Drechsler, Jörg; Kiesl Hans

Working Paper

Beat the heap: An imputation strategy for valid inferences from rounded income data

IAB-Discussion Paper, No. 2/2014

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Drechsler, Jörg; Kiesl Hans (2014) : Beat the heap: An imputation strategy for valid inferences from rounded income data, IAB-Discussion Paper, No. 2/2014, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg

This Version is available at:
https://hdl.handle.net/10419/103068

Institute for Employment Research
The Research Institute of the Federal Employment Agency

IAB-Discussion Paper 2/2014

Articles on labour market issues

Beat the Heap – An Imputation Strategy for Valid Inferences from Rounded Income Data

Jörg Drechsler
Hans Kiesl

ISSN 2195-2663

Beat the Heap – An Imputation Strategy for Valid Inferences from Rounded Income Data

Jörg Drechsler (IAB)
Hans Kiesl (OTH Regensburg, Department of Computer Science and Mathematics, Postfach 12 03 27,93025 Regensburg, Germany)

Contents

Abstract 4

Zusammenfassung 4

1 Introduction 6

2 Potential Bias from Rounding 7

3 Correcting the Rounding Error Through Imputation 9
3.1 Correction methods if income is reported in intervals 10
3.1.1 The model 10
3.1.2 Estimation of the parameters 10
3.1.3 Imputation of exact income values 11
3.2 Correction methods if the true rounding interval is unknown 12
3.2.1 The model 12
3.2.2 Estimation of the parameters 12
3.2.3 Imputation of exact income values 14

4 Simulation Study 15
4.1 Generating the Population for the Simulation 15
4.2 Constructing the Rounded Income 15
4.3 The Repeated Sampling Design 16
4.4 Results 17

5 Application to the Panel Study “Labour Market and Social Security” 18
5.1 Evaluation of the Model Assumptions 19
5.1.1 The Income Model 20
5.1.2 The Rounding Behaviour Model 20
5.2 Results 21

6 Conclusion 22

References 25

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 generally tend to be lower than response rates for other non-sensitive questions. On the other hand respondents usually don't remember their exact income and thus tend to provide a rounded estimate. The negative effects of item nonresponse are well studied and most statistical agencies have developed sophisticated imputation methods to correct for this potential source of bias. However, to our knowledge the effects of rounding are hardly ever considered in practice, despite the fact that several studies have found strong evidence that most of the respondents round their reported income values.

In this paper we illustrate the substantial impact that rounding can have on important measures derived from the income variable such as the poverty rate. To obtain unbiased estimates, we propose a two stage imputation strategy that estimates the posterior probability for rounding given the observed income values at the first stage and re-imputes the observed income values given the rounding probabilities at the second stage. A simulation study shows 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 "Labour Market and Social Security."

Zusammenfassung

Befragungen zu Einkommensverhältnissen sind typischerweise von zwei Fehlerquellen betroffen, die zu Verzerrungen führen können, wenn sie bei der Analyse nicht berücksichtigt werden: Auf der einen Seite gilt das Einkommen als sensible Information und die Antwortraten zum Einkommen liegen in der Regel niedriger als Antwortraten bei anderen nicht sensiblen Fragen. Auf der anderen Seiten können sich die Befragten in aller Regel nicht genau an ihr exaktes Einkommen erinnern und geben daher einen gerundeten Wert an. Die negativen Auswirkungen des Antwortausfalls sind bereits gründlich untersucht worden und die meisten datenbereitstellenden Institutionen haben bereits Imputationsmethoden implementiert um möglichen Verzerrungen durch den Ausfall entegegenzuwirken. Im Gegensatz dazu werden die Auswirkungen des Rundens nach unserer Kenntnis bisher in der Praxis weitestgehend vernachlässigt, obwohl etliche Studien deutlich gezeigt haben, dass die meisten Befragten Ihrer Einkommensangaben runden.

In diesem Papier veranschaulichen wir den starken Einfluss, den dieses Runden auf wichtige Kennziffern wie die Armutsquote haben kann. Um unverzerrte Schätzergebnisse zu erhalten, stellen wir ein zweistufiges Imputationsverfahren vor, bei dem in einem ersten Schritt gegeben das beobachtete Einkommen die a posteriori Wahrscheinlichkeit zu Runden geschätzt wird. In einem zweiten Schritt wird dann das tatsächliche Einkommen unter den bestimmten Rundungswahrscheinlichkeiten imputiert. Anhand einer Simulationsstudie illustrieren wir, dass es mit diesem Verfahren möglich ist, unverzerrte Schätzergebnisse zu gewinnen. Darüberhinaus präsentieren wir Ergebnisse auf Basis der IAB Längsschnittstudie "Panel Arbeitsmarkt und Soziale Sicherung (PASS)".

