

A Pseudoscore Estimator for Regression Problems With Two-Phase Sampling

Nilanjan CHATTERJEE, Yi-Hau CHEN, and Norman E. BRESLOW

Two-phase stratified sampling designs yield efficient estimates of population parameters in regression models while minimizing the costs of data collection. In measurement error problems, for example, error-free covariates are ascertained only for units selected in a validation sample. Estimators proposed heretofore for such designs require all units to have positive probability of being selected. We describe a new semiparametric estimator that relaxes this assumption and that is applicable to, for example, case-only or control-only validation sampling for binary regression problems. It uses a weighted empirical covariate distribution, with weights determined by the regression model, to estimate the score equations. Implementation is relatively easy for both discrete and continuous outcome data. For designs that are amenable to alternative methods, simulation studies show that the new estimator outperforms the currently available weighted and pseudolikelihood methods and often achieves efficiency comparable to that of semiparametric maximum likelihood. The simulations also demonstrate the vulnerability of the case-only or control-only designs to model misspecification. These results are illustrated by the analysis of data from a population-based case-control study of leprosy.

KEY WORDS: Measurement error; Missing data; Pseudolikelihood; Response selective sampling; Restricted sampling; Semiparametric inference.

1. INTRODUCTION

Two-phase designs, introduced originally by Neyman (1938) as a technique for stratification, are currently used to estimate regression parameters β in a model $f_\beta(y|x, z)$, where y is a response variable and x and z are covariates. At phase one, N subjects with random variables $(Y_i, X_i, Z_i)_{i=1}^N$ are sampled at random from a population described by the joint density $f_\beta(y|x, z)dG(x|z)dH(z)$, where G and H denote arbitrary (nonparametric) conditional and marginal covariate distributions. Y_i and Z_i are observed for all N subjects, but X_i is observed only for those in a phase-two subsample selected according to a random mechanism. Let R_i denote the indicator of whether ($R_i = 1$) or not subject i is selected at phase two. We assume that $(R_i, Y_i, X_i, Z_i), i = 1 \dots N$, are iid random vectors and that

$$P(R = 1|Y, X, Z) = P(R = 1|Y, Z) \equiv \pi(Y, Z); \quad (1)$$

that is, the X 's are missing at random in the sense of Rubin (1976). Such designs are used in complex survey sampling (e.g., Skinner, Holt, and Smith 1989, sec. 1.6), where a finite population drawn from the superpopulation model (f_β, G, H) constitutes the "phase-one sample," in validation sampling for measurement error problems (Carroll, Ruppert, and Stefanski 1995, sec. 9) and in stratified case-control sampling in epidemiology (Breslow 1996).

If X_i were known for all subjects, then maximum likelihood estimation of β would involve solving the score equations

$$O = U(\beta) = \sum_{i=1}^N S_\beta(Y_i|X_i, Z_i) \equiv \sum_{i=1}^N \frac{\partial \log f_\beta(Y_i|X_i, Z_i)}{\partial \beta}. \quad (2)$$

When data are missing, the scores are replaced by their conditional expectations given the data that are observed. Such maximum likelihood techniques have been described for estimation of parameters in fully parametric models for complex survey data, including two-phase designs (Breckling, Chambers, Dorfman, Tam, and Welsh 1994). But misspecification of the parametric covariate distribution can lead to inconsistent estimates of the regression parameters (Pepe and Fleming 1991). Semiparametric efficient inference (Robins, Hsieh, and Newey 1995) alleviates this problem, but may be difficult to implement. When the outcome Y is continuous, it involves numerical solution of an infinite-dimensional integral equation. As far as we know, semiparametric efficient inference has been fully implemented only when Y (and often also Z) is discrete (Robins, Rotnitzky, and Zhao 1994; Robins et al. 1995; Breslow and Holubkov 1997; Scott and Wild 1997). Lawless, Kalbfleisch, and Wild (1999) recommended discretization of continuous phase-one data to achieve an easily calculable maximum profile likelihood estimator. As we show later, however, such data reduction can itself entail a substantial loss of efficiency.

In view of the computational complexity of efficient inference, most applications to date have involved simpler, inefficient methods. One strategy is to recognize that for fixed β , the efficient score (2) is a finite population sum. Hence it may be estimated from the phase-two sample by inverse probability weighting based on π (Horvitz and Thompson 1952; Cochran 1977, sec. 9). Skinner et al. (1989, sec. 3.4.4) presented a discussion of this strategy related to survey sampling, and Flanders and Greenland (1991) explored applications to epidemiology. A second approach is to consider the "complete data likelihood" generated by the weighted distributions of the phase-two observations, conditional on the event that they were sampled at phase two. Authors who have investigated this approach include Breslow and Cain (1988) for applications to case-control studies, Krieger and Pfeiffermann (1992)

Nilanjan Chatterjee is Principal Investigator, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Rockville, MD 20852. Yi-Hau Chen is Assistant Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, ROC. Norman Breslow is Professor, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195. Part of this research was Chatterjee's doctoral thesis work at the Department of Statistics, University of Washington, Seattle. The research of Chatterjee and Breslow was supported by National Institutes of Health grant R01-CA40644; that of Chen, by the National Science Council of Taiwan, ROC. The authors especially thank the associate editor for input on robustness issues of the proposed method.

© 2003 American Statistical Association
Journal of the American Statistical Association
March 2003, Vol. 98, No. 461, Theory and Methods
DOI 10.1198/016214503388619184

for sample surveys, and Carroll et al. (1995, sec. 9) for measurement error problems. Finally, some authors have used the fact that when G is known, the likelihood contribution for a subject with missing X is obtained by integrating the usual likelihood using $dG(x|Z)$. An estimated likelihood or, equivalently, an estimated score is obtained by substituting an empirical version of G (Pepe and Fleming 1991; Carroll and Wand 1991; Hu and Lawless 1997). Breslow and Chatterjee (1999) and Lawless, Kalbfleisch and Wild (1999) evaluated the efficiency of various pseudo-likelihood methods relative to the fully efficient SPMLE methods when both Y and Z are discrete.

We propose a new type of estimated likelihood score, the *pseudoscore* (PS), which uses the postulated parametric regression model to provide a smoother, consistent estimate of G and consequently improves the efficiency of estimation of β . When Z is discrete (which is assumed in much of what follows), the new method is computationally simple for both discrete and continuous outcomes. An interesting feature of the PS estimator is that it may be applied in restricted situations where other estimators cannot. Most formulations of the two-phase sampling problem include the requirement that each individual selected at phase one has a positive probability of being sampled at phase two. Practical considerations, however, may preclude obtaining validation samples for subjects with certain outcomes. In case-control studies, for example, if invasive medical tests are required to accurately measure the covariate, then controls may be unwilling to participate at the validation stage.

The PS estimator may be applied with such restricted designs, provided that subjects with any possible combination of (X, Z) values have a positive probability of representation at phase two. Other semiparametric methods of which we are aware cannot be applied. Of course, the lack of validation subjects with certain Y values means that there is only limited facility for model checking using study data alone. Thus, although restricted designs should ordinarily be avoided, the PS estimator provides the investigator willing to make the necessary model assumptions with the flexibility to use them when necessary.

Section 2 presents the new system of estimating equations and an iteratively reweighted regression algorithm that may be used to solve them using standard software packages. It also gives the asymptotic properties of the estimator. Section 3 reports the results of simulation experiments that evaluate the small-sample behavior of the proposed method, compare its efficiency with alternative methods, and study its robustness under model misspecification. Section 4 demonstrates the new method using data from a case-control study of leprosy. The final section discusses the advantages and limitations of the PS estimator and mentions some possible extensions.

2. METHODS

2.1 The Pseudoscore Function

We start with the assumption that Z , the vector of phase-one covariates known for everyone, is discrete. Although the PS function is well defined more generally, its estimation for continuous Z would require nonparametric regression methods. This generalization will be pursued elsewhere.

The joint probability law of the data vector (R, Y, X, Z) for the just-described problem is governed by the four parameters (β, π, G, H) , some of which may be infinite dimensional. We denote the true values by $(\beta_0, \pi_0, G_0, H_0)$, the corresponding probability law by P_0 , and expectation under P_0 by E_0 . We assume that in a neighborhood of the true parameter values (β_0, π_0) ,

$$(A) \int \pi(y, Z) dy > 0 \quad \text{and} \quad (B) \quad f_\beta(Y|X, Z) > 0 \quad (3)$$

almost surely for all (Y, X, Z) in the sample space. Condition A is weaker than requiring $\pi(y, z) > 0$ everywhere. It ensures that for each value of z , $\pi(y, z) > 0$ for at least one value of y if y is categorical and in an interval of y if y is continuous. This, together with condition B, further ensures that

$$q_\beta^\pi(X, Z) \equiv P(R=1|X, Z) = \int \pi(y, Z) f_\beta(y|X, Z) dy > 0 \quad (4)$$

almost surely.

