MI double feature: multiple imputation to address nonresponse and rounding errors in income questions simultaneously

Joerg Drechsler[1] and Hans Kiesl[2]

[1]Institute for Employment Research.
Regensburger Str. 104, 90478 Nuremberg, Germany, joerg.drechsler@iab.de
[2]Regensburg University of Applied Sciences,
Fakultaet IM, Postfach 12 03 27, 93025 Regensburg, Germany, hans.kiesl@hs-regensburg.de

Abstract

Questions on income in surveys are prone to two sources of errors that can cause bias if not addressed adequately at the analysis stage. On the one hand, income is considered sensitive information and response rates on income questions in general tend to be lower than response rates for other non-sensitive questions. If the missing data mechanism for income is not missing completely at random (MCAR) – and there is ample of research indicating that it isn't – results based only on the observed income values will be biased. On the other hand respondents tend to round their income. Depending on the respondent and on the value of income, the magnitude of rounding can range from rounding to the closest 5 dollar value to rounding to the closest 10000 dollar value. While this kind of measurement error will not affect statistics like the mean of the income variable if the rounding is random, analysis regarding the distribution of income such as the median income will be biased. This can be especially problematic if income thresholds are used to establish program eligibility.
In this paper we propose a two stage imputation strategy that estimates the posterior probability for rounding given the observed values on the first stage and re-imputes the observed income values given the rounding probabilities on the second stage. Missing values are also imputed at this stage. We provide a simulation study that illustrates that the proposed imputation model can help overcome the possible negative effects of rounding. We also present results based on the household income variable from the German panel study "Labor Market and Social Security".

Keywords: Measurement Error, Multiple Imputation, Income, Nonresponse, Rounding

1. Introduction

Obtaining reliable information on individual and household level income is important for various reasons. This information can for example help to estimate the rate of people at risk of poverty, measure the inequality of a society and provide information on discrimination, etc. Many political decisions such as the establishment or elimination of social security programs heavily rely on information regarding the income distribution. For these reasons, many household surveys collect income information, but measuring income in surveys is a difficult task. On the one hand income is considered sensitive information and many survey respondents are unwilling to reveal their personal income. Thus, income related questions consistently show the highest nonresponse rates among all variables in a survey. Additionally, there is ample of research indicating that the income respondents are not a random subset of the sampled units, i.e. the missing mechanism is not missing completely at random (MCAR) and thus estimates based on the observed data alone are not only less efficient but also biased.

On the other hand even if the respondent is willing to provide his or her income, he or she will often find it difficult to report the exact amount of income. This is especially true if the respondent is asked to report his or her total income including income from savings, rent, alimony, etc or if a direct estimate for the total household income is requested. Usually, the respondent tends to round the reported income to some extent. Depending on the respondent, the reporting period, and on the value of income, the magnitude of rounding can range from rounding to the closest 5 dollar value to rounding to the closest 10,000 dollar value. As a result, the reported income data have several spikes at even income values. For example, Czajka and Denmead (2008) find that regarding income for the year 2002 "28 to 30 percent of earners report amounts divisible by $5,000, and 16 to 17 percent report amounts divisible by $10,000" in the Current Population Survey (CPS) and the American Community Survey (ACS).

However, this phenomenon is not limited to those surveys that ask for the yearly income directly. Even if the monthly income is requested, respondents tend to round although obviously the typical rounding base will only vary between 5 and 1,000 dollars in this case. As an illustration Table 1 provides the percentage of the reported income values that are divisible by a given number for the year 2008/2009 of the German panel study "Labor Market and Social Security" that we use in Section 7 to illustrate our approach.

Table 1: Percentage of reported income values that are divisible by a given 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

It is obvious that most of the reported data are rounded to some extent. More than 60 percent of the reported income values are divisible by 100 and only about 15 percent of the data are not divisible by 5. Based on these results it is evident that treating the reported income as a continuous variable is more than questionable. Furthermore, the rounded income values can lead to biased results if the analyst doesn't account for the rounding as we will illustrate in Section 4.

