A ROBUST ESTIMATION METHOD IN THE PRESENCE OF NONIGNORABLE NONRESPONSE

Jean-François Beaumont, Statistics Canada
Household Survey Methods Division, Statistics Canada, Ottawa (Ontario), Canada K1A 0T6

Key Words: Nonignorable Nonresponse, Imputation, Maximum Likelihood, Estimating Equations

1. INTRODUCTION

Most surveys suffer from total and/or item nonresponse. Total nonresponse is experienced when all the survey items are missing for at least one unit in the sample and is generally compensated for by using nonresponse adjustment weights. Item nonresponse occurs when some but not all items are missing for one or more sampled units. It is generally compensated for by using imputation methods. The focus of this paper will only be on item nonresponse and imputation.

When the probability of responding to a given item in a survey does not depend on unobserved values, as it is the case when it depends on an auxiliary variable observed for all units of the sample, then nonresponse is said to be ignorable. In all other cases, nonresponse is nonignorable. It happens when the response probability of a given unit depends on the value of the variable of interest only observed for part of the sample. Readers are referred to Little (1982) for a more formal definition of the concepts of nonignorable and ignorable nonresponse. In practice, it is often assumed that nonresponse is ignorable. This assumption can, however, be erroneous. For example, when income is measured in a survey, it is realistic to believe that individuals with a low income are more likely to be nonrespondents to an income item than individuals with a high income, or vice versa.

In estimating a mean or a total, using techniques appropriate for ignorable nonresponse when nonresponse is actually nonignorable can lead to a severe bias. Few statistical methods dealing with nonignorable nonresponse are available. For ratio imputation, Rancourt, Lee and Särndal (1994) propose simple correction factors that reduce bias when estimating a mean. Greenlees, Reece and Zieschang (1982) model the response probability and the variable of interest jointly, using the maximum likelihood method to estimate parameters of the models. This method, however, is based on the assumption that errors, in the model involving the variable of interest, are normally distributed. This assumption can be difficult to test when nonresponse is nonignorable.

This paper will propose a simple way of testing the normality assumption and will present an estimation method that is robust against bad specification of the error distribution and, to some extent, against bad specification of the model involving the variable of interest.

In section 2, the problem is more precisely defined and some notation is introduced. The third section contains different estimators of a population mean under different hypotheses concerning the response mechanism and the data distribution. In section 4, the results of a simulation study comparing the estimators described in the third section are presented. Finally, the last section contains a brief discussion.

2. NOTATION

In the following, we want to estimate the mean of the variable Y for a population P. First, a sample S is selected from the population. Then, the survey is conducted and the variable Y is only observed for part of the sample S. The sample of respondents is denoted by R and the sample of nonrespondents is denoted by O. It is also assumed that we have one variable, X, observed for all units in the sample S and correlated with Y.

The estimator of the population mean, $\mu = \sum_{i\in P} Y_i / N$, where N is the population size, can be obtained by imputing missing values:

$$\mu_I^* = \frac{\sum_{i\in R} w_i Y_i + \sum_{i\in O} w_i Y_i^*}{\sum_{i\in O} w_i}, \qquad (2.1)$$

where w_i are the sampling weights corresponding to the inverse of the selection probabilities and Y_i^* are imputed values for the nonresponding units. For simplicity, it will be assumed, in the following, that the sampling weights are constant for all the units in the population. The sampling weights w_i can thus be removed from equation (2.1).

3. ESTIMATION METHODS

This section is devoted to developing expression (2.1) under different hypotheses concerning the response mechanism and the data distribution and to describe appropriate estimation methods. Section (3.1) considers the case of an ignorable response mechanism and section (3.2) deals with a nonignorable response mechanism.

