Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: Adapting to a changing world

Modeling Self-enumeration and Follow-up Response Indicators as Discrete-time Survival

A. Demnati[1]

Abstract

Collecting information from sampled units over the Internet or by mail is much more cost-efficient than conducting interviews. These methods make self-enumeration an attractive data-collection method for surveys and censuses. Despite the benefits associated with self-enumeration data collection, in particular Internet-based data collection, self-enumeration can produce low response rates compared with interviews. To increase response rates, nonrespondents are subject to a mixed mode of follow-up treatments, which influence the resulting probability of response, to encourage them to participate. Factors and interactions are commonly used in regression analyses, and have important implications for the interpretation of statistical models. Because response occurrence is intrinsically conditional, we first record response occurrence in discrete intervals, and we characterize the probability of response by a discrete time hazard. This approach facilitates examining when a response is most likely to occur and how the probability of responding varies over time. The nonresponse bias can be avoided by multiplying the sampling weight of respondents by the inverse of an estimate of the response probability. Estimators for model parameters as well as for finite population parameters are given. Simulation results on the performance of the proposed estimators are also presented.

Key Words: Event history analysis; Longitudinal data; Maximum likelihood; Mixed-mode surveys; Partially classified units.

1. Introduction

Mixing modes of follow-up and data collection offers the possibility of offsetting the disadvantages of one mode with the advantages of another. For example, recognizing that the Internet, unlike mail, offers the ability to move data capture and editing closer to the respondent, many statistical agencies are now offering electronic questionnaires as a voluntary option to both improve quality of statistical processes and reduce survey costs. This potential increase in survey quality in combination with the fact that collecting information from sampled units over the Internet or by mail is much more cost-effective than conducting interviews makes self-enumeration an attractive data-collection method for surveys and censuses. Although there are benefits associated with self-enumeration, in particular Internet-based surveys, as well as an expected wider application of this approach in future, self-enumeration brings particular difficulties to surveys and censuses. Observed values of typical variable of interest y might depend on the variable y_m associated with mode m of data collection, $m = 1,..., M$, where M is the number of modes of data collection under consideration for a given survey. In principle, each unit k of the population P of size N has all responses, i.e., a response $y_{m;k}$ that would have resulted if it had chosen mode m. Since each unit receives or chooses only one mode, only one response is observed. If the variable of interest is defined uniquely and independently from each mode, then $y_{m;k}$ represents the value the unit believes is the correct answer for y_k to y, resulting from the medium of mode m in which the question is presented to the unit. Suppose that the model mean of the response y_k is specified by $E_M(y_k) = \mu_k(\boldsymbol{\chi}_k^T \boldsymbol{\Theta})$, where $\boldsymbol{\chi}_k = (\chi_{1k},..., \chi_{pk})^T$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\Theta} = (\Theta_1,..., \Theta_p)^T$ is the $p \times 1$ vector of model parameter and E_M denotes model expectation. We assume that the finite population parameter associated with the vector model parameter $\boldsymbol{\Theta}$ is defined as solution to an estimating equation of the form

$$\boldsymbol{S}(\boldsymbol{\Theta}) = \sum_k \boldsymbol{s}(y_k; \boldsymbol{\Theta}) - \boldsymbol{v}(\boldsymbol{\Theta}) = \boldsymbol{0}, \tag{1.1}$$

[1] A. Demnati, Business Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6 (Abdellatif.Demnati@statcan.gc.ca).

where $\sum_k$ is the sum over all the finite population units, the known function $\boldsymbol{s}(y_k;\boldsymbol{\Theta})$ is a p-dimensional vector-valued function of y_k and the know function $\boldsymbol{v}(\boldsymbol{\Theta})$ allows for explicitly defined parameters. For linear and logistic regression models, $\boldsymbol{s}(y_k;\boldsymbol{\Theta}) = \boldsymbol{\chi}_k(y_k - \mu_k(\boldsymbol{\chi}_k^T\boldsymbol{\Theta}))$ and $\boldsymbol{v}(\boldsymbol{\Theta}) = \mathbf{0}$. For the special case of the finite population total $Y = \sum_k y_k$, $\boldsymbol{s}(y_k;\boldsymbol{\Theta}) = y_k$, $\boldsymbol{v}(\boldsymbol{\Theta}) = \Theta_N$ and $\Theta_N = Y$.

