omit the details in this article. To construct a confidence interval for Δ , we adopt a nonparametric bootstrap method by repeatedly sampling subjects with replacement from the observed data. It is known that when the sample size is large and the statistic is well behaved, the nonparametric bootstrap method generally works well. Let $\hat{\Delta}^1, \dots, \hat{\Delta}^B$ be the estimated treatment effect in B bootstrap repetitions. Then the 95% confidence interval of Δ is given by the 2.5th and 97.5th percentiles of $\{\hat{\Delta}^1, \dots, \hat{\Delta}^B\}$.

4. MONTE CARLO SIMULATIONS

To evaluate the performance of the proposed estimators, we conduct a series of simulation studies to compare them with alternative methods, including the DTL estimator, $\hat{\mu}_{1DTL}$; the HT estimator, and the Hájek estimator given by

$$\hat{\mu}_{1HJ} = \left\{ \sum_{i=1}^n \frac{R_i D_{1i} y_{1i}}{\delta w_1(z_i, x_{1i})} \right\} \left\{ \sum_{i=1}^n \frac{R_i D_{1i}}{\delta w_1(z_i, x_{1i})} \right\}^{-1}.$$

As a reference value, we also calculate the sample mean of all posttest measures for each treatment group, including those that are set to missing. For all simulation studies, we generate 2,000 simulated data sets, each with 400 subjects. The subjects are randomized to treatment ($R = 1$) and control ($R = 0$) with equal probabilities ($\delta = .5$).

For the first set of simulation studies, we generate the baseline covariate Z from a $N(0, 1)$ distribution. We generate the intermediate measures by assuming $X_1 = 2 + Z + \varepsilon_1$ for treatment and $X_0 = 1 + Z + \varepsilon_0$ for control. Moreover, we simulate posttest measures by assuming $Y_1 = 1 + 2X_1 - .2X_1^2 + \varepsilon_1^*$ for treatment and $Y_0 = 1 + X_0 - .1X_0^2 + \varepsilon_0^*$ for control. Thus the intermediate X is related to Z , and, given X , Z is irrelevant to Y . For the second set of simulation studies, we generate the baseline covariate Z from an exponential distribution with mean 1. We generate the intermediate measures by assuming $X_1 = 1 + Z + \varepsilon_1$ for treatment and $X_0 = Z + \varepsilon_0$ for control. Moreover, we simulate posttest measures by assuming $Y_1 = 1 + 2X_1 - .2X_1^2 + X_1 \varepsilon_1^*$ for treatment and $Y_0 = 1 + X_0 - .1X_0^2 + X_0 \varepsilon_0^*$ for control. The values ε_R and ε_R^* , $R = 0, 1$, are generated from an $N(0, 1)$ distribution, and Z , ε_R , and ε_R^* are mutually independent. In all simulations, the posttest measure is independent of the baseline covariate after conditioning on the intermediate measurement. The first set of simulations assumes a constant error variance for the posttest measure, whereas the second set of simulations assumes heteroscedasticity errors.

The probability of missing in each group is determined by a logistic regression model in which $P(D_R = 1 | z, x_R) = w_R(z, x_R) = 1 / \{1 + \exp(\theta_{0R} + \theta_1 z + \theta_2 x_R)\}$. We consider scenarios in which the posttest measures are MAR with $\theta_{01} = -2$, $\theta_{00} = 0$, $\theta_1 = 2$, and $\theta_2 = -1$, as well as scenarios in which the response is missing completely at random (MCAR) with $\theta_{01} = \theta_{00} = 1$ and $\theta_1 = \theta_2 = 0$.

In the simulation study, the missing probability function w_R is estimated by fitting a separate logistic regression with Z and X_R as covariates. We choose the conjectured mean functions h_R and g_R by fitting two different sets of working regression models,

$$Y_R = a_{0R} + a_{1R}z + \varepsilon_{1R} \text{ and } Y_R = b_{0R} + b_{1R}x_R + b_{2R}z + \varepsilon_{2R},$$

designated working model (I), and

$$Y_R = a_{0R} + a_{1R}z + a_{2R}z^2 + \varepsilon_{1R} \text{ and } Y_R = b_{0R} + b_{1R}x_R + b_{2R}x_R^2 + \varepsilon_{2R},$$

designated working model (II), for $R = 1, 0$, to the observed data in each group. Note that the second set of working models is closer to the underlying relationship and thus is expected to yield more efficient estimates.