3.1 Ignorable response mechanism

A very simple ignorable response mechanism is one in which all the units in the sample have the same response probability and is called the uniform response mechanism. Under such a response mechanism, even imputing the mean of respondents to the nonrespondents will produce a valid estimator of the population mean. This estimator can, however, be strongly biased in the presence of a more realistic response mechanism such as one in which the response probability depends on the variable X (which is correlated with Y). In this case, it is preferable to use the variable X as supplementary information for the estimation of the population mean μ. For example, the usual ratio imputation estimator (RAT) imputes missing values by using the following model:

$$Y_i = \beta X_i + \varepsilon_i , \tag{3.1}$$

where β is an unknown parameter and ε_i is a random error term uncorrelated with X_i with zero mean and variance $\sigma^2 X_i$. The imputed values are given by $Y_i^* = B^* X_i$, where $B^* = \sum_{i\in R} Y_i / \sum_{i\in R} X_i$ is an estimator (obtained by the weighted least-squares method and the responding units) of B which is itself an estimator of β. Actually, B is the estimator that we would have obtained (by the weighted least-squares method) if we had observed all the units in the sample S. Note finally that all the models considered in this paper are assumed to be valid for all the units in the sample S.

3.2 Nonignorable response mechanism

The usual ratio imputation estimator, RAT, can be strongly biased when the response mechanism depends on unobserved values of Y (nonignorable response mechanism). Using model (3.1) with responding units leads to an inconsistent estimator of the slope β because the expectation of the random error ε_i is not zero, given that i is a responding unit. The RAT estimator is thus not satisfactory in the presence of a nonignorable response mechanism. In such a situation, a better approach consists of simultaneously estimating the parameter of equation (3.1) and some parameters used to model the response probability. For example, this probability could be modeled in the following way:

$$P(R_i = 1 \mid Y_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 Y_i)]} , \tag{3.2}$$

where α_0 and α_1 are unknown parameters and R_i is a binary variable indicating whether unit i responds ($R_i = 1$) or not ($R_i = 0$).

In the next three subsections, three estimation methods developed for the case of a response mechanism that depends on Y will be presented.

3.2.1 Maximum likelihood method

This method can be found in Greenlees, Reece and Zieschang (1982). It consists of using models (3.1) and (3.2) and requires the additional assumption that random errors ε_i are normally distributed (or any other distribution relevant for the type of data analyzed) and mutually independent. The natural logarithm of the likelihood function, l, can be written:

$$l = \sum_{i\in R} \ln[p(Y_i) f(Y_i \mid X_i)] + \sum_{i\in O} \ln[1 - \mathrm{E}(p(Y_i) \mid X_i)] , \tag{3.3}$$

where $p(Y_i) = P(R_i = 1 \mid Y_i)$ and $f(Y_i \mid X_i)$ is the probability density function of Y_i given X_i. The maximum likelihood method consists of finding the parameter values which maximize l. To calculate $\mathrm{E}(p(Y_i) \mid X_i)$, the following approximation has been used (Zeger, Liang and Albert, 1988):

$$\mathrm{E}(p(Y_i)|X_i) \approx \frac{1}{1 + \exp\{-k[\alpha_0 + \alpha_1(\beta X_i)]\}} , \tag{3.4}$$

where $k = 1/\sqrt{c^2\alpha_1^2\sigma^2 X_i + 1}$ and $c = 16\sqrt{3}/15\pi$. This approximation is based on the assumption that errors are normally distributed with variance $\sigma^2 X_i$. The maximization of (3.3) was done by using a Newton-Raphson type algorithm and the NLIN procedure of the SAS software (SAS Institute Inc.,1990).

Once the unknown parameters have been estimated, imputed values Y_i^* can be determined so as to minimize $\sum_{i\in S} e_i^2 / X_i$ under the constraint $\sum_{i\in S} e_i = 0$, where $e_i = Y_i - \beta^* X_i$, for $i \in R$, $e_i = Y_i^* - \beta^* X_i$, for $i \in O$, and β^* is the estimate of β. The estimator of the population mean can thus be written: $\mu_I^* = \beta^* \sum_{i\in S} X_i / n$, where n is the sample size of S, and will be denoted by ML. The rationale behind this approach is that the preceding constraint would have been respected if we had observed the variable Y for all the units in the sample S and if we had modeled this variable by the use of model (3.1).

