STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2836–2852
Published online 29 November 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2754

Pseudoscore-based estimation from biased observations

X. Joan Hu[1,*,†], R. Jason Schroeder[2], Winfred C. Wang[3] and James M. Boyett[4]

[1]*Department of Statistics and Actuarial Science, Simon Fraser University, Vancouver, BC V5A 1S6, Canada*
[2]*Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, MD 20850, U.S.A.*[‡]
[3]*Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN 38105, U.S.A.*
[4]*Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, U.S.A.*

SUMMARY

There are many practical situations where observation of the primary variable Y for individuals in a population is incomplete and depends on some auxiliary variables X that are potentially correlated with Y. We consider parameter estimation for the distribution of Y with the incomplete data, without specifying the underlying association between Y and X. The approach is based on a class of pseudoscore functions using available information of X. We demonstrate the consistency and asymptotic normality of the estimators and study their finite-sample properties in various situations *via* simulation. The methodology is illustrated by an example involving kindergarten readiness skills in children with sickle cell disease. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: auxiliary information; estimating function; missing data; non-ignorable non-response

1. INTRODUCTION

There has been substantial interest in statistical analysis with biased observations from epidemiology, reliability and other areas. For example, Suzuki [1], Kalbfleisch and Lawless [2] and Hu and Lawless [3] consider response-biased failure-time data, Breslow and Cain [4] and Reilly and Pepe [5] study incomplete data in case-control and other epidemiologic problems, Wild [6]

*Correspondence to: X. Joan Hu, Department of Statistics and Actuarial Science, Simon Fraser University, Vancouver, BC V5A 1S6, Canada.
†E-mail: joanh@stat.sfu.ca
‡The views presented in this article do not necessarily reflect those of the Food and Drug Administration, U.S.A.

Contract/grant sponsor: National Institute of Allergy and Infectious Diseases of U.S.A.; contract/grant number: R01 AI 56995
Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada; contract/grant number: RGPIN 262823

Received 8 February 2006
Accepted 2 October 2006

deals with choice-based sampling with categorical responses, and Robins *et al.* [7, 8] present weighted estimating functions for analyses with regression models in the presence of missing data. More recently, Lawless *et al.* [9] considers a general situation where the observation of the response Y and covariates X depends on which of the finite number of strata (Y, X) belongs to and Chatterjee *et al.* [10] presents a pseudoscore estimator for regression problems with two-phase sampling. Related papers include Qin *et al.* [11], which proposes a likelihood-based estimation with survey data under non-ignorable non-response, and Wang *et al.* [12], which considers a joint conditional likelihood estimator in logistic regression with missing covariate data.

There is an assumption that young children with sickle cell disease are at risk of brain damage. A study was conducted to assess the risk of cognitive impairment of sickle cell children, using a teacher-administered test on kindergarten-appropriate skills [13]. The study team approached the families of the 125 children identified from a sickle cell disease database for consent to review the test results of the children. Roughly 62 per cent (77 out of 125) of the families could not be contacted at all, either because the family had moved to another home address or had changed their telephone number. Of those 48 children whose families were successfully contacted, one family declined to participate in the study, test data for 13 of the children were not found in the database, and one child's record was not complete. This led to a rather low proportion of available test scores ($33/125 \approx 26.4$ per cent). No significant difference was detected between the consented children with and without test scores. However, a preliminary analysis indicated that the median family incomes (as determined by zip code) were significantly lower for the $125 - 33 = 92$ sickle cell children without available test scores than for the 33 children with available test scores. Since socioeconomic factors, such as family income and living environment, may very much affect children's cognitive development, the available test data are likely biased. An analysis of the study data with an adjustment for the possible bias was desirable to the medical investigators. Given the available information, it seemed safe to assume that the missingness of test scores is dependent on the test scores through some other variables, say family income. On the other hand, no model appeared convincing for the specification of the association between test scores and family income. These considerations motivated the approach presented in this paper.

Suppose the distribution of a response random variable Y is $f(y; \theta)$. Our primary interest is in estimating θ with the available observations from a group of independent subjects: $\{y_i : r_i = 1, i = 1, \ldots, N\}$, where r_i is the indicator of whether y_i is observed or not. We consider the situations where the available data are not a random sample from the population, that is, $P(R = 1|Y) \neq P(R = 1)$, with R the indicator for an arbitrary individual. Suppose further that the observation mechanism is associated with other variables X, which are potentially correlated with Y. Here X are continuous or discrete, and readily obtained for all subjects. We assume that (Y_i, R_i, X_i), $i = 1, \ldots, N$, are iid and that

$$P(R = 1|Y, X) = P(R = 1|X) \equiv \pi(X) \tag{1}$$

with $\pi(x) > 0$ for all possible values of X. Model (1) is used in many published papers. See Chatterjee *et al.* [10] for an example. In the current context, with the introduction of the variables X, the missingness mechanism associated with the Y observation is referred to as missing at random in the literature (e.g. Little and Rubin [14]). We assume further that $P(R = 1|Y = y) > 0$ for all possible y.

It is apparent that an estimator for θ using only the observed Y_i could be seriously biased. Utilizing the available X_i, which provide some auxiliary information about Y_i, may address this problem under the assumption (1). To do so, we may specify $Y|X = x \sim f(y|x; \beta)$ and

DOI: 10.1002/sim

assume $X \sim G(\cdot)$. By noting $f(y;\theta) = \int f(y|x;\beta)\,\mathrm{d}G(x)$, estimation of θ may be derived from the likelihood with the available data

$$L(\beta, G) = \prod_{i:R_i=1} f(Y_i|X_i;\beta)\,\mathrm{d}G(X_i) \prod_{i:R_i=0} \mathrm{d}G(X_i)$$

Alternatively, we may specify $(Y, X) \sim g(x|y;\alpha) f(y;\theta)$ and consider the estimation based on the likelihood

$$L(\theta, \alpha) = \prod_{i:R_i=1} g(X_i|Y_i;\alpha) f(Y_i;\theta) \prod_{i:R_i=0} \int g(X_i|y;\alpha) f(y;\theta)\,\mathrm{d}y$$

These two approaches are valid. However, implementing them is in general difficult. Besides the anticipatory difficulty confronted in specifying an appropriate regression model in either of the two approaches, the computation involved could be rather intensive.

Note that, if Y were known for all subjects, the score function

$$S(\theta) = \sum_{i=1}^{N} \partial \log f(Y_i;\theta)/\partial\theta \tag{2}$$

would lead to the usual maximum likelihood estimator (MLE) of θ. When the conditional distribution $Y|X$ is fully specified, the projection of $S(\theta)$ in (2) to the current data space

$$\sum_{i=1}^{N} R_i \partial \log f(Y_i;\theta)/\partial\theta + (1 - R_i) E\{\partial \log f(Y_i;\theta)/\partial\theta | X_i\} \tag{3}$$

may yield efficient inferences on θ. However, the conditional expectation in (3) is not necessarily easy to calculate. In addition, in many situations, such as for the study on young children with sickle cell disease described above, we are not confident in specifying the conditional distribution $Y|X$. Model checking for $Y|X$ is difficult with the current incomplete data. Possible misspecification of $Y|X$ may lead to potentially heavy bias in inference. The estimation based on (3) when $Y|X$ is either well-specified or misspecified will be further discussed in the next section *via* simulation.