Some details about the algorithm and the software program used for the simulation studies are available at http://www.amstat.org/publications/jasa/supplemental_materials. Simulation results are summarized in Tables 2 and 3. In general, the DTL estimator and the EL estimator perform better than the HT estimator and the Hájek (HJ) estimator, because the former incorporate baseline and intermediate covariate information in the estimation

procedure. The EL estimator outperforms the DTL estimator in terms of having a smaller mean squared error (MSE) under MAR (I) in Table 2 and under MAR (I) and MAR (II) in Table 3. These differences are statistically significant after accounting for simulation variation. In contrast, the two estimators yield similar results when the working model is close to the truth or the data are MCAR. We also report the simulation results for the new double-robust estimator $\tilde{\mu}_{ELDR}$. It can be observed that the proposed double-robust estimator yields similar MSEs to the EL estimator. Table 4 reports the average of the 10th, 50th, and 90th percentile estimates in each group, which are available only with the EL method. It is easy to see that the EL estimates are close to the true percentiles of the response distribution, except perhaps for the 10th percentile under heteroscedastic errors with MAR.

5. DATA ANALYSIS

We apply the proposed method to data from patients enrolled in the AIDS Clinical Trials Group (ACTG) protocol 175 (Hammer et al. 1996), a randomized, double-blind trial designed to evaluate nucleoside monotherapy versus combination therapy in human immunodeficiency virus (HIV)-infected individuals. A total of 2,139 HIV-positive patients with CD4 counts between 200 and 500 cells/mm³ were equally randomized to 4 antiretroviral regimens: zidovudine (ZDV) monotherapy, ZDV plus didanosine (ddI), ZDV plus zalcitabine (ddC), and ddI monotherapy. Following DTL (2005), we are interested comparing the CD4 counts measured at 96 ± 5 weeks postrandomization (Y) between patients who received ZDV monotherapy (control group) and those who received other therapies (treatment group). The baseline covariates, Z , used in the data analysis include weight, HIV symptoms, previous experience with antiretroviral drug therapy, Karnofsky score (a measure of ability to perform activities of daily living), CD4 count, CD8 count, and the quadratic terms of CD4 and CD8 counts. The intermediate measures, X , at 20 ± 5 weeks postrandomization include treatment status, CD4 count, CD8 count, and the quadratic terms of CD4 and CD8 counts. The follow-up response, Y , is the CD4 count at 96 ± 5 weeks. Of the 2,139 participants, 37% had a missing CD4 count at 96 ± 5 weeks.

Following DTL (2005), we first fit separate linear models to the observed data in each treatment group by using the ordinary least squares method to obtain guesses of $E[Y_1|X_1, Z]$ and $E[Y_1|Z]$ for the treatment group and of $E[Y_0|X_0, Z]$ and $E[Y_0|Z]$ for the control group. We denote the fitted models by $\hat{g}_1(z, x)$ and $\hat{h}_1(z)$ for treatment and by $\hat{g}_0(z, x)$ and $\hat{h}_0(z)$ for control. Next, we fit logistic regressions to estimate the missing probabilities using the aforementioned 13 predictors, fit separately for treatment and control, and denote the fitted model by $\hat{w}_1(z, x)$ and $\hat{w}_0(z, x)$. For computing the EL estimator, we set $\mathbf{v}(z, x) = \{1 - \hat{w}_1(z, x)\}(\hat{h}_1(z), \hat{g}_1(z, x))$ and $\mathbf{u}(z) = \hat{h}_1(z)$ for treatment ($R = 1$) and set $\mathbf{v}(z, x) = \{1 - \hat{w}_0(z, x)\}(\hat{h}_0(z), \hat{g}_0(z, x))$ and $\mathbf{u}(z) = \hat{h}_0(z)$ for control ($R = 0$).

