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

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),

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

(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*, $f[Y_i \mid X_i, i \in S; \alpha]$, 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*),

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

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,

$$sc_R(\lambda) = \tfrac{\partial}{\partial\lambda}\log[L_R(\lambda)] = E[\tfrac{\partial}{\partial\lambda}\log L_S(\lambda) \mid \underbrace{Y_R, X_S, I_S}_{Observed}].$$

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,

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

Then, assuming for simplicity independence of the sample outcomes (Poisson sampling),

$$f(Y_R, I_S \mid X_S = x_S) =$$
$$\prod_{i\in R} p(Y_i, x_i) f[Y_i \mid X_i = x_i, i \in S] \prod_{j\in R^c} [1 - \int p(y, x_j) f(y \mid X_j = x_j, j \in S) dy] =$$
$$\prod_{i\in R} \frac{f(Y_i \mid X_i = x_i, i \in R)}{\int p^{-1}(y, x_i) f(y \mid X_i = x_i, i \in R) dy} \times$$
$$\prod_{j\in R^c} [1 - \int \frac{f(y \mid X_i = x_j, i \in R)}{\int p^{-1}(y, x_j) f(y \mid X_i = x_j, i \in R) dy} dy] =$$
$$\prod_{i\in R} \frac{f(Y_i \mid X_i = x_i, i \in R)}{\int p^{-1}(y, x_i) f(y \mid X_i = x_i, i \in R) dy} \prod_{j\in R^c} [1 - \frac{1}{\int p^{-1}(y, x_j) f(y \mid X_i = x_j, i \in R) dy}].$$

The Full Likelihood can be defined as

$$f(Y_R, I_S \mid X_S = x_S; \beta, \gamma) = \prod_{i \in R} \frac{f(Y_i \mid X_i = x_i, i \in R; \beta)}{\int p^{-1}(y, x_i; \gamma) f(y \mid X_i = x_i, i \in R; \beta) dy} \times$$

$$\prod_{j \in R^c} [1 - \frac{1}{\int p^{-1}(y, x_j; \gamma) f(y \mid X_i = x_j, i \in R; \beta) dy}], \qquad \textbf{(B)}$$

where $f(Y_i \mid X_i, i \in R; \beta)$ is a model of the outcome distribution *after non-response* and it refers to the fully *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, β and γ, are estimated simultaneously, or by estimating β *based on the observed data* and then maximizing $f(Y_R, I_S \mid X_S; \hat{\beta}, \gamma)$ over γ.

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,

$$f(Y_S, I_S \mid X_S = x_S) =$$

$$\prod_{i \in R} p(Y_i, x_i) f[Y_i \mid X_i = x_i, i \in S] \prod_{j \in R^c} \{[1 - p(Y_j, x_j)] f(Y_j \mid X_j = x_j, j \in S)\} =$$

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

$$\prod_{j \in R^c} \{[1 - p(Y_j, x_j)] \frac{p^{-1}(Y_j, x_j) f(Y_j \mid X_j = x_j, j \in R)}{\int p^{-1}(y, x_j) f(y \mid X_i = x_j, i \in R) dy}\} =$$

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

$$\prod_{j \in R^c} \{[p^{-1}(Y_j, x_j) - 1] \frac{f(Y_j \mid X_j = x_j, j \in R)}{\int p^{-1}(y, x_j) f(y \mid X_i = x_j, i \in R) dy}\}.$$

Therefore the log-likelihood *before non-response* can be defined as,

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

$$\sum_{j \in R^c} \log\{[p^{-1}(Y_j, x_j; \gamma) - 1] \frac{f(Y_j \mid X_j = x_j, j \in R; \beta)}{\int p^{-1}(y, x_j; \gamma) f(y \mid X_i = x_j, i \in R; \beta) dy}\}$$

Then, following MIP, the score function based on the likelihood *after non-response* can be written as,

$$sc_R(\gamma,\beta)=\tfrac{\partial}{\partial(\gamma,\beta)}\log[L_R(\gamma,\beta)]=E[\tfrac{\partial}{\partial(\gamma,\beta)}\log L_S(\gamma,\beta)\,|\,\underbrace{Y_R=y_R,X_S=x_S,I_S=i_S}_{Observed}]=$$

$$\sum_{i\in R}\tfrac{\partial}{\partial(\gamma,\beta)}\log[\frac{f(Y_i\mid X_i=x_i,i\in R;\beta)}{\int p^{-1}(y,x_i;\gamma)f(y\mid X_i=x_i,i\in R;\beta)dy}]+$$

