Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

LIKELIHOOD AND PSEUDO LIKELIHOOD ESTIMATION BASED ON RESPONSE-BIASED OBSERVATION

J.F. Lawless
University of Waterloo

ABSTRACT

Response-biased observation refers to situations where the probability a unit is observed depends on the value of a response associated with that unit. We discuss the construction of estimating equations for parametric regression models through likelihood and pseudo likelihoods, for situations in which responses are stratified and sampling is stratum-specific. Properties of the resulting estimators are reviewed and an illustration involving field reliability data is presented.

1 Introduction

In many observational studies the probability that a specific individual or unit is observed or selected in a sample depends upon responses or covariates associated with that unit. That is, if units in some population have associated response variables y and covariates x, then the probability unit i is selected depends upon the values (y_i, x_i) for that unit. When the probability of selection depends upon y_i, we call the observation scheme response-selective, or response-biased.

For simplicity of exposition I will focus mainly on situations where the probability of selection depends solely on y_i. However, as described at the end of Section 2 and in Section 4, situations where the probability of selection depends on both y_i and x_i may also be handled using the methods considered here.

Examples of response-selective observation are abundant. In socio-economic studies based on samples drawn from administrative records, selection is often response-related (e.g. Hausman and Wise 1981). Similarly, in a study of factors affecting low birth weight of humans, one might select newborns over a period of time and measure covariates so as to over-sample babies with

low birth weights. Extreme forms of response-selection are embodied in case-control or choice-based sampling used in epidemiology and economics (e.g. Breslow and Cain 1988; Hsieh, Manski and McFadden 1985); in this case the response variable is categorical and covariates are observed for samples of individuals selected from each response category. Examples of selection bias in more complicated settings are given by Hoem (1985) and Kalbfleisch and Lawless (1988a), who discuss the observation of life history events for human populations.

When the probability of selection $p(y)$ for a unit is a known function of y, methods that weight log likelihood or estimating function components according to $p(y)^{-1}$ are often used. Several other approaches can also be used in various contexts. This article discusses four methods of estimation for a rather broad class of situations. The approaches described are not new, but questions remain about their properties. Our purpose is to review the four methods and recent investigations into their properties, and to indicate connections with other areas.

It is assumed that (y, x) values for individuals or units in some "population" from which units will be sampled are generated from a probability distribution with density or mass function

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

and that our objective is to estimate the p-dimensional parameter θ. We wish to avoid strong parametric assumptions about $g(x)$ and the corresponding distribution function $G(x)$, as is common in regression modelling. Sampling is response-selective in the following sense: the range of y is partitioned into strata $S_1, ..., S_k$ and if $y_i \in S_j$ then unit i is sampled (selected) with probability p_j. In other words, $p(y_i) = \sum_{j=1}^{k} p_j I(y_i \in S_j)$, where $I(A)$ is the indicator function which equals 1 if event A is true and 0 otherwise.

More specifically, we assume that a finite population of N units has values (y_i, x_i), $i = 1, ..., N$ generated as independent realizations from (1). In survey sampling terminology, we have a stratified population and wish to estimate parameters in the superpopulation model (1). Samples may be selected in various ways. We consider two, termed basic stratified sampling (BSS) and basic variable probability sampling (BVPS). In BSS a simple random sample of specified size n_j is selected from units in the j'th stratum (i.e. with $y_i \in S_j$). In BVPS each unit in the population is considered for selection independent of every other unit. The j'th stratum size (number of units with $y_i \in S_j$) is denoted by N_j $(j = 1, ..., k)$. In the case of BSS the fixed sample size from stratum j is $n_j = N_j p_j$, whereas in the case of BVPS the size of the sample from the j'th stratum is random. It is an important feature of our framework that the stratum sizes $N_1, ..., N_k$ are known, or observable. This latter feature is not present in some applications, e.g. for

many instances of case-control or choice-based sampling.

The remainder of the paper is as follows. Section 2 makes the observational framework precise and then describes methods of estimation based both on likelihood and on pseudo likelihood functions. Section 3 discusses asymptotic properties and variance estimation, and Section 4 illustrates the methodology. Section 5 concludes with remarks on extensions and relationships with missing data methods.

