*University of Michigan School of Public Health*

The University of Michigan Department of Biostatistics Working Paper Series

*Year* 2010 *Paper* 88

# AN ANALYSIS OF NONIGNORABLE NONRESPONSE IN A SURVEY WITH A ROTATING PANEL DESIGN

Caterina Giusti* Roderick J. Little†

*University of Pisa, caterina_giusti@ds.unifi.it

†University of Michigan, rlittle@umich.edu

# AN ANALYSIS OF NONIGNORABLE NONRESPONSE IN A SURVEY WITH A ROTATING PANEL DESIGN

Caterina Giusti and Roderick J. Little

**Abstract**

Missing values to income questions are common in survey data. When the probabilities of nonresponse are assumed to depend on the observed information and not on the underlining unobserved amounts, the missing income values are missing at random (MAR), and methods such as sequential multiple imputation can be applied. However, the MAR assumption is often considered questionable in this context, since missingness of income is thought to be related to the value of income itself, after conditioning on available covariates. In this article we describe a sensitivity analysis based on a pattern-mixture model for deviations from MAR, in the context of missing income values in a rotating panel survey. The sensitivity analysis avoids the well-known problems of underidentification of parameters of non-MAR models, is easy to carry out using existing sequential multiple imputation software and has a number of novel features.

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

Caterina Giusti[1], Roderick J.A. Little[2]

**Abstract.** Missing values to income questions are common in survey data. When the probabilities of nonresponse are assumed to depend on the observed information and not on the underlining unobserved amounts, the missing income values are missing at random (MAR), and methods such as sequential multiple imputation can be applied. However, the MAR assumption is often considered questionable in this context, since missingness of income is thought to be related to the value of income itself, after conditioning on available covariates. In this article we describe a sensitivity analysis based on a pattern-mixture model for deviations from MAR, in the context of missing income values in a rotating panel survey. The sensitivity analysis avoids the well-known problems of underidentification of parameters of non-MAR models, is easy to carry out using existing sequential multiple imputation software and has a number of novel features.

*Keywords*: 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 (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

[1] Caterina Giusti is Research Fellow in Statistics at Department of Statistics and Mathematics Applied to Economics, University of Pisa, Via Ridolfi 10, 56124 Pisa, Italy (email: caterina.giusti@ec.unipi.it)
[2] Roderick J.A. Little is Richard D. Remington Collegiate Professor, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor MI 48105, U.S.A. (email: rlittle@umich.edu)

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.

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, Smith and Welch (1986) fitted a non-MAR 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, Little, Samuhel and Triest, 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 missing data that are not missing at random (NMAR) 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 NMAR 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, Reece and Zieschang (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, chap. 15), Little (1993), Kenward and Carpenter (2008) or 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, Rotnitzky and Robins (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 NMAR income nonresponse in this article adopts this approach, based on a pattern-mixture model for the data.

Published sensitivity analyses based on NMAR 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 non-MAR 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, 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. 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 NMAR analysis.

We describe here an analysis that addresses these features, 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 covariates using MAR sequential regression methods (Van Buuren and Oudshoorn, 1999; Raghunathan, Lepkowski, Van Hoewyk and Solenberger, 2001) that allow us to condition on covariate information, include income data from other quarters if available. For another application of sequential MI of income in a cross-sectional survey, see Schenker, Raghunathan, Chiu, Makuc, Zhang and Cohen (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

The labor force survey of the Municipality of Florence in Italy is an important source of information on the employment rate, the proportion of persons in search for a job and income for employed people in the Florentine area. The survey collects data in four waves every year (April, July, October, January) to produce quarterly 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 re-enters again for two waves. To determine this timing, each subject is randomly assigned to a "panel group"; the strata of the sample are equally represented in the panel groups. In any given wave, one quarter of the sample is on the first interview, one quarter on the second, one quarter on the third and one on the fourth interview. Thus, there is a 50% overlap after three and 12 months and a 25% overlap after nine and 15 months. However, the number of individuals interviewed for just one wave is usually higher because of

failure to contact some respondents more than once. In this case, substitutes from the same population stratum are interviewed.