3.2.2 Robust estimation method

The maximum likelihood method requires that the model (3.1) be appropriate and the errors be normally distributed. When one or both assumptions are violated, it is preferable to use a more robust estimation method. If the response probabilities were known and greater than zero for all the units in the population, a robust estimation method (against bad specification of the error distribution as well as against bad specification of the model (3.1)) could be obtained by minimizing the sum of squared errors weighted by the inverse of the error variance times the inverse of the response probabilities $p(Y_i)$. This minimization is equivalent to solving the system of equations

$$\sum_{i\in R} \frac{1}{p(Y_i)}(Y_i - \beta X_i) = 0 \ . \qquad (3.5)$$

By similar reasoning, if $f(Y_i \mid X_i)$ were known, we could estimate α_0 and α_1 of model (3.2) by the maximum likelihood method and solve the system of equations

$$\sum_{i\in R} \frac{\partial}{\partial\alpha_k} \ln[p(Y_i)] + \sum_{i\in O} \frac{\partial}{\partial\alpha_k} \ln[1 - \mathrm{E}(p(Y_i) \mid X_i)] = 0 \ , \qquad (3.6)$$

for $k \in \{0,1\}$. Parameter estimates of β, α_0 and α_1 can thus be obtained by solving unbiased estimating equations (3.5) and (3.6). An algorithm allowing the discovery of the solution consists of alternatively solving (3.5) and (3.6) until convergence is reached. To achieve this, $\mathrm{E}(p(Y_i) \mid X_i)$ must be calculated. Calculating this expectation, however, requires knowledge of the error distribution (which is likely to be unknown). To cope with this problem, we can assume that errors are normally distributed and use approximation (3.4), but this approximation was found to be not very robust against bad specification of model (3.1). A globally better and simpler approximation to $\mathrm{E}(p(Y_i) \mid X_i)$ allowing us to get rid of the normality assumption is obtained by using the first term of a Taylor series expansion:

$$\mathrm{E}(p(Y_i) \mid X_i) \approx p(\beta X_i) \ . \qquad (3.7)$$

An interesting property of (3.7) is that alternatively solving (3.5) and (3.6) is obtained with the following algorithm:

1. Set initial values for α_0 and α_1 (for example, set $p(Y_i) = 1$ for all responding units);
2. Solve (3.5) with current values for α_0 and α_1 by using a weighted linear regression procedure;
3. Impute missing values by $Y_i^{*(j)} = \beta^{*(j)} X_i$, where superscript *(j)* indicates the iteration number (one iteration consists of passing from step 2 to step 4) and $\beta^{*(j)}$ is obtained through step 2;
4. Solve (3.6) by using a logistic regression procedure and the imputed values calculated in step 3;
5. Return to step 2 or stop if convergence is reached.

Thus, it suffices to have only a linear regression procedure and a logistic regression procedure to obtain the desired estimates. In practice, this algorithm is very efficient for finding the solution but, in some cases, it can take many iterations to reach convergence. This is the reason for the use of the Newton-Raphson algorithm in the simulation study of the next section. Note that the algorithm above bears some similarities to the EM algorithm of Dempster, Laird and Rubin (1977), except that it is not conceived to maximize a likelihood function.

As in section (3.2.1), once the unknown parameters have been estimated, imputed values Y_i^* can be determined so as to minimize $\sum_{i\in S} e_i^2 / X_i$ under the constraint $\sum_{i\in S} e_i = 0$, where $e_i = Y_i - \beta^* X_i$, for $i\in R$, $e_i = Y_i^* - \beta^* X_i$, for $i\in O$. The estimator of the population mean can still be written by $\mu_I^* = \beta^* \sum_{i\in S} X_i / n$ and will be denoted by R-T1.

A slight modification of the algorithm above can be obtained by changing step 3 so as to impute missing values as described in the preceding paragraph but using $\beta^{*(j)}$, the current value of β^* at iteration j, instead of β^*. This modification, however, was not retained in this paper because it was found to be not globally superior to R-T1 and could sometimes require too many iterations and much computer time before converging.