This paper considers an approach based on different pseudoscore functions of θ, which may also be viewed as unbiased estimators of the score function $S(\theta)$ given in (2) using the current data. Our approach can be taken as an extension of an approach discussed by Lawless *et al.* [9], which is applicable in principle when X are discrete, especially when the number of distinct values of X is small relatively to $n = \sum_{i=1}^{N} r_i$, the number of individuals with observed Y. We will compare our approach with a direct application of the approach presented by Lawless *et al.* [9] *via* simulation in Section 3. As pointed out by Reviewers, our approach is closely related to the one presented by Robins *et al.* [8] and Robins and Rotnitzky [15], when θ is, say, the expectation of Y. This observation in fact indicates an extension of the current approach to the situations where the distribution of Y is not specified and estimating the moments of Y is of primary interest. Our work is different from published papers on related problems in several respects. We focus on parametric estimation of the distribution of Y from a possibly biased sample. Some available auxiliary information, i.e. in the form of observations on X, is used to address the observation mechanism without specifying the underlying association between Y and X. Also, we intend to accommodate the situations with continuous X.

The rest of the paper is organized as follows. Section 2 presents pseudoscore-based estimation. It also derives asymptotic properties and describes implementation of the approach. Section 3 reports

DOI: 10.1002/sim

simulation results. The simulation was conducted to examine the finite sample performances, to compare the efficiency with other methods, and to study the robustness against model misspecification. Section 4 illustrates the method using data from the cognitive impairment study that motivated this research. The final section discusses the advantages and limitations of our approach and mentions some possible extensions.

2. PSEUDOSCORE-BASED ESTIMATION

We start with presenting a pseudoscore function of θ with the biased observations described in Section 1, assuming $\pi(x)$ in (1) is known. An estimation procedure derived from the pseudoscore function along with its asymptotic properties is presented. We then describe implementation of the procedure when the observation probability $\pi(x)$ is unknown. The associated asymptotic results and variance estimation are also presented.

2.1. Pseudoscore function

Notice that the complete data score function $S(\theta)$ is not evaluable with the current data. Instead, we consider

$$PS(\theta;\pi)=\sum_{i=1}^{N}\frac{R_i}{\pi(X_i)}\frac{\partial \log f(Y_i;\theta)}{\partial\theta} \tag{4}$$

a pseudoscore function of θ for a given π. It is in fact an unbiased estimator of $S(\theta)$ conditional on the complete data $\{(Y_i, X_i) : i=1,\ldots,N\}$ for any given θ and π. The evaluation of $PS(\theta;\pi)$ in (4) does not require specifying the conditional distribution $Y|X$. Further, it does not rely on a specification of the distribution of X. Notice that $PS(\theta;\pi)$ is the sum of iid terms, each with mean 0 and a common variance, which we denote by $\Sigma(\theta;\pi)$. Under some regularity conditions, $PS(\theta;\pi)/N\to 0$ in probability and $PS(\theta;\pi)/\sqrt{N}\to \mathrm{N}(0,\Sigma(\theta;\pi))$ in distribution as $N\to\infty$. Plus it is easy to verify that for given θ and π

$$\hat{\Sigma}(\theta;\pi)=\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi(X_i)^2}\left(\frac{\partial \log f(Y_i;\theta)}{\partial\theta}\right)\left(\frac{\partial \log f(Y_i;\theta)}{\partial\theta}\right)' \tag{5}$$

is an unbiased estimator for $\Sigma(\theta;\pi)$.

2.2. Estimation procedure with known $\pi(x)$

Since $E\{PS(\theta;\pi)\}=0$, the estimation equation $PS(\theta;\pi)=0$ is unbiased for the parameter θ, for a given π. Denote the estimator derived from the estimation equation by $\hat{\theta}$. The proposition below establishes the consistency and asymptotic normality of $\hat{\theta}$.

Proposition 1

Under some regularity conditions, as $N\to\infty$ and for a given π, $\hat{\theta}\to\theta$ in probability and $\sqrt{N}(\hat{\theta}-\theta)$ converges in distribution to the normal distribution with mean 0 and variance in the form

$$I(\theta)^{-1}\Sigma(\theta;\pi)I(\theta)^{-1} \tag{6}$$

where $I(\theta)=E\{-\partial^2 \log f(Y;\theta)/\partial\theta^2\}$.

DOI: 10.1002/sim

A proof of the above result may be obtained by utilizing approaches similar to those of Hu and Lawless [16], for example. Note that the asymptotic variance in (6) is

$$I(\theta)^{-1} + I(\theta)^{-1} E\left\{\left(\frac{1}{\pi(X)} - 1\right)\left(\frac{\partial \log f(Y;\theta)}{\partial\theta}\right)\left(\frac{\partial \log f(Y;\theta)}{\partial\theta}\right)'\right\} I(\theta)^{-1}$$

The first term above is the asymptotic variance associated with the MLE of θ if the data are complete, and the second term represents the loss in efficiency in estimating θ due to the incompleteness of the data. The loss is proportional to $\pi(x)$ and depends on the association between Y and X. Further, we see that

$$\hat{I}(\theta;\pi) = -\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi(X_i)}\frac{\partial^2 \log f(Y_i;\theta)}{\partial\theta^2} \tag{7}$$

is an unbiased estimator for $I(\theta)$. Combining (5) and (7) gives an consistent estimator of (6), the asymptotic variance with fixed $\pi(\cdot)$, provided that $\partial^2 \log f(y;\theta)/\partial\theta^2$ is a continuous function of θ.

2.3. Estimation procedure with unknown $\pi(x)$

The probability $\pi(x) = P(R=1|X=x)$ may be predetermined in a study so that the data collection follows the chosen sampling scheme. This mostly occurs in studies where there is a design for data collection at the beginning and good adherence to the design afterwards. However, in many situations, especially in observational studies, $\pi(x)$ is unknown. We then propose to estimate θ by solving $PS(\theta;\hat{\pi}) = 0$, where $\hat{\pi}(\cdot)$ is an appropriate estimator of $\pi(\cdot)$. The estimating function $PS(\theta;\hat{\pi})$ should perform similarly to $PS(\theta;\pi)$ in (4) if $\hat{\pi}(x)$ bears some necessary properties. In particular, we consider specifying a parametric form for $\pi(x)$ in the following.

Assume that $\pi(x)$ is known up to a parameter vector α, and denote this by $\pi(x;\alpha)$. An appropriate estimator for $\pi(x)$ can then be obtained by estimating α through the likelihood

$$\prod_{i=1}^{N}\pi(X_i;\alpha)^{R_i}\{1-\pi(X_i;\alpha)\}^{1-R_i} \tag{8}$$

A logistic regression model for $\pi(\cdot)$ may be employed in many situations. Note that (X_i, R_i) are available for all individuals in the population. The various techniques of regression model checking for binary response situations may be applied to diagnose the model specification. This will be further discussed at some length in Section 4 through the practical example.

