On Modeling and Estimation of Response Probabilities
When Missing Data are Not Missing at Random  October 2010

Michail Sverchkov
Bureau of Labor Statistics & SAGE Computing,
2 Massachusetts Avenue, NE, Suite 1950, Washington, DC. 20212,
Sverchkov.Michael@bls.gov

Abstract
Most methods that deal with the estimation of response probabilities assume either explicitly or implicitly that the missing data are ‘missing at random’ (MAR). However, in many practical situations this assumption is not valid, since the probability of responding often depends on the outcome value or on latent variables related to the outcome. The case where the missing data are not MAR (NMAR) can be treated by postulating a parametric model for the distribution of the outcomes under full response and a model for the response probabilities. The two models define a parametric model for the joint distribution of the outcome and the response indicator, and therefore the parameters of this model can be estimated by maximization of the likelihood corresponding to this distribution. Modeling the distribution of the outcomes under full response, however, can be problematic since no data are available from this distribution. Sverchkov (2008) proposed two approaches that permit estimating the parameters of the model for the response probabilities without modelling the distribution of the outcomes under full response. The approaches utilize relationships between the population, the sample and the sample-complement distribution derived in Pfeffermann and Sverchkov (1999) and Sverchkov and Pfeffermann (2004). The present paper extends one of these approaches.

Key words: sample distribution, sample-complement distribution, full likelihood, missing information principle, model selection

1. Definitions and the result
Let \{Y_i, X_i : i \in U\} be a finite population from unknown pdf \(f(Y_i | X_i)\) where “pdf” is the probability density function when \(Y_i\) is continuous or the probability function when \(Y_i\) is discrete. Let \(\{Y_i, X_i : i \in S\}\) be a sample drawn from finite population \(U\) with known inclusion probabilities \(\pi_i = \Pr(i \in S)\). Let \(Y_i\) be the target outcome variable and \(X_i = (X_i^1, ..., X_i^K)\) be covariates (assumed to be fully observed). Denote by \(R\) a sample
of respondents (the sample with observed outcome values) and by \( R^c = S - R \) the corresponding sample of non-respondents. It is assumed that the response occurs stochastically, independently between units.

The observed sample of respondents can be viewed therefore as the result of a two-phase sampling process: in the first phase the parent sample is selected with known inclusion probabilities and in the second phase the final sample is ‘self selected’ with unknown response probabilities (Särndal and Swensson, 1987).

If \( p(Y_i, X_i) = \Pr(i \in R \mid Y_i, X_i, i \in S) \) were known then the sample of respondents could be considered as a sample from the finite population with known selection probabilities \( \tilde{\pi}_i = \pi_i p(Y_i, X_i) \) and population model parameters (or finite population parameters) could be estimated as if there was no non-response.

Also, if known, the response probabilities could be used for imputation via the relationship between the sample and sample-complement distributions (Sverchkov & Pfeffermann 2004),

\[
f(Y_i \mid X_i = x, i \in R^c) = \frac{[p^{-1}(Y_i, x) - 1] f(Y_i \mid X_i = x, i \in R)}{\int [p^{-1}(y, x) - 1] f(y \mid X_i = x, i \in R) dy}.
\]

(Here and in what follows if the outcome variable \( Y_i \) is discrete then the integrals have to be replaced by sums).

Note that \( f(Y_i \mid X_i = x, i \in R) \) refers to the observed data and therefore can be estimated by use of classical statistical inference.

Most methods of estimation in the presence of non-response assume (explicitly or implicitly) that the missing data are ‘missing at random’ (MAR) (Rubin, 1976; Little, 1982), \( \Pr(i \in R \mid Y_i, X_i, i \in S) = \Pr(i \in R \mid X_i, i \in S) \). In many practical situations this assumption is violated: the probability of responding may depend directly on the outcome value. In this case methods that assume MAR can lead to large biases of parameter estimators and large imputation bias.

