Statistica Sinica **27** (2017), 261-285
doi:http://dx.doi.org/10.5705/ss.2014.174

A SEMIPARAMETRIC INFERENCE TO REGRESSION ANALYSIS WITH MISSING COVARIATES IN SURVEY DATA

Shu Yang and Jae Kwang Kim

North Carolina State University and Iowa State University

Abstract: Parameter estimation in parametric regression models with missing covariates is considered under a survey sampling setup. Under missingness at random, a semiparametric maximum likelihood approach is proposed which requires no parametric specification of the marginal covariate distribution. By drawing from the von Mises calculus and V-Statistics theory, we obtain an asymptotic linear representation of the semiparametric maximum likelihood estimator (SMLE) of the regression parameters, which allows for a consistent estimator of asymptotic variance. An EM algorithm for computation is then developed to implement the proposed method using fractional imputation. Simulation results suggest that the SMLE method is robust, whereas the fully parametric method is subject to severe bias under model misspecification. A rangeland study from the National Resources Inventory (NRI) is used to illustrate the practical use of the proposed methodology.

Key words and phrases: Asymptotic linearization representation, fractional imputation, nonparametric maximum likelihood estimator, nonresponse.

1. Introduction

Analyzing survey data to make inference about superpopulation models from finite populations is an area of major interest in survey sampling. When the sampling design is informative, survey data obtained from complex sampling do not follow the distribution of the finite population, sampling weights are incorporated into the estimation procedure to obtain valid inferences about the superpopulation model. Skinner, Holt and Smith (1996), Korn and Graubard (1999), Chambers and Skinner (2003), and Fuller (2009, Ch.6) provide comprehensive overviews on this topic. Regression analysis under informative sampling is, in particular, an important topic in this area. See Chambers (2003), Pfeffermann and Sverchkov (2009), Scott and Wild (2011), Kim and Skinner (2013), and references therein.

When the covariates in the regression have missing values, however, the existing methods for regression analysis under complex sampling cannot be directly applied. Adjustments need to be made to obtain consistent estimation. Little

(1992), Horton and Laird (1999), and Ibrahim et al. (2005) provided comprehensive literature reviews on the regression problem with missing covariates under non-survey sampling setups. Under complex survey sampling, the literature is somewhat sparse. Skinner, Holt and Smith (1996) used a pseudo maximum likelihood method to handle missing covariate under complex sampling. Moore et al. (2009) used response propensity weighting to obtain doubly robust estimation for a logistic regression model. The methods considered in Ibrahim et al. (2005) and Skinner, Holt and Smith (1996) are fully parametric in the sense that the marginal distribution of the covariates is assumed in addition to the conditional distribution of the response variable given the covariates. In the usual regression analyses, the marginal distribution of the covariates need not be assumed under complete response. Only missingness in the covariates calls for such extra assumption.

Semiparametric inference based on an efficient score function has become more popular recently. The semiparametric efficient estimator of Robins, Rotnitzky and Zhao (1994) and Robins, Hsieh and Newey (1995) achieves the semiparametric information bound. Zhao, Lipsitz and Lew (1996) proposed a joint estimating equation approach for missing covariates by modeling the response mechanism. Wang and Paik (2006) and Didelez (2002) provided comparison of the aforementioned semiparametric efficient estimators. In the context of the missing covariate problem, the marginal distribution of the covariates can be viewed as a nuisance parameter. If the nuisance parameter is infinite-dimensional but the regression model itself is parametric, the joint model becomes semiparametric. Zhang and Rockette (2005) considered the problem with a single covariate and obtained a semiparametric efficient estimator of the regression parameters but did not discuss an extension to complex survey sampling.

In this paper we consider, under a complex sampling setup, a semiparametric approach of imputing the missing covariate using nonparametric maximum likelihood estimates of the covariate distribution, which does not require parametric specification of the marginal covariate distribution and therefore enjoys robustness. The proposed method can be implemented using a version of the fractional imputation of Kim (2011), semiparametric fractional imputation, where the imputed values for each missing value are from observed values. Fractional weights of the imputed values are calculated by incorporating the regression model and the nonparametric maximum likelihood estimates of the covariate distribution.

Section 2 provides the basic setup and introduction of the proposed method. In Section 3, main results are presented by drawing from the von Misses calculus and V-statistics theory. In Section 4, the computational aspect of the proposed method is discussed in light of semiparametric fractional imputation. Section 5 shows the results from three simulation studies. A rangeland study from the

National Resources Inventory (NRI) using our method is presented in Section 6, and concluding remarks are made in Section 7.

2. Basic Setup and the Proposed Method

Suppose we are interested in estimating θ in a regression model $f(y \mid x;\theta)$ for $\theta \in \Theta \subset \mathbb{R}^d$, when the covariate x has missing values. The finite population is assumed to be a random sample from a model with a joint density $f(y \mid x;\theta)g(x)$, where $g(x)$, the marginal density of x, is completely unspecified. Let $U = \{1, 2, \ldots, N\}$ be the index set of the finite population and $A \subset U$ be the index set of the sample obstained by a probability sampling. Without loss of generality, we assume that $A = \{1, \cdots, n\}$. Let w_i be the sampling weight of unit i in the sample such that $N^{-1}\sum_{i\in A} w_i y_i$ is consistent to the population mean $\mu_Y = N^{-1}\sum_{i=1}^{N} y_i$. Let $\delta_i = 1$ if x_i is observed and $\delta_i = 0$ if x_i is missing. We assume that the missing mechanism is missing at random (MAR) in the sense that

$$P(\delta = 1 \mid x, y) = P(\delta = 1 \mid y).$$

Under complete response, the pseudo maximum likelihood estimator (PMLE) of θ can be obtained as a solution to

$$\sum_{i=1}^{n} w_i S(\theta; x_i, y_i) = 0, \tag{2.1}$$

where $S(\theta; x, y) = \partial \ln f(y \mid x;\theta)/\partial\theta$ is the score function of θ. Godambe and Thompson (1986) and Chambers et al. (2012) have built a solid theoretical base for the PMLE under complex sampling. In the presence of missing data, the PMLE of θ can be obtained by solving

$$\sum_{i=1}^{n} w_i \delta_i S(\theta; x_i, y_i) + \sum_{i=1}^{n} w_i (1-\delta_i) E\{S(\theta; X, y_i) \mid y_i\} = 0. \tag{2.2}$$

The conditional expectation in (2.2) can be written as

$$E\{S(\theta; X, y_i) \mid y_i\} = \frac{\int S(\theta; x, y_i) f(y_i \mid x;\theta) dP^X(x)}{\int f(y_i \mid x;\theta) dP^X(x)}, \tag{2.3}$$

where P^X is the (unknown) marginal distribution of X. In the context of measurement error models, Pepe and Fleming (1991) used a nonparametric estimate of (2.3) for discrete data and Carroll, Knickerbocker and Wang (1995) used a kernel method to estimate the conditional distribution $f(x \mid y)$. On the other hand, Ibrahim, Chen and Lipsitz (1999) considered a fully parametric method in modeling both the marginal distribution of x and the conditional distribution $f(y \mid x;\theta)$. The nonparametric methods described above have limited applicability due to the curse of dimensionality. The fully parametric approach is very sensitive to departures from the parametric modeling assumptions.

We consider a semiparametric model in the sense that we assume a parametric model for the conditional distribution $f(y \mid x; \theta)$, but the marginal distribution of x, often not a major interest of the study, is completely unspecified. We start a formal discussion of the semiparametric approach with the population-level log-likelihood of θ and P^X as

$$l(\theta, P^X) = \sum_{i=1}^{N} \delta_i \left\{ \log f(y_i \mid x_i; \theta) + \log P^X(x_i) \right\} + \sum_{i=1}^{N} (1 - \delta_i) \log f(y_i; \theta, P^X),$$

where $f(y_i; \theta, P^X) = P^X f(y_i|x; \theta) = \int f(y_i|x; \theta) dP^X(x)$.

Thus, the observed pseudo log-likelihood of θ and P^X is

$$l_{obs}(\theta, P^X) = \sum_{i=1}^{n} w_i \delta_i \left\{ \log f(y_i|x_i; \theta) + \log P^X(x_i) \right\} + \sum_{i=1}^{n} w_i (1 - \delta_i) \log f(y_i; \theta, P^X). \tag{2.4}$$

The global maximization of $l_{obs}(\theta, P^X)$ over the parameter space $\Theta \times \mathcal{G}$ is infinite-dimensional, where $\Theta \subset \mathbb{R}^d$ is the parameter space for θ and $\mathcal{G}$ is the set of all probability measures on X. For a simpler maximization, we restrict the support of P^X to belong to the set of the observed values of X. For simplicity of notation, assume that we have full response in the first r units and partial response in the remaining $n - r$ units. Let $\pi_k = P(x = x_k)$ be the point mass assigned to the observed x_k such that $\sum_{k=1}^{r} \pi_k = 1$. We focus on the observed pseudo log-likelihood for θ and $\pi = (\pi_1, \ldots, \pi_r)$ given by

$$l_{obs}(\theta, \pi) = \sum_{i=1}^{r} w_i \left\{ \log f(y_i|x_i; \theta) + \log \pi_i \right\} + \sum_{i=r+1}^{n} w_i \log \{ \sum_{j=1}^{r} f(y_i|x_j; \theta) \pi_j \}, \tag{2.5}$$

where $\sum_{j=1}^{r} f(y_i|x_j; \theta)\pi_j$ in (2.5) can be viewed as an approximation to $f(y_i; \theta, P^X) = \int f(y_i|x; \theta) dP^X(x)$ in (2.4). The observed pseudo log-likelihood is semiparametric because we have a parametric component θ and a non-parametric component π. Such semiparametric models have been considered in Lawless, Kalbfleisch and Wild (1999), Scott and Wild (2001), Scott and Wild (2002), and Breslow and Wellner (2007) mostly under two-phase sampling setups. Maximizing the observed pseudo log-likelihood in (2.5) with respect to (θ, π) subject to $\sum_{i=1}^{r} \pi_i = 1$ leads to the semiparametric maximum likelihood estimator (SMLE) of (θ, π). The asymptotic properties of the SMLE of θ will be discussed in the next section.