$$\sum_{j\in R^c}E[\tfrac{\partial}{\partial(\gamma,\beta)}\log\{[p^{-1}(Y_j,x_j;\gamma)-1]\times$$

$$\frac{f(Y_j\mid X_j=x_j,j\in R;\beta)}{\int p^{-1}(y,x_j;\gamma)f(y\mid X_i=x_j,i\in R;\beta)dy}\mid X_j=x_j,j\in R^c;\beta]\overset{by\ (A)}{=}$$

$$\sum_{i\in R}\tfrac{\partial}{\partial(\gamma,\beta)}\log[\frac{f(Y_i\mid X_i=x_i,i\in R;\beta)}{\int p^{-1}(y,x_i;\gamma)f(y\mid X_i=x_i,i\in R;\beta)dy}]+$$

$$\sum_{j\in R^c}\int dy\left(\frac{[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)}{\int[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)dy}\times\right.$$

$$\left.\tfrac{\partial}{\partial(\gamma,\beta)}\log\{\frac{[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)}{\int p^{-1}(y,x_j;\gamma)f(y\mid X_i=x_j,i\in R;\beta)dy}\}\right)=$$

$$\sum_{i\in R}\tfrac{\partial}{\partial(\gamma,\beta)}\log[\frac{f(Y_i\mid X_i=x_i,i\in R;\beta)}{\int p^{-1}(y,x_i;\gamma)f(y\mid X_i=x_i,i\in R;\beta)dy}]+$$

$$\sum_{j\in R^c}\int dy\left(\frac{\int p^{-1}(Y_j,x_j;\gamma)f(y\mid X_i=x_j,i\in R;\beta)dy}{\int[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)dy}\times\right.$$

$$\left.\tfrac{\partial}{\partial(\gamma,\beta)}\{\frac{[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)}{\int p^{-1}(y,x_j;\gamma)f(y\mid X_i=x_j,i\in R;\beta)dy}\right)=$$

$$\sum_{i\in R}\tfrac{\partial}{\partial(\gamma,\beta)}\log[\frac{f(Y_i\mid X_i=x_i,i\in R;\beta)}{\int p^{-1}(y,x_i;\gamma)f(y\mid X_i=x_i,i\in R;\beta)dy}]+$$

$$\sum_{j\in R^c}\frac{\int p^{-1}(y,x_j;\gamma)f(y\mid X_i=x_j,i\in R;\beta)dy}{\int[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)dy}\times$$

$$\int dy\left(\tfrac{\partial}{\partial(\gamma,\beta)}\{\frac{[p^{-1}(y,x_j;\gamma)-1]f(y\mid X_i=x_j,i\in R;\beta)}{\int p^{-1}(y,x_j;\gamma)f(y\mid X_i=x_j,i\in R;\beta)dy}\right).\qquad\textbf{(C)}$$

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),** $sc_R(\gamma,\beta)=0$, in this case both sets of parameters, β and γ are estimated simultaneously, or by estimating β based on the observed data and then solving $sc_R(\gamma,\hat{\beta})=0$ over γ.

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

Beaumont, J.F. (2000). An estimation method for nonignorable nonresponse, *Survey Methodology*, **26**, 131-136.

Breckling J.U., Chambers R.L., Dorfman A.H., Tam S.M., and Welsh A.H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review*, **62**, 349-363.

Cepillini, R., Siniscialco, M., and Smith, C.A.B. (1955). The estimation of gene frequencies in a random mating population, *Annals of Human Genetics*, **20**, 97-115.

Chambers R.L., Dorfman, A.H., and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B*, **60**, 397-411.

Chambers, R. L. (2003). In: R. L. Chambers and C. Skinner (eds.). *Analysis of survey data*. New York: Wiley.

Greenlees, J.S. Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed, *Journal of the American Statistical Association*, **77** , 251-261.

Little, R.J.A. (1982). Models for nonresponse in sample surveys, *Journal of the American Statistical Association* **77** , 237-250.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, **88** , 125-134.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data,* New York: Wiley.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: theory and application, *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697-715.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data, *Sankhya*, 61, 166-186.

Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581-590.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys,* New York: Wiley

Qin, J., Leung, D. and Shao, J. (2002). Estimation with Survey data under nonignorable nonresponse or informative sampling, *Journal of the American Statistical Association* **97**, 193-200.

Sarndal C.E., and Swensson B. (1987). A general view of estimation for two fases of selection with applications to two-phase sampling and nonresponse, I*nternational Statistical Review* **55**, 279-294.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution, *Survey Methodology* **30**, 79-92.

Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random. *2008 JSM Meetings, Proceedings of the Section on Survey Methods Research*, 867-874

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