Let $\hat{\alpha}$ be the MLE from (8). Recall that $\sqrt{N}(\hat{\alpha}-\alpha) \to \mathrm{N}(0, \Gamma(\alpha)^{-1})$ in distribution, where $\Gamma(\alpha) = E\{-\partial^2 \log h(X;\alpha)/\partial\alpha^2\}$ with $h(X;\alpha) = \pi(X;\alpha)^R\{1-\pi(X;\alpha)\}^{1-R}$. The following proposition presents the asymptotic properties of $PS(\theta;\hat{\pi})$ with $\hat{\pi}(x) = \pi(x;\hat{\alpha})$ and the estimator $\tilde{\theta}$ derived from $PS(\theta;\hat{\pi}) = 0$.

Proposition 2

Under some regularity conditions, the following results hold.

(i) For $N \to \infty$, $PS(\theta;\hat{\pi})/\sqrt{N}$ converges in distribution to a normal distribution with mean 0 and variance $\Xi(\theta;\alpha)$ in the form

$$\Xi(\theta;\alpha) = \Sigma(\theta;\alpha) + \Pi(\theta;\alpha) + 2\Psi(\theta;\alpha) \tag{9}$$

DOI: 10.1002/sim

where $\Sigma(\theta; \alpha)$ is $\Sigma(\theta; \pi)$ given in Section 2.1 with $\pi(\cdot)$ replaced with $\pi(\cdot; \alpha)$, $\Pi(\theta; \alpha)$ and $\Psi(\theta; \alpha)$ are presented in Appendix A.

(ii) The estimator $\tilde{\theta}$ is consistent and $\sqrt{N}\{\tilde{\theta}-\theta\}$ converges in distribution to the normal distribution with mean 0 and variance

$$I(\theta)^{-1}\Xi(\theta; \alpha)I(\theta)^{-1} \tag{10}$$

A consistent estimator of the asymptotic variance (10) is given in Appendix A. Appendix B presents a proof of the proposition.

When X takes only a finite number of values, say, $a_1, \ldots, a_K$, the current data observation reduces to a special case of what is considered in Lawless *et al.* [9]. We can use $\hat{\pi}(a_k) = n_k/N_k$ with N_k and n_k the sizes of $\{i : X_i = a_k\}$ and $\{i : R_i = 1; X_i = a_k\}$, respectively. Then $PS(\theta; \hat{\pi})$ is the derivative of the weighted log-pseudo-likelihood

$$\sum_{k=1}^{K} \frac{N_k}{n_k} \sum_{i:X_i=a_k} R_i \log f(Y_i; \theta) \tag{11}$$

We can easily obtain the asymptotic properties of the estimator based on (11) from *Proposition* 2 by viewing $\pi(\cdot)$ as being specified upon a K-dimensional parameter. Weighted pseudo-likelihoods implement the Horvitz–Thompson approach common in survey sampling. See, for example, [17, 18]. They have been used in problems involving response-selective observation by, for example, [3, 9, 19–21].

For the situations with continuous X, or discrete X but with a large number of distinct values, we may apply the estimation derived from the weighted log-pseudo-likelihood (11) by grouping the values of X. Specifically, we may stratify the subjects according to their values of X into a finite number of strata, say K^* strata, and then use $\hat{\pi}(x) = n_k/N_k$ for x in stratum $k = 1, \ldots, K^*$, where N_k is the size of stratum k, and n_k is the number of observed Y_i. Here K^* should be small relative to N. When N and K^* are both large enough, this $\hat{\pi}(x)$ can be a good approximation to $\pi(x)$. In general, when $\hat{\pi}(x)$ is uniformly consistent, the estimating function $PS(\theta; \hat{\pi})/N \to 0$ in probability. Thus, the estimator derived from $PS(\theta; \hat{\pi}) = 0$ should be consistent. This will be discussed further in Section 3 based on some empirical results.

3. SIMULATION

To study the finite-sample properties of the proposed estimation, we performed the following simulation. We chose the group size to be $N = 100$ and generated independent X_i, $i = 1, \ldots, N$, from the lognormal distribution with $E(\log X_i) = 1$ and $\text{var}(\log X_i) = 1$. Then the primary variables Y_i were generated as $Y_i = \beta_0 + \beta_1 \log X_i + e_i$, where e_i are independent of X_i and iid from $\text{N}(0, 0.5^2)$. We took $\beta_0 = 1$ through all the simulation cases, and varied β_1 from 0.3, 1 to 3 to simulate different levels of dependence of the primary variable Y on the auxiliary variable X. The marginal distribution of Y was then normal with mean $\mu_Y = 1 + \beta_1$ and variance $\sigma_Y^2 = \beta_1^2 + 0.5^2$. The parameter to be estimated in the simulation was $\theta = (\mu_Y, \sigma_Y)$.

The observation mechanism for Y was simulated as follows. Given $X_i = x_i$, the indicator R_i for whether Y_i was observed or not was generated from the Bernoulli distribution with the success probability $\pi(x)$ following $\text{logit}(\pi(x)) = \alpha_0 + \alpha_1 \log x + \alpha_2 (\log x)^2$. We chose α_1 to be 0, 0.5 or 1,

DOI: 10.1002/sim

and α_2 to be 0 or 0.2, to simulate varying levels of dependence of the observation mechanism on X. The value of α_0 was chosen accordingly to α_1 and α_2 to vary incompleteness of the Y observation from moderate to heavy: the marginal observation rates n/N ranged from 75, 50 to 25 per cent.

Both estimators $\hat{\theta}$ and $\tilde{\theta}$ in Sections 2.2–3 were evaluated using the simulated data. For evaluating the estimator $\tilde{\theta}$ in Section 2.3, we employed the following estimators for $\pi(x)$:

(a) We took the logistic regression model $\text{logit}(\pi(x)) = \alpha_0 + \alpha_1 \log x + \alpha_2 (\log x)^2$ and estimated the parameters by (i) their maximum likelihood estimates from the likelihood in (8), and (ii) fixing $\alpha_2 = 0$ and estimating α_0 and α_1 from (8) under the reduced model. The resulting estimates for θ are labelled in the tables summarizing the simulation results by 'PSE($\hat{\pi}$)' and 'PSE($\tilde{\pi}$)', respectively.

(b) We classified the simulated individuals into K strata according to their values of X: individual i was classified into the kth stratum if $\log x_i$ was in between the normal quantiles of $(k-1)100/K$ and $k100/K$. The probability $\pi(x)$ is estimated by $\hat{\pi}(x) = n_k/N_k$ for x in the kth interval, where N_k is the size of the kth stratum and n_k the number of observed Y within the kth stratum. The simulation used $K = 5$ and 10. The corresponding estimates are labelled in the tables reporting the simulation results by 'WPSE ($K = 5$)' and 'WPSE ($K = 10$)', respectively.

We also evaluated the MLE based on (2) from the complete data and, ignoring the possible bias in the data, the MLE from the incomplete data (labelled by 'NAIVE' in the tables). The MLE is not evaluable with the data structure discussed in this paper. The naive approach assumes the observed Y_i from a random sample of the population, which is violated in the situations of interest. We included them in the simulation in order to explore the extent to which the proposed approach recovers the missing information.