For fixed G , the (conditional) likelihood of the observable data is proportional to

$$L(\beta; G) = \prod_{i \in V} f_\beta(Y_i|X_i, Z_i) \prod_{j \in \bar{V}} \int f_\beta(Y_j|x, Z_j) dG(x|Z_j), \quad (5)$$

where $V = \{i : R_i = 1\}$ denotes the validation or phase-two sample. Assuming that the scores [see (2)] and the integrals to follow all exist, the score function is

$$S(\beta; G) = \frac{\partial \log L(\beta; G)}{\partial \beta} = \sum_{i \in V} S_\beta(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} \frac{\int S_\beta(Y_j|x, Z_j) f_\beta(Y_j|x, Z_j) dG(x|Z_j)}{\int f_\beta(Y_j|x, Z_j) dG(x|Z_j)}. \quad (6)$$

An estimate of $G(\cdot|z)$ can be substituted into (6) to estimate the score function. Such an estimate generally cannot be obtained directly from the validation data, however, due to the biased sampling. What one can obtain directly are estimates of the conditional distributions $P(X|z, R=1)$, hereafter denoted $G^*(\cdot|z)$. Specifically, the empirical estimate

$$G_N(x|z) = \frac{\sum_i I_{[X_i \leq x, Z_i = z, R_i = 1]}}{\sum_i I_{[Z_i = z, R_i = 1]}}, \quad (7)$$

where I_A denotes the indicator function of the event A , is consistent for $G_0^*(x|z)$. When the selection probabilities π depend only on z (Pepe and Fleming 1991), or when $f_{\beta_0}(y|x, z)$ is free of x , $G_0^* = G_0$. Otherwise, $G_0^* \neq G_0$, and naively substituting G_N for G will produce a biased estimate of the score function. Some modifications of (7) are therefore needed to accommodate the general situation.

From Bayes's theorem, when $P(R=1|X, Z) > 0$ almost surely,

$$dG(x|Z) = \frac{dP(X \leq x|Z, R=1)P(R=1|Z)}{P(R=1|X=x, Z)}. \quad (8)$$

Substituting the right side of (8) for dG in (6) yields the PS function

$$S_{PS}(\beta; G^*, \pi) = \sum_{i \in V} S_{\beta}(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} \frac{\int S_{\beta}(Y_j|x, Z_j) h_{\beta}^{\pi}(Y_j, x, Z_j) dG^*(x|Z_j)}{\int h_{\beta}^{\pi}(Y_j, x, Z_j) dG^*(x|Z_j)}, \quad (9)$$

where

$$h_{\beta}^{\pi}(y, x, z) = \frac{f_{\beta}(y|x, z)}{q_{\beta}^{\pi}(x, z)}.$$

Although G^* implicitly depends on β and π , we fix $G^* = G_0^*$ and propose estimating G_0^* directly using G_N . Unbiasedness of the PS function follows, because at the true parameter values ($\beta = \beta_0, G^* = G_0^*, \pi = \pi_0$), it equals the likelihood score. It is not a true score function, however, in the sense that there is no associated log-likelihood or log-pseudolikelihood whose β derivative is given by (9). This follows from the classical theory of the calculus because, as discussed later, $\partial S_{PS}(\beta; G^*, \pi)/\partial \beta^T$ is in general an asymmetric matrix. For the PS function to be unbiased, the full distribution of $[Y|X, Z]$ must be specified correctly through the model $f_{\beta}(Y|X, Z)$. Whereas for most standard regression models, the score function $S_{\beta}(Y|X, Z)$ involves only a few lower-order moments of the distribution of $[Y|X, Z]$, computation of $h_{\beta}^{\pi}(y, x, z)$ in the PS function involves the full density function $f_{\beta}(Y|X, Z)$. This could be a particularly important issue for continuous outcomes, because different choices of $f_{\beta}(Y|X, Z)$ may give rise to the same lower-order moments. In limited simulations (see sec. 3.3), however, we found that the PS method was quite robust for regression inference on lower-order moments even when the full distribution was moderately misspecified.

2.2 Estimation

By substituting G_N for G^* in expression (9), we obtain the estimating function

$$S_{PS}(\beta; G_N, \pi) = \sum_{i \in V} S_{\beta}(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} \sum_{i \in V} \frac{S_{\beta}(Y_j|X_i, Z_j) h_{\beta}^{\pi}(Y_j, X_i, Z_j) I(Z_j = Z_i)}{\sum_{i \in V} h_{\beta}^{\pi}(Y_j, X_i, Z_j) I(Z_j = Z_i)}. \quad (10)$$

Provided that $\pi = \pi_0$ is known (as it generally will be in a two-phase study), one can estimate β by solving the equations $S_{PS}(\beta; G_N, \pi_0) = 0$. Whether or not π_0 is known, however, we propose to replace it by a consistent estimate $\hat{\pi}$ from a correct model, because this results in more efficient estimates of the regression parameters (Pierce 1982; Robins et al. 1994). It is common practice to model $P(R = 1|Y, Z) = \pi(Y, Z)$ parametrically (by, e.g., logistic regression) and to estimate the corresponding regression parameters from the data on (R, Y, Z) . When both Y and Z are discrete, one can use a saturated model for π , in which case the $\hat{\pi}$ are the observed sampling fractions.

The estimating equations defined by $S_{PS}(\beta; G_N, \hat{\pi}) = 0$ can be solved by a standard Newton–Raphson algorithm. But the form of (10) immediately suggests the following iterated

reweighting algorithm, which gives better insight into the new approach:

1. Start with an initial estimate β^0 and call it the current value β_c .
2. Use the units in the validation sample V as they are. For each $j \in \bar{V}$, construct a set of filled-in data, $\{(Y_j, X_i, Z_j), i \in V_{Z_j}\}$, where V_{Z_j} denotes the subset of validation units with $Z = Z_j$.
3. For each filled in observation $(Y_j, X_i, Z_j)_{j \in \bar{V}, i \in V_{Z_j}}$ calculate an associated weight $w_{ji}(\beta_c)$ defined by

$$w_{ji}(\beta_c) = \frac{h_{\beta_c}^{\hat{\pi}}(Y_j, X_i, Z_j)}{\sum_{\ell \in V_{Z_j}} h_{\beta_c}^{\hat{\pi}}(Y_j, X_{\ell}, Z_j)}.$$

4. Obtain a new current estimate β_c by fitting the parametric regression model to the combined set of data (i.e., the original data for the validation sample and the filled-in data for the nonvalidation sample), using a program that allows for prior weights. Assign weights of unity to the validation data and weights as defined earlier to the filled-in data. For popular regression models, such as logistic or linear regression, standard software can be used at this step.
5. Repeat steps 3 and 4 until convergence.

It is easy to see that if the algorithm converges, it converges to a solution of $S_{PS}(\beta; G_N, \hat{\pi}) = 0$. Proposition 1 in Section 2.5 states that under certain regularity conditions, the PS estimating equations have a unique consistent sequence of solutions. Chatterjee (1999) showed that under additional regularity conditions and starting from a known consistent estimate, the iterated reweighted algorithm converges to this consistent solution. In simulations we have found that, except for some nonidentifiable situations that arise with restricted designs, this algorithm always converges to a unique solution irrespective of the starting value. Proof of such global convergence is not available, however. In contrast, the Newton–Raphson algorithm occasionally fails to converge, even with good starting values and appropriate scaling.

Regardless of which algorithm is used, computation of the quantities $q_{\beta}^{\pi}(X_i, Z_j)$ is needed at each iteration. When Y is categorical, integration is simply replaced by summation in (4). When Y is continuous, evaluation of the integrals may involve some work, depending on the form of $\pi(y, z)$ and $f_{\beta}(y|x, z)$. For the two-phase stratified design, however, the task may be simplified. For example, when Y is continuous and univariate, the phase-two sampling typically stratifies on class intervals of Y . Suppose that for $Z = z$, the range of Y is partitioned into the disjoint intervals $\{I_1(z), \dots, I_M(z)\}$ and that $\pi(y, z) = \pi_m(z)$ if $y \in I_m(z)$, $m = 1, \dots, M$. Then

$$\int \pi(y, z) f_{\beta}(y|x, z) dy = \sum_{m=1}^M \pi_m(z) [F_{\beta}\{b_m(z)|x, z\} - F_{\beta}\{a_m(z)|x, z\}],$$

where $b_m(z)$ and $a_m(z)$ are the upper and lower endpoints of $I_m(z)$ and $F_{\beta}(y|x, z)$ is the cumulative distribution function corresponding to $f_{\beta}(y|x, z)$. Thus the PS method is easy to implement for two-phase stratified samples involving either continuous or discrete outcomes.

