Pseudolikelihood Estimation in a Class of Problems with Response-Related Missing Covariates
Author(s): X. Joan Hu and Jerald F. Lawless
Source: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, Vol. 25, No. 2 (Jun., 1997), pp. 125-142
Published by: Statistical Society of Canada
Stable URL: http://www.jstor.org/stable/3315727
Accessed: 03/10/2008 17:25

The Canadian Journal of Statistics
Vol. 25, No. 2, 1997, Pages 125–142
La Revue Canadienne de Statistique

Pseudolikelihood estimation in a class of problems with response-related missing covariates*

X. Joan HU and Jerald F. LAWLESS

Harvard School of Public Health and *University of Waterloo*

Key words and phrases: Estimating function, incomplete data, supplementary information.
AMS 1991 subject classifications: 62F11; 62F12.

ABSTRACT

Many practical situations involve a response variable Y and covariates $\mathbf{X}$, where data on $(Y, \mathbf{X})$ are incomplete for some portion of a sample of individuals. We consider two general types of pseudolikelihood estimation for problems in which missingness may be response-related. These are typically simpler to implement than ordinary maximum likelihood, which in this context is semiparametric. Asymptotics for the pseudolikelihood methods are presented, and simulations conducted to investigate the methods for an important class of problems involving lifetime data. Our results indicate that for these problems the two methods are effective and comparable with respect to efficiency.

RÉSUMÉ

Plusieurs situations pratiques impliquent une variable de réponse Y et des covariables $\mathbf{X}$, où les données sur $(Y, \mathbf{X})$ sont incomplètes pour une certaine portion d'un échantillon d'individus. Nous considérons deux types généraux d'estimation de pseudo-vraisemblance pour des problèmes dans lesquels l'absence de données peut être reliée à la réponse. Ceux-ci sont typiquement plus faciles à mettre en pratique que le maximum de vraisemblance ordinaire, qui, dans ce contexte, est semiparamétrique. Nous présentons les propriétés asymptotiques pour les méthodes de pseudo-vraisemblance, ainsi que les simulations effectuées pour étudier les méthodes pour une classe importante de problèmes impliquant des données de durée de vie. Nos résultats indiquent que, pour ces problèmes, les deux méthodes sont efficaces et comparables quant à leur efficacité.

1. INTRODUCTION

1.1. General Introduction.

There are many practical situations involving a response variable Y and covariates $\mathbf{X}$ where the data on $(Y, \mathbf{X})$ are incomplete for some portion of a sample of individuals. This may happen because of systematic features of the data collection process, or because some observations are randomly missing. For example, see Little and Rubin (1987) for a general discussion and many examples of missing data, Kalbfleisch and Lawless (1988a) and Hu and Lawless (1996) for response-biased failure-time data, Breslow and Cain (1988) and Reilly and Pepe (1995) for incomplete data in case-control and other epidemiologic problems, Wild (1991) for categorical response data, and Rubin (1987)

*This research was supported in part by a fellowship to the first author and by grants to the second author from the Natural Sciences and Engineering Research Council of Canada. This paper is based on the first author's Ph.D. thesis, which received the Pierre Robillard Award for 1996.

for missing data in sample surveys. Heitjan and Rubin (1991) define the concept of coarsening to encompass a wide variety of incomplete data scenarios, including missing data, grouped data and censored data. There are several types of estimation procedures for incomplete data, including maximum-likelihood or Bayesian methods (e.g., Little and Rubin 1987, Heitjan and Rubin 1991), various types of pseudolikelihoods or estimating functions (e.g., Breslow and Cain 1988; Kalbfleisch and Lawless 1988a, b; Wild 1991; Reilly and Pepe 1995; Robins *et al.* 1994, 1995), and data augmentation or imputation (e.g., Rubin 1987; Tanner and Wong 1987). Carroll *et al.* (1995, Ch. 9) provide a lucid review of recent work.

Aside from a few important topics like retrospective or choice-based sampling with categorical responses (e.g., Hsieh *et al.* 1985, Breslow and Cain 1988, Wild 1991), there have been few comparisons of methods, and little guidance is available concerning best methods for specific types of problems. Robins *et al.* (1994, 1995) described a general methodology for asymptotically efficient estimation when covariate data are missing, but implementation and investigation of their methods in many practical contexts is difficult. The purpose of this article is to describe a general framework that includes many missing-data problems, and to discuss two quite general types of pseudolikelihood estimation. We then study a class of problems that arise in connection with lifetime data analysis, and compare ordinary and pseudolikelihood methods.

Suppose that (response, covariates) pairs $(Y, \mathbf{X})$ are generated from a distribution with density or probability function

$$f(y|\mathbf{x};\boldsymbol{\theta})g(\mathbf{x}), \qquad y \in \mathcal{Y}, \quad \mathbf{x} \in \mathcal{X}, \tag{1}$$

and that our objective is to make inferences about the $p \times 1$ parameter $\boldsymbol{\theta} \in \Theta$. We wish to avoid parametric assumptions about $g(\mathbf{x})$ and the corresponding distribution function $G(\mathbf{x})$, as is common in regression problems. We consider situations in which for a randomly generated set of observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, some or all of the covariate vector $\mathbf{x}_i$ is missing for a portion of the observations. We also allow for some incompleteness in the observation of Y: the observation on the ith response is assumed to be of the form $Y_i \in B_i \subseteq \mathcal{Y}$, where B_i may or may not depend on $\mathbf{x}_i$. If B_i depends on $\mathbf{x}_i$ and some or all components of $\mathbf{x}_i$ are missing, then B_i may not be fully known; this occurs in the example below and in the model of Section 1.2.

Example (Censored lifetime data with missing covariates). There are numerous situations involving lifetime data where covariate values are missing for individuals whose lifetimes y_i are censored. For example, in reliability studies of manufactured products under warranty, a product which fails at time y_i after sale is reported to the manufacturer and covariates $\mathbf{z}_i$ are recorded, provided that y_i is less than a censoring time τ_i which depends on the date of sale of the product and the length of the warranty coverage (e.g., Kalbfleisch and Lawless 1988a, Hu and Lawless 1996). However, if $y_i > \tau_i$, the values of the covariates $\mathbf{z}_i$ (and possibly also τ_i) are not known. Similar problems arise in epidemiology, for example when ages at onset of a childhood disease and associated covariates are obtained for children with onset before a certain time, but not for others. In the notation used above, $B_i = \{y : y > \tau_i\}$ for a censored lifetime. If τ_i is not observed, then B_i is not known.

Let O_1 denote the index set of observations for which information about $\mathbf{x}_i$ is not missing, and O_2 the index set of observations for which it is. Then $\{1, \ldots, n\} = O_1 \cup O_2$. We assume that the following likelihood function is valid:

$$L(\boldsymbol{\theta}, G) = \prod_{i \in O_1} \mathrm{P}(Y_i \in B_i|\mathbf{x}_i;\boldsymbol{\theta})\, dG(\mathbf{x}_i) \prod_{i \in O_2} \mathrm{P}(Y_i \in B_i|\mathbf{x}_i^{\mathrm{obs}};\boldsymbol{\theta})\, dG(\mathbf{x}_i^{\mathrm{obs}}), \tag{2}$$

where for $i \in O_2$, $\mathbf{x}_i$ is partitioned as $\mathbf{x}_i = (\mathbf{x}_i^{\mathrm{obs}}, \mathbf{x}_i^{\mathrm{mis}})$ to correspond to the parts of $\mathbf{x}_i$ that are observed and missing, respectively. For simplicity we assume that all observations of units in O_2 have the same components of $\mathbf{x}$ missing, and use $G(\cdot)$ to denote the distribution of either $\mathbf{x}_i$ or $\mathbf{x}_i^{\mathrm{obs}}$, it being clear from the argument which is referred to. Here and elsewhere we also often suppress notationally the fact that $\mathrm{P}(Y_i \in B_i)$ and other quantities depend on $G(\cdot)$ as well as $\boldsymbol{\theta}$. For a discussion of conditions under which (2) is valid, see Heitjan and Rubin (1991). Note that with regard to the observations with $i \in O_2$ we have

$$\mathrm{P}(Y_i \in B_i | \mathbf{x}_i^{\mathrm{obs}}; \boldsymbol{\theta}) = \int \mathrm{P}(Y_i \in B_i | \mathbf{x}_i; \boldsymbol{\theta})\, dG(\mathbf{x}_i^{\mathrm{mis}} | \mathbf{x}_i^{\mathrm{obs}}), \tag{3}$$

where $G(\mathbf{x}_i^{\mathrm{mis}} | \mathbf{x}_i^{\mathrm{obs}})$ is the conditional distribution of $\mathbf{x}_i^{\mathrm{mis}}$ given $\mathbf{x}_i^{\mathrm{obs}}$.

Although we are interested in estimation of $\boldsymbol{\theta}$, the terms (3) depend on the distribution $G(\cdot)$ of the covariates, and by writing $L(\boldsymbol{\theta}, G)$ in (2) we recognize this dependency. Semiparametric maximum likelihood, in which (2) is maximized jointly with respect to $\boldsymbol{\theta}$ and $G(\cdot)$, is usually complicated, and we seek other methods. We now describe two approaches that may be applied in many settings. First, write the logarithm of (2) as

$$l(\boldsymbol{\theta}; G) = \sum_{i \in O_1} \log \mathrm{P}(Y_i \in B_i | \mathbf{x}_i) + \sum_{i \in O_2} \log \mathrm{P}(Y_i \in B_i | \mathbf{x}_i^{\mathrm{obs}}) \tag{4}$$

plus a constant that depends only on $G(\cdot)$ and the $\mathbf{x}_i$'s. The semicolon in $l(\boldsymbol{\theta}; G)$ indicates that we wish to estimate $\boldsymbol{\theta}$ but that $G(\cdot)$ is a nuisance parameter. If $G(\cdot)$ were known, then we could use the estimating equation $\mathbf{S}(\boldsymbol{\theta}; G) = \partial l(\boldsymbol{\theta}; G)/\partial \boldsymbol{\theta} = \mathbf{0}$ to estimate $\boldsymbol{\theta}$. The first approach is to obtain an empirical estimate $\tilde{G}$ of G from the available data, and to employ the estimating equation

$$\mathbf{PS}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\theta}; \tilde{G}) = \mathbf{0}. \tag{5}$$

Generalizing the terminology of Gong and Samaniego (1981) for parametric problems, we refer to $l(\boldsymbol{\theta}; \tilde{G})$ as a pseudologlikelihood and to $\mathbf{PS}(\boldsymbol{\theta})$ of (5) as a pseudoscore function. This idea has been used in various contexts (e.g., Carroll and Wand 1991; Pepe and Fleming 1991).

The second approach is to replace the second term in (4) with another type of estimate, by considering the complete data loglikelihood

$$l_C(\boldsymbol{\theta}) = \sum_{i \in O_1} \log \mathrm{P}(Y_i \in B_i | \mathbf{x}_i) + \sum_{i \in O_2} \log \mathrm{P}(Y_i \in B_i | \mathbf{x}_i). \tag{6}$$

Some of the information needed for the second term is missing, so we estimate it, given the observed data. In many cases this is done by estimating the expectation of the term with missing data, given the observed data. This idea has also been employed in various contexts (e.g., Suzuki 1985a, b, Kalbfleisch and Lawless 1988a, b, Wild 1991, Reilly and Pepe 1995, Robins *et al.* 1994, 1995, Hu and Lawless 1996).

The incomplete data must arise in an appropriate way in order for either of the methods to be used. Moreover, there are different ways to implement the general approaches described, and conditions that ensure the proper behaviour of estimators will vary according to the model and type of data. The objective of this article is to study one type of problem in some detail: the situation considered is one in which a response time Y and associated covariates $\mathbf{X}$ for individuals in a population or sample are observed differentially according to whether Y satisfies a certain condition (e.g., is less than a certain value). This problem is worth studying for several reasons: it is of considerable practical importance; it allows approaches of the two general types described above; there are feasible "ordinary" likelihood methods with which the pseudolikelihood estimation techniques may be compared.

1.2. A Specific Model Involving Missing Data.

We now describe a framework that covers problems like that in the preceding section. For each individual i in a random sample or population $\mathcal{P} = \{1, \dots, M\}$, there is an associated set A_i, which may depend on $\mathbf{x}_i$. If y_i is not in A_i, then both y_i and $\mathbf{x}_i$ are observed; if y_i is in A_i then we know this, but we do not know the values of y_i, $\mathbf{x}_i$, and perhaps A_i. In many applications $A_i = \{y : y > \tau_i\}$, where τ_i is a censoring time that may be considered part of the covariate vector $\mathbf{x}_i$. Assuming that the pairs $(y_i, \mathbf{x}_i)$, $i = 1, \dots, M$, come from (1), these data, referred to as the primary data, produce the likelihood function

$$L_1(\boldsymbol{\theta}, G) = \prod_{i \in O_1} f(y_i|\mathbf{x}_i; \boldsymbol{\theta})\, dG(\mathbf{x}_i) \prod_{i \in O_2} \mathrm{P}(Y_i \in A_i; \boldsymbol{\theta}). \tag{7}$$

We denote by $|O_1| = m$ and $|O_2| = M - m$ the sizes of O_1 and O_2, respectively.

The primary data are often supplemented by an additional sample that provides information about $G(\cdot)$, especially in applications when m/M is small (Kalbfleisch and Lawless 1988a, b, Prentice 1986). We consider two types of supplementary data: (i) *followup samples*, where a random subsample S^f of n^f units in O_2 is selected, and their $\mathbf{x}_i$ values obtained; (ii) *independent samples*, where a random independent sample $S^s = \{\mathbf{x}_j : j = 1, \dots, n^s\}$ is obtained from $G(\mathbf{x})$ in some way. Both approaches are commonly used in fields such as reliability and epidemiology (e.g., Hsieh *et al.* 1985, Suzuki 1985a, b, Kalbfleisch and Lawless 1988a, b). The combined primary data and supplementary data produce the likelihood functions

$$L^c(\boldsymbol{\theta}, G) = L_1(\boldsymbol{\theta}, G) \prod_{i \in S^f} \frac{\mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})\, dG(\mathbf{x}_i)}{\mathrm{P}(Y_i \in A_i; \boldsymbol{\theta})} \tag{8}$$

$$= \prod_{i \in O_1} f(y_i|\mathbf{x}_i; \boldsymbol{\theta})\, dG(\mathbf{x}_i)$$

$$\times \prod_{i \in S^f} \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})\, dG(\mathbf{x}_i) \prod_{i \in O_2 \setminus S^f} \mathrm{P}(Y_i \in A_i; \boldsymbol{\theta}) \tag{9}$$

and

$$L^s(\boldsymbol{\theta}, G) = L_1(\boldsymbol{\theta}, G) \prod_{j \in S^s} dG(\mathbf{x}_j), \tag{10}$$

respectively, for the cases of followup and independent supplementary samples. Note that (9) and (10) are both special cases of the likelihood (2).

The remainder of the paper is as follows. In Section 2 we describe how to make inferences about $\boldsymbol{\theta}$ based on the two types of data. In particular, pseudolikelihood estimating equations are obtained for each situation, and the possibilities for maximum-likelihood estimation are also noted. Section 3 provides asymptotic theory for the pseudolikelihood methods. In Section 4 we consider a class of censored response-time problems like those described in the preceding example, in which case $A_i = \{y : y > \tau_i\}$; simulation is used to study and compare different methods of estimation. Section 5 concludes with some comments on related work.

2. LIKELIHOOD AND PSEUDOLIKELIHOOD ESTIMATION

We now describe several approaches to the estimation of $\boldsymbol{\theta}$ under the observation schemes described in Section 1.2. We consider the case where there are no supplementary data, and then each type of supplementary data.

2.1. No Supplementary Data.

If there are no supplementary data, then one way to estimate $\boldsymbol{\theta}$ is to use only the data of the items in O_1, and consider a likelihood function based on the distribution of Y_i, given that $Y_i \notin A_i$ and $\mathbf{x}_i$:

$$L_T(\boldsymbol{\theta}) = \prod_{i \in O_1} \frac{f(y_i|\mathbf{x}_i;\, \boldsymbol{\theta})}{\mathrm{P}(Y_i \notin A_i|\mathbf{x}_i;\, \boldsymbol{\theta})}. \tag{11}$$

In many contexts, the estimates obtained by maximizing (11) are too imprecise to be very useful (e.g., Kalbfleisch and Lawless 1988a, Hu and Lawless 1996), and in some cases (e.g., Wild 1991) they are completely uninformative.

Another approach is to maximize (7) jointly with respect to $\boldsymbol{\theta}$ and $G(\cdot)$. The profile likelihood $L_1(\boldsymbol{\theta}, \hat{G}(\cdot;\boldsymbol{\theta}))$, where $\hat{G}(\cdot;\boldsymbol{\theta})$ is the nonparametric maximum-likelihood estimate of $G(\cdot)$ based on (7) for given $\boldsymbol{\theta}$, can be used for inferences about $\boldsymbol{\theta}$. This estimation procedure is computationally complex and in contexts like those of Section 4 does not perform much better than (11).

In many practical situations it is highly desirable to have supplementary information about $G(\cdot)$, in order to get reasonably precise inferences about $\boldsymbol{\theta}$. We now discuss the two situations introduced in Section 1.2.

2.2. With Supplementary Data.

(a) Construction of pseudolikelihood estimating functions.

In Section 1.1 two approaches to the construction of estimating functions were described; we now apply them to the problems described in Section 1.2. The first approach is the one represented by (5). We take the score function for $\boldsymbol{\theta}$ based on (7), $\mathbf{S}(\boldsymbol{\theta};G)$, and insert an estimate $\tilde{G}(\cdot)$ for $G(\cdot)$ that is based on observed data, but not necessarily obtained by maximization of any of (7), (9) or (10). This produces an estimating function of the form (5):

$$\begin{aligned}\mathbf{PS}(\boldsymbol{\theta}) &= \mathbf{S}(\boldsymbol{\theta};\tilde{G}) \\ &= \sum_{i \in O_1} \frac{\partial \log f(y_i|\mathbf{x}_i;\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &\quad + \sum_{i \in O_2} \frac{\partial \log \int \mathrm{P}(Y_i \in A_i|\mathbf{x}_i = \mathbf{x};\, \boldsymbol{\theta})\, d\tilde{G}(\mathbf{x})}{\partial \boldsymbol{\theta}}. \end{aligned} \tag{12}$$

Note that $\mathbf{S}(\boldsymbol{\theta};G)$ is also the $\boldsymbol{\theta}$-score for the likelihood (10) based on the primary and independent supplementary data. It is not the $\boldsymbol{\theta}$-score for the analogous likelihood (9) in the case of supplementary followup data. However, because it has a simpler form and performs well, we use (12) with both types of supplementary data.

The second approach discussed in Section 1.1 is based on estimation of a "complete" data loglikelihood (6), obtained by replacing the terms $\mathrm{P}(Y_i \in A_i; \boldsymbol{\theta})$, $i \in O_2$, in (7) with $\mathrm{P}(Y_i \in A_i | \mathbf{x}_i; \boldsymbol{\theta})$. This gives the score function

$$\mathbf{CS}(\boldsymbol{\theta}) = \sum_{i \in O_1} \frac{\partial \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{i \in O_2} \frac{\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{13}$$

$$= \sum_{i \in O_1} \left(\frac{\partial \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)$$

$$+ \sum_{i \in \mathcal{P}} \frac{\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{14}$$

where the second term of (14), denoted by $\mathbf{U}(\boldsymbol{\theta})$, involves missing data. An estimating function of the form

$$\mathbf{ES}(\boldsymbol{\theta}) = \sum_{i \in O_1} \left\{ \frac{\partial \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\} + \tilde{\mathbf{U}}(\boldsymbol{\theta}) \tag{15}$$

is then used, where

$$\tilde{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{P}} \int \frac{\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i = \mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \, d\tilde{G}(\mathbf{x}) \tag{16}$$

with $\tilde{G}(\cdot)$ an estimate of $G(\cdot)$, as earlier.

(b) Implementation with supplementary followup data.

A random followup sample of individuals in O_2 can be taken in various ways. We consider a simple random sample without replacement of $n^f = p^f(M - m)$ items from the $M - m$ ones in O_2; p^f is a prespecified positive constant. The methods below may be adapted to other schemes.

To implement (12), we consider the estimator

$$\tilde{G}^f(\mathbf{x}) = \frac{m}{M} \hat{G}(\mathbf{x}|Y_i \notin A_i) + \frac{M - m}{M} \hat{G}_f(\mathbf{x}|Y_i \in A_i), \tag{17}$$

where $\hat{G}(\mathbf{x}|Y_i \notin A_i)$ and $\hat{G}_f(\mathbf{x}|Y_i \in A_i)$ are the empirical distribution functions based on $\{\mathbf{x}_i : i \in O_1\}$ and $\{\mathbf{x}_i : i \in S^f\}$, respectively. Inserting $\tilde{G}^f(\cdot)$ in (12) yields $\mathbf{PS}(\boldsymbol{\theta})$.

To implement (15) we use (16) with $\tilde{G}(\cdot)$ given by (17). This gives the same result as if we estimated the second term of (13) by the average over S^f of $\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$. The pseudoscore function obtained is

$$\mathbf{ES}^f(\boldsymbol{\theta}) = \sum_{i \in O_1} \frac{\partial \log f(y_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{M - m}{n^f} \sum_{i \in S^f} \frac{\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{18}$$

Maximum-likelihood methods can also be used to estimate $\boldsymbol{\theta}$. One approach is to maximize (9) jointly with respect to $\boldsymbol{\theta}$ and $G(\cdot)$. This is rather complex, though an extension of the method of Scott and Wild (1997) may be applied. An alternative (e.g., Kalbfleisch and Lawless 1988a, Wild 1991) is obtained by noting that a likelihood function is available for a related situation, and may be used more generally to give estimating functions. If each item in O_2 were sampled with probability p^f independently [Kalbfleisch and Lawless (1988a) call this a Bernoulli sampling scheme], we could obtain an exact conditional likelihood based on the probability of the ith datum, given that item i is in either O_1 or S^f:

$$L^f(\boldsymbol{\theta}) = \prod_{i \in \mathcal{P}} \frac{f(y_i|\mathbf{x}_i)^{R_{1i}}\{p^f \mathrm{P}(Y_i \in A_i|\mathbf{x}_i)\}^{R_{2i}}}{\{\mathrm{P}(Y_i \notin A_i|\mathbf{x}_i) + p^f \mathrm{P}(Y_i \in A_i|\mathbf{x}_i)\}^{R_i}}, \tag{19}$$

where $R_i = R_{1i} + R_{2i}$, and R_{1i} and R_{2i} are indicators of whether i is in O_1 and S^f, respectively. It can be verified that $\partial \log L^f/\partial\boldsymbol{\theta}$ remains an unbiased estimating function for $\boldsymbol{\theta}$ under random sampling without replacement. Inferences on $\boldsymbol{\theta}$ can then be drawn from it.

(c) Implementation with independent supplementary data.

Suppose that $\mathbf{x}_j, j \in S^s$, is a random sample arising from $G(\mathbf{x})$, with size $n^s = p^s M > 0$. Note that p^s may be greater than 1; $n^s \longrightarrow \infty$ leads to the ideal case where $G(\cdot)$ is known.

For the first approach we replace the $\tilde{G}(\cdot)$ in (12) with the empirical distribution

$$\tilde{G}^s(\mathbf{x}) = \frac{1}{n^s} \sum_{j \in S^s} \mathrm{I}(\mathbf{x}_j \leq \mathbf{x}). \tag{20}$$

For the second approach we use the independent supplementary sample to estimate the sample total $\mathbf{U}(\boldsymbol{\theta})$ as in (14), and thus obtain an estimating function from (15).

We do not actually require independence of S^s and $\mathcal{P}$; S^s could be selected randomly from $\mathcal{P}$ (e.g., Prentice 1986). Other sampling mechanisms, such as stratified sampling, also work when suitable modifications to estimates of $G(\cdot)$ are made.

An alternative likelihood method would be to consider the profile likelihood $L^s(\boldsymbol{\theta}, \hat{G}(\cdot;\boldsymbol{\theta}))$ based on the combined likelihood function (10), where $\hat{G}(\mathbf{x};\boldsymbol{\theta})$ is the nonparametric maximum-likelihood estimate of $G(\cdot)$ from (10).

In the interests of brevity we will present theoretical and simulation results below only for the case of supplementary followup data. Results for independent supplementary data are available in Hu (1995) or from the authors.

3. ASYMPTOTIC PROPERTIES AND VARIANCE ESTIMATION

3.1. General Asymptotic Results for Pseudolikelihood.

First we present general results for estimators obtained from (12) or (15). Suppose that $\boldsymbol{\theta}_0$ is the true value of the parameter and that it lies in some open subset $\mathcal{A}$ of Θ. Assume that $0 < \mathrm{P}(Y_i \in A_i;\boldsymbol{\theta}) < 1$ if $\boldsymbol{\theta} \in \mathcal{A}$ and that $\tilde{\boldsymbol{\theta}}_{\mathrm{PS}}$ and $\tilde{\boldsymbol{\theta}}_{\mathrm{ES}}$, obtained from $\mathbf{PS}(\boldsymbol{\theta}) = \mathbf{0}$ and $\mathbf{ES}(\boldsymbol{\theta}) = \mathbf{0}$, are in $\mathcal{A}$ almost surely. Finally, denote $\mathrm{P}(\mathbf{x};\boldsymbol{\theta}) = \mathrm{P}(Y_i \in A_i|\mathbf{x}_i = \mathbf{x};\boldsymbol{\theta})$ and $\mathrm{P}(\boldsymbol{\theta}) = \mathrm{P}(Y_i \in A_i;\boldsymbol{\theta})$; both are independent of $i = 1,\ldots,M$, since the $(Y_i, \mathbf{X}_i)$'s are assumed to be i.i.d.

The theorems below establish the consistency and asymptotic normality of $\tilde{\boldsymbol{\theta}}_{\mathrm{PS}}$ and $\tilde{\boldsymbol{\theta}}_{\mathrm{ES}}$; proofs are sketched in the Appendix. In the following we use $\partial^l h/\partial\boldsymbol{\theta}^l$ to represent the lth derivative of $h(\boldsymbol{\theta})$, where $\partial^0 h/\partial\boldsymbol{\theta}^0 = h(\boldsymbol{\theta})$, and $Cov(\mathbf{W}, V)$ to represent the vector

$$(Cov(W_1, V),\ Cov(W_2, V), \ldots, Cov(W_q, V))^{\mathrm{T}},$$

where $\mathbf{W} = (W_1, W_2, \ldots, W_q)^{\mathrm{T}} \in \mathcal{R}^q$ is a random vector ($\{\cdot\}^{\mathrm{T}}$ denotes the matrix transpose), and V a random variable.

THEOREM 1. *Under assumptions* (A1)–(A5) *of the Appendix, if an estimate* $\tilde{G}(\cdot)$ *satisfies*

$$\frac{\partial^l \tilde{\mathrm{P}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^l} - \frac{\partial^l \mathrm{P}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^l} \xrightarrow{\text{a.s.}} \mathbf{0}, \qquad \text{as} \quad M \longrightarrow \infty, \tag{21}$$

for $l = 0, 1, 2, 3$, *where* $\tilde{\mathrm{P}}(\boldsymbol{\theta}) = \int \mathrm{P}(\mathbf{x}; \boldsymbol{\theta})\, d\tilde{G}(\mathbf{x})$, *then* $\tilde{\boldsymbol{\theta}}_{\mathrm{PS}} \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$.

THEOREM 2. *In addition to assumptions* (A1)–(A5) *and the condition* (21), *suppose that*

$$\sqrt{M}\left(\frac{\partial^l \tilde{P}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^l} - \frac{\partial^l P(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^l}\right) \xrightarrow{\text{a.s.}} N(\mathbf{0}, \boldsymbol{\Pi}_{\mathrm{PS},l}) \qquad \text{as} \quad M \longrightarrow \infty, \tag{22}$$

with $l = 0, 1$. *Then* $\sqrt{M}(\tilde{\boldsymbol{\theta}}_{\mathrm{PS}} - \boldsymbol{\theta}_0)$ *is asymptotically normal as* $M \longrightarrow \infty$, *with covariance matrix*

$$\boldsymbol{\Sigma}_{\mathrm{PS}}^{-1} + \boldsymbol{\Sigma}_{\mathrm{PS}}^{-1}(2\boldsymbol{\Xi}_{\mathrm{PS}} + \boldsymbol{\Pi}_{\mathrm{PS}})\boldsymbol{\Sigma}_{\mathrm{PS}}^{-1}, \tag{23}$$

where

$$\boldsymbol{\Sigma}_{\mathrm{PS}} = \lim_{M\to\infty} \frac{1}{M}\, \mathcal{V}ar\ \mathbf{S}(\boldsymbol{\theta}_0; G), \qquad \boldsymbol{\Xi}_{\mathrm{PS}} = \boldsymbol{\Xi}_{\mathrm{PS},1} - \boldsymbol{\Xi}_{\mathrm{PS},0}\mathbf{B}(\boldsymbol{\theta}_0)^{\mathrm{T}}$$

with $\boldsymbol{\Xi}_{\mathrm{PS},l} = \lim_{M\to\infty} \mathcal{C}ov\{\mathbf{S}(\boldsymbol{\theta}_0; G), \partial^l \tilde{P}(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^l\}$ *and* $\mathbf{B}(\boldsymbol{\theta}_0) = \partial \log P(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$, *and*

$$\boldsymbol{\Pi}_{\mathrm{PS}} = \boldsymbol{\Pi}_{\mathrm{PS},1} - 2\boldsymbol{\Pi}_{\mathrm{PS},01}\mathbf{B}(\boldsymbol{\theta}_0)^{\mathrm{T}} + \boldsymbol{\Pi}_{\mathrm{PS},0}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)^{\mathrm{T}}$$

with $\boldsymbol{\Pi}_{\mathrm{PS},l} = \lim_{M\to\infty} M\ \mathcal{V}ar\{\partial^l \tilde{P}(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^l\}$ *and* $\boldsymbol{\Pi}_{\mathrm{PS},01} = \lim_{M\to\infty} M\ \mathcal{C}ov\{\partial \tilde{P}(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}, \tilde{P}(\boldsymbol{\theta}_0)\}$.

THEOREM 3. *Under assumptions* (A1)–(A5) *of the Appendix, if*

$$\frac{\partial^l \boldsymbol{\Delta}_{\mathrm{ES}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^l} \xrightarrow{\text{a.s.}} \mathbf{0} \qquad \text{as} \quad M \longrightarrow \infty, \tag{24}$$

with $l = 0, 1, 2$, *where*

$$\boldsymbol{\Delta}_{\mathrm{ES}}(\boldsymbol{\theta}) = \frac{1}{M}\{\mathbf{ES}(\boldsymbol{\theta}) - \mathbf{CS}(\boldsymbol{\theta})\}, \tag{25}$$

then $\tilde{\boldsymbol{\theta}}_{\mathrm{ES}} \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0$.

THEOREM 4. *In addition to assumptions* (A1)–(A5) *and the condition* (24), *suppose that*

$$\sqrt{M}\boldsymbol{\Delta}_{\mathrm{ES}}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Pi}_{\mathrm{ES}}) \qquad \text{as} \quad M \longrightarrow \infty. \tag{26}$$

Then $\sqrt{M}(\tilde{\boldsymbol{\theta}}_{\mathrm{ES}} - \boldsymbol{\theta}_0)$ *is asymptotically normal as* $M \longrightarrow \infty$, *with covariance matrix*

$$\boldsymbol{\Sigma}_{\mathrm{ES}}^{-1} + \boldsymbol{\Sigma}_{\mathrm{ES}}^{-1}[2\boldsymbol{\Xi}_{\mathrm{ES}} + \boldsymbol{\Pi}_{\mathrm{ES}}]\boldsymbol{\Sigma}_{\mathrm{ES}}^{-1}, \tag{27}$$

where

$$\boldsymbol{\Sigma}_{\mathrm{ES}} = \lim_{M\to\infty} \frac{1}{M}\, \mathcal{V}ar\ \mathbf{CS}(\boldsymbol{\theta}_0), \qquad \boldsymbol{\Xi}_{\mathrm{ES}} = \lim_{M\to\infty} \mathcal{C}ov\{\mathbf{CS}(\boldsymbol{\theta}_0), \boldsymbol{\Delta}_{\mathrm{ES}}(\boldsymbol{\theta}_0)\},$$

and

$$\mathbf{\Pi}_{\mathbf{ES}} = \lim_{M\to\infty} M \,\mathcal{V}ar\, \mathbf{\Delta}_{\mathbf{ES}}(\boldsymbol{\theta}_0).$$

We remark that the second term in (23) and (27) may be interpreted as the price for using $\tilde{G}(\cdot)$ and $\tilde{U}(\cdot)$, respectively. It vanishes in the ideal cases where either $G(\cdot)$ is known or all $\mathbf{x}_i$, $i \in \mathcal{P}$, are observed.

3.2. Specific Results and Variance Estimation.

In this subsection, we check the conditions for the asymptotic properties of the estimators proposed in Section 2.2, and consider estimation of their asymptotic variances.

(a) Estimators $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$ Based on $\mathbf{PS}(\boldsymbol{\theta})$ of (12)

First, when $\tilde{G}(\cdot)$ is either $\tilde{G}^f(\cdot)$ of (17) or $\tilde{G}^s(\cdot)$ of (20), it is easy to check that the conditions (21) and (22) are satisfied. This is because the two estimates of $G(\cdot)$ are a linear combination of empirical distributions and an empirical distribution, respectively, and $n^f = p^f(M-m)$ and $n^s = p^s M$. The consistency and the asymptotic normality of the estimator $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$ based on (12) for the two special cases are then ensured by Theorems 1 and 2.

To estimate the asymptotic variance (23) of $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$, we notice that $\mathbf{\Sigma}_{\mathbf{PS}}$ and $\partial^l P(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^l$, $l = 1, 2$, can be consistently estimated by

$$\tilde{\mathbf{\Sigma}}_{\mathbf{PS},\tilde{G}}(\boldsymbol{\theta}) = -\frac{1}{M}\frac{\partial \mathbf{PS}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \quad\text{and}\quad \frac{\partial^l \tilde{\mathrm{P}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^l} = \frac{\partial^l \int \mathrm{P}(Y_i \notin A_i|\mathbf{x}_i = \mathbf{x};\boldsymbol{\theta})\, d\tilde{G}(\mathbf{x})}{\partial\boldsymbol{\theta}^l},$$

respectively, evaluated at $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$, and with $\tilde{G}(\mathbf{x})$ replaced by $\tilde{G}^f(\mathbf{x})$ or $\tilde{G}^s(\mathbf{x})$. In the following, we give consistent estimators for $\mathbf{\Pi}_{\mathbf{PS},l}$, $\mathbf{\Pi}_{\mathbf{PS},01}$, and $\mathbf{\Xi}_{\mathbf{PS},l}$, $l = 0, 1$, in (23) for the case of followup data. Consistent estimators for the asymptotic covariance matrices (23) of the estimators are thus obtained. We use the notation

$$\mathcal{V}ar\ \mathbf{W}|_{G=\tilde{G}} = \int \mathbf{W}\mathbf{W}^{\mathrm{T}}\, d\tilde{G} - \int \mathbf{W}\, d\tilde{G}\int \mathbf{W}^{\mathrm{T}}\, d\tilde{G}$$

and

$$\mathcal{C}ov(\mathbf{W},\mathbf{V})|_{G=\tilde{G}} = \int \mathbf{W}\mathbf{V}^{\mathrm{T}}\, d\tilde{G} - \int \mathbf{W}\, d\tilde{G}\int \mathbf{V}^{\mathrm{T}}\, d\tilde{G},$$

where $\mathbf{W}$ and $\mathbf{V}$ are functions of a random vector with distribution $G(\cdot)$.

Variance estimates for $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$ are somewhat messy. We have

$$\frac{\partial^l \tilde{P}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^l} = \frac{1}{M}\sum_{i\in\mathcal{P}}\left(R_{i1} + \frac{R_{i2}}{p^f}\right)\mathbf{v}_{i,l}(\boldsymbol{\theta}), \tag{28}$$

where $\mathbf{v}_{i,l}(\boldsymbol{\theta}) = \partial^l \mathrm{P}(Y_i \in A_i|\mathbf{x}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}^l$, $l = 0, 1$; $R_{i1} = \mathrm{I}(i \in O_1)$ and $R_{i2} = \mathrm{I}(i \in S^f)$, $i \in \mathcal{P}$. We can see that $\mathbf{\Pi}_{\mathbf{PS},l}$, $\mathbf{\Pi}_{\mathbf{PS},01}$ and $\mathbf{\Xi}_{\mathbf{PS},l}$ are estimated consistently by $\tilde{\mathbf{\Pi}}_{\mathbf{PS},l}(\tilde{\boldsymbol{\theta}}_{\mathbf{PS}})$, $\tilde{\mathbf{\Pi}}_{\mathbf{PS},01}(\tilde{\boldsymbol{\theta}}_{\mathbf{PS}})$ and $\tilde{\mathbf{\Xi}}_{\mathbf{PS},l}(\tilde{\boldsymbol{\theta}}_{\mathbf{PS}})$, where

$$\tilde{\mathbf{\Pi}}_{\mathbf{PS},l}(\boldsymbol{\theta}) = \mathcal{V}ar\{\mathbf{V}_l(\boldsymbol{\theta})\}|_{G=\tilde{G}^f} + \left(1 - \frac{m}{M}\right)\frac{1-p^f}{p^f}\,\widetilde{\mathrm{var}}\{\mathbf{V}_l(\boldsymbol{\theta})|Y_i \in A_i\} \tag{29}$$

with $\mathbf{V}_l(\boldsymbol{\theta}) = \mathbf{V}_l(Y_i, \mathbf{X}_i;\boldsymbol{\theta}) = \partial^l \mathrm{P}(Y_i \notin A_i|\mathbf{X}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}^l$, $\widetilde{\mathrm{var}}\{\mathbf{V}_l(\boldsymbol{\theta})|Y_i \in A_i\} = \sum_{i\in S^f}\{\mathbf{v}_{i,l}(\boldsymbol{\theta}) - \bar{\tilde{\mathbf{v}}}_l(\boldsymbol{\theta})\}\{\mathbf{v}_{i,l}(\boldsymbol{\theta}) - \bar{\tilde{\mathbf{v}}}_l(\boldsymbol{\theta})\}^{\mathrm{T}}/(n^f - 1)$, and $\bar{\tilde{\mathbf{v}}}_l(\boldsymbol{\theta}) = \sum_{i\in S^f}\mathbf{v}_{i,l}(\boldsymbol{\theta})/n^f$, $l = 0, 1$;

$$\tilde{\Pi}_{\mathrm{PS},01}(\boldsymbol{\theta}) = Cov\{\mathbf{V}_1(\boldsymbol{\theta}), V_0(\boldsymbol{\theta})\}|_{G=\tilde{G}^f} + \left(1-\frac{m}{M}\right)\frac{1-p^f}{p^f}\,\widetilde{\mathrm{cov}}\{\mathbf{V}_1(\boldsymbol{\theta}), V_0(\boldsymbol{\theta})|Y_i \in A_i\} \quad (30)$$

with $\widetilde{\mathrm{cov}}\{\mathbf{V}_1(\boldsymbol{\theta}), V_0(\boldsymbol{\theta})|Y_i \in A_i\} = \sum_{i\in S^f}\{\mathbf{v}_{i,1}(\boldsymbol{\theta}) - \bar{\mathbf{v}}_1(\boldsymbol{\theta})\}\{\mathbf{v}_{i,0}(\boldsymbol{\theta}) - \bar{\mathbf{v}}_0(\boldsymbol{\theta})\}^{\mathrm{T}}/(n^f-1)$;

$$\tilde{\boldsymbol{\Xi}}_{\mathrm{PS},l}(\boldsymbol{\theta}) = \frac{1}{M-1}\Bigg[\sum_{i\in O_1}\mathbf{u}_i(\boldsymbol{\theta})\mathbf{v}_{i,l}(\boldsymbol{\theta})^{\mathrm{T}} + \tilde{\mathbf{u}}(\boldsymbol{\theta})\sum_{i\in S^f}\frac{1}{p^f}\mathbf{v}_{i,l}(\boldsymbol{\theta})^{\mathrm{T}} - \frac{1}{M}\mathbf{PS}(\boldsymbol{\theta})\sum_{i\in\mathcal{P}}\left(R_{i1}+\frac{R_{i2}}{p^f}\right)\mathbf{v}_{i,l}(\boldsymbol{\theta})^{\mathrm{T}}\Bigg] \quad (31)$$

with $\mathbf{u}_i(\boldsymbol{\theta}) = \partial \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and $\tilde{\mathbf{u}}(\boldsymbol{\theta}) = \partial \log \tilde{P}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\,|_{\tilde{G}=\tilde{G}^f}$, $l = 0, 1$.

(b) Estimators $\tilde{\boldsymbol{\theta}}_{\mathrm{ES}}$ based on $\mathbf{ES}(\boldsymbol{\theta})$ *of* (15).

It is easy to check that $\tilde{G}^f(\cdot)$ and $\tilde{G}^s(\cdot)$ also satisfy the conditions (24) and (26) when $p^f \neq 0$ and $p^s \neq 0$, respectively, since $\boldsymbol{\Delta}_{\mathrm{ES}}$ in (25) is a linear combination of sums of i.i.d. random variables in each of the two cases. Then the estimator $\tilde{\boldsymbol{\theta}}_{\mathrm{ES}}$ is consistent under the limit $M \longrightarrow \infty$, and $\sqrt{M}(\tilde{\boldsymbol{\theta}}_{\mathrm{ES}} - \boldsymbol{\theta}_0)$ are asymptotically normal with asymptotic variance matrix of form (27).

Notice that $\boldsymbol{\Sigma}_{\mathrm{ES}}$ in (27) can be consistently estimated by

$$-\frac{1}{M}\frac{\partial\mathbf{ES}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_{\mathrm{ES}}}$$

with $\tilde{G}(\cdot)$ replaced by either $\tilde{G}^f(\cdot)$ or $\tilde{G}^s(\cdot)$, and $\tilde{\boldsymbol{\theta}}_{\mathrm{ES}}$ by a corresponding estimate, respectively. We give consistent estimators of $\boldsymbol{\Pi}_{\mathrm{ES}}$ and $\boldsymbol{\Xi}_{\mathrm{ES}}$ for the case of followup data.

We have $\sqrt{M}\,\boldsymbol{\Delta}_{\mathrm{ES}}(\boldsymbol{\theta}) = M^{-\frac{1}{2}}\sum_{i\in\mathcal{P}}(R_{i2}/p^f - 1)(1-R_{i1})\partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Its asymptotic variance matrix $\boldsymbol{\Pi}_{\mathrm{ES}}$ is consistently estimated by $\tilde{\boldsymbol{\Pi}}_{\mathrm{ES}^f}(\tilde{\boldsymbol{\theta}}_{\mathrm{ES}})$ with

$$\tilde{\boldsymbol{\Pi}}_{\mathrm{ES}^f}(\boldsymbol{\theta}) = \left(1-\frac{m}{M}\right)\frac{1-p^f}{p^f}\,\widetilde{\mathrm{var}}\{\mathbf{V}(\boldsymbol{\theta})|Y_i \in A_i\}, \quad (32)$$

where $\widetilde{\mathrm{var}}\{\mathbf{V}(\boldsymbol{\theta})|Y_i \in A_i\} = \sum_{i\in S^f}\{\mathbf{v}_i(\boldsymbol{\theta}) - \bar{\mathbf{v}}(\boldsymbol{\theta})\}\{\mathbf{v}_i(\boldsymbol{\theta}) - \bar{\mathbf{v}}(\boldsymbol{\theta})\}^{\mathrm{T}}/(n^f-1)$ with $\mathbf{V}(\boldsymbol{\theta}) = \partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $\mathbf{v}_i(\boldsymbol{\theta}) = \partial \log \mathrm{P}(Y_i \in A_i|\mathbf{x}_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and $\bar{\mathbf{v}}(\boldsymbol{\theta}) = \sum_{i\in S^f}\mathbf{v}_i(\boldsymbol{\theta})/n^f$. Because $\mathrm{E}\{\boldsymbol{\Delta}_C(\boldsymbol{\theta})|(y_i,\mathbf{x}_i) : i \in \mathcal{P}\} = \mathbf{0}$, we have $\boldsymbol{\Xi}_{\mathrm{ES}} = \mathbf{0}$ in this case.

4. A COMPARISON OF METHODS

To study the behaviour of the proposed estimators, and make a comparison of the several methods in Section 2, we performed some simulations for a class of incomplete response-time problems. This is a special case of the setup in Section 1.2. In the following, we describe the class of problems first, and then present the simulation results.

4.1. An Incomplete Response-Time Problem.

We consider the situation where y_i represents a failure time, and y_i and covariates $\mathbf{z}_i$ are observed if and only if y_i does not exceed an associated censoring time τ_i. If $y_i > \tau_i$, then this fact is known but the values of y_i, τ_i and $\mathbf{z}_i$ are not. Problems falling into this framework arise in epidemiology and reliability. For example, one way in which the field reliability of manufactured products is assessed is through failures reported under warranty. Suppose item i is sold at calendar time d_i and covered by a warranty scheme which ensures that a failure occurring before time $d_i + W_i$ is reported. If we consider a population of M items sold by time D, then y_i and associated covariates are observed if and only if $y_i \leq \tau_i$, where $\tau_i = \min(D - d_i, W_i)$. Because dates of sale for most products d_i are obtained only when a failure is reported, we do not know either τ_i or covariate values if $y_i > \tau_i$.

This problem corresponds to that described in Section 1.2 if we set $\mathbf{x}_i = (\tau_i, \mathbf{z}_i)$ and $A_i = \{y : y > \tau_i\}$. In many applications y_i and τ_i are conditionally independent, given $\mathbf{z}_i$, so that $f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = f(y_i|\mathbf{z}_i;\boldsymbol{\theta})$. We assume that a population of M observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, M$, is generated from the distribution

$$f(y|\mathbf{x};\, \boldsymbol{\theta})\, dG(\mathbf{x}) = f(y|\mathbf{z};\, \boldsymbol{\theta})\, dG(\tau, \mathbf{z}), \qquad y > 0, \quad \tau > 0, \quad \mathbf{z} \in \mathcal{Z}.$$

Under the observation scheme described, the likelihood function (7) for $\boldsymbol{\theta}$ is

$$L_1(\boldsymbol{\theta}) = \prod_{i \in O_1} f(y_i|\mathbf{z}_i;\, \boldsymbol{\theta})\, dG(\tau_i, \mathbf{z}_i) \left(\int \bar{F}(\tau|\mathbf{z};\, \boldsymbol{\theta})\, dG(\tau, \mathbf{z}) \right)^{M-m}, \tag{33}$$

where $\bar{F}(\tau|\mathbf{z};\boldsymbol{\theta}) = \mathrm{P}(Y > \tau|\mathbf{z};\boldsymbol{\theta})$.

Hu and Lawless (1996) give one method of estimation for this situation, which, in our terminology, is an application of the method based on $\mathbf{ES}(\boldsymbol{\theta})$ of (15). We consider the two pseudolikelihood methods discussed in Section 2.2 under followup supplementary sampling, and compare these with likelihood methods.

4.2. Simulations.

The simulations we performed are based on the Weibull proportional-hazards model, a widely used parametric lifetime regression model. We took censoring times and lifetimes to be independent and varied their distributions to give heavy to moderate censoring ($m/M = 0.05$ to 0.25). Sample size, parameter values and censoring were chosen to be realistic for warranty and field-failure applications like that discussed in Section 5 of Hu and Lawless (1996).

We chose $M = 4000$, and generated lifetimes y_i, $i = 1, \ldots, M$, from the p.d.f.

$$f(y|z_i;\, \beta_0, \beta_1, \delta) = \delta y^{\delta-1} e^{\beta_0+\beta_1 z_i} \exp(-y^{\delta} e^{\beta_0+\beta_1 z_i}), \qquad y > 0, \tag{34}$$

where the z_i's were independent and took on values 0 and 1 each with probability 0.5. A random sample u_i, $i = 1, \ldots, M$, from $N(\mathcal{T}^0, \sigma^2)$ with $\mathcal{T}^0 = 300.0$ and $\sigma = 80.0$ was then drawn, and censoring time τ_i obtained by $\tau_i = u_i \wedge \mathcal{T}^0$. The set $\{(y_i, \tau_i, z_i) : y_i \leq \tau_i\}$ is taken as the observed (primary) data; the size of this set is denoted by m. In addition, supplementary followup data were simulated as follows. We randomly selected items with probability p^f from the $M - m$ items with $y_i > \tau_i$. The (τ_i, z_i)'s for the selected items play the role of a followup sample. We considered p^f equal to 0.05, 0.10 and 0.20.

We present simulation results for three models, given by (34) with

$$(\beta_0, \beta_1, \delta) = (-8.6,\, 1.0,\, 1.0),\ (-15.8,\, 1.0,\, 2.5),\ (-17.6,\, 1.0,\, 2.5).$$

TABLE 1: Estimated standard deviations of the estimates in three cases.[a]

		Case I			Case II			Case III		
Method	p	$\hat\sigma_0$[b]	$\hat\sigma_1$	$\hat\sigma_2$	$\hat\sigma_0$	$\hat\sigma_1$	$\hat\sigma_2$	$\hat\sigma_0$	$\hat\sigma_1$	$\hat\sigma_2$
L_T		0.560	1.640	0.095	0.626	0.615	0.125	6.294	4.279	0.176
L_1		0.304	0.123	0.051	0.407	0.078	0.072	0.984	0.163	0.173
L_C		0.303	0.118	0.051	0.402	0.067	0.071	0.982	0.159	0.172
L^f	0.20	0.307	0.134	0.051	0.415	0.093	0.073	0.984	0.172	0.173
	0.10	0.311	0.153	0.051	0.426	0.118	0.075	0.989	0.188	0.173
	0.05	0.316	0.178	0.051	0.435	0.148	0.076	0.997	0.210	0.174
PS	0.20	0.305	0.134	0.051	0.422	0.094	0.074	0.986	0.173	0.173
	0.10	0.308	0.151	0.051	0.432	0.116	0.075	0.990	0.186	0.174
	0.05	0.315	0.188	0.052	0.489	0.168	0.085	1.023	0.222	0.180
ES	0.20	0.306	0.134	0.051	0.420	0.094	0.074	0.984	0.173	0.173
	0.10	0.309	0.154	0.051	0.441	0.121	0.077	0.992	0.189	0.174
	0.05	0.310	0.179	0.051	0.471	0.153	0.082	1.013	0.212	0.178

[a] Cases I, II, III: $(\beta_0, \beta_1, \delta) = (-8.6, 1.0, 1.0), (-15.8, 1.0, 2.5), (-17.6, 1.0, 2.5)$.
[b] $\widehat{\text{var}}$ = variance estimate: $\hat\sigma_0 = \sqrt{\widehat{\text{var}}(\hat\beta_0)}$; $\hat\sigma_1 = \sqrt{\widehat{\text{var}}(\hat\beta_1)}$; $\hat\sigma_2 = \sqrt{\widehat{\text{var}}(\hat\delta)}$.

These give approximate proportions $m/M = 0.10, 0.25, 0.05$, respectively. For each case, the two approaches based on the estimating functions (12) and (15) were used to estimate the parameters and to compute estimates of the asymptotic covariance matrices for $\sqrt{M}(\hat\beta_0 - \beta_0, \hat\beta_1 - \beta_1, \hat\delta - \delta)$, as discussed in Section 3.2. Newton's method was used to solve $\mathbf{PS}(\boldsymbol\theta) = \mathbf{0}$ and $\mathbf{ES}(\boldsymbol\theta) = \mathbf{0}$.

In our simulations we also investigated maximum-likelihood estimation based on the truncated data likelihood $L_T(\boldsymbol\theta)$ (11), the likelihood of the primary data $L_1(\boldsymbol\theta, G)$ (7) assuming $G(\cdot)$ known, the conditional likelihood of the primary data with a followup sample $L^f(\boldsymbol\theta)$ (19), and the censored data likelihood $L_C(\boldsymbol\theta)$, which the score function $\mathbf{CS}(\boldsymbol\theta)$ in (13) is based on. Estimation based on $L_1(\boldsymbol\theta, G)$ and $L_C(\boldsymbol\theta)$ cannot be used in the situations we consider in the previous sections, since they require knowledge of $G(\cdot)$ and the covariates $\mathbf{x}_i$ for all $i = 1, \ldots, M$, respectively. However, we included them in the simulation in order to investigate the extent to which the methods discussed in the paper recover this missing information. Table 1 shows variance estimates of the likelihood estimators, $\hat{\boldsymbol\theta}_{\mathbf{PS}}$, and $\hat{\boldsymbol\theta}_{\mathbf{ES}}$ based on a single simulated sample for each set of parameter values $(\beta_0, \beta_1, \delta)$. We can see that the likelihood $L_T(\boldsymbol\theta)$, based solely on the truncated data, is uninformative about the regression parameters compared to the other methods, which utilize supplementary information. This agrees with results of Kalbfleisch and Lawless (1988a) in a simpler setting, and indicates the utility of supplementary data. We also found in this example and in other simulations that the likelihood method based on $L^f(\boldsymbol\theta)$ behaved similarly to the pseudolikelihood estimating-function methods. Consequently, to conserve space we will present the simulation results only for the two likelihood methods based on $L_1(\boldsymbol\theta, G)$ and $L_C(\boldsymbol\theta)$, and the estimating-function methods based on (12) and (15).

Tables 2–5 show results based on 100 repetitions of the simulation for each set of parameter values. The sample means of the parameter estimates obtained by different methods are shown in Table 2. Table 3 gives the sample standard deviations of the estimates. We see that all of the estimators have small bias relative to the standard deviations. Table 3 reveals that the estimating-function methods are effective. Using a supplementary sample of only 5% ($p^f = 0.05$) gives very informative inferences, and once the supplementary-sample fraction rises to 20%, the estimates for β_0 and δ

TABLE 2: Sample means of the estimates in three cases.[a]

		Case I			Case II			Case III		
Method	p	β_0	β_1	δ	β_0	β_1	δ	β_0	β_1	δ
L_1		−8.637	0.995	1.006	−15.762	0.999	2.494	−17.608	1.003	2.535
L_C		−8.641	1.000	1.006	−15.769	0.997	2.496	−17.685	0.989	2.518
PS	0.20	−8.643	0.999	1.007	−15.762	0.989	2.495	−17.708	0.983	2.521
	0.05	−8.641	1.007	1.006	−15.789	0.988	2.501	−17.686	0.977	2.517
ES	0.20	−8.641	0.997	1.007	−15.768	0.990	2.497	−17.718	0.984	2.523
	0.05	−8.634	1.003	1.005	−15.794	0.991	2.501	−17.749	0.989	2.529

[a] Cases I, II, III: $(\beta_0, \beta_1, \delta) = (-8.6, 1.0, 1.0), (-15.8, 1.0, 2.5), (-17.6, 1.0, 2.5)$.

TABLE 3: Sample standard deviations of the estimates in three cases.[a]

		Case I			Case II			Case III		
Method	p	$\hat{\sigma}_0$[b]	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
L_1		0.336	0.113	0.058	0.370	0.076	0.066	0.879	0.153	0.154
L_C		0.334	0.107	0.058	0.363	0.063	0.065	0.875	0.141	0.146
PS	0.20	0.338	0.122	0.058	0.388	0.101	0.069	0.859	0.155	0.147
	0.05	0.345	0.176	0.058	0.484	0.160	0.087	0.908	0.195	0.155
ES	0.20	0.334	0.122	0.057	0.402	0.102	0.071	0.875	0.154	0.150
	0.05	0.342	0.180	0.057	0.492	0.159	0.087	0.963	0.202	0.165

[a] Cases I, II, III: $(\beta_0, \beta_1, \delta) = (-8.6, 1.0, 1.0), (-15.8, 1.0, 2.5), (-17.6, 1.0, 2.5)$.
[b] $\overline{\text{var}}$ = Sample variance from the 100 repetitions: $\hat{\sigma}_0 = \sqrt{\overline{\text{var}}(\hat{\beta}_0)}$, $\hat{\sigma}_1 = \sqrt{\overline{\text{var}}(\hat{\beta}_1)}$, $\hat{\sigma}_0 = \sqrt{\overline{\text{var}}(\hat{\delta})}$.

based on (12) and (15) are nearly as efficient as those based on $L_1(\boldsymbol{\theta}, G)$ and $L_C(\boldsymbol{\theta})$, which require full information about censoring times and covariates. The estimates of the regression coefficient β_1 have standard deviations about 25% higher than those from $L_C(\boldsymbol{\theta})$ when m/M is smaller (cases I and III) and about 80–90% higher when m/M is larger (case II). This is consistent with the simulation results presented in Kalbfleisch and Lawless (1988a). We also observe that no matter whether censoring is heavy (case III with $m/M = 0.05$) or moderate (case II with $m/M = 0.25$), the relative behaviour of the proposed methods is similar. Table 3 also shows that increasing the size of the supplementary sample (in particular, going from $p = 0.05$ to $p = 0.20$) makes a bigger difference for estimation of the regression coefficient β_1 than for the parameters β_0 and δ in (34).

Tables 4 and 5 investigate the adequacy of variance estimation (Section 3.2) and the asymptotic normal approximations, which one would use to obtain confidence intervals or tests for parameters. We give in Table 4 the sample variances of the parameter estimates, and the sample means of the variance estimates described in Section 3.2, from the 100 repetitions obtained for each of the cases on which Tables 2 and 3 are based. To conserve space we include $L_C(\boldsymbol{\theta})$ but exclude $L_1(\boldsymbol{\theta}, G)$, since they give similar results. Also, we give results only for $p = 0.05$; results for $p = 0.20$ are comparable. The sample variances are close to the corresponding sample means of the variance estimates. In Table 5, for the same cases we present the frequencies with which the standardized statistics $Z_0 = (\hat{\beta}_0 - \beta_0)/\hat{\sigma}_0$, $Z_1 = (\hat{\beta}_1 - \beta_1)/\hat{\sigma}_1$, $Z_2 = (\hat{\delta} - \delta)/\hat{\sigma}_2$ are less than or equal to the standard-normal-distribution 5, 25, 50, 75 and 95th percentiles. Here, $\hat{\sigma}_0$, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the estimated standard derivations for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\delta}$ described in Section 3.2. Given the small number of samples (100) for each case, the frequencies are reasonably close to

TABLE 4: Comparison of sample variances and variance estimates.[a]

Method	Quantity[b]	Case I β_0	Case I β_1	Case I δ	Case II β_0	Case II β_1	Case II δ	Case III β_0	Case III β_1	Case III δ
L_C	$\overline{\text{var}}$	0.1152	0.0115	0.0034	0.1320	0.0040	0.0042	0.7655	0.0200	0.0214
	$\overline{\widehat{\text{var}}}$	0.0979	0.0145	0.0028	0.1617	0.0046	0.0050	0.9174	0.0251	0.0283
PS($p = 0.05$)	$\overline{\text{var}}$	0.1189	0.0310	0.0034	0.2339	0.0256	0.0075	0.8236	0.0382	0.0240
	$\overline{\widehat{\text{var}}}$	0.1064	0.0342	0.0029	0.2149	0.0247	0.0065	0.9669	0.0451	0.0298
ES($p = 0.05$)	$\overline{\text{var}}$	0.1172	0.0325	0.0032	0.2423	0.0252	0.0076	0.9274	0.0408	0.0271
	$\overline{\widehat{\text{var}}}$	0.1066	0.0342	0.0029	0.2395	0.0254	0.0073	0.9854	0.0468	0.0304

[a] Cases I, II, III: $(\beta_0, \beta_1, \delta) = (-8.6, 1.0, 1.0), (-15.8, 1.0, 2.5), (-17.6, 1.0, 2.5)$.
[b] $\overline{\widehat{\text{var}}}$ = sample mean of the variance estimates from the 100 repetitions; $\overline{\text{var}}$ = sample variance from the 100 repetitions.

TABLE 5: Frequencies[a] of $Z_0, Z_1, Z_2 \leq q(5), q(25), q(50), q(75), q(95)$.

Method	Quantity[b]	Case I 5	25	50	75	95	Case II 5	25	50	75	95	Case III 5	25	50	75	95
L_C	Z_0	6	28	54	77	92	2	20	41	75	98	2	28	50	80	96
	Z_1	5	20	48	81	98	3	30	53	77	96	5	25	58	79	98
	Z_2	7	21	48	71	94	2	25	55	80	97	2	20	48	75	99
PS($p = 0.05$)	Z_0	4	27	58	77	93	4	26	48	70	94	3	26	53	78	96
	Z_1	4	20	50	74	94	6	24	53	82	95	7	26	57	79	97
	Z_2	7	24	49	73	92	8	26	52	73	96	2	20	43	76	98
ES($p = 0.05$)	Z_0	4	27	56	78	93	4	24	48	76	95	7	26	55	79	97
	Z_1	7	24	53	72	94	2	28	53	83	93	5	26	55	77	97
	Z_2	7	23	49	74	93	5	28	53	74	96	2	19	43	74	94

[a] $q(5), q(25), \ldots$ are quantiles of the standard normal distribution.
[b] $Z_0 = (\hat{\beta}_0 - \beta_0)/\hat{\sigma}_0$, $Z_1 = (\hat{\beta}_1 - \beta_1)/\hat{\sigma}_1$, $Z_2 = (\hat{\delta} - \delta)/\hat{\sigma}_2$.

the expected values 5, 25, 50, 75 and 95. In addition, Q-Q plots of the 100 standardized statistics for each case show the pivotals to be approximately normal. These results suggest that the asymptotic approximations are, for the cases considered, sufficiently accurate for practical purposes.

5. CONCLUDING REMARKS

In this paper we have noted the general structure of two approaches to pseudolikelihood estimation with missing data, and have studied their theoretical properties and finite-sample behaviour for a specific class of censored response-time problems. Our results indicate that for these problems the two types of methods are effective and comparable with respect to efficiency. They also perform well relative to "ordinary" maximum-likelihood estimation procedures. We observe that, in the situations with supplementary followup data, the methods based on $\mathbf{PS}(\boldsymbol{\theta})$ of (12) are somewhat more complicated to implement than those using $\mathbf{ES}(\boldsymbol{\theta})$ of (15), primarily because variance estimation (see Section 3.2) is more complicated. On the other hand, for such situations with no covariate and independent censoring, $\mathbf{ES}(\boldsymbol{\theta})$ requires observation of censoring times for units in O_1, but $\mathbf{PS}(\boldsymbol{\theta})$ does not. However, both methods are relatively straightforward.

As noted earlier, variation of these approaches have been applied previously in connection with missing covariate data; Carroll *et al.* (1995) contains a partial review. Extensive comparisons of methods have not been made except for some situations involving categorical responses. However, these results and theoretical results of Robins *et al.* (1994, 1995) suggest that the pseudolikelihood methods could be rather inefficient relative to full semiparametric maximum likelihood, particularly when the missing data mechanism is known and the missing and observed covariates are highly correlated. Robins *et al.* (1994, 1995) discuss improved estimating-function methods for such cases. In our simulations, τ_i was independent of (y_i, z_i) and the pseudolikelihood methods were quite efficient. Broader investigations would be useful.

APPENDIX. PROOFS OF ASYMPTOTIC RESULTS

The proofs below utilize approaches similar to those of Inagaki (1973), and Gong and Samaniego (1981).

Let

$$\Phi(y_i, \mathbf{x}_i; \boldsymbol\theta) = \mathrm{I}(y_i \notin A_i) \log f(y_i|\mathbf{x}_i; \boldsymbol\theta) + \mathrm{I}(y_i \in A_i) \log P(Y_i \in A_i; \boldsymbol\theta),$$

$$\Psi(y_i, \mathbf{x}_i; \boldsymbol\theta) = \mathrm{I}(y_i \notin A_i) \log f(y_i|\mathbf{x}_i; \boldsymbol\theta) + \mathrm{I}(y_i \in A_i) \log P(Y_i \in A_i|\mathbf{x}_i; \boldsymbol\theta).$$

We assume the following regularity conditions hold, where partial derivatives are denoted by using subscript notation, for example, $\Phi_{\boldsymbol\theta\boldsymbol\theta} = \partial^2\Phi/\partial\boldsymbol\theta^2$; also let $\mathrm{P}(\mathbf{x};\boldsymbol\theta) = \mathrm{P}(Y_i \in A_i|\mathbf{x}_i = \mathbf{x};\boldsymbol\theta)$ and $\mathrm{P}(\boldsymbol\theta) = \mathrm{P}(Y_i \in A_i;\boldsymbol\theta) = \int \mathrm{P}(\mathbf{x};\boldsymbol\theta)\, dG(\mathbf{x})$:

(A1) For all $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$, the following partial derivatives exist: $\Phi_{\boldsymbol\theta}$, $\Phi_{\boldsymbol\theta\boldsymbol\theta}$, $\Phi_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}$, $\Psi_{\boldsymbol\theta}$, $\Psi_{\boldsymbol\theta\boldsymbol\theta}$, $\Psi_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}$, $\forall\boldsymbol\theta \in \mathcal{A}$.

(A2) Interchanges of differentiation and integration of f are valid for first, second and third derivatives with respect to $\boldsymbol\theta$.

(A3) $\boldsymbol\Sigma_{\mathrm{PS}}(\boldsymbol\theta_0) = \mathcal{E}_{\boldsymbol\theta_0,G}\, \Phi_{\boldsymbol\theta}\Phi_{\boldsymbol\theta}^{\mathrm{T}}$ and $\boldsymbol\Sigma_{\mathrm{ES}}(\boldsymbol\theta_0) = \mathcal{E}_{\boldsymbol\theta_0}\Psi_{\boldsymbol\theta}\Psi_{\boldsymbol\theta}^{\mathrm{T}}$ exist and are positive definitive, $\forall\boldsymbol\theta \in \mathcal{A}$.

(A4) Each element of $\Phi_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}$ and $\Psi_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}$ is bounded by an integrable function for all $\boldsymbol\theta \in \mathcal{A}$. We symbolize this as

$$|\Phi_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}| \le \mathbf{H}(y, \mathbf{x}), \quad |\Psi_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}| \le \mathbf{H}(y, \mathbf{x}) \qquad \forall\boldsymbol\theta \in \mathcal{A},$$

where $\mathcal{E}\{\mathbf{H}(Y, \mathbf{X})\}$ exists.

(A5) $\mathrm{P}(\mathbf{x};\boldsymbol\theta)$, $\mathrm{P}(\mathbf{x};\boldsymbol\theta)_{\boldsymbol\theta}$, $\mathrm{P}(\mathbf{x};\boldsymbol\theta)_{\boldsymbol\theta\boldsymbol\theta}$, $\mathrm{P}(\mathbf{x};\boldsymbol\theta)_{\boldsymbol\theta\boldsymbol\theta\boldsymbol\theta}$ are bounded measurable functions with respect to $\mathbf{x}$, and uniformly continuous functions with respect to $\boldsymbol\theta \in \mathcal{A}$.

Proof of Theorem 1. Let $\boldsymbol\theta \in \mathcal{A}$ be fixed. We have by Taylor expansion of $PS(\boldsymbol\theta)$ of (12) about the point $\boldsymbol\theta_0$:

$$\mathbf{PS}(\boldsymbol\theta) = \mathbf{PS}(\boldsymbol\theta_0) + \left.\frac{\partial\mathbf{PS}(\boldsymbol\lambda)}{\partial\boldsymbol\lambda}\right|_{\boldsymbol\lambda=\boldsymbol\theta_0}(\boldsymbol\theta - \boldsymbol\theta_0) + \tfrac{1}{2}(\boldsymbol\theta - \boldsymbol\theta_0)^{\mathrm{T}}\left.\frac{\partial^2\mathbf{PS}(\boldsymbol\lambda)}{\partial\boldsymbol\lambda^2}\right|_{\boldsymbol\lambda=\boldsymbol\xi}(\boldsymbol\theta - \boldsymbol\theta_0), \quad (35)$$

with $\boldsymbol\xi \in \mathcal{A}$ and $\|\boldsymbol\xi - \boldsymbol\theta_0\| \le \|\boldsymbol\theta - \boldsymbol\theta_0\|$. Notice that

$$\boldsymbol\Delta_{\mathrm{PS}}(\boldsymbol\theta) = \frac{1}{M}\{\mathbf{PS}(\boldsymbol\theta) - \mathbf{S}(\boldsymbol\theta; G)\} = \frac{M-m}{M}\left(\frac{\partial \log \tilde{P}(\boldsymbol\lambda)}{\partial\boldsymbol\lambda} - \frac{\partial \log P(\boldsymbol\lambda)}{\partial\boldsymbol\lambda}\right)\Bigg|_{\boldsymbol\lambda=\boldsymbol\theta}.$$

Because of (A5) and (21), each element of $\partial^l\boldsymbol\Delta_{\mathrm{PS}}(\boldsymbol\theta_0)/\partial\boldsymbol\theta^l$, $l = 0, 1, 2$, converges almost surely to zero when $M \longrightarrow \infty$, since $(M - m)/M \le 1$. By recalling the properties of the score function $\mathbf{S}(\boldsymbol\theta; G)$, and noting (A4) and (A5), we know that, as $M \longrightarrow \infty$,

$$\frac{1}{M}\mathbf{PS}(\boldsymbol\theta_0) \xrightarrow{\text{a.s.}} \mathbf{0}, \qquad \frac{1}{M}\left.\frac{\partial\mathbf{PS}(\boldsymbol\lambda)}{\partial\boldsymbol\lambda}\right|_{\boldsymbol\lambda=\boldsymbol\theta_0} \xrightarrow{\text{a.s.}} -\boldsymbol\Sigma_{\mathrm{PS}}(\boldsymbol\theta_0),$$

and

$$\frac{1}{M}\left|\frac{\partial^2 \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\xi}}\right| \le \mathcal{E}\{\mathbf{H}(Y,\mathbf{X})\} + \underline{I}, \qquad \text{a.s.,} \tag{36}$$

when $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$ is small enough, where $\underline{I}$ is a matrix of ones with the same dimension as $\mathbf{H}(y,\mathbf{x})$. Now let $\epsilon > 0$ be given, such that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \epsilon\underline{1}$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_0 + \epsilon\underline{1}$ lie in $\mathcal{A}$, (36) holds, and $\epsilon < \|\underline{1}^{\mathrm{T}}\mathcal{E}\mathbf{H}(Y,\mathbf{X}) + \underline{I}\}\underline{1}\|/2\|\boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\underline{1}\|$, where $\underline{1}$ is the p-dimensional vector with every component being 1. Then, by (35),

$$\begin{aligned}\left\|\frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_1) - \epsilon\boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\underline{1}\right\| &\le \left\|\frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_0)\right\| + \epsilon\left\|\left(\frac{1}{M}\frac{\partial \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\theta}_0} + \boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\right)\underline{1}\right\| \\ &\quad + \frac{\epsilon^2}{2}\left\|\underline{1}^{\mathrm{T}}\left(\frac{1}{M}\frac{\partial^2 \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\xi}_1}\right)\underline{1}\right\| \\ &\le \frac{3\epsilon}{4}\|\boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\underline{1}\|, \qquad \text{a.s.,}\end{aligned}$$

and

$$\begin{aligned}\left\|\frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_2) + \epsilon\boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\underline{1}\right\| &\le \left\|\frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_0)\right\| + \epsilon\left\|\left(\frac{1}{M}\frac{\partial \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\theta}_0} + \boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\right)\underline{1}\right\| \\ &\quad + \frac{\epsilon^2}{2}\left\|\underline{1}^{\mathrm{T}}\left(\frac{1}{M}\frac{\partial^2 \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\xi}_2}\right)\underline{1}\right\| \\ &\le \frac{3\epsilon}{4}\|\boldsymbol{\Sigma}_{\mathbf{PS}}(\boldsymbol{\theta}_0)\underline{1}\|, \qquad \text{a.s.,}\end{aligned}$$

if M is sufficiently large. For such M, the interval

$$\left[\min\left\{\frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_2), \frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_1)\right\}, \max\left\{\frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_2), \frac{1}{M}\mathbf{PS}(\boldsymbol{\theta}_1)\right\}\right]$$

contains the point 0 almost surely, and hence, by the continuity of $\mathbf{PS}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, the interval $[\boldsymbol{\theta}_0 - \epsilon\underline{1}, \boldsymbol{\theta}_0 + \epsilon\underline{1}]$ contains a solution of $\mathbf{PS}(\boldsymbol{\theta}) = \mathbf{0}$ almost surely. We have thus proved the consistency of $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$. □

Proof of Theorem 2. Let $\boldsymbol{\theta}$ in (35) be $\tilde{\boldsymbol{\theta}}_{\mathbf{PS}}$. Thus, we have

$$\mathbf{PS}(\boldsymbol{\theta}_0) = -\left(\frac{\partial \mathbf{PS}(\lambda)}{\partial \boldsymbol{\lambda}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\theta}_0} + \frac{1}{2}(\tilde{\boldsymbol{\theta}}_{\mathbf{PS}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\frac{\partial^2 \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\xi}}\right)(\tilde{\boldsymbol{\theta}}_{\mathbf{PS}} - \boldsymbol{\theta}_0).$$

Notice that

$$\frac{1}{\sqrt{M}}\mathbf{PS}(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{M}}\mathbf{S}(\boldsymbol{\theta}; G) + \frac{M-m}{M}\sqrt{M}\left(\frac{\partial \log \tilde{P}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} - \frac{\partial \log P(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right),$$

and

$$\begin{aligned}\sqrt{M}\left(\frac{\partial \log \tilde{P}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \log P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) &= \frac{1}{\tilde{P}(\boldsymbol{\theta})}\sqrt{M}\left(\frac{\partial \tilde{P}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) \\ &\quad - \frac{1}{\tilde{P}(\boldsymbol{\theta})}\frac{\partial \log P(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\sqrt{M}\{\tilde{P}(\boldsymbol{\theta}) - P(\boldsymbol{\theta})\}.\end{aligned}$$

By the strong law of large numbers, the condition (21), and the asymptotic normality of the score function $\mathbf{S}(\boldsymbol{\theta}_0; G)$, we have

$$\frac{1}{\sqrt{M}}\mathbf{PS}(\boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{PS}}), \qquad \frac{M-m}{M} \xrightarrow{\text{a.s.}} P(\boldsymbol{\theta}_0) \xrightarrow{\text{a.s.}} \tilde{P}(\boldsymbol{\theta}_0)$$

as $M \longrightarrow \infty$. Furthermore, because of the condition (22), we know that $\mathbf{PS}(\boldsymbol{\theta}_0)/\sqrt{M}$ is asymptotically normal as $M \longrightarrow \infty$. On the other hand, applying the results of Theorem 1 and its proof, we have, as $M \longrightarrow \infty$,

$$\frac{1}{M}\left.\frac{\partial \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\theta}_0} + \frac{1}{2}(\tilde{\boldsymbol{\theta}}_{\mathrm{PS}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \left.\frac{\partial^2 \mathbf{PS}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\right|_{\boldsymbol{\lambda}=\boldsymbol{\xi}} \xrightarrow{\text{a.s.}} -\boldsymbol{\Sigma}_{\mathrm{PS}},$$

since $\|\boldsymbol{\xi} - \boldsymbol{\theta}_0\| \leq \|\tilde{\boldsymbol{\theta}}_{\mathrm{PS}} - \boldsymbol{\theta}_0\| \longrightarrow 0$ almost surely. The theorem now follows in a straightforward way. □

Proofs of Theorems 3 and 4 may be given similarly; Appendix 1 of Hu and Lawless (1996) outlines the derivation.

ACKNOWLEDGEMENT

We thank the referees for helpful suggestions.

REFERENCES

Breslow, N.E., and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11–20.

Carroll, R.J., and Wand, M.P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B*, 53, 573–585.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

Gong, G., and Samaniego, F.J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Ann. Statist.*, 9, 861–869.

Heitjan, D.F., and Rubin, D.B. (1991). Ignorability and coarse data. *Ann. Statist.*, 19, 2244–2253.

Hsieh, D.A., Manski, C.F., and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Amer. Statist. Assoc.*, 80, 651–662.

Hu, X.J. (1995). Estimation from truncated data with supplementary information, with application to field reliability. Ph.D. Thesis, University of Waterloo, Canada.

Hu, X.J., and Lawless, J.F. (1996). Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, 83, 747–761.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Ann Inst. Statist. Math.*, 25, 1–26.

Kalbfleisch, J.D., and Lawless, J.F. (1988a). Estimation of reliability from field performance studies (with discussion). *Technometrics*, 30, 365–388.

Kalbfleisch, J.D., and Lawless, J.F. (1988b). Likelihood analysis of multistate models for disease incidence and mortality. *Statist. Med.*, 7, 149–160.

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

Pepe, M.S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, 79, 355–365.

Pepe, M.S., and Fleming, T.R. (1991). A nonparametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.*, 86, 108–113.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73, 1–11.

Reilly, M., and Pepe, M.S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82, 299–314.

Robins, J.M., Rotnitzky, A., Zhao, L.-P., and Lipsitz, S. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89, 846–866.

Robins, J.M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. Roy. Statist. Soc. Ser. B*, 57, 409–424.

Rubin, D.B., (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57–71.

Suzuki, K. (1985a). Estimation method of lifetime based on the record of failures during the warranty period. *J. Amer. Statist. Assoc.*, 80, 68–72.

Suzuki, K. (1985b). Nonparametric estimation of lifetime distribution from incomplete field data. *Technometrics*, 27, 263–271.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, 82, 528–540.

Wild, C.J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, 78, 705–717.

Received 2 October 1996
Revised 6 January 1997
Accepted 6 January 1997

Department of Biostatistics
Harvard School of Public Health
Boston, Massachusetts
U.S.A. 02115

Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1