Instead of doing step 4 of the algorithm presented in this section, we could also solve the following two equations:

$$\sum_{i\in S} R_i - p(Y_{i\bullet}) = 0$$

and

$$\sum_{i\in S} (R_i - p(Y_{i\bullet})) Y_i^{*(j)} = 0 \ ,$$

where $Y_{i\bullet} = Y_i$, for $i\in R$, and $Y_{i\bullet} = Y_i^{*(j)}$, for $i\in O$. The estimator of the population mean obtained this way will be denoted by R-T1M. Note that step 4 is very similar to the preceding two equations, except that in step 4, $Y_{i\bullet}$ replaces $Y_i^{*(j)}$.

In order to use the maximum likelihood method, we can be interested in verifying whether or not the errors of model (3.1) seem to be normally distributed. To achieve this, we can do the graph of standardized residuals $e_i/\sigma^* X_i^{1/2}$ versus $\Phi^{-1}[F^*(e_i/\sigma^* X_i^{1/2})]$, for $i\in R$, where σ^* is an estimate of σ, $\Phi(.)$ is the distribution function of a standard normal random variable, and $F^*(.)$ is the estimated empirical distribution function. When errors are normally distributed, the points on this graph should approximately be along a line of slope 1 passing through the origin. Since units in the sample respond with unequal probabilities, the estimated empirical distribution function can be given by (Särndal, Swensson and Wretman, 1992, p. 199):

$$F^*(e_i/\sigma^* X_i^{1/2}) = \frac{\sum_{j:j\in R \text{ and } e_j/X_j^{1/2} \le e_i/X_i^{1/2}} 1/p^*(Y_j)}{\sum_{j\in R} 1/p^*(Y_j)} .$$

It should be noted that, in the above equation, the response probabilities are estimated, as opposed to the Särndal, Swensson and Wretman formula in which the selection probabilities are known.

3.2.3 Corrected ratio imputation method

This method has been developed by Rancourt, Lee and Särndal (1994) for the case of ratio imputation (model 3.1). It consists of simply multiplying imputed values obtained in the case of an ignorable response mechanism (see section 3.1) by a correction factor, C. The authors studied the behaviour of 8 such factors. In this paper, only one of these 8 factors, which is globally very good with respect to bias, will be considered:

$$C = 1 - \left(\left(\frac{\sum_{i\in O} W_i/(n-n_R)}{\sum_{i\in S} W_i/n} \right)^2 - 1 \right) \left(\hat{R}^2 - 1 \right) ,$$

where W_i is the rank of X_i, for $i\in S$, n_R is the respondent sample size and $\hat{R}$ is the estimated coefficient of correlation between X and Y based on the respondent data. This method will be denoted by RAT-C. It possesses the advantage of being simple although the correction factors are only available for model (3.1).

4. SIMULATION STUDY

In order to compare the estimators of the population mean presented in the preceding section, a simulation study was carried out. Four populations of size 1000 have first been generated according to the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \sigma X_i^{\lambda/2} \varepsilon_i , \qquad (4.1)$$

where X_i, for $i\in P$, are mutually independent and exponentially distributed random variables with mean 3 and ε_i, for $i\in P$, are mutually independent random variables, uncorrelated with X_i, with zero mean and variance 1. Two populations have ε_i distributed according to the standard normal distribution ($\varepsilon_i \to N(0,1)$) and the two other populations have ε_i distributed according to the standard exponential distribution centered at zero ($\varepsilon_i \to \text{Exp(mean} = 1) - 1$). For each of these two distributions, one population is in agreement with model (3.1), that is $\beta_0 = 0$, $\beta_1 = 1.5$, $\sigma^2 = 4.5$ and $\lambda = 1$ (ratio model) and the other population is generated according to a linear model with non-zero intercept and constant variance, that is $\beta_0 = 0.5$, $\beta_1 = 1.5$, $\sigma^2 = 13.5$ and $\lambda = 0$ (non-ratio model). The squared theoretical coefficient of correlation between X and Y is 60% for the four populations considered in this simulation study. Also, for each of these four populations, 5 response mechanisms have been tested; one in which all the units in the population have the same response probability, called UNIFORM, one in which the response probability decreases when X increases, called DEC-X, one in which the response probability increases when X increases, called INC-X, one in which the response probability decreases when Y increases, called DEC-Y, and, finally, one in which the response probability increases when Y increases, called INC-Y. For each of these 5 response mechanisms the expected response rate is 70%. For the four non-uniform response mechanisms, the response variable is generated by using the following response probability model:

$$p(Z_i) = \exp(-\exp(\alpha_0 + \alpha_1 Z_i)) , \qquad (4.2)$$

where $Z_i = X_i$ for DEC-X and INC-X, $Z_i = Y_i$ for DEC-Y and INC-Y, $\alpha_1 = 0.1$ for DEC-X and DEC-Y and $\alpha_1 = -0.1$ for INC-X and INC-Y. The parameter α_0 is determined so as to obtain an expected response rate of 70%. Note that the form of the response probability (4.2) differs from the form (3.2) used for the ML, R-T1 and R-T1M estimators.

For each of the 20 combinations of the 5 response mechanisms and the 4 populations, 1000 samples of respondents have been generated. For each of these 1000 samples of respondents, the 5 estimates of the population mean described in the preceding section have been calculated. Then, for each of the 20 combinations, the mean and the variance of the 1000 estimates have been calculated, denoted by $\bar{\mu}_I^*$ and V_I^* respectively. Finally, we calculated an estimate of the relative bias (RB) given by $[(\bar{\mu}_I^* - \mu)/\mu]$ 100% and an estimate of the standard error of the relative bias (SE) given by $(100/\mu)\,(V_I^*/1000)^{1/2}$.

Table 1 shows the results of the simulation study comparing the bias of the 5 estimators described in the preceding section. The best method with respect to the relative bias is highlighted for each situation. Not surprisingly, the ML estimator performs very well when the response mechanism is nonignorable (DEC-Y and INC-Y), the error distribution is normal and the model is in agreement with (3.1). When one or more of these hypotheses does not hold true, though, the bias can be substantial. In particular, this method is very sensitive to the validity of the model. As we cannot very often have full confidence in a model, great care must be taken before using this estimator.

For the uniform response mechanism, all the estimators perform very well, except the ML estimator for the non-ratio model. For the other ignorable response mechanisms (DEC-X and INC-X) and apart from the ML estimator, the RAT estimator always performs better than the RAT-C estimator which usually (7 cases out of 8) performs better than the R-T1 estimator which itself always performs better than the R-T1M estimator. For the nonignorable response mechanisms (DEC-Y and INC-Y), the case in which we are especially interested in this paper, the opposite tendency is observed; the R-T1M estimator always performs better than the R-T1 estimator which always performs better than the RAT-C estimator which itself always performs better than the RAT estimator.

5. DISCUSSION

In the cases where the nonresponse rate is very low or the correlation between X and Y is very high, all the estimators studied in this paper will likely have a low bias and the choice of an estimator should be made according to the simplicity criterion which favours the usual RAT estimator. This estimator is also favoured when nonresponse is ignorable.

For nonignorable response mechanisms, which may be more realistic in practice, the robust estimation method presented in this paper seems to be the best method, according to the simulation study, with respect to bias. If we have confidence in a model and the normality assumption seems to hold true, the maximum likelihood method can also be used. But, in practice, these two assumptions are often violated.

If we do not want to assume anything at all about the response mechanism, the RAT-C estimator is preferable to the other 4 estimators. It seems to be a compromise between the usual RAT estimator, which works well in the case of ignorable nonresponse, and the robust estimators R-T1 and R-T1M, which work well in the case of nonignorable nonresponse. It possesses the advantage of being simple but, on the other hand, it cannot be extended, like other methods, to more general regression situations.

The robust estimation method, as well as the maximum likelihood method, require modeling the response probability of each unit in the sample of respondents. So, instead of imputing missing values, we could weight the respondent data by the inverse of the response probabilities. Note also that models (3.1) and (3.2) could be modified to include other independent variables, if available.