One of the main objectives of the mixed-mode of data collection is to influence the unit to get its cooperation, regardless of its preference for data-collection mode. The overall coverage rate for unit k for the combined modes can be defined as $r_k = 1 - \prod_{m=1}^{M}(1 - r_{m;k}) = \sum_{m=1}^{M} r_{m;k}$ and the overall probability of response can be represented in the mixture form $\xi_k = \sum_{m=1}^{M} \phi_m \xi_{m;k}$, where ϕ_m is the proportion of the population using mode m, $\xi_{m;k} = E_r(r_{m;k})$ is the probability of response associated with mode m of data collection and E_r denotes expectation with respect to the response mechanism. If the mixed-mode can increase the overall response rates, we will, of course, be pleased to quantify and examine the contribution of each mode of the response probability. In reality, self-enumeration can produce low response rates in comparison to interviews. To gain nonrespondents' cooperation and therefore maximize survey quality, nonrespondents are assigned to a mixed-mode of follow-up treatments. Different costs are associated with different treatments. For example, face-to-face follow-up is more expensive than telephone follow-up. Currently, in some business surveys, to reduce the total cost of data collection, follow-up for nonresponse is performed on only a portion of nonrespondents. These units are often identified in a deterministic way based, for example, on their expected contribution to the estimate. In addition, since a significant number of units are never followed up for nonresponse, the final response rate can be very low. The nonresponse bias can be avoided by multiplying the sampling weight of respondents by the inverse of the response probability. Since the response probability is unknown, estimated probability is used. As noted by Rosenbaum (1987) and others, estimators using the estimated response probability can be more efficient than estimators using the true response probability.

Given the above issues, one question should first be of particular interest to statistical agencies: How should both the response probabilities under mixed-mode and the influence of the follow-up treatments on the resulting probability of response be modelled? Other relevant questions include the following: If one factor of the mixed-mode is improved, what will be the effect on the performance of the response mechanism? How can we estimate the response probability due to a particular mixed-mode factor of interest with the presence of the other mixed-mode factor? As intensive follow-up is expensive, a follow-up strategy is needed to make optimal use of resources compared to the quality of the estimates. Since a follow-up treatment could produce estimates with better quality, the strategy should consist of allocating nonrespondents to different treatments while controlling for data collection costs. In an attempt to discuss some of these and other issues, and given the page limitation of this paper, the first part of our work below is organized as follows: in Section 2, the response probability is first characterized by a discrete-time hazard, then factors and interactions are considered through regression analysis; in Section 3, estimators of the regression (or nuisance) parameter as well as estimators of the parameter of interest under the previously mentioned setup are studied; and, in Section 4, simulation results are presented.

2. Formalizing a response model

2.1 Discrete-time hazard

Consider a homogeneous sample of units, each at risk of experiencing a single target event, − responding. The target event is nonrepeatable. To record response occurrence in discrete intervals, we divide continuous time into a sequence of continuous time periods: 1, 2, and so on. Suppose the duration of data collection is made up of I time periods. Let t represent the discrete random variable that indicates the time period i when the response occurs for a randomly selected unit from the sample. Then each unit k is observed until some period I_k, with $I_k \leq I$. Observation of the unit could be discontinued for two reasons: 1) the unit responds; or 2) the survey ends. In the first case, $t = I_k$. In the second case, it is only known that $t > I$. Units with $t > I$ are right-censored—it is unknown whether they respond. Because response occurrence is intrinsically conditional, we characterize t by its conditional probability density function − the distribution of the probability that a response will occur in each time period given that it has not already occurred in a previous time period − known as the discrete-time hazard function. Discrete-time

hazard, $h_{ki}(\boldsymbol{x}_k, \boldsymbol{\beta})$, h_{ki} for short, is defined as the conditional probability that unit k will respond in time period i, given that the unit did not respond prior to i:

$$h_{ki} = \Pr(t = i \mid t \geq i), \tag{2.1}$$

where $\boldsymbol{x}_k$ refers to both time-invariant and time-varying explanatory variables and $\boldsymbol{\beta}$ is the unknown vector parameter to be estimated. For unit with $t = i$, the probability of obtaining a response at time period i could be expressed in terms of the hazard as

$$\Pr(t = i) = h_{ki} \prod_{j=1}^{i-1}(1 - h_{kj}). \tag{2.2}$$

For units with $t > i$, the probability of obtaining a response can be expressed as

$$\Pr(t > i) = \prod_{j=1}^{i}(1 - h_{kj}). \tag{2.3}$$

We assume that every unit in the sample lives through each successive discrete time period until the unit responds or is censored by the end of data collection. The use of mixed-mode modifies the expression for the hazard function in (2.1) as $h_{ki} = \sum_{m=1}^{M} \phi_m h_{ki|m}$, where $h_{ki|m}$ is the discrete-hazard function for mode m. The marginal probability of obtaining a response after I time periods is given by

$$\xi_k = 1 - \prod_{i=1}^{I}(1 - h_{ki}) = \sum_{i=1}^{I} \xi_k^{(i)}. \tag{2.4}$$

where $\xi_k^{(i)} = h_{ki} \prod_{j=1}^{i-1}(1 - h_{kj})$. It is easily seen from (2.4) that ξ_k increases (or stays the same) as the level of effort increases, where the level of effort is seen in terms of follow-up treatments and data-collection periods. This suggests that costs and benefits of increasing the level of effort should be explored given that, in some circumstances, there are a number of follow up treatments made with a high percentage of cost expanded to get values from a few nonrespondents.

2.2 Influence of follow-up on response probability

We now express the inverse link function of the hazard rate as a function of explanatory variables $\boldsymbol{x}_{ki}$ and a vector parameter $\boldsymbol{\beta}$ to be estimated. For units under self-enumeration data collection, the inverse link form of the hazard-rate is expressed as

$$g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}^{(0)}, \boldsymbol{\beta}^{(0)}), \tag{2.5}$$

for known function $\eta(.)$, where $\boldsymbol{x}_{ki}^{(0)}$ is the vector of explanatory variables for self-enumeration, $\boldsymbol{\beta}^{(0)}$ is the associated unknown vector parameter to be estimated, $\boldsymbol{x}_{ki} = \boldsymbol{x}_{ki}^{(0)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ and $g(.)$ is a link function—although the link function is generally used to transform (or to link) the conditional mean to the linear predictor $\boldsymbol{x}_{ki}^T\boldsymbol{\beta}$. For example, $g(a) = a$ with $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^T\boldsymbol{\beta}$ gives a linear regression model and $g(a) = \exp(a)/\{1 + \exp(a)\}$ with $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^T\boldsymbol{\beta}$ gives a logistic regression model for binary responses r_{ki}, where r_{ki} is a sequence of response indicators defined for each unit k whose values are defined as $r_{ki} = 1$ if the unit does respond in period i and $r_{ki} = 0$ if the unit does not respond in period i.

Additional influences on response probability can be investigated by adding further predictors to the initial discrete-time hazard model. For instance, the following model differs from the model in (2.5) by the inclusion of the time-variant predictor follow-up $\gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)}$, the influence of which is captured by the parameter $\boldsymbol{\beta}^{(1)}$:

$$g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}^{(0)}, \boldsymbol{\beta}^{(0)}; \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)}, \boldsymbol{\beta}^{(1)}), \tag{2.6}$$

