Austrian Journal of Statistics
April 2015, Volume 44, 59–71.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.77

MI Double Feature: Multiple Imputation to Address Nonresponse and Rounding Errors in Income Questions

Jörg Drechsler
Institute for Employment Research

Hans Kiesl
OTH Regensburg

Matthias Speidel
Institute for Employment Research

Abstract

Obtaining reliable income information in surveys is difficult for two reasons. On the one hand, many survey respondents consider income to be sensitive information and thus are reluctant to answer questions regarding their income. If those survey participants that do not provide information on their income are systematically different from the respondents (and there is ample of research indicating that they are) results based only on the observed income values will be misleading. On the other hand, respondents tend to round their income. Especially this second source of error is usually ignored when analyzing the income information.

In a recent paper, Drechsler and Kiesl (2014) illustrated that inferences based on the collected information can be biased if the rounding is ignored and suggested a multiple imputation strategy to account for the rounding in reported income. In this paper we extend their approach to also address the nonresponse problem. We illustrate the approach using the household income variable from the German panel study "Labor Market and Social Security".

Keywords: heaping, measurement error, multiple imputation, nonresponse, poverty rate.

1. Introduction

Reliable information on individual and household income is difficult to obtain. Most administrative data sources contain only specific sources of income such as income from earnings or program participation and often only cover a subset of the population (self-employed are usually not included). Thus, most agencies rely on household surveys to collect information on total income. However, inferences based on the collected income information might be biased for two reasons: First, income is considered sensitive information and many survey participants are reluctant to answer questions on their personal income. Second, most respondents do not remember their exact income, especially if they are asked to provide an estimate for their total income including income from earnings, assets, transfers, etc. Respondents often round their income in this case, implicitly incorporating their uncertainty regarding the true value.

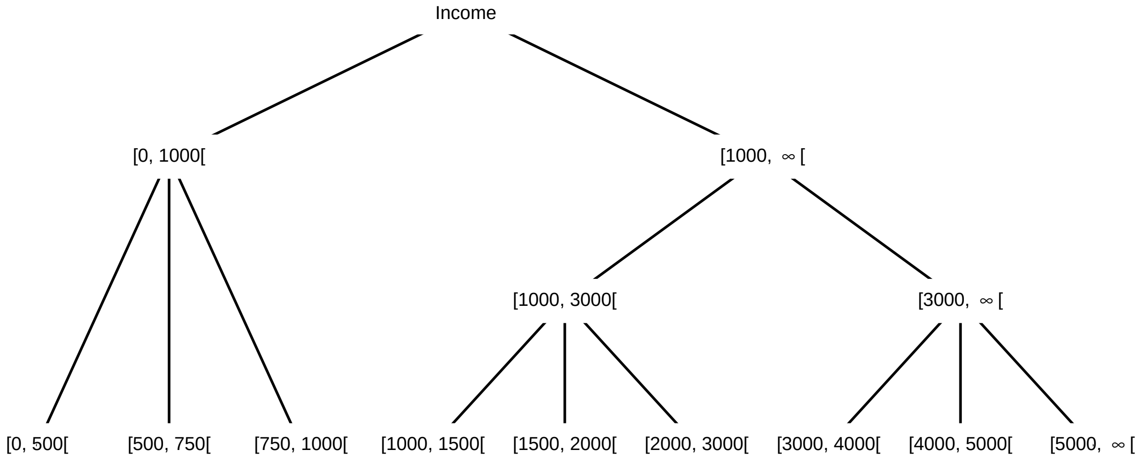

Figure 1: Implied income intervals based on partial income information collected from respondents unwilling to provide their exact income.

Nonresponse can bias inferences if the respondents are systematically different from the nonrespondents. For example, it seems plausible to assume that younger survey respondents are less concerned with confidentiality violations and the protection of sensitive information ("generation Facebook") and thus, their response rates to income questions will be higher. Since income usually increases with age, individuals with lower income will be over-represented among the respondents in this case and the average income of the population will be underestimated if only the observed income values are used.

To reduce the risk of nonresponse bias, many surveys try to obtain at least partial income information for those survey participants that are unwilling or unable to provide exact income information by asking whether the income lies in certain pre-specified intervals. Often subsequent questions further narrow down the interval in which the true income falls. Figure 1 provides an example how (partial) income information is collected in the German panel study "Labor Market and Social Security" (PASS) (Trappmann, Gundert, Wenzig, and Gebhardt 2010). Respondents are first asked for an estimate of their total household income. If they are unwilling or unable to provide this information, the interviewer provides a first threshold ($1,000$ euros) and asks whether the income is above or below that threshold. Depending on the answer to this question the survey participant is asked to choose from three specific intervals (if the respondent reported an income below $1,000$ euros for the first question) or a new threshold ($3,000$ euros) is provided and the respondent is asked again whether his or her income is above or below this threshold. If the respondent provides an answer to the second threshold question, three different income intervals are offered for both response options and the respondent is asked to pick the interval in which his or her income falls. Figure 1 illustrates the decision steps and the corresponding income intervals that are implied by the responses to each of the questions. The interview process could terminate in any of the nodes of the decision tree. For example, a respondent might refuse to provide the exact income information but might be willing to provide the information that his or her income is larger than $1,000$ but less than $3,000$ euros. However, he or she might be unwilling to further specify whether the income is in the interval $[1,000,\ 1,500[$ or $[1,500,\ 2,000[$ or $[2,000,\ 3,000[$.

Asking those respondents that are unwilling to provide their exact income for information regarding the interval in which their income falls is a successful strategy to reduce the nonresponse rate. For example, in wave six of the PASS survey, 76.96% of the respondents who are unwilling or unable to provide their exact income provided some information on the interval in which their income falls, reducing the initial nonresponse rate from 4.56% to 1.05%.

Following this procedure, the collected income information consists of exact information for those respondents that are willing to answer the exact income question and interval informa-

Table 1: Percentage of reported monthly household income values that are divisible by a given round number in the PASS survey for the year 2008/2009.

Income divisible by	1,000	500	100	50	10	5
Relative frequency (%)	13.97	23.94	61.57	69.58	80.71	84.13

tion of different lengths for those individuals that answer (some of) the interval questions. Directly obtaining valid inferences from this type of data is not straightforward, especially if refusal to answer any of the income questions should also be taken into account. In this paper we will present an imputation approach that simplifies the analysis of the collected income data. The multiple imputation methodology is not only used to impute the missing values; plausible exact income values are also generated for those respondents that only provided interval information regarding their income. The obtained imputed income data can be analyzed as if the exact income would have been obtained for all respondents. The additional uncertainty implied by the fact that only partial information is available for some of the respondents is correctly reflected through the multiple imputation procedure.

The negative effects of nonresponse are well known. However, the impacts of heaping, i.e., rounding to certain numbers such as multiples of 5, 10, 100, etc., are less studied. Rounding is a common phenomenon in surveys. Most quantitative variables such as questions on expenditure or subjective beliefs (*How likely is it that...*) show some form of rounding (Manski and Molinari 2010). But also questions on timing of events (Huttenlocher, Hedges, and Bradburn 1990) or smoking behavior typically are affected (Wang and Heitjan 2008). In a recent experimental study Ruud, Schunk, and Winter (2013) demonstrated that the amount of rounding increases with the level of uncertainty the respondent feels regarding the quantity he or she is asked for. Regarding questions on income the level of uncertainty is usually very high. Most respondents do not know their income from earnings to the exact euro amount (especially if the earnings before taxes is requested) and exact values for other sources such as monthly income from savings are even more difficult to provide. Thus, it is not surprising that questions on income usually show a large degree of rounding. Table 1 provides the percentage of the reported monthly income values that are divisible by a given round number obtained from the PASS survey for the year 2008/2009 (see Section 4 for a description of the survey). It seems that most of the reported data are rounded to some extent. More than 60% of the reported income values are divisible by 100 and only about 15% of the data are not divisible by 5.

Drechsler and Kiesl (2014) illustrate that heaping in income data can cause substantial bias in important measures such as the poverty rate. They also suggest a strategy for dealing with the problem and demonstrate its merits through simulations and real data applications. The basic idea is to model the rounding behaviour given the reported income value and then to replace the reported value by multiple plausible candidates for the true value that would have been observed if the respondent had not have rounded his or her income. A related idea has been proposed by Heitjan and Rubin (1990) for heaped age data and has later been applied in a number of papers to model the smoking behaviour based on reported cigarette counts (Heitjan 1994; Wang and Heitjan 2008; Wang, Shiffman, Griffith, and Heitjan 2012). The major advantage of the approach is that the imputed values can be treated as true values in any analysis following the imputation, i.e., it is not necessary to develop adjustment methods for each type of analysis separately. The analyst only needs to repeat the analysis of interest on each imputed dataset using standard analysis techniques. The final inferences are obtained using standard multiple imputation combining rules (Rubin 1978, 1987).

In this paper we extend the approach by Drechsler and Kiesl (2014) in order to address (partial) nonresponse and heaping simultaneously. We review the approach of Drechsler and Kiesl (2014) in Section 2 and discuss the necessary extensions to incorporate the interval information and to adjust for nonresponse in Section 3. In Section 4 we illustrate the approach based on data from the PASS survey. The paper concludes with some final remarks.

2. Strategies to adjust for rounding errors

This section discusses the imputation approach suggested by Drechsler and Kiesl (2014) which itself is based on an idea by Heitjan and Rubin (1990). In their paper Heitjan and Rubin (1990) proposed to use multiple imputation to correct for heaped reported age values of young children in Tanzania. The section borrows heavily from Drechsler and Kiesl (2014) and we refer the reader to this paper for a more detailed discussion of the methodology.

To obtain imputed income values that are adjusted for potential rounding, we need two models: one for the true income and one for the rounding behaviour. Following common practice, we model the conditional distribution of the household income Y given some covariates X by a log-normal distribution (see, for example, Clementi and Gallegati (2005) for a motivation for this model):

$$\log(Y)|X \sim N(X'\beta,\ \sigma^2). \tag{1}$$

We only consider rounding to the nearest multiple of c, which corresponds to the rounding function $f_c : x \mapsto c \cdot \lfloor x/c + 1/2 \rfloor$ and which we call rounding of degree c . Other rounding models could be considered: for example, Heitjan and Rubin (1990) suggest a model in which some age values are truncated and not rounded. However, we feel that rounding to the nearest multiple of c is the most plausible rounding strategy for income data. In our model, no rounding at all will be called rounding of degree 0. We assume that there are p possible degrees of rounding $c_1 < ... < c_p$. Typically, the set of c_i's consists of values such as 0, 1, 5, 10, 50, 100. For a given household, our model for the degree of rounding is an ordered probit model, i.e., we assume a normally distributed latent variable G which may (linearly) depend on the logged income $\log(Y)$ and some covariates Z (where some or all components of Z might be in X and vice versa):

$$G|\log(Y), Z \sim N(\gamma_0 + \gamma_1 \cdot \log(Y) + Z'\gamma_2,\ \tau^2)$$

Rounding of degree c_1 occurs, if $G < k_1$; rounding of degree c_i $(1 < i < p)$ occurs, if $G \in [k_{i-1}, k_i[$; rounding of degree c_p occurs, if $G \geq k_{p-1}$. The $p - 1$ threshold values $k_1 < k_2 < ... < k_{p-1}$ are unknown model parameters.

We assume that given X, $\log(Y)$ and Z are independent, and analogously, given Z, G and X are independent. Under these assumptions $\log(Y)$ and G have the following bivariate normal distribution given X and Z:

$$\log(Y), G|X, Z \sim N(\mu,\ \Omega), \quad \text{where}$$

$$\mu = \begin{pmatrix} X'\beta \\ \gamma_0 + X'\gamma_1\beta + Z'\gamma_2 \end{pmatrix}, \tag{2}$$

$$\Omega = \begin{pmatrix} \sigma^2 & \gamma_1\sigma^2 \\ \gamma_1\sigma^2 & \tau^2 + \gamma_1^2\sigma^2 \end{pmatrix}. \tag{3}$$

To impute true income values based on these models, it is necessary to derive the likelihood for all the unknown parameters $\Psi = (\beta,\ \sigma^2,\ \gamma_1,\ \gamma_2,\ k_1,\ ...,\ k_{p-1})$ (we need to fix γ_0 at 0 and τ^2 at 1 to make the ordered probit model identifiable). Let s_i be the observed income of household i. It can be shown that this likelihood is given as (see Drechsler and Kiesl (2014) for details)

$$\begin{aligned} L(\Psi|s,\ x,\ z) &= \prod_i f(s_i,\ x_i,\ z_i|\Psi) \\ &= \prod_i f(x_i,\ z_i) \cdot \prod_i f(s_i|x_i,\ z_i,\ \Psi) \\ &\propto \prod_i \iint_{A(s_i)} f(g,\ \log(y)|x_i,\ z_i,\ \Psi) d\log(y) dg, \end{aligned} \tag{4}$$

where $A(s_i)$ is the set of $(g,\ \log(y))$ that are consistent with an observed s_i.

Maximizing this likelihood will provide the parameter vector Ψ necessary for the imputations. To approximate a draw from the posterior distribution of $f(\Psi|s,\ x,\ z)$ under the assumption of flat priors for all parameters, we can draw from

$$\Psi^* \sim MVN(\hat{\Psi}_{ML},\ I(\hat{\Psi}_{ML})),$$

where $\hat{\Psi}_{ML}$ contains the maximum likelihood estimates of Ψ, and $I(\hat{\Psi}_{ML})$ is the negative inverse of the Hessian matrix of the log-likelihood with $\hat{\Psi}_{ML}$ plugged in.
To impute exact income values, Drechsler and Kiesl (2014) suggest a simple rejection sampling approach:

1. Draw candidate values for $(\log(y_i)^{imp},\ g_i)$ from a truncated bivariate normal distribution with mean vector (2) and covariance matrix (3) (using parameters from Ψ^*), where the truncation points are given by the maximal possible degree of rounding given the observed income s_i (for example, for an observed income value 850 with possible degrees of rounding 1, 5, 10, 50, 100, 500, and 1,000, $\log(y_i)$ is bounded by $\log(825)$ and $\log(875)$ and g_i has to be in $]-\infty,\ k_4^*[$).

2. Accept the drawn values if they are consistent with the observed rounded income, i.e., rounding the drawn income value according to the drawn rounding indicator gives the observed income s_i, and impute $\exp(\log(y_i)^{imp})$ as the exact income value.

3. Otherwise draw again.

Repeating this procedure m times provides m imputed datasets that properly reflect the uncertainty from imputation.

3. Extensions for (partial) nonresponse

As discussed in the introduction, many agencies ask respondents who refuse to answer the exact income question whether they would be willing to provide information in which given interval their income falls. This partial information can be used to improve the inferences regarding the income variable. In this paper we suggest to use this partial information when setting up the likelihood and then to impute plausible true income values for each reported income interval. The approach is related to the approach to account for rounding described in the previous section with the only difference that the interval in which the true income must fall is known in advance and does not need to be estimated from the observed data.

Let r_i, $r_i \in \{0,\ 1,\ ...,\ R+1\}$, be a random variable that identifies to which income response group individual i, $i = 1, ..., n$ belongs. Let $r_i = 0$ represent exact income information (which might still be affected by rounding) and let $r_i = 1,\ ...,\ R$ identify the R different income intervals that could be selected from the predefined intervals provided by the agency. For example, according to Figure 1 $R = 13$ in the PASS survey. Finally, let $r_i = R+1$ represent refusal to provide any income information at all. Let I_i^r be an indicator function that equals 1 if individual i belongs to income response group r and equals 0 otherwise. Let l^r and u^r be the upper and lower bound of the income interval for response group r. We set $l^0 = y = u^0$ and $l^{R+1} = -\infty$ and $u^{R+1} = +\infty$. All other bounds are defined by the income intervals provided by the agency. We extend the definition of s_i to also include all reported income intervals, i.e., s_i is a single value for all individuals that reported the exact income, but is an interval for all individuals that only provided the information in which interval their income falls. The extended likelihood that also takes the interval information into account is given by

$$\begin{aligned} L(\Psi|s,\ x,\ z) &= \prod_i f(x_i,\ z_i) \cdot \prod_i f(s_i|x_i,\ z_i,\ \Psi) \\ &\propto \prod_i \{ (\iint_{A(s_i)} f(g,\ \log(y)|x_i,\ z_i,\ \Psi) d\log(y) dg)^{I_i^0} \\ &\cdot \prod_{r=1}^{R+1} [F(\log(u_i^r),\ \mu_i = x_i'\beta,\ \sigma^2) - F(\log(l_i^r),\ \mu_i = x_i'\beta,\ \sigma^2)]^{I_i^r} \}. \end{aligned} \tag{5}$$

Once estimates for all parameters are obtained by maximizing the likelihood in (5), imputation of the plausible values for the true income Y is straightforward. The first imputation step is similar to Section 2: Approximate a draw from the posterior distribution of the parameters by drawing from a multivariate normal with mean equal to the maximum likelihood estimates of the parameters and variance equal to the negative inverse of the Hessian matrix of the log-likelihood. The second step depends on the type of data that is imputed. The true income for all exact reporters is imputed as described in Section 2 to account for potential rounding in the reported income values. The true income for the interval respondents is imputed by drawing from a truncated normal distribution $N_t(\mu, \sigma^2)$ with $\mu = X'\beta^*$, $\sigma^2 = (\sigma^*)^2$, where β^* and $(\sigma^*)^2$ are the drawn parameters from step one. The truncation points are given by the bounds of the reported income interval. Finally, imputations for those respondents that refused to provide any information regarding their income are obtained by drawing from a normal distribution with parameters $\mu = X'\beta^*$ and $\sigma^2 = (\sigma^*)^2$.