JEL classification:C42, D31

Keywords: Heaping, Measurement error, Multiple imputation, Poverty rate

Acknowledgements: We are thankful to Stephanie Eckman, Frauke Kreuter, Jerry Reiter, and Hans Schneeweiß for very useful comments on an earlier version of the manuscript.

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

1 Introduction

Obtaining reliable income information in surveys is important for numerous reasons. The collected data regularly form the basis for important indicators of inequality, such as the proportion of people at risk of poverty, and many political decisions such as the establishment or elimination of social security programmes rely heavily on estimates of the income distribution. For these reasons, many household surveys collect income data, but measuring income in surveys is a difficult task. Firstly, 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 research indicating that the survey participants who are willing to answer income related questions 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 (see for example Bollinger/Hirsch (2013) for a recent discussion of the topic).

Secondly, even if the respondent is willing to provide his or her income, he or she will often find it difficult to report the exact income amount. 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 amount of income, the magnitude of rounding can range from rounding to the closest 5 euro value to rounding to the closest 10,000 euro value. As a result the reported income data have several spikes at certain income values. For example, Czajka/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 when monthly income is requested, respondents tend to round although obviously the typical rounding base will only vary between 5 and 1,000 in this case.

As an illustration Table 1 provides the percentage of the reported monthly income values that are divisible by a given round number obtained from the German panel study "Labour Market and Social Security (PASS)" (Trappmann et al., 2010) for the year 2008/2009 (see Section 5 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. Based on these results it is evident that treating the reported income as a direct measure for the true income is inappropriate. As we will illustrate below, the rounded income values can lead to biased inferences if the

analyst doesn't account for the rounding.

Dealing with rounded income values would be easy, if the rounding mechanism were known completely. For example, if respondents are asked to report their yearly income rounded to the nearest multiple of 1,000 and the reported income is 30,000, we could infer that the true income must fall in the interval [29,500;30,500]. In this case standard techniques for interval data could be applied (see for example, Schenker et al. (2006)). However, in most surveys the underlying rounding mechanism is unknown. To continue the example, a reported income of 30,000 could be the result of no rounding, rounding to the closest 10, 100, 1,000, or even 10,000. Therefore, rounded data cannot simply be seen as a special form of interval data and alternative techniques are required.

To address the nonresponse problem discussed above, many agencies already provide public use files in which the missing values in the income variable are imputed. For example the income in the ACS and in the CPS is imputed using sophisticated hot deck imputation methods (U.S. Census Bureau, 2009, 2013). Thus for many surveys, the burden of dealing with item nonresponse in the income variable is already bourn by the data providers. However, the problems stemming from rounding the income information are still widely ignored.

In this paper we propose to use multiple imputation – widely accepted nowadays as a straightforward tool to obtain valid inferences from data subject to nonresponse – to reduce biases introduced by respondent rounding. 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/Rubin (1990) for heaped age data.

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 regularly computed from income data: the poverty rate. In Section 3 we discuss our imputation approach for dealing with rounding errors. Section 4 contains a simulation study that demonstrates that the proposed imputation approach can correct the rounding bias. In Section 5 we apply the approach to the German panel study "Labour Market and Social Security." The paper concludes with a discussion of future research topics.

2 Potential Bias from Rounding

The impact of rounding has been studied for many years. Sheppard (1898) was the first to investigate to what extent rounding affects the estimation of different moments of a continuous variable. He showed that under his assumptions regarding the underlying rounding mechanism the first moment of the distribution was only slightly biased whereas higher moments were severely affected. These results led him to suggest the well known Sheppard's correction to get an approximate estimate for the variance of the underlying continuous

