Journal of Official Statistics, Vol. 27, No. 2, 2011, pp. 211–229

An Analysis of Nonignorable Nonresponse to Income in a Survey with a Rotating Panel Design

Caterina Giusti[1] *and Roderick J.A. Little*[2]

In a rotating panel survey, individuals are interviewed in some waves of the survey but are not interviewed in others. We consider the treatment of missing income data in the labor force survey of the Municipality of Florence in Italy, a survey with a rotating panel design where recipiency and amount of income are missing for waves where individuals are not interviewed, and amount of income is missing for waves where individuals are interviewed but refuse to answer the income amount question. It is thus a question of a multivariate missing data problem with two missing-data mechanisms, one by design and one by refusal, and varying sets of covariates for imputation depending on the wave of the survey. Existing methods for multivariate imputation such as sequential regression multiple imputation (SRMI) can be applied, but assume that the missing income values are missing at random (MAR). This assumption is reasonable when missing data arise from the rotating panel design, but less reasonable when the missing data arise from refusal to answer the income question, since in this case missingness of income is generally thought to be related to the value of income itself, after conditioning on available covariates. In this article we describe a sensitivity analysis to assess the impact of departures from MAR for refusals, based on SRMI for a pattern-mixture model. The sensitivity analysis avoids the well-known problems of underidentification of parameters of missing not at random models, is easy to carry out using existing sequential multiple imputation software, and takes into account the different mechanisms that lead to missing data.

Key words: Missing data; pattern-mixture models; multiple imputation; sensitivity analysis.

1. Introduction

Missing data on income questions is an important concern in labor force surveys, given the inability or unwillingness of some individuals to report income information. An important early example methodologically is the hot deck imputation method of the Income Supplement of the U.S. Current Population Survey (CPS) (Ono and Miller 1969; U.S. Bureau of the Census 2002). The CPS Hot Deck creates adjustment cells based on recorded information for respondents and nonrespondents, and then imputes income amounts from a randomly chosen respondent in the same cell as the nonrespondent. This method assumes that the income variables are missing at random (MAR) (see e.g., Little and Rubin 2002), in the sense that missingness depends only on observed characteristics, and not on the missing values of the income variables themselves.

[1] Department of Statistics and Mathematics Applied to Economics, University of Pisa, Via Ridolfi 10, 56124 Pisa, Italy. Email: caterina.giusti@ec.unipi.it
[2] Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor MI 48105, U.S.A., and Associate Director for Research and Methodology, U.S. Census Bureau. Email: rlittle@umich.edu

The MAR assumption in the context of income nonresponse has been questioned by many analysts, who argue that nonresponse is more likely among individuals with low or high incomes than among individuals with incomes in the middle of the income distribution. In particular, Lillard et al. (1986) fitted a missing not at random (MNAR) model for income that attempts to correct for selection bias, based on models initially developed by Heckman (1976) and others. They concluded that the incomes of nonrespondents imputed by the CPS Hot Deck were being severely underestimated. However, these methods have been criticized on the grounds of their sensitivity to structural assumptions (Rubin 1983; Little 1985), and empirical work based on a match of the CPS to IRS data showed no evidence against the MAR assumption (David et al. 1986). Despite this study, the potential bias from assuming that missing incomes are MAR remains a concern, particularly in situations where there is limited covariate information to characterize differences between respondents and nonrespondents.

The treatment of MNAR missing data is a difficult problem, given the absence of empirical data to characterize differences between respondents and nonrespondents that are not captured by observed covariates. From a likelihood-based perspective, a model is needed for the joint distribution of the survey variables Y and the matrix M, which indicates which values are observed and which are missing. Most early work on MNAR models was based on selection models, which factor this joint distribution into the marginal distribution of Y (the "complete-data model") and the conditional distribution of M given Y (the "model for the missing-data mechanism"). Applications of this approach to income data include Greenlees et al. (1988) and Lillard et al. (1986). More recently, there has also been interest in pattern-mixture models, which factor the joint distribution into the marginal distribution of M (the distribution of each missing-data pattern) and the conditional distribution of Y given M (the model for Y within each pattern). (For discussions of the relative merits of these approaches see Little and Rubin 2002, Chapter 15; Little 1993; Kenward and Carpenter 2008; Little 2008.) Both approaches share severe problems of underidentification of parameters, essentially because the data provide no direct information about differences in Y between respondents and nonrespondents that are not accounted for by observed data. Thus, it has been argued (e.g., Rubin 1977; Little 1994; Scharfstein et al. 1999) that the most scientific approach is to assess sensitivity to non-MAR missing data, by considering the effect of a range of plausible differences between respondents and nonrespondents after adjusting for the available covariates. The analysis of MNAR income nonresponse in this article adopts this approach, based on a pattern-mixture model for the data.

Published sensitivity analyses based on MNAR models have been largely limited to the relatively simple problem where missing values are confined to a single variable. In this article we propose a sensitivity analysis to MNAR nonresponse in the setting of missing income information in a labor force survey conducted by the Municipality of Florence. This problem has a number of interesting complicating features. Specifically, there are missing data due to income nonresponse, which is potentially not MAR; the missing data pattern is multivariate, because quarterly income measures are recorded repeatedly over time, and the survey has a rotating panel design, which means that individuals are interviewed for some waves of the survey and not interviewed for others. The rotating panel design induces a designed missing data aspect, both for income recipiency and

income, which is essentially MAR (but not quite, since some individuals who are not interviewed in a wave might refuse if they were interviewed). Income recipiency and amount need to be considered for each quarter, since earned income is zero when individuals do not have a job (see Table 1). For both types of missing data, the amount of observed income information from other waves varies markedly from one individual to another, and this aspect should be appropriately reflected in the MNAR analysis.

We describe here an analysis that addresses these features. It is based on multiple imputation (MI) (Rubin 1987), an important approach for handling item nonresponse, particularly in public use data files. Initially, we multiply impute missing quarterly income values and missing values on occupational status and covariates using MAR sequential regression methods (Van Buuren and Oudshoorn 1999; Raghunathan et al. 2001), also known as chained equation methods. These allow us to condition on covariate information, including income data from other quarters if available. (For another application of sequential MI of income in a cross-sectional survey see Schenker et al. 2006.) We then describe two sensitivity analyses to deal with potential non-MAR missing income data. In contrast to approaches based on selection models, these methods are relatively simple to implement and provide useful information about the potential impact of deviations from MAR in the missing income items.