2 Likelihood and Pseudo Likelihood Estimating Functions

Suppose that individual pairs (y_i, x_i), $i = 1, ..., N$ are generated independently from (1), and that N_j units have $y_i \in S_j$ $(j = 1, ..., k)$. Units are selected by either BSS or BVPS as described in Section 1 and (y_i, x_i) observed. Let $R_i = I(\text{unit } i \text{ is selected})$, and $D_j = \{i : \ R_i = 1, y_i \in S_j\}$ denote the units selected from stratum j, where $|D_j| = n_j$. As is customary, Y and X are used to represent the random variables of which y_i and x_i are realizations.

For simplicity I consider only situations where the population size N and stratum sizes N_j are fixed at the time units are selected. It is also assumed that the N_j's are known and further, that for units not selected all that is known is which stratum they are in. In some contexts such as the birth weight study, the values of N and the N_j's may be unknown until the end of the sampling period and the y_i's (but not the x_i's) may be known for units not selected. Such features may also be dealt with via the methods discussed here; Lawless, Wild and Kalbfleisch (1997) consider a variety of response-selective sampling schemes.

Under BSS or BVPS in the framework described, the data include the stratum sizes $N_1, ..., N_k$, the pairs (y_i, x_i), $i \in D_j$ $(j = 1, ..., k)$ for the selected units, and for BVPS, the sample sizes $n_1, ..., n_k$. In either case the probability density function for the observed data defines a likelihood function for the unknown parameters θ and G (the distribution function of X) which is proportional to

$$L_F(\theta, G) = \prod_{j=1}^{k} \{Q_j(\theta, G)^{N_j - n_j} \prod_{i \in D_j} f(y_i | x_i; \theta) dG(x_i)\}, \tag{2}$$

where

$$\begin{aligned} Q_j(\theta, G) &= Pr(Y \in S_j; \theta, G) \\ &= \int Pr(Y \in S_j | x; \theta) dG(x) \end{aligned} \tag{3}$$

It is recognized in the notation that (2) depends on both θ and G. From our point of view G is a nuisance parameter but because of (3) it is necessary to estimate it in order to estimate θ by maximizing (2). One approach is to maximize the semiparametric likelihood (2) jointly with respect to θ and G. This is feasible when Y is categorical (Wild 1991, Scott and Wild 1997) or when G is discrete with relatively few points of support (Hsieh et al. 1985), and recent work suggests it is feasible quite generally.

A second line of attack is to maximize $L_F(\theta, \tilde{G})$, where $\tilde{G}$ is a simple nonparametric estimate of G; this is an extension of the parametric pseudo likelihood idea of Gong and Samaniego (1981) to a semiparametric setting. Noting that

$$G(x) = \sum_{j=1}^{k} Pr(X_i \leq x | Y_i \in S_j) Pr(Y_i \in S_j),$$

we propose to use the estimate

$$\tilde{G}(x) = \sum_{j=1}^{k} \tilde{G}_j(x)(\frac{N_j}{N}), \tag{4}$$

where $\tilde{G}_j(x)$ is the empirical cumulative distribution function (cdf) based on the x_i's for units $i \in D_j$ sampled from the j'th stratum. Inserting (4) into (3) and taking $\partial \log L(\theta, \tilde{G})/\partial\theta$, we obtain the pseudo score function

$$\begin{aligned} S_P(\theta) &= \sum_{j=1}^{k} \sum_{i \in D_j} \frac{\partial \log f(y_i|x_i;\theta)}{\partial\theta} \\ &\quad + \sum_{j=1}^{k} (N_j - n_j) \left\{ \frac{\sum_{\ell=1}^{k} p_\ell^{-1} \sum_{i \in D_\ell} \partial Q_j^*(x_i;\theta)/\partial\theta}{\sum_{\ell=1}^{k} p_\ell^{-1} \sum_{i \in D_\ell} Q_j^*(x_i;\theta)} \right\}, \end{aligned} \tag{5}$$

where $p_\ell = n_\ell/N_\ell$ and

$$Q_j^*(x;\theta) = Pr(Y \in S_j | x;\theta). \tag{6}$$

We estimate θ by solving the equation $S_P(\theta) = 0$. This idea has been used in other contexts by Pepe and Fleming (1991), and Hu and Lawless (1997).