In each of the four survey waves considered here (April, July, October 2002 and January 2003) around 1200 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 3209.

The labor force survey questionnaire begins with a question defining the occupational status. An individual is considered as employed if (s)he declares himself as such or if (s)he has worked during the preceding week; this employment 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 concerning the job search.

In this article we will focus on the missing data to the questions about occupational status and earned income for employed people. When a person is interviewed, the question defining the occupational status is always observed. Employed individuals are asked the question "What is your monthly net income?", and this question suffers from nonresponse. Note that, due to the questionnaire structure, the income considered here is only the earned income from the current job; other sources of income are not included in this survey. Table 1 summarizes the number of employed people and the corresponding percentages of missing values to the income question, separately for the panel groups.

The rates of missing values to the question on the monthly income are high compared with those of other questions in the survey, which all have rates of less than 3%. However, these percentages are comparable with those of other surveys about income, assets, expenditures and financial variables (Heeringa, Little and Raghunathan, 2002). Note that the zeros in Table 1 derive from the rotation of the panel: if the respondents were interviewed at these times, we would observe the number of employed and the percentages of missing income values also in these waves. Thus,

the rotation of the panel yields an additional source of missing data for both the occupational status and the income values, in addition to the "true" nonresponses to the income question in Table 1. Finally, note that if a person is not asked the income question because he is not employed, then the corresponding income value should be considered to be zero, not missing. For a discussion of alternative approaches to modeling financial variables with a proportion of zeros, see Buntin and Zaslavsky (2004).

**Table 1**. Number of employed people (N) and percentage of missing values (% missing) for the monthly income. The zeros in the table derive from the rotation of the survey scheme.

| Panel | April 2002 | | July 2002 | | October 2002 | | January 2002 | |
|---|---|---|---|---|---|---|---|---|
| Group | 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 |

Let $Z_{hij} = 0,1\ (h = 1,...,H, i = 1,...,n_h, j = 1,...,J)$ be the indicator of the occupational status for subject $i$ in stratum $h$ and wave $j$ of the year 2002, 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 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:

$$\hat{\bar{Y}}_{..j} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} Y_{hij} Z_{hij} w_h}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} Z_{hij} w_h}. \qquad (1)$$

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 $\hat{\bar{Y}}_{..j}$ 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:

$$\hat{\bar{Y}}_{hi.} = \frac{\sum_{j=1}^{4} Y_{hij} Z_{hij}}{\sum_{j=1}^{4} Z_{hij}}, \quad \sum_{j=1}^{4} Z_{hij} > 0 \qquad (2)$$

and then derive the overall monthly estimate as:

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\hat{\bar{Y}}_{hi.}w_h}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_h} \quad (3)$$

The results from these analyses of available data are displayed in Table 2.

**Table 2**. Number of employed people (N), mean 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 | 1327 |
| Mean estimate | 1195.2 | 1186.8 | 1309.0 | 1234.3 | 1221.2 |
| Standard error | 31.3 | 26.6 | 33.3 | 26.8 | 22.7 |

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:

$$Pr(M_{hi} \mid Y_{hi}, Z_{hi}, X_{hi}, \psi) = Pr(M_{hi} \mid Y_{\text{obs},hi}, Z_{\text{obs},hi}, X_{hi}, \psi) \quad (4)$$

where $M_{hi}$ represents the vector of missing data indicators for subject $i$ 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 income 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, $Y_{\text{mis},hi}$ and $Z_{\text{mis},hi}$, and the missing covariates under the assumption that all the values are MAR. In MI, $m$ complete datasets are produced, with missing values replaced 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 $Y$ and $Z$ proceeds as follows. A regression model is chosen for each variable with missing values: in our case 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 $ln(Y_{\text{obs},hij})$ on the $X_{hi}$ and the completed $Z_{hij}$ is fitted and also the $Y_{\text{mis},hij}$ are imputed, setting imputed values to zero when the corresponding $Z_{hij} = 0$. 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).