2. The Labor Force Survey

Our methods are motivated by missing data in the Labor Force Survey of the Municipality of Florence in Italy, an important source of information on the employment rate and income for employed people in the Florentine area. The survey collects data in four waves every year (April, July, October, and January) to produce quarterly estimates, which are then combined to yield annual estimates. A random sample of individuals is drawn from the municipal register of Florence, stratified by sex, age-class, and zone of residence. The survey has a rotating panel design, where each subject enters in the sample for two consecutive waves, exits for two and then reenters for two waves, with a 50% overlap after three and twelve months and a 25% overlap after nine and 15 months. To determine this timing, each subject is randomly assigned into one of eight "panel groups", in each of which sample strata are equally represented. Sampled individuals who cannot be contacted are

Table 1. Status of the variables Z (occupational status) and Y (monthly income) in the four quarters, for each panel group: observed (Obs) or missing (Mis)

	April 2002		July 2002		October 2002		January 2003	
Panel group	*Z*	*Y*	*Z*	*Y*	*Z*	*Y*	*Z*	*Y*
Group 1	Obs	Obs/Mis	Mis	Mis	Mis	Mis	Mis	Mis
Group 2	Mis	Mis	Obs	Obs/Mis	Mis	Mis	Mis	Mis
Group 3	Mis	Mis	Mis	Mis	Obs	Obs/Mis	Mis	Mis
Group 4	Mis	Mis	Mis	Mis	Mis	Mis	Obs	Obs/Mis
Group 5	Obs	Obs/Mis	Mis	Mis	Mis	Mis	Obs	Obs/Mis
Group 6	Obs	Obs/Mis	Obs	Obs/Mis	Mis	Mis	Mis	Mis
Group 7	Mis	Mis	Obs	Obs/Mis	Obs	Obs/Mis	Mis	Mis
Group 8	Mis	Mis	Mis	Mis	Obs	Obs/Mis	Obs	Obs/Mis

replaced by substitutes from the same stratum. In each of the four survey waves considered here (April, July, October 2002, and January 2003) around 1,200 people were interviewed in Florence. Depending on the "panel group" assignment, each subject was surveyed one or two times. The total number of distinct respondents in the four waves is 3,209.

The questionnaire begins with a question directed toward defining the individual's occupational status. An individual is considered as employed if he/she declares himself/herself as such or if he/she has worked during the preceding week; this definition includes both dependent and self-employed positions. The questionnaire proceeds with questions regarding the type of job and income for employed people, while for those not employed the survey asks questions about the job search.

In this article we focus on missing values for the questions about occupational status and earned income for employed people. (Missing data rates for other questions are less than 3%, and hence a minor issue.) Two distinct mechanisms lead to missing data on these variables. The first arises because occupational status and earned income are not recorded for waves where an individual is not interviewed, because of the rotating panel design. That is, if the individual data from different waves are combined into a single longitudinal file, variables are missing for the waves where an individual is not interviewed. We treat the unobserved data arising from the rotation of the panel as "designed missing data," since they meet the definition of missing data in Little and Rubin (2002) as unobserved values that are meaningful in the analysis. Also, imputation of these data using values from observed waves is useful since it increases the efficiency of the estimates.

The second missing-data mechanism is refusal to answer the income question in waves where an individual is interviewed. The question defining the occupational status is always observed in these waves, but some employed individuals refuse to answer the question "What is your monthly net income?" The questionnaire is structured so that "income" refers only to earned income from the current job; other sources of income are excluded. Accordingly, if a person is not asked the income question because he or she is not employed, then the corresponding income value is considered to be zero, not missing. For approaches to modeling financial variables with a proportion of zeros, see Buntin and Zaslavsky (2004). An important feature of our proposed analysis is that it treats these two mechanisms of missing data differently, confining MNAR methods to the refusal component.

Table 1 summarizes the status, observed or missing, for the occupational status (Z) and the monthly income (Y) in each of the quarters and for each panel group.

Table 2 reports the number of employed people and the corresponding percentages of missing values to the income question, for each panel group. The rates of missing values to the question on the monthly income are comparable with those of other surveys about income, assets, expenditures, and financial variables (Heeringa et al. 2002). Note that the zeros in Table 2 derive from the rotation of the panel: if the respondents were interviewed at these times, their occupational status would be recorded, some would report their income, and others would refuse to provide their income, as in other waves.

Let $Z_{hij} = 0, 1\,(h = 1, \ldots, H,\ \ i = 1, \ldots, n_h,\ \ j = 1, \ldots, J)$ be the indicator of the occupational status for Subject i in Stratum h and Wave j, and let Y_{hij} be the corresponding monthly net income from a job in Euros. If a subject is not employed ($Z_{hij} = 0$), then the income is zero ($Y_{hij} = 0$). Let X_{hij} denote the matrix containing personal characteristics for

Table 2. Number of employed people (N) and percentage of missing values (% missing) for the monthly income Y. The zeros in the table derive from the rotation of the survey scheme

Panel group	April 2002		July 2002		October 2002		January 2003	
	N	% missing	*N*	% missing	*N*	% missing	*N*	% missing
Group 1	286	31.47	0	0	0	0	0	0
Group 2	0	0	195	37.95	0	0	0	0
Group 3	0	0	0	0	174	36.21	0	0
Group 4	0	0	0	0	0	0	272	39.34
Group 5	118	31.36	0	0	0	0	119	26.05
Group 6	244	24.59	245	31.43	0	0	0	0
Group 7	0	0	239	38.49	239	36.82	0	0
Group 8	0	0	0	0	263	36.50	264	31.44
Total	648	28.86	679	35.79	676	36.54	655	33.74

Subject i in Stratum h and Wave j. These characteristics include information fixed during all the survey waves, such as sex, age-class, educational level and civil status, and information which may change depending on the occupational status in a given wave, like the type of job (employee or self-employee). Finally, let w_h be the sampling weight for individuals in Stratum h. The stratification is defined by three of the X variables: sex, age-class, and zone of residence.

Define the missingness indicator M_{hij} such that $M_{hij} = 0$ if occupational status and income are observed; $M_{hij} = 1$ if occupational status and income are both missing, as when the subject belongs in a panel group that is not interviewed in Wave j; and $M_{hij} = 2$ if occupational status is observed but income is missing, as when an individual is interviewed but refuses to answer the income question. For simplicity of notation we treat all the characteristics X_{hij} as fully observed, although a few covariate values of these variables are missing in each wave. These values are imputed using the MAR sequential MI procedure described below. The weights w_h are all observed.