Remark. We can easily extend the above setup to a multiple regression problem where one covariate has missing values and other covariates are completely observed. To illustrate this point, we consider a multiple regression problem $y_p = \beta_0 + \beta_1 y_1 + \cdots + \beta_{p-1} y_{p-1} + \beta_p x + \epsilon$, where x has missing values and

covariates $y_1, \ldots, y_{p-1}$ are completed observed in the sample. Let y now be a p-dimensional vector $y = (y_1, \ldots, y_p)$. Under our setup, we assume the distribution of $y \mid x$ is a product of a series of one-dimensional conditional distributions

$$f(y \mid x; \theta) = f_1(y_1 \mid x; \theta_1) f_2(y_2 \mid x, y_1; \theta_2) \cdots f_p(y_p \mid x, y_1, \ldots, y_{p-1}; \theta_p), \quad (2.6)$$

where θ_k is the parameter in the kth conditional distribution of y_k given x and $y_1, \ldots, y_{k-1}$. Therefore, θ_p is the parameter of primary interest in the multiple regression problem. In some situations, it is difficult to find a natural distribution of $y \mid x$. Consider, for example, that y contains a continuous variable y_1 and a binary variable y_2. A suitable distribution of y given x may be obtained from (2.6) by specifying a normal distribution of $y_1 \mid x$ and a logistic regression model for $y_2 \mid y_1, x$ treating y_1, x as covariates in the logistic regression model. This seems to be a natural specification of the distribution of (y_1, y_2) given x in this setting. See Section 5.3 for an illustration.

3. Main Theoretical Results

In this section, we establish the asymptotic properties of the proposed SMLE $\hat{\theta}$. We present the theory here and leave proofs to Appendices.

Let

$$h(\theta, P^X; z) = \delta \log f(y|x; \theta) + (1 - \delta) \log P^X f(y|X; \theta),$$

where $z = (\delta, \delta x, y)$, $P^X g(X) = \int g(x) dP^X(x)$, and P^X is the marginal distribution of X. Define the population empirical distribution induced by $Z_1, \ldots, Z_N$, potentially available for all units in the population as $P_N^Z = N^{-1} \sum_{i=1}^N \delta_{Z_i}$, where δ_{Z_i} is the Dirac function of Z_i. Thus, given a measurable function $g(Z)$, the expectation of g under P_N^Z is $P_N^Z g(Z) = N^{-1} \sum_{i=1}^N g(Z_i)$. In the survey sampling context, a sample is selected according to a probability sampling scheme. Define the sample empirical measure by $P_n^Z = N^{-1} \sum_{i=1}^n w_i \delta_{Z_i}$ so that $P_n^Z g(Z) = N^{-1} \sum_{i=1}^n w_i g(Z_i)$. Let I_i be the sampling indicator of unit i, $I_i = 1$ if unit i is selected in the sample and 0 otherwise. Note that $P_n^Z g(Z) = N^{-1} \sum_{i=1}^N w_i I_i g(Z_i)$ and we assume that $E P_n^Z g(Z) = E P_N^Z g(Z) = P_0^Z g(Z)$, where we use the subscript 0 to index the true probability measure. Let P_n^X represent the nonparametric distribution of X indexed by π, so that $P_n^X = \sum_{i=1}^n \delta_i \pi_i \delta_{X_i}$. Thus, given a measurable function $g(X)$, the expectation of g under P_n^X is $P_n^X g(X) = \sum_{i=1}^n \delta_i \pi_i g(X_i)$. Therefore, the SMLE $\hat{\theta}$ maximizes $P_n^Z h(\theta, P_n^X; z)$.

We now make the following assumptions.

(C1) *There exists a positive number* ξ_1 *such that* $P(\delta = 1 \mid x, y) = P(\delta = 1 \mid y) > \xi_1 > 0$.

(C2) *There exists a positive number* ξ *such that* $P(I_i = 1 \mid X_i, Y_i, \delta_i) > \xi > 0$, *and the sampling design is consistent in the sense that* $P_n^Z g(Z)$ *is consistent to* $P_N^Z g(Z)$ *for any bounded and measurable function* $g(Z)$.

(C3) *For some* $\epsilon_0 > 0$, $\{\delta \log f(Y|X;\theta) : \theta \in \Theta\}$ *and* $\{(1-\delta)\log P^X f(Y|X;\theta) : \theta \in \Theta, ||P^X - P_0^X|| < \epsilon_0\}$ *are* P^Z *Glivenko-Cantelli.*

(C4) $\log\{f(Y|X;\theta)f(X)\}$ *is dominated by an integrable function* $F(X,Y)$, *i.e.,* $|\log\{f(Y|X;\theta)f(X)\}| < F(X,Y)$ *for any* $\theta \in \Theta$ *and* $E\{F(X,Y)\} < \infty$.

Lemma 1. *Let* $\hat{\theta}$ *be the SMLE of* θ *and* θ_0 *be the true parameter value of* θ, *interior to the compact parameter space* Θ. *Let* P_0^X *be the true probability measure of* X. *Under the MAR assumption and* (C1)–(C4), $\hat{\theta} - \theta_0 \to 0$ *in probability as* $n \to \infty$.

A proof of Lemma 1 is in Appendix A. Condition (C1) requires that the missing mechanism be MAR in the sense of Rubin (1976). Condition (C2) is commonly used in survey sampling. It means that the sampling design is consistent in the sense that the sample empirical measure of any bounded, measurable function is consistent to the population empirical measure. See Fuller (2009) for some sufficient conditions for (C2). In Condition (C3), the Glivenko-Cantelli property is imposed on a family of functions for which the uniform strong law of large numbers holds (Van Der Vaart and Wellner (1996, page 81)). Conditions (C3) and (C4) are the usual regularity conditions for the consistency of the SMLE of the regression models in a simple random sample (See, for example, Van der Vaart (2000, Chap. 25)), which applies to such as the linear regression model, the logistic regression model, and the Poisson regression model. Therefore, if we treat the finite population as a simple random sample from a superpopulation, the consistency of the SMLE obtained based on the finite population follows by a similar argument as in Van der Vaart (2000). Condition (C2) then preserves the consistency of the SMLE obtained based on the survey sample.

(C5) $\delta \log f(Y|X;\theta)$ *and* $(1-\delta)\log P^X f(Y|X;\theta)$ *are continuously twice differentiable with respect to* θ *and* $\{\delta\partial^2 \log f(Y|X;\theta)/\partial\theta\partial\theta^T : \theta \in \Theta\}$ *and* $\{(1-\delta)\cdot \partial^2 \log P^X f(Y|X;\theta)/\partial\theta\partial\theta^T : \theta \in \Theta, ||P^X - P_0^X|| < \epsilon_0\}$ *are* P^Z *Glivenko-Cantelli, and non-singular at* θ_0.

(C6) $E\{S(\theta, P^X; z)^3\} < \infty$, *where* $S(\theta, P^X; z) = \delta\partial \log f(y|x;\theta)/\partial\theta + (1-\delta)\partial \log P^X f(y|X;\theta)/\partial\theta$.

Theorem 1. *Under* (C1)–(C6), $\hat{\theta}$ *has the asymptotic linear representation,*

$$\hat{\theta} - \theta_0 = \sum_{i=1}^{n} w_i \kappa(z_i; \theta_0, P_0^X) + o_p(n^{-1/2}), \tag{3.1}$$

where $\kappa(z_i; \theta_0, P_0^X)$ *is defined in* (C.4) *in Appendix C. Thus*

$$\Sigma^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_{d\times d})$$

as $n \to \infty$, *with* $\Sigma = Var\left\{\sum_{i=1}^n w_i \kappa(Z_i; \theta_0, P_0^X)\right\}$, *where* $I_{d\times d}$ *is the* $d\times d$ *identity matrix.*

The proof of Theorem 1 is in Appendix C; which relies on the von Mises calculus (Fernholz (1983)) and V-statistic theory (von Mises (1947); Hoeffding (1948)). Even for the complete response problem, it appears difficult to formulate a single set of conditions that cover most sampling designs of interest. Instead, we establish the V-statistic theory for Poisson samples in Appendix B. The generalization of the V-statistic theory for general complex sampling designs can be established under similar regularity conditions. Fuller (1998) considered Poisson sampling in a two-phase sampling problem and argued that Poisson sampling is a good approximation. Condition (C5) requires that the function h be sufficiently smooth. Together with (C2), it implies that $P_n^Z \partial^2 h(\theta, P^X; z)/\partial\theta\partial\theta^T$ converges uniformly to $E\{\partial^2 h(\theta, P^X; z)/\partial\theta\partial\theta^T\}$ for $\theta \in \Theta$ and $||P^X - P_0^X|| < \epsilon_0$. Condition (C6) is a moment condition for the Central Limit Theorem.