4. Application to the panel study Labor Market and Social Security

We illustrate the application of our approach using data from the German panel study "Labor Market and Social Security" (PASS). To enable a comparison of our extended approach with the approach of Drechsler and Kiesl (2014) that only focuses on rounding, we use the same models for the income and rounding behaviour and also use the poverty rate to evaluate which impacts the adjustments have on important measures that are regularly computed from income data. The poverty rate is defined as the percentage of persons with an income less than a fixed percentage of the median income. For example, in the European countries the poverty rate is defined as the proportion of persons with an income less than 60% of the median income.

Before presenting the results, we provide a description of the data and a short summary of the imputation models borrowed from Drechsler and Kiesl (2014). The interested reader is referred to this paper for more details.

The PASS survey started in 2006 and conducted yearly ever since, aims at measuring the social effects of labour market reforms. The survey consists of two different samples, each containing roughly 6,000 households. The first sample is drawn from the Federal Employment Agency's register data containing all persons in Germany receiving unemployment benefit for long time unemployment. The second sample is drawn from the MOSAIC database of housing addresses collected by the commercial data provider, microm. This sample is representative for the resident population in Germany. The stratified sampling design for this sample oversamples low-income households. The major benefit of this combination of two different samples lies in the fact that control groups for the benefit recipients can easily be constructed. The panel contains a large number of socio-demographic characteristics (for example, age, gender, marital status, religion, migration background), employment-related characteristics (for example, status of employment, working hours, income from employment, employment history), benefit-related characteristics (for example, benefit history, amount of

Table 2: Covariates included in the income model.

variable	characteristics
household size	5 categories (household sizes> 4 set to “5 or more”)
deprivation index	range: 0–21
living space	range: 7–903 square meters
type of household	8 categories
amount of debt	7 categories
income from savings	yes/no
age of respondent	range: 15–99
amount of savings	8 categories (not available for wave 1)
unemployment benefits	yes/no
weight	range: 24.95–186,000

benefits, participation in training measures), and subjective indicators (for example, fears and problems, employment orientation, subjective social position). A detailed description of the survey can be found in Trappmann *et al.* (2010).

To model the true income, we assume a log-normal distribution for income conditional on a set of covariates X. Details about the covariates included in the model are contained in Table 2.

All variables are standardized, some sparsely populated categories in X are collapsed and influential outliers are removed to ensure convergence of the maximisation procedure (see Drechsler and Kiesl (2014) for details). For the rounding behaviour, we assume that the tendency to round only depends on the true income.

4.1. Evaluation of the model assumptions

Since the proposed rounding adjustment strategy is purely model based, an evaluation of the model assumptions is essential. We follow the approach of Drechsler and Kiesl (2014) to check whether the model assumptions are reasonable. They suggest to use posterior predictive simulations (Gelman, Carlin, Stern, and Rubin 2004, Chap. 6) for the evaluations since the true income and the rounding behaviour are never observed which complicates the evaluation.

The income model

For the income model evaluation we generate a very large number of imputations for the true income based on the parameters obtained from maximizing the likelihood in (5) at the last iteration of the sequential regression imputation procedure (see Section 4.2 for details). The rounding behaviour is completely ignored here, i.e., imputations are generated for all observations based on the marginal income model described in (1). The obtained imputations can be seen as samples from the posterior predictive distribution of the income for each observation according to the model. To evaluate the model fit we can check whether these posterior distributions cover the observed income values from the original data. Of course many of the observed income values are subject to rounding, so we limit the evaluation to those records for which we can be sure that the reported value is only rounded to the next euro (i.e., all records for which the reported value is only divisible by 1). If the imputation model is correct, the true (observed) income should be covered in the region between the empirical $\alpha/2$ quantile and the $1-\alpha/2$ quantile of the imputed values with a probability of $1-\alpha$. Thus, as a measure for the model fit we calculate the fraction of unrounded income values from the observed data that are covered by this interval computed from the imputed values and compare this fraction to the expected coverage rates. Results based on $m = 1,000$ imputations are presented in Table 3. The empirical coverages are generally close to the nominal coverages: except for wave 2 and 5 the empirical coverages never differ more than

Table 3: Percentage of true income values from the PASS survey that are covered in the defined regions of the posterior distribution of the imputed income values.

Expected	Empirical Coverage (in %)					
Cov. (in %)	wave 1	wave 2	wave 3	wave 4	wave 5	wave 6
99.00	97.65	93.76	97.31	97.19	95.43	96.87
95.00	95.06	91.63	93.34	93.57	92.69	93.66
90.00	91.91	89.00	89.72	89.31	88.55	89.53

Table 4: Percentage of income values that are divisible by a given round number (but not by any of the larger numbers) in the observed PASS data, the unrounded data, and the re-rounded data.