A third possibility is to weight score contributions for sampled units to give the weighted pseudo score

$$S_W(\theta) = \sum_{j=1}^{k} p_j^{-1} \sum_{i \in D_j} \frac{\partial \log f(y_i|x_i;\theta)}{\partial\theta}. \tag{7}$$

It is apparent that $S_W(\theta)$ is unbiased (has mean 0) under BSS or BVPS, where expectation is with respect to (1) and the selection scheme. Weighting is common in survey sampling (eg. Holt, Smith and Winter 1980, Binder 1983, Binder and Patak 1994) and has been considered for maximum likelihood methods by Hsieh et al. (1985), Scott and Wild (1986), Kalbfleisch and Lawless (1988ab) and others. Robins et al. (1994, 1995) consider weighted pseudo score functions when the selection probabilities $p(y_i)$ may depend on unknown parameters.

The final method considered is based on the observation that in the case of BVPS the distribution of the observed responses, conditional on the values of R_i and x_i $(i = 1, ..., N)$, yields the conditional, or selection-biased likelihood

$$\begin{aligned} L_C(\theta) &= \prod_{i=1}^{N} \left\{ \frac{f(y_i|x_i;\theta)p(y_i)}{Pr(R_i = 1;\theta)} \right\}^{R_i} \\ &= \prod_{j=1}^{k} \prod_{i \in D_j} \left\{ \frac{p_j f(y_i|x_i;\theta)}{\sum_{\ell=1}^{k} p_\ell Q_\ell^*(x_i;\theta)} \right\}. \end{aligned} \tag{8}$$

The corresponding score function is

$$S_C(\theta) = \sum_{j=1}^{k} \sum_{i \in D_j} \left\{ \frac{\partial \log f(y_i|x_i;\theta)}{\partial \theta} - \frac{\sum_{\ell=1}^{k} p_\ell \partial Q_\ell^*(x_i;\theta)/\partial\theta}{\sum_{\ell=1}^{k} p_\ell Q_\ell^*(x_i;\theta)} \right\}. \tag{9}$$

Straightforward calculation shows that $S_C(\theta)$ is unbiased under BSS as well as under BVPS.

If $p_1 = \cdots = p_k$ then $S_W(\theta) = 0$ and $S_C(\theta) = 0$ yield the same estimate, but otherwise the estimators obtained from the estimating equations $S_P(\theta) = 0$, $S_W(\theta) = 0$ and $S_C(\theta) = 0$ appear to be distinct. General results concerning the relative efficiencies of the estimators $\hat{\theta}_P$, $\hat{\theta}_W$ and $\hat{\theta}_C$ obtained from these equations, and $\hat{\theta}_F$ obtained by maximization of $L_F(\theta, G)$ with respect to θ and G, are not available, though results of Robins et al. (1994, 1995) show that the pseudo-likelihood methods are asymptotically inefficient. Another important point is that in the case of BVPS it is preferable to use $p_j = n_j/N_j$ rather than prior selection probabilities. In Section 3 we discuss asymptotic properties of the estimates and in Sections 4 and 5 some limited simulation results.

We conclude this section by noting that analogous estimating functions may be given for cases in which the probability of selection for a unit depends

on both y and x. In particular, suppose that k strata are defined according to whether $(y_i, x_i) \in S_j$, where $S_1, ..., S_k$ partitions $\mathcal{Y} \times \mathcal{X}$. The likelihood function based on the observed data is now given by (2), with $Q_j(\theta, G)$ redefined as

$$\begin{aligned} Q_j(\theta, G) &= Pr\{(Y, X) \in S_j; \theta, G\} \\ &= \int Pr\{(Y, x) \in S_j | x; \theta\} dG(x). \end{aligned} \tag{10}$$