Quarterly estimates of the monthly earned income are currently based on available information, dropping cases for which income is not observed. The estimated mean in Wave j, accounting for the stratification weights, is:

$$\hat{\bar{Y}}_{\cdot j} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{hij} Z_{hij} w_h}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} Z_{hij} w_h} \tag{1}$$

The associated estimate of the standard error is obtained using the SAS Proc Surveymeans software, which uses a Taylor series expansion method.

Besides the quarterly estimates, an estimate of the monthly income aggregated over the whole year 2002 is also of interest. This estimate could be computed by averaging the $\hat{\bar{Y}}_{\cdot j}$ over the J waves; however in this estimate some subjects contribute to only one wave mean, other subjects to two wave means. Alternatively, we can estimate the average monthly income during 2002 using one value for each subject in each stratum, represented

by the mean of observed monthly income estimates

$$\hat{\bar{Y}}_{hi.} = \frac{\sum_{j=1}^{4} Y_{hij} Z_{hij}}{\sum_{j=1}^{4} Z_{hij}}, \quad \sum_{j=1}^{4} Z_{hij} > 0,$$

and then derive the overall monthly estimate as:

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \hat{\bar{Y}}_{hi.} w_h}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_h} \tag{2}$$

The results from these analyses of available data are displayed in Table 3.

From these results we note that the monthly income estimates increase in the last two quarters of the year, especially in the third. The lowest value is for the second quarter, observed in the month of July. The monthly income estimate referring to the whole year ($\hat{\bar{Y}}$) is higher than the first two quarterly estimates, lower than the remaining two.

This approach makes the strong assumption that the missing values for each month are missing completely at random (MCAR) – that is, are unrelated to the missing income values or the observed covariates. This assumption is justified for missingness attributable to the rotating panel design, but is a strong assumption for missingness of income because of refusal to answer the income question. It is generally preferable to develop consistent estimates under the weaker MAR assumption, which allows the conditional distribution of the missing data indicators to depend on the observed data (Little and Rubin 2002). The MAR assumption in our setting is:

$$Pr\,(M_{hi.}|Y_{hi.}, Z_{hi.}, X_{hi.}, \psi) = Pr\,(M_{hi.}|Y_{\text{obs},hi.}, Z_{\text{obs},hi.}, X_{hi.}, \psi) \tag{3}$$

where $M_{hi.}$ represents the vector of missing data indicators for Subject i in Stratum h over the survey waves, $Y_{hi.}$, $Z_{hi.}$, $X_{hi.}$ represent the vectors of values of income, income recipiency and covariates over all survey waves, and $Y_{\text{obs},hi.}$ and $Z_{\text{obs},hi.}$ are the observed components of $Y_{hi.}$, $Z_{hi.}$; we define the corresponding missing components as $Y_{\text{mis},hi.}$, $Z_{\text{mis},hi.}$. We now describe an MI analysis that imputes the missing values under the MAR assumption.

3. Multiple Imputation Under MAR

In this section we multiply-impute the missing values of occupational status and monthly income, $Z_{\text{mis},hi.}$ and $Y_{\text{mis},hi.}$, and the missing covariates under the assumption that all the values are MAR. In MI, m complete datasets are produced, with missing values replaced

Table 3. Number of employed people (N), monthly income estimates and standard errors for the monthly income (in Euros) with the complete-case analysis

Estimates	$\hat{\bar{Y}}_{..1}$	$\hat{\bar{Y}}_{..2}$	$\hat{\bar{Y}}_{..3}$	$\hat{\bar{Y}}_{..4}$	$\hat{\bar{Y}}$
N	461	436	429	434	1,327
Mean estimate	1,195.2	1,186.8	1,309.0	1,234.3	1,221.2
Standard error	31.3	26.6	33.3	26.8	22.7

by draws from their posterior predictive distribution under an imputation model. In order to address the multivariate nature of the missing and observed data and condition fully on the observed information, we applied the sequential regression multivariate approach to MI (Raghunathan et al. 2001; Van Buuren and Oudshoorn 1999). This approach avoids the specification of a full joint multivariate model for the variables, which can be difficult when these variables are numerous and have different distributional forms. Under the MAR assumption, it is not necessary to distinguish whether an income value Y_{hij} is missing because Subject i of Stratum h was not interviewed in Wave j, or because the subject was interviewed but refused to answer.

Under the MAR hypothesis, the sequential regression MI for variables Z and Y proceeds as follows. A regression model is chosen for each variable with missing values: here a logit regression for the dummy variable measuring the occupational status and a linear regression for the logarithm of the income. Diffuse prior distributions are assumed for the parameters of the regressions. At the first step a regression of $Z_{\text{obs},hij}$ on the covariates $X_{hi.}$ is fitted and the missing values $Z_{\text{mis},hij}$ are imputed from the corresponding posterior predictive distribution; next, a regression of $log(Y_{\text{obs},hij})$ on the $X_{hi.}$ and the completed Z_{hij} is fitted and also the $Y_{\text{mis},hij}$ are imputed when $Z_{hij} = 1$, while the income is set to zero when $Z_{hij} = 0$. In the same way the missing values of the $X_{hi.}$ variables are imputed based on their regression on $Z_{hi.}$ and $Y_{hi.}$. Then the procedure begins to cycle with each regression fitted again using as predictors the covariates and all the previously imputed values, until stable imputations for all the variables are obtained. A Gibbs sampler algorithm is necessary, since the missing data pattern is not monotone (Raghunathan et al. 2001).

The distributions for the log income amounts are all assumed normal, and the prior distributions of the parameters are noninformative $g(\beta_j, \sigma_{jj}) = \sigma_{jj}^{-1/2}$. To ensure approximate normality for the continuous income variables, we also considered Box-Cox family transformations (Box and Cox 1964). The power transformation estimated by the method of maximum likelihood was near to zero (log transformation) for each of the four income variables. Thus, we chose this transformation though the transformed variables show a kurtosis higher than that for the normal distribution. A refinement would replace the normal by a longer-tailed distribution like the multivariate t, but the focus here is on the MNAR sensitivity analysis discussed below.