2.3 Comparison With the Horvitz–Thompson Estimator of G

Comparison of the proposed PS approach with two closely related methods provides further insight. One alternative is to directly estimate the scores (6) by substituting for G the weighted (Horvitz–Thompson) empirical distribution function

$$G_N^{\text{HT}}(\pi_0) = \frac{\sum_{i \in V} I_{[X_i \leq x, Z_i = z]} / \pi_0(Y_i, Z_i)}{\sum_{i \in V} I_{[Z_i = z]} / \pi_0(Y_i, Z_i)}. \quad (11)$$

Horvitz–Thompson estimators are common in survey data analysis, having been used to estimate a distribution function (Rao et al. 1990), for example. When the sampling weights are highly variable, however—as will be the case for an efficient two-phase design—the weighted estimator typically has a large variance (Pfeffermann 1996; Korn and Graubard 1999, sec. 4.4). Simulation studies reported herein show that imprecise estimation of the nuisance parameter G using this approach can cause serious loss of efficiency in estimation of β . The estimating function that we propose in (10) can also be viewed as an estimate of the score function (6) obtained by substituting a weighted empirical estimator for G . It exploits the regression model $f_\beta(y|x, z)$, however, to define a more efficient set of weights. Because $h_\beta^\pi(Y, X, Z) = f_\beta(y|x, z)/q_\beta^\pi(X, Z)$ in (10), $1/q_\beta^\pi(X, Z)$ can be viewed as a new set of inverse probability weights for estimating $G(X|Z)$ from the validation data for use in (6). Furthermore, because $q_\beta^\pi(X, Z) = P(R = 1|X, Z) = E\{\pi(Y, Z)|X, Z\}$, one would expect the new set of weights to be less variable, and hence more efficient, than the Horvitz–Thompson weights $1/\pi(Y, Z)$.

2.4 Comparison With the Mean Score Estimator

The contribution of nonvalidation subjects to the score (6) is $E\{S_\beta(Y|X, Z)|Y, Z\}$. For the situation where both Y and Z are discrete, Reilly and Pepe (1995) proposed estimating $E\{S_\beta(Y|X, Z)|Y, Z\}$ for an incomplete unit by $\int S_\beta(Y|x, Z) dP_N(x|Y, Z)$, where $P_N(\cdot|Y, Z)$ is the empirical distribution of $[X|Y, Z]$ in the validation sample. Their purely empirical “mean score” approach is valid because $[X|Y, Z] = [X|Y, Z, R = 1]$. However, it ignores the fact that the distribution of $[X|Z, Y, R = 1]$ is partially determined by the parametric model. More explicitly,

$$dP(X|Y, Z, R = 1) = \frac{dP(Y|X, Z, R = 1) dG^*(x|Z)}{\int dP(Y|x, Z, R = 1) dG^*(x|Z)},$$

where $dP(Y|X, Z, R = 1)$ is related to the regression model and the selection probabilities by the formula

$$dP(Y|X, Z, R = 1) = \frac{\pi(Y, Z) f_\beta(Y|X, Z)}{\int \pi(y, Z) f_\beta(y|X, Z) dy} \equiv f_\beta^\pi(Y|X, Z).$$

Thus one would expect to gain efficiency by estimating only $P(X|Z, R = 1)$ empirically, determining the remainder of $P(X|Y, Z, R = 1)$ from the model. This is precisely what the PS (9) accomplishes. Note that $\pi(y, z) h_\beta^\pi(y, x, z) = f_\beta^\pi(y|x, z)$ and that $\pi(y, z)$ may be inserted in both numerator and denominator of the contributions of nonvalidation subjects without affecting (9). The density of $[Y|X, Z]$ in the

validation sample, $f_\beta^\pi(y|x, z)$, defines the “complete-data likelihood” used elsewhere as a basis for estimation (Breslow and Cain 1988; Krieger and Pfeffermann 1992; Carroll et al. 1995, sec. 9).

2.5 Asymptotic Properties

Some additional notation will be useful to describe the asymptotic properties of the new estimator. Define $\Psi_N(\beta; G_0^*, \pi_0) = S_{\text{PS}}(\beta; G_0^*, \pi_0)/N$, $\Psi(\beta; G_0^*, \pi_0) = E_0 \Psi_N(\beta; G_0^*, \pi_0)$, $\Psi_\beta(\beta, G_0^*, \pi_0) = \partial \Psi(\beta; G_0^*, \pi_0) / \partial \beta$, $S_{\beta_0, G_0}(y|z) = E_0\{S_{\beta_0}(y|X, z)|y, z\}$, and $D(y, x, z) = S_{\beta_0}(y|x, z) - S_{\beta_0, G_0}(y|z)$. Assuming sufficient regularity to allow interchange of expectation and differentiation, some calculation gives $\Psi_\beta(\beta_0, G_0^*, \pi_0) = -\mathcal{J}_\beta(\beta_0, G_0) - C_\beta$, where $\mathcal{J}_\beta(\beta, G) = -\partial E_0 S(\beta; G) / \partial \beta$ denotes the expected Fisher information for the true likelihood and

$$C_\beta = E_0\{[1 - \pi_0(Y, Z)] \tilde{C}_\beta(Y, Z)\}, \quad (12)$$

with

$$\tilde{C}_\beta(Y, Z) = \text{cov}_0 \left\{ S_{\beta_0}(Y|X, Z), \frac{\partial \log q_{\beta_0}^{\pi_0}(X, Z)}{\partial \beta} \Big| Y, Z \right\}.$$

Note that $\Psi_\beta(\beta_0, G_0^*, \pi_0)$ may not be symmetric due to the presence of the covariance term.

When selection probabilities are estimated from the data, we assume that they are estimated using a parametric regression model. Let \mathcal{A} denote the set of (y, z) such that $\pi(y, z) > 0$. Given $(y, z) \in \mathcal{A}$, suppose that we have a regression model $E(R|y, z) = \pi(y, z; \alpha)$, which we abbreviate as $\pi(\alpha)$. The maximum likelihood estimator, $\hat{\alpha}$, satisfies the score equations $\sum_{i=1}^N S_\alpha(R_i|Y_i, Z_i) = 0$, where

$$S_\alpha(R|Y, Z) = \frac{\partial \pi(\alpha)}{\partial \alpha} \frac{I_{\mathcal{A}}(Y, Z)}{\pi(\alpha)\{1 - \pi(\alpha)\}} \{R - \pi(\alpha)\}, \quad (13)$$

and the corresponding information matrix is

$$\mathcal{J}_\alpha(\alpha) = E_0 \left[\left\{ \frac{\partial \pi(\alpha)}{\partial \alpha} \right\}^2 \frac{I_{\mathcal{A}}(Y, Z)}{\pi(\alpha)\{1 - \pi(\alpha)\}} \right]. \quad (14)$$

Let $\hat{\pi} = \pi(\hat{\alpha})$. When Y and Z are discrete and the model is saturated, $\hat{\pi}$ equals the observed sampling fractions. If we define $\Psi_\alpha(\beta, G^*, \alpha) = \partial \Psi\{\beta; G^*, \pi(\alpha)\} / \partial \alpha$ and α_0 the true value of α , then it can be shown that

$$\Psi_\alpha \equiv \Psi_\alpha(\beta_0, G_0^*, \alpha_0) = -E_0\{[1 - \pi_0(Y, Z)] \tilde{C}_\alpha(Y, Z)\}, \quad (15)$$

where

$$\tilde{C}_\alpha(Y, Z) = \text{cov}_0 \left\{ S_{\beta_0}(Y|X, Z), \frac{\partial \log q_{\beta_0}^{\pi(\alpha_0)}(X, Z)}{\partial \alpha} \Big| Y, Z \right\}.$$