In addition, we evaluated estimators derived from (3) based on either correctly specified or misspecified models for $Y|X$, denoted by 'CPSE' and 'CPSEI' in the tables, respectively. We calculated $E\{\partial \log f(Y_i; \theta)/\partial\theta | X_i\}$ using $E(Y|X) - \mu_Y$ as $\beta_1^*(X - \gamma)$ instead of the true value $E(Y|X) - \mu_Y = \beta_1\{\log X - E(\log X)\}$ for 'CPSEI', to simulate a possible misspecification of $Y|X$, where β_1^* and γ were chosen to obtain the same mean and variance using the 'misspecified' $E(Y|X) - \mu_Y$ as the true one.

The simulation settings were chosen for studying different situations where the location parameter μ_Y varies. To focus, we present only the simulation results regarding estimation of the mean μ_Y. Tables I–IV summarize the results based on 1000 repetitions of the simulation in the cases described above, excluding the ones with $n/N = 50$ per cent. The results for the cases with $n/N = 50$ per cent do not indicate anything beyond what we can see from the tables. Specifically, Tables I and II present the sample mean square errors of the estimates for the parameter μ_Y in cases $\alpha_2 = 0$ and 0.2, respectively.

Overall, the simulation results indicate that the pseudoscore estimators perform well relative to the maximum likelihood estimation with the complete data. They obviously out-perform the naive approach, particularly for the situations where there is a strong dependence of the observation mechanism on Y through X and where the proportion of unobserved Y is large, regardless of using the true or estimated $\pi(\cdot)$. For example, Table I shows that, in the cases of $\alpha_1 = 1$ and $\beta_1 = 3$, the simulated situations in which there is a strong dependence of the observation of Y on X, the ratios of the sample mean square errors of the pseudoscore estimators *vs* MLE were 1.09 to 1.39 (when $n/N = 75$ per cent) and from 1.93 to 3.06 (when $n/N = 25$ per cent), while the ratios of

DOI: 10.1002/sim

Table I. Sample mean square errors ($\times 1000$) of the estimates for μ_Y based on 1000 repetitions when $\alpha_2 = 0$.

	β_1	0.3			1			3		
Estimator	α_1	0	0.5	1	0	0.5	1	0	0.5	1
(a) n/N (observation rate) $= 75$ per cent										
MLE		58	58	60	113	114	111	295	294	301
NAIVE		67	67	73	131	147	182	336	491	847
WPSE ($K = 5$)		65	65	69	120	121	120	301	306	309
WPSE ($K = 10$)		65	67	72	119	121	121	298	303	323
PSE(π)		67	69	74	131	138	152	336	351	419
PSE($\tilde{\pi}$)		65	66	72	118	122	130	296	304	368
PSE($\hat{\pi}$)		65	66	72	118	122	126	296	304	329
CPSE		70	68	70	148	141	135	387	361	355
CPSEI		70	70	75	150	154	161	394	422	503
(b) n/N (observation rate) $= 25$ per cent										
MLE		57	57	60	110	107	109	295	308	307
NAIVE		120	130	141	226	343	581	631	1661	3897
WPSE ($K = 5$)		119	121	133	166	183	236	379	464	709
WPSE ($K = 10$)		126	124	129	180	205	257	458	572	922
PSE(π)		120	130	144	226	249	328	631	696	939
PSE($\tilde{\pi}$)		109	116	137	149	183	263	337	473	720
PSE($\hat{\pi}$)		112	116	135	168	172	220	352	418	594
CPSE		153	144	131	406	342	293	1182	1052	846
CPSEI		157	155	152	427	390	385	1227	1265	1192

NAIVE *vs* MLE were 2.81 and 12.69 in the two cases, respectively. As the simulation results show that the standard deviations of different estimators are all close to the true values, the large mean square errors of the NAIVE estimates in Tables I and II were mostly due to the bias. Moreover, the pseudoscore estimators and the estimator based on (3) with correct specification of $Y|X$, i.e. CPSE in the table, are of comparable performance across the various simulation cases. The relative performance of the estimator from (3) with a misspecified $Y|X$ (i.e. CPSEI in the tables) to the PSE estimators in the simulation suggests the PSE estimators are preferable alternatives to CPSE, especially when we are not confident in specifying the conditional distribution $Y|X$.

Using $\alpha_2 = 0.2$ and different estimates of $\pi(x)$ for the evaluations of the estimators for θ, we examined the robustness of the pseudoscore estimator $\tilde{\theta}$ against the situations where $\pi(x)$ is misspecified. The simulation results in Table II indicate that the estimator even with a misspecified $\pi(x)$ is preferable over the naive approach, and the estimator with a well-specified $\pi(x)$ improves the naive approach a great deal. We also noted that the computing associated with PSE is in general less intense than for CPSE/CPSEI.

The relative behaviours of the PSE estimates, labelled by PSE($\tilde{\pi}$) and PSE($\hat{\pi}$) in Tables I and II, are consistent with our understanding about them. The pseudoscore estimator becomes more sensitive to the misspecification of π when the degree of the incompleteness is increasing, i.e. when n/N is decreasing. The performance of the estimator PSE(π) appeared somehow counter-intuitive at first glance. It is natural to think that PSE(π) should be the best among the three PSE estimators, since it uses the true $\pi(\cdot)$. The simulation results, however, show that PSE($\hat{\pi}$) performed the best.

DOI: 10.1002/sim

Table II. Sample mean square errors ($\times 1000$) of the estimates for μ_Y based on 1000 repetitions when $\alpha_2 = 0.2$.

	β_1	0.3			1			3		
Estimator	α_1	0	0.5	1	0	0.5	1	0	0.5	1
(a) n/N (observation rate) $= 75$ per cent										
MLE		58	58	60	109	111	115	306	310	308
NAIVE		68	72	77	136	165	222	434	713	1183
WPSE ($K = 5$)		66	70	73	115	120	125	320	326	327
WPSE ($K = 10$)		66	71	75	114	121	130	313	323	348
PSE(π)		67	72	77	124	137	161	349	373	452
PSE($\tilde{\pi}$)		65	72	86	113	132	181	312	360	519
PSE($\hat{\pi}$)		65	71	74	112	120	127	310	334	341
CPSE		69	71	71	134	138	136	387	381	374
CPSEI		70	74	76	143	156	174	413	469	571
(b) n/N (observation rate) $= 25$ per cent										
MLE		60	58	58	108	112	111	308	309	308
NAIVE		128	143	152	336	591	771	1578	3767	5545
WPSE ($K = 5$)		122	136	144	181	231	269	428	576	826
WPSE ($K = 10$)		129	131	138	192	247	290	486	717	1094
PSE(π)		118	136	159	217	279	341	610	722	975
PSE($\tilde{\pi}$)		124	171	191	242	355	416	593	1045	1151
PSE($\hat{\pi}$)		116	129	139	163	198	230	363	464	589
CPSE		141	127	118	331	310	275	1020	889	825
CPSEI		158	146	138	412	403	372	1262	1177	1133