Repeating this process m times, m completed datasets are produced. Then, the subsequent steps are: conduct separate analyses on the m complete datasets with traditional techniques to obtain, for example, the estimates of a parameter θ; combine these estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ together with their associated variances $\hat{U}_1, \ldots, \hat{U}_m$ through MI combining rules (Rubin 1987). In particular, the MI estimate of θ is: $\hat{\theta} = \sum_{k=1}^{m} \hat{\theta}_k / m$, with variance $\hat{V} = \bar{U} + (1 + m^{-1})B$, where $\bar{U} = \sum_{k=1}^{m} \hat{U}_k / m$ is the within-imputation variance and $B = \sum_{k=1}^{m} (\hat{\theta}_k - \hat{\theta})^2 / (m - 1)$ is the between-imputation variance.

The sequential regression approach to MI is flexible and makes good use of the available information, but has some limitations. The conditional distributions of the variables with missing values may be incoherent, in the sense that they cannot be derived by a single joint multivariate distribution (Little and Rubin 2002). Theoretically it is possible that the Gibbs sampler for these imputation models does not converge stochastically to a draw from the joint distribution. However, the method appears to work well in practice (Van Buuren et al. 2006; Heeringa et al. 2002).

4. Results Under the MAR Model

We chose to impute $m = 25$ datasets with the software package IVEware (Raghunathan et al. 2002). Smaller values of m suffice when the rate of missing values is very low (Rubin 1987), but here a higher value is required since the rotating panel design leads to a high rate of missingness (see Table 4). The number $m = 25$ yielded a stable estimate of the between-imputation component of the MI variance.

The MAR imputation scheme (3) requires choosing a set of covariates X to condition in the imputation model. To keep the imputation model as general as possible, besides the occupational status and income in the different waves, we conditioned here on the personal characteristics fixed during all the survey waves, namely sex, age-class, number of household members, zone of residence in the Municipality of Florence, educational level, civil status. Also, we conditioned on some characteristics available for the quarters when the subject is interviewed and employed – that is, the type of job (employee or self-employee), the number of household members perceiving income, and the involvement in a second job. Note that since these characteristics are not available for some quarters, due to the rotating scheme, we needed to impute them under our MAR model. Finally, we included the survey weights as covariates in the imputation model. We imputed using the option MINRSQD of IVEware, specifying a minimum marginal R-squared for a step-wise regression equal to 0.005. Checking the details of the imputation procedure we found that the chosen covariates were included as predictors in the sequential regressions.

The imputations of occupational status are highly influenced by the observed covariate information. For example, if a subject was interviewed in two waves and declared himself/herself as (not) employed in both, then his/her occupational status is imputed as (not) employed in the remaining two waves with a 95% probability (mean value across the 25 MIs). When the occupational status changes in the two observed waves, the imputations are more changeable. Otherwise, when there is only one observed value, the same occupational status is imputed in the remaining three waves for approximatively 85% of the cases. The average number of employed people across the 25 imputed datasets and the corresponding percentages of missing income values are shown in Table 4. Of course, when the occupational status is missing because of the rotation of the panel, the corresponding income is always missing. Considering all the panel groups, the percentage of income values to be imputed in each wave is very high, around 75%.

Concerning the imputation of the income, we compared the relationship between the pairs of observed income values with that between one observed and one imputed value (due to refusal to answer) for individuals interviewed in two waves (data referring to panel groups from 5 to 8, see Table 1). A scatterplot of these couples of values is shown on the left in Figure 1; for ease of comparison, only observed and imputed values below 5,000 Euros are included. We note that the positive correlation between the observed income values is well preserved by the imputations; results are quite similar in all the imputed datasets.

To check the fit of the imputation models for the income variables, we display bivariate scatterplots that plot the residuals against the predicted values for the observed income values in each wave (Abayomi et al. 2008; Su et al. 2009). Figure 1 reports on the right the scatterplot referring to the monthly income in April 2002 in the log scale: as we can see, no

Table 4. Number of employed people (N) and percentage of missing values (% missing) for the monthly income across the 25 MAR multiple imputations

	April 2002		July 2002		October 2002		January 2003	
Panel group	*N*	% missing	*N*	% missing	*N*	% missing	*N*	% missing
Group 1	286	31.47	302	100	304	100	298	100
Group 2	187	100	195	37.95	194	100	191	100
Group 3	162	100	166	100	174	36.21	168	100
Group 4	265	100	274	100	279	100	272	39.34
Group 5	118	31.36	126	100	126	100	119	26.05
Group 6	244	24.59	245	31.43	248	100	239	100
Group 7	228	100	239	38.49	239	36.82	229	100
Group 8	258	100	273	100	263	36.50	264	31.44
Total	1,748	73.63	1,820	76.04	1,827	76.52	1,780	75.61

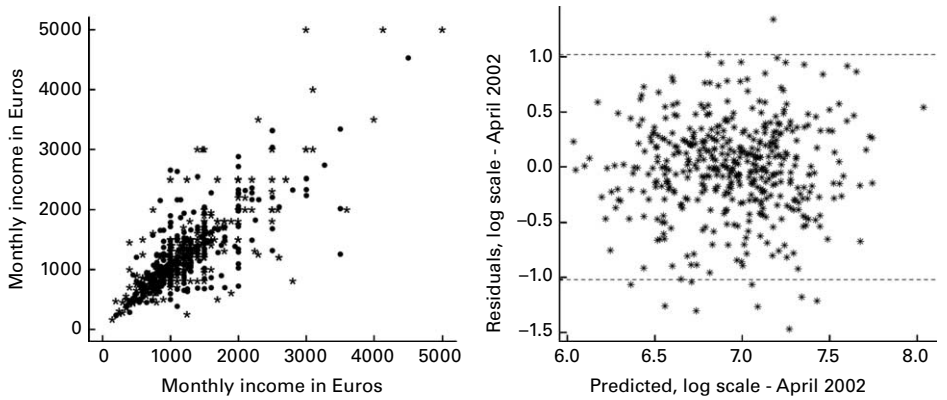

Fig. 1. On the left, scatterplot of the couples of observed income values (stars) and of observed and imputed income values (dots) for one randomly chosen datasets. On the right, scatterplot of residuals vs predicted values, with 95% error bounds

patterns are present and almost all the points fall within the 95% error bounds, demonstrating the appropriateness of the imputation model.

Finally, as an additional diagnostic tool, we compared the empirical densities of some of the 25 MAR imputed income distributions with those of the corresponding observed values (Figure 2). This visual examination may identify potential problems when imputing in a multivariate setting (Abayomi et al. 2008). For each of the four income distributions we never observed dramatic differences between the empirical density before and after the MAR imputations. The observed differences depend on the covariate information in the MAR imputation model.

