Survey Methodology, December 2005
Vol. 31, No. 2, pp. 227-231
Statistics Canada, Catalogue No. 12-001-XPB

On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment

Jean-François Beaumont [1]

Abstract

Nonresponse weight adjustment is commonly used to compensate for unit nonresponse in surveys. Often, a nonresponse model is postulated and design weights are adjusted by the inverse of estimated response probabilities. Typical nonresponse models are conditional on a vector of fixed auxiliary variables that are observed for every sample unit, such as variables used to construct the sampling design. In this note, we consider using data collection process variables as potential auxiliary variables. An example is the number of attempts to contact a sample unit. In our treatment, these auxiliary variables are taken to be random, even after conditioning on the selected sample, since they could change if the data collection process were repeated for a given sample. We show that this randomness introduces no bias and no additional variance component in the estimates of population totals when the nonresponse model is properly specified. Moreover, when nonresponse depends on the variables of interest, we argue that the use of data collection process variables is likely to reduce the nonresponse bias if they provide information about the variables of interest not already included in the nonresponse model and if they are associated with nonresponse. As a result, data collection process variables may well be beneficial to handle unit nonresponse. This is briefly illustrated using the Canadian Labour Force Survey.

Key Words: Nonresponse bias; Nonresponse model; Nonresponse variance; Number of attempts; Paradata; Response probability.

1. Introduction

Unit nonresponse is often handled in surveys by using a nonresponse weight adjustment method. The basic principle that is often chosen is to adjust the design weights by the inverse of estimated response probabilities (see, for example, Ekholm and Laaksonen 1991). These estimated response probabilities are obtained by postulating a model for the unknown nonresponse mechanism, which we call the nonresponse model. Key to reducing the nonresponse bias and variance as much as possible is to condition on a vector of auxiliary variables that are observed for every sample unit and that are good predictors of both nonresponse and the variables of interest (Little and Vartivarian 2005). Usually, the auxiliary variables are treated as being fixed both unconditionally and conditionally on the selected sample.

In this note, we consider using Data Collection Process (DCP) variables as potential auxiliary variables to be included in the nonresponse model. An example is the number of attempts to contact a sample unit. Such type of data is sometimes called paradata (see Couper and Lyberg 2005 for a recent reference on paradata) and has been used to deal with unit nonresponse by Holt and Elliott (1991), among others. In our treatment, contrary to Holt and Elliott (1991), DCP variables are taken to be random, even after conditioning on the selected sample, since they could change if the data collection process were repeated for a given sample.

DCP variables may be particularly useful in cross-sectional surveys where the auxiliary variables available to handle unit nonresponse are often limited to variables used to construct the sampling design. Although such design variables are not useless, they are often neither very good predictors of nonresponse nor the variables of interest. The additional information from data collection process may be welcome in these cases. In longitudinal surveys, there is a wealth of potential auxiliary variables to deal with wave nonresponse. DCP information may thus not have the same importance to compensate for wave nonresponse than the importance it has to compensate for unit nonresponse in cross-sectional surveys. However, we have yet to study this in any depth. It may turn out that, at change points, DCP variables may matter greatly.

In section 2, we introduce notation and theory concerning the effect of using random auxiliary variables in the nonresponse model when estimating population totals. This issue of the randomness of DCP auxiliary variables was raised and debated at Statistics Canada's Advisory Committee on Statistical Methods after the paper by Alavi and Beaumont (2004) was presented. The goal of section 2 is thus to shed some light on this issue. The use of DCP variables to adjust design weights for nonresponse is briefly illustrated in section 3, using the Canadian Labour Force

1. Jean-François Beaumont, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@statcan.ca.

Survey (CLFS). The last section, section 4, contains a brief summary of the paper.

2. Theory

Let us assume that we are interested in estimating the population total $t_y = \sum_{k \in U} y_k$ of a variable of interest y for a certain fixed population U of size N. From this population, a random sample s of size n is selected according to a probability sampling design $p(s \mid \mathbf{D})$, where $\mathbf{D}$ is a $N-$row matrix containing $\mathbf{d}'_k$ in its k^{th} row and $\mathbf{d}$ is the vector of design variables. Let also assume that, in the absence of nonresponse, we would use the Horvitz-Thompson estimator $\hat{t}_y = \sum_{k \in s} w_k y_k$, where $w_k = 1/\pi_k$ is the design weight of unit k and $\pi_k = P(k \in s)$ is its selection probability.

Usually, due to a number of reasons, unit nonresponse occurs so that the variable y is only observed for a subset s_r of s, the respondents. Along with s_r, a random vector $\mathbf{z}$ of DCP variables is also observed for every sample unit according to a joint mechanism $\# q(\mathbf{Z}_s,\ s_r \mid s,\ \mathbf{Y},\ \mathbf{D},\ \mathbf{X})$. As mentioned in the introduction, the number of attempts to contact a sample unit is an example of a DCP variable. The vector $\mathbf{z}$ of DCP variables and the set of respondents s_r are random after conditioning on the selected sample since these quantities would likely take different values if the data collection process were repeated for a given sample. The quantity $\mathbf{Z}_s$ is a $n-$row matrix containing $\mathbf{z}'_k$ in its k^{th} row, $\mathbf{Y}$ is a $N-$element vector containing y_k in its k^{th} element and $\mathbf{X}$ is a $N-$row matrix containing $\mathbf{x}'_k$ in its k^{th} row. The vector $\mathbf{x}$ is a vector of additional fixed auxiliary variables. For instance, these auxiliary variables could come from an administrative file or, in a longitudinal survey, they could be the variables of interest observed at the previous wave. As a result, the vector $\mathbf{x}$ may not be available for nonsample units. Table 1 summarizes the availability of the different types of variables for the respondents, nonrespondents and nonsample units.

Table 1
Availability of Variables

	y	z	x	d
Respondents: s_r	YES	YES	YES	YES
Nonrespondents: $s - s_r$	NO	YES	YES	YES
Nonsample units: $U - s$	NO	NO *	YES **	YES

* The vector **z** is not even defined for nonsample units.
** The vector **x** may not always be available for nonsample units.

The joint mechanism $\# q(\mathbf{Z}_s,\ s_r \mid s,\ \mathbf{Y},\ \mathbf{D},\ \mathbf{X})$ can be factorized into two distinct random mechanisms: i) $\#(\mathbf{Z}_s \mid s,\ \mathbf{Y},\ \mathbf{D},\ \mathbf{X})$ and ii) $q(s_r \mid s,\ \mathbf{Y},\ \mathbf{D},\ \mathbf{X},\ \mathbf{Z}_s)$. The former is called the DCP mechanism while the latter is called the nonresponse mechanism. This factorization will be useful later to obtain properties of our nonresponse-weight-adjusted estimator defined in equation (2.2) below. We assume that

$$q(s_r \mid s,\ \mathbf{Y},\ \mathbf{D},\ \mathbf{X},\ \mathbf{Z}_s) = q(s_r \mid s,\ \mathbf{D}_s,\ \mathbf{X}_s,\ \mathbf{Z}_s), \quad (2.1)$$

where $\mathbf{D}_s$ and $\mathbf{X}_s$ are the sample portions of $\mathbf{D}$ and $\mathbf{X}$ respectively. This assumption implies that the nonresponse mechanism is independent of (or unconfounded with) $\mathbf{Y}$, after conditioning on s, $\mathbf{D}_s$, $\mathbf{X}_s$ and $\mathbf{Z}_s$, and that the data are missing at random. However, we make no explicit simplifying assumption about the DCP mechanism so that it may well depend on $\mathbf{Y}$, even after conditioning on s, $\mathbf{D}$ and $\mathbf{X}$.

To compensate for unit nonresponse, we consider the nonresponse-weight-adjusted estimator

$$\hat{t}_y^{\text{NWA}} = \sum_{k \in s_r} \frac{w_k}{p_k(\hat{\boldsymbol{\alpha}})} y_k, \quad (2.2)$$

where $p_k(\boldsymbol{\alpha}) = P(k \in s_r \mid s,\ \mathbf{D}_s,\ \mathbf{X}_s,\ \mathbf{Z}_s;\boldsymbol{\alpha})$ is the conditional response probability for a unit $k \in s$ and $\hat{\boldsymbol{\alpha}}$ is an estimator of the vector of unknown nonresponse model parameters $\boldsymbol{\alpha}$. Note that a nonresponse model is a set of assumptions about the unknown nonresponse mechanism $q(s_r \mid s, \mathbf{Y},\ \mathbf{D},\ \mathbf{X},\ \mathbf{Z}_s)$; one of them being assumption (2.1). We assume that $\hat{\boldsymbol{\alpha}}$ is implicitly defined by the equation $\mathbf{U}_1(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$, where $\mathbf{U}_1(.)$ is a vector of $q-$unbiased estimating functions for $\boldsymbol{\alpha}$; that is, $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) \mid s,\ \mathbf{Y},\ \mathbf{D},\ \mathbf{X},\ \mathbf{Z}_s\} = \mathbf{0}$. Therefore, $\mathbf{U}_1(.)$ is also $p\# q-$unbiased for $\boldsymbol{\alpha}$. In the remaining of the paper, we remove everywhere the conditioning on $\mathbf{Y}$, $\mathbf{D}$ and $\mathbf{X}$ when taking expectations and variances since these vectors are always treated as being fixed. For instance, we will write $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) \mid s,\ \mathbf{Z}_s\} = \mathbf{0}$ instead of $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) \mid s, \mathbf{Y}, \mathbf{D},\ \mathbf{X},\ \mathbf{Z}_s\} = 0$. This will simplify considerably the notation.

Note that the nonresponse-weight-adjusted estimator (2.2) is implicitly defined by the equation

$$U_2(\hat{\boldsymbol{\alpha}},\ \hat{t}_y^{\text{NWA}}) = \hat{t}_y^{\text{NWA}} - \sum_{k \in s_r} \frac{w_k}{p_k(\hat{\boldsymbol{\alpha}})} y_k = 0. \quad (2.3)$$

If the nonresponse model is correctly specified and, in particular, if assumption (2.1) is satisfied, then the estimating function $U_2(.,.)$ is $p\# q-$unbiased for t_y; that is, $\mathrm{E}_{p\# q}\{U_2(\boldsymbol{\alpha}, t_y)\} = 0$. To make assumption (2.1) as plausible as possible, it is important that the nonresponse model be conditional on design, auxiliary and DCP variables that are well correlated with y, provided that these variables are also associated with nonresponse. This recommendation should be useful to control the magnitude of the nonresponse bias, which may be unavoidable in real surveys.

This is also in line with the recommendation given in Little and Vartivarian (2005). Therefore, if DCP variables contain information about y above the information already contained in $\mathbf{d}$ and $\mathbf{x}$, then the use of DCP variables may be useful to reduce the nonresponse bias if they are associated with nonresponse.

Now, let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', t_y)'$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}', \hat{t}_y^{\text{NWA}})'$ and $\mathbf{U}(\tilde{\boldsymbol{\theta}}) = \{\mathbf{U}_1'(\tilde{\boldsymbol{\alpha}}), U_2(\tilde{\boldsymbol{\alpha}}, \tilde{t}_y)\}'$, for some vector $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}', \tilde{t}_y)'$. As noted above, $\hat{\boldsymbol{\theta}}$ is implicitly defined by the equation $\mathbf{U}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and the estimating function $\mathbf{U}(.)$ is $p\#q-$unbiased for $\boldsymbol{\theta}$ since $\mathbf{E}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$. Using a first-order Taylor approximation (see Binder 1983), we have $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} - \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{U}(\boldsymbol{\theta})$, where $\mathbf{H}(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{p\#q}\{\partial \mathbf{U}(\tilde{\boldsymbol{\theta}}) / \partial \tilde{\boldsymbol{\theta}}'\}$. The matrix $\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}$ is thus given by

$$\{\mathbf{H}(\boldsymbol{\theta})\}^{-1} = \begin{pmatrix} \{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & \mathbf{0} \\ -\mathbf{H}_{21}(\boldsymbol{\theta})\{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & 1 \end{pmatrix}, \tag{2.4}$$

where $\mathbf{H}_{i1}(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{p\#q}(\partial \mathbf{U}_i(\tilde{\boldsymbol{\theta}}))/(\partial \tilde{\boldsymbol{\alpha}}')$, for $i = 1, 2$. Using conditions similar to those of Binder (1983), $\hat{\boldsymbol{\theta}}$ is asymptotically normal and asymptotically $p\#q-$unbiased for $\boldsymbol{\theta}$. As a result, $\hat{t}_y^{\text{NWA}}$ is asymptotically normal and asymptotically $p\#q-$unbiased for t_y. Therefore, using DCP variables in the nonresponse model does not introduce any bias in the nonresponse-weight-adjusted estimator $\hat{t}_y^{\text{NWA}}$ provided that the nonresponse model (specification of $q(s_r \mid s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s)$ and assumption 2.1) holds. Also, if the true unknown nonresponse mechanism depends on the sample portion of $\mathbf{Y}$, $\mathbf{Y}_s$, after conditioning on s, $\mathbf{D}_s$ and $\mathbf{X}_s$, then conditioning on a vector $\mathbf{z}$ of DCP variables is likely to reduce the nonresponse bias if the DCP mechanism depends on $\mathbf{Y}_s$, after conditioning on s, $\mathbf{D}_s$ and $\mathbf{X}_s$, which means that the DCP variables contain information about y not already contained in $\mathbf{d}$ and $\mathbf{x}$.

Continuing our Taylor linearization, and using the fact that

$$\begin{aligned}\mathbf{V}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{V}_p\mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) \mid s\} \\ + \mathbf{E}_p\mathbf{V}_{\#}\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\} \\ + \mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\},\end{aligned}$$

the $p\#q-$variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, $\mathbf{V}_{p\#q}(\hat{\boldsymbol{\theta}})$, is approximated by

$$\begin{aligned}\dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}}) = \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{V}_p\mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) \mid s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1} \\ + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_p\mathbf{V}_{\#}\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1} \\ + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}.\end{aligned} \tag{2.5}$$

The first term on the right-hand side of equation (2.5) is called the sampling variance of $\hat{\boldsymbol{\theta}}$, the second term is called the DCP variance of $\hat{\boldsymbol{\theta}}$ and the third term is called the nonresponse variance of $\hat{\boldsymbol{\theta}}$. The variance $\mathrm{V}_{p\#q}(\hat{t}_y^{\text{NWA}})$ is approximated by the value in the last row and in the last column of equation (2.5). Using expression (2.4) and the fact that $\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\} = (\mathbf{0}', t_y - \hat{t}_y)'$, the approximate variance (2.5) reduces to

$$\begin{aligned}\dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{V}_p(\hat{t}_y) \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} \\ + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}.\end{aligned} \tag{2.6}$$

The second matrix on the right-hand side of equation (2.6) corresponds to the DCP variance of $\hat{\boldsymbol{\theta}}$ and contains 0 for all its elements. Therefore, using random auxiliary (DCP) variables in the nonresponse model does not introduce any additional term of variance, as opposed to using only fixed auxiliary variables, when the nonresponse model is properly specified. Since DCP variables are likely to reduce the nonresponse bias if they are associated with y, then it seems beneficial to take advantage of them when handling unit nonresponse through a weight adjustment. Also, as pointed out by Little and Vartivarian (2005), adding auxiliary variables in the nonresponse model that are associated with y tends to reduce the nonresponse variance. The mean squared error can therefore be reduced on both counts.