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 $\theta$; combine these estimates $\hat{\theta}_1,...,\hat{\theta}_m$ together with their associated variances $\hat{U}_1,...,\hat{U}_m$ through MI combining rules (Rubin 1987). In particular, the MI estimate of $\theta$ is: $\hat{\theta} = \sum_{k=1}^{m} \frac{\hat{\theta}_k}{m}$, with variance $\hat{V} = \overline{U} + (1 + m^{-1})B$, where $\overline{U} = \sum_{k=1}^{m} \frac{\hat{U}_k}{m}$ is the within-imputation variance and $B = \sum_{k=1}^{m} \frac{(\hat{\theta}_k - \hat{\theta})^2}{m-1}$ is the between-imputation variance.

We assume that the conditional distribution of each quarterly income amount may depend on the income amounts and the occupational status for all the other quarters. Specifically, at iteration $t$ of the algorithm, the imputations for the log-income $ln(Y_{hij})$ at wave $j$ for individual $i$ in stratum $h$ are drawn respectively from the distributions:

$$
\begin{aligned}
&f[ln(Y_{hi1}) \mid Z_{hi1}^{t}, ln(Y_{hi2})^{t-1}, Z_{hi2}^{t-1}, ln(Y_{hi3})^{t-1}, Z_{hi3}^{t-1}, ln(Y_{hi4})^{t-1}, Z_{hi4}^{t-1}, X_{hi}, \sigma_{11}^{t}, \beta_1^{t}] \\
&f[ln(Y_{hi2}) \mid ln(Y_{hi1})^{t}, Z_{hi1}^{t}, Z_{hi2}^{t}, ln(Y_{hi3})^{t-1}, Z_{hi3}^{t-1}, ln(Y_{hi4})^{t-1}, Z_{hi4}^{t-1}, X_{hi}, \sigma_{22}^{t}, \beta_2^{t}] \\
&f[ln(Y_{hi3}) \mid ln(Y_{hi1})^{t}, Z_{hi1}^{t}, ln(Y_{hi2})^{t}, Z_{hi2}^{t}, Z_{hi3}^{t}, ln(Y_{hi4})^{t-1}, Z_{hi4}^{t-1}, X_{hi}, \sigma_{33}^{t}, \beta_3^{t}] \\
&f[ln(Y_{hi4}) \mid ln(Y_{hi1})^{t}, Z_{hi1}^{t}, ln(Y_{hi2})^{t}, Z_{hi2}^{t}, ln(Y_{hi3})^{t}, Z_{hi3}^{t}, Z_{hi4}^{t}, X_{hi}, \sigma_{44}^{t}, \beta_4^{t}]
\end{aligned}
$$

Here, $Z_{hij}^{t} = Z_{hij}$ and $Y_{hij}^{t} = Y_{hij}$ if the values are observed ( $M_{hij} = 0$), and the conditioning on $Z_{hij}$ implies $Y_{hij} = 0$ if $Z_{hij} = 0$. 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 0 (log transformation) for each of the four income variables. Thus, we chose this transformation for our subsequent analyses, 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 NMAR sensitivity analysis discussed below.

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, Brand, Groothuis-Oudshoorn and Rubin 2006; Heeringa et al. 2002).

## 4 Results under the MAR model

We chose to impute $m = 25$ datasets with the software package IVEware (Raghunathan, Solenberger and Van Hoewyk 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 3). The number $m = 25$ yielded a stable estimate of the between-imputation component of the MI variance.

The MAR imputation scheme (equation (4)) 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 for each subject, 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 which are 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 a source of 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 as (not) employed in both, then his 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 in Table 3. 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%.