A pseudo score $S_P(\theta)$ corresponding to (5) is obtained by estimating (10) via (4), where $\tilde{G}(x)$ is as before the empirical cdf based on the x_i's for units $i \in D_j$. A weighted pseudo score function is given by (7) once again, and a conditional score is given by (9) with $Q^*_\ell(x_i; \theta)$ replaced by

$$Q^*_\ell(x; \theta) = Pr\{(Y, x) \in S_l | x; \theta\}. \tag{11}$$

3 Asymptotic Properties, Variance Estimates and Confidence Limits

By taking limits as $N \to \infty$ and with $p_j = n_j/N_j$ $(j = 1, ..., k)$ fixed positive values, we may show that under mild conditions the estimators $\hat{\theta}$ obtained by solving $S_W(\theta) = 0$ or $S_C(\theta) = 0$ are consistent and asymptotically normal. Special cases have been considered by Kalbfleisch and Lawless (1988b), Wild (1991) and Scott and Wild (1997). Asymptotics under BVPS may also be obtained.

A rigorous development of asymptotics for the case of full maximum likelihood based on (2) or for the estimating equation $S_P(\theta) = 0$ is more difficult. For the former Wild (1991) and Scott and Wild (1997) deal with the case of categorical responses, and for the latter Hu and Lawless (1997) deal with the special problem described in Section 4.

We outline the asymptotic normal results for $S_W(\theta)$ and $S_C(\theta)$ given by (7) and (9), respectively; Lawless et al. (1997) give a fuller treatment. Both (7) and (9) may be written in the form

$$S(\theta) = \sum_{i=1}^{N} R_i U(Y_i; X_i; \theta). \tag{12}$$

For $S_W(\theta)$, for example,

$$U(Y_i; X_i; \theta) = \sum_{j=1}^{k} p_j^{-1} I(Y_i \in S_j) \partial \log f(Y_i | X_i; \theta) / \partial \theta. \tag{13}$$

Assume $p\lim(N_j/N) = \pi_j > 0$ as $N \to \infty$ and define

$$A_N(\theta) = \tfrac{1}{N}(-\frac{\partial S}{\partial \theta'}) \quad B_N(\theta) = \mathrm{Var}\{N^{-1/2}S(\theta)\}$$

$$A(\theta) = p\lim A_N(\theta) \quad B(\theta) = \lim B_N(\theta).$$

Under mild regularity conditions on the model (1), we have

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1}\right),$$

where θ_0 represents the true value of θ.

The asymptotic variance of $\hat{\theta}$ may be estimated as $\hat{A}^{-1}\hat{B}\hat{A}^{-1}$, where $\hat{A}$ and $\hat{B}$ are consistent estimates of $A(\theta_0)$ and $B(\theta_0)$. The matrix $\hat{A} = A_N(\hat{\theta})$ may be used to estimate $A(\theta_0)$. For $S_C(\theta)$ in the case of BVPS the estimating function is a likelihood score function, so $A(\theta_0) = B(\theta_0)$. For the other cases we require a consistent estimator $\hat{B}$, which is not hard to obtain. For $S_W(\theta)$, for example, extending the approach of Kalbfleisch and Lawless (1988b) and defining $v_i(\theta) = \partial \log f(y_i|x_i;\theta)/\partial\theta$ and $\bar{\bar{v}}^{(j)}(\theta) = \sum_{i:y_i\in S_j} v_i(\theta)/N_j$, we get

$$\begin{aligned}\mathrm{Var}\{S_W(\theta)\} &= \mathrm{Var}_{Y|X}E_{R|Y,X}\{S_W(\theta)\} + E_{Y|X}\mathrm{Var}_{R|Y,X}\{S_W(\theta)\} \\ &= E(\frac{-\partial S_W}{\partial\theta}) + C(\theta), \end{aligned} \tag{14}$$

where for BSS we have

$$C(\theta) = E_{Y|X}\sum_{j=1}^{k}\frac{n_j(1-p_j)}{p_j^2(N_j-1)}\sum_{i:y_i\in S_j}[v_i(\theta) - \bar{\bar{v}}^{(j)}(\theta)][v_i(\theta) - \bar{\bar{v}}^{(j)}(\theta)]'. \tag{15}$$

Since $N^{-1}E(-\partial S_W/\partial\theta) = A(\theta)$, equations (14) - (15) indicate that $B(\theta)$ may be estimated by $\hat{B} = \hat{A} + \hat{C}$, where

$$\hat{C} = \frac{1}{N}\sum_{j=1}^{k}\frac{(1-p_j)}{p_j^2}(\frac{n_j}{n_j-1})\sum_{i\in D_j}[v_i(\hat{\theta}) - \bar{v}^{(j)}(\hat{\theta})][v_i(\hat{\theta}) - \bar{v}^{(j)}(\hat{\theta})]', \tag{16}$$

where $\bar{v}^{(j)}(\theta) = \sum_{i\in D_j} v_i(\theta)/n_j$.