Income divisible by	1	5	10	50	100	500	1,000
Observed income (%)	14.94	4.05	11.58	7.74	37.34	10.29	14.06
Unrounded income (%)	80.05	9.98	7.97	1.00	0.79	0.11	0.10
Re-rounded income (%)	9.67	2.93	12.10	9.49	45.79	10.08	9.94

2.2 percentage points from the nominal coverages. The largest differences are observed for the expected 99% coverage rate for wave 2 (difference of 5.24 percentage points) and wave 5 (3.57 percentage points). But even for these waves the nominal coverages never differ more than 1.5 percentage points from the expected 90% coverage rate. Overall the results indicate a reasonable fit for the income model.

The rounding behaviour model

To evaluate the quality of the rounding behaviour model, we repeatedly re-round the imputed (unrounded) income variable based on the obtained likelihood parameters and compare it to the originally observed data. Specifically, we repeatedly ($m = 100$) generate unrounded income data that are consistent with the original data according to the joint model for income and rounding behaviour. Then, we repeatedly round each of the obtained exact income variables (100 times for each of the generated income variables) according to the rounding probabilities based on the parameters from the rounding behaviour model. Since we have no direct measure for the rounding behaviour we use a proxy for the evaluation. We compare the share of the income values that are divisible by values that are typically used as rounding bases. Table 4 lists these shares for the original data, the re-rounded data (computed as the average across the 10,000 generated datasets) and the unrounded data (computed as the average across the $m = 100$ replicates). Each column reports the percentage of records for which the given number represents the maximum possible rounding base, i.e., these records would not be divisible by any of the larger rounding bases listed in the table. The results are pooled across all waves of the PASS data for readability. Similar results were obtained when looking at each wave individually.

As expected the percentages differ substantially between the observed income and the unrounded income. Most of the values (80.05%) in the unrounded data (second row in the table) are only divisible by one and the percentages decrease quickly as the rounding base increases (note that we assume that values in the unrounded data are always rounded to the nearest euro). This is different for the observed data (first row). Only 14.94% of the data are only divisible by 1 and 37.34% of the records have a maximum rounding base of 100. The divisibility of the re-rounded data (third row) is reasonably close to the observed data. Again, most records are in the category with a maximum rounding base of 100, although the percentage of records that fall into this category is slightly overestimated (45.79%). This overestimation leads to a slight underestimation of the percentage of records that are only divisible by one (9.67%). For most of the remaining categories the percentages based on the re-rounded data are fairly close to the percentages based on the observed data: the difference in percentage points is less than 1.2 for the rounding bases 5, 10, and 500. The percentage of records with maximum rounding bases of 50 and 1,000 differ somewhat more between the observed and the re-rounded data (1.75 and 4.12 percentage points respectively). Overall the results indicate a reasonable fit of the rounding behaviour model.

4.2. Results

We compare three different approaches to estimate the poverty rates from the six waves of the PASS survey that are available so far. In the first approach we treat the reported income as the true income and only use the information from those respondents that answered the exact income question. To keep the results consistent with the second approach described below, we also exclude the respondents that provided an answer to the exact income question but did not provide an answer for at least one of the covariates listed in Table 2. This approach assumes that the reported income is never rounded and implies that the respondents to the exact income question are not systematically different regarding their income from those that only provide income intervals, completely refuse to provide any information regarding their income, or have missings in the list of covariates, i.e., this approach assumes that the income information is missing completely at random (MCAR) in the terminology of Rubin (1976). In the second approach we use the methodology of Drechsler and Kiesl (2014) to account for the rounding but still only use the data from respondents who provided an answer to the exact income question and all the covariates, i.e., we still assume MCAR. The final approach is the extended approach described in this paper which also takes the information from the interval respondents into account and imputes the missing information in the covariates and missing income information for those survey participants that completely refused to provide any information regarding their income. We note that this approach uses more information to estimate the parameters in the imputation model and only assumes that the income information is missing at random (MAR), i.e., the missingness can be explained by the covariates included in the imputation model.

We apply the models described above separately for each year (the variable *amount of savings* is not available in the first wave of the survey and is thus excluded from the income model in that year). For the third approach the imputation routine for the true income is incorporated into a sequential regression multivariate imputation (SRMI, Raghunathan, Lepkowski, van Hoewyk, and Solenberger (2001)) procedure to impute missing values in any of the covariates. With the SRMI approach missing values in any of the variables are imputed by iteratively drawing from the conditional distributions of each variable given all the other variables. The process of iteratively drawing from the conditional distributions can be viewed as a Gibbs sampler that will converge to draws from the theoretical joint distribution of the data if this joint distribution exists. This is not guaranteed in practice. However, Liu, Gelman, Hill, Su, and Kropko (2013) show that consistent results can still be obtained if the conditional models are correctly specified.