where the value of $\gamma_{ki}^{((1)}$ is set to 1 if the first follow-up treatment is started, or set to 0 if this is not the case, with $\boldsymbol{x}_{ki} = (\boldsymbol{x}_{ki}^{(0)T}, \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)T})^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\beta}^{(1)T})^T$. Note that (2.6) can be used to define different slopes and intercepts, in which case the parameter $\boldsymbol{\beta}^{(1)}$ reflects the changes in the intercepts and in the slopes associated with changing from self-enumeration only to self-enumeration followed by the first follow-up treatment. For example, in the specification $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^{(0)T}\boldsymbol{\beta}^{(0)} + \boldsymbol{x}_{ki}^{(1)T}\boldsymbol{\beta}^{(1)}$, $i = 1,\ldots,I$, with $\boldsymbol{x}_{ki}^{(0)T}\boldsymbol{\beta}^{(0)} = \alpha_{0i}^{(0)} + x_{ki}\alpha_{1i}^{(0)}$ and $\boldsymbol{x}_{ki}^{(1)T}\boldsymbol{\beta}^{(1)} = \gamma_{ki}^{(1)}(\alpha_{0i}^{(1)} + x_{ki}\alpha_{1i}^{(1)})$, the regression parameters $\alpha_{0i}^{(0)}$, $\alpha_{1i}^{(0)}$ and the values x_{ki} represent respectively the intercept, the slope and the predictor associated with self-enumeration data collection in time period i. We have $\boldsymbol{x}_{ki}^{(0)} = (D_{k1}^{(0)},..., D_{kI}^{(0)}, x_{k1}^{(0)},..., x_{kI}^{(0)})^T$ and $\boldsymbol{\beta}^{(0)} = (\alpha_{01}^{(0)},..., \alpha_{0I}^{(0)}, \alpha_{11}^{(0)},..., \alpha_{1I}^{(0)})^T$, where $D_{ki}^{(0)} = 1$, $x_{ki}^{(0)} = x_{ki}$, $D_{kj}^{(0)} = 0$ and $x_{kj}^{(0)} = 0$ for $j \neq i$. The vector predictor follow-up is given by $\boldsymbol{x}_{ki}^{(1)} = (D_{k1}^{(1)},..., D_{kI}^{(1)}, x_{k1}^{(1)},..., x_{kI}^{(1)})^T$ and the changes due to the follow-up in the intercepts and slopes are

reflected by vector parameter $\boldsymbol{\beta}^{(1)} = (\alpha_{01}^{(1)},...,\alpha_{0I}^{(1)},\alpha_{11}^{(1)},...,\alpha_{1I}^{(1)})^T$, where $D_{ki}^{(1)} = 1$, $x_{ki}^{(1)} = x_{ki}$, $D_{kj}^{(1)} = 0$ and $x_{kj}^{(1)} = 0$ for $j \neq i$. To increase response rates, nonrespondents are subject to intensive multiple follow-ups by telephone or other treatments to encourage them to participate. A treatment can take the form of mailed reminders, emailed reminders, telephone calls or in-person interviews. The follow up process through treatments is conducted using data collection calendars with a specific strategy for each sampled unit. In case of $1+J$ follow-up treatments, the inverse link form of the hazard-rate can be expressed as $g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta})$, where $\boldsymbol{x}_{ki} = (\boldsymbol{x}_{ki}^{(0)T}, \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)T},..., \gamma_{ki}^{(J)}\boldsymbol{x}_{ki}^{(J)T})^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\beta}^{(1)T},..., \boldsymbol{\beta}^{(J)T})^T$.

Consider the case of $J = 1$ where T_1 consists of intensive follow up and T_0 consists of sending the questionnaire, and suppose for simplicity the case in which the response outcome is instant. After collecting the response from self-enumeration respondents, follow-up is performed in a deterministic way—nonrespondents with $u_k \geq \kappa$ are assigned to treatment T_1, where κ is a predetermined constant and u is an auxiliary variable with values available for all sampled units. Suppose all units under T_1 responded, while the other units have still not responded. We have $\xi_k = h_{k1} + (1-h_{k1})1 = 1$ for unit k with $u_k \geq \kappa$ and $\xi_k = h_{k1} + (1-h_{k1})0 = h_{k1}$ for units with $u_k < \kappa$. This highlights the significant effect of follow-up on the probabilities of response.