Of course a sophisticated analyst will still be able to obtain valid results from the reported income despite the nonresponse and rounding for example by applying some combination of directly modeling the likelihood of the observed data (see for example Little and Rubin, 2002) to account for the nonresponse and modeling the interval probabilities given the observed values to account for the rounding. However, many users will not have the statistical background to apply these methods and hence will have to rely on the data providing agency to prepare the data for them in a way so that they can apply their standard methods ignoring any deficiencies in the data. Additionally, the agency might be in a better position to adjust for nonresponse and rounding errors since there might be additional information available to the agency that might help to model the nonresponse and the rounding probability which might not be included in the public use files for confidentiality reasons. Furthermore, it is reasonable to assume that the staff responsible for the survey has a better understanding of the reasons behind the data deficiencies and might thus do a better job to correct for those deficiencies than the applied analyst. For these reasons many agencies already provide public use files for which the missing values in the income variable are imputed. For example the income in the CPS is imputed using a sophisticated hot deck imputation method. Thus, the burden of dealing with the item nonresponse in the income variable is already lifted from the applied researcher for many surveys. However, the problems stemming from the rounding of the provided income are still widely ignored.

In this paper we propose a unified approach to account for nonresponse and rounding simultaneously. We suggest to use multiple imputation – widely accepted nowadays as a straightforward tool to obtain valid inferences from data subject to nonresponse – not only to account for nonresponse but also to correct the potential biases from rounding. The basic idea is to model the probability for rounding given the reported value and then to replace the reported value by multiple plausible candidates for the true value that would have been observed if the respondent would not have rounded his or her income.

The remainder of the paper is organized as follows. In Section 2 we illustrate the potential bias from rounding using one of the most influential and highly political estimates that is computed from income data: the poverty rate. In Section 3 we illustrate our imputation approach for dealing with rounding and nonresponse simultaneously. Section 4 contains a simulation study that demonstrates that the imputation approach can correct the rounding bias for the poverty rates. In Section 5 we apply the approach to the German panel study "Labor Market and Social Security". The paper concludes with a discussion of future research topics.

2. Potential bias from rounding

At first thought the damaging effect of rounding might not be that obvious. If the rounding process is completely at random measures like the average income will not be affected. However, the income distribution will change and all measures that are based on the percentiles of the distribution will be biased. This is also true for one of the most prominent figures that are routinely calculated from income data: the rate of persons that are at risk of poverty (poverty rate). This rate is usually defined as the percentage of persons with an income less than a fixed percentage of the median income. For example in the European Union member states the poverty rate is defined as the rate of

persons with an income less than 60% of the median income. This quantity intended to measure the inequality in the wealth distribution is of great political importance since it allows a direct comparison between regions and countries but also because many political decisions like establishing new labor market programs are directly influenced by this measure. For this reason even small changes in the estimated poverty rate will be followed by substantial political debates and might also have a direct impact on future political decisions. It is therefore essential that the poverty rate is estimated with the highest accuracy possible. In practice the rate is computed from the disposable household income adjusted for household size and number of children (in the U.S., a slightly different measure of poverty is used; see http://www.census.gov/hhes/www/poverty/methods/measure.html on how the U.S. Census Bureau calculates the measure).

In this section we illustrate the potential bias for the estimated poverty rate obtained from rounded income data in a simplified setting based on a simulated dataset. We don't discuss the negative effects here that nonresponse can have on the estimates especially if the missingness is not completely at random. This effect has been discussed in the literature extensively elsewhere (see for example Little and Rubin, 2002). For our simulation, we assume that the true monthly income follows a log-normal distribution with $\mu=8$ and $\sigma=0.47$. Modeling income with a log-normal distribution is standard in the economic literature and the parameters of the distribution are chosen somewhat arbitrarily to obtain an income variable that provides reasonable poverty rates from a German perspective. We further assume that the probabilities for rounding to the closest 1, 10, 100, or 1000 dollars are equal to 0.1, 0.4, 0.4, and 0.1 respectively.