Confidence intervals for parameters may be obtained by treating $\sqrt{N}(\hat{\theta} - \theta)$ as approximately normal with a suitably estimated covariance matrix. An alternative would be to use some form of bootstrap. Investigation of specific problems by simulation is needed to gain insight into the adequacy of confidence interval procedures for different sample and population sizes.

As noted in Section 5, relatively little is known about the efficiency of $S_W(\theta)$, $S_C(\theta)$ and $S_P(\theta)$ in general situations; this too deserves investigation.

4 An Example

Kalbfleisch and Lawless (1988b), Hu and Lawless (1996) and others have considered problems in epidemiology and reliability in which a response time $y_i \geq 0$ and covariates z_i for an individual or unit in some population are always observed if y_i does not exceed an associated censoring time $\tau_i \geq 0$. The number of units for which the response time is censored (i.e. $y_i > \tau_i$) is known but the values of τ_i and z_i are not, so a fraction p_2 of the censored units are sampled, and their τ_i and z_i values are obtained. The objective is to estimate the distribution $f(y_i|z_i;\theta)$, where it is assumed that Y_i and τ_i are independent, given z_i.

This problem involves response-selective sampling of the type described at the end of Section 2. In particular, let $x_i = (\tau_i, z_i)$ be an extended covariate vector representing the censoring time τ_i and covariates z_i for unit i. The data for N units $i = 1, ..., N$ are assumed to come from (1), where $f(y_i|x_i;\theta) = f(y_i|z_i;\theta)$. Consider two strata for (y, τ, z) defined by $S_1 = \{(y, \tau, z) : y \leq \tau\}$ and $S_2 = \{(y, \tau, z) : y > \tau\}$. Units with $(y_i, x_i) \in S_1$ are selected with probability $p_1 = 1$ and those with $(y_i, x_i) \in S_2$ are selected with probability $p_2 \leq 1$.

We extend the four estimation procedures of Section 2 slightly to deal with the stratification on both y and x, as described at the end of Section 2, and to reflect the fact that for units $i \in D_2$ we know only that $y_i > \tau_i$, and not y_i's exact value. We obtain the likelihood function corresponding to (2) as

$$L_F(\theta, G) = \prod_{i \in D_1} f(y_i|x_i;\theta) dG(x_i) \prod_{i \in D_2} \bar{F}(\tau_i|x_i;\theta) dG(x_i) Q_2(\theta, G)^{N-n_1-n_2}, \tag{17}$$

where $\bar{F}(\tau|x;\theta) = \int_\tau^\infty f(y|x;\theta) dy$ and

$$Q_2(\theta, G) = \int \bar{F}(\tau|x;\theta) dG(x).$$

The pseudo score $S_P(\theta)$ corresponding to (5) is then

$$\begin{aligned} S_P(\theta) &= \sum_{i \in D_1} \frac{\partial \log f(y_i|x_i;\theta)}{\partial \theta} + \sum_{i \in D_2} \frac{\partial \log \bar{F}(\tau_i|x_i;\theta)}{\partial \theta} \\ &+ (N - n_1 - n_2) \left\{ \frac{\sum_{i \in D_1} \frac{\partial \bar{F}(\tau_i|x_i;\theta)}{\partial \theta} + \frac{1}{p_2} \sum_{i \in D_2} \frac{\partial \bar{F}(\tau_i|x_i;\theta)}{\partial \theta}}{\sum_{i \in D_1} \bar{F}(\tau_i|x_i;\theta) + \frac{1}{p_2} \sum_{i \in D_2} \bar{F}(\tau_i|x_i;\theta)} \right\} \end{aligned}$$