For variance estimation, let

$$\hat{\Sigma} = \sum_{i=1}^n \sum_{j=1}^n \Delta_{ij} \kappa(z_i; \hat{\theta}, P_n^X)\kappa(z_j; \hat{\theta}, P_n^X)^T + \hat{V}\{\sum_{i=1}^N k(z_i; \theta, P_0^X)\}, \tag{3.2}$$

which is a consistent estimator for the variance of $\hat{\theta}$, where Δ_{ij} are the coefficients for variance estimation. For example, under simple random sampling, $\Delta_{ij} = -1/\{n^2(n-1)\}$ for $i \neq j$ and $\Delta_{ii} = 1/n^2$. The second term in (3.2) is a consistent estimator of $V\{\sum_{i=1}^N k(z_i; \theta, P_0^X)\}$, taking into account the second term in (6.2.9) of Fuller (2009) for the case $n/N = O(1)$. The second term is needed since, under Condition (C2), we have $n/N = O(1)$. The linearization method may involve specialized programming for different models. In contrast, the Jackknife method of variance estimation can be easily implemented (See Appendix D).

Our setup has a broad scope, including multi-stage sampling design, stratified sampling, and cluster sampling, among others. For illustration, we provide a detailed description of our method under a two-stage sampling design in Appendix E.

4. Computation

We propose an EM algorithm to compute the SMLE of θ. Assume that x has observed values on the realized sample support $S_x = \{x_1, \ldots, x_r\}$. Maximizing the observed log-likelihood

$$l_{obs}(\theta,\pi)=\sum_{i=1}^{r} w_i\{\log f(y_i|x_i;\theta)+\log\pi_i\}+\sum_{i=r+1}^{n} w_i \log\left\{\sum_{j=1}^{r}\delta_j f(y_i|x_j;\theta)\pi_j\right\} \quad (4.1)$$

subject to $\sum_{i=1}^{r}\pi_i = 1$ with respect to (θ,π) can be obtained by applying the Lagrange multiplier method. The solution to this optimization is given by solving

$$\sum_{i=1}^{r} w_i S(\theta;x_i,y_i)+\sum_{i=r+1}^{n} w_i\left\{\frac{\sum_{j=1}^{r}\pi_j f(y_i\mid x_j;\theta)S(\theta;x_j,y_i)}{\sum_{j=1}^{r}\pi_j f(y_i\mid x_j;\theta)}\right\}=0, \quad (4.2)$$

$$\pi_k=\frac{w_k+\sum_{i=r+1}^{n} w_i w_{ik}^*(\theta)}{\sum_{i=1}^{n} w_i}, \quad (4.3)$$

where $w_{ij}^*(\theta)=\pi_j f(y_i|x_j;\theta)/\sum_{k=1}^{r}\pi_k f(y_i|x_k;\theta)$.

To obtain the solution to (4.2) and (4.3), an EM algorithm using fractional imputation can be applied:

Step 0. For each unit with $\delta_i = 0$, r imputed values of x are assigned with $x_{ij}^* = x_j$. Let $\pi_k^{(0)} = 1/r$ and $\theta^{(0)}$ be the PMLE of θ using only respondents.

Step 1. At the tth EM iteration, compute the fractional weight

$$w_{ij(t)}^*=\frac{f(y_i\mid x_{ij}^*;\theta^{(t)})\pi_j^{(t)}}{\sum_{k=1}^{r} f(y_i\mid x_{ik}^*;\theta^{(t)})\pi_k^{(t)}}.$$

Step 2. Use $w_{ij(t)}^*$ and (x_{ij}^*, y_i) to update the parameters by solving the imputed score equation

$$\sum_{i=1}^{r} w_i S(\theta;x_i,y_i)+\sum_{i=r+1}^{n} w_i\sum_{j=1}^{r} w_{ij(t)}^* S(\theta;x_{ij}^*,y_i)=0, \quad (4.4)$$

$$\pi_k^{(t+1)}=\frac{w_k+\sum_{i=r+1}^{n} w_i w_{ik(t)}^*}{\sum_{i=1}^{n} w_i}. \quad (4.5)$$

Step 3. Set $t = t + 1$ and go to Step 1. Continue until convergence.

Step 1 is the E-step in the EM algorithm. Step 2 is the M-step that uses (4.4) and (4.5) to update the parameters. An important property of the EM algorithm is that, $l_{obs}(\theta^{(t+1)},\pi^{(t+1)}) \geq l_{obs}(\theta^{(t)},\pi^{(t)})$. To see this, write

$$\begin{aligned} &l_{obs}(\theta^{(t+1)},\pi^{(t+1)})-l_{obs}(\theta^{(t)},\pi^{(t)})\\ &=\sum_{i=1}^{r} w_i\left[\log\{f(x_i,y_i;\theta^{(t+1)})\pi_i^{(t+1)}\}-\log\{f(x_i,y_i;\theta^{(t)},\pi^{(t)})\pi_i^{(t)}\}\right] \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=r+1}^{n} w_i \left[\log \left\{ \frac{\sum_{j=1}^{r} f(y_i|x_j;\theta^{(t+1)})\pi_j^{(t+1)}}{\sum_{j=1}^{r} f(y_i|x_j;\theta^{(t)})\pi_j^{(t)}} \right\} \right] \\
&\geq \sum_{i=1}^{r} w_i \left[\log\{f(x_i,y_i;\theta^{(t+1)})\pi_i^{(t+1)}\} - \log\{f(x_i,y_i;\theta^{(t)},\pi^{(t)})\pi_i^{(t)}\} \right] \\
&\quad + \sum_{i=r+1}^{n} w_i \left[\sum_{j=1}^{r} \log \left\{ \frac{f(y_i|x_j;\theta^{(t+1)})\pi_j^{(t+1)}}{f(y_i|x_j;\theta^{(t)})\pi_j^{(t)}} \right\} \frac{f(y_i|x_j;\theta^{(t)})\pi_j^{(t)}}{\sum_{k=1}^{r} f(y_i|x_k;\theta^{(t)})\pi_k^{(t)}} \right] \\
&= \sum_{i=1}^{r} w_i \left[\log\{f(x_i,y_i;\theta^{(t+1)})\pi_i^{(t+1)}\} - \log\{f(x_i,y_i;\theta^{(t)},\pi^{(t)})\pi_i^{(t)}\} \right] \\
&\quad + \sum_{i=r+1}^{n} w_i \sum_{j=1}^{r} w_{ij(t)}^{*} \left[\log\{f(y_i|x_j;\theta^{(t+1)})\pi_j^{(t+1)}\} - \log\{f(y_i|x_j;\theta^{(t)})\pi_j^{(t)}\} \right] \\
&= \sum_{i=1}^{r} w_i \log\{f(x_i,y_i;\theta^{(t+1)})\} + \sum_{i=r+1}^{n} w_i \sum_{j=1}^{r} w_{ij(t)}^{*} \log\{f(y_i|x_j;\theta^{(t+1)})\} \\
&\quad - \sum_{i=1}^{r} w_i \log\{f(x_i,y_i;\theta^{(t)})\} - \sum_{i=r+1}^{n} w_i \sum_{j=1}^{r} w_{ij(t)}^{*} \log\{f(y_i|x_j;\theta^{(t)})\} \\
&\quad + \sum_{i=1}^{r} w_i \log \pi_i^{(t+1)} + \sum_{i=r+1}^{n} w_i \sum_{j=1}^{r} w_{ij(t)}^{*} \log \pi_j^{(t+1)} \\
&\quad - \sum_{i=1}^{r} w_i \log \pi_i^{(t)} - \sum_{i=r+1}^{n} w_i \sum_{j=1}^{r} w_{ij(t)}^{*} \log \pi_j^{(t)} \geq 0,
\end{aligned}
$$

where the first inequality follows by Jensen's inequality and the last line follows by the M-step of the EM algorithm. Thus, the sequence $\{l_{obs}(\theta^{(t)},\pi^{(t)})\}$ is monotone increasing, bounded above if the SMLE exists, and converges to some value l^*. In most cases, l^* is a stationary value in the sense that $l^* = l_{obs}(\theta^*,\pi^*)$ for some (θ^*,π^*) at which $\partial l_{obs}(\theta,\pi)/\partial(\theta,\pi) = 0$. Under fairly weak conditions, the EM sequence $\{(\theta^{(t)},\pi^{(t)})\}$ converges to a stationary point (θ^*,π^*). Furthermore, if $l_{obs}(\theta,\pi)$ is uni-modal with (θ^*,π^*) the only stationary point, $\{(\theta^{(t)},\pi^{(t)})\}$ converges to the unique maximizer of l_{obs}. Further convergence details can be found in Wu (1983) and McLachlan and Krishnan (2007).