Proposition 1. Under regularity conditions A0–A5 listed in the Appendix, the following results hold:

a. The estimating equations $S_{\text{PS}}\{\beta; G_N^*, \hat{\pi}\} = 0$ have a unique, consistent sequence of solutions, $\{\hat{\beta}_N^{\text{PS}}\}_{N \geq 1}$.

b.

$$\sqrt{N}(\hat{\beta}_N^{PS} - \beta_0) = -\Psi_\beta^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \Phi(R_i, Y_i, X_i, Z_i) + o_p(1),$$

where

$$\Phi(R, Y, X, Z) = a_0(R, Y, X, Z) + a_1(R, X, Z) + a_2(R, Y, Z),$$

with

$$\begin{aligned} a_0(R, Y, X, Z) &= RS_{\beta_0}(Y|X, Z) + (1 - R)S_{\beta_0, G_0}(Y|Z), \\ a_1(R, X, Z) &= RE_0 \left[\frac{1 - \pi_0(Y, Z)}{q_{\beta_0}^{\pi_0}(X, Z)} D(Y, X, Z) | X, Z \right], \end{aligned} \tag{16}$$

and

$$a_2(R, Y, Z) = -\Psi_\alpha J_\alpha^{-1} S_{\alpha_0}(R|Y, Z), \quad J_\alpha = J_\alpha(\alpha_0).$$

c. If $\text{var}_0 \Phi < \infty$, then $\sqrt{N}(\hat{\beta}_N^{PS} - \beta_0) \rightarrow N(0, \Omega)$ in distribution, where

$$\Omega = (J_\beta + C_\beta)^{-1} (J_\beta + \Sigma) (J_\beta^T + C_\beta^T)^{-1}$$

and

$$\Sigma = \text{var}_0\{a_1(R, X, Z)\} + C_\beta + C_\beta^T - \Psi_\alpha J_\alpha^{-1} \Psi_\alpha^T.$$

Here $a_0(R, Y, X, Z)$ is the likelihood score for a single observation for fixed $G = G_0$, $a_1(R, X, Z)$ adjusts for estimation of G^* and is nonzero only for units with complete data ($R = 1$), and $a_2(R, Y, Z)$ adjusts for estimation of π . Because it contributes the last term in the expression for Σ , which is a nonnegative definite matrix, estimation of π improves efficiency even when $\pi = \pi_0$ is known. An outline of the proof of the proposition is given in the Appendix under the assumption that Z , Y , and X are discrete. A more complete argument that relaxes these conditions has been given in Chatterjee's (1999) dissertation.

A "plug-in" approach can be used to estimate Ω . First, Σ can be consistently estimated by

$$\hat{\Sigma} = \widehat{\text{var}}\{\hat{a}_1(R, X, Z)\} + \hat{C}_\beta + \hat{C}_\beta^T - \hat{\Psi}_\alpha \hat{J}_\alpha^{-1} \hat{\Psi}_\alpha^T,$$

where $\widehat{\text{var}}$ denotes the empirical variance. The estimate $\hat{a}_1(R, X, Z)$ is obtained by plugging in estimates for the true parameter values and estimating $S_{\beta_0, G_0}(y|z) = E_0\{S_{\beta_0}(y|X, z)|y, z\}$ by taking the expectation of $S_{\beta_0}(y|X, z)$ with respect to the conditional density

$$d\hat{P}(X|y, z) = \frac{h_\beta^{\hat{\pi}}(y, X, z) dG_N(X|z)}{\int h_\beta^{\hat{\pi}}(y, x, z) dG_N(x|z)}.$$

Similarly, \hat{C}_β , \hat{J}_α , and $\hat{\Psi}_\alpha$ are obtained from (12), (14), and (15), by plugging in the estimates for the true parameter values in the corresponding formulas, estimating the covariance term in these expressions with respect to the conditional distribution $\hat{P}(X|y, z)$, and estimating the expectations with respect to the

distribution of (Y, Z) by the corresponding empirical versions. Further, \hat{J}_β can be estimated as

$$\hat{J}_\beta = \frac{1}{N} \sum_{i \in V} I_\beta(Y_i|X_i, Z_i) + \frac{1}{N} \sum_{j \in \bar{V}} \hat{E}\{I_\beta(Y_j|X, Z_j)|Y_j, Z_j\},$$

where $I_\beta(Y|X, Z) = -\partial S_\beta(Y|X, Z)/\partial \beta$ and \hat{E} denotes the expectation with respect to the conditional distribution $\hat{P}(X|Y, Z)$.

3. SIMULATION STUDIES

3.1 Finite-Sample Performance

The finite-sample performance of the proposed estimator of β was investigated by simulation. Two models for $f_\beta(Y|X, Z)$, logistic and linear, were considered. For the logistic model, the 0–1 outcome variable Y was generated from logit $P(Y = 1|X, Z) = \beta_0 + \beta_1 X$ with logit $p = \log\{p/(1 - p)\}$. Here X is standard normal and Z is a surrogate for X such that $Z = I(X + \epsilon > 0)$ with $\epsilon \sim N(0, 1)$ independent of X and Y . For the linear model, the continuous outcome Y was generated from $Y = \beta_0 + \beta_1 X + \eta$, with η standard normal. X and its surrogate Z were generated as for the logistic model. Let $\tilde{Y} = Y$ for the logistic model case and $\tilde{Y} = I(Y > 1)$ for the linear model case. Several designs for validation sampling were considered by varying the selection probabilities, $\pi = \{\pi(Y, Z)\} = \{\pi(\tilde{Y}, Z)\} = \{\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)\}$. These variations include (a) simple random sampling with $\pi(\tilde{Y}, Z)$ constant, (b) stratified sampling with $\pi(\tilde{Y}, Z) > 0$ for each (\tilde{Y}, Z) , and (c) restricted sampling with $\pi(1, Z) = 0$ but $\pi(0, Z) > 0$. To implement the PS method, we used the observed sampling fractions as estimates of the selection probabilities. For all sampling designs, the total sample size was $N = 300$. Results based on 500 replications for various values of β and π are displayed in Table 1. Overall, the simulation means of $\hat{\beta}$ were close to their true values, means of estimated variances were close to the respective empirical variances, and the observed coverage probabilities for 95% confidence intervals based on the estimates and the estimated standard errors were close to the nominal value of .95. As would be expected for a \sqrt{n} -consistent estimator, the biases of the PS estimates were always of smaller order than the respective standard errors.

3.2 Efficiency Comparisons

Breslow and Holubkov (1997), Scott and Wild (1997), and Lawless et al. (1999) studied the semiparametric maximum likelihood estimator (SPMLE) in situations where the phase-one data, including both covariates and outcomes, are discrete and there is a positive probability of sampling from each stratum. Breslow, McNeney, and Wellner (2003) showed that the SPMLE is asymptotically efficient and equivalent to the estimate that solves the efficient score equations (Robins et al. 1995). Lawless et al. (1999) conducted a series of simulation experiments in this setting of discrete phase-one data to compare the performance of various pseudolikelihood methods relative to the SPMLE. We used their simulation setups to compare the efficiency of the proposed PS method with the SPMLE and other pseudolikelihood methods for both the "surrogate covariate" and "expensive covariate" problems. For

