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

Michail Sverchkov

Bureau of Labor Statistics
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 a new approach that permits estimating the parameters of the model for the response probabilities without modelling the distribution of the outcomes under full response. The approach utilizes 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 investigates how this approach can be used for testing whether response is MAR or NMAR.

Key words: sample distribution, complement-sample distribution, prediction under informative sampling and non-response, estimating equations, missing information principle, non-parametric estimation

1. Introduction

There is almost no survey without nonresponse, but in practice most methods that deal with this problem assume either explicitly or implicitly that the missing data are 'missing at random' (MAR, Rubin, 1976; Little, 1982). However, in many practical situations this assumption is not valid, since the probability of responding often depends directly on the outcome value. In this case, the use of methods that assume that the nonresponse is MAR can lead to large biases of population parameter estimators and large imputation bias.

The case where the missing data are not MAR (NMAR) can be treated by postulating a parametric model for the distribution of the outcomes *before non-response* and a model for the response mechanism. These two models define a parametric model for the joint distribution of the outcomes and response indicators, and therefore the parameters of these models can be estimated by maximization of the likelihood based on this joint distribution. See, Greenlees *et al*. (1982), Rubin (1987), Little (1993), Beaumont (2000), Little and Rubin (2002) and Qin *et al*. (2002).

Modeling the distribution of the outcomes *before non-response* can be problematic since it refers to the partly unobserved data. Qin *et al.* (2002) suggests using a non-parametric model for this distribution (empirical likelihood approach). Sverchkov (2008) suggests an alternative approach that allows one to estimate the parameters of the response model by independent parametric or non-parametric estimation of the outcomes distribution *after non-response* (which can be done by use of classic statistical inferences since the latter refers to the observed data) and then by solving

estimating equations obtained from the census likelihood function of the response indicators. The derivation of these estimating equations utilizes the relationships between the population, the sample and the sample-complement distributions, as in Pfeffermann and Sverchkov (1999, 2003), Sverchkov and Pfeffermann (2004). Even under this approach one needs to assume a model for the response mechanism which cannot be checked from the observed data in case of NMAR. Therefore it is important to know whether the response is MAR or NMAR. The present paper investigates how the Sverchkov (2008) approach can be used for testing whether response is MAR or NMAR.

2. Notation

Let Y_i denote the value of an outcome variable Y associated with unit i belonging to a sample $S=\{1,...,n\}$, drawn from a finite population $U=\{1,...,N\}$. Let X_i denote the corresponding values of covariates $X_i=(X_{1i},...,X_{Ki})'$. In what follows we assume that the population outcome values are independent realizations from distributions with unknown probability density functions (*pdf*), $f(Y_i|X_i)$. We use the abbreviation *pdf* for the probability density function when Y_i is continuous and the probability function when Y_i is discrete. Let $R=\{1,...,n_r\}$ define the sample of respondents (the sample with observed outcome values), and $R^c=\{n_r+1,...,n\}$ define the sample of nonrespondents. The response process is assumed to occur stochastically, independently between units. The observed sample of respondents can be viewed therefore as the result of a two-phase sampling process where in the first phase the sample S is selected from U with known inclusion probabilities $\pi_i=\Pr(i\in S)$ and in the second phase the sample R is 'self selected' with unknown response probabilities (Särndal and Swensson, 1987).

Denote by $p(Y_i,X_i)=\Pr(i\in R|Y_i,X_i,i\in S)$ and let u_i and v_i be any random vectors such that (u_i,v_i) and response indicators, R_i ($R_i=1$ if $i\in R$ and 0 otherwise), are independent given $(Y_i,X_i,i\in S)$. For example, u_i and v_i are functions of (Y_i,X_i), or the responses are completely defined by (Y_i,X_i). In what follows we use the following relationships between population and sample distribution (Pfeffermann and Sverchkov 1999, 2003 and Sverchkov and Pfeffermann 2004) which can be written in terms of response probabilities as,

$$E(u_i|v_i,i\in S)=\frac{E(p^{-1}(Y_i,X_i)u_i|v_i,i\in R)}{E(p^{-1}(Y_i,X_i)|v_i,i\in R)}, \tag{2.1}$$

$$E(u_i|v_i,i\in R^c)=\frac{E\{[p^{-1}(Y_i,X_i)-1]u_i|v_i,i\in R)}{E\{[p^{-1}(Y_i,X_i)-1]|v_i,i\in R)}. \tag{2.2}$$