variable in this case. Since then a number of papers on statistical inference from rounded data have been published (see for example Dempster/Rubin (1983) or Liu et al. (2010) for a discussion of the effects of rounding in linear regression). Schneeweiss/Komlos/Ahmad (2010) recently provided a concise review of the body of work in the area.
Nevertheless, most of the rounding literature assumes symmetric rounding intervals that can be derived directly from the reported value. For example, if weight is reported in kilograms, it is assumed that the true weight must be in the interval *reported weight* $\pm$ *500 grams*. However, this does not generally hold for heaping. With heaping certain values, say multiples of fives, are preferred over other values and respondents tend to round to these values. As discussed in the introduction, the interval for the true value can no longer be directly derived from the reported value if different potential rounding bases (rounding to the closest 5, 10, 50, etc.) are possible. Dan Heitjan addressed this special form of rounding in several papers (Heitjan (1989, 1994); Heitjan/Rubin (1990, 1991); see also Manski/Molinari (2010) for a different perspective on the same topic). For this type of rounding behaviour even the first moment of the heaped data may differ substantially from the first moment of the underlying continuous variable. Furthermore, as with any type of rounding, the marginal distribution will change and all measures that are based on the percentiles of the distribution will be biased.
This effect can also be observed for one of the most prominent estimates that is routinely calculated from income data: the proportion 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 countries the poverty rate is defined as the proportion of persons with an income less than 60% of the median income. This statistic is of great political importance because it allows a direct comparison between regions and countries but also because many political decisions such as establishing new labour market programmes 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. However, to our knowledge no measures are taken to adjust for the fact that the reported income might be heaped.

To illustrate the potential for bias in the estimated poverty rate obtained from rounded income data, we generate data in a simplified setting. For our small simulation, we assume that the true income follows a log-normal distribution with $\mu = 8$ and $\sigma = 0.47$. Approximating the income distribution 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 1,000 euros 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 (defined as 60% of the median income) from the sample before and after rounding. The results are displayed in Figure 1. There are obvious spikes in the rounded income data at the round numbers. The poverty

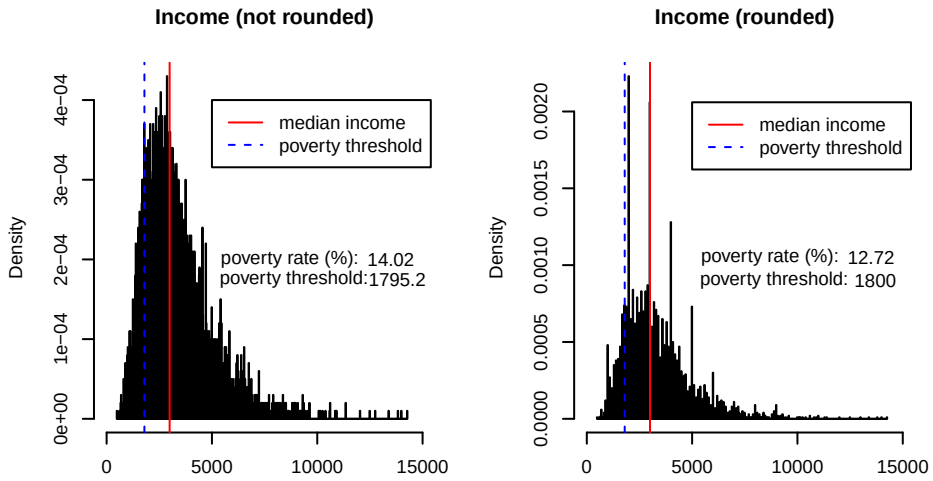

Figure 1: Simulated income distribution, poverty rate, and poverty threshold before and after rounding.

rate drops from 14.02% before rounding to 12.72% after rounding while at the same time the poverty threshold slightly increases from 1795 EUR to 1800 EUR. 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 substantial and analyzing the reported data without any adjustments to account for 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 might lack the capacity to deal with the problem adequately. Third, the analyst has his own science to worry about and thus the burden of correctly handling data deficiencies 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 by Heitjan/Rubin (1990) who proposed to use multiple imputation to correct for heaped reported age values of young children in Tanzania. The basic idea is to estimate the rounding probabilities given the observed data and to impute the missing exact income based on the observed data and the estimated rounding probabilities. Van der Laan/Kuijvenhoven (2011) have recently followed a related multiple imputation approach in the context of rounded unemployment durations; however, they assume constant rounding probabilities within certain intervals,

while our approach is more general (in the context of income values, the assumption of uniform rounding is too restrictive).