The case where the missing data are not missing at random (NMAR) can be treated by postulating a parametric model for the distribution of the outcomes before non-response, \( f[Y_i | X_i, i \in S; \alpha] \), and a model for the response probabilities, \( p(Y_i, X_i; \gamma) \), the two models define a parametric model for the joint distribution of the outcomes and the response indicators, therefore the parameters of these models can be estimated by maximization of the likelihood based on the joint distribution (Full Likelihood),

\[
f(Y_i, I_S \mid X_S; \alpha, \gamma) = \prod_{i \in R} p(Y_i, X_i; \gamma) f(Y_i \mid X_i, i \in S; \alpha) \prod_{j \in R^c} [1 - p(X_j; \alpha, \gamma)],
\]

where \( I_S = \{I_k; k \in S\} \) is the set of response indicators, \( p(X_j; \alpha, \gamma) = \int p(y, X_j; \gamma) f[y \mid X_j, j \in S; \alpha] dy \) and the sample outcomes are assumed to be independent. See, Greenlees et al. (1982), Rubin (1987), Little (1993), Beaumont (2000), Little and Rubin (2002) and Qin et al. (2002).
Another way of defining the full likelihood is by application of the Missing Information Principle (MIP, Cipillini et al. 1955, Orchard and Woodbury 1972). The basic idea is to express the score function after non-response as the conditional expectation of the score function before non-response, given the observed data.

Following Chambers (2003, Ch. 2), define the likelihood after non-response as, 
\[ L_R(\lambda) = f(Y_R, X_S, I_S; \lambda), \]
the corresponding likelihood before non-response as 
\[ L_s(\lambda) = f(Y_s, X_S, I_S; \lambda). \]
Then the MIP is, 
\[ sc_R(\lambda) = \frac{d}{d\lambda} \log[L_R(\lambda)] = E[\frac{d}{d\lambda} \log L_s(\lambda) | Y_R, X_S, I_S]. \]

A similar identity defines the relationship between the information matrix after non-response and the information matrix before non-response, which allows estimating the variances of the estimators. See Breckling et al. (1994) and Chambers et al. (1998).

Both approaches face the difficulty that modeling the distribution of the outcomes before non-response refers to partly unobserved data.

**The main result:** Full Likelihood or MIP combined with the relationships between the population, sample and sample-complement distributions derived in Pfeffermann & Sverchkov 1999 and Sverchkov & Pfeffermann 2004 allow us to estimate the parameters of the response model without modeling the distribution of the outcomes before non-response. We indicate how in Section 2 and 3.

### 2. Full Likelihood without need to model the distribution of the outcomes before non-response

Pfeffermann and Sverchkov (1999) derived the relationship between the population distribution and the sample distribution which in the case of non-response can be written as, 
\[ f(Y_j | X_j = x_j, i \in S) = \frac{p^{-1}(Y_j, x_j) f(Y_j | X_j = x_j, i \in R)}{\int p^{-1}(y, x_j) f(y | X_j = x_j, i \in R) dy}. \]

Then, assuming for simplicity independence of the sample outcomes (Poisson sampling),
\[ f(Y_R, I_S | X_S = x_S) = \prod_{i \in R} p(Y_i, x_i) f(Y_i | X_i = x_i, i \in S) \prod_{j \in R} [1 - \int p(y, x_j) f(y | X_j = x_j, j \in S) dy] = \]
\[ \prod_{i \in R} \int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy \times \]
\[ \prod_{j \in R} [1 - \int p^{-1}(y, x_j) f(y | X_j = x_j, i \in R) dy] = \]
\[ \prod_{i \in R} \int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy \]
\[ \prod_{j \in R} [1 - \int p^{-1}(y, x_j) f(y | X_j = x_j, i \in R) dy] = \]
\[ \prod_{i \in R} \int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy \prod_{j \in R} [1 - \int p^{-1}(y, x_j) f(y | X_j = x_j, i \in R) dy]. \]

The Full Likelihood can be defined as
\[ f(Y_r, I_s | X_s = x_s; \beta, \gamma) = \prod_{i \in R} \frac{f(Y_i | X_i = x_i, i \in R; \beta)}{\int p^{-1}(y, x_i; \gamma) f(y | X_i = x_i, i \in R; \beta) dy} \times \]

\[ \prod_{j \in R} \left[ 1 - \frac{p^{-1}(y, x_i; \gamma) f(y | X_i = x_i, i \in R; \beta) dy}{\int p^{-1}(y, x_i; \gamma) f(y | X_i = x_i, i \in R; \beta) dy} \right], \tag{B} \]

where \( f(Y_i | X_i, i \in R; \beta) \) is a model of the outcome distribution \textit{after non-response} and it refers to the fully \textit{observed data}! and therefore can be estimated by use of classical statistical inference.