The weighted pseudo score corresponding to (7) is

$$S_W(\theta) = \sum_{i \in D_1} \frac{\partial \log f(y_i|x_i;\theta)}{\partial \theta} + \frac{1}{p_2} \sum_{i \in D_2} \frac{\partial \log \bar{F}(\tau_i|x_i;\theta)}{\partial \theta}.$$

Finally, the conditional (pseudo) score corresponding to (9) is

$$S_C(\theta) = \sum_{i \in D_1} \frac{\partial \log f(y_i|x_i;\theta)}{\partial \theta} + \sum_{i \in D_2} \frac{\partial \log \bar{F}(\tau_i|x_i;\theta)}{\partial \theta}$$

$$- \sum_{i \in D_1 \cup D_2} \left\{ \frac{\partial F(\tau_i|x_i;\theta)/\partial \theta + p_2 \partial \bar{F}(\tau_i|x_i;\theta)/\partial \theta}{F(\tau_i|x_i;\theta) + p_2 \bar{F}(\tau_i|x_i;\theta)} \right\}$$

where $F(\tau|x;\theta) = 1 - \bar{F}(\tau|x;\theta)$.

Hu and Lawless (1996, 1997) illustrate the use of the four estimation methods on problems involving automobile warranty data, and compare the methods in a simulation study. In their context N is large ($N = 4000$ in the simulation) and the proportion of the population falling into stratum 1 is .25 or smaller. They found that with selection probabilities p_2 in the range .05 - .20, the four estimation methods were all close to unbiased and gave estimators with roughly the same variance. In addition, normal approximations for $\hat{\theta}$ were adequate for the range of population and sample sizes considered. The estimators based on $S_W(\theta)$ and $S_C(\theta)$ are easier to deal with in terms of variance estimation, and $S_W(\theta)$ has the added convenience of being computable with standard censored lifetime data software that allows variable case weights.

5 Additional Remarks

This article reviews several approaches to estimation of parameters when sampling is response-selective with known selection probabilities. Discussion was restricted to two common selection procedures (BSS and BVPS), but extensions to other schemes are possible. For example, in some applications, such as the birth weight study mentioned in Section 1, quota sampling may be used so that the total size N of the population assumed to be generated by (1) is random.

Information about the relative efficiencies of the different estimation procedures is at present quite limited. For the scenario described in Section 4 Hu and Lawless (1997) found all four methods to be comparable. However, in the case of binary responses Wild (1991) and others have found that estimators based on the weighted pseudo score $S_W(\theta)$ can be considerably less efficient than those based on $S_C(\theta)$ or on $L_F(\theta, G)$. Wild also found that $S_C(\theta)$ gave estimates very nearly as efficient as those obtained

from the full likelihood $L_F(\theta, G)$. A limited simulation study by Robins et al. (1994) in a binary response problem, however, revealed situations where all of $S_W(\theta), S_C(\theta)$ and $S_P(\theta)$ were rather inefficient relative to $L_F(\theta, G)$. Further investigation is desirable.

It should be mentioned that another feature of weighted pseudo scores is their applicability in more complex probability sampling situations (e.g. Binder 1983, Binder and Patak 1994) and in situations where only moments of Y given X are modelled, rather than $f(y|x;\theta)$ (O'Hara Hines 1997). However, other more efficient pseudo likelihood methods can also be developed (Robins et al. 1994).

There is a close connection between the methods discussed here and methods for dealing with missing data. Indeed, the present framework can be viewed as one in which covariate values are missing for units that are not selected. The approaches to estimation used here may also be applied with more general missing data problems. Robins et al. (1994, 1995) and Carroll et al. (1995, Chapter 9) provide wide-ranging discussions. Hu and Lawless (1997) also provide general discussion, and some simulation results. Robins et al. deal with very general problems in which the probability an observation is incomplete (has data missing) may depend upon unknown parameter values. They obtain asymptotically optimal estimators of θ within semiparametric models but their methods are generally difficult to implement. As remarked earlier, it is of considerable interest to compare the various approaches in more detail. Lawless et al. (1997) give some results.