We can recompute the estimates of interest, Quantities (1) and (2), and an additional annual income estimate, using data imputed using this method. Considering individuals employed in every wave of year 2002 ($Z_{hij} = 1$ for $j = 1, \ldots, 4$) and referring each quarterly estimate to the preceding three months, define the personal estimate of the annual income in year 2002 as $\hat{Y}_{hi\,2002} = \sum_{j=1}^{4} \tilde{Y}_{hij}$*3. Then, the overall annual income estimate is:

$$\hat{Y}_{2002} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \hat{Y}_{hi\,2002} w_h}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_h} \tag{4}$$

Using Rubin's rules, we combined the monthly and annual estimates computed in the 25 multiply imputed datasets. For the two estimates referring to the whole year 2002, $\hat{\bar{Y}}$ and

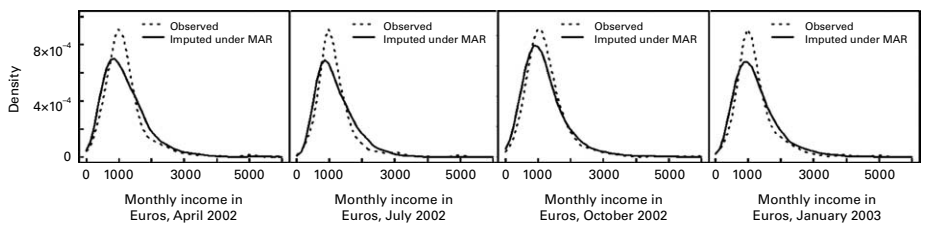

Fig. 2. Empirical densities of the observed income values in the four quarters (dotted lines) and of the imputed income values (solid lines) under the MAR model

$\hat{Y}_{2002}$, we also computed the median, 20th and 80th percentiles of the distribution. To calculate these estimates' variances in each dataset we used the bootstrap resampling technique, drawing 200 samples by random sampling with replacement, separately in each sampling stratum.

We also computed the fraction of missing information, which measures how the missing data contribute to inferential uncertainty about θ, the estimate of interest. The fraction of missing information can be computed as $\hat{\lambda} = r + 2/(v+3)/(r+1)$ where $r = (1 + m^{-1})B/\bar{U}$ and $v = (m-1)[1 + (\bar{U}/(1 + m^{-1})B)]^2$ (Schafer 1997), and where U and B are respectively the within and between variances across the m imputations.

The results for the quarterly and annual income estimates under the MAR model are shown in Table 5. The differences in the distribution of income between the waves are reduced under the MAR sequential imputations, compared with the MCAR results (Table 3). However, the estimates referring to the last two quarters of year 2002 are still higher, though the number of employed people does not increase. The fraction of missing information is also different between the quarters.

These differences depend on some really high observed values in the first and third waves, which contributed to increase the between variance of the multiple imputed estimates in July. However, the fraction of missing information is lower than the fraction of missing values (Table 4) for all the other quarters, reflecting the information incorporated into the imputations via the sequential regression model. Moreover, if we measure the relative efficiency of the MI estimates using $m = 25$ with using an infinity number of imputations, that is the quantity $1 + /(1+\hat{\lambda}/m)$ (Rubin 1987), we obtain an efficiency between the 97–98% for all the estimates. Therefore, the choice $m = 25$ seems a reasonable one in the current setting.

The results for the two annual estimates, $\hat{\bar{Y}}$ and $\hat{Y}_{2002}$, are shown in Table 6. As we can see, the monthly income estimate for the whole year 2002 is slightly lower under the MAR method (1,198.1 Euros) than under the MCAR method (1,221.2 Euros, see Table 3). For both methods the estimated median is lower than the estimated mean, reflecting a positive skew in the income distribution.

5. Sensitivity Analysis for Deviations from MAR

We now describe modifications of the MAR analysis of the previous section to examine sensitivity to MNAR missing-data mechanisms. The MNAR mechanism is modeled via the joint distribution of Y_{hij}, Z_{hij} and M_{hij} given the observed variables, including

Table 5. Number of employed people (N), monthly income estimates and standard errors (in Euros) and fraction of missing information (% missing info) across the 25 MAR multiple imputations

Estimates	$\hat{\bar{Y}}_{..1}$	$\hat{\bar{Y}}_{..2}$	$\hat{\bar{Y}}_{..3}$	$\hat{\bar{Y}}_{..4}$
MI N	1,748	1,820	1,827	1,780
MI mean estimate	1,210.09	1,188.21	1,280.90	1,249.83
MI standard error	25.48	28.56	27.67	25.99
MI % missing info	62.46	80.38	53.59	66.14

Table 6. Number of employed people (N), annual income estimates and standard errors (in Euros) and fraction of missing information (% missing info) across the 25 MAR multiple imputations

Estimates	$\hat{\bar{Y}}$	$\hat{Y}_{2002}$
MI N	2,420	1,086
MI mean estimate	1,198.12	15,532.00
MI standard error	17.30	234.49
MI % missing info	68.62	59.53
MI median estimate	1,091.18	14,405.16
MI median standard error	16.55	257.64
MI 20th percentile estimate	783.04	10,755.72
MI 20th percentile standard error	12.51	243.83
MI 80th percentile estimate	1,535.71	19,585.32
MI 80th percentile standard error	27.24	384.12

covariates X_{hi} and observed income information in other waves, which we write generically as $C_{\text{obs},hi.}$. We first factorize this distribution as:

$$f[Y_{hij}, Z_{hij}, M_{hij}|C_{\text{obs},hij}] = f[Y_{hij}, Z_{hij}|M_{hij}, C_{\text{obs},hij}] \times f[M_{hij}|C_{\text{obs},hij}]$$

which is a pattern-mixture factorization of the joint distribution (Little 1993). We assume:

$$f[Y_{hij}, Z_{hij}|M_{hij} = 1, C_{\text{obs},hij}] = f[Y_{hij}, Z_{hij}|M_{hij} \neq 1, C_{\text{obs},hij}] \quad (5)$$

which expresses the fact that the distribution of Y_{hij}, Z_{hij} is the same for individuals who are or are not interviewed because of the rotation group design. Further, for the missing income values due to refusal we assume that:

$$f[Y_{hij}|Z_{hij} = 1, M_{hij} = 2, C_{\text{obs},hij}] \neq f[Y_{hij}|Z_{hij} = 1, M_{hij} = 0, C_{\text{obs},hij}]$$