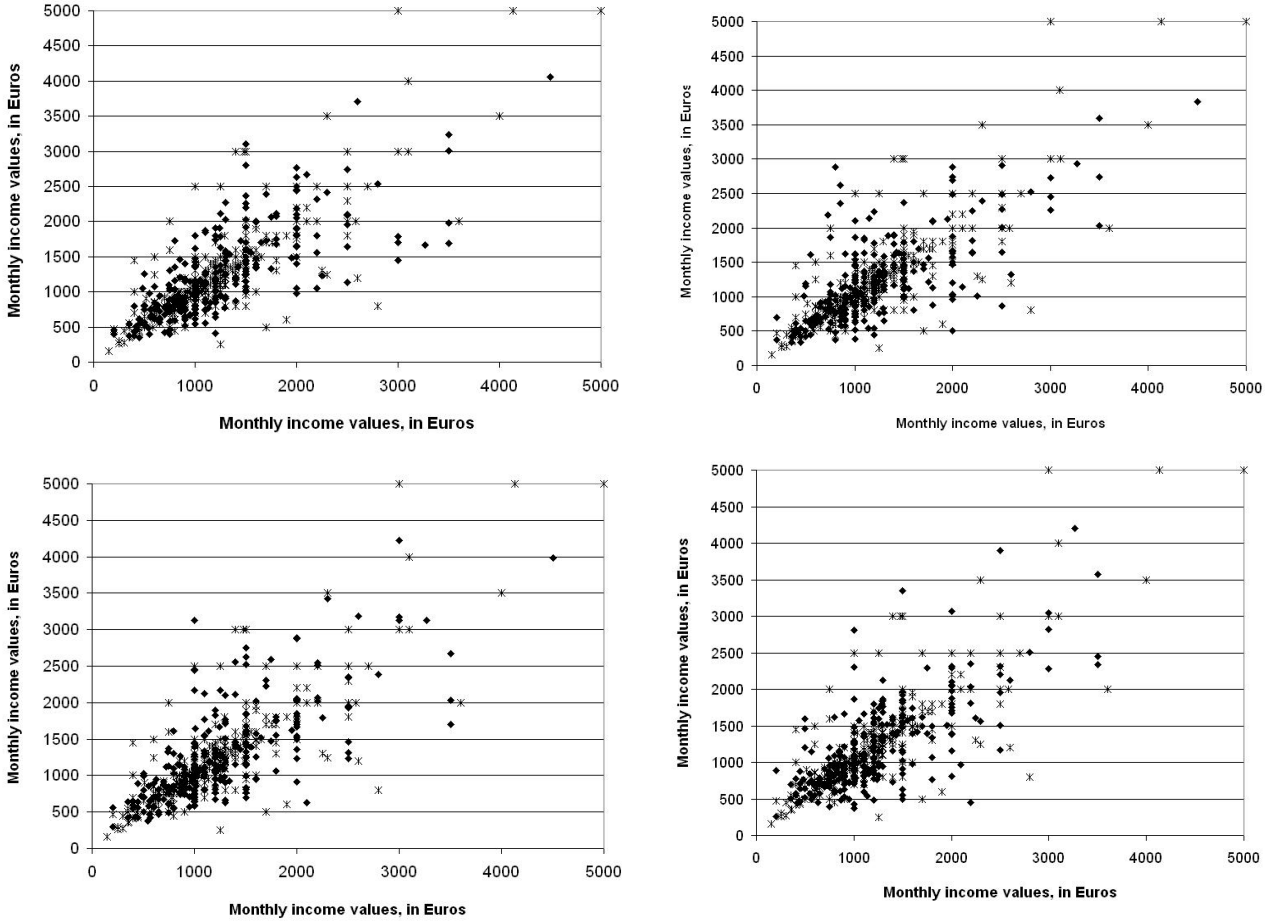
**Table 3**. Number of employed people (N) and percentage of missing values (% missing) for the monthly income across the 25 MAR multiple imputations.

| Panel | April 2002 | | July 2002 | | October 2002 | | January 2002 | |
|---|---|---|---|---|---|---|---|---|
| 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 | 0 | 0 | 0 | 0 |
| 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 | 1748 | 73.63 | 1819 | 76.04 | 1827 | 76.52 | 1780 | 75.61 |

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. Scatterplots of the two observed income values and of the observed versus imputed values are shown in Figure 1; for ease of comparison, only