The estimated treatment effects, $\hat{\Delta}$, given by the HT, HJ, DTL, EL, and the new double-robust estimators are 53.76, 54.69, 57.24, 57.15, and 57.22 cells per mm³. To obtain standard errors, we repeatedly sample 2,139 subjects with replacement from the observed data 1,000 times and apply the four estimation procedures to each resampling data set. Table 5 gives the estimates of standard errors (SEs) and 95% confidence intervals (CIs) based on 1,000 bootstrap estimates for each estimation method. The 95% bootstrap percentile CIs for the five methods are (29.91, 76.87), (31.48, 77.37), (38.04, 77.30), (37.70, 77.13), and (37.88, 77.19). Thus all of the methods suggest that the mean CD4 count at 96 ± 5 weeks is significantly higher in the treatment group than in the control group. The EL estimator, the proposed double-robust estimator, and the DTL estimator give similar estimates, and their SE estimates are smaller than those of the HT and HJ estimators.

The cumulative distribution of the observed CD4 counts at 96 ± 5 weeks for the control and treatment groups are given in Figure 1 (solid lines). As suggested by DTL, the missing-data mechanism is not MCAR, and thus the complete cases are not representative. To see how the proposed method corrects the bias, Figure 1 shows the EL-based estimates of the distribution function. For the treatment group at the 10th, 50th, and 90th percentiles, the observed CD4 counts are 24, 15, and 7 cells larger than the EL estimates. For the control group, the analogous differences are 15, 22, and 9 cells. Because all differences are positive, this suggests that sicker patients are more likely to have missing observations. Because $22 > 15$, it is not surprising that the EL method yields a greater estimate of treatment effect (57.15 cells/mm^3) than the crude comparison (53.83 cells/mm^2). Similar behavior is seen for the median, with an EL estimate of 55 (95% CI, 39–78), whereas the observed difference is 47. The median, of course, has appeal as a robust estimate of treatment effect with skewed data.

6. DISCUSSION

In this article, we have applied the EL method to the pretest–posttest study with missing data by incorporating information from the baseline and posttreatment covariates to improve efficiency. Under MCAR or a correct specification of the underlying relationship between the posttest outcome and baseline covariates, the EL estimator of the treatment effect is asymptotically equivalent to the doubly robust estimator studied by LTD (2003); both achieve the semiparametric efficiency bound. In practice, however, the exact relationship is unknown. We have shown that under MAR, the EL estimator can be more efficient than the LTD estimator if the prespecified, conjectured working model is not the optimal choice. The EL method also gives an asymptotically unbiased estimate of the response distribution, and, as demonstrated in the data analysis of the AIDS clinical trial, the proposed method can be very useful for comparing treatments when the response distribution is skewed.

In the special case where there are no missing data (or under MCAR) and where the conjectured function is $a(z) = a_0 + a_1 z$, the influence function of the EL estimator is given by

$$\frac{RY_1}{\delta} - (\rho_1 + \rho_2 a_0) \frac{R - \delta}{\delta} - \rho_2 a_1 \frac{R - \delta}{\delta} Z.$$

Straightforward derivation shows that if a_0 and a_1 are regression coefficients estimated from the data, then the asymptotic representation of the EL estimate has coefficients $(\rho_1, \rho_2) = (0, 1)$, the same as the asymptotic representation of the LTD estimator. Thus, without missing data (or under MCAR) and with an estimated linear regression working model, the EL estimator and the LTD estimator are asymptotically equivalent even when the function $E[Y_1|Z]$ is misspecified. But such a result does not apply to situations where the posttest measure is MAR. When the working models for $E[Y_1|Z]$ and $E[Y_1|Z, X_1]$ are misspecified, the EL estimator and the DTL estimator usually have different asymptotic representations in the MAR setting.

Acknowledgments

The authors thank Professor Marie Davidian for providing ACTG 175 data. They also thank the referees, an associate editor, and the joint editor's for helpful comments and suggestions.

APPENDIX: PROOF OF THEOREM 1

Substituting the p_j 's and r_j 's given in (5) and (6) back in the log-likelihood function (2), we obtain the profile log-likelihood

$$\ell = - \sum_{i=1}^{n_1} \log[1 + \lambda\{a(z_i) - a\}] - \sum_{i=n_1+1}^n \log[1 + \tau\{a(z_i) - a\}].$$

We then maximize ℓ by taking derivative with respect to a and apply the constraints in (3) to obtain $n_1\lambda + n_0\tau = 0$.

Let $\boldsymbol{\psi}(z_i) = (1, a(z_i))^T$ and denote $\boldsymbol{\xi} = (\xi_1, \xi_2)$, where $\xi_1 = n_1(1 - \lambda a)/n$ and $\xi_2 = n_1\lambda/n$. We can reparameterize p_i and r_i as

$$p_i = \frac{1}{n} \times \frac{1}{\boldsymbol{\xi}^T \boldsymbol{\psi}(z_i)}, i=1, \dots, n_1,$$

and

$$r_i = \frac{1}{n} \times \frac{1}{1 - \boldsymbol{\xi}^T \boldsymbol{\psi}(z_i)}, i=n_1+1, \dots, n.$$

The constraints (3) and (4) imply $\sum_{i=1}^n R_i p_i \psi(z_i) = \sum_{i=1}^n (1 - R_i) r_i \psi(z_i)$. By plugging the p_i 's and r_i 's into (3) and (4), we obtain the following estimating equation for $\boldsymbol{\xi}$:

$$\mathcal{U}(\boldsymbol{\xi}) = \frac{1}{n} \sum_{i=1}^n \left[\frac{R_i}{\boldsymbol{\xi}^T \boldsymbol{\psi}(z_i)} - \frac{1 - R_i}{1 - \boldsymbol{\xi}^T \boldsymbol{\psi}(z_i)} \right] \boldsymbol{\psi}(z_i) = \mathbf{0}.$$

For ease of notation, we use $\mathbf{0}$ to denote zero vectors. Denote $\boldsymbol{\xi}_0^T = (\delta, 0)$. It follows from $\boldsymbol{\xi}_0^T \boldsymbol{\psi}(z_i) = \delta$ that $E[\mathcal{U}(\boldsymbol{\xi}_0)] = \mathbf{0}$. Let $\hat{\boldsymbol{\xi}}$ be the solution of $\mathcal{U}(\boldsymbol{\xi}) = \mathbf{0}$. Assuming $E\|a(Z)\|^{-3} < \infty$ and arguing as in the proof of lemma 1 of Qin and Lawless (1994), we can show by a Taylor series expansion that

$$\sqrt{n}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) = \mathbf{A}^{-1} \times \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{B}_i + o_p(1),$$

where $\mathbf{A} = \delta^{-1}(1 - \delta)^{-1} E\{\boldsymbol{\psi}(Z)\boldsymbol{\psi}(Z)^T\}$ and $\mathbf{B}_i = \delta^{-1}(1 - \delta)^{-1} \times (R_i - \delta)\boldsymbol{\psi}(z_i)$. Under the assumed regularity conditions in Theorem 1, it follows from the central limit theorem for iid random vectors that as $n \rightarrow \infty$, $n^{-1/2} \sum_{i=1}^n \mathbf{B}_i$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{A}$. The EL-based estimator for μ_1 is given by $\sum_{i=1}^{n_1} \hat{p}_i y_{1i}$, where $\hat{p}_i$ is derived by replacing $\boldsymbol{\xi}$ with $\hat{\boldsymbol{\xi}}$ in (5). Following a first-order expansion of $\hat{p}_i$ about $\boldsymbol{\xi}_0$, the EL estimator has the following asymptotic representation:

$$\hat{\mu}_{1EL} = \frac{1}{n} \sum_{i=1}^n \frac{R_i y_{1i}}{\hat{\boldsymbol{\xi}}^T \boldsymbol{\psi}(z_i)} = \frac{1}{n} \sum_{i=1}^n \frac{R_i y_{1i}}{\delta} - \frac{1}{n} \sum_{i=1}^n \frac{R_i y_{1i}}{\delta^2} \boldsymbol{\psi}(z_i)^T (\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0) + o_p(n^{-1/2}) = \frac{1}{n} \sum_{i=1}^n \frac{R_i y_{1i}}{\delta} - \delta^{-1} E\{Y_1 \boldsymbol{\psi}(Z)^T\} \mathbf{A}^{-1} \times \frac{1}{n} \sum_{i=1}^n \mathbf{B}_i + o_p(n^{-1/2}).$$