3.1 Correction methods if income is reported in intervals

If the interval in which the true income must fall is known given the reported income, imputing the unknown true income is straightforward. Such a situation would arise, for example, if income is collected as interval data only (as in the German Microcensus). In the following, we describe how maximum likelihood estimates for the parameters of the income distribution could be obtained in this case and how these estimates could be used to impute the true income given the observed income intervals (a similar approach has been applied by Schenker et al. (2006) for the National Health Interview Survey in the United States).

3.1.1 The model

We model the distribution of the household income Y by a log-normal distribution. (There is huge evidence that the log-normal distribution is an adequate model for the largest part of the income distribution; see, for example, Clementi/Gallegati (2005).) Consequently, $\log(Y)$ is normally distributed. We allow the mean of $\log(Y)$ to depend on some covariates X and assume a constant variance. (Allowing for heterogeneity would be a straightforward extension, which we don't regard as necessary for our data.) Thus, our data model is

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

Next, we define R as an indicator variable with $R = 0$ if the household income is known exactly, and $R = 1$ if it is only known to be in some interval $[L, U]$ (with U possibly being ∞). We assume that the conditional distribution of Y given $Y \in [L, U]$ is independent of R.

3.1.2 Estimation of the parameters

Let y_i be the true income for sample household i, $i = 1, ..., n$ (so that $\log(y_i)$ is the respective logged income value), and let l_i and u_i be the lower and upper bound of the known interval for y_i (with $l_i = u_i$ if y_i is known exactly). Under the model described above, the likelihood function (assuming independent observations) is given by

$$\begin{aligned} L(\beta, \sigma^2|y, l, u, r, x) &= \prod_i f(\log(y_i), \mu_i = x_i'\beta, \sigma^2)^{1-r_i} \times \\ &\quad \prod_i [F(\log(u_i), \mu_i = x_i'\beta, \sigma^2) - F(\log(l_i), \mu_i = x_i'\beta, \sigma^2)]^{r_i}, \end{aligned}$$

where f and F are the density and the cdf of a normal distribution. If the data stem from a complex survey, the assumption of independent obervations might be questionable. The usual alternatives in this situation may then be applied (for example, calculating a weighted likelihood function or adding survey design variables to the vector of covariates).

3.1.3 Imputation of exact income values

Given the maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}^2$ obtained by maximizing the likelihood function, and assuming flat priors for all parameters, multiple imputations of the missing exact income values could be obtained as follows:

1. Approximate a draw from the posterior distribution of β and σ^2 by drawing from a multivariate normal distribution with mean vector $\hat{\mu} = (\hat{\beta}, \hat{\sigma}^2)$ and covariance matrix $\hat{\Sigma}$, where $\hat{\Sigma}$ is the negative inverse of the Hessian of the log-likelihood function (with the maximum likelihood estimates plugged in):

$$(\beta^*, \sigma^{2*}) \sim MVN(\hat{\mu}; \hat{\Sigma}).$$

2. Estimate a logged income value z_i^{imp} for all records with $r_i = 1$ by drawing from a truncated log-normal distribution given the known truncation points $\log(l_i)$ and $\log(u_i)$ and parameters $\mu_i = x_i'\beta^*$ and $\sigma^2 = \sigma^{2*}$.

3. Impute the exact income value $y_i^{imp} = \exp(z_i^{imp})$.

Repeating this procedure m times would yield m imputed datasets that properly reflect the uncertainty from imputation. Valid inferences from these data could be obtained using the standard multiple imputation combining rules (Rubin, 1978).

We stress that this approach assumes that households pretending to report their income values exactly do this without any degree of rounding. Rounded income values could be dealt with, as far as the rounding mechanism is completely known. If it were known, for example, that all reported income values s_i are rounded to the nearest number divisible by 100, we could conclude that $[l_i, u_i] = [s_i - 50, s_i + 50]$ for every i. However, in most income surveys, we neither know which values are rounded nor the individual rounding degree. We therefore extend our model in order to additionally estimate the rounding degree given observed income values.