Note that (2.1) implies

$$E[p^{-1}(Y_i,X_i)|i\in R]=1/E[p(Y_i,X_i)|i\in S]. \tag{2.3}$$

Remark 2.1 In the following sections we concentrate on estimation of the response probabilities $p(Y_i,X_i)$. Note that if the response probabilities or their estimates are known then the sample respondents can be considered as a sample from the finite population with known, $\tilde{\pi}_i=\pi_i p(Y_i,X_i)$, or estimated selection probabilities, $\hat{\tilde{\pi}}_i=\pi_i\hat{p}(Y_i,X_i)$. Then population model parameters (or finite population parameters) can be estimated as if there were no non-response with these new inclusion probabilities, see Särndal and Swensson (1987). One can use these probabilities for imputation also using the relationship between the sample and sample-complement distributions derived in Sverchkov and Pfeffermann (2004),

$$f(u_i|v_i,i\in R^c)=\frac{[p^{-1}(Y_i,X_i)-1]f(u_i|v_i,i\in R)}{E\{[p^{-1}(Y_i,X_i)-1]|v_i,i\in R\}}. \tag{2.4}$$

3. Estimation of the Response Probabilities when Non-Response is NMAR (Sverchkov 2008)

Let $p(Y_i, X_i; \gamma) = \Pr(i \in R \mid Y_i, X_i, i \in S; \gamma)$ be a parametric set of pdf's and suppose that $p(Y_i, X_i; \gamma)$ is differentiable with respect to (vector) parameter γ.

For simplicity we consider the following scenario: The covariates are observed for all non-respondents, i.e. Observed Data= $\{Y_i, i \in R, X_k, k \in S\}$.

Under this scenario, if the missing data were later observed, γ could be estimated by solving the likelihood equations,

$$\sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} = 0. \qquad (3.1)$$

Similarly to the Missing Information Principle (Cipillini *et al*, 1955, Orchard and Woodbury 1972), since the outcome values are missing for $j \in R^c$, we propose to solve instead,

$$0 = E[\sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \mid \text{Observed Data}], \text{ i.e.,}$$

$$0 = E[\sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \mid \{Y_i, i \in R, X_k, k \in S\}]$$

$$= \sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} E[\frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \mid i \in R^c, \{X_k, k \in S\}]$$

$$= \sum_{i \in R} \frac{\partial \log p(Y_i, X_i; \gamma)}{\partial \gamma} + \sum_{i \in R^c} \frac{E\{[p^{-1}(Y_i, X_i; \gamma) - 1] \frac{\partial \log[1 - p(Y_i, X_i; \gamma)]}{\partial \gamma} \mid X_i, i \in R\}}{E\{[p^{-1}(Y_i, X_i; \gamma) - 1] \mid X_i, i \in R\}} \qquad (3.2a)$$

$$= \sum_{i \in R} \frac{\partial p(Y_i, X_i; \gamma)}{\partial \gamma} p^{-1}(Y_i, X_i; \gamma) - \sum_{i \in R^c} \frac{\int p^{-1}(Y_i, X_i; \gamma) \frac{\partial p(Y_i, X_i; \gamma)}{\partial \gamma} f(Y_i \mid X_i, i \in R) dY_i}{\int p^{-1}(Y_i, X_i; \gamma) f(Y_i \mid X_i, i \in R) dY_i - 1}. \qquad (3.2b)$$

The third equation follows from (2.2) where we assume for simplicity that $p(Y_i, X_i; \gamma)$ and $(X_k, k \in S)$ are independent given X_i. Note that the second sum in (3.2a) and (3.2b) predicts the unobserved second sum in (3.1). Note also that if $p(Y_i, X_i; \gamma)$ is a function of X_i and γ only (missing data are MAR) then (3.2b) reduces to a common system of log-likelihood equations,

$$\sum_{i \in R} \frac{\partial \log p(X_i; \gamma)}{\partial \gamma} - \sum_{i \in R^c} \frac{\partial \log[1 - p(X_i; \gamma)]}{\partial \gamma} = 0. \qquad (3.3)$$

Estimating functions (3.2b) suggest the following two-step estimation procedure:

Step 1. Fit the model $f_r(Y_i \mid X_i) = f(Y_i \mid X_i, i \in R)$. Note that this *pdf* refers to the respondents' sample and therefore can be identified and estimated from the observed data using classic statistical inference.

Step 2. Approximate (3.2b) by replacing $f_r(Y_i \mid X_i)$ by its estimate, $\hat{f}_r(Y_i \mid X_i)$, and solve (3.2b) for γ.