We draw a sample of 5,000 records from the specified distribution and compute the income distribution, the poverty rate and the poverty threshold from the sample before and after rounding. The results are displayed in Figure 1.

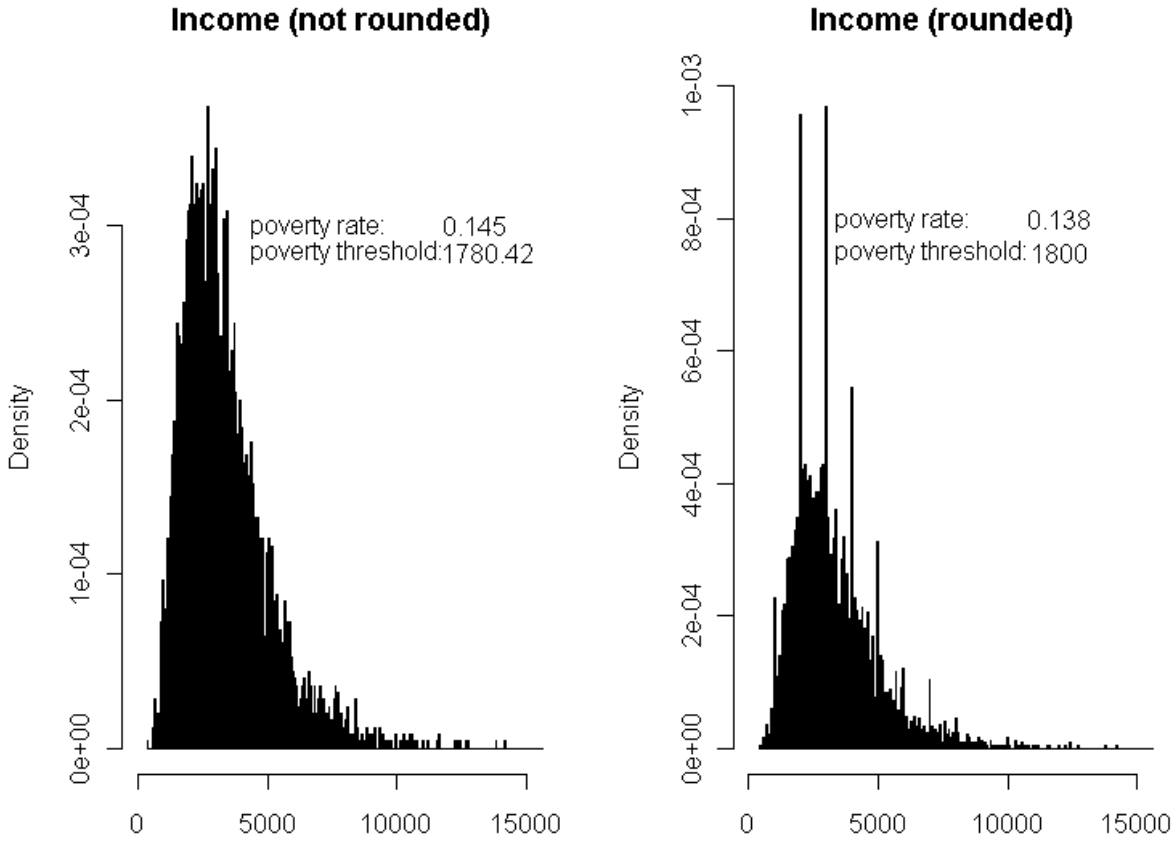

Figure 1: Income distribution, poverty rate, and poverty threshold before and after rounding

There are obvious spikes in the rounded income data at the even numbers. The poverty rate drops from 14.5 percent before rounding to 13.8 percent after rounding while at the same time the poverty threshold increases from 1780 dollars to 1800 dollars. Given that small changes in the poverty rate usually cause tremendous political debate and noting that our rounding probabilities seem to be conservative compared to the findings in Table 1, the effect on the poverty measures is relevant and using the reported data ignoring the rounding will lead to biased results.