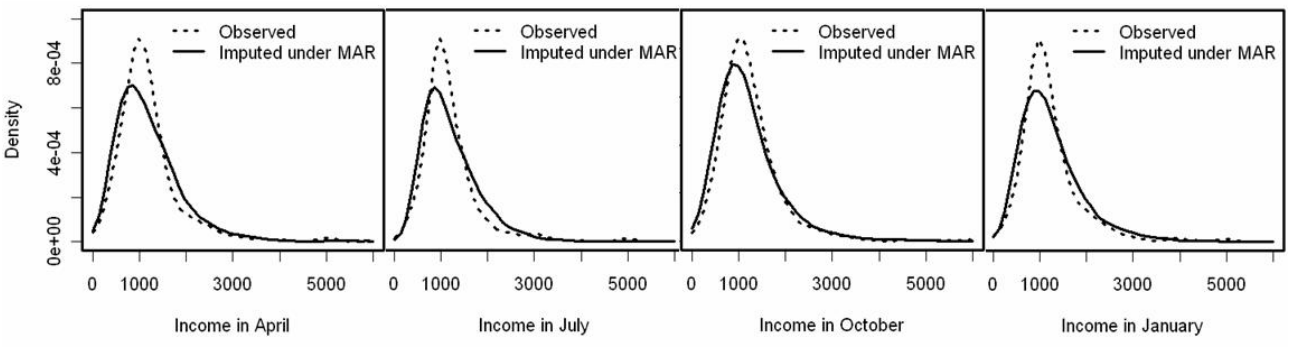
observed and imputed values under 5000 Euros are included. As we can see, the positive correlation between the observed income values is well preserved by the imputations; results are quite similar in all the imputed datasets.

**Figure 1**. Scatterplots of the couples of observed income values (stars) and of observed and imputed income values (dots) for four randomly chosen datasets.




As an additional diagnostic tool, we plotted the empirical densities of some of the 25 MAR imputed income distributions, comparing them with the density of the corresponding observed values (Figure 2). The visual examination of the empirical densities may identify potential problems when imputing in a multivariate setting (Abayomi, Gelman and Levy 2008). For each of the four income distributions we never observe 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.

**Figure 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.




We can recompute the estimates of interest, quantities (1) and (3), 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 $j = 1,...,4$ ) and referring each quarterly estimate to the preceding three months, define the personal estimate of the annual income in year 2002 as:

$$\hat{Y}_{hi\,2002} = \sum_{j=1}^{4} \hat{\bar{Y}}_{hij} * 3. \tag{5}$$

Then, the overall annual income estimate is:

$$\hat{Y}_{2002} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} \hat{Y}_{hi\,2002} w_h}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_h}. \tag{6}$$

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 $\hat{Y}_{2002}$, 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 from each imputed dataset 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 $\theta$, the estimate of interest. The fraction of missing

information can be computed as $\hat{\lambda} = \frac{r + 2/(v+3)}{r+1}$ where $r = \frac{(1+m^{-1})B}{\overline{U}}$ and $v = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B}\right]^2$ (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 4.

**Table 4**. Number of employed people (N), quarterly 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 | 1748 | 1819 | 1827 | 1780 |
| MI mean estimate | 1210.09 | 1188.21 | 1280.90 | 1249.83 |
| MI standard error | 25.48 | 28.56 | 27.67 | 25.99 |
| MI % missing info | 62.46 | 80.38 | 53.59 | 66.14 |

The differences in the distribution of income between the waves are reduced under the MAR sequential imputations, compared with the MCAR results (Table 2). 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 3) 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 that 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 in Table 5.

As we can see, the monthly income estimate for the whole year 2002 is slightly lower under the MAR method (1198.1 Euros) than under the MCAR method (1221.2 Euros, see Table 2). For

both methods the estimated median is lower than the estimated mean, reflecting a positive skew in the income distribution.