The goal of this paper consisted of developing a robust estimation method against bad specification of the error distribution and, to some extent, against bad specification of the model involving the variable of interest allowing a reduction in the bias introduced by a nonignorable response mechanism. In future research, it would be interesting to evaluate simple variance estimation methods in the presence of imputed data when the robust estimation method presented in this paper is used.

ACKNOWLEDGEMENTS

The author is grateful to the Small Area and Administrative Data Division of Statistics Canada. The author would also like to thank Éric Rancourt, Nancy Forget, Phil Reiss and Geoffrey Eden for their useful comments.

REFERENCES

Dempster, A.P., Laird, N.M., and Rubin, R.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.

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.

Rancourt, É., Lee, H., and Särndal, C.-É. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse. *Survey Methodology*, 20, 137-147.

Särndal, C.-É., Swensson, B., and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New-York: Springer-Verlag.

SAS Institute Inc. (1990). *SAS/STAT User's Guide*, Volume 2, Version 6, Fourth Edition, Cary, NC: SAS Institute Inc.

Zeger, S.L., Liang, K., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

TABLE 1: RESULTS OF THE SIMULATION STUDY

RESP. MECHAN.	EST.	POPULATION							
		EXPONENTIAL				NORMAL			
		RATIO		NON-RATIO		RATIO		NON-RATIO	
		RB(%)	SE	RB(%)	SE	RB(%)	SE	RB(%)	SE
DEC-Y	RAT	-13.9	0.05	-10.6	0.06	-10.8	0.06	-7.1	0.05
	RAT-C	-7.5	0.07	-3.5	0.09	-2.3	0.08	1.1	0.09
	ML	-3.8	0.08	10.2	0.13	1.4	0.07	33.7	0.17
	R-T1	-4.3	0.14	-2.4	0.08	-1.6	0.08	-0.6	0.06
	R-T1M	-3.6	0.11	-1.1	0.09	0.2	0.09	0.5	0.07
INC-Y	RAT	5.8	0.04	6.2	0.04	6.3	0.04	6.2	0.04
	RAT-C	3.2	0.04	3.4	0.04	3.9	0.05	3.7	0.04
	ML	-12.0	0.50	-42.4	0.17	-0.1	0.06	-34.8	0.13
	R-T1	1.1	0.04	1.4	0.05	1.9	0.06	1.7	0.06
	R-T1M	-0.0	0.05	-0.1	0.06	0.4	0.07	0.3	0.07
DEC-X	RAT	-0.2	0.06	1.2	0.06	0.9	0.06	1.9	0.05
	RAT-C	4.3	0.08	6.2	0.08	5.7	0.07	7.1	0.08
	ML	6.8	0.08	15.2	0.15	8.4	0.07	31.8	0.18
	R-T1	12.5	0.20	7.2	0.08	7.3	0.08	5.9	0.06
	R-T1M	13.0	0.15	8.3	0.08	8.4	0.08	6.4	0.06
INC-X	RAT	0.1	0.05	-0.7	0.05	-0.4	0.05	-1.1	0.04
	RAT-C	-1.8	0.05	-2.6	0.05	-2.4	0.05	-2.9	0.05
	ML	-6.5	0.13	-51.7	0.24	-5.8	0.06	33.9	0.15
	R-T1	-3.5	0.05	-4.2	0.05	-4.4	0.06	-4.8	0.06
	R-T1M	-4.5	0.05	-5.2	0.06	-5.9	0.07	-6.0	0.07
UNIFORM	RAT	0.0	0.06	0.0	0.05	0.0	0.05	0.0	0.05
	RAT-C	0.1	0.06	0.1	0.06	0.1	0.06	0.0	0.06
	ML	-0.4	0.07	-24.8	0.97	0.0	0.07	-25.4	1.06
	R-T1	0.1	0.07	0.1	0.06	0.0	0.06	0.0	0.06
	R-T1M	0.1	0.07	0.1	0.06	0.0	0.07	-0.1	0.06