3. Estimation

3.1 Likelihood function and the EM algorithm

The mode of data collection is known for a unit that responds at any time of data collection, while it is unknown for a unit that is censored. Let's define a vector of indicator variables as $\ell_{m;k} = 1$ if unit k uses mode m, and $\ell_{m;k} = 0$ if not, where $\ell_k = (\ell_{1k},..., \ell_{Mk})^T$ are realizations of independent and identically distributed random variables according to a multinomial distribution, $Mult_M(1, \boldsymbol{\varphi})$, and $\boldsymbol{\varphi} = (\phi_1,..., \phi_M)^T$ is the vector of the population proportions with $\sum_{m=1}^{M} \phi_m = 1$. Consequently the likelihood for the complete data is given by

$$L_c(\boldsymbol{\Phi}) = \prod_k \prod_{m=1}^{M} \{\phi_m f_m(t_k)\}^{\ell_{m;k}}, \tag{3.1}$$

and the log-likelihood for the complete data is given by $l_c(\boldsymbol{\Phi}) = \sum_k \sum_{m=1}^{M} \ell_{m;k} \{\log \phi_m + \log f_m(t_k)\}$, where $\boldsymbol{\Phi} = (\phi_1,..., \phi_{M-1}, \boldsymbol{\beta}^T)^T$, $f_m(t_k)$ is $f(t_k)$ for mode m,

$$f(t_k) = \Pr(t = I_k)^{\delta_k} \Pr(t > I_k)^{1-\delta_k}, \tag{3.2}$$

$\delta_k = 1$ if unit k is uncensored and $\delta_k = 0$ if unit k is censored. When unit k is censored, either unit k will respond at some future time period $t > I$ or the unit will never respond. To derive maximum likelihood estimates, we use the Expectation Maximization (EM) algorithm introduced by Hartley (1958)—formalized and termed by Dempster et al. (1977)—which has become a major tool for finding maximum likelihood estimates in situations considered practically intractable such as missing data. Using some initial value for $\boldsymbol{\Phi}$, say $\boldsymbol{\Phi}^{(b)}$, the E-step of the EM algorithm requires the calculation of a function of $\boldsymbol{\Phi}$, $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(b)})$, such that $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(b)}) = E\{l_c(\boldsymbol{\Phi}) \mid \boldsymbol{I}^{(o)}, \boldsymbol{\Phi}^{(b)}\}$, where $\boldsymbol{\Phi}$ is the parameter of interest, $\boldsymbol{\Phi}^{(b)}$ is the value of $\boldsymbol{\Phi}$ in the previous iteration, and $\boldsymbol{I}^{(o)}$ is the observed data. Then, the M step of the EM algorithm intent to choose the value of $\boldsymbol{\Phi}$, say $\boldsymbol{\Phi}^{(b+1)}$, that maximizes $Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(b)})$. If we iterate the E-step and M-step until convergence, under regularity conditions, the algorithm converges to the maximum likelihood estimate. We have

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(b)}) = \sum_k \sum_{m=1}^{M} \tau_{m;k}^{(b)} \log \phi_m + \sum_k \sum_{m=1}^{M} \tau_{m;k}^{(b)} \log f_m(t_k), \tag{3.3}$$

that is, each indicator variable $\ell_{m;k}$ is replaced by $\tau_{m;k}^{(b)} = E(\ell_{m;k} \mid \boldsymbol{I}^{(o)}, \boldsymbol{\Phi}^{(b)})$, its expectation conditional on the observed data $\boldsymbol{I}_k^{(o)}$. The E step of the algorithm involves creating a set of "pseudo-data" in which the respondents are left intact and the nonrespondents are fractionated into M partially complete pseudo-observations. The weight assigned to this pseudo-observation is the conditional probability that the unit belongs to an associated population. The conditional expectation $\tau_{m;k}$ is the probability that unit k uses mode m given the observed data and prior estimation of the parameters. Once $\boldsymbol{\Phi}^{(b)}$ has been obtained, estimates of the conditional probabilities can be formed for each k. The conditional probability that a nonrespondent k uses mode m is given by

$\tau_{m;k}^{(b)} = \{\sum_n \phi_n^{(b)} f_n^{(b)}(t_k)\}^{-1} \phi_m^{(b)} f_m^{(b)}(t_k)$, where $f_m^{(b)}(t_k)$ is $f(t_k, \boldsymbol{\beta}^{(b)})$ for mode m. Estimation of $\boldsymbol{\Phi}$ and $\tau_{m;k}$ are alternated repeatedly, where in their subsequent execution, the initial fit $\boldsymbol{\Phi}^{(b)}$ is replaced by the current fit $\boldsymbol{\Phi}^{(b+1)}$ for $\boldsymbol{\Phi}$. From the first term of the right hand of (3.3), we get the subsequent estimate $\phi_m^{(b+1)}$ of ϕ_m using

$$S_c(\phi_m) = \sum_k \{\tau_{m;k}^{(b)} - \phi_m \sum_m \tau_{m;k}^{(b)}\} = 0 . \tag{3.4}$$