The weights $w_{ij}^{*}(\hat{\theta})$ assigned to imputed values can be called fractional weights. Imputed values are not changed, only fractional weights are updated for each EM iteration. The proposed method is an application of the parametric fractional imputation of Kim (2011), but instead of assuming a parametric model for the marginal distribution of x, we used a nonparametric model. Paik (2000) proposed the same method in the context of missing covariates in logistic regression.

5. Simulation Study

To evaluate the performance of the proposed estimator, we conducted three Monte Carlo simulation studies.

5.1. Simulation one - linear model

Finite populations of size $N = 2,000$ were generated according to

$$x_i \sim \text{Beta } (0.5, 1), \tag{5.1}$$

$$y_i | x_i \sim \text{Normal } (\beta_0 + \beta_1 x_i, \sigma^2), \tag{5.2}$$

where $(\beta_0, \beta_1, \sigma^2) = (0, 5, 1)$. Each unit in the finite population was associated with a size variable, $z_i \sim \text{Gamma } (x_i + |y_i| + 1, 1)$. For each finite population, we generated a sample of size $n = 100$ by the probability proportional to size (PPS) sampling method. Let p_i be the selection probability for PPS sampling, where $p_i = nz_i / \sum_{i=1}^{N} z_i$. If $p_i > 1$, we set p_i to be 1. The sampling weight was $w_i = 1/p_i$. The sampling mechanism was informative by construction.

We also generated δ_i, the response indicator variable of x_i, from Bernoulli(ϕ_i) with $\phi_i = 0.75$(MCAR) and $\text{logit}(\phi_i) = -1 + 2y_i$(MAR), such that the response rate was 0.75. Interest was estimating the regression parameters β_0 and β_1.

We compared the proposed semiparametric maximum likelihood estimator (SMLE) over 1,000 datasets with three other estimators: CC, the complete case analysis discarding the cases with missing values; PMLE_w, the pseudo maximum likelihood estimator obtained by solving (2.2) assuming $x_i \sim N(\mu_x, \sigma_x^2)$; PMLE_t, the pseudo MLE assuming $x_i \sim \text{Beta}(\alpha, \beta)$, as in (5.1). In PMLE_w the covariate distribution was wrongly specified, whereas in PMLE_t the covariate distribution was correctly specified. In SMLE, PMLE_w, and PMLE_t, the regression model was correctly specified as in (5.2). For variance estimation, we considered the conventional delete-one Jackknife variance estimator.

Table 1 presents numerical results for the linear regression under MCAR and MAR. Each method and parameter combination has a point estimate and a variance estimate. The Monte Carlo bias (Bias) and variance (Var) are the bias and variance for the point estimates over the Monte Carlo samples. $E(\widehat{\text{Var}})$ is the Monte Carlo mean of the variance estimates over the Monte Carlo samples.

Under MCAR, CC is unbiased in estimating all parameters. However, it is inefficient compared with other methods. Under MAR, CC is shown to be invalid as it is associated with a large bias in the regression parameters considered. This indicates that analysis ignoring missing values can be misleading.

If the covariate distribution is correctly specified, PMLE (PMLE_t) is both unbiased and efficient. However, if the covariate distribution is misspecified,

Table 1. Linear regression estimation under MCAR and MAR. CC: the complete case estimator; PMLE_t: pseudo MLE under the true model; PMLE_w: pseudo MLE under model misspecification; SMLE: Semiparametric MLE.

		Bias		Var		$E(\widehat{\text{Var}})$	
Setup	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
MCAR	CC	0.00	0.00	0.0086	0.0395	0.0089	0.0415
	PMLE_t	0.00	0.00	0.0073	0.0330	0.0075	0.0343
	PMLE_w	0.00	0.00	0.0075	0.0343	0.0078	0.0353
	SMLE	0.00	0.00	0.0073	0.0335	0.0075	0.0345
MAR	CC	0.10	-0.07	0.0093	0.0384	0.0096	0.0404
	PMLE_t	0.00	0.00	0.0072	0.0319	0.0074	0.0333
	PMLE_w	0.04	-0.08	0.0074	0.0317	0.0076	0.0327
	SMLE	0.00	0.01	0.0072	0.0322	0.0074	0.0332

PMLE (PMLE_w) can be biased. Under MAR, PMLE_w is biased in estimating the parameters of interest.

In all scenarios, SMLE is unbiased. On the other hand, PMLE_t is more efficient than SMLE, but the efficiency gain is not significant. In practice, misspecification of covariate distribution for PMLE is a big concern since it is often difficult to specify a correct parametric model when missing data are present. SMLE is attractive since it avoids error-prone model speculation. Variance estimation of the SMLE is also nearly unbiased in this simulation.

5.2. Simulation two - Poisson regression model

We considered a Poisson regression model with a canonical link including an intercept. The complete-data pseudo log-likelihood was

$$l_{obs}(\theta, P^X) = \sum_{i \in A} w_i \{y_i(\beta_0 + \beta_1 x_i) - \exp(\beta_0 + \beta_1 x_i)\} + \sum_{i \in A} w_i \log\{P^X(x_i)\}.$$

The data generating process was the same as in Simulation One except for the conditional distribution, $y_i|x_i \sim$ Poisson (μ_i), where $\log \mu_i = \log\{E(y_i|x_i)\} = \beta_0 + \beta_1 x_i$ with $\beta_0 = 0$ and $\beta_1 = 1$. As in Simulation One, $x_i \sim$ Beta $(0.5, 1)$. Intereste was estimating the regression parameters, but $n = 100$ here. Table 2 summarizes numerical results obtained for the Poisson regression with MCAR and MAR. The results are in line with Simulation One, with similar conclusions drawn.

5.3. Simulation three - multiple regression model

We considered the populations to consist of $N_I = 100$ clusters of size M_i, where $M_i \sim$ Binom $(50, 0.5) + 50$. Thus, the cluster size ranged from 50 to

Table 2. Poisson regression estimation under MCAR and MAR. CC: the complete case estimator; PMLE_t: pseudo MLE under the true model; PMLE_w: pseudo MLE under model misspecification; SMLE: Semiparametric MLE.

		Bias		Var		$E(\widehat{\text{Var}})$	
Setup	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
MCAR	CC	0.00	-0.01	0.0267	0.0962	0.0274	0.0997
	PMLE_t	0.00	-0.01	0.0234	0.0884	0.0236	0.0914
	PMLE_w	-0.02	-0.02	0.0253	0.0957	0.0254	0.0988
	SMLE	0.00	0.00	0.0237	0.0901	0.0239	0.0928
MAR	CC	0.11	-0.05	0.0257	0.0841	0.0258	0.0851
	PMLE_t	0.00	-0.02	0.0227	0.0852	0.0227	0.0903
	PMLE_w	0.00	-0.05	0.0260	0.0966	0.0262	0.1000
	SMLE	0.00	-0.01	0.0229	0.0867	0.0230	0.0916

100. We considered two-stage cluster sampling to generate samples with the final sample size $n = 100$. In the first stage of the cluster sampling we selected $n_I = 10$ clusters using PPS sampling with selection probability proportional to M_i, $p_i = M_i/(\sum_{i=1}^{100} M_i)$, and in the second stage within each selected cluster, we sampled $m_i = 10$ units by simple random sampling. We generated three values in the population according to $x_{1ij} \sim$ Beta $(0.5, 1)$, $x_{2ij}|x_{1ij} \sim$ Normal $(\alpha_0 + \alpha_1 x_{1ij}, \sigma_x^2)$, and $y_{ij}|x_{1ij}, x_{2ij} \sim$ Normal $(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}, \sigma^2)$, where i indexes cluster, j indexes unit within clusters, $\alpha_0 = 0$, $\alpha_1 = 5$, $\sigma_x^2 = 1$, $\beta_0 = 0$, $\beta_1 = -1$, $\beta_2 = 2$, $\sigma^2 = 1$. The regression parameters of Y on X_1 and X_2 were of primary interest. We took only X_1 to have missing values. We assumed $f(y, x_2|x_1; \theta) = f(y|x_1, x_2; \beta) f(x_2|x_1; \alpha)$, where $\beta = (\beta_0, \beta_1, \beta_2, \sigma^2)$ and $\alpha = (\alpha_0, \alpha_1, \sigma_x^2)$, with the latter being a nuisance parameter in the sense that we were not directly interested in estimating α; however we needed to estimate it in order to estimate $\beta = (\beta_0, \beta_1, \beta_2, \sigma^2)$. We let δ_{ij} be the response indicator variable of x_{1ij} generated from Bernoulli(r_{ij}), with $r_{ij} = 0.75$ (MCAR) and $logit(r_{ij}) = \phi_0 + \phi_1 y_{ij}$ (MAR), where $(\phi_0, \phi_1) = (0, 0.3)$. The standard errors were calculated using the customary delete-a-cluster Jackknife variance estimator of Rao, Wu and Yue (1992), see Appendix E for details. Table 3 summarizes numerical results for the multiple regression with MCAR and MAR. The results are consistent with Simulation One and Simulation Two in that SMLE is unbiased under both MCAR and MAR (in contrast to CC, which is biased under MAR) and robust (in constrast to PMLE, which is biased under a misspecified covariate distribution).