3. Correcting the rounding error through imputation

Instead of accounting for the rounding at the analysis stage we suggest to account for the rounding at the data processing stage. We see several benefits from this approach. First, the correction can be performed by the data producer who will in general have more information available for the correction than the data user. Second, the data user often lacks the capacity to deal with the problem appropriately. Third, the analyst has his own problems to worry about and thus the burden of correctly handling deficiencies in the data should be kept as small as possible. And finally, correcting the data at the processing stage will guarantee consistent results between different researchers that might otherwise include different correction methods in their analysis.

The multiple imputation strategy that we suggest is related to the approach suggested by Heitjan and Rubin (1990) who proposed to use multiple imputation to correct for heaped ages for young children in Tanzania. The basic idea is to estimate the rounding probabilities given the observed data and to impute the "missing" true income based on the observed data and the estimated rounding probabilities.

If the interval in which the true income must fall would be known given the reported income y_i for each record, imputing the missing income would be straightforward. Assuming a log-normal distribution for income, the parameters of the distribution could be estimated by maximizing the likelihood:

$$L(\mu,\sigma^2;y)=\prod_{r_i=0} f(y_i;\mu,\sigma^2)\prod_{r_i\neq 0}[F(u_i;\mu,\sigma^2)-F(l_i;\mu,\sigma^2)],$$

where r_i is a rounding indicator, which equals zero if no rounding occurred and l_i and u_i are the lower and upper bound of the rounding interval for record i. Given the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$ and assuming flat priors for all parameters, we could multiply impute the missing income as follows:

1. Approximate a draw from the posterior distribution of μ and σ^2 by drawing from a bivariate normal distribution with mean $\mu=(\hat{\mu},\hat{\sigma}^2)$ and variance Σ, where Σ is the inverse of the hessian matrix obtained from the ML estimation.
2. Impute an income value for all those records with $r_i\neq 0$ by drawing from a truncated log-normal distribution given the known truncation points l_i and u_i and the parameters drawn in step 1.

Repeating steps 1 and 2 m times would yield m imputed datasets that properly reflect the uncertainty from imputation. Again, a sophisticated analyst might prefer to use $\hat{\mu}$ and $\hat{\sigma}^2$ directly to obtain his or her estimate of interest, e.g. the poverty rate. However, providing a dataset with income multiply imputed would allow the user to obtain valid inferences without the need of accounting for the deficiencies due to rounding by simply applying the standard multiple imputation combining rules (Rubin, 1987).

Unfortunately, the rounding interval is unknown in our setting since for example a reported income of $2,500 might be a result of no rounding, or rounding to the nearest $5, $10, $50, $100, or $500. Thus, the probability for rounding also needs to be estimated.

Let $d_i=j$, with $j=0,\ldots,J$ be a divisibility indicator. For example, if we assume the reported income in a dataset might be a result of rounding to the nearest integer, rounding to the nearest $10, $100, or $1,000, all records with a reported amount divisible by 1,000 would have $d_i=3$ whereas all records with a reported income not divisible by 10 would have $d_i=0$. Using this definition, the joint likelihood of μ and σ^2 and the rounding probabilities p_j, $j=0,1,\ldots,J$ is given by:

$$L(\mu,\sigma^2,p_j;y)=\prod_{d_i=0} p_0 f(y_i,\mu,\sigma^2)\times\prod_{d_i\neq 0}\{\sum_{s=1}^{d_i} p_s[F(u_{s,i},\mu,\sigma^2)-F(l_{s,i},\mu,\sigma^2)]\}, \qquad (1)$$

where $l_{s,i}$ and $u_{s,i}$ are the lower and upper bound of the rounding interval for record i if the true "degree of rounding" would be s.