A closer look at the construction of PSE(π), compared to that of PSE($\hat{\pi}$) and PSE($\tilde{\pi}$), may explain. Because the R_i were random variables generated from the Bernoulli distribution with the success probabilities $\pi(x_i)$, the estimates $\hat{\pi}(x_i)$ provide observation rates for the generated subjects closer to the true ones given R_i than the nominal observation probabilities $\pi(x_i)$. This may be further verified by examining the construction of the asymptotic variance (9). Note that the first term on the right hand side of (9) with PSE($\hat{\pi}$) can be very close to the one associated with PSE(π), and the sum of the remaining two terms can be negative for PSE($\hat{\pi}$) when X is positively correlated with Y but is zero for PSE(π). This relative performance of PSE($\hat{\pi}$) and PSE(π) suggests that an appropriate estimate of $\pi(x)$ may lead to a better estimate. It indicates that knowledge of the form of $\pi(x)$ may be more important for the PSE estimation. In addition, for studies where $\pi(\cdot)$ is predetermined and PSE(π) is evaluable, the estimator PSE($\hat{\pi}$) provides an alternative which may perform better than PSE(π).

The results of WPSE($K = 5$) and WPSE($K = 10$) suggest that the approximation method described at the end of Section 2.3 works quite well. Classifying the population into more strata may not necessarily give us more efficient inference, particularly in the situations where the number of observed Y_i is small. This is quite intuitive. For example, in the case of $n/N = 25$ per cent in our simulation, there were on average only two to three available Y_i in each of the strata when $K = 10$, and the two to three 'observed' Y_i may not provide a good representation of the whole stratum due to variability in Y.

DOI: 10.1002/sim

Table III. Comparison of sample means of variance estimates ($\bar{sd}$) and sample standard deviations ($\hat{sd}$) of estimates for μ_Y based on 1000 repetitions when $\alpha_2 = 0$.

	β_1	0.3			1			3		
Estimator	α_1	0	0.5	1	0	0.5	1	0	0.5	1
(a) n/N (observation rate) $= 75$ per cent										
MLE	$\bar{sd}$	0.058	0.058	0.058	0.111	0.111	0.111	0.301	0.301	0.302
	$\hat{sd}$	0.058	0.058	0.060	0.113	0.114	0.111	0.295	0.294	0.301
PSE(π)	$\bar{sd}$	0.067	0.068	0.074	0.128	0.134	0.148	0.347	0.362	0.411
	$\hat{sd}$	0.067	0.069	0.074	0.131	0.138	0.152	0.336	0.350	0.419
PSE($\tilde{\pi}$)	$\bar{sd}$	0.065	0.065	0.068	0.115	0.116	0.120	0.302	0.305	0.316
	$\hat{sd}$	0.065	0.066	0.072	0.118	0.122	0.130	0.296	0.304	0.368
PSE($\hat{\pi}$)	$\bar{sd}$	0.070	0.069	0.069	0.138	0.131	0.123	0.381	0.352	0.322
	$\hat{sd}$	0.065	0.066	0.072	0.118	0.122	0.126	0.296	0.303	0.328
(b) n/N (observation rate) $= 25$ per cent										
MLE	$\bar{sd}$	0.058	0.058	0.058	0.111	0.111	0.111	0.302	0.301	0.302
	$\hat{sd}$	0.057	0.057	0.060	0.110	0.107	0.109	0.295	0.308	0.307
PSE(π)	$\bar{sd}$	0.115	0.119	0.126	0.218	0.233	0.256	0.591	0.649	0.704
	$\hat{sd}$	0.120	0.130	0.144	0.226	0.249	0.324	0.629	0.696	0.912
PSE($\tilde{\pi}$)	$\bar{sd}$	0.103	0.106	0.112	0.138	0.154	0.180	0.325	0.369	0.449
	$\hat{sd}$	0.109	0.116	0.137	0.149	0.183	0.261	0.337	0.472	0.701
PSE($\hat{\pi}$)	$\bar{sd}$	0.103	0.104	0.106	0.151	0.159	0.167	0.363	0.397	0.419
	$\hat{sd}$	0.112	0.116	0.134	0.168	0.171	0.214	0.352	0.412	0.548

Another interesting finding from the simulation is about the relative performance of the PSE and the WPSE estimators. When $\alpha_2 = 0$ and 0.2, the advantage of specifying $\pi(x)$ into a parametric form becomes more apparent as n/N is getting smaller, i.e. the degree of the incompleteness is increasing. Note that specifying $\pi(x)$ may link the available information in different strata and overcome the small size of the available data within each stratum. The results also indicate the importance of appropriately specifying $\pi(x)$. We will discuss about how to choose an appropriate model for $\pi(x)$ in the next section.

In addition, we investigated the adequacy of the variance estimation and the asymptotic normal approximations for the pseudoscore estimators, both of which would be used to obtain confidence intervals or tests for parameters. Tables III and IV present the sample means of the standard deviation estimates ($\bar{sd}$) and the sample standard deviations ($\hat{sd}$) of the PSE estimates along with the ones associated with the MLE in various cases. The evaluations of $\bar{sd}$ and $\hat{sd}$ are reasonably close to each other, except for the PSE with misspecified $\pi(\cdot)$ in the cases of heavy missingness. Thus, caution should be used in those situations.

We also examined the Q–Q plots of the 1000 standardized statistics $(\hat{\mu}_Y - \mu_Y)/\hat{sd}(\hat{\mu}_Y)$. The plots showed that the pivotals were approximately normal. At least for the cases considered here, these results indicate that the asymptotic approximations are sufficiently accurate for practical purposes.

DOI: 10.1002/sim

Table IV. Comparison of sample means of variance estimates ($\bar{sd}$) and sample standard deviations of estimates ($\hat{sd}$) for μ_Y based on 1000 repetitions when $\alpha_2 = 0.2$.

	β_1	0.3			1			3		
Estimator	α_1	0	0.5	1	0	0.5	1	0	0.5	1
(a) n/N (observation rate) $= 75$ per cent										
MLE	$\bar{sd}$	0.058	0.058	0.058	0.111	0.111	0.111	0.302	0.302	0.301
	$\hat{sd}$	0.058	0.058	0.060	0.109	0.111	0.115	0.306	0.310	0.308
PSE(π)	$\bar{sd}$	0.067	0.070	0.076	0.126	0.136	0.156	0.344	0.370	0.429
	$\hat{sd}$	0.067	0.072	0.077	0.124	0.137	0.161	0.349	0.373	0.452
PSE($\tilde{\pi}$)	$\bar{sd}$	0.065	0.068	0.072	0.116	0.118	0.129	0.310	0.311	0.330
	$\hat{sd}$	0.065	0.072	0.086	0.113	0.131	0.179	0.312	0.357	0.505
PSE($\hat{\pi}$)	$\bar{sd}$	0.079	0.074	0.073	0.176	0.144	0.133	0.491	0.399	0.358
	$\hat{sd}$	0.065	0.071	0.074	0.112	0.120	0.127	0.310	0.333	0.339
(b) n/N (observation rate) $= 25$ per cent										
MLE	$\bar{sd}$	0.058	0.058	0.058	0.111	0.111	0.111	0.302	0.302	0.300
	$\hat{sd}$	0.060	0.058	0.058	0.108	0.112	0.111	0.308	0.309	0.308
PSE(π)	$\bar{sd}$	0.116	0.126	0.135	0.213	0.246	0.269	0.572	0.668	0.746
	$\hat{sd}$	0.118	0.136	0.159	0.217	0.278	0.337	0.609	0.721	0.960
PSE($\tilde{\pi}$)	$\bar{sd}$	0.103	0.114	0.120	0.139	0.184	0.194	0.342	0.451	0.496
	$\hat{sd}$	0.123	0.169	0.190	0.232	0.340	0.414	0.523	0.908	1.113
PSE($\hat{\pi}$)	$\bar{sd}$	0.108	0.113	0.116	0.178	0.192	0.195	0.455	0.488	0.514
	$\hat{sd}$	0.116	0.129	0.139	0.163	0.197	0.221	0.363	0.457	0.538

4. AN EXAMPLE

For illustration, this section presents an analysis of the data from the study described in Section 1 (see also [13]), using the approach discussed in the previous sections.

Young children's performance on a standardized school test provides important information about their cognitive ability. The study was designed to achieve its objective by comparing sickle cell children with healthy children based on the results of a test on kindergarten-appropriate skills, including mathematics, language, memory and auditory discrimination. Out of the 125 children initially identified by the study, 47 children were consented and the test scores of 33 children among them were then obtained. For each of the 33 children with test scores available, two healthy children were selected from the database of the city school system as controls with matched gender, date of birth, and school (residence area). The test scores of the controls were then collected. The medical paper [13] reported a comparison between the 33 sickle cell children and the controls by the paired t-test, where the average of the scores of the two controls was taken as the score of the control for each sickle cell child. The naive approach was used in the analysis.

The low response rate in sickle cell children ($33/125 \approx 26.4$ per cent) led to a further investigation. T-test or Pearson's chi-square test did not detect any significant differences in age, gender, or diagnosis between the 33 sickle cell children with test scores available (Group I) and the remaining $125 - 33 = 92$ sickle cell children (Group II). However, based on the zip codes of residence and USA census data, median family income of the residence areas of Group I was \$30 977, much

DOI: 10.1002/sim

higher than that of Group II, which had a median family income of \$22 120. To explore the data further, the study subjects were stratified into five groups, according to the median family incomes of their residence areas: below \$10 000, between \$10 000 and \$20 000, between \$20 000 and \$30 000, between \$30 000 and \$40 000, and above \$40 000. The distributions of Groups I and II in the five strata are significantly different (Chi-square test: p-value < 0.001). This revealed a dependence of the test score's observation mechanism on median family income/residence area. Since socioeconomic factors, such as family income and living area, may affect children's cognitive development, the available test data were very likely biased. By noting this, the medical investigators of the study hesitated to draw any definitive conclusions [13]. In the following, we present a re-analysis of the data by applying the pseudoscore approach discussed in the previous sections. Inferences made in our analysis may be more acceptable than those obtained in [13].

Let Y_i be the difference in test score between subject i and his control for a particular test, and R_i indicate whether Y_i is observed or not, $i = 1, \ldots, 125$. We assumed that $Y \sim \mathrm{N}(\mu, \sigma^2)$ and $P(R = 1|Y, X) = \pi(X)$ with X representing the median family income of the subject's residence area. To evaluate the pseudoscore estimator for μ, we considered $\mathrm{logit}(\pi(x))$ to be (a) $\alpha_0 + \alpha_1 x$, (b) $\alpha_0 + \alpha_1 x + \alpha_2 x^2$, or (c) $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$, and estimated the parameters correspondingly based on the likelihood (8). The resulting estimates for $\pi(x)$ are given in Figure 1, labelled by 'logit linear', 'logit quadratic', and 'logit cubic', respectively. Likelihood ratio tests showed the following: (i) α_1 in Model (a) was significantly different from zero, which indicates a dependence of the observation pattern on median family income; (ii) by the likelihood ratio tests (LRT), the fitting with Model (b) improved the one with Model (a) significantly, but Model (c) did not improve much on Model (b). These are confirmed by Figure 1. The two curves of 'logit quadratic' and 'logit cubic' in the figure are very close. The common trend of the curves suggests that the study tended to obtain the test results more likely from subjects living in areas with high median family income. Figure 1 also presents the proportions of observation *vs* median family income (the dots), and a non-parametric estimate (a robust locally linear estimate of $\pi(x)$, obtained by the SPlus function 'lowess'). The non-parametric estimate is similar to both 'logit quadratic' and 'logit cubic'. This indicates that Model (b) (the logit quadratic model) was quite appropriate to the data.

We evaluated the pseudoscore estimator in Section 2.3 using the estimate of $\pi(x)$ under Model (b). Table V presents the estimates of the mean differences in test score (for both raw score and percentile) between sickle cell children and their healthy controls in mathematics, language, memory, and auditory discrimination ($\hat{\mu}$), the estimated standard deviations (se), and the p-values of the associated Z-tests of $H_0 : \mu = 0$. The results were obtained using the variance estimation given in Appendix A and the asymptotic normality of the estimator. For comparison, Table V also shows the corresponding results from the available data without any adjustment for the possibly sampling bias, i.e. using the naive approach.

Our analysis outcome, labelled with PSE in the table, indicates that on average the sickle cell children had lower test scores than the controls, across the different tests and regardless of raw score or national percentile. The difference was rather significant in auditory discrimination, but not so in mathematics, language, and memory. We see from the p-values that PSE and NAIVE in comparing sickle cell children and controls are quite consistent, except for national percentile for the language test. However, the estimated differences between sickle cell children and healthy controls by PSE tend to be larger than the ones by NAIVE. This suggests that the sickle cell children with observed test scores perform likely better in the tests than those without observed test scores, which was suspected before the analysis, since they tended to come from relatively stabler families or better residence areas. This indicates the necessity of the adjustment for the observation mechanism.

DOI: 10.1002/sim

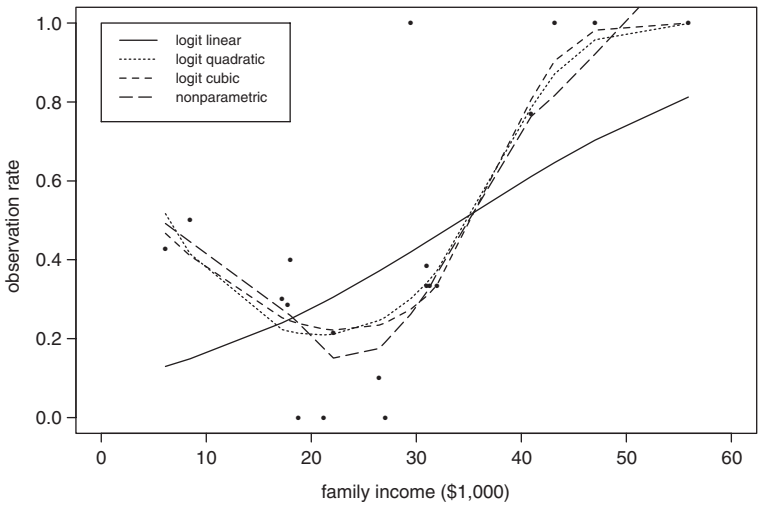

Figure 1. Estimates of observation probability conditional on family income: the dots (•) are the observed rates.