Table 1. Finite-Sample Properties of $\hat{\beta}$

π	$bias \times 10^2$		$var \times 10^2$		$\widehat{var} \times 10^2$		95% CP		$bias \times 10^2$		$var \times 10^2$		$\widehat{var} \times 10^2$		95% CP	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
<i>Logistic model: logit P = $\beta_0 + \beta_1 X$</i>																
	$\beta_0 = -1, \beta_1 = 0$								$\beta_0 = -1, \beta_1 = 0.7$							
(.2,.2,.2,.2)	-.982	1.01	2.10	5.06	1.87	4.47	93%	95%	-3.65	6.28	3.19	8.27	3.18	7.67	95%	94%
(.2,.4,.2,.4)	-.862	.745	2.05	3.87	1.81	3.63	93%	95%	-3.25	4.93	3.02	6.23	2.82	5.48	95%	95%
(.2,.2,.7,.7)	-.117	.010	2.03	3.59	1.76	3.02	93%	94%	-1.37	2.89	2.42	5.14	2.17	4.42	94%	94%
(.1,.3,.5,.7)	-.226	-.187	2.02	3.57	1.77	3.19	93%	94%	-1.48	3.25	2.64	5.19	2.29	4.35	94%	93%
(.3,.8,0,0)	-2.47	1.03	2.24	6.02	2.09	6.09	94%	96%	-6.76	5.19	8.32	11.1	11.0	11.9	95%	94%
(.1,.5,1,1)	-.077	-.367	2.02	2.71	1.74	2.58	93%	95%	-1.06	1.75	2.33	3.34	2.04	3.03	94%	95%
<i>Linear model: Y = $\beta_0 + \beta_1 X + \eta$, $\eta \sim N(0, 1)$</i>																
	$\beta_0 = 0, \beta_1 = 0.5$								$\beta_0 = 0, \beta_1 = 1$							
(.2,.2,.2,.2)	.514	-.191	.577	.821	.552	.805	95%	94%	.305	.327	1.07	1.12	1.05	1.02	94%	93%
(.2,.4,.2,.4)	.428	-.017	.480	.662	.483	.664	96%	96%	.237	.408	.807	.823	.796	.775	95%	95%
(.2,.2,.7,.7)	.491	-.203	.486	.639	.475	.658	96%	95%	-.228	.018	.740	.778	.754	.757	95%	95%
(.1,.3,.5,.7)	.022	.273	.524	.729	.525	.706	95%	94%	-1.07	1.05	.977	1.11	1.02	1.06	95%	93%
(.3,.8,0,0)	.593	-.221	.409	.605	.439	.590	96%	95%	1.98	1.54	.663	.833	.668	.665	94%	93%
(.1,.5,1,1)	.175	.286	.468	.536	.457	.542	94%	95%	-.685	.951	.812	.928	.807	.846	94%	94%

NOTE: Total sample size $N = 300$, $X \sim N(0, 1)$, and $Z = I(X + \epsilon > 0)$ with $\epsilon \sim N(0, 1)$. $\pi = \{\pi(\tilde{Y}, Z)\} = \{\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)\}$, where $\tilde{Y} = Y$ for the logistic model case and $\tilde{Y} = I(Y > 1)$ for the linear model case. Here \widehat{var} refers to the simulation mean of the estimated variance of $\hat{\beta}$, and CP is the coverage probability, all based on 500 replications.

the surrogate covariate problem (Table 2), we first generated a covariate $X \sim N(0, 1)$ and an outcome Y from the logistic regression model $\text{logit } P(Y = 1|X) = \alpha + \beta X$. For each value of β , we chose α such that the marginal probability of observing $Y = 1$ was .05. We then generated $W \sim N(0, 1)$ such that $\text{cor}(X, W) = .9$ and obtained Z , a surrogate for X , by collapsing W into six levels, with the levels defined by the corresponding hexiles. The phase-two sample was selected in two ways: all cases ($Y = 1$) and 5% of the controls ($Y = 0$), or 20% of the cases and 1% of the controls. The total sample sizes (n) for these two cases were $n = 9,000$ and $n = 45,000$, so that in both situations the size of the phase-two sample was approximately 900, with balanced numbers of cases and controls.

The same setup was used for the expensive covariate problem, except that $\text{cor}(X, W) = .3$ and Z , obtained by discretizing W into six levels, was itself considered a covariate of interest. The corresponding logistic regression model was $\text{logit } P(Y = 1|X, Z) = \alpha + \beta X + \gamma Z$, with γ fixed at .5. We implemented the PS method with both known (PSI) and estimated

(PSII) selection probabilities. Lawless et al. (1999) had found an “estimated pseudolikelihood” approach, equivalent to estimating the scores (6) using the weighted empirical covariate distribution (11), to be the “most promising” of the pseudolikelihood methods. We also implemented this method with both known (ELI) and estimated (ELII) selection probabilities.

Several conclusions can be drawn from the results given in Tables 2 and 3. First, for both models and for all parameters, the PS methods, particularly PSII, had remarkably high efficiency and often achieved the same efficiency as the SPMLE. Second, in all cases the PS method outperformed the estimated pseudolikelihood method, particularly when the regression effect was strong. Third, the efficiency of the PS method slowly declined as the regression effect increased and the sampling fraction decreased. Finally, the use of estimated

Table 2. Efficiencies of $\hat{\alpha}$ and $\hat{\beta}$ Relative to Maximum Likelihood for the Six-Level Fine Surrogate in Logistic Regression

β	$n = 9,000$ $\pi(1, z) = 1.0, \pi(0, z) = .05$				$n = 45,000$ $\pi(1, z) = .2, \pi(0, z) = .01$			
	ELI	ELII	PSI	PSII	ELI	ELII	PSI	PSII
$\hat{\alpha}$								
0	100	100	100	100	100	100	100	100
.5	99	99	100	100	96	96	100	100
1.0	94	93	100	100	87	86	100	100
1.5	83	82	99	99	70	71	95	99
$\hat{\beta}$								
0	92	92	100	100	99	99	100	100
.5	80	80	99	100	78	78	100	100
1.0	69	68	98	99	55	54	96	96
1.5	67	67	94	98	48	50	85	95

Table 3. Efficiencies of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ Relative to ML for the Expensive Covariate Data Problem in Logistic Regression

β	$n = 9,000$ $\pi(1, z) = 1.0, \pi(0, z) = .05$				$n = 45,000$ $\pi(1, z) = .2, \pi(0, z) = .01$			
	ELI	ELII	PSI	PSII	ELI	ELII	PSI	PSII
$\hat{\alpha}$								
0	100	100	100	100	105	105	105	105
.5	97	96	100	100	87	87	103	103
1.0	87	87	99	100	77	78	95	99
1.5	79	80	93	99	72	72	93	98
$\hat{\gamma}$								
0	98	97	100	100	102	102	107	107
.5	95	95	100	100	82	82	99	101
1.0	84	84	99	100	62	63	91	99
1.5	77	77	94	98	54	55	85	99
$\hat{\beta}$								
0	92	91	101	100	100	100	109	108
.5	85	84	98	100	86	85	100	102
1.0	81	81	93	100	77	77	82	94
1.5	75	76	80	99	70	71	66	89

Table 4. Efficiencies of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}$ Relative to PSI for the Surrogate Covariate Problem in the Linear Regression Model

	β	WL	ELI	SPMLE (reduced data)
$\hat{\alpha}$	0.0	2	100	≈ 29
	0.5	5	57	≈ 63
	1.0	20	72	≈ 167
$\hat{\beta}$	0.0	3	98	≈ 33
	0.5	10	59	≈ 83
	1.0	38	73	≈ 211
$\hat{\sigma}$	0.0	4	100	≈ 80
	0.5	5	93	≈ 83
	1.0	13	85	≈ 163

selection probabilities (observed sampling fractions) improved the efficiency of the PS method.

In a second experiment, we simulated data from the continuous outcome setup described by Lawless et al. (1999). We generated X from a $N(0, 1)$ distribution and Y from the linear regression model $Y = \alpha + \beta X + \epsilon$, $\epsilon \sim N(0, \sigma^2)$. We generated Z following the same setup of Table 2. To select the phase-two sample, we defined three strata— $Y < C_1$, $C_1 \leq Y < C_2$ and $Y \geq C_2$ —so that $P(Y < C_1) = P(Y > C_2) = .05$; 25% of the units in the tail strata and only 1.4% in the middle stratum were selected. To implement the SPMLE, Lawless et al. (1999) defined six strata, splitting each of the three strata used to select the sample, and assumed that only stratum information on the outcome was available at phase one. We implemented ELI and PSI for this problem. Because both methods easily handle continuous outcome data, however, we used the exact outcome information at both phases. We also implemented a weighted likelihood (WL) approach, performing a weighted regression analysis on the units selected at phase two and using the inverses of the known selection probabilities as weights. Because Lawless et al. (1999) reported the efficiency of this complete-case method relative to their SPMLE on the reduced data, estimation of the efficiency of WL relative to PSI enabled us to approximate the efficiency of their SPMLE relative to PSI.

The results are reported in Table 4. PSI estimated nonnull regression effects more efficiently than did ELI. The relative efficiency of the SPMLE for the reduced data, computed as the ratio of the relative efficiency of WL to PSI and that of WL to SPMLE as reported by Lawless et al. (1999), was severely low for both the intercept and the slope parameter when there was no association between Y and X . This suggests that discretizing the outcome information for the sake of implementing SPMLE may result in substantial loss of power when the true regression relationship is weak or moderate. However, for $\beta = 1$, which represents a very strong regression relationship between Y and X , the SPMLE was substantially more efficient than the PS estimator, even though it used only the reduced data.

3.3 Robustness and Model Checking

In this section we study robustness of the PS estimator when the underlying regression model is misspecified. The issue of robustness becomes particularly important for case-only and control-only and other restricted designs because they offer only limited scope for checking model assumptions.