Note that instead of estimation of f_r in (3.2b) one can estimate the expectations in (3.2a) non-parametrically, and after substituting the estimates in (3.2a) solve them for γ. For example, for discrete X-s and an arbitrary function g, $E[g(Y_j, X_j, \gamma) \mid X_j = x, j \in R]$ can be estimated by the

respective mean, $\left(\sum_{j\in R:X_j=x} 1\right)^{-1} \sum_{j\in R:X_j=x} g(Y_j, X_j, \gamma)$. For continuous X-s let $m(x,\gamma)$ be an estimator of $E(g(Y_j, X_j, \gamma) \mid X_j = x, j \in R)$, for example the Nadaraya-Watson estimator, $m(x,\gamma) = \dfrac{\sum_{j\in R} K[(x - X_j)/h] g(Y_j, X_j, \gamma)}{\sum_{j\in R} K[(x - X_j)/h]}$, where h and K are a scale-factor and a kernel. Estimating the respective conditional expectations in the second sum of (3.2a) by $m(x,\gamma)$ one obtains the following *estimating equations*,

$$\sum_{i\in R} \frac{\partial p(Y_i, X_i;\gamma)}{\partial \gamma} p^{-1}(Y_i, X_i;\gamma) - \sum_{j\in R^c} \frac{\sum_{k\in R} K[(X_k - X_j)/h] p^{-1}(Y_k, X_k;\gamma) \dfrac{\partial p(Y_k, X_k;\gamma)}{\partial \gamma}}{\sum_{k\in R} K[(X_k - X_j)/h][p^{-1}(Y_k, X_k;\gamma) - 1]} = 0, \qquad (3.4)$$

which defines an estimator of γ.

Estimating equations (3.4) do not require any knowledge of the model for the respondents. On the other hand one can expect that the estimates obtained by solving (3.4) will be less stable than the estimates obtained from (3.2b) by the above two step estimation procedure when the model for the respondents can be fitted well.

4. Is it MAR or NMAR?

The proposed approach requires knowledge of the parametric form of the response model which refers to the unobserved data in the case of NMAR. On the other hand, if response is MAR, the propensity score, $p(X_i;\alpha) = \Pr(i \in R \mid X_i, i \in S;\alpha)$, can be estimated from the observed data for example by solving a common system of log-likelihood equations (3.3). Note that the latter estimator much more stable than the estimators assuming NMAR. Therefore it is very important to know whether response is MAR or NMAR. We suggest using the following procedure for testing the latter:

Step 1. Fit the model for propensity score, $p(X_i;\alpha) = \Pr(i \in R \mid X_i, i \in S;\alpha)$, and estimate the parameter α from the observed data assuming MAR.

Step 2. Define a class of models for $p(Y_i, X_i;\gamma) = \Pr(i \in R \mid Y_i, X_i, i \in S;\gamma)$, $\gamma \in \Gamma$, in such way that for some $\tilde{\gamma} \in \Gamma$, $p(Y_i, X_i;\tilde{\gamma}) = p(X_i;\alpha)$. It recommended to use models that include the *Y*-component in a simple form, for example, if $\text{logit}[p(X_i;\alpha)] = g(X_i;\alpha)$ then one can consider $\text{logit}[p(Y_i, X_i;\gamma)] = \text{g}(\text{X}_i;\alpha) + cY_i$, $\gamma = (\alpha, c)$, so in this case for $\tilde{\gamma} = (\alpha, 0)$, $p(Y_i, X_i;\tilde{\gamma}) = p(X_i;\alpha)$.

Step 3. Obtain estimating equations (3.2a) based on the class of models defined in Step 2.

Step 4.1. Solve them and check whether *Y*-component is significant (in which case the response is for sure NMAR) or not (the response is MAR or "not very informative").

The latter can be done by a bootstrap procedure: one can take *B* simple random samples with replacement from the original sample and repeat steps 1 – 4 above in order to get a variance estimate for the *Y*-component.

Remark 4.1. Since the parametric family defined in Step 2 does not necessarily include the true response probability $\Pr(i \in R \mid Y_i, X_i, i \in S)$, even if the *Y*-component is insignificant we cannot conclude for sure that response is MAR. Nevertheless, we recommend to use propensity score assuming MAR in this case. If response is very informative then one can expect that the *Y*-component will be significant even in a simplified model.

Instead of Step 4.1 one can do
Step 4.2. Substitute $\tilde{\gamma}$ from Step 2 into (3.2a-b) obtained in Steps 1 - 3 and check whether the result is significantly non-zero (response is NMAR) or not (response "seems to be" MAR since $\tilde{\gamma}$ corresponds to the propensity score). The latter can also be done by use of a bootstrap.

Acknowledgements: The opinions expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics. 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.
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.
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.
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.
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.
Pfeffermann, D., and Sverchkov, M. (2003). "Fitting generalized linear models under informative probability sampling", In: *Analysis of Survey Data*, eds. C. J. Skinner and R. L. Chambers. New York: John Wiley & Sons. pp 175-195.
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.