Table 3. Multiple regression estimation under MCAR and MAR. CC: the complete case estimator; PMLE_t: pseudo MLE under the true model; PMLE_w: pseudo MLE under model misspecification; SMLE: Semiparametric MLE.

		Bias			Var			$E(\widehat{\text{Var}})$		
Setup	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
MCAR	CC	0.00	-0.01	0.00	0.033	0.514	0.014	0.033	0.558	0.015
	PMLE_t	0.00	0.00	0.00	0.025	0.482	0.012	0.024	0.451	0.015
	PMLE_w	0.02	0.02	0.00	0.027	0.475	0.012	0.026	0.443	0.011
	SMLE	0.00	0.00	0.00	0.025	0.487	0.013	0.025	0.527	0.014
MAR	CC	0.13	0.00	-0.02	0.044	0.518	0.016	0.044	0.566	0.017
	PMLE_t	0.00	0.00	0.00	0.025	0.478	0.012	0.024	0.436	0.011
	PMLE_w	0.06	-0.15	0.01	0.030	0.446	0.011	0.029	0.414	0.010
	SMLE	0.00	0.00	0.00	0.025	0.487	0.013	0.025	0.528	0.014

6. Data Example

The National Resources Inventory (NRI) is a stratified, two-stage area sample of non-federal lands in the United States conducted by the Natural Resources Conservation Service (NRCS) of the U.S. Department of Agriculture (USDA). One of the NRI onsite surveys is a longitudinal study on rangeland, with rangeland defined as a land/use category in which the plant cover is composed of native grass, grass-like plants, and forbs for grazing and browsing. The characteristic of interest in this study is the percentage of the non-native plant species. We focus on the samples with repeated measures in years 2005 and 2013, denoted by (x_i, y_i), where i indexes the sampling unit called segment, x_i is the observation in 2005, and y_i is the observation in 2013. Since we have repeated measurements on segments over time, we can estimate change in rangeland characteristics.

All variables in the sample data are completely observed. The original scale of x and y was coded from 1 to 6. We first transformed the original variables using $\log[(x/6.1)/\{1 - x/6.1\}]$ and $\log[(y/6.1)/\{1 - y/6.1\}]$. Hereafter, we use x and y to denote the variables on a transformed scale.

To evaluate the performance of the SMLE, we generated missingness for x_i intentionally. Specifically, we considered x_i to be subject to missingness and y_i is completely observed. We generated the response indicator δ_i for x_i, which equals 1 if x_i is available and 0 otherwise.

We created four scenarios by different missing mechanisms (MCAR and MAR) and response rates (70% and 50%). Under MAR, we generated δ_i from a Binomial distribution with probability

$$Pr(\delta_i = 1) = \frac{\exp(\phi_0 + \phi_1 y_i)}{1 + \exp(\phi_0 + \phi_1 y_i)},$$

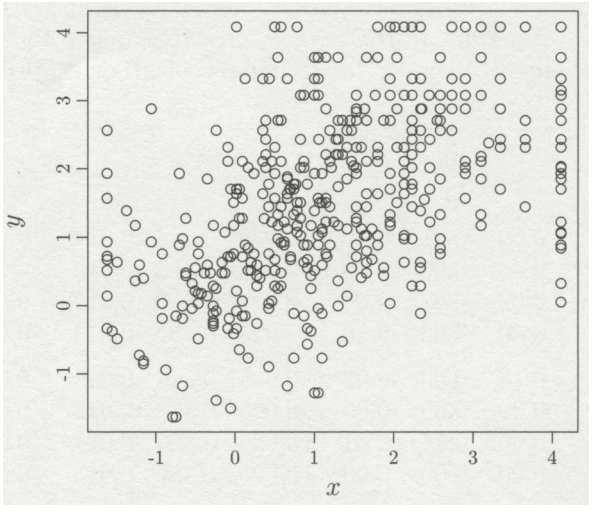

Figure 1. Scatter plot of Y against X based on the full sample (on a transformed scale).

where $(\phi_0, \phi_1) = (-1, 1.3)$ and $(\phi_0, \phi_1) = (-2, 1.15)$ correspond to 70% and 50% response rate, respectively.

We assumed $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. Figures 1, 2, and 3 show the scatter plot of Y against X, the scatter plot of the residuals from the linear regression of Y on X based on the full data, and the QQ plot of residuals, respectively. Figures 1 and 2 suggest a linear regression of Y on X with constant variance of the error term. In Figure 3, points lie closely to the diagonal line, which suggests that the normality assumption of the error term is adequate.

We were interested in estimating the regression parameters β_0, β_1 and σ^2. Table 4 shows the results for the full sample (FULL) estimator, the complete case (CC) sample estimator, the pseudo maximum likelihood estimator further assuming a normal distribution of x_i (PMLN), and the semiparametric maximum likelihood estimator (SMLE). The standard errors were calculated using the customary delete-a-cluster Jackknife variance estimator of Rao, Wu and Yue (1992).

Table 4 presents the numerical results for the NRI rangeland study. Under MCAR, CC is close to FULL but associated with larger standard errors. In such situations, the CC analysis is valid but loses efficiency due to discarding incomplete cases. Under MAR, compared to FULL, CC is associated with a large bias in estimating the intercept and slope. Considering the missing mechanism, the units with larger outcomes are more likely to respond, and therefore the points on the bottom left corner of the (x, y) plane are more likely to be excluded from the CC analysis, which explains the fact that the CC estimate of the intercept

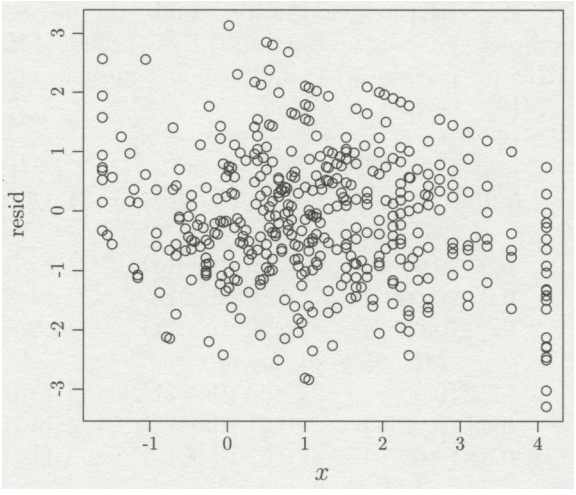

Figure 2. Scatter plot of residuals from linear regression of Y on X against X based on the full sample (on a transformed scale).

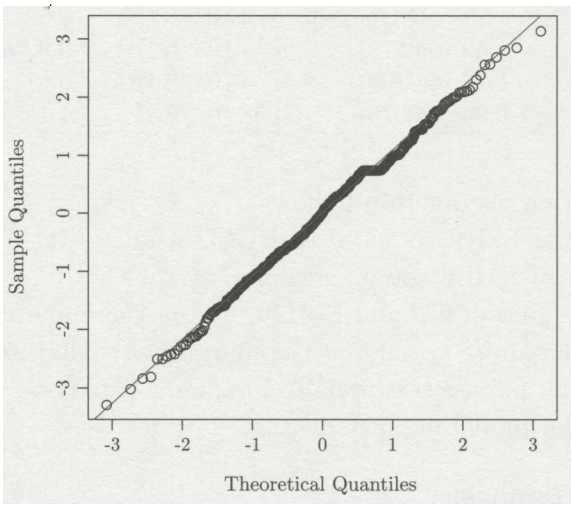

Figure 3. QQ plot of the residuals from a linear regression of Y on X based on the full sample (on a transformed scale).

tends to be larger than the FULL estimate, while the CC estimate of the slope tends to be smaller than the FULL estimate. As the response rate decreases, the bias increases.

Under MAR, there is a remarkable difference in the point estimates between FULL and PMLN, suggesting that PMLN is probably biased due to model misspecification. Due to the difficulty in correctly specifying a parametric model, practitioners often choose to use a normal model. That would lead to biased

Table 4. Results for NRI rangeland study. Full: the full sample estimator; CC: the complete case estimator; PMLE_t: pseudo MLE under the true model; PMLE_w: pseudo MLE under model misspecification; SMLE: Semiparametric MLE.