The simplest method for analyzing data from two-phase stratified designs is the standard survey approach based on WL. Although this method is inefficient when the assumed model is correct, sometimes seriously so (Table 4), its appeal has been a type of robustness. Even if the model $f_{\beta}(y|x)$ is incorrect, the WL estimator is consistent for a population parameter of interest. Manski and Thompson (1989) showed that for a binary regression model $P(Y = 1|X = x) = p_{\beta}(x)$, the WL estimate is consistent for the parameter value B that minimizes the expectation of the loss function $-\log\{1 - |Y - p(X)|\}$ within the class of all predictors of the form $p(x) = p_{\beta}(x)$. Scott and Wild (1986) suggested that the robustness of other estimators could be assessed by examining their bias as estimators of B under misspecification.

Following Scott and Wild (1986), we generated binary outcome data using the logistic regression model $\text{logit } P(Y = 1|X = x) = \beta_0 + \beta_1 x + \delta x^2$ but assumed the working model $\text{logit } P(Y = 1|X = x) = \beta_0 + \beta_1 x$ for analysis. Data on X were generated from the standard normal distribution. We considered two types of surrogates Z for the data available at phase one: a two-level crude surrogate generated using the model described in Section 3.1 and a six-level fine surrogate generated using the model described in Section 3.2. We considered three alternative phase-two designs: case-only, control-only, and case-control as described in Table 5. For two different underlying true models, Table 5 shows the bias, mean squared error (MSE), and the 95% confidence interval (CI) coverage probability of $\hat{\beta}_1$ as an estimate of B_1 , the almost sure limit of the maximum likelihood estimates for β_1 in the working model based on the complete simulated data.

Several interesting observations can be made from Table 5. First, the PS estimator overestimated B_1 irrespective of the design. Second, for each design, the bias of the PS estimator was substantially lower when the phase-one data consisted of the fine surrogate. Third, the bias in estimation of B_1 was smallest for the control-only design and largest for the case-only design. For the case-only design with a crude surrogate at phase one, the bias in the PS estimator of β_1 was as great as 60% of the magnitude of B_1 (see model b). Thus we see that under gross violation of the linearity assumptions, the bias in the PS estimator with the case-only design can be unacceptably high. Finally, for the case-control design, although the WL method had minimal bias, it had large variability. As a result, except for the situations where the PS estimator was severely biased, the MSEs for the WL estimator were often significantly higher than those of the PS estimates.

When the case-control design is used at phase two, departure from linearity can be tested in the usual fashion by including quadratic or other higher-order terms in the model and assessing the significance of the corresponding regression coefficients. For restricted designs, however, there may be only limited scope for such model checking. Consider, for example, the case-only design. Clearly, the regression relationship between Y and X is not identifiable from the phase-two data alone. Thus the phase-one data are needed both to correct for sampling bias in selection of the phase-two sample and to identify the regression relationship. Intuitively, for estimating a quadratic trend, at least three distinct values of Z are required. We tested this assertion using the simulation setup

Table 5. Simulation Results Based On 500 Replications

	Crude surrogate				Fine surrogate			
	PS			WL	PS			WL
	Case-only	Control-only	Case-control		Case-only	Control-only	Case-control	
Model a: $\text{logit } P(Y = 1 x) = -4 + x - .15x^2$ ($B_1 = .783$)								
Bias	.269	.097	.133	.061	.114	.029	.071	.061
MSE	.270	.040	.052	.097	.038	.019	.023	.097
$\widehat{\text{Std}}$ *	.244	.173	.177	.263	.155	.136	.142	.263
95% CI coverage	.920	.934	.924	.952	.919	.949	.964	.952
Model b: $\text{logit } P(Y = 1 x) = -4 + x - .3x^2$ ($B_1 = .636$)								
Bias	.405	.136	.197	.060	.169	.038	.101	.060
MSE	.261	.053	.077	.140	.055	.019	.029	.104
$\widehat{\text{Std}}$.287	.188	.188	.264	.173	.140	.150	.264
95% CI coverage	.818	.918	.870	.964	.892	.967	.938	.964

NOTE: The size of the phase-one sample was 5,000. In the case-only design, all of the cases, with expected numbers of 113 and 96 for models a and b, are selected at phase two. In the control-only design, 113 controls and 96 controls are selected at phase two. In the case-control design, cases and controls are selected in 1:1 ratio, so that the expected total size of the phase-two sample is 113 and 96 for models a and b.

*Std: mean of estimated standard errors.

described earlier. When we tried to fit the quadratic model with the crude two-level surrogate, the lack of identifiability manifested itself by PS equations with multiple roots. Using the EM-type algorithm described in Section 2.2, we obtained different estimates depending on the starting values. When we used a three-level surrogate instead, this multiplicity occurred much less frequently and disappeared with increasing sample size. In general, we anticipate that the richness of the model that can be fitted using a restricted design will depend on the richness of the phase-one data and the extent of correlation between phase-one and phase-two covariates. In any given application, a possible way to check for identifiability problems would be to vary the starting values for the EM-type algorithm over a wide range of the parameter space and see whether the algorithm converges to different estimates for different starting values. Even if the true model is identifiable for a given study design, the statistical power to discriminate it from submodels may be quite low.

These studies of robustness involved misspecification of the regression model. A separate robustness issue, especially for continuous outcomes, occurs when the regression model for the lower-order moments of scientific interest is correctly specified, but the density function $f_{\beta}(Y|X, Z)$ is not. We simulated this situation in the linear regression setting of Table 4. The error ϵ was generated from a t -distribution with 5 degrees of freedom but was assumed normal. Regardless of the strength of the regression relationship, the PS method yielded unbiased estimates for both the slope and intercept parameters of the linear regression model (data not shown). Further exploration of the robustness of the PS method is needed, however, under more general misspecification of the error distribution.

4. CASE-CONTROL STUDY OF LEPROSY

The leprosy data shown in Table 6 were taken from Scott and Wild (1997, table 1). Clayton and Hills (1993, p. 156) have provided a detailed description. In short, these data resulted from case-control sampling of a population survey of people under 35 in Northern Malawi. Cases were all new cases

of leprosy. Controls were random samples from those without leprosy. The variable Age refers to the age-group midpoints, and Scar represents the presence or absence of a BCG vaccination scar (1, present; 0, absent). The known population totals in each category classified by leprosy status Y and age group Z constituted the phase-one data. Scar was observed only in the validation (i.e., case-control) sample. As suggested by Scott and Wild (1997), the linear logistic model

$$\text{logit}P = \beta_0 + \beta_1 T + \beta_2 \text{Scar},$$

where $T = 100(\text{Age} + 7.5)^{-2}$, was chosen to fit the data.

Results for PSII and the SPMLLE proposed by Scott and Wild (1997) are given in Table 7. The PS method gave similar point estimates and standard errors to those of SPMLLE. To illustrate the potential applicability of the PS method to restricted designs, a “case-only” analysis was conducted by dropping data on Scar for controls and pretending that the validation sample contained only the cases. Similarly, a “control-only” analysis was performed with data on Scar for cases deleted. Compared with the previous results, these two analyses yielded similar regression coefficients but had substantial loss of efficiency for estimation of the BCG (Scar) effect.

Table 6. The Leprosy Data

Age	Case-control (validation) sample				Population totals	
	Scar = 0		Scar = 1		Case	Control
	Case	Control	Case	Control		
2.5	1	24	1	31	2	19,367
7.5	11	22	14	39	25	17,388
12.5	28	23	22	27	50	13,222
17.5	16	5	28	22	44	10,352
22.5	20	9	19	12	39	8,047
27.5	36	17	11	5	47	6,003
32.5	47	21	6	3	53	6,503

Source: Table 1 of Scott and Wild (1997).

Table 7. Results of Analyses of Table 6 Data: $T = 100(\text{Age} + 7.5)^{-2}$

	Poststratified analysis				Case-only analysis: PS		Control-only analysis: PS	
	PS		MLE		Coefficient	Standard error	Coefficient	Standard error
	Coefficient	Standard error	Coefficient	Standard error				
Intercept	-4.484	.113	-4.481	.114	-4.423	.171	-4.477	.128
T	-4.092	.448	-4.091	.449	-3.976	.527	-4.040	.478
Scar	-.415	.169	-.421	.178	-.574	.368	-.460	.311

5. DISCUSSION

Our proposed new method for analyzing two-phase data is computationally simple and yet, at least in the examples that we have considered, highly efficient. For typical two-phase designs, where each subject has positive probability of being selected at phase two, the PS estimator had remarkably high efficiency compared to alternative pseudolikelihood estimators of comparable computational complexity. Moreover, except for extreme parameter values, it often achieved full or nearly full efficiency, in comparison with the fully efficient SPMLE. This is encouraging, because semiparametric efficient methods, particularly for continuous Y , can be complex and difficult to implement.