By double expectation, we have

$$\text{cov}\left(\frac{R}{\delta}Y_1, \frac{R_i - \delta}{\delta}\psi(Z)\right) = \frac{(1 - \delta)}{\delta}E\{Y_1\psi(Z)\}.$$

Thus as $n \rightarrow \infty$, $\sqrt{n}(\hat{\mu}_{1EL} - \mu_1)$ converges weakly to a normal distribution with mean 0 and variance σ_1^2 , where

$$\sigma_1^2 = \left\{ \frac{1}{\delta}E(Y_1^2) - \mu_1^2 \right\} - \frac{1 - \delta}{\delta}E\{Y_1\psi(Z)^T\}[E\{\psi(Z)\psi(Z)^T\}]^{-1}E\{Y_1\psi(Z)\}.$$

REFERENCES

- Bonate, PL. Analysis of Pretest–Posttest Designs. New York: Chapman & Hall/CRC; 2003.
- Brogan DR, Kutner MH. Comparative Analyses of Pretest–Posttest Research Design. The American Statistician. 1980; 34:229–232.
- Chen J, Qin J. Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information. Biometrika. 1993; 80:107–116.
- Chen J, Wu C. Estimation of Distribution Function and Quantiles Using the Model-Calibrated Pseudo-Empirical Likelihood Method. Statistica Sinica. 2002; 12:1223–1239.
- Chen J, Sitter RR, Wu C. Using Empirical Likelihood Methods to Obtain Range-Restricted Weights in Regression Estimators for Surveys. Biometrika. 2002; 89:230–237.
- Davidian M, Tsiatis AA, Leon S. Semiparametric Estimation of Treatment Effect in a Pretest–Posttest Study With Missing Data. Statistical Science. 2005; 3:261–301. (with discussion). [PubMed: 19081743]
- Fleiss, JL. The Design and Analysis of Clinical Experiments. New York: Wiley; 1986.
- Follmann D. The Effect of Screening on Some Pretest–Posttest Test Variances. Biometrics. 1991; 47:763–771. [PubMed: 1912270]
- Hájek, J. Sampling From a Finite Population. New York: Marcel Dekker; 1981.
- Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC. for The AIDS Clinical Trials Group Study 175 Study Team. A Trial Comparing Nucleoside Monotherapy With Combination Therapy in HIV–Infected Adults With CD4 Cell Counts From 200 to 500 per Cubic Millimeter. New England Journal of Medicine. 1996; 335:1081–1090. [PubMed: 8813038]
- Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association. 1952; 47:663–685.
- Laird N. Further Comparative Analyses of Pretest–Posttest Research Design. The American Statistician. 1983; 37:329–340.
- Leon S, Tsiatis AA, Davidian M. Semiparametric Estimation of Treatment Effect in a Pretest–Posttest Study. Biometrics. 2003; 59:1046–1055. [PubMed: 14969484]
- Liang H, Wang SJ, Carroll R. Partially Linear Models With Missing Response Variables and Error-Prone Covariates. Biometrika. 2007; 94:1–14.
- Owen AB. Empirical Likelihood Ratio Confidence Intervals for a Single Functional. Biometrika. 1988; 75:237–249.
- Owen AB. Empirical Likelihood Ratio Confidence Regions. The Annals of Statistics. 1990; 18:90–120.
- Owen, AB. Empirical Likelihood. Boca Raton, FL: Chapman & Hall/CRC; 2001.
- Qin J, Lawless J. Empirical Likelihood and General Estimating Equations. The Annals of Statistics. 1994; 22:300–325.
- Qin J, Zhang B. Empirical–Likelihood–Based Inference in Missing Response Problems and Its Application in Observational Studies. Journal of the Royal Statistical Society, Ser. B. 2007; 69:101–122.

- Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Small CG, McLeish DL. Projection as a Method for Increasing Sensitivity and Eliminating Nuisance Parameters. *Biometrika*. 1989; 76:693–703.
- Stanek EJ. Choosing a Pretest–Posttest Analysis. *The American Statistician*. 1988; 42:178–183.
- Tan Z. A Distributional Approach for Causal Inference Using Propensity Scores. *Journal of the American Statistical Association*. 2006; 101:1619–1637.
- Wang D, Chen SX. Empirical Likelihood for Estimating Equation With Missing Values. *The Annals of Statistics*. 2006 to appear.
- Wang Q, Rao JNK. Empirical Likelihood–Based Inference Under Imputation for Missing Response Data. *The Annals of Statistics*. 2002; 30:896–924.
- Wang Q, Linton O, Härdle W. Semiparametric Regression Analysis With Missing Response at Random. *Journal of the American Statistical Association*. 2004; 99:334–345.
- Wang X, Zhou H. A Semiparametric Empirical Likelihood Method for Biased Sampling Schemes With Auxiliary Covariates. *Biometrics*. 2006; 62:1149–1160. [PubMed: 17156290]
- Wu C, Sitter RR. A Model-Calibration Approach to Using Complete Auxiliary Information From Survey Data. *Journal of the American Statistical Association*. 2001; 96:185–193.
- Yang L, Tsiatis AA. Efficiency Study of Estimators for a Treatment Effect in a Pretest–Posttest Trial. *The American Statistician*. 2001; 55:314–321.
- Zhou Y, Liang H. Empirical-Likelihood–Based Semiparametric Inference for the Treatment Effect in the Two-Sample Problem With Censoring. *Biometrika*. 2005; 92:271–282. [PubMed: 20179776]

Figure 1.

Estimated cdf's for (a) treatment and (b) control. The solid lines represent the empirical distribution functions; dashed lines represent the EL estimators. The dotted lines indicate the observed 10th, 50th, and 90th percentile CD4 count (treatment, 120, 330, and 550; control, 55, 283, and 500) and the EL-based estimates (treatment, 96, 315, and 543; control, 40, 261, and 491) at 96 ± 5 week postrandomization.

Table 1

The observed data in a pretest–posttest study

Subject i	Group R	Baseline Z	Intermediate		Missing identity		Final response	
			X_1	X_0	D_1	D_0	Y_1	Y_0
1	1	z_1	x_{11}	??	1	??	y_{11}	??
:	:	:	:	:	:	:	:	:
n_1	1	z_{n_1}	x_{1n_1}	??	1	??	y_{1n_1}	??
$n_1 + 1$	1	z_{n_1+1}	x_{1n_1+1}	??	0	??	?	??
:	:	:	:	:	:	:	:	:
n_1	1	z_{n_1}	x_{1n_1}	??	0	??	?	??
$n_1 + 1$	0	z_{n_1+1}	??	x_{0n_1+1}	??	1	??	y_{0n_1+1}
:	:	:	:	:	:	:	:	:
$n_1 + n_0$	0	$z_{n_1+n_0}$	??	$x_{0n_1+n_0}$	??	1	??	$y_{0n_1+n_0}$
$n_1 + n_0 + 1$	0	$z_{n_1+n_0+1}$	??	$x_{0n_1+n_0+1}$	??	0	??	?
:	:	:	:	:	:	:	:	:
n	0	z_n	??	x_{0n}	??	0	??	?

NOTE: ?? denotes “counterfactual” missing; ? denotes traditional MAR.