**Table 5**. Number of employed people (N), 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 | 2420 | 1086 |
| MI mean estimate | 1198.12 | 15532.00 |
| MI standard error | 17.30 | 234.49 |
| MI % missing info | 68.62 | 59.53 |
| MI median estimate | 1091.18 | 14405.16 |
| MI median standard error | 16.55 | 257.64 |
| MI 20$^{th}$ percentile estimate | 783.04 | 10755.72 |
| MI 20$^{th}$ percentile standard error | 12.51 | 243.83 |
| MI 80$^{th}$ percentile estimate | 1535.71 | 19585.32 |
| MI 80$^{th}$ percentile standard error | 27.24 | 384.12 |

## 5 Sensitivity analysis for deviations from MAR

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

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

which is a pattern-mixture factorization of the joint distribution (Little, 2003). We assume:

$$f[Y_{hij}, Z_{hij} \mid M_{hij} = 1, C_{\text{obs},hij}] = f[Y_{hij}, Z_{hij} \mid M_{hij} \neq 1, C_{\text{obs},hij}], \quad (7)$$

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:

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

This is NMAR 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

$$E[\log Y_{hij} \mid Z_{hij} = 1, M_{hij} = 2, C_{\text{obs},hij}] = E[\log Y_{hij} \mid Z_{hij} = 1, M_{hij} = 0, C_{\text{obs},hij}] + k\sigma_{hj} \quad (8)$$

where $\sigma_{hj}$ is the 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 variance $\sigma_{hj}$. 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, Boshuizen and Knook (1999).

To illustrate this NMAR model, consider an individual in panel group 5, where an individual is part of the rotating panel in waves 1 and 4, but not in the panel 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 $\sigma_{hj}$.

This model is implemented as follows:

(A) The MAR multiple imputations were 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 $NMAR_1$. We also present results under an alternative assumption (denoted $NMAR_2$), 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 $NMAR_2$ model applies an offset to the imputed values for the first and fourth wave 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 $NMAR_2$ mechanism is clearly closer to MAR than model $NMAR_1$. We think of $NMAR_1$ and $NMAR_2$ as bounding a range of plausible combinations of these models, for any given choice of $k$.

## 6 Results under the NMAR models

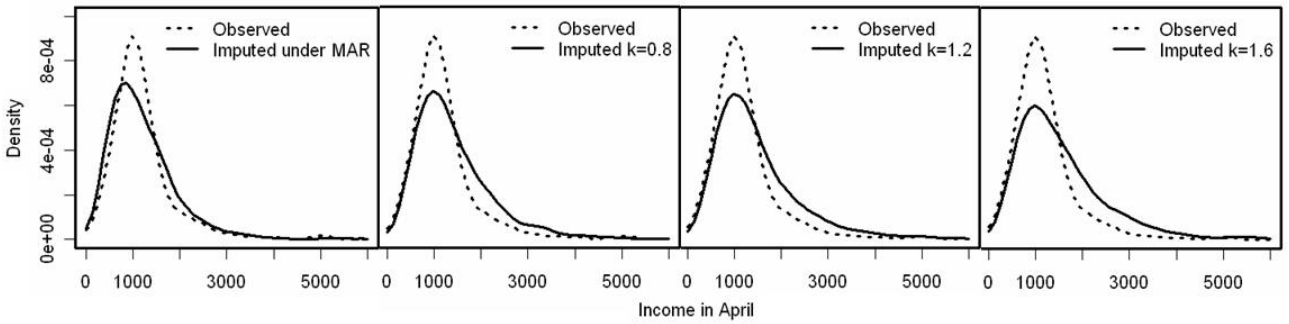
To evaluate the impact of the NMAR increments on the income distributions referring to the four quarters we plotted again the empirical densities of some of the 25 imputed income distributions, comparing them with the density of the corresponding observed values. In Figure 3 the empirical density of the observed income distribution in the first quarter (April) and the corresponding densities obtained after the MAR and $NMAR_1$ 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 $NMAR_2$ imputations, not shown here, are very similar to those in Figure 3, with the increments under the $NMAR_2$ model causing a lower shift for the distributions.

We then computed the estimates of interest for the NMAR imputed income variables. The quarterly income estimates under the $NMAR_1$ and $NMAR_2$ models are in Table 6.