In some practical applications, it may be impossible or very expensive to collect detailed covariate data for subjects with certain values of Y . The assumption used by existing semiparametric methods—namely, that all subjects have positive probability of selection into the validation sample—renders them useless in such circumstances. The new methodology proposed in this article provides a way of analyzing data from such restricted designs. But restriction of the validation sample to subjects with certain Y values implies heavy reliance on the model assumptions, with limited opportunity for model checking. Thus, even though the proposed method allows analyses of data from restricted designs, we do not recommend their use unless absolutely necessary for practical reasons. Nevertheless, if such a design is inevitable, the new method provides at least a way of extracting some information from the data. In particular, if the underlying model assumption is not too unrealistic and if the phase-one covariate data contain substantial information about the full covariate data of interest, then the PS method could be a valuable tool for making regression inference.

In this article, we assessed the robustness of the PS estimator under model misspecification. In doing so we followed conventional wisdom, that the parameter estimate from WL is useful for prediction even if the wrong model is fitted. Thus we assessed the performance of the PS estimates relative to the WL estimates. Although they supported this viewpoint earlier (Scott and Wild 1986), Scott and Wild (2002) argued more recently that this is not always justified. They showed that if a linear logistic regression model is fitted when the underlying true model is quadratic, then the WL slope estimator is consistent for the tangent to the underlying quadratic curve at a point near the tail of the X distribution, where most of the

cases occur. By contrast, the SPMLE slope estimates the tangent at a more moderate or central value of X . Thus the choice between WL and SPML estimators depends on the covariate values of greatest interest to the researcher. With this alternative view in mind, we also compared the bias of the PS estimator relative to that of SPMLE. For the case-control design, where all three were applicable, the PS estimates were much closer to the SPMLEs than to the WL estimates.

The PS estimator is quite flexible and can handle data from two-phase stratified designs and other missing-covariate data problems for a large class of regression models. In his dissertation, Chatterjee (1999) considered several extensions of the basic approach outlined here. He showed that the method easily accommodates designs that generate non-iid data, for example, studies with case-control sampling at phase one (Breslow and Holubkov 1997). It also accommodates the more complex definition of strata considered by Lawless et al. (1999) and certain types of censored-data regression problems arising in reliability studies (Hu and Lawless 1996). A kernel-smoothing approach for estimating (9) in the presence of continuous Z , in the spirit of Carroll and Wand (1991), was found to perform well. These and other related results will be presented in a later publication. Of course, such smoothing techniques are limited by the “curse of dimensionality” to problems involving a small number of continuous Z variables. Further research is warranted on extensions of the method to various other designs that reduce costs by limiting the evaluation of expensive covariates to selected subjects. General multiphase designs, partial questionnaire designs (Wacholder, Carroll, Pee, and Gail 1994), and the case-cohort and nested case-control designs used in survival analysis are some examples of particular interest.

APPENDIX:

Regularity Conditions for Proposition 1: Z and Y Are Discrete

Let z_1, \dots, z_K be the possible values of Z and let y_1, \dots, y_S be the possible values of Y , and set $\gamma = (\beta, \alpha)$. The following assumptions are sufficient for the conclusion of Proposition 1 (consistency and asymptotic normality of the PS estimator):

(A0) Conditions A and B in (3) hold.

(A1) $\beta \rightarrow \log f_{\beta}(y|x, z)$ is thrice differentiable with respect to β , and the third derivatives are bounded by $M(y, x, z)$, an integrable function of (y, x, z) , for all β 's in a neighborhood of β_0 .

(A2) $\alpha \rightarrow r \log \pi(y, z; \alpha) + (1 - r) \log \{1 - \pi(y, z; \alpha)\}$ satisfies the analogous classical smoothness assumptions (as stated in A1) for α in a neighborhood of α_0 .

(A3) $\Psi_\beta = J_\beta + C_\beta$ is nonsingular.

(A4) For all (s, k) ,

$$0 < \int h_{\beta_0}^{\pi_0}(y_s|x, z_k) dG_{k0}^*(x) < \infty, \int |S_{\beta_0}(y_s|x, z_k)| \\ \times h_{\beta_0}^{\pi_0}(y_s|x, z_k) dG_{k0}^*(x) < \infty.$$

(A5) For all (s, k) , the functions $h_{\beta_0}^{\pi_0}(y_s|x, z_k)$ and $h_{\beta_0}^{\pi_0}(y_s|x, z_k)S_{\beta_0}(y_s|x, z_k)$ are twice differentiable with respect to γ , with the second derivatives uniformly integrable with respect to $G_0^*(x)$ for all γ in a neighborhood of γ_0 .

Outline of Proof of Proposition 1: Z, Y , and X all Discrete

a. Consistency of $\hat{\beta}_N^{\text{PS}}$ can be proved using arguments given by Foutz (1977). The key condition, the unbiasedness of the PS functions $S_{\text{PS}}(\beta; G_0^*, \pi_0)$, was shown in Section 2.1.

b. Let $G_{k0}^*(x) = P(X \leq x|Z = z_k, R = 1)$, for $k = 1, \dots, K$ and set $G_0^* = (G_1^*, \dots, G_K^*)$. Define G_{Nk} to be the empirical distribution function of $[X|Z = z_k]$ from the validation sample. Under appropriate stochastic equicontinuity conditions and smoothness assumptions (see, e.g., van der Vaart and Wellner 1996, p. 310), the following expansion can be shown to be valid:

$$\sqrt{N}(\hat{\beta}_N^{\text{PS}} - \beta_0) = -\Psi_\beta(\beta_0, G_0^*, \pi_0) \\ \times \sqrt{N} \left\{ \Psi_N(\beta_0; G_0^*, \pi_0) + \sum_{k=1}^K \Psi_{G_k^*}[G_{Nk} - G_{k0}^*] + \Psi_\alpha[\hat{\alpha} - \alpha_0] \right\} \\ + o_p(1). \quad (\text{A.1})$$

In (A.1), if X is discrete with possible values in $\{x_1, \dots, x_L\}$ and G_k^* is defined by $p_k^* = (p_{1k}^*, \dots, p_{Lk}^*)^T$, where $p_{lk}^* = \text{pr}(X = x_l|Z = z_k, R = 1)$, then $\Psi_{G_k^*}$ is the $\text{dim}(\beta) \times L$ derivative matrix $\partial\Psi(\beta_0; G_k^*, \pi_0)/\partial p_k^{*T}$ and $[G_{Nk} - G_{k0}^*]$ is the $L \times 1$ vector of $[\hat{p}_k^* - p_{k0}^*]$. When X is continuous, $\Psi_{G_k^*}$ is an operator representing the Fréchet derivative of $\Psi(\beta_0; G_k^*, \pi_0)$ with respect to the infinite-dimensional parameter G_k^* , and $[G_{Nk} - G_{k0}^*]$ is an empirical process. For simplicity, we continue the proof assuming that X is discrete. The steps for the continuous case are similar, although the details involve the more sophisticated arguments developed in Chatterjee's dissertation.

From (A.1), the proof of part b follows from standard linearization argument that shows

$$\sqrt{N} \sum_{k=1}^K \Psi_{G_k^*}[G_{Nk} - G_{k0}^*] = N^{-1/2} \sum_{i=1}^N a_1(R_i, X_i, Z_i) + o_p(1) \quad (\text{A.2})$$

and

$$\sqrt{N} \Psi_\alpha[\hat{\alpha} - \alpha_0] = \frac{1}{\sqrt{N}} \sum_{i=1}^N a_2(R_i, Y_i, Z_i) + o_p(1), \quad (\text{A.3})$$

where the functions $a_1(r, x, z)$ and $a_2(r, y, z)$ are defined in Proposition 1.

c. Asymptotic normality follows from standard application of the central limit theorem. To derive the given form of the asymptotic variance, it is enough to prove the following:

- (C1) $E_0\{a_0(R, Y, X, Z)a_2(R, Y, Z)\} = 0$.
 (C2) $E_0\{a_0(R, Y, X, Z)a_1(R, X, Z)\} = C_\beta$.
 (C3) $E_0\{a_1(R, X, Z)a_2(R, Y, Z)\} = \Psi_\alpha J_\alpha^{-1} \Psi_\alpha^T$.