3.2 Correction methods if the true rounding interval is unknown

3.2.1 The model

As above, we model the distribution of the household income Y by a log-normal distribution (given some covariates X):

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

Now we also need a model for the rounding behaviour. A rounding function in general is a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ which is piecewise constant within given intervals. 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. No rounding at all will be called rounding of degree 0. If $c > t$ we say that f_c has a higher degree of rounding than f_t. 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. Thus, $\log(Y)$ and G have a bivariate normal distribution given X and Z:

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

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

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

3.2.2 Estimation of the parameters

We fix γ_0 at 0 and τ^2 at 1 to make the ordered probit model identifiable. The remaining set of parameters to be estimated is therefore given by $\Psi = (\beta, \sigma^2, \gamma_1, \gamma_2, k_1, ..., k_{p-1})$. Let

s_i be the observed income of household i. The density of the observed income (given covariates x_i and z_i) may be written as follows:

$$f(s_i, x_i, z_i|\Psi) = f(s_i|x_i, z_i, \Psi) \cdot f(x_i, z_i)$$

with

$$\begin{aligned} f(s_i|x_i, z_i, \Psi) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(g, \log(y), s_i|x_i, z_i, \Psi) d\log(y) dg \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(s_i|g, \log(y), x_i, z_i, \Psi) \cdot f(g, \log(y)|x_i, z_i, \Psi) d\log(y) dg \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(s_i|g, \log(y)) \cdot f(g, \log(y)|x_i, z_i, \Psi) d\log(y) dg. \end{aligned}$$

Now observe that s_i is uniquely determined given g and $\log(y)$. Thus, $f(s_i|g, \log(y))$ is simply an indicator function with $f(s_i|g, \log(y)) = 1$ if g and $\log(y)$ are consistent with s_i, and 0 otherwise.

If we write $A(s_i)$ for the set of $(g, \log(y))$ that are consistent with an observed s_i, the conditional density becomes

$$f(s_i|x_i, z_i, \Psi) = \iint_{A(s_i)} f(g, \log(y)|x_i, z_i, \Psi) d\log(y) dg.$$

To give an example, let us assume the observed income is $s_i = 850$. If the possible degrees of rounding are 1, 5, 10, 50, 100, 500, and 1000 (which seems reasonable given Table 1), we can conclude that (g, y) lies in $]-\infty, k_1[\times[849.5, 850.5[$ (rounding to the nearest integer), in $[k_1, k_2[\times[847.5, 852.5[$ (rounding to the nearest multiple of 5), in $[k_2, k_3[\times[845, 855[$ (rounding to the nearest multiple of 10), or in $[k_3, k_4[\times[825, 875[$ (rounding to the nearest multiple of 50). The conditional density of s_i given x_i, z_i, Ψ is then

$$\begin{aligned} f(s_i|x_i, z_i, \Psi) &= \int_{-\infty}^{k_1} \int_{\log(849.5)}^{\log(850.5)} f(g_i, \log(y)|x_i, z_i, \Psi) d\log(y) dg + \\ &+ \int_{k_1}^{k_2} \int_{\log(847.5)}^{\log(852.5)} f(g, \log(y)|x_i, z_i, \Psi) d\log(y) dg + \\ &+ \int_{k_2}^{k_3} \int_{\log(845)}^{\log(855)} f(g, \log(y)|x_i, z_i, \Psi) d\log(y) dg + \\ &+ \int_{k_3}^{k_4} \int_{\log(825)}^{\log(875)} f(g, \log(y)|x_i, z_i, \Psi) d\log(y) dg. \end{aligned}$$

The likelihood function for Ψ given the observed income values s_i and covariates x_i, z_i

(assuming independent observations) may then be written as

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

The parameter vector Ψ may now be estimated by maximizing $L(\Psi|s,x,z)$ using numerical methods.

3.2.3 Imputation of exact income values

Assuming flat priors for all parameters and independence between the prior distributions of (β,σ^2) and $(\gamma_1,\gamma_2,k_1,...,k_{p-1})$ we can approximate a draw from the posterior distribution of $f(\Psi|s,x,z)$ by drawing 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 we 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 1000, $\log(y_i)$ 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 gives m imputed datasets that properly reflect the uncertainty from imputation. Again valid inferences from these data could be obtained using the standard multiple imputation combining rules (Rubin, 1978).

4 Simulation Study

To illustrate that valid inferences can be obtained using the approach described above we implement a repeated simulation design, i.e., we repeatedly sample from a population with known characteristics and evaluate whether valid inferences can be obtained based on the rounded income in the sample. Before the actual simulation study can be conducted, a population with plausible distributional characteristics and realistic rounding behaviour needs to be constructed. In the following we first describe how we obtained this population based on information from the wave 2006/2007 of the panel study "Labour Market and Social Security (PASS)" (described in detail in Section 5). We then present the results of the simulation study.