This is MNAR because the distribution of Y_{hij} given Z_{hij} and $C_{\text{obs},hij}$ is different for refusers and responders. Note that this distribution conditions on Z_{hij} since that variable is observed for cases with $M_{hij} = 0$ or 2. Specifically, we model the difference by assuming

$$E[log(Y_{hij})|Z_{hij}=1, M_{hij}=2, C_{\text{obs},hij}] = E[log(Y_{hij})|Z_{hij}=1, M_{hij}=0, C_{\text{obs},hij}] + k\sigma_{hj} \quad (6)$$

where σ_{hj} is the residual standard deviation of the distribution of $log(Y_{hij})$ for respondents given $Z_{hij}=1$ and $C_{\text{obs},hij}$, and k is a positive predetermined multiplier. The effect is to increase the mean of the distribution for refusers relative to that for respondents by a value $k\sigma_{hj}$ that depends on the choice of k and the predictive power of $C_{\text{obs},hij}$, as reflected in the residual standard deviation σ_{hj}. Note that the shift in the distribution for nonrespondents is applied after fitting the MAR model, and is not part of the imputation algorithm. This is because we do not want the increment to be amplified by the iterations of the imputation scheme, a point discussed in Van Buuren et al. (1999).

To illustrate this MNAR model, consider panel group 5, where an individual is part of the rotating panel in Waves 1 and 4, but not in Waves 2 and 3 (see Table 1). This results in four possible patterns for M_{hij}, namely 0110, 2110, 0112, 2112. People belonging to pattern 0110 reported their income when interviewed, while people in pattern 2110 refused

to answer (indicator equal to 2) at the first but not at the fourth wave, and so on. Missing values of income in Waves 2 and 3 are imputed using the corresponding distributions for individuals in the sample (for respondents and refusers, since individuals not interviewed might refuse if interviewed). For the refusals in Waves 1 or 4, we apply the offset for non-MAR missing data. The size of the offset for refusals in the first wave is larger for pattern 2112 than for pattern 2110, since the latter allows the missing income at Wave 1 to condition on the observed income value at Wave 4, thereby reducing the value of σ_{hj}.

This model is implemented as follows:

(A) The MAR multiple imputations are created as before;
(B) A value of k is chosen (0.8, 1.2 or 1.6, which we consider to reflect small, medium, and large deviations from MAR). The offsets are then applied to the imputations for refusals;
(C) For each of the m sets of multiple imputations, the imputations for the refusals are treated as known, and the sequential multiple imputation method is applied to reimpute the missing values of Y and Z for months not in the rotation group. This allows these imputations to condition on the offsetted values of the refusals, reflecting the fact that individuals not in the rotation group may also refuse.

We label this imputation model MNAR_1. We also present results under an alternative assumption (denoted MNAR_2), where missing values for cases with at least one income value reported can be regarded as MAR. The offset is thus restricted to cases with no observed income values. Considering again a subject belonging to panel group 5, the MNAR_2 model applies an offset to the imputed values for the first and fourth waves in pattern 2112, when both the income values are refusals, but does not apply an offset to the imputed values for patterns 2110 or 0112, when one of the income values is observed. The MNAR_2 mechanism is clearly closer to MAR than model MNAR_1. We think of MNAR_1 and MNAR_2 as bounding a range of plausible combinations of these models, for any given choice of k.

6. Results Under the MNAR Models

To evaluate the impact of the MNAR increments on the income distributions referring to the four quarters we plotted again the empirical densities of some of the 25 imputed income distributions, comparing them with those of the corresponding observed values. In Figure 3 the empirical densities of the observed income distributions in the first quarter (April) and the corresponding ones obtained after the MNAR_1 imputations are represented.

From the visual representations of the empirical densities we can appreciate the impact of the proposed imputation models on the income distribution in April. As expected, higher k values cause a more pronounced shift for the corresponding density. The same plots referring to the remaining three quarters and to the MNAR_2 imputations, not shown here, are very similar to those in Figure 3, with the increments under the MNAR_2 model causing a lower shift for the distributions.

We then computed the estimates of interest for the MNAR imputed income variables. The quarterly income estimates under the MNAR_1 and MNAR_2 models are shown in Table 7. The MNAR offsets result in larger estimates than those under MAR, especially

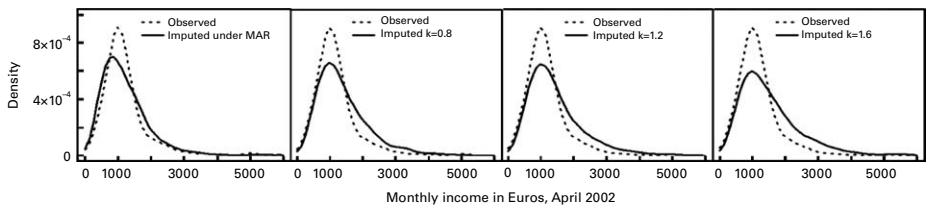

Fig. 3. Empirical densities of the observed income values (dotted lines) and of imputed income values (solid lines) in the first quarter under the MNAR$_1$ model

for larger values of k. As expected, the MNAR$_1$ assumption leads to larger increases than the MNAR$_2$ assumption, especially for $k = 1.2$ and $k = 1.6$. As under the MCAR and MAR hypothesis, the monthly income estimates in the first and second quarters are lower than those in the remaining two quarters, both under MNAR$_1$ and MNAR$_2$ and for each value of k.

When it comes to the percentage increase of these estimates as compared to the estimates obtained under the MAR assumption, when $k = 0.8$ the percentage increase of the quarterly income estimates is around the 10% and the 7% under the MNAR$_1$ and MNAR$_2$ mechanisms respectively. For $k = 1.2$ and $k = 1.6$ we observe a more pronounced impact of the MNAR$_1$ mechanism, especially for the monthly income estimate in the third quarter.