The last term of the right hand of (3.3) corresponds to the response-time distribution and involves only $\boldsymbol{\beta}$. Substituting (2.2) and (2.3) into (3.2), and taking the logarithm yields

$$\log f(t_k) = \delta_k \log\{h_{kl_k}/(1-h_{kl_k})\} + \sum_{i=1}^{I_k} \log(1-h_{ki}) . \tag{3.5}$$

Using the sequence of response indicators r_{ki}, expression (3.5) can be rewritten (Allison, 1982) as $\log f(t_k) = \sum_{i=1}^{I_k} r_{ki} \log\{h_{ki}/(1-h_{ki})\} + \sum_{i=1}^{I_k} \log(1-h_{ki})$. Taking the derivatives, we get

$$\partial \log f(t_k)/\partial \boldsymbol{\beta} = \sum_{i=1}^{I_k} \dot{\boldsymbol{h}}_{ki}(\boldsymbol{\beta})(r_{ki} - h_{ki})\{h_{ki}(1-h_{ki})\}^{-1} , \tag{3.6}$$

where $\dot{\boldsymbol{h}}_{ki}(\boldsymbol{\beta}) = \partial h_{ki}/\partial \boldsymbol{\beta}$. Under a census case, the subsequent estimate $\boldsymbol{\beta}^{(b+1)}$ of $\boldsymbol{\beta}$ is the value that satisfies

$$\boldsymbol{S}_c(\boldsymbol{\beta}) = \sum_k \sum_{m=1}^{M} \tau_{m;k}^{(b)} \partial \log f_m(t_k)/\partial \boldsymbol{\beta} = \boldsymbol{0} , \tag{3.7}$$

where $\partial \log f(t_k)/\partial \boldsymbol{\beta}$ is given by (3.6).

3.2 Estimation of the parameters from survey data

Suppose a probability sample s is selected from the finite population and let $d_k(s)$ denote the sampling weight attached to unit k with $d_k(s) = 0$ if $k \notin s$. We use the design weights to estimate the estimating equations of the EM algorithm. The weighted estimating equation associated with (3.4) and (3.7) is given by

$$\hat{\boldsymbol{S}}_c(\boldsymbol{\Phi}) = \sum_k d_k(s) \boldsymbol{s}_{c;k}(\boldsymbol{\Phi}) = \boldsymbol{0} ,$$

where $\boldsymbol{S}_c(\boldsymbol{\Phi}) = (\boldsymbol{S}_c^T(\phi_1), ..., \boldsymbol{S}_c^T(\phi_{M-1}), \boldsymbol{S}_c^T(\boldsymbol{\beta}))^T$. The parameters of interest must still be estimated. An estimator $\hat{\boldsymbol{\Theta}}$ of the parameter $\boldsymbol{\Theta}$ associated by (1.1) is the solution of the weighted estimating equation

$$\hat{\boldsymbol{S}}(\boldsymbol{\Theta}) = \sum_k d_k(s)(r_k/\hat{\xi}_k)\boldsymbol{s}(y_k;\boldsymbol{\Theta}) - \boldsymbol{v}(\boldsymbol{\Theta}) = \boldsymbol{0} ,$$

with $r_k = 1 - \prod_{i=1}^{I_k}(1-r_{ki}) = \sum_{i=1}^{I_k} r_{ki}$ and $\hat{\xi}_k = \xi_k(\hat{\boldsymbol{\beta}})$. Since respondents in one mode cannot be considered a random subsample of the whole population, an estimator $\hat{\boldsymbol{\Theta}}_m$ of the parameter of interest $\boldsymbol{\Theta}_m$ associated with mode m is the solution to the weighted estimating equation

$$\hat{\boldsymbol{S}}(\boldsymbol{\Theta}_m) = \sum_k d_k(s)(r_k/\hat{\xi}_k)(l_{m;k}/\hat{\phi}_{m;k})\boldsymbol{s}(y_{m;k};\boldsymbol{\Theta}_m) - \boldsymbol{v}(\boldsymbol{\Theta}_m) = \boldsymbol{0} .$$