**Figure 3**. Empirical densities of the observed income values in the first quarter (dotted lines) and of imputed income values (solid lines) under the MAR and NMAR1 models.




The NMAR offsets result in larger estimates than those under MAR, especially for larger values of $k$. As expected, the $NMAR_1$ assumption leads to larger increases than the $NMAR_2$ 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 $NMAR_1$ and $NMAR_2$ and for each value of $k$.

In terms of percentage increase of these estimates with respect 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 $NMAR_1$ and $NMAR_2$ mechanisms respectively. For $k = 1.2$ and $k = 1.6$ we observe a more pronounced impact of the $NMAR_1$ mechanism, especially for the monthly income estimate in the third quarter. Note that this greater increase depends on some high income values observed in the first and third quarters, already noted for the MAR model in (see Table 4); 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 3) which are incremented under the NMAR models.

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

| Model | Estimates | $\hat{\bar{Y}}_{..1}$ | $\hat{\bar{Y}}_{..2}$ | $\hat{\bar{Y}}_{..3}$ | $\hat{\bar{Y}}_{..4}$ |
|---|---|---|---|---|---|
| $NMAR_1$, $k=0.8$ | MI N | 1754 | 1791 | 1814 | 1774 |
| | MI mean estimate | 1316.9 | 1306.8 | 1421.5 | 1369.5 |
| | MI standard error | 23.9 | 20.4 | 24.2 | 24.0 |
| | MI % missing info | 44.6 | 47.3 | 28.1 | 51.2 |
| $NMAR_1$, $k=1.2$ | MI N | 1755 | 1796 | 1819 | 1770 |
| | MI mean estimate | 1390.0 | 1383.1 | 1518.7 | 1452.4 |
| | MI standard error | 25.7 | 25.6 | 31.2 | 23.7 |
| | MI % missing info | 41.4 | 59.0 | 48.0 | 35.4 |
| $NMAR_1$, $k=1.6$ | MI N | 1756 | 1780 | 1812 | 1777 |
| | MI mean estimate | 1475.9 | 1465.1 | 1605.0 | 1526.8 |
| | MI standard error | 31.3 | 28.0 | 32.6 | 30.7 |
| | MI % missing info | 50.3 | 57.2 | 44.3 | 52.4 |
| $NMAR_2$, $k=0.8$ | MI N | 1751 | 1791 | 1813 | 1771 |
| | MI mean estimate | 1290.2 | 1263.3 | 1375.7 | 1342.0 |
| | MI standard error | 24.5 | 19.9 | 24.3 | 21.0 |
| | MI % missing info | 51.1 | 51.0 | 33.1 | 38.5 |
| $NMAR_2$, $k=1.2$ | MI N | 1749 | 1787 | 1814 | 1776 |
| | MI mean estimate | 1343.3 | 1320.2 | 1439.4 | 1399.5 |
| | MI standard error | 25.4 | 20.1 | 27.1 | 26.9 |
| | MI % missing info | 47.1 | 41.9 | 38.1 | 56.3 |
| $NMAR_2$, $k=1.6$ | MI N | 1738 | 1784 | 1811 | 1784 |
| | MI mean estimate | 1416.8 | 1366.0 | 1509.2 | 1468.4 |
| | MI standard error | 27.9 | 21.5 | 27.2 | 26.0 |
| | MI % missing info | 45.2 | 39.5 | 27.9 | 41.1 |

The estimates referring to the whole year 2002 under the two NMAR hypotheses are in Table 7.

The increases for the annual estimate $\hat{\bar{Y}}$ are similar to those of the quarterly estimates (10% and 8% respectively under $NMAR_1$ and $NMAR_2$), 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 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 NMAR annual income estimates $\hat{Y}_{2002}$ are all between 15,000 and 19,000 Euros. This range is consistent from 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 16070 Euros for employees, 24400 for self-employee workers. Considering that the employee workers represent approximately the 72% in 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 18404 Euros. This value is coherent with the estimates and standard errors we obtain for $\hat{Y}_{2002}$ under the $NMAR_1$ model with $k = 1.2$ and $k = 1.6$, and for model $NMAR_2$ with $k = 1.6$.