Table 7. Number of employed people (N), monthly income estimates and standard errors (in Euros) and fraction of missing information (% missing info) across the 25 MNAR multiple imputations

Model	Estimates	$\hat{\bar{Y}}_{..1}$	$\hat{\bar{Y}}_{..2}$	$\hat{\bar{Y}}_{..3}$	$\hat{\bar{Y}}_{..4}$
MNAR$_1$, $k = 0.8$	MI N	1,754	1,791	1,814	1,774
	MI mean estimate	1,316.9	1,306.8	1,421.5	1,369.5
	MI standard error	23.9	20.4	24.2	24.0
	MI % missing info	44.6	47.3	28.1	51.2
MNAR$_1$, $k = 1.2$	MI N	1,755	1,796	1,819	1,770
	MI mean estimate	1,390.0	1,383.1	1,518.7	1,452.4
	MI standard error	25.7	25.6	31.2	23.7
	MI % missing info	41.4	59.0	48.0	35.4
MNAR$_1$, $k = 1.6$	MI N	1,756	1,780	1,812	1,777
	MI mean estimate	1,475.9	1,465.1	1,605.0	1,526.8
	MI standard error	31.3	28.0	32.6	30.7
	MI % missing info	50.3	57.2	44.3	52.4
MNAR$_2$, $k = 0.8$	MI N	1,751	1,791	1,813	1,771
	MI mean estimate	1,290.2	1,263.3	1,375.7	1,342.0
	MI standard error	24.5	19.9	24.3	21.0
	MI % missing info	51.1	51.0	33.1	38.5
MNAR$_2$, $k = 1.2$	MI N	1,749	1,787	1,814	1,776
	MI mean estimate	1,343.3	1,320.2	1,439.4	1,399.5
	MI standard error	25.4	20.1	27.1	26.9
	MI % missing info	47.1	41.9	38.1	56.3
MNAR$_2$, $k = 1.6$	MI N	1,738	1,784	1,811	1,784
	MI mean estimate	1,416.8	1,366.0	1,509.2	1,468.4
	MI standard error	27.9	21.5	27.2	26.0
	MI % missing info	45.2	39.5	27.9	41.1

Note that this greater increase depends on some high income values observed in the first and third quarters, already noted for the MAR model; these values cause a bigger residual standard deviation for the corresponding log-normal regression model. Moreover, in the third quarter we also observe a slightly higher percentage of nonresponses (see Table 4) which are incremented under the MNAR models.

The estimates referring to the whole year 2002 under the two MNAR hypotheses are shown in Table 8. The increases for the annual estimate $\hat{\bar{Y}}$ are similar to those for the quarterly estimates (10% and 8% respectively under MNAR_1 and MNAR_2), while those for $\hat{Y}_{2002}$ are slightly lower (8% and 5.4% respectively). The percentage increases are slightly lower in terms of median values, as it is for the estimates of the 20th percentiles.

The monthly estimate referring to all the year 2002, $\hat{\bar{Y}}$, is always higher than the first and second quarter estimates, and lower than the third and fourth quarter estimates, as in the MAR analysis. The MNAR annual income estimates $\hat{Y}_{2002}$ are all between 15,000 and

Table 8. Number of employed people (N), annual income estimates and standard errors (in Euros) and fraction of missing information (% missing info) across the 25 MNAR_1 and MNAR_2 multiple imputations

		MNAR_1		MNAR_2	
k value	Estimates	$\hat{\bar{Y}}$	$\hat{Y}_{2002}$	$\hat{\bar{Y}}$	$\hat{Y}_{2002}$
$k = 0.8$	MI N	2,129	1,405	2,129	1,410
	MI mean estimate	1,322.3	16,762.0	1,285.0	16,381.0
	MI standard error	13.9	216.3	14.6	216.2
	MI % missing info	32.1	47.4	42.2	49.0
	MI median estimate	1,198.3	15,319.0	1,163.2	14,923.7
	MI median standard error	16.5	236.0	14.8	250.8
	MI 20th percentile	854.8	11,259.4	836.3	11,002.8
	MI 20th percentile standard error	13.7	194.4	12.6	177.2
	MI 80th percentile	1,705.4	21,467.5	1,650.8	20,827.2
	MI 80th percentile standard error	29.0	386.5	25.9	399.5
$k = 1.2$	MI N	2,137	1,403	2,119	1,414
	MI mean estimate	1,398.3	17,837.0	1,350.2	16,921
	MI standard error	17.3	252.6	16.5	247
	MI % missing info	47.4	52.1	45.2	52.9
	MI median estimate	1,258.9	16,145.8	1,211.7	15,245.6
	MI median standard error	17.3	266.7	17.7	248.8
	MI 20th percentile	886.5	11,774.9	859.2	11,087.3
	MI 20th percentile standard error	14.8	203.1	12.3	204.7
	MI 80th percentile	1,817.5	22,863.4	1,747.5	21,725.4
	MI 80th percentile standard error	32.1	480.3	30.4	425.9
$k = 1.6$	MI N	2,119	1,414	2,114	1,418
	MI mean estimate	1,484.1	18,772.0	1,420.5	17,590.5
	MI standard error	20.4	257.6	16.6	257.6
	MI % missing info	55.3	46.7	29.9	47.1
	MI median estimate	1,322.6	16,917.4	1,258.3	15,764.3
	MI median standard error	18.7	279.4	19.2	264.1
	MI 20th percentile	924.5	12,122.9	880.8	11,241.4
	MI 20th percentile standard error	14.4	225.0	12.5	186.7
	MI 80th percentile	1,941.0	24,241.4	1,850.4	22,730.9
	MI 80th percentile standard error	35.2	511.6	32.5	510.7

19,000 Euros. This range is consistent with data coming from independent sources. In particular, the estimate of annual net income from job (the same estimate we are considering) resulting from a survey on tax records in the Municipality of Florence in 2002 is equal to 16,070 Euros for employees and 24,400 for the self-employeed. Considering that the employees represent approximately 72% of the population under study (mean value across the quarters and multiple imputations), the annual net income estimated using the tax record data is equal to 18,404 Euros. This value is coherent with the estimates and standard errors we obtain for $\hat{Y}_{2002}$ under the MNAR_1 model with $k = 1.2$ and $k = 1.6$, and for model MNAR_2 with $k = 1.6$.

Our results are also consistent with the estimates resulting from a national survey conducted by the Italian National Institute of Statistics (ISTAT) – the Survey on Income and Living Conditions 2004 – which links to tax reports in the case of nonresponse. This survey estimated an annual mean net income from employment in 2003 in the region of Florence, Tuscany, of 15,727 Euros, with the corresponding median estimate equal to 13,284 Euros. However, the confidence intervals for the mean and median estimates referring to the Municipality of Florence, though rather wide since based on around 200 units, suggest that the Florentine area is richer than the Tuscany region as a whole, as reflected in our estimates.

These external references suggest that the value $k = 1.6$ can be considered as a maximum for our proposed MNAR models. Broadly speaking, we can say that the MNAR deviations from the MAR estimates are moderate, especially under the MNAR_1 model.