To improve the quality of the imputations we included some additional variables in the imputation models for the covariates. We treated the first 100 iterations of the Gibbs sampler in each wave as the burn-in phase to ensure convergence and stored every 5th iteration after the burn in phase as one imputed dataset. Traceplots of all variable means and variances and the Heidelberger&Welch diagnostic (Heidelberger and Welch 1983) indicated that all Gibbs samplers converged after 90 iterations and autocorrelation plots showed no significant correlation after 3 iterations.

Table 5 presents the poverty rates for the different waves. The estimated poverty rate is based on the disposable income, i.e., the reported income is adjusted for the number of household members and the age of the household members as suggested by the OECD (see, for example, Eurostat (2014a)). The first column contains the number of cases for the available case procedures of approach one and two. The second column contains sample sizes if all missing

Table 5: Estimated poverty rates from the PASS survey (with 95% confidence intervals reported in brackets).

Wave	n_{obs}	n_{imp}	Original data	Rounding adjustment	Nonresponse and rounding adjustment
Wave 1	10,214	12,791	17.29 (15.81;18.77)	16.35 (15.14;17.55)	16.60 (15.48;17.71)
Wave 2	7,311	8,428	16.91 (15.79;18.03)	16.98 (15.69;18.27)	16.39 (15.15;17.63)
Wave 3	8,169	9,534	14.27 (12.28;16.27)	15.40 (13.91;16.90)	15.66 (14.35;16.97)
Wave 4	6,538	7,845	14.89 (13.44;16.35)	14.61 (13.40;15.81)	14.81 (13.61;16.02)
Wave 5	8,623	10,232	16.34 (14.81;17.87)	15.75 (14.41;17.10)	15.82 (14.35;17.29)
Wave 6	8,267	9,508	15.95 (14.49;17.42)	16.27 (14.81;17.72)	15.78 (14.47;17.09)

or partially observed values are imputed. The results based on the original data without any adjustments are presented in the third column while the results for the multiply imputed true income accounting for rounding are included in column 4. The fifth column contains the results based on all data. All imputation results are based on $m = 10$ imputations. The 95% confidence intervals reported in brackets are based on bootstrap variance estimates. We used the normal approximation to compute the confidence intervals based on the estimated variances.

Generally, the impacts of the different adjustment methods are modest. Given the large amount of uncertainty in the estimates, the 95% confidence intervals mostly overlap. Still, there is some evidence that the impact from rounding is stronger than the impact due to (partial) nonresponse in most years. While the differences between the poverty rates based on the unadjusted point estimates and the estimates that account for the rounding (column three compared to column four) range from -1.13 to $+0.94$ percentage points, the differences between the adjusted estimates and the estimates that also account for the nonresponse (column four and column five) only range from -0.26 to $+0.59$ percentage points. The nonresponse adjustments only have a stronger impact in waves 2 and 6 in which the poverty rate hardly changes between the naïve direct estimate and the adjusted estimate. The smaller impact of the nonresponse is to be expected given that only 13–20% of the records are imputed to adjust for nonresponse compared to approximately 85% of the records that are imputed for rounding adjustments. Still, the differences in the poverty rates albeit small indicate that income is not missing completely at random and ignoring the nonresponse results in biased inferences.

5. Conclusions and Outlook

Obtaining reliable income information from surveys is notoriously difficult. Income is considered sensitive information and survey respondents often find it difficult to remember their exact income. In this paper we suggested a strategy to address two common potential sources of bias: nonresponse and rounding. Our multiple imputation approach tackles both problems simultaneously and provides a simple tool to incorporate interval information when making inference based on the collected data. The application to the PASS survey showed that adjusting for these two factors can have a direct impact on politically important measures such as the poverty rate. We found that rounding has a higher impact on the results than nonresponse at least for our study. The changes in the poverty rates that we found in our empirical evaluation are modest although an increase of the poverty rate by 1.4% as observed for wave 3 of the PASS survey would likely cause some political discussions. We believe that the main reason for the relatively small changes lies in the robustness of the poverty measure which is based on the median of the income distribution. It would be an interesting area of future research to evaluate the impacts on less robust measures such as the income quintile share ratio (see, for example, Eurostat (2014b)) which computes the ratio of the 80% and the 20% quantile of the income distribution as a measure of income inequality.