Table V. Analysis outcomes of the test scores from the sickle cell study.

	Mathematics			Language			Memory			Auditory discrimination		
	$\hat{\mu}^*$	se†	$p^\ddagger$	$\hat{\mu}$	se	p	$\hat{\mu}$	se	p	$\hat{\mu}$	se	p
Raw score												
NAIVE	−1.090	1.638	0.253	−1.47	1.179	0.106	−1.09	1.230	0.188	−2.98	1.092	0.003
PSE	−0.915	1.638	0.288	−1.55	1.219	0.102	−1.18	1.321	0.186	−3.47	1.219	0.002
Percentile												
NAIVE	−1.430	1.592	0.185	−2.45	1.469	0.048	−3.18	2.082	0.063	−6.84	1.908	<0.001
PSE	−2.400	4.184	0.283	−5.43	4.219	0.099	−9.32	6.061	0.062	−16.38	5.781	0.002

*Estimate of difference ($\mu = EY$) in test score between sickle cell children and their healthy controls.
†Estimate of standard deviation of the difference estimator ($\sqrt{\text{var}(\hat{\mu})}$).
‡P-value of Z-test based on $\hat{\mu}$ and se for $H_0 : \mu = 0$.

The standard errors associated with NAIVE, on the other hand, are similar to or smaller than the ones with PSE. This is due to the standard errors with PSE consisting of two parts, one of which accounts for the loss in efficiency resulting from the non-ignorable non-response. These two findings together may explain the consistency of PSE and NAIVE in comparing sickle cell children and controls by the Z-test, which is rather coincidental.

We noted that the majority of Group II, 77 out of 92 children (84.0 per cent), were those whose parents were not contacted successfully by the study staff or did not respond to the inquiry. We had no test results at all for this group. Their families may be relatively unstable, compared to the

DOI: 10.1002/sim

others. It is arguable to use the available test scores of Group I to recover the missing information for the whole Group II by any means. Thus, the assumption (1) may need to be justified further in the current situation. This can be addressed if we collect additional information about Group II, such as test results from matched controls of the group. Information such as parent education level for each sickle cell child may also help for further investigation of differences in cognitive development between sickle cell children and healthy controls.

5. CONCLUDING REMARKS

This paper presents pseudoscore-based estimation from possibly biased observations on the primary variable Y. We have studied the asymptotic properties and the finite sample behaviour of the estimation procedures. Our results indicate that the procedures perform well relative to the maximum likelihood estimation with the complete data. In addition, our approach can be easily implemented. We use the information contained in auxiliary variables X to adjust for possible bias in the observation of Y, by modelling how the Y observation mechanism (i.e. R) depends on X. It does not require specifying the underlying relationship between the primary variable Y and the auxiliary variables X. Model checking on the regression assumption $Y|X$ is hard in the situations with incomplete Y data. There are, however, various techniques available for checking models on $R|X$.

Several further investigations would be worthwhile. One is to study the performance of the approach described at the end of Section 2.3, which discretizes the continuous X and then applies the method for discrete covariates. It may be viewed as using a non-parametric estimator of $\pi(\cdot)$ to implement $PS(\theta; \pi)$ when π is unknown. The simulation results in Section 3 showed that this approach performs quite well. It would be of some technical difficulty to study the asymptotic properties of the resulting estimator for θ and its variance estimation without specifying $\pi(\cdot)$ into a parametric form.

Another one is to consider testing procedures rather than the Z-test type constructed based on the estimator and its estimated standard deviation, as used in Section 4. We have noted that a testing procedure can be derived directly from the pseudoscore function. However, it would not be straightforward for the situations involving nuisance parameters. Notice that the estimator of μ obtained from the approach when $Y \sim \mathrm{N}(\mu, \sigma^2)$ is of the form $\hat{\mu} = \sum_{i=1}^{N} \frac{R_i}{\pi(X_i)} Y_i / \sum_{i=1}^{N} \frac{R_i}{\pi(X_i)}$. This estimator and a slightly different version of it, $\tilde{\mu} = \sum_{i=1}^{N} \frac{R_i}{\pi(X_i)} Y_i / N$, are both consistent non-parametric estimators for $E(Y)$. A non-parametric testing procedure for test on population mean may be derived based on them.

The assumption (1) is crucial to the approach. Methods are needed for diagnosing the assumption. With little modification, the approach may be extended to situations where (1) does not hold but $P(R = 1|Y, X)$ is known.

Variations of pseudoscore-based estimation have been applied previously in connection with response-related incomplete data in regression problems. See, for example, [8–10]. The approach discussed in the paper can also be extended to estimating the regression parameters. Another extension is to consider estimating functions of the moment parameters of Y without specifying the distribution of Y, similar to the ones in [8, 15] for regression parameters. Finally, as suggested by a Referee, we may further study the efficiency of the proposed approach and consider the optimal estimating functions in the situations, following the discussion in [15] on the regression analysis.

DOI: 10.1002/sim

APPENDIX A: SOME TECHNICAL DETAILS

The (k, l) element of $\Pi(\theta; \alpha)$ in $\Xi(\theta; \alpha)$ (9) is

$$E\left\{-b_k(Y;\theta)\frac{R}{\pi(X;\alpha)}c(X;\alpha)'\right\}\Gamma(\alpha)^{-1}E\left\{-b_l(Y;\theta)\frac{R}{\pi(X;\alpha)}c(X;\alpha)\right\}$$

with $b_k(Y;\theta)$ the kth component of $\partial \log f(Y;\theta)/\partial\theta$ and $c(X;\alpha) = \partial \log \pi(X;\alpha)/\partial\alpha$, and the (k, l) element of $\Psi(\theta; \alpha)$ in (9) is

$$E\left\{b_k(Y;\theta)\frac{R}{\pi(X;\alpha)}\left(\frac{\partial \log h(X;\alpha)}{\partial\alpha}\right)'\right\}\Gamma(\alpha)^{-1}E\left\{-b_l(Y;\theta)\frac{R}{\pi(X;\alpha)}c(X;\alpha)\right\}$$