The response model can be estimated either by maximizing the Full Likelihood (B), in which case both sets of parameters, \( \beta \) and \( \gamma \), are estimated simultaneously, or by estimating \( \beta \) \textit{based on the observed data} and then maximizing \( f(Y_r, I_s | X_s; \hat{\beta}, \gamma) \) over \( \gamma \).

### 3. Likelihood based on MIP without modeling the distribution of the outcomes before non-response

(Again, for simplicity assume Poisson sampling design.)

By use Pfeffermann and Sverchkov (1999) pdf of the outcomes before non-response can be expressed as,

\[ f(Y_s, I_s | X_s = x_s) = \prod_{i \in k} p(Y_i, x_i) f(Y_i | X_i = x_i, i \in S) \prod_{j \in \bar{R}} \left\{ [1 - p(Y_j, x_j)] f(Y_j | X_j = x_j, j \in S) \right\} = \]

\[ \prod_{i \in k} \frac{f(Y_i | X_i = x_i, i \in R)}{\int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy} \times \]

\[ \prod_{j \in \bar{R}} \left\{ [1 - p(Y_j, x_j)] \frac{p^{-1}(Y_j, x_j) f(Y_j | X_j = x_j, j \in R)}{\int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy} \right\} = \]

\[ \prod_{i \in k} \frac{f(Y_i | X_i = x_i, i \in R)}{\int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy} \times \]

\[ \prod_{j \in \bar{R}} \left\{ [p^{-1}(Y_j, x_j) - 1] \frac{f(Y_j | X_j = x_j, j \in R)}{\int p^{-1}(y, x_i) f(y | X_i = x_i, i \in R) dy} \right\}. \]

Therefore the log-likelihood \textit{before non-response} can be defined as,

\[ L_s(\gamma, \beta) = \sum_{i \in k} \log \left[ \frac{f(Y_i | X_i = x_i, i \in R; \beta)}{\int p^{-1}(y, x_i; \gamma) f(y | X_i = x_i, i \in R; \beta) dy} \right] + \]

\[ \sum_{j \in \bar{R}} \log \left[ \frac{[p^{-1}(Y_j, x_j; \gamma) - 1]}{\int p^{-1}(y, x_i; \gamma) f(y | X_i = x_i, i \in R; \beta) dy} \right] \]

Then, following MIP, the score function based on the likelihood \textit{after non-response} can be written as,
Again, the likelihood after non-response is a function of the response model and the distribution of the outcomes after non-response (and the latter can be modeled or estimated from the observed data).

As in Section 2, the response model can be estimated either by solving the likelihood equations based on (C), \( \text{sc}(\gamma, \beta) = 0 \), in this case both sets of parameters, \( \beta \) and \( \gamma \) are estimated simultaneously, or by estimating \( \hat{\beta} \) based on the observed data and then solving \( \text{sc}(\gamma, \hat{\beta}) = 0 \) over \( \gamma \).
Sverchkov (2008) consider another estimating procedure similar to the above also based on MIC.

4. Remarks

Based on the full likelihood (B) or the score function (C) one can define the classical information criteria like Akaike AIC, Schwarz BIC, etc. which can be used for selecting the response model. Also, one can define the information matrix based on (B) or (C) and therefore estimate the variance of the parameter estimators. The latter allow checking whether response is NMAR or MAR: if parameter estimates connected with the outcome variable are insignificant then the response is rather MAR (see Sverchkov 2008) and therefore the simpler methods that assume MAR can be applied for estimating the response probabilities.

Acknowledgements

Some of this research is supported by a grant from the United Sates-Israel Binational Science Foundation (BSF). The author thanks Alan Dorfman and Danny Pfeffermann for useful discussions.

References


The opinions expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics.