

SSC CASE STUDY 2002 HANDLING MISSING DATA IN THE 1994 NATIONAL POPULATION HEALTH SURVEY

Francisco Aguirre and Tao Sun¹

ABSTRACT

This case study analyzes a sub-sample of the 1994 National Population Health Survey with missing responses. Various plots were used to assess the relationship of the response (Health Status) with the predictors. A logistic regression was fitted to assess the mechanism of non-response which was found to be missing at random (MAR). Three multiple imputation tools namely, TRANSCAN in S-Plus, PROC MI in SAS, and an ad-hoc multiple imputation algorithm were used to impute the missing responses. A linear model was applied to the multiple imputed data sets and the estimated coefficients were compared against the coefficients obtained from a linear model with case-wise deletion of missing values. It was concluded that TRANSCAN yielded better results than PROC MI and the ad-hoc on this particular data set. The comparison also showed how the exclusion of the missing observations from the analysis caused a significant bias in the coefficient of the main predictor that was related to the non-response mechanism.

KEY WORDS: Missing at random (MAR); Missing data; Multiple imputation; National Population Health Survey.

RÉSUMÉ

Cette étude de cas analyse un sous-échantillon de l'enquête de santé de population du national 1994 avec des réponses manquantes. De divers traçages ont été utilisés aux ânes le rapport de la réponse (état de santé) avec les prédiseurs. Une régression logistique a été adaptée aux ânes le mécanisme de la non réaction qui est avérée manquante au hasard (MARS). Trois outils multiples d'imputation notamment, TRANSCAN dans S-Plus, PROC MI dans SAS, et un algorithme multiple ad-hoc d'imputation ont été utilisés pour imputer les réponses manquantes. Un modèle linéaire a été appliqué aux Modem imputés multiples et les coefficients estimés ont été comparés contre les coefficients obtenus à partir d'un modèle linéaire à la suppression cas-sage des valeurs manquantes. On l'a conclu que TRANSCAN a donné de meilleurs résultats que PROC MI et l'ad-hoc sur ce Modem particulier. La comparaison a également montré comment l'exclusion des observations manquantes de l'analyse a causé une polarisation significative dans le coefficient du prédiseur principal qui a été lié au mécanisme de non réaction.

MOTS CLÉS : Données manquantes, Enquête nationale de santé de population, imputation multiple, manquer au hasard (MARS).

1. INTRODUCTION

Survey non-response in public health research often poses a challenge to the analysts. When missing responses are not handled properly, less efficient and biased estimates are likely to be obtained, especially when complete-data methods with case-wise deletion of missing values are used. Multiple imputation methods have been widely advertised in the literature as the most adequate approach to address missing data vis-à-vis single imputation methods. Nowadays most advanced statistical software packages such as S-plus and SAS offer routines to conduct multiple imputations easily, due to the power and flexibility of personal computers. This case study uses a

sub-sample of the 1994 National Population Health Survey (NPHS) with simulated non-response. The objectives of this case study are three, namely: To assess the response mechanism present in the data; compare the results obtained from three different multiple imputation software routines applied to it, and to assess the main relationships between health status and the health predictor variables.

1.1 Background

Survey statisticians generally consider three types of response mechanisms related to missing data: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Informally speaking,

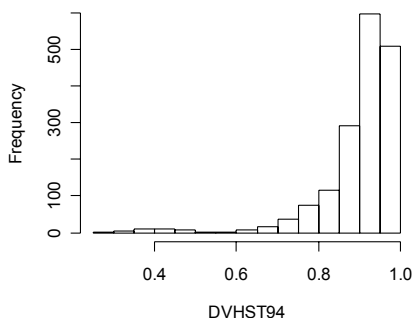
¹ Francisco J. Aguirre and Tao Sun, Department of Mathematics and Statistics, York University, faguirre@mathstat.yorku.ca, tsun@mathstat.yorku.ca.

a MCAR response mechanism is one for which the probability of response does not depend on the variable of interest nor any other particular variable. A response mechanism is MAR if the probability of response depends on some auxiliary variables but not on the variable of interest; and a NMAR response mechanism is a mechanism where the probability of response depends on the variable of interest and potentially on other variables that are not observed.

1.2 Description of the data.

This case study uses a sub-sample of the 1994 National Population Health Survey (NPHS). Missing values were included to simulate non response, however the actual “missing” data values in the data sample were removed for this case study although they are, in reality, present in the public use micro-data files. The National Population Health Survey (NPHS) used the Labour Force Survey sampling frame to draw the initial sample of approximately 20,000 households. The survey is designed to collect information on the health of the Canadian population and related socio-demographic information. The questionnaires include content related to health status, use of health services, determinants of health, a health index, chronic conditions and activity restrictions. The use of health services is probed through visits to health care providers, both traditional and non-traditional, and the use of drugs and other medications. Health determinants include smoking, alcohol use and physical activity. As well, a section on self-care has also been included in this cycle. The socio-demographic information includes age, sex, education, ethnicity, household income and labour force status. The data represent persons, aged 20-65, living in a private household in the Prairie Provinces. (Pregnant women were excluded in this analysis.)

Figure 1 – Histogram of DVHST94

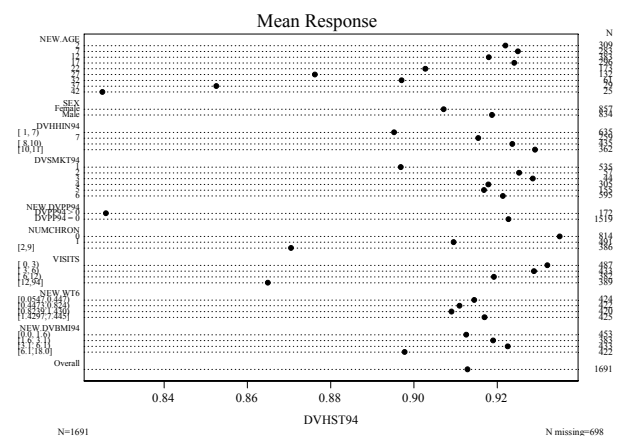


2. PRELIMINARY ANALYSIS

There are a total of 2,389 observations from which 698 have missing responses for health status. This is measured with either the general health question (GH_Q1) or the Health Utilities Index (DVHST94). GH_Q1 is an ordinal variable that takes values from 1 to 5, whereas DVHST94 is a continuous variable that takes values between zero and one. We chose the latter as our dependant variable in order to fit a linear regression model. Figure 1 exhibits a histogram of the response which appears to be highly skewed to the left.

The predictors are age in grouped cohorts (AGEGRP), the respondents’ gender (SEX), the derived total household income from all sources in the past 12 months (DVHHIN94), the derived body mass index (DVBMI94), the derived smoker type (DVSMKT94), the derived predicted probability of depression (DVPP94), the total number of chronic conditions diagnosed to the respondent out of 20 possible conditions (NUMCHRON), the number of visits or consultations made to a health care specialist in the last 12 months (VISITS), and the sampling weights (WT6). There were 1,097 males and 1,292 females in the sample (45.9% and 54.1% respectively). Individual analysis of the predictor variables revealed a small number of missing values in the predictor DVBMI94. All of these happened to be males, therefore were imputed with the mean value for males given the small number of missing values (3.9% of the male sample). Given that the relation of this variable with the response is not monotonic (i.e values between 20 and 24 are consider normal and any significant deviation from this range either above or below is related to poor health) it was recoded into NEW.DVBMI94 as the absolute distance from the nominal mid range value (22). The Variable AGEGRP was recoded into NEW.AGE taking the mid point of each age cohort centered around 20 to avoid a negative intercept given that the lowest age group starts from 20. Also, DVPP94 was

Figure 2 – Visualizing the relationship between the predictors and the response



recoded as a dichotomous variable to differentiate the cases with zero probability (91% of the observations) from the cases with strictly positive probability. The sampling weights were also recoded as NEW.WT6, having them rescaled so that they add up to the total sample size.

Figure 2 gives a basic idea of the relationship between the predictors and the response. This plot was obtained in S-Plus using a special summary function from the Hmisc library developed by Frank E. Harrell. The plot shows that the variables NEW.AGE, NUMCHRON and VISITS have a negative relationship with the response which appears to be linear to some degree; in the other hand, DVHHIN94 appears to have a positive relationship with the response.

In order to further assess the strength of the relationship between the predictors and the response, the non-monotonic (quadratic in ranks) generalization of the Spearman rank correlation coefficient was computed and plotted in Figure 3 using a function also provided in the Hmisc library. The computed correlations revealed a weak relationship between each individual covariate and the response, being probability of depression, the number of chronic conditions, the number of visits and the age cohort the most important in descending order.

3. ASSESING THE RESPONSE MECHANISM

A logistic model was fitted to the data to assess whether the probability of non response depended on any of the predictors. For this purpose a new indicator variable called IS.Y.MISS was created to differentiate the missing responses from the observed responses. The model

Figure 3 – Assessing the strength of the relationship between the predictors and the response

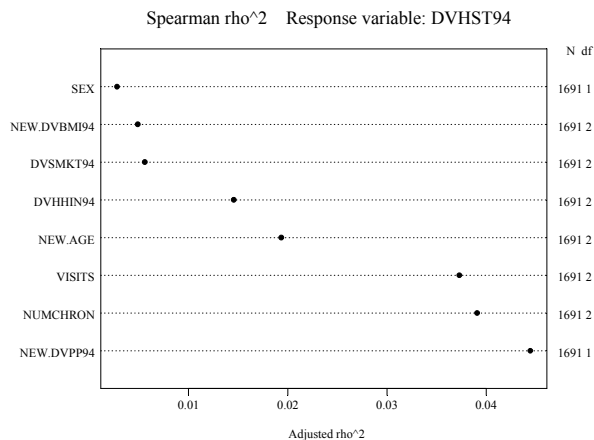


Table 1 – Output from the logistic regression model

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.058793	0.367083	-13.781	< 2e-16

NEW.AGE	0.181625	0.007524	24.140	< 2e-16

SEXMa1e	-0.847947	0.131475	-6.450	1.12e-10

DVHHIN94	0.047828	0.028768	1.787	0.0740
DVSMKT94	-0.015131	0.031662	-0.478	0.6327
NEW.DVPP94 = 0	0.233188	0.226732	1.028	0.3037
NUMCHRON	-0.087992	0.048783	-1.804	0.0713
VISITS	0.012483	0.008563	1.902	0.0572
NEW.WT6	-0.043935	0.077407	-0.568	0.5703
NEW.DVBM194	-0.015622	0.017299	-0.903	0.3665

revealed that the missing values depended on the age cohort and the gender of the respondent with a significance level of 95%. Also, the variables VISITS, NUMCHRON and DVHHIN94 appeared to be marginally significant in the model (See Table 1).

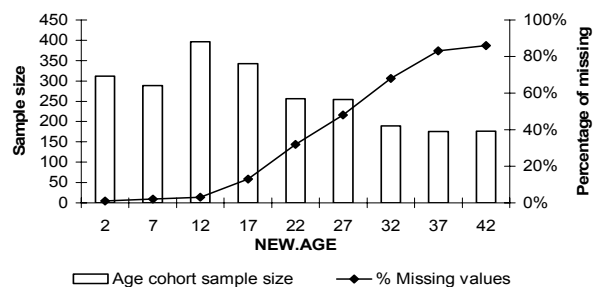
From these results we were able to conclude that the response mechanism was not MCAR but MAR depending mostly on the NEW.AGE predictor. (See figure 4.)

4. MULTIPLE IMPUTATION

4.1 Brief review on multiple imputation.

The paper “Multiple Imputation After 18+ Years” (Rubin 1996) provides a clear overview of multiple imputation. The basic idea, first proposed by Rubin (1977), is to simulate several sets of plausible values for the missing observations. Each set of imputations is used to create a complete data set, each of which is analyzed using standard complete-data methods. A fundamental aspect of multiple imputation is repeated imputation (Rubin 1987). The estimated statistics obtained from these analyses are combined together to form one repeated-imputation inference that appropriately accounts for the uncertainty due to the missing values.

Figure 4 – Relation between the missing responses and the age cohort



Assuming the values of the estimates obtained from the complete-data analyses are $\hat{\beta}_{*1}, \dots, \hat{\beta}_{*m}$ with variance-covariance U_{*1}, \dots, U_{*m} , the repeated-imputation estimates are calculated as:

$$\bar{\beta} = \sum_{i=1}^m \hat{\beta}_{*i} / m$$

The associated variance-covariance of $\bar{\beta}$ is,

$$V = \bar{U} + \frac{m+1}{m} B$$

Where,

$$\bar{U} = \sum_{i=1}^m U_{*i} / m$$

is the within-imputation variability and

$$B = \sum_{i=1}^m (\hat{\beta}_{*i} - \bar{\beta})(\hat{\beta}_{*i} - \bar{\beta})' / (m-1)$$

is the between-imputation variability.

The derivation of these expressions follows from the Bayesian perspective of treating the estimates as random variables with normal conditional distributions (For details see Rubin 1987).

For this case study, a simple linear regression model was fitted to each of the m complete data sets; therefore the estimates of interest are the regression coefficients, and their standard errors. Also the mean R-squared statistic was calculated by combining the individual values obtained from each analysis.

4.2 Multiple imputation routines

For this case study we used three different multiple imputation routines, two of which are readily available in two popular statistical software packages, namely: the *Transcan* function from the *Hmisc* library in S-Plus developed by Frank Harrell and the PROC MI procedure available in SAS. We also developed an ad-hoc routine to carry out random hot-deck multiple imputation using the bootstrap.

4.2.1 Transcan from Hmisc library in S-Plus.

Transcan imputes the missing values with an algorithm that closely resembles Rubin's Bayesian bootstrap. The algorithm first transforms continuous and categorical variables to have maximum correlation with the best linear combination of the predictors. Assuming we have n

missing responses and r observed responses in the data, the imputation works as follows: A multiple regression model is fitted to the r observed records yielding r residuals. Then, a sample of size n with replacement is drawn from these r residuals. The imputed values are obtained by fitting the model to the predictor variables of the missing responses and then adding to the fitted values the residuals from the previous step. By repeating this process m times, we obtain m multiple imputed complete data sets. The imputed data sets are analyzed with the function *fit.mult.impute* that provides valid estimates for the coefficients.

4.2.2 PROC MI in SAS.

This procedure assumes that the data follows a multivariate normal distribution. **PROC MI** generates five imputation values by default for each missing value using MCMC (Markov Chain Monte Carlo), which repeatedly simulates the distribution from which the imputed values are drawn. The Initial values for the MCMC are determined via the EM algorithm. As for the analysis step, the linear model is fitted using **PROC REG** which simultaneously fits the model to the five complete data sets. Then, **PROC MIANALYZE** is used to produce valid statistics for the coefficients.

4.2.3 Hot-deck multiple imputation with bootstrap.

An ad-hoc procedure was programmed in S-Plus to impute the missing responses using the observed responses. The observations were randomly selected with replacement, using probabilities according to the sampling weights. Thus, we run 1000 bootstrap complete-data samples which were subsequently analyzed with a linear model. The results from these models were pooled to produce valid estimates of the coefficients according to the formulas presented before.

4.3 Comparison of the routines.

Table 2 shows the estimated coefficients of the linear regression model fitted to the multiple imputed data sets obtained with each of the routines along with their standard errors (The significant coefficients at a 95% confidence level display an asterisk). The last column also shows the coefficients obtained by fitting a linear model with case-wise deletion of the missing observations. Based on the R-Squared statistic, we concluded that *Transcan* provided the best multiple imputation for this particular data set. The main advantage of this routine is that it does not require the assumption of normality or symmetry of residuals and it can take into account shrinkage to prevent over fitting.

Table 2 – Comparison of the estimated regression coefficients

	S-plus TRANSCAN	SAS PROC MI	Random hot deck bootstrap	Non-missing only
INTERCEPT	0.8495 (0.0135) *	0.9281 (0.01) *	0.8711 (0.0128) *	0.861 (0.012) *
NEW.AGE	-0.0039 (0.0004) *	-0.0016 (0.0004) *	-0.0006 (0.0002) *	-0.0013 (0.0003) *
SEX (Male=1)	0.0045 (0.0045)	0.0023 (0.0045)	0.0031 (0.0055)	0.0037 (0.0049)
DVHHIN94	0.0083 (0.0016) *	0.0061 (0.0012) *	0.0029 (0.0007) *	0.0051 (0.0011) *
NEW.DVBM194	-0.0001 (0.0007)	-0.0005 (0.0008)	-0.0006 (0.0005)	-0.0007 (0.0007)
DVSMKT94	0.0012 (0.0014)	0.0009 (0.0013)	0.0019 (0.0008) *	0.0012 (0.0012)
NEW.DVPP94(=0)	0.0904 (0.0085) *	0.0717 (0.0092) *	0.0531 (0.0089) *	0.0686 (0.0081) *
NUMCHRON	-0.0174 (0.0022) *	-0.0123 (0.0023) *	-0.0079 (0.0013) *	-0.013 (0.0021) *
VISITS	-0.0026 (0.0003) *	-0.0023 (0.0003) *	-0.0017 (0.0002) *	-0.0023 (0.0003) *
Mean R-square	0.33	0.193	0.093	0.183

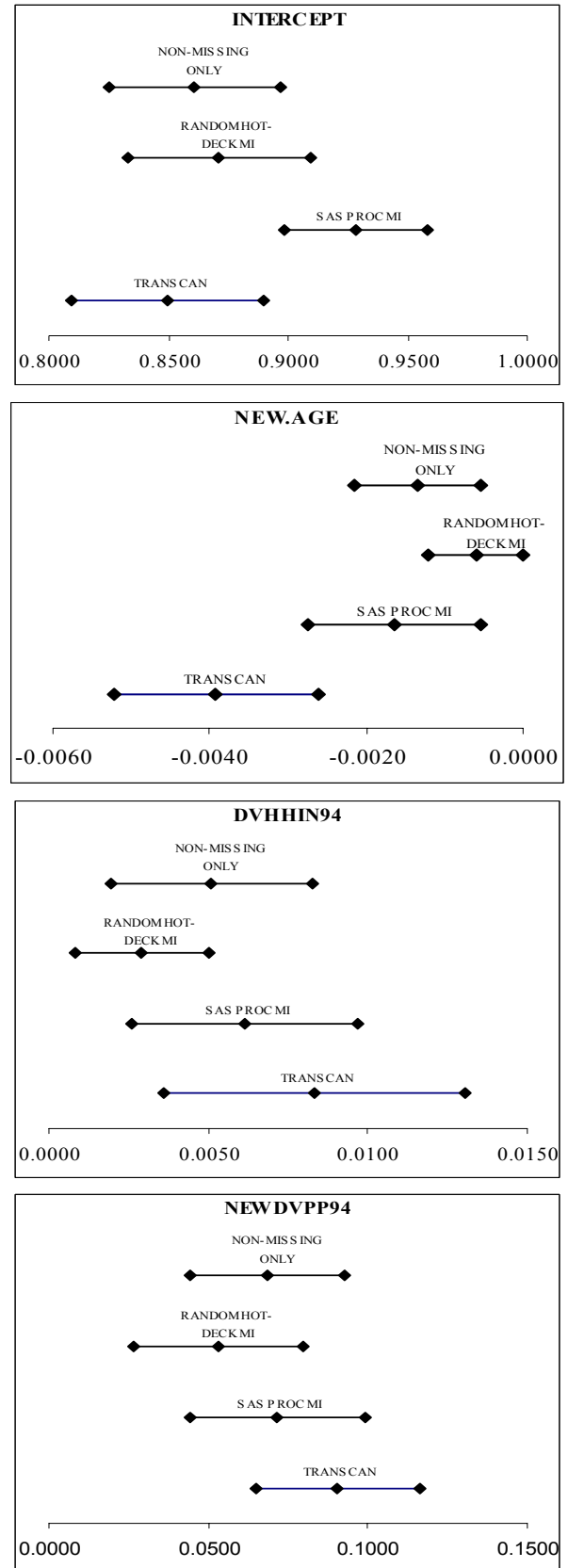
The only drawback of this procedure is that instead of using different estimates of coefficients for each imputation, it uses the same model to impute all the missing values.