A more detailed expression for the nonresponse variance term in equation (2.6) as well as a sampling and a nonresponse variance estimator can be obtained similarly as in Beaumont (2005). Beaumont (2005) also discusses the effect of estimating the nonresponse model parameters on the variance of an estimator of a population total.

3. The Example of the Canadian Labour Force Survey

The goal of this example is not to provide every detail of the analysis that was conducted on the Canadian Labour Force Survey (CLFS) data but simply to describe some issues related to the choice of the nonresponse model and to the estimation of response probabilities. With these points in mind, we then go on to discuss the main conclusions that were reached. Greater detail about the results of the investigations in the CLFS, implementation of the new method and a comparison with the previous method can be found in Alavi and Beaumont (2004).

The CLFS is a monthly survey with a stratified multi-stage sampling design (Gambino, Singh, Dufour, Kennedy and Lindeyer 1998). The information used to construct the sampling design and to draw a sample of dwellings is essentially geographic. The sample is divided into six representtative rotation groups and each sampled dwelling stays in the sample for six consecutive months. One rotation group contains dwellings for which the members are interviewed for the first time; another rotation group contains dwellings for which the members are interviewed for the second time and so on. Thus, for five rotation groups out of six, the sampled dwellings are common from one month to the next. Computer-assisted interviews are used to collect the survey information for every person in the selected households. With computer-assisted interviews, a large amount of DCP information is obtained for both responding and non-responding households.

A logistic nonresponse model has been considered to model the unknown nonresponse mechanism $q(s_r \mid s, \mathbf{D}_s, \mathbf{Z}_s)$. With this model, the unknown response probability for household k is expressed as $p_k(\boldsymbol{\alpha}) = \{1 + \exp(-\boldsymbol{\alpha}' (\mathbf{zd})_k)\}^{-1}$ and sampled households are assumed to respond independently of one another. The vector $\mathbf{zd}$ is a vector that contains DCP variables $\mathbf{z}$, fixed design variables $\mathbf{d}$ as well as interactions between these two types of variables. No additional vector $\mathbf{x}$ of auxiliary variables was available. Two DCP variables were used: the number of attempts to contact a sampled household, which was divided into five categories, and the time of the last attempt, which was also divided into five categories. The design variables used were mainly geographic and also included the rotation group indicator. Due to potential interviewer and clustering effects, the above model may not be entirely realistic. It was used for its simplicity and because it appeared reasonable and an improvement over the previous method. Also, the estimated response probabilities resulting from this model were used only to provide a score and were not used directly to adjust design weights, as described below in this section.

The unknown vector $\boldsymbol{\alpha}$ was estimated by the maximum likelihood method using the $q-$unbiased estimating function

$$\mathbf{U}_1(\boldsymbol{\alpha}) = \sum_{k \in s} \{r_k - p_k(\boldsymbol{\alpha})\}(\mathbf{zd})_k, \qquad (3.1)$$

where $r_k = 1$, if $k \in s_r$, and $r_k = 0$, otherwise. Note that a design-weighted estimating function was not considered. This follows the practice recommended in Little and Vartivarian (2003) and can be justified by noting that the interest is in modelling the nonresponse mechanism only for sampled households $k \in s$ (not for the whole population) and that this mechanism is conditional on *s*. Also, the DCP variables are not even defined outside the sample. The use of design weights does thus not make sense in this context and increases the variance of $\hat{\boldsymbol{\alpha}}$ if the nonresponse model is correctly specified. Also, it is not clear that using a design-weighted estimating function would systematically bring robustness in this case. However, note that we do not ignore design information since it is included in the nonresponse model. This can be paralleled to the recommendation of including design information in imputation models (see, for example, Rubin 1996).

Stepwise logistic regression was performed for several months in order to determine appropriate design and DCP variables to be included in the final nonresponse model. In all months considered, the variable ‘number of attempts’ was the first to enter in the model and thus the most useful for explaining nonresponse. This variable was also correlated with the main variables of interest ‘employment’ and ‘unemployment’. For instance, people belonging to respondent households with a large number of attempts, i.e. those that are difficult to reach, had a tendency to be more often employed (see Alavi and Beaumont 2004). Households with a large number of attempts had also a tendency to be nonrespondents. Therefore, it seems appropriate to give a larger weight adjustment to the responding households for which the number of attempts is large since their propensity to respond is lower and they are more likely to have characteristics similar to the nonrespondents.

The final nonresponse model chosen fit reasonably well the CLFS data in most months considered, according to the Hosmer-Lemeshow goodness-of-fit test. Nevertheless, the score method of Little (1986) was used to obtain some robustness against undetected model failures. The above logistic nonresponse model was first used to obtain an estimated response probability for every sampled household and then the sample was divided into about 50 homogeneous classes with respect to this estimated response probability using the clustering algorithm implemented in the procedure FASTCLUS of SAS. This large number of classes was possible given the large CLFS sample size. It was chosen so as to reduce the nonresponse bias not only at the population level but also for smaller domains. The nonresponse weight adjustment for a responding household *k* within a given class *c* was simply computed as the inverse of the unweighted response rate within class *c*. A threshold on the nonresponse weight adjustment was set to 2.5 to control the nonresponse variance of the nonresponse-weight-adjusted estimator. When needed, the application of this threshold was necessary only for a very small number of classes. These were the classes with the smallest estimated response probabilities. Without this threshold, nonresponse weight adjustments around 4 could occasionally be observed.

Another nonresponse model was considered in which the response probability for a household *k* is modelled as the product of the probability that household *k* be contacted, times the probability that this household respond, given it is contacted. The latter two probabilities were modelled separately. Although this model seems to be a better approximation of reality and gave slightly better results in the sense that it better explained nonresponse, the gains were not deemed sufficient to add this complexity in the nonresponse adjustment method. It may deserve further study.

4. Conclusion

An important contribution of this paper is that DCP information must be treated as being random when used in a nonresponse model. We then have shown that the use of such information to handle unit nonresponse through a weight adjustment does not introduce any bias and that there is no additional variance component in the estimates of population totals when the nonresponse model is properly specified. Moreover, we have argued that if DCP information is associated with the variables of interest and with nonresponse, then its use is likely to reduce the nonresponse bias when the nonresponse mechanism depends directly on the variables of interest. We have also illustrated through the CLFS example that such information can be useful for dealing with unit nonresponse in a major survey.

The full response estimator that we have considered is the Horvitz-Thompson estimator. Our conclusions would have remained the same had we used instead a generalized regression estimator. We have used the Horvitz-Thompson estimator for its simplicity and because it was sufficient to show our main point.

Acknowledgements

I would like to thank the members of Statistics Canada's Advisory Committee on Statistical Methods for raising issues with the application of the proposed method in the Labour Force Survey and, particularly, J.N.K. Rao and Chris Skinner for useful discussions following the presentation to the Committee. I would also like to sincerely thank the Associate Editor for his comments and suggestions. They were very useful and improved the clarity of the paper. Finally, I am much indebted to Asma Alavi and Cynthia Bocci of Statistics Canada for preparing computer programs used to analyse Labour Force Survey data.

References

Alavi, A., and Beaumont, J.-F. (2004). Nonresponse adjustment plans for the Labour Force Survey. Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, May 2-3, 2004.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, Series B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Couper, M., and Lyberg, L. (2005). The use of paradata in survey research. *Bulletin of the International Statistical Institute* (to appear).

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 7, 325-337.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue number 71-526.

Holt, D., and Elliott, D. (1991). Methods of weighting for unit non-response. *The Statistician*, 40, 333-342.

Little, R.J. (1986). Survey nonresponse adjustment for estimate of means. *International Statistical Review*, 54, 139-157.

Little, R.J., and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology* , 31, 161-168.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.