**Table 7**. Number of employed people (N), income estimates and standard errors (in Euros) and fraction of missing information (% missing info) across the 25 $NMAR_1$ and $NMAR_2$ multiple imputations.

| | | $NMAR_1$ | | $NMAR_2$ | |
|---|---|---|---|---|---|
| $k$ value | Estimates | $\hat{\bar{Y}}$ | $\hat{Y}_{2002}$ | $\hat{\bar{Y}}$ | $\hat{Y}_{2002}$ |
| $k = 0.8$ | MI N | 2129 | 1405 | 2129 | 1410 |
| | MI mean estimate | 1322.3 | 16762.0 | 1285.0 | 16381.0 |
| | 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 | 1198.3 | 15319.0 | 1163.2 | 14923.7 |
| | MI median standard error | 16.5 | 236.0 | 14.8 | 250.8 |
| | MI 20th percentile | 854.8 | 11259.4 | 836.3 | 11002.8 |
| | MI 20th percentile standard error | 13.7 | 194.4 | 12.6 | 177.2 |
| | MI 80th percentile | 1705.4 | 21467.5 | 1650.8 | 20827.2 |
| | MI 80th percentile standard error | 29.0 | 386.5 | 25.9 | 399.5 |
| $k = 1.2$ | MI N | 2137 | 1403 | 2119 | 1414 |
| | MI mean estimate | 1398.3 | 17837.0 | 1350.2 | 16921 |
| | MI standard error | 17.3 | 252.6 | 16.5 | 247 |
| | MI % missing info | 47.4 | 52.1 | 45.2 | 52.9 |
| | MI median estimate | 1258.9 | 16145.8 | 1211.7 | 15245.6 |
| | MI median standard error | 17.3 | 266.7 | 17.7 | 248.8 |
| | MI 20th percentile | 886.5 | 11774.9 | 859.2 | 11087.3 |
| | MI 20th percentile standard error | 14.8 | 203.1 | 12.3 | 204.7 |
| | MI 80th percentile | 1817.5 | 22863.4 | 1747.5 | 21725.4 |
| | MI 80th percentile standard error | 32.1 | 480.3 | 30.4 | 425.9 |
| $k = 1.6$ | MI N | 2119 | 1414 | 2114 | 1418 |
| | MI mean estimate | 1484.1 | 18772.0 | 1420.5 | 17590.5 |
| | 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 | 1322.6 | 16917.4 | 1258.3 | 15764.3 |
| | MI median standard error | 18.7 | 279.4 | 19.2 | 264.1 |
| | MI 20th percentile | 924.5 | 12122.9 | 880.8 | 11241.4 |
| | MI 20th percentile standard error | 14.4 | 225.0 | 12.5 | 186.7 |
| | MI 80th percentile | 1941.0 | 24241.4 | 1850.4 | 22730.9 |
| | MI 80th percentile standard error | 35.2 | 511.6 | 32.5 | 510.7 |



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 case of nonresponse. This survey estimated an annual mean net income from job 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 being based on around 200 units, suggest that the Florentine area is richer than 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 NMAR models. Broadly speaking, we can say that the NMAR deviations from the MAR estimates are moderate, especially under the $\text{NMAR}_1$ model.

## 7 Conclusion

We have described the use of sequential multiple imputation 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 NMAR model is based on a pattern-mixture factorization, and extends existing NMAR 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 NMAR 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, chap. 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 if they are thought appropriate; for example the standard deviation of the predictive distribution of log income might be increased if it was thought that the income values for nonrespondents are more dispersed than those predicted under the MAR model.

**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, 475–92.

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

Kenward, M. G., and Carpenter, J. R. (2008), Multiple Imputation, in *Longitudinal Data Analysis*,

eds. G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, New York: CRC Press, pp. 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*, eds. G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, New York: CRC Press, pp. 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 *American Statistical Association Proceedings of the Social Statistics Section*, Washington: American Statistical Association, pp. 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, Univeristy 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, in *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.

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