	$\hat{\beta}_0(\widehat{\text{Var}})$	$\hat{\beta}_1(\widehat{\text{Var}})$	$\hat{\sigma}^2(\widehat{\text{Var}})$
FULL	**0.95** (0.0046)	**0.59** (0.0010)	**1.18** (0.0053)
	MCAR, response rate=70%		
CC	0.93 (0.0065)	0.59 (0.0014)	1.15 (0.0071)
PMLN	0.93 (0.0061)	0.59 (0.0012)	1.14 (0.0067)
SMLE	0.92 (0.0056)	0.59 (0.0011)	1.12 (0.0066)
	MCAR, response rate=50%		
CC	0.91 (0.0088)	0.60 (0.0021)	1.13 (0.0102)
PMLN	0.89 (0.0078)	0.61 (0.0017)	1.14 (0.0096)
SMLE	0.88 (0.0070)	0.61 (0.0016)	1.11 (0.0099)
	MAR, response rate=70%		
CC	1.44 (0.0076)	0.48 (0.0013)	1.01 (0.0058)
PMLN	1.45 (0.0074)	0.48 (0.0013)	1.00 (0.0057)
SMLE	0.97 (0.0054)	0.58 (0.0014)	1.18 (0.0093)
	MAR, response rate=50%		
CC	1.69 (0.0115)	0.44 (0.0017)	0.92 (0.0066)
PMLN	1.70 (0.0106)	0.44 (0.0016)	0.91 (0.0064)
SMLE	0.96 (0.0076)	0.58 (0.0018)	1.16 (0.0144)

parameter estimation, as our result shows.

In all scenarios, SMLE is close to FULL. Under MCAR, although CC is unbiased, the use of SMLE gains efficiency over CC. Under MAR, the use of SMLE corrects the bias of CC and PMLN. Again, the main advantage of SMLE is that it does not require specifying the marginal distribution of x and is thus robust, whereas the parametric pseudo maximum likelihood method is subject to severe bias under model misspecification.

7. Concluding Remarks

In this paper, a semiparametric maximum likelihood procedure is proposed to handle missing covariates in survey data. The proposed method does not require a parametric specification of the covariate distribution, and is thus robust compared with the fully parametric methods. We also provide an EM type of computation algorithm, which results in a fractionally imputed data set. Our simulation compares the proposed method (SMLE) with the complete case (CC) analysis, and the pseudo maximum likelihood method (PMLE) based on parametric models for the covariate distribution. The CC analysis tends to lose efficiency and introduce bias in estimation under MAR. The PMLE is efficient under the true model but can be severely biased under a wrong model. The proposed

SMLE produces valid inference with good efficiency in the simulation study. It is usually difficult, if not impossible, to specify a passable covariate distribution. Thus, the proposed semiparametric approach has promise for handling missing covariates in practice.

The proposed method is based on the MAR assumption. When MAR does not hold, method that brings in the exponential titling technique (Kim and Yu (2011)) can be developed accordingly. This will be investigated in the future.

Acknowledgement

We thank an anonymous referee, the AE, and the Editor for their constructive comments which have greatly improved the quality of the paper. The research of the second author is partially supported by a grant from NSF (MMS-121339) and by a Cooperative Agreement between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

Appendix

Appendix A. Proof of Lemma 1

The positivity condition (C1) and condition (C3) ensure that $\{h(\theta, P^X; Z) : \theta \in \Theta, ||P^X - P_0^X|| < \epsilon_0\}$ is P^Z Glivenko-Cantelli, and therefore $\sup_{\theta, ||P^X - P_0^X|| < \epsilon_0} |P_N^Z h(\theta, P^X; Z) - E\{h(\theta, P^X; Z)\}| \to 0$, as $N \to \infty$. Moreover, by the design consistency condition in (C2), we have

$$|P_n^Z h(\theta, P^X; Z) - P_N^Z h(\theta, P^X; Z)| \to \infty$$

for $\theta \in \Theta$ and $||P^X - P_0^X|| < \epsilon_0$. Together, we have

$$\sup_{\theta, ||P^X - P_0^X|| < \epsilon_0} |P_n^Z h(\theta, P^X; z) - E\{h(\theta, P^X; z)\}| \to 0, \tag{A.1}$$

as $n \to \infty$. Condition (C4) implies that

$$E\{h(\theta, P_n^X; z) - h(\theta, P_0^X; z)\} \to 0, \tag{A.2}$$

for any P_n^X such that $||P_n^X - P_0^X|| \to 0$ as $n \to \infty$. Note that

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \sum_{i=1}^n w_i h(\theta, P_n^X; z_i) \\ &= \arg\max_\theta \left\{ \sum_{i=1}^n w_i h(\theta, P_n^X; z_i) - Eh(\theta, P^X; z)|_{P^X = P_n^X} \right\} + Eh(\theta, P^X; z)|_{P^X = P_n^X}\end{aligned}$$

$$= o_p(1) + \arg\max_{\theta} \left\{ Eh(\theta, P^X; z)|_{P^X = P_n^X} - Eh(\theta, P_0^X; z) \right\} + Eh(\theta, P_0^X; z)$$
$$= o_p(1) + \arg\max_{\theta} Eh(\theta, P_0^X; z)$$
$$= o_p(1) + \theta_0,$$

where the third and fourth equalities follow from (A.1) and (A.2), respectively. Therefore, $\hat{\theta}$ converges to θ_0 in probability as $n \to \infty$.

Appendix B. The V-statistic theory for Possion sampling

We first establish the V-statistic theory for Poisson sampling. In Poisson sampling, the sampling indicator is independently generated with a known sampling probability, which preserves the i.i.d. structure of the observations $\{I_i, Z_i \equiv (\delta_i, \delta_i X_i, Y_i)\}$. The V-statistic is

$$V_N(v) = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j I_i I_j v(Z_i, Z_j),$$

where the function, $v : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, is a measurable symmetric function. Then $w_i w_j I_i I_j v(Z_i, Z_j)$ is the symmetric kernel of the V-statistic. We use Hoeffding decomposition (Hoeffding (1948)), defining functions $v_1(I_i, Z_i)$ and $v_2(I_i, Z_i, I_j, Z_j)$ as

$$\begin{aligned} v_1(I_i, Z_i) &= E\{w_i w_j I_i I_j v(Z_i, Z_j) \mid (I_i, Z_i)\} - v_0 \\ &= w_i I_i E\{v(Z_i, Z_j) \mid (I_i, Z_i)\} - v_0, \end{aligned} \tag{B.1}$$

$$v_2(I_i, Z_i, I_j, Z_j) = w_i w_j I_i I_j v(Z_i, Z_j) - v_1(I_i, Z_i) - v_1(I_j, Z_j) - v_0, \tag{B.2}$$

where $v_0 = E\{w_i w_j I_i I_j v(Z_i, Z_j)\} = E\{v(Z_i, Z_j)\}$. Therefore, we have

$$w_i w_j I_i I_j v(Z_i, Z_j) = v_0 + v_1(I_i, Z_i) + v_1(I_j, Z_j) + v_2(I_i, Z_i, I_j, Z_j). \tag{B.3}$$

Here v_1 and v_2 satisfy $Ev_1(I_i, Z_i) = 0$, and $E\{v_2(I_i, Z_i, I_j, Z_j)|(I_i, Z_i)\} = E\{v_2(I_i, Z_i, I_j, Z_j)|(I_j, Z_j)\} = 0$, $\forall (I_i, Z_i), (I_j, Z_j)$. Then $v_2(I_i, Z_i, I_j, Z_j)$ is called a degenerated kernel. From (B.3), by some algebra, we obtain the Hoeffding decomposition of the V-statistic

$$V_N(v) = v_0 + \frac{2}{N} \sum_{i=1}^{N} v_1(I_i, Z_i) + V_N(v_2).$$

In this way, we decompose $V_N(v)$ into a sum of a constant term v_0, an average of $v_1(I_i, Z_i)$, and a V-statistic with a degenerate kernel v_2. The terms $v_1(I_i, Z_i)$, $1 \le i \le N$ and $v_2(I_i, Z_i, I_j, Z_j)$, $1 \le i < j \le N$ are all mutually uncorrelated. Thus,

$$\operatorname{var}\Big\{\frac{2}{N}\sum_{i=1}^{N} v_1(I_i, Z_i)\Big\} = \frac{4}{N}\operatorname{var}\{v_1(I_1, Z_1)\},$$

$$\operatorname{var}\{V_N(v_2)\} = \frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N} E\{v_2(I_i, Z_i, I_j, Z_j)^2\}$$
$$\leq \frac{1}{N^2}\max\left[\operatorname{var}\{v_2(I_1, Z_1, I_1, Z_1)\}, \operatorname{var}\{v_2(I_1, Z_1, I_2, Z_2)\}\right].$$

It follows that $V_N(v_2) = O_p(N^{-1})$ and we can obtain the V-statistic central limit theorem from the Central Limit Theorem for partial sums of i.i.d. random variables, $\sqrt{N}\{V_N(v) - v_0\}$ has an asymptotically normal distribution.