We obtain a consistent estimator of the asymptotic variance of $\sqrt{N}(\tilde{\theta} - \theta)$ (10) as follows. Consider the form

$$\hat{I}(\theta;\alpha)^{-1}\hat{\Xi}(\theta;\alpha)\hat{I}(\theta;\alpha)^{-1}$$

with θ and α replaced by their estimators $\tilde{\theta}$ and $\hat{\alpha}$, respectively, where $\hat{I}(\theta;\alpha)$ is as $\hat{I}(\theta;\pi)$ given in (7) with $\pi(x)$ replaced by $\pi(x;\alpha)$, and $\hat{\Xi}(\theta;\alpha)$ is $\hat{\Sigma}(\theta;\alpha) + \hat{\Pi}(\theta;\alpha) + 2\hat{\Psi}(\theta;\alpha)$ with $\hat{\Sigma}(\theta;\alpha)$ as $\hat{\Sigma}(\theta;\pi)$ given in (5) with $\pi(x) = \pi(x;\alpha)$, the (k, l) element of $\hat{\Pi}(\theta;\alpha)$ as

$$(\hat{B}_k(\theta;\alpha))'\hat{\Gamma}(\alpha)^{-1}\hat{B}_l(\theta;\alpha)$$

and the (k, l) element of $\hat{\Psi}(\theta;\alpha)$ as

$$(\hat{D}_k(\theta;\alpha))'\hat{\Gamma}(\alpha)^{-1}\hat{B}_l(\theta;\alpha)$$

with

$$\hat{B}_k(\theta;\alpha) = -\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi(X_i;\alpha)}b_k(Y_i;\theta)c(X_i;\alpha)$$

and

$$\hat{D}_k(\theta;\alpha) = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi(X_i;\alpha)}b_k(Y_i;\theta)\frac{\partial \log h(X_i;\alpha)}{\partial\alpha}$$

APPENDIX B: PROOF OF PROPOSITION 2

For simplicity, in the following we consider the case of one-dimensional θ. The multivariate extension should be straightforward.

Notice that the difference of the two pseudo-score functions $PS(\theta;\hat{\pi})$ and $PS(\theta;\pi)$ is

$$\Delta(\theta;\pi,\hat{\pi}) = (\alpha - \hat{\alpha})'\sum_{i=1}^{N}\frac{R_i}{\pi(X_i;\alpha)\pi(X_i;\hat{\alpha})}\frac{\partial\pi(X_i;\xi)}{\partial\alpha}b(Y_i;\theta)$$

DOI: 10.1002/sim

with ξ in between α and $\hat{\alpha}$. Provided that $\partial\pi(x;\alpha)/\partial\alpha$ is continuous with respect to α, we can show that $\Delta(\theta;\pi,\hat{\pi})/\sqrt{N}$ converges in distribution to the normal distribution with mean 0 and variance

$$\Pi(\theta;\alpha)=E\left\{-\frac{R}{\pi(X;\alpha)}c(X;\alpha)b(Y;\theta)\right\}'\Gamma(\alpha)^{-1}E\left\{-\frac{R}{\pi(X;\alpha)}c(X;\alpha)b(Y;\theta)\right\}$$

This, along with the asymptotic normality of $PS(\theta;\pi)/\sqrt{N}$, establishes the asymptotic normality of $PS(\theta;\hat{\pi})/\sqrt{N}$, of which the asymptotic variance is given in (9). The third term in (9) is two times the covariance of $PS(\theta;\pi)/\sqrt{N}$ and $\Delta(\theta;\pi,\hat{\pi})/\sqrt{N}$, which is obtained by noting

$$\begin{aligned}&\mathrm{Cov}\left\{\frac{1}{\sqrt{N}}PS(\theta;\pi),\frac{1}{\sqrt{N}}\Delta(\theta;\pi,\hat{\pi})\right\}\\&=\mathrm{Cov}\left\{\frac{1}{\sqrt{N}}PS(\theta;\pi),(\sqrt{N}(\hat{\alpha}-\alpha)')\left(-\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi(X_i;\alpha)\pi(X_i;\hat{\alpha})}\frac{\partial\pi(X_i;\xi)}{\partial\alpha}b(Y_i;\theta)\right)\right\}\end{aligned}$$

which converges to

$$\lim_{N\to\infty}\mathrm{Cov}\left\{\frac{1}{\sqrt{N}}PS(\theta;\pi),\sqrt{N}(\hat{\alpha}-\alpha)'\right\}E\left\{-\frac{R}{\pi(X;\alpha)}b(Y;\theta)c(X;\alpha)\right\}$$

and

$$\begin{aligned}&\mathrm{Cov}\left\{\frac{1}{\sqrt{N}}PS(\theta;\pi),\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial\log h(X_i;\alpha)}{\partial\alpha}\right\}\\&=\mathrm{Cov}\left\{\frac{1}{\sqrt{N}}PS(\theta;\pi),(\sqrt{N}(\hat{\alpha}-\alpha)')\left(-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2\log h(X_i;\xi)}{\partial\alpha^2}\right)\right\}\end{aligned}$$

converges to

$$E\left\{\frac{R}{\pi(X;\alpha)}\frac{\partial\log h(X;\alpha)}{\partial\alpha}b(Y;\theta)\right\}$$

and

$$\lim_{N\to\infty}\mathrm{Cov}\left(\frac{1}{\sqrt{N}}PS(\theta;\pi),\sqrt{N}(\hat{\alpha}-\alpha)'\right)\Gamma(\alpha)$$

We can then show (ii) of *Proposition* 2 by using approaches similarly to those used by Hu and Lawless [16].

ACKNOWLEDGEMENTS

This research was supported in part by grants from the National Institute of Allergy and Infectious Diseases of USA and from the Natural Sciences and Engineering Research Council of Canada. The authors are grateful to the associate editor and two anonymous referees for their constructive comments and suggestions, which lead to the current version of this paper.

DOI: 10.1002/sim

REFERENCES

1. Suzuki K. Estimation method of lifetime based on the record of failures during the warranty period. *Journal of the American Statistical Association* 1985; **80**:68–72.
2. Kalbfleisch JD, Lawless JF. Estimation of reliability from field performance studies (with Discussion). *Technometrics* 1988; **30**:365–388.
3. Hu XJ, Lawless JF. Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika* 1996; **83**:747–761.
4. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988; **75**:11–20.
5. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 1995; **82**:299–314.
6. Wild CJ. Fitting prospective regression models to case-control data. *Biometrika* 1991; **78**:705–717.
7. 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.
8. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
9. Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:413–438.
10. Chatterjee N, Chen YH, Breslow NE. A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* 2003; **98**:158–168.
11. Qin J, Leung D, Shao J. Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association* 2002; **97**:193–200.
12. Wang CY, Chen J, Lee S, Ou S. Joint conditional likelihood estimator in logistic regression with missing covariate data. *Statistica Sinica* 2002; **12**:555–574.
13. Steen RG, Hu XJ, Elliott VE, Miles MA, Jones S, Wang WC. Kindergarten readiness skills in children with sickle cell disease: evidence of early neurocognitive damage? *Journal of Child Neurology* 2002; **17**:111–116.
14. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, 2002.
15. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**:122–129.
16. Hu XJ, Lawless JF. Pseudolikelihood estimation in a class of problems with response-related missing covariates. *The Canadian Journal of Statistics* 1997; **25**:125–142.
17. Holt D, Smith TMF, Winter PD. Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, *Series B* 1980; **143**:474–487.
18. Binder DA, Patak Z. Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association* 1994; **89**:1035–1043.
19. Hsieh DA, Manski CF, McFadden D. Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association* 1985; **80**:651–662.
20. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society*, *Series B* 1986; **48**:170–182.
21. Whittemore AS. Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society*, *Series B* 1997; **59**:589–602.

DOI: 10.1002/sim