4. Simulation study

We conducted a small simulation study to illustrate the performances of the estimators of the model parameters as well as estimators of the finite population totals related to the mixed-mode of data collection. We generate values for each unit k of a finite population of size $N = 2000$ independently from the model

$$\begin{pmatrix} y_k \\ u_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_u \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

where y is the variable of interest, u is the auxiliary variable, $\mu_y = \mu_u = 2$ and $\rho = .8$. We set $M = 2$ and we created the proportions of mail and Internet populations using $\ell_k = (\ell_{M;k}, \ell_{I;k})^T \sim Mult_2(1, \boldsymbol{\varphi})$ with $\boldsymbol{\varphi} = (\phi_M, \phi_I)^T = (.6, .4)^T$ and values of the variable of interest associated with each mode of data collection are generated from

$$y_{m;k} \mid y_k \sim N\left(\mu_m + \rho_m(y_k - \mu_y), (1-\rho_m^2)\right) \text{ for } m \in \{M, I\} ,$$

with $\mu_M = E_m(y_M) = 3$, $\mu_I = E_m(y_I) = 4$, $\rho_M = \rho(y_M, y) = .3$ and $\rho_I = \rho(y_I, y) = .7$. We maintained the population values $(y_k, u_k, y_{M;k}, y_{I;k}, \ell_{M;k}, \ell_{I;k})$ fix for $k = 1, ..., N$, and we selected $A = 200$ Bernoulli samples each of sampling fraction $f = .5$ from the generated population. We set $I = 7$ and $J = 2$. Nonrespondents at time period 2 with

$u_k \geq \mu_u$ are assigned to T_1, while nonrespondents at time period 4 are assigned to T_2. In this case, we have in total two strategies S_1 and S_2, where $S_1 = (0,0,0,T_2,0,0,0)$ and $S_2 = (0,T_1,0,T_2,0,0,0)$. We created respondents and nonrespondents using $r_{ki} \sim B(1,h_{ki})$, with $logit(h_{ki}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}$, $x_{ki}^T \boldsymbol{\beta} = F(T_0) + F(T_1) + F(T_2)$, $F(T_0) = (\ell_{M;k}\alpha_M^{(0)} + \ell_{I;k}\alpha_I^{(0)})t$, $F(T_1) = 1(u_k \geq \mu_u)1(t \geq 3)(\ell_{M;k}\alpha_M^{(1)} + \ell_{I;k}\alpha_I^{(1)})t_1$ and $F(T_2) = 1(t \geq 5)\alpha^{(2)}t_2$, where $1()$ is the truth function and t_j is the time period since treatment T_j has been started. Table 4-1 displays values of the response model parameters. The first component $\boldsymbol{\varphi}_N$ of the vector parameter $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_N^T, \boldsymbol{\beta}^T)^T$ corresponds to the mode of data collection and involves the finite population proportions $\boldsymbol{\varphi}_N = (\phi_{M;N}, \phi_{I;N})^T = N^{-1}\sum_k (\ell_{M;k}, \ell_{I;k})^T$, while the second component $\boldsymbol{\beta}$ of the vector parameter corresponds to the response-model and involves $\boldsymbol{\beta} = (\alpha_M^{(0)}, \alpha_I^{(0)}, \alpha_M^{(1)}, \alpha_I^{(1)}, \alpha^{(2)})^T$. We estimated $\boldsymbol{\Phi}$ using the EM algorithm. Let $\hat{\theta}$ denote an estimator of the parameter of interest θ. We calculated $\hat{\theta}$, from each repetition a ($a = 1,..., A$) and their averages $\bar{\hat{\theta}} = A^{-1}\sum_{a=1}^{A} \hat{\theta}_a$, where $\hat{\theta}_a$ is the value of $\hat{\theta}$ for the a^{th} sample. The simulated bias, relative bias and mean squared error (MSE) of $\hat{\theta}$ are calculated as $B(\hat{\theta}) = (\bar{\hat{\theta}} - \theta)$, $RB(\hat{\theta}) = B(\hat{\theta})/\theta$ and $MSE(\hat{\theta}) = A^{-1}\sum_{a=1}^{A}(\hat{\theta}_a - \theta)^2$ respectively. The overall response rate for the 200 repetitions varies from a minimum of .72 to a maximum of .79 with mean around .76. The minimum, maximum and mean of the EM iterations are 12, 17 and 15 respectively. All cases converged. We calculated $B(\hat{\theta})$ for regression parameters, and those values are reported in Table 4-1. Table 4-1 clearly demonstrates that the bias is small for each regression parameter. We also considered the estimation of the finite population total: $\boldsymbol{\Theta} = (\sum_k y_k, \sum_k y_{m;k}; m \in \{M, I\})^T$. We used two sets of weights: the first set of weights uses estimated model parameters, $(d_k(s)(r_k/\hat{\xi}_k), d_k(s)(r_k/\hat{\xi}_k)(l_{m;k}/\hat{\phi}_{m;k}); m \in \{M, I\})^T$; while the second set of weights uses the true model parameter, $(d_k(s)(r_k/\xi_k), d_k(s)(r_k/\xi_k)(l_{m;k}/\phi_{m;k}); m \in \{M, I\})^T$. We calculated $RB(\hat{\theta})$ and MSE ratios for each estimator $\hat{\theta}$ with $\hat{Y} = \sum_k d_k(s)(r_k/\hat{\xi}_k)y_k$ and those values are reported in Table 4-2. Table 4-2 clearly indicates that all relative biases are small. The estimator using an estimated regression parameter is more efficient than an estimator using the true regression parameter. For comparison, Table 4-2 also provides results for calibrated estimators to the population size, which indicate that calibration-to-population size is highly efficient.