Table 2

Summary of simulation studies with constant errors in posttest measurements

Missing-data mechanism	Working model	Method	μ_1			μ_0			Δ		
			Bias	MSE	MSE	Bias	MSE	MSE	Bias	MSE	MSE
MAR	(I)	SM	-.75	2.10	1.18	-.38	1.18	-.37	3.24		
		HT	-.09	6.27	.37	3.83	-.46	10.00			
		HJ	2.93	11.60	1.95	5.11	.98	16.50			
		DTL	.71	5.86	.34	2.46	.36	7.71			
		EL	.18	3.10	-.59	1.81	.77	4.21			
		ELDR	3.30	2.62	1.41	1.57	1.89	3.49			
MAR	(II)	SM	.02	2.14	-.124	1.20	1.26	3.35			
		HT	.62	7.06	.33	3.39	.29	10.45			
		HJ	3.56	12.09	2.17	4.7	1.39	16.55			
		DTL	-.16	2.42	-.76	1.74	.59	3.71			
		EL	.20	2.77	-.89	1.90	1.09	4.15			
		ELDR	-1.34	2.36	-1.66	1.71	.32	3.52			
MCAR	(I)	SM	-.71	2.10	-.38	1.17	-.33	3.20			
		HT	-.69	2.33	-.36	1.39	-.33	3.73			
		HJ	-.68	2.33	-.37	1.40	-.32	3.73			
		DTL	-.46	2.04	-.41	1.19	-.05	2.72			
		EL	-.37	2.05	-.24	1.21	-.12	2.74			
		ELDR	.27	2.03	.05	1.19	.22	2.70			
MCAR	(II)	SM	-.12	2.10	.02	1.15	-.15	3.24			
		HT	-.29	2.30	-.05	1.41	-.25	3.69			
		HJ	-.30	2.31	-.05	1.41	-.25	3.69			
		DTL	-.19	1.86	-.04	1.23	-.16	2.56			
		EL	.20	1.91	.16	1.26	.04	2.65			
		ELDR	-.18	1.86	-.02	1.23	-.16	2.57			

NOTE: Working model (I) is $hR(z) = \hat{a}_0R + \hat{a}_1Rz$ and $gR(x, z) = \hat{h}_0R + \hat{b}_1RvR + \hat{b}_2Rz$. Working model (II) is $hR(z) = \hat{a}_0R + \hat{a}_1Rz + \hat{a}_2Rz^2$ and $gR(x, z) = \hat{b}_0R + \hat{b}_1R^2x_R + \hat{b}_2R^2x_R^2$. Bias and MSE are empirical bias ($\times 100$) and mean squared errors ($\times 100$) based on 2,000 estimates using the sample mean (SM), and the HT, HJ, DTL, EL, and ELDR estimators.

Table 3
Summary of simulation studies with heteroscedasticity errors in posttest measurements

Missing-data mechanism	Working model	Method	$\hat{\mu}_1$			$\hat{\mu}_0$			Δ		
			Bias	MSE		Bias	MSE		Bias	MSE	
MAR	(I)	SM	1.29	4.46	.06	2.05	1.23	6.66			
		HT	1.44	8.41	-.18	12.33	1.62	21.16			
		HJ	5.05	16.71	2.76	9.56	2.29	26.53			
		DTL	2.55	11.25	.37	12.70	2.18	24.15			
		EL	5.18	5.49	1.95	2.64	3.23	7.89			
		ELDR	6.72	5.32	2.85	2.75	3.88	7.38			
MAR	(II)	SM	.65	4.16	.76	2.08	-.11	6.25			
		HT	.78	8.23	1.31	8.54	-.53	16.79			
		HJ	4.62	17.50	3.39	8.48	1.23	25.92			
		DTL	.37	8.50	.95	5.80	-.58	14.35			
		EL	2.70	5.47	1.22	2.92	1.47	8.13			
		ELDR	-.85	5.64	-.36	3.22	-.49	8.64			
MCAR	(I)	SM	1.02	4.33	1.07	2.10	-.05	6.51			
		HT	.81	5.61	1.04	2.80	-.23	8.57			
		HJ	.80	5.61	1.04	2.80	-.24	8.56			
		DTL	.90	5.44	1.23	2.76	-.34	8.08			
		EL	.86	5.45	1.19	2.77	-.33	7.97			
		ELDR	1.92	5.47	1.85	2.75	.07	7.93			
MCAR	(II)	SM	.32	4.19	.64	1.99	-.32	6.09			
		HT	.49	5.52	.44	2.62	.04	8.10			
		HJ	.47	5.51	.44	2.62	.03	8.09			
		DTL	.43	5.37	.67	2.62	-.24	7.66			
		EL	.79	5.34	.65	2.58	.14	7.60			
		ELDR	.64	5.49	.63	2.67	.01	7.81			