Note that p_j is the unconditional probability that the true degree of rounding equals j. However, for the imputation step we need the probability for rounding given the reported income y_i, more precisely, we need the probability for rounding to a specific rounding base given the observed divisibility d_i, i.e. we need $P(DR_i=j|d_i)$, where DR_i is the degree of rounding for record i and $j=0,\ldots,J$. Thus in our example above, $P(DR_i=3|d_i=3)$ would be the probability that the reported value was rounded to the next \$1,000 given that it is divisible by \$1,000. Using Bayes' Theorem we have

$$P(DR_i = j \mid d_i) = \frac{P(d_i \mid DR_i = j)P(DR_i = j)}{P(d_i)} \propto P(d_i \mid DR_i = j)P(DR_i = j)$$

Continuing the example above, the probability for rounding given that we observe a zero as the last digit of y (but not for the second last digit) is given by

$$P\begin{pmatrix} DR_i = 0 \mid d_i = 1 \\ DR_i = 1 \mid d_i = 1 \\ DR_i = 2 \mid d_i = 1 \\ DR_i = 3 \mid d_i = 1 \end{pmatrix} \propto P\begin{pmatrix} d_i = 1 \mid DR_i = 0 \\ d_i = 1 \mid DR_i = 1 \\ d_i = 1 \mid DR_i = 2 \\ d_i = 1 \mid DR_i = 3 \end{pmatrix} P\begin{pmatrix} DR_i = 0 \\ DR_i = 1 \\ DR_i = 2 \\ DR_i = 3 \end{pmatrix} = \begin{pmatrix} 0.09\,p_1 \\ 0.9\,p_2 \\ 0 \\ 0 \end{pmatrix}, \qquad (2)$$

if we approximate $P(d_i=1|DR_i)$ by assuming that the last digits of the rounded y-values are uniformly distributed.

Given the maximum likelihood estimates $\hat{\mu}, \hat{\sigma}^2, \hat{p}_0, \hat{p}_1, \ldots, \hat{p}_J$ and assuming flat priors for all parameters, we can impute the true income in four steps:

1. Approximate a draw from the posterior distribution of $\mu, \sigma^2, p_0, p_1, \ldots, p_J$ by drawing from a $(J+3)$- variate normal distribution with $\mu = (\hat{\mu}, \hat{\sigma}^2, \hat{p}_0, \hat{p}_1, \ldots, \hat{p}_J)$ and Σ, where Σ is the inverse of the hessian matrix obtained from the ML estimation.
2. Compute the posterior probabilities for rounding according to (2) given the drawn probabilities from step 1.
3. Draw a new rounding interval given the probabilities from step 2 and compute the according interval bounds l_i and u_i.
4. Impute an income value for all those records with $d_i \neq 0$ by drawing from a truncated log-normal distribution given the known truncation points l_i and u_i and the parameters drawn in step 1.

4. Illustrative simulations

To illustrate that valid inferences can be obtained using the approach described above, we use a repeated simulation design. We generate a population of $N=1,000,000$ income records by randomly drawing from $Y \sim logN(8,0.47)$. From this population we repeatedly draw simple random samples of size $n=5,000$. Again, we assume that the true probabilities for rounding are given by $p=\{0.1,0.4,0.4,0.1\}$, where $p_0,\ldots,p_3$ stand for no rounding, rounding to the nearest \$10, \$100, and \$1000 respectively. Each sample is rounded according to these probabilities and this would be the sample that would be available for the user to analyze.

To impute the true income we use two strategies. With the first strategy that we call the naïve strategy, we assume that the reported income is always rounded to the maximum possible interval given the observed income, e.g. if the last digit is zero, we assume that the income was rounded to nearest \$10, if the last two digits are zero we always assume that the income was rounded to the nearest \$100 etc. This strategy serves to illustrate that such a simplified imputation technique can lead to biased inferences based on the imputed data. With the second strategy that we call the improved strategy, the rounding probabilities are estimated and the true income is imputed according to the steps described above. We assume that the population quantity of interest is the poverty rate (*pr*) defined as the percentage of units with an income less than 60 percent of the median income in the population. To estimate the variance of the poverty measure computed from the sample, we use the variance estimator suggested by Preston (1995). We repeat the whole process of sampling, imputing and analyzing the data 5,000 times. The results are summarized in Table 2.