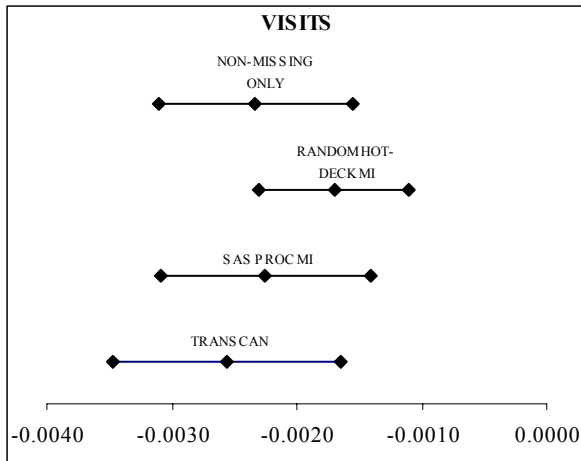
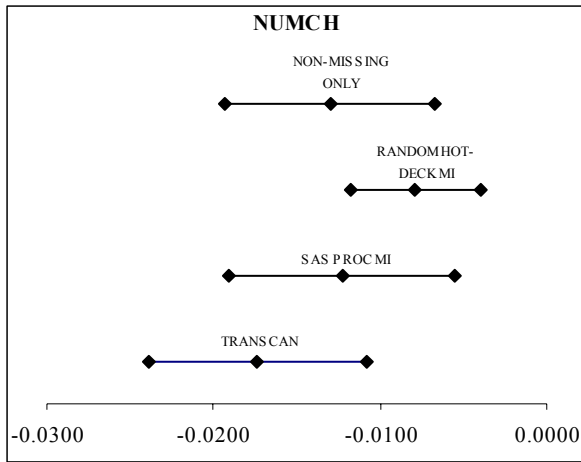
The multiple imputations obtained with PROC MI did not provide a significant improvement in the fit of the model vis-à-vis the model with case-wise deletion of missing values. Most of the estimates were very similar to the ones obtained by fitting the non-missing cases only. This routine did not perform well due to the violation of the underlying assumption of normality. The preliminary analysis revealed that the response had a highly skewed distribution.

The ad-hoc random hot deck imputation routine yielded unsatisfactory results. The fit of the model turned out to be worse than the model fitted with the non-missing cases only. The main flaw of this approach lies in the fact that the algorithm does not make use of the information provided by the predictors. Given that the response mechanism is MAR, this information is crucial and the algorithm did not include it.

The plots in Figure 5 allow to easily compare the 95% confidence intervals of the significant coefficients for each of the multiple imputation routines. The first conclusion drawn from these plots is that Transcan consistently yielded wider confidence intervals accounting for the additional uncertainty derived from the imputation process. The second conclusion is that PROC MI produced almost the same estimates and confidence intervals as if the model had been fitted with only non-missing responses. The third conclusion is that the ad-hoc routine produced overly narrow confidence intervals. The last conclusion, and perhaps the most interesting one, is that the confidence interval for NEW.AGE obtained from the data imputed by Transcan, practically does not overlap

Figure 5 – Comparison of the 95% confidence intervals of the significant coefficients.





with any of the other confidence intervals from the other routines. It is important to recall that this variable was found to be closely related to the non-response mechanism. This result provides evidence of the kind of bias that is likely to occur when the missing values are excluded from the analysis given a MAR response mechanism.

The fitted coefficients allowed us to confirm relationships between the predictors and the response that were rather obvious: the health status of the population decreases with age; higher income is related to better health; people with poorer health demand more medical services; people with more chronic conditions have lower health status and people with zero derived probability of depression appear as having better health.

5. CONCLUSIONS

When missing values are present in a data set, a careful assessment of the response mechanism must be conducted to avoid bias and model inaccuracy. Multiple imputation is a powerful tool that should be used whenever there is a significant number of missing values in the data, doing so will allow to obtain more valid statistical inferences. There are several software routines available to perform multiple imputation, it is up to the analyst to choose the one that best suits his needs bearing in mind the underlying assumptions of each one of them. As for this data set, we concluded that the *Transcan* routine implemented by Harrell in his *Hmisc* library for S-plus, is the best option to solve the problem of “filling in” the missing values.

ACKNOWLEDGMENTS

We want to thank Dr. Peggy Ng and Dr. Georges Monette for their kind help and support. Also, we want to thank our other two group members, Lena Zhang and Yaqing Chen, who contributed their work in the presentation of this case study at the Conference in Hamilton. Our sincere gratitude to the staff from Statistics Canada who organized this case study: Julie Bernier, David Haziza, Karla Nobrega and Patricia Whitridge.

REFERENCES

- Donald B. Rubin. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience.
- Donald B. Rubin. (1996). *Multiple imputation after 18+ years*. Journal of the American Statistical Association, 476,477.
- Frank E. Harrell, Jr. (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York : Springer.
- SAS Institute. *Documentation for the 8.2 release of the experimental MI and MIANALYZE procedures*. Available at: <http://www.sas.com/rnd/app/papers/miv802.pdf>
<http://www.sas.com/rnd/app/papers/mianalyzev802.pdf>