Appendix C. Proof of Theorem 1

Let

$$h(\theta, P^X; z) = \delta \log f(y|x;\theta) + (1-\delta)\log P^X f(y|X;\theta),$$

where $z = (\delta, \delta x, y)$. The corresponding score function is

$$S(\theta, P^X; z) = \frac{\partial}{\partial\theta} h(\theta, P^X; z) = \delta\frac{f_\theta(y|x;\theta)}{f(y|x;\theta)} + (1-\delta)\frac{P^X f_\theta(y|X;\theta)}{P^X f(y|X;\theta)},$$

where $f_\theta(y|x;\theta) = \partial f(y|x;\theta)/\partial\theta$. The SMLE $\hat\theta$ solves $P_n^Z S(\theta, P_n^X; z) = 0$, where

$$P_n^Z S(\theta, P^X; z) = N^{-1}\Big\{\sum_{i=1}^{n} w_i\delta_i\frac{f_\theta(y_i|x_i;\theta)}{f(y_i|x_i;\theta)} + \sum_{i=1}^{n} w_i(1-\delta_i)\frac{P^X f_\theta(y_i|X;\theta)}{P^X f(y_i|X;\theta)}\Big\}.$$

By Lemma 1, $\hat\theta - \theta_0 \to 0$ in probability as $n \to \infty$, so we can apply a Taylor expansion method to get

$$0 = P_n^Z S(\hat\theta, P_n^X) = P_n^Z S(\theta_0, P_n^X, z) + P_n^Z S_\theta(\theta_0, P_n^X, z)(\hat\theta - \theta_0) + o_p(\hat\theta - \theta_0),$$

where $S_\theta = \partial S/\partial\theta^T$. Then

$$\begin{aligned}\hat\theta - \theta_0 &= E(-S_\theta)^{-1} P_n^Z S(\theta_0, P_n^X) + o_p(\hat\theta - \theta_0)\\ &= E(-S_\theta)^{-1}\left[P_n^Z S(\theta_0, P_0^X) + P_n^Z\left\{S(\theta_0, P_n^X) - S(\theta_0, P_0^X)\right\}\right] \quad \text{(C.1)}\\ &\quad + o_p(\hat\theta - \theta_0).\end{aligned}$$

The quantity $P_n^Z\left\{S(\theta_0, P_n^X) - S(\theta_0, P_0^X)\right\}$ quantifies the discrepancy between P_n^X and the true distribution P_0^X in the score function. To calculate this term, let δ_x be the Dirac function with point mass one at x. By a Taylor expansion and the von Mises calculus (Fernholz (1983)), we have

$$S(\theta_0, P_n^X) - S(\theta_0, P_0^X)$$

$$
\begin{aligned}
&= S_P(\theta_0, P_0^X)(P_n^X - P_0^X) + o_p(n^{-1/2}) \\
&= S_P(\theta_0, P_0^X)(\sum_{i=1}^{n} \pi_i \delta_{x_i} - P_0^X) + o_p(n^{-1/2}) \\
&= \sum_{i=1}^{n} \frac{dS(\theta_0, (1-t)P_0^X + t\pi_i\delta_{x_i})}{dt}|_{t=0} + o_p(n^{-1/2}), \qquad \text{(C.2)}
\end{aligned}
$$

where $S_P(\theta_0, P_0^X) = \partial S(\theta, P^X)/\partial P^X|_{(\theta=\theta_0, P^X=P_0^X)}$. Since

$$
\begin{aligned}
&S(\theta, (1-t)P_0^X + t\pi_i\delta_{x_i}) \\
&= \delta \frac{f_\theta(y|x;\theta)}{f(y|x;\theta)} + (1-\delta)\frac{\{(1-t)P_0^X + t\pi_i\delta_{x_i}\} f_\theta(y|x;\theta)}{\{(1-t)P_0^X + t\pi_i\delta_{x_i}\} f(y|x;\theta)},
\end{aligned}
$$

we have

$$
\begin{aligned}
&\frac{dS\left(\theta_0, (1-t)P_0^X + t\pi_i\delta_{x_i}\right)}{dt}|_{t=0} \\
&= (1-\delta)\frac{(\pi_i\delta_{x_i} - P_0^X) f_\theta(y|x;\theta_0)}{P_0^X f(y|x;\theta_0)} - (1-\delta)\frac{(\pi_i\delta_{x_i} - P_0^X) f(y|x;\theta_0) P_0^X f_\theta(y|x;\theta_0)}{\{P_0^X f(y|x;\theta_0)\}^2}.
\end{aligned}
$$

Thus, (C.2) becomes

$$
(1-\delta)\left[\frac{(P_n^X - P_0^X) f_\theta(y|x;\theta_0)}{P_0^X f(y|x;\theta_0)} - \frac{(P_n^X - P_0^X) f(y|x;\theta_0) P_0^X f_\theta(y|x;\theta_0)}{\{P_0^X f(y|x;\theta_0)\}^2}\right] + o_p(n^{-1/2}),
$$

and

$$
\begin{aligned}
&P_n^Z\{S(\theta_0, P_n^X) - S(\theta_0, P_0^X)\} \\
&= N^{-1}\sum_{i=1}^{N} w_i I_i (1-\delta_i)\left[\frac{(P_n^X - P_0^X) f_\theta(y_i|x;\theta_0)}{P_0^X f(y_i|x;\theta_0)}\right. \\
&\quad \left. - \frac{(P_n^X - P_0^X) f(y_i|x;\theta_0) P_0^X f_\theta(y_i|x;\theta_0)}{\{P_0^X f(y_i|x;\theta_0)\}^2}\right] \\
&= N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j I_i I_j (1-\delta_i)\delta_j \frac{1}{w_j}\left[\frac{f_\theta(y_i|x_j;\theta_0)\pi_j}{P_0^X f(y_i|x;\theta_0)}\right. \\
&\quad \left. - \frac{f(y_i|x_j;\theta_0)\pi_j P_0^X f_\theta(y_i|x;\theta_0)}{\{P_0^X f(y_i|x;\theta_0)\}^2}\right] + o_p(n^{-1/2}) \\
&= N^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j v(z_i, z_j) + o_p(n^{-1/2}) \\
&= N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j I_i I_j v(z_i, z_j) + o_p(n^{-1/2})
\end{aligned}
$$

$$\equiv V_N + o_p(n^{-1/2}),$$

where the second equality follows from evaluating

$$P_n^X f(y_i|x;\theta_0) = \sum_{j=1}^{n} \delta_j f(y_i|x_j;\theta_0)\pi_j$$

and

$$P_n^X f_\theta(y_i|x;\theta_0) = \sum_{j=1}^{n} \delta_j f_\theta(y_i|x_j;\theta_0)\pi_j.$$

Thus,

$$V_N = N^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j I_i I_j v(z_i, z_j)$$

is a V-statistics with the kernel function $w_i w_j I_i I_j v(z_i, z_j)$, where

$$v(z_i, z_j) = \frac{1}{2}\Big\{\frac{(1-\delta_i)\delta_j f_\theta(y_i|x_j;\theta_0)\pi_j}{w_j P_0^X f(y_i|x;\theta_0)} - \frac{(1-\delta_i)\delta_j f(y_i|x_j;\theta_0)\pi_j P_0^X f_\theta(y_i|x;\theta_0)}{w_j\{P_0^X f(y_i|x;\theta_0)\}^2}$$
$$+\frac{(1-\delta_j)\delta_i f_\theta(y_j|x_i;\theta_0)\pi_i}{w_i P_0^X f(y_j|x;\theta_0)} - \frac{(1-\delta_j)\delta_i f(y_j|x_i;\theta_0)\pi_i P_0^X f_\theta(y_j|x;\theta_0)}{w_i\{P_0^X f(y_j|x;\theta_0)\}^2}\Big\}.$$

Let

$$\begin{aligned} v_1(z_i;\theta_0,P^X) &= E\{v(z_i,z_j)|z_i\} \\ &= \frac{1}{2}P(\delta_j=1)E\Bigg[E\Bigg\{\frac{f_\theta(y_i|x_j;\theta_0)\pi_j}{w_j P_0^X f(y_i|x;\theta_0)} \\ &\quad -\frac{f(y_i|x_j;\theta_0)\pi_j P_0^X f_\theta(y_i|x;\theta_0)}{w_j\{P_0^X f(y_i|x;\theta_0)\}^2}|\delta_j=1\Bigg\}|\delta_i=0,y_i\Bigg] \\ &\quad +\frac{1}{2w_i}P(\delta_j=0)E\Bigg[E\Bigg\{\frac{f_\theta(y_j|x_i;\theta_0)\pi_i}{P_0^X f(y_j|x;\theta_0)} \\ &\quad -\frac{f(y_j|x_i;\theta_0)\pi_i P_0^X f_\theta(y_j|x;\theta_0)}{\{P_0^X f(y_j|x;\theta_0)\}^2}|\delta_j=0\Bigg\}|\delta_i=1,x_i\Bigg]. \end{aligned}$$

From the theory of V-statistics, we have

$$V_n = N^{-1}\sum_{i=1}^{N} w_i I_i\{2v_1(z_i;\theta_0,P_0^X)\} + o_p(n^{-1/2}). \tag{C.3}$$

Combining (C.1) and (C.3), we have

$$\hat\theta - \theta_0 = N^{-1}\sum_{i=1}^{N} w_i I_i \kappa(z_i;\theta_0,P_0^X) + o_p(n^{-1/2}),$$

where

$$\kappa(z_i;\theta_0,P_0^X)=E(-S_\theta)^{-1}\left\{S(\theta_0,P_0^X;z_i)+2v_1(z_i;\theta_0,P_0^X)\right\}. \tag{C.4}$$

Therefore, $\Sigma^{-1/2}(\hat{\theta}-\theta_0)\xrightarrow{d}N(0,I_d)$, where $\Sigma=Var\{\sum_{i=1}^n w_i\kappa(Z_i;\theta_0,P_0^X)\}$.