Table 2: Simulation results

	$mean(\widehat{pr})$	$var(\widehat{pr})$	$mean(\widehat{var(\widehat{pr})})$	*Variance ratio*	*95% coverage rate*
Org. sample	13.85	$2.268*10^{-5}$	$2.376*10^{-5}$	1.048	95.50
Rounded sample	12.89	$2.205*10^{-5}$	$2.342*10^{-5}$	1.062	48.06
Naïve imputation	13.99	$3.001*10^{-5}$	$2.782*10^{-5}$	0.927	93.20
Improved imp.	13.88	$2.204*10^{-5}$	$2.664*10^{-5}$	1.209	96.98

The second column contains the average point estimates across the 5,000 simulation runs. Given that the poverty rate in the population is 13.92 percent, we find that only the estimate of the poverty rate in the original sample and the estimate using the improved imputation method are unbiased. For the rounded sample the poverty rate is underestimated by more than 1 percent on the absolute scale, the naïve imputation slightly overestimates the true poverty rate. The third column contains the true variance of the estimated poverty rates across the 5,000 simulation runs, whereas the fourth column contains the average of the estimated variances. If the variance estimate is unbiased the ratio of the average estimated variance and the true variance reported in column 5 should be close to one. Only the variance estimate for the original sample and the variance estimate for the rounded data are unbiased. The naïve imputation method slightly underestimates and the improved imputation method slightly overestimates the true variance.

The last column reports the percentage of times the 95% confidence interval around the estimated poverty rate contains the true poverty rate. The confidence interval of the rounded sample clearly has less than nominal coverage rate, the other coverage rates are close to the nominal coverage rate with a small undercoverage for the naïve imputation method. As a result of the overestimation of the variance for the improved imputation method, the coverage rate is slightly above the nominal coverage level indicating a conservative variance estimate. This might be a result of using an approximation of the posterior distribution when drawing new rounding probabilities during the imputation step. Nevertheless, from the results it is obvious that only the improved imputation method provides unbiased point estimates and a confidence interval with at least nominal coverage rate. Still, the bias in the naïve imputation is rather small and would probably be considered acceptable at least for this simulation.

5. Application to the panel study "Labor Market and Social Security"

In this section we apply our imputation approach to the German panel survey "Labor Market and Social Security (PASS)". This study was launched in 2006 and is conducted yearly ever since. It aims at measuring the social effects of labor market reforms and at getting insights into the living conditions of low income households. 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. One of the major benefits of this combination of two different samples is the easiness of generating control groups for the benefit recipients. PASS contains a large number of socio-demographic characteristics (e.g. age, gender, marital status, religion, migration background) , employment-related characteristics (e.g., status of employment, working hours, income from employment, employment history), benefit-related characteristics (e.g. benefit history, amount of benefits, participation in training measures), and subjective indicators (e.g. fears and problems, employment orientation, subjective social position). A detailed description of the survey can be found in Trappmann et al. (2010).

One of the income related questions of the survey asks the head of household to provide an estimate of the total household income per month. Since income is a sensitive question, nonresponse rates on this question range between 6.6 percent and 11.5 percent for the waves 2006/2007 to 2008/2009. To obtain at least partial information for those respondents that are not willing or are unable to provide an exact estimate for their income, those respondents are asked whether their income is less or at least 1,000 EUR. If the respondent is willing to answer this question, further questions follow that try to obtain narrower income brackets based on the answers from previous questions. With this approach more than 98 percent of the participants are willing to provide some income information. However, as discussed in the introduction (see Table 1) the exact reported income seems to be subject to rounding. More than 80 percent of the reported income values are divisible by 10 and more than 60 percent are divisible by 100.