Table 4-1
Bias for Model Parameter Estimates

Parameter θ:	$\phi_{M;N}$	$\phi_{I;N}$	$\alpha_M^{(0)}$	$\alpha_I^{(0)}$	$\alpha_M^{(1)}$	$\alpha_I^{(1)}$	$\alpha^{(2)}$
Value	.61	.39	-.7	-.4	.7	.1	.1
Estimate	.60	.40	-.7	-.4	.7	.09	.09
$B(\hat{\theta})$	-.0016	.0016	.0007	-.0003	.0012	-.0055	-.006

Table 4-2
Relative Bias and Mean Square Error Ratios for Finite Population Totals

Parameter of interest	Weights	Relative bias and (MSE ratios) Model parameter: Estimated	Model parameter: True
$\theta = \sum_k y_k$	Design	.0002 (1.00)	-.0004 (1.56)
	Calibration*	-.0002 (.32)	-.0005 (.35)
$\theta_M = \sum_k y_{M;k}$	Design	-.0024 (2.76)	.0063 (6.46)
	Calibration*	.0035 (.81)	-.0036 (.83)
$\theta_I = \sum_k y_{I;k}$	Design	.0031 (3.66)	-.0111 (14.98)
	Calibration*	.0037 (.85)	.0036 (.91)

*Calibration to the population size

5. Concluding remarks

This paper has introduced discrete-time hazard to the analysis of response indicators in surveys and censuses. The proposed approach facilitates examination of the shape of the hazard function. Since inspection of the shape of the hazard function indicates when a response is most likely to occur, and how the probability varies over time, the description of the shapes of hazard function have an important role to play in survey quality and cost. We used regression analysis to investigate the effect of mixed-mode on the response probability. Estimators of model parameters as well as estimators of finite population under mixed-mode surveys were given.

References

Allison, P. D. (1982). "Discrete-time methods for the analysis of event histories." S. Leinhardt (ed.). *Sociological Methodology*. San Francisco: Jossey–Bass. pp. 61–98.

Dempster, A. P., N.M. Laird and D.B. Rubin. (1977). "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society B.* 39, pp. 1–38.

Hartley, H. O. (1958). "Maximum likelihood estimation from incomplete data." *Biometrics*. 4, pp. 174–194.

Rosenbaum, P. R. (1987). "Model-based direct adjustment." *Journal of the American Statistical Association*. 82, pp. 387–394.