The proof of C1 follows easily from the fact that $E[a_0(R, Y, X, Z)\{R - \pi(Y, Z; \alpha)\}|Y, Z] = 0$. Let E_0^z denote the expectation operator given $Z = z$. To prove C2, note that

$$E_0^z\{a_0(R, Y, X, z)a_1(R, Y, z)\} \\ = E_0^z \left\{ RS_{\beta_0}(Y|X, z) \int \frac{1 - \pi_0(y, z)}{q_{\beta_0}^{\pi_0}(X, z)} D(y, X, z) f_{\beta_0}(y|X, z) dy \right\} \\ = \int [1 - \pi_0(y, z)] \\ \times E_0^z \left\{ \frac{\pi_0(Y, z)}{q_{\beta_0}^{\pi_0}(X, z)} S_{\beta_0}(Y|X, z) D(y, X, z) f_{\beta_0}(y|X, z) \right\} dy.$$

The proof now follows from the fact that

$$\frac{1}{q_{\beta_0}^{\pi_0}(X, Z)} E_0\{S_{\beta_0}(Y|X, Z)\pi_0(Y, Z)|X, Z\} = \frac{\partial}{\partial \beta} \log q_{\beta_0}^{\pi_0}(X, Z).$$

To prove C3, it is enough to show that $E_0\{a_1(R, X, Z)S_{\alpha_0}(R|Y, Z)\} = \Psi_\alpha$. If we write $a_1(R, X, Z)$ as $Rm(X, Z)$, then it easily follows that

$$E_0\{a_1(R, X, Z)S_{\alpha_0}(R|Y, Z)\} = E_0 \left[\frac{\partial \pi(Y, Z; \alpha_0)}{\partial \alpha} E\{m(X, Z)|Y, Z\} \right].$$

Next, rewrite the formula for Ψ_α given in (15) as

$$-E_0 E_0^Z E_0^{Y, Z} \left[\begin{aligned} & \{1 - \pi_0(Y, Z)\} \frac{D(Y, X, Z)}{q_{\beta_0}^{\pi_0}(X, Z)} \\ & \times \int \frac{\partial \pi(y, Z; \alpha_0)}{\partial \alpha} f_{\beta_0}(y|X, Z) dy \end{aligned} \right].$$

Then, by changing the order of the expectations and the integration, the foregoing expression can be written as $E_0\{\{\partial \pi(Y, Z; \alpha_0)/\partial \alpha\}h(Y, Z)\}$, where

$$h(y, Z) \\ = \frac{1}{dP(y|Z)} E_0^Z E_0^{Y, Z} \left[\{1 - \pi_0(Y, Z)\} \frac{D(Y, X, Z)}{q_{\beta_0}^{\pi_0}(X, Z)} f_{\beta_0}(y|X, Z) \right] \\ = \frac{1}{dP(y|Z)} \int \left[\begin{aligned} & \{1 - \pi_0(Y, Z)\} \\ & \times \int \frac{D(Y, x, Z) f_{\beta_0}(Y|x, Z) f_{\beta_0}(y|x, Z)}{[q_{\beta_0}^{\pi_0}(x, Z)]^2} dG^*(x|Z) \end{aligned} \right] dY \\ = \frac{1}{dP(y|Z)} \int \frac{f_{\beta_0}(y|x, Z)}{[q_{\beta_0}^{\pi_0}(x, Z)]^2} \int \{1 - \pi_0(Y, Z)\} D(Y, x, Z) f_{\beta_0}(Y|x, Z) dY \\ dG^*(x|Z) \\ = E\{m(X, Z)|Y = y, Z\}.$$

[Received April 2001. Revised March 2002.]

REFERENCES

- Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M., and Welsh, A. H. (1994), "Maximum Likelihood Inference From Sample Survey Data," *International Statistical Review*, 62, 349–363.
 Breslow, N. E. (1996), "Statistics in Epidemiology: The Case-Control Study," *Journal of the American Statistical Association*, 91, 14–28.
 Breslow, N. E., and Cain, K. C. (1988), "Logistic Regression for Two-Stage Case-Control Data," *Biometrika*, 75, 11–20.
 Breslow, N. E., and Chatterjee, N. (1999), "Design and Analysis of Two Phase Studies With Binary Outcome Applied to Wilms Tumor Prognosis," *Applied Statistics*, 48, 457–468.
 Breslow, N. E., and Holubkov, R. (1997), "Maximum Likelihood Estimation of Logistic Regression Parameters Under Two-Phase, Outcome-Dependent Sampling," *Journal of the Royal Statistical Society, Ser. B*, 59, 447–461.
 Breslow, N. E., McNeney, B., and Wellner, J. (2002), "Large Sample Theory for Semiparametric Regression Models With Two-Phase, Outcome-Dependent Sampling," *The Annals of Statistics*, (in press).

- Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.
- Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, Ser. B, 53, 573-587.
- Chatterjee, N. (1999), "Semiparametric Inference Based on Estimating Equations in Regression Models for Two-Phase Outcome-Dependent Sampling," unpublished doctoral dissertation, University of Washington, Seattle.
- Clayton, D., and Hills, M. (1993), *Statistical Models in Epidemiology*, London: Chapman and Hall.
- Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.
- Flanders, T. R., and Greenland, S. (1991), "Analytic Methods for Two-Stage Case-Control Studies and Other Stratified Designs," *Statistics in Medicine*, 10, 739-747.
- Foutz, R. V. (1977), "On the Unique Consistent Solution to the Likelihood Equations," *Journal of the American Statistical Association*, 72, 147-148.
- Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663-685.
- Hu, X. J., and Lawless, J. F. (1996), "Estimation From Truncated Lifetime Data With Supplementary Information on Covariates and Censoring Time," *Biometrika*, 83, 747-761.
- Korn, E. L., and Graubard, B. I. (1999), *Analysis of Health Surveys*, New York: Wiley.
- Krieger, A. M., and Pfeiffermann, D. (1992), "Maximum Likelihood Estimation From Complex Sample Surveys," *Survey Methodology*, 18, 225-239.
- Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999), "Semiparametric Methods for Response-Selective and Missing Data Problems in Regression," *Journal of the Royal Statistical Society*, Ser. B, 61, 413-438.
- Manski, C. F., and Thompson, T. S. (1989), "Estimation of best predictors of binary response," *Journal of Econometrics*, 40, 97-123.
- Neyman, J. (1938), "Contribution to the Theory of Sampling From Human Populations," *Journal of the American Statistical Association*, 33, 101-116.
- Pepe, M. S., and Fleming, T. R. (1991), "A Non-Parametric Method for Dealing With Mismeasured Covariate Data," *Journal of the American Statistical Association*, 86, 108-113.
- Pfeiffermann, D. (1996), "The Use of Sampling Weights for Survey Data," *Statistical Methods in Medical Research*, 5, 239-261.
- Pierce, D. A. (1982), "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics," *The Annals of Statistics*, 10, 475-478.
- Rao, J. N. K., Kovar, J. G., and Mantel, H. J. (1990), "On Estimating Distribution Functions and Quantiles From Survey Data Using Auxiliary Information," *Biometrika*, 77, 365-375.
- Reilly, M., and Pepe, M. S. (1995), "A Mean Score Method for Missing and Auxiliary Covariate Data in Regression Models," *Biometrika*, 82, 299-214.
- Robins, J. M., Hsieh, F., and Newey, W. (1995), "Semiparametric Efficient Estimation of a Conditional Density With Missing or Mismeasured Covariates," *Journal of the Royal Statistical Society*, Ser. B, 57, 409-424.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846-866.
- Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 72, 497-412.
- Scott, A. J., and Wild C. J. (1986), "Fitting Logistic Models Under Case-Control or Choice-Based Sampling," *Journal of the Royal Statistical Society*, Ser. B, 48, 170-182.
- (1997), "Fitting Regression Models to Case-Control Data by Maximum Likelihood," *Biometrika*, 84, 57-71.
- (2002), "The Robustness of the Weighted Methods for Fitting Models to Case-Control Data," *Journal of the Royal Statistical Society*, Ser. B, 64, 207-219.
- Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, Chichester, U.K.: Wiley.
- van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.
- Wacholder, S., Carroll, R. J., Pee, D., and Gail, M. H. (1994), "The Partial Questionnaire Design for Case-Control Studies," *Statistics in Medicine*, 13, 623-634.