Finally, we note another method of estimation that is suggested by the use of the EM algorithm to maximize $L_F(\theta, G)$. By considering the "complete" data log likelihood based on knowledge of all x_i's $(i = 1, ..., N)$,

$$\begin{aligned} \ell_{\text{com}}(\theta, G) &= \sum_{i=1}^{N} R_i\{\log f(y_i|x_i;\theta) + \log dG(x_i)\} \\ &+ (1 - R_i)\{\log Q_{ji}^*(x_i;\theta) + \log dG(x_i)\}, \end{aligned}$$

we obtain the following E-M algorithm, which leads to a stationary point of $\ell_F(\theta, G)$: let $x_1^*, ..., x_m^*$ denote the distinct x_i's observed, and denote $g_r = dG(x_r^*)$. Then we have

E-step: Given current estimates $\tilde{\theta}$, $\tilde{G} \equiv (\tilde{g}_1, ..., \tilde{g}_m)$, compute

$$\tilde{w}_{rj} = \frac{\tilde{g}_r Q_j^*(x_r^*;\tilde{\theta})}{\sum_{s=1}^{m} \tilde{g}_s Q_j^*(x_s^*;\tilde{\theta})}. \tag{18}$$

M-step: Obtain the updated estimate of θ by solving

$$\sum_{j=1}^{k} \sum_{i \in D_j} \frac{\partial \log f(y_i|x_i;\theta)}{\partial \theta} + \sum_{j=1}^{k} (N_j - n_j) \sum_{r=1}^{m} \tilde{w}_{rj} \frac{\partial \log Q_j^*(x_r^*;\theta)}{\partial \theta} = 0. \tag{19}$$

Obtain the updated estimate of G from

$$g_r = \frac{d_r + \sum_{j=1}^{k}(N_j - n_j)\tilde{w}_{rj}}{N},$$

where $d_r = \sum_{i=1}^{N} I(R_i = 1, x_i = x_r^*)$.

If instead of (18) we use the empirical estimates

$$\tilde{w}_{rj} = \sum_{i \in D_j} \frac{I(x_i = x_r^*)}{n_j}$$

in (19), we obtain an estimating equation $S_M(\theta) = 0$, where

$$S_M(\theta) = \sum_{j=1}^{k} \sum_{i \in D_j} \frac{\partial \log f(y_i|x_i;\theta)}{\partial \theta} + \sum_{j=1}^{k} \frac{(N_j - n_j)}{n_j} \sum_{i \in D_j} \frac{\partial \log Q_j^*(x_i;\theta)}{\partial \theta}.$$

A similar idea has been used in a different context by Reilly and Pepe (1995), and it would be of interest to see how it performs in the current framework. It is easily seen that for the example in Section 4, $S_M(\theta)$ is identical with $S_W(\theta)$ and it is also identical when y is categorical with strata corresponding to categories. More generally, however, it is different.

Acknowledgements

This work was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* 51, 279-292.

Binder, D.A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *J. Amer. Statist. Assoc. 89*, 1035-1043.

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

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

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

Hausman, J.A. and Wise, D.A. (1981). Stratification on endogenous variables and estimation: The Gary Income Maintenance Experiment. In *Structural Analysis of Discrete Data with Econometric Applications*, eds. C.F. Manski and D. McFadden, Cambridge, MA: MIT Press, pp. 364-391.

Hoem, J.M. (1985). Weighting, misclassification and other issues in the analysis of survey samples of life histories. Chapter 5 in *Longitudinal Analysis of Labor Market Data*, eds. J.J. Heckman and B. Singer. Cambridge, UK: Cambridge University Press.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *J. Roy. Statist. Soc. A* 143, 474-487.

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

Hu, X.J. and Lawless, J.F. (1997). Pseudo likelihood estimation in a class of problems with response-related missing covariates. To appear in *Canad. J. Statistics.*

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

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

Lawless, J.F., Wild, C.J. and Kalbfleisch, J.D. (1997). Likelihood and pseudo likelihood estimation for response-stratified data. U. of Waterloo Technical Report Stat-97-07.

O'Hara Hines, R.J. (1997). Fitting generalized linear models to retrospectively sampled clusters with categorical responses. To appear in *Canad. J. Statistics.*

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

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., and Zhao, L.P.(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. B* 57, 409-424.

Scott, A.J. and Wild, C.J. (1986). Fitting logistic regression models under case-control or choice-based sampling. *J. Roy. Statist. Soc. B* 48, 170-182.

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

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