NOTE: Working model (I) is $hR(z) = \hat{a}_0R + \hat{a}_1Rz$ and $gR \times R \cdot z = \hat{b}_0R + \hat{b}_1R \times R + \hat{b}_2Rz$. Working model (II) is $hR(z) = \hat{a}_0R + \hat{a}_1Rz + \hat{a}_2Rz^2$ and $gR(z) = \hat{b}_0R + \hat{b}_1R + \hat{b}_2Rz + \hat{b}_3Rz^2$. Bias and MSE are empirical bias ($\times 100$) and mean squared errors ($\times 100$) based on 2,000 estimates using the SM, HT, HJ, DTL, EL, and ELDR estimators.

Table 4

Summary of percentile estimates

Missing-data mechanism	Working model	R = 1 percentiles			R = 0 percentiles		
		10	50	90	10	50	90
Constant error model							
True values							
MAR	(I)	1.0799	4.0782	6.1289	−.3116	1.8056	3.5683
		1.0842	3.9935	6.0710	−.3231	1.7562	3.5094
MAR	(II)	.9951	3.9946	6.0801	−.3502	1.7535	3.5181
MCAR	(I)	1.0323	4.0617	6.0972	−.3370	1.7942	3.5398
MCAR	(II)	1.0356	4.0646	6.1036	−.3427	1.7914	3.5343
Nonconstant error model							
True values							
MAR	(I)	.8420	3.3678	7.5261	−.1479	1.3195	4.0797
		1.0733	3.3223	7.3784	−.0574	1.2950	3.9627
MAR	(II)	1.0427	3.2863	7.3705	−.1039	1.2840	3.9524
MCAR	(I)	.8043	3.3496	7.4511	−.1942	1.3247	4.0367
MCAR	(II)	.8077	3.3500	7.4464	−.2022	1.3212	4.0250

NOTE: Working model (I) is $hR(z) = \hat{a}_0R + \hat{a}_1Rz$ and $gR^{\text{vs}}R \cdot z) = \hat{b}_0R + \hat{b}_1R^{\text{vs}}R + \hat{b}_2Rz$. Working model (II) is $hR(z) = \hat{a}_0R + \hat{a}_1Rz + \hat{a}_2Rz^2$ and $gR^{\text{vs}}R \cdot z) = \hat{b}_0R + \hat{b}_1R^{\text{vs}}R + \hat{b}_2Rz + \hat{b}_3R^2$.

Table 5
Treatment effect estimates for CD4 count at 96 ± 5 weeks postrandomization in ACTG 175

Method	$\hat{\mu}_1$	$SE(\hat{\mu}_1)$	$\hat{\mu}_0$	$SE(\hat{\mu}_0)$	$\hat{\Delta}$	$SE(\hat{\Delta})$	95% CI
HT	325.56	5.39	271.80	10.74	53.76	11.97	(29.91, 76.87)
HJ	325.85	5.33	271.16	10.67	54.69	11.80	(31.48, 77.37)
DTL	324.46	5.17	267.21	9.58	57.24	10.47	(38.04, 77.30)
EL	325.13	5.18	267.98	9.38	57.15	10.33	(37.70, 77.13)
ELDR	324.92	5.18	267.70	9.40	57.22	10.34	(37.88, 77.19)

NOTE: $SE(\hat{\mu}_1)$, $SE(\hat{\mu}_0)$, and $SE(\hat{\Delta})$, the standard deviations of 2,000 bootstrap estimates of μ_0 , μ_1 , and Δ ; 95% CI, (2.5%, 97.5%) quantiles of the 2,000 estimates of Δ .