Thus, the income variable shows three forms of deficiencies. For some respondents, the income is missing completely, for some respondents who answered only the bracketed income questions, only an interval is available in which the true income must fall. Depending on the level of detail the respondent was willing to give, the interval might be smaller or wider, but the exact upper and lower bound of the interval are always known. Finally, part of the reported income is rounded with unknown rounding intervals that need to be estimated from the data. To obtain estimates for the true income for this dataset, we proceed in two steps: In the estimation step we use the reported income for those units that provided the exact income information to estimate the parameters of the income distribution and the rounding probabilities as described in Section 3. With this approach we assume that the reported income of the exact reporters does not differ systematically from the income of those respondents that answered only the bracketed income questions or did not provide any income information at all. This is a strong assumption since the missingness mechanism for the income is generally assumed not to be missing completely at random (MCAR). A simple extension of our application that would weaken this assumption to some extent would be to include the bracketed information when maximizing the likelihood function. By including this information we would assume that the income for those units that provided at least partial information regarding their income is not systematically different from those less than 2 percent that did not provide any income information at all. To further reduce the MCAR assumption we could also include the rich set of covariates to estimate the conditional income distribution given the set of covariates. We leave this for future research.

Inspecting the spikes in the reported income distribution reported in Table 1, we assume that respondents round to the nearest 5, 10, 50, 100, 500, or 1,000 Euros. The results from the estimation step are summarized in Table 3.

Table 3: Estimated parameters for the household income. μ and σ^2 are the parameters of the log-normal distribution. $p_{0,\ldots,}\ p_6$ are the estimated probabilities for not rounding and rounding to the nearest 5, 10, 50, 100, 500, or 1000 Euros respectively.

μ	σ^2	p_0	p_1	p_2	p_3	p_4	p_5	p_6
7.218	0.669	0.199	0.029	0.105	0.133	0.392	0.106	0.036

We find that the probability to round to the nearest 100 Euros is almost 40 percent and the probability for rounding to the nearest 500 Euros is still above 10 percent. Once the income distribution parameters and the rounding probabilities are estimated, we calculate the conditional rounding probabilities and draw rounding intervals for the reported exact income.

In the imputation step we repeatedly draw ($m=5$) from a truncated log-normal distribution with the drawn truncation points for the exact reporters and given truncation points for the interval reporters. For those units for which no income information is available, the truncation points are set to $[0,\infty[$.

The analysis of interest again is the estimated poverty rate. To estimate the poverty rate we compute the disposable income that corrects for the number of household members and the age of the household members as suggested by the OECD. We compare the poverty rate based on the reported income for the exact income reporters with the poverty rate obtained from the imputed dataset. The poverty rate changes from 14.7 percent using only the reported income to 16.67 using the imputed income. Of course it should be noted that this application only serves illustrative purposes and that the imputation model is overly simplified so that the dramatic change for the poverty measure should be interpreted with caution.

6. Discussion

The application described above serves to illustrate how to implement the suggested imputation approach in practice in a simplified setting. We could improve the imputation model in several ways. First, we could include covariates X in the model, i.e. we could set $\mu=X\beta$ in (1) and estimate the parameters β jointly with the rounding probabilities. This is an area of current research. So far, we experienced numerical instabilities with our optimization algorithm once we try to jointly estimate all the unknown parameters.

Another even more restrictive assumption in our application is that the rounding probabilities are constant for all units. Intuitively it seems reasonable that the probability to round depends on the income. However, this would imply that the rounding mechanism is no longer ignorable since it depends on the unobserved true income. This might require treating the rounding indicator as missing data and estimate its distribution using EM type algorithms

as described in Little and Rubin (2002). A useful alternative might be to use the reported income as a proxy for the true income along with other potentially explanatory variables when modeling the rounding probabilities for example by using a multinomial logit model.