7. Conclusion

We have described the use of SRMI to impute missing income amounts in a rotating panel survey, where values of income recipiency and amount are missing for quarters when the individual is not interviewed, and amounts are also missing because of refusal or inability to answer the amount question. Compared with other approaches, this analysis conditions imputations on available information, and hence is particularly attractive when information on income is available for some waves. However, this approach makes the MAR assumption. Thus, we have also described a sensitivity analysis for deviations from MAR, based on offsets applied to the imputations from the MAR model, defined as a fraction k of the residual standard deviation from the log-normal regression model on observed income values and covariates. The sensitivity analysis suggested that income amounts are moderately sensitive in this application, for a range of plausible values of k.

The MNAR model is based on a pattern-mixture factorization, and it extends existing MNAR models in a number of useful ways. First, it distinguishes between the two types of missing data in this application, one of which is essentially MCAR (the rotation group design) and one of which may not be MAR (refusal). This approach operationalizes the recommendation in Little (1995) to tailor the model for nonresponse according to the reason that a value is missing. It also limits the scope of the sensitivity analysis to the missing values likely to deviate from MAR, thus avoiding an overstatement of the additional uncertainty from nonresponse. The idea of modeling MNAR by adding offsets to the mean of the respondent distribution has the advantage of being easy to implement, involving simpler adjustments to the MAR imputations, and the deviations from MAR are

readily understood. Rubin (1977) expressed the need for simple sensitivity analyses for deviations from MAR as follows:

> "In special cases, it may be possible to estimate the effect of nonrespondents under accepted models. More often, the investigator has to make subjective judgments about the effect of nonrespondents. Given this situation, it seems reasonable to try to formulate these subjective notions so that they can be easily stated, communicated, and compared."

The advantage of pattern-mixture models in terms of simplicity is noted in Kenward and Carpenter (2008). In contrast, deviations from MAR in selection models require more complex computations and are harder to explain to practitioners, since the predictive distribution of the missing values is being modeled indirectly (Little and Rubin 2002, Chapter 15; Kenward and Carpenter 2008). We suggest that specifying an offset is more realistic than attempting to estimate selection bias using structural assumptions, since in practice the evidence in the data to estimate deviations from MAR is very limited.

Another novel aspect of our analysis is to choose the offset as a fraction of the residual standard deviation from the regression of the missing variable on observed covariates. This approach takes into account relationships with known covariates, which is particularly important in our application given the potential to use income amounts from other quarters as covariates: clearly these values carry considerable information for the value being imputed. With income modeled on the log scale, the offset can be interpreted approximately as a percentage change on the raw scale, which is easy to interpret.

In the application we perturbed the values by making them larger, on the assumption that missingness is positively related to the actual income value. Other deviations from MAR can also be considered in our proposed sensitivity model, if they are thought appropriate. For example, an alternative sensitivity analysis that has the effect of changing the income distribution in both tails might be specified by increasing the standard deviation of the predictive distribution of log income, under the assumption that income values for nonrespondents are more dispersed that those predicted under the MAR model.

8. References

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for Multivariate Imputations. Journal of the Royal Statistical Society, Series C, 57, 273–291.

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society, Series B, 26, 211–252.

Buntin, M.B. and Zaslavsky, A.M. (2004). Too Much Ado about Two-part Models and Transformation? Comparing Methods of Modeling Medicare Expenditures. Journal of Health Economics, 23, 525–542.

David, M., Little, R.J.A., Samuhel, M.E., and Triest, R.K. (1986). Alternative Methods for CPS Income Imputation. Journal of the American Statistical Association, 81, 29–41.

Greenlees, J.S., Reece, W.S., and Zieschang, K.D. (1988). Imputation of Missing Values When the Probability of Response Depends on the Varible being Imputed. Journal of the American Statistical Association, 77, 251–261.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation. Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. Annals of Economic and Social Measurement, 5, 475–492.

Heeringa, S.G., Little, R.J.A., and Raghunathan, T.E. (2002). Multivariate Imputation of Coarsened Survey Data on Household Wealth. In Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, 357–371.

Kenward, M.G. and Carpenter, J.R. (2008). Multiple Imputation. In Longitudinal Data Analysis, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds). New York: CRC Press, 477–500.

Lillard, L., Smith, J.P., and Welch, F. (1986). What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation. Journal of Political Economy, 94, 489–506.

Little, R.J.A. (1985). A Note about Models for Selectivity Bias. Econometrica, 53, 1469–1474.

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. (1994). A Class of Pattern-Mixture Models for Normal Incomplete Data. Biometrika, 81, 471–483.

Little, R.J.A. (1995). Modeling the Drop-out Mechanism in Longitudinal Studies. Journal of the American Statistical Association, 90, 1112–1121.

Little, R.J.A. (2008). Selection and Pattern-Mixture Models. In Longitudinal Data Analysis, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds). New York: CRC Press, 409–432.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.

Ono, M. and Miller, H.P. (1969). Income Nonresponses in the Current Population Survey. In Proceedings of the American Statistical Association, Social Statistics Section. Washington: American Statistical Association, 277–288.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, 27, 85–95.

Raghunathan, T.E., Solenberger, P.W., and Van Hoewyk, J. (2002). IVEware: Imputation and Variance Estimation Software User Guide. Survey Methodological Program, Survey Research Center, Insitute for Social Research, University of Michigan.

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. Journal of the American Statistical Association, 72, 538–543.

Rubin, D.B. (1983). Imputing Income in the CPS. The Measurement of Labor Cost. Chicago: University of Chicago Press.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Sample Surveys. New York: Wiley.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman and Hall.

Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models (with Discussion). Journal of the American Statistical Association, 94, 1096–1146.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., and Cohen, A.J. (2006). Multiple Imputation of Missing Income Data in the National Health Interview Survey. Journal of the American Statistical Association, 101, 924–933.

Su, Y.S., Gelman, A., Hill, J., and Yajima, M. (2009). Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box, to appear in the Journal of Statistical Software.

U.S. Bureau of the Census (2002). Current Population Survey: Design and Methodology. Technical Report, 63RV. U.S. Bureau of Labor Statistics.

Van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. Statistics in Medicine, 18, 681–694.

Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., and Rubin, D. (2006). Fully Conditional Specification in Multivariate Imputation. Journal of Statistical Computation and Simulation, 76, 1049–1064.

Van Buuren, S. and Oudshoorn, K. (1999). Flexible Multivariate Imputation by MICE. Technical Report, 54. Netherlands Organization for Applied Scientific Reserch (TNO).

Received March 2010
Revised January 2011