Appendix D. The Jackknife variance estimator

The Jackknife variance estimator provides a useful tool to calculate variances under complex sampling designs. The goal is to replicate the design in a series of subsamples that reflect the overall sample. Each of these subsamples retains the features of the original design.

To implement the Jackknife variance estimation, first consider $w_i^{[k]}$ as the kth replication weight such that

$$\hat{V}_{rep}=\sum_{k=1}^{L}c_k(\hat{Y}^{[k]}-\hat{Y})^2$$

is consistent for the variance of $\hat{Y}=\sum_{i\in A}w_iy_i$, where L is the replication size, c_k is the kth replication factor depending on the replication method and the sampling mechanism, and $\hat{Y}^{[k]}=\sum_{i\in A}w_i^{[k]}y_i$ is the kth replicate of $\hat{Y}$. In delete-one Jackknife variance estimation, $L=n$ and $c_k=(n-1)/n$.

For the replication method, we first obtain the kth replicate SMLE $\hat{\theta}^{[k]}$ of $\hat{\theta}$ by solving

$$\sum_{i=1}^{r}w_i^{[k]}S(\theta;x_i,y_i)+\sum_{i=r+1}^{n}w_i^{[k]}\left\{\frac{\sum_{j=1}^{r}\pi_j f(y_i\mid x_j;\theta)S(\theta;x_j,y_i)}{\sum_{j=1}^{r}\pi_j f(y_i\mid x_j;\theta)}\right\}=0,$$

$$\pi_j=\frac{w_j^{[k]}+\sum_{i=r+1}^{n}w_i^{[k]}w_{ij}^*(\theta)}{\sum_{i=1}^{n}w_i^{[k]}},$$

where $w_{ij}^*(\theta)=\pi_j f(y_i|x_j;\theta)/\sum_{k=1}^{r}\pi_k f(y_i|x_k;\theta)$ and $w_i^{[k]}$ is the replication weight. The replication variance estimator of $\hat{\theta}$ is obtained as

$$\hat{V}_{rep}(\hat{\theta})=\sum_{k=1}^{L}c_k(\hat{\theta}^{[k]}-\hat{\theta})^2.$$

Appendix E. Illustration with two-stage sampling design

Under a two-stage sampling design, let A_I be the index set of the primary sampling units (PSUs) in the sample. Let A_i be the index set of units selected in PSU $i\in A_I$. The final sample of units is indexed by $A=\cup_{i\in A_I}A_i$. Let π_{Ii} be the selection probability of the PSU i and $\pi_{k|i}$ be the conditional selection

probability of unit k given that PSU i is selected in the first stage where $k \in$PSU i. Thus, the first order inclusion probability of unit $k \in$PSU i is $P(k \in A) = P(k \in A_i | i \in A_I) P(i \in A_I) = \pi_{k|i}\pi_{Ii}$ and the sampling weight for this unit is $w_{ik} = 1/(\pi_{k|i}\pi_{Ii})$. Let (x_{ik}, y_{ik}) be the covariate and the outcome variable for unit $k \in A_i$.

The point estimation of θ can be obtained by solving the imputed score equation

$$\sum_{i \in A_I} \left[\sum_{k \in R_i} w_{ik} S(\theta; x_{ik}, y_{ik}) + \sum_{k \in M_i} w_{ik} \left\{ \frac{\sum_{j \in A_I} \sum_{l \in R_j} \pi_{jl} f(y_{ik} \mid x_{jl}; \theta) S(\theta; x_{jl}, y_{ik})}{\sum_{j' \in A_I} \sum_{l' \in R_{j'}} \pi_{j'l'} f(y_{ik} \mid x_{j'l'}; \theta)} \right\} \right] = 0,$$

$$\pi_{jl} = \frac{w_{jl} + \sum_{i \in A_I} \sum_{k \in M_i} w_{ik} w^*_{ik,jl}(\theta)}{\sum_{i \in A_I} \sum_{k \in A_i} w_{ik}},$$

where R_i and M_i are the index sets for the respondents and the nonrespondents in PSU i, i.e., $A_i = R_i \cup M_i$ and $w^*_{ik,jl}(\theta) = \pi_{jl} f(y_{ik} | x_{jl}; \theta) / \{\sum_{j' \in A_I} \sum_{l' \in R_{j'}} \pi_{j'l'} f(y_{ik} \mid x_{j'l'}; \theta)\}$. The EM algorithm described in Section 4 can be implemented to obtain the solution.

For variance estimation, we consider the Jackknife variance estimation considered in Rao, Wu and Yue (1992) for multi-stage sampling:

$$\hat{V}_{rep}(\hat{\theta}) = \sum_{i \in A_I} \frac{n_i - 1}{n_i} \sum_{k \in A_i} (\hat{\theta}^{[k]} - \hat{\theta})^2,$$

where $n_i = |A_i|$ is the sample size in PSU i, $\hat{\theta}^{[k]}$ is computed by omitting unit $k \in A_i$ and by modifying the weights so that $\pi^{-1}_{j|i}$ is replaced by $n_i \pi^{-1}_{j|i}/(n_i - 1)$ for all $j \in A_i$ and weight stays unaltered for all other $j \notin A_i$.

References

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.* **34**, 86-102.

Carroll, R. J., Knickerbocker, R. K. and Wang, C. Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *Ann. Statist.* **23**, 161-181.

Chambers, M. (2003). The asymptotic efficiency of cointegration estimators under temporal aggregation. *Econom. Theory* **19**, 49-77.

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of Survey Data.* Wiley, New York.

Chambers, R. L., Steel, D. G., Wang, S. and Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys.* Chapman & Hall / CRC, Boca Raton, FL.

Didelez, V. (2002). Ml and semiparametric estimation in logistic models with incomplete covariate data. *Statist. Neerlandica* **56**, 330-345.

Fernholz, L. T. (1983). *von Mises Calculus For Statistical Functionals.* Springer Verlag.

Fuller, W. A. (1998). Replication variance estimation for two-phase samples. *Statist. Sinica* **8**, 1153-1164.

Fuller, W. A. (2009). *Sampling Statistics.* Wiley, Hoboken.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.* **54**, 127-138.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293-325.

Horton, N. and Laird N. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statist. Methods Medical Res.* **8**, 37-50.

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrika* **55**, 591-596.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **33**, 290-299.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98**, 119-132.

Kim, J. K. and Skinner, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika* **100**, 385-398.

Kim, J. K. and Yu, C. L. (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data. *J. Amer. Statist. Assoc.* **106**, 157-165.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys.* Wiley, New York.

Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-select data problems in regression. *J. Roy. Statist. Soc. Ser. B* **61**, 413-438.

Little, R. J. A. (1992). Regression with missing x's: a review. *J. Amer. Statist. Assoc.* **87**, 1227-1237.

McLachlan, G. and Krishnan, T. (2007). *The EM Algorithm and Extensions.* John Wiley & Sons.

Moore, C. J., Lipstiz, S. R., Addy, C. L., Hussey, J. R., Fitzmaurice, G. and Natarjan, S. (2009). Logistic regression with incomplete covariate data in complex survey sampling. *Epidemiology* **20**, 382-390.

Paik, M. C. (2000). Methods for missing covariates in logistic regression. *Communications in Statistics - Simulation and Computation* **20**, 1-19.

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

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. *Handbook of Statistics; Sample Surveys: Inference and Analysis* **29**.

Rao, J. N. K., Wu. C. F. J. and Yue. K. (1992). Some recent work on resampling methods for complex surveys. *Surv. Methodol.* **18**, 209-217.

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

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Scott, A. and Wild, C. (2001). The analysis of clustered case-control studies. *J. Roy. Statist. Soc. Ser. C* **50**, 389-401.

Scott, A. and Wild, C. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *J. Roy. Statist. Soc. Ser. B* **64**, 207-220.

Scott, A. and Wild, C. (2011). Fitting regression models with response-biased samples. *Canad. J. Statist.* **39**, 519-536.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1996). Regression analysis for complex survey data with missing values of a covariate. *J. Roy. Statist. Soc. Ser. A* **159**, 265-274.

Van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge university press.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18**, 309-348.

Wang, C. and Paik, M. C. (2006). Efficiencies of methods dealing with missing covariates in regression analysis. *Statist. Sinica* **16**, 1169-1192.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

Yoshihara, K.-I. (1976). Limiting behavior of u-statistics for stationary, absolutely regular processes. *Probab. Theory Related Fields* **35**, 237-252.

Zhang, Z. and H. Rockette (2005). On maximum likelihood estimation in parametric regression with missing covariates. *J. Statist. Plann. Inference* **134**, 206-223.

Zhao, L. P., Lipsitz, S. and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165-82.

Department of Statistics, North Carolina State University, Raleigh, NC 27606, USA.

E-mail: syang24@ncsu.edu

Department of Statistics, Iowa State University, Ames, IA, 50010, USA.

E-mail: jkim@iastate.edu

(Received May 2014; accepted December 2015)