It should also be noted that our application is based on a screener variable for the total household income, i.e. the head of household is asked to estimate the total household income. However, there is a common agreement that the screener variable approach leads to a high measurement error since it will be difficult for the survey respondent to know the exact income amounts or even only to remember all income sources for all members of the household. For this reason researchers tend to prefer the individual income component approach for which each individual in the household is interviewed and is asked to report all his income sources. The final household income is then derived by aggregating the different income sources of all household members. Official poverty rates are usually based on this approach, too.

The amount of rounding in the household income variable should generally be higher for the screener variable approach for two reasons. First, it is reasonable to assume that the tendency to round is positively related to the amount of uncertainty the respondent feels regarding the estimate he or she is asked for and this uncertainty should be higher for the total household income compared to the respondent's own income components. Second the tendency to round will likely increase with the requested amount. Thus, the individual income components might show less rounding compared to the total family income. The findings by Czajka and Denmead (2008) seem to support this hypothesis. Looking only at individuals with a total family income below $52,500, the authors find that in the National Health Interview Survey (NHIS) which uses the screener variable approach, 35.6 percent of the individuals reported an income divisible by $5,000 and 20.9 percent reported an income divisible by $10,000. In the CPS (ACS) those numbers reduced to 11.0 (16.2) percent and 6.2 (9.5) percent respectively. Thus, the income based on the screener variable approach seems to be more affected by rounding.

However, despite the fact that our findings in this paper are based on a screener income variable, we strongly believe that the effect of rounding should not be neglected even if the family income variable is based on a large number of individual components for two reasons. First, we believe that most of the survey respondents will not have an income in all those categories. To the contrary we believe that for a substantial number of respondents the income from earnings will be the only relevant source of income. And even if the respondent has more than one source of income, the income from earnings will be the dominant one and the reported earnings might still be rounded. Thus, we might not see large spikes in the derived total household income because small amounts of income from other sources mask the rounding of the income from earnings. Nevertheless, the derived income distribution will be biased unless the rounding in the earnings variable is corrected.

It is also important to note that the 11.0 percent of values divisible by $5,000 for the CPS which in itself indicates a non negligible amount of rounding uses a large rounding base. We expect that the percentage of values divisible by $1,000 is substantially larger and rounding to the next $1,000 will already negatively affect the inferences obtained from the rounded data. Thus, while we admit that the effect of rounding on poverty measures might be lower than the effect we find in our evaluations, we still strongly believe that the effect will be substantial enough to have an important impact on the results and our paper serves to illustrate the potential gains from the suggested approach. If a reasonable model can be specified to estimate the probabilities for rounding and the parameters of the distribution, imputing the true income during the data processing stage will considerably reduce the efforts necessary for the data analyst to obtain valid inferences from the provided data. The analyst can simply run his or her standard models without needing to worry how to best address the potential biases from nonresponse and rounding. All the analyst needs to do is to repeat his or her analysis on the multiple copies of the dataset and then to apply the simple multiple imputation combining rules to obtain a final point and variance estimate for his or her quantity of interest. We believe that this approach can be highly relevant in practice since many applied researchers currently either shy away from using data with high nonresponse and rounding rates since they don't know how to deal with these deficiencies accordingly, or they simply analyze the data using their standard set of tools risking considerable bias in their results.

References

Czajka, J.L., Denmead, G. (2008): Income Data for Policy Analysis: A Comparative Assessment of Eight Surveys, *Mathematica Policy Research Working Paper* No. PR08-62.

Heitjan D.F., Rubin D.B. (1990): Inference from coarse data via multiple imputation with application to age heaping, *Journal of the American Statistical Association* **85**, 304–314.

Little, R.J.A., Rubin, D.B. (2002): *Statistical Analysis with Missing Data, 2nd edition.* New York: Wiley.

Preston, I. (1995): Sampling distributions of relative poverty statistics, *Applied Statistics (Journal of the Royal Statistical Society: Series C)* **44**, 91-99.

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

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.