Of course the adjustments proposed in this paper are based on several assumptions and it is important to critically review these assumptions. First, the correction methods are based on models and the underlying model assumptions need to be evaluated. Alternative models for the income distribution have been suggested in the literature. For example, Graf and Nedyalkova (2013) suggested to model the income distribution using the generalized beta distribution of the second kind. However, it is not straightforward to incorporate covariates in this model. Furthermore, we feel that our model evaluations in Section 4.1 indicate a good fit of the log-linear model for the conditional income distribution. Second, we assume that the income information is missing at random (MAR), i.e., the nonresponse can be explained by the variables included in the imputation model. This is a crucial assumption in most imputation models and this assumption can never be tested based on the observed data. We believe that the covariates in our model such as age of the respondent, deprivation index, or household size should help to explain the nonresponse in the data. However, if the MAR assumption does not hold, results from our imputation strategy will be biased and imputation models such as the non-ignorable models proposed in Little and Rubin (2002, Chap. 15) need to be considered. Finally, nonresponse and rounding might not be the only sources of bias in the data. Several studies found that individuals with low earnings tend to overreport their income while individuals with high income tend to underreport their income (see, for example, Pischke (1995)). Incorporating this additional measurement error into the adjustment strategy would be an interesting area of future research.

Acknowledgements: We thank one anonymous referee and the editor for thoughtful suggestions which helped to improve the paper. This work was partially supported by the DFG grants DR 831/2-1 and KI 1368/1-1.

References

Clementi F, Gallegati M (2005). "Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States." In A Chatterjee, S Yarlagadda, B Chakrabarti (eds.), *Econophysics of wealth distributions*, pp. 3–14. Milan: Springer.

Drechsler J, Kiesl H (2014). "Beat the Heap – An Imputation Strategy for Valid Inferences from Rounded Income Data." *IAB Discussion Paper 2/2014*, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Eurostat (2014a). "Glossary: Equivalised Disposable Income - Statistics Explained (2014/11/07)." http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Equivalised_disposable_income.

Eurostat (2014b). "Glossary: Income Quintile Share Ratio - Statistics Explained (2014/11/07)." http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:S80/S20_ratio.

Gelman A, Carlin J, Stern H, Rubin D (2004). *Bayesian Data Analysis*. Second edition. London: Chapman and Hall.

Graf M, Nedyalkova D (2013). "Modeling of Income and Indicators of Poverty and Social Exclusion Using the Generalized Beta Distribution of the Second Kind." *Review of Income and Wealth, online first.*

Heidelberger P, Welch P (1983). "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research*, **31**, 1109–1144.

Heitjan D (1994). "Ignorability in General Incomplete-Data Models." *Biometrika*, **81**, 701–708.

Heitjan D, Rubin D (1990). "Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping." *Journal of the American Statistical Association*, **85**, 304–314.

Huttenlocher J, Hedges LV, Bradburn NM (1990). "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**(2), 196–213.

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data.* Second edition. New York: John Wiley and Sons.

Liu J, Gelman A, Hill J, Su YS, Kropko J (2013). "On the Stationary Distribution of Iterative Imputations." *Biometrika*, p. (online first).

Manski C, Molinari F (2010). "Rounding Probabilistic Expectations in Surveys." *Journal of Business & Economic Statistics*, **28**, 219–231.

Pischke JS (1995). "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study." *Journal of Business & Economic Statistics*, **13**(3), 305–314.

Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models." *Survey Methodology*, **27**, 85–96.

Rubin DB (1976). "Inference and Missing Data." *Biometrika*, **63**, 581–590.

Rubin DB (1978). "Multiple imputations in sample surveys." In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 20–34. Alexandria, VA: American Statistical Association.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Ruud PA, Schunk D, Winter JK (2013). "Uncertainty Causes Rounding: An Experimental Study." *Experimental Economics*, pp. 1–23.

Trappmann M, Gundert S, Wenzig C, Gebhardt D (2010). "PASS: A Household Panel Survey for Research on Unemployment and Poverty." *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, **130**, 609–622.

Wang H, Heitjan D (2008). "Modeling Heaping in Self-Reported Cigarette Counts." *Statistics in medicine*, **27**(19), 3789–3804.

Wang H, Shiffman S, Griffith SD, Heitjan DF (2012). "Truth and Memory: Linking Instantaneous and Retrospective Self-Reported Cigarette Consumption." *The annals of applied statistics*, **6**(4), 1689–1706.

Affiliation:

Jörg Drechsler
Department for Statistical Methods
Institute for Employment Research
Regensburger Straße 104
90478 Nürnberg, Germany
E-mail: joerg.drechsler@iab.de

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 44 *Submitted:* 2014-09-30
April 2015 *Accepted:* 2014-11-21