4.1 Generating the Population for the Simulation

The population is generated in three steps: Since the observed income in the PASS survey is subject to rounding, we first impute one replicate of the unobserved true income based on the approach described above (see Section 5 for details regarding the models used for imputation) and treat this replicate as the true income in the survey. We next generate a population of $N = 1,000,000$ records by sampling with replacement from the survey using the sampling weights to define the probability of selection for each record. In a final step we fit the following linear regression model to the imputed income in the population:

$$log(income) = \alpha + \beta_1 \cdot hhsize + \beta_2 \cdot age + \beta_3 \cdot unemp\_ben + \beta_4 \cdot livspace + \varepsilon,$$

where $hhsize$ is the household size, age is the age of the respondent, $unemp\_ben$ is an indicator, whether the household receives unemployment benefits, and $livspace$ is the living space of the household. This model is used to generate a new set of imputed income values by randomly drawing from

$$inc_{final} \sim \log N(X'\beta, \sigma^2),$$

where inc_{final} is the final income that will be used for the simulation, $X = \{1, hhsize, age, unemp\_ben, livspace\}$, $\beta = \{\alpha, \beta_1, \beta_2, \beta_3, \beta_4\}$, and σ^2 is the residual variance from the linear regression. The final step is necessary to ensure that all parameters that govern the income distribution are known in the population.

4.2 Constructing the Rounded Income

To achieve the second goal of a realistic rounding behaviour, we model the rounding behaviour in the PASS survey and use the estimated parameters to round the income in the population. Specifically, we assume that the tendency to round the reported income

value only depends on the true income y and can be modeled as a seven categories ordered probit model, i.e., we assume the latent rounding behaviour can be modeled as $g \sim N(\gamma \cdot \log(y); \sigma^2 = 1)$. The seven categories reflect the different rounding bases, i.e., rounding to the nearest 1, 5, 10, 50, 100, 500, and 1,000 euro value. By maximizing the likelihood in (4) for the PASS survey we obtain estimates for γ and the thresholds that determine the different rounding categories. We use these estimates to calculate rounding probabilities given the predicted rounding behaviours $E(\hat{g}_i|X_i)$ for each record i, $i = 1, ...N$. In the final step a rounding base is determined for each record by randomly picking one of the rounding categories according to the given rounding probability. The reported income is obtained by rounding the generated income according to the drawn rounding base.

4.3 The Repeated Sampling Design

From the constructed population we repeatedly draw simple random samples of size $n = 5,000$. We assume that only the rounded income is observed in the sample. To impute the true income we use two strategies. For the first strategy, which we call the naïve imputation strategy, we assume that the interval in which the true income must fall is always defined as the maximum possible interval given the observed income, for example, if the last digit is zero, we assume that the income was rounded to the nearest 10 euro value, if the last two digits are zero, we always assume that the income was rounded to the nearest 100 euro value, etc. Under this assumption the true income could be imputed according to the simplified approach described in Section 3.1. Implementing this strategy serves to illustrate that such a simplified imputation technique can lead to biased inferences based on the imputed data if the true rounding base is unknown. For the second strategy, which we call the improved imputation strategy, the rounding probabilities are estimated and the true income is imputed according to the steps described in Section 3.2. We assume that the models to impute the two missing variables G and $income$ are correctly specified, i.e. the underlying model assumptions are:

$$\begin{aligned} income &\sim \log N(X'\beta, \sigma^2), \\ G &\sim N(\gamma \cdot \log(income); \sigma^2 = 1). \end{aligned}$$

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% of the median income in the population. We note however, that any other quantity could be estimated from the imputed data using standard procedures without the need to account for the rounding effects. To estimate the variance of the estimated poverty measure, we use the variance estimator suggested by Preston (1995). We repeat the whole process of sampling, imputing and analyzing the data 1,000 times.

Table 2: Simulation results. The poverty rate (pr) in the population is 18.46%.

	mean($\hat{pr}$)	Var($\hat{pr}$)	mean($\widehat{Var}(\hat{pr})$)	Variance ratio	95% Coverage rate
True income	18.44	$2.49*10^{-5}$	$2.62*10^{-5}$	1.05	95.3
Rounded income	19.20	$3.27*10^{-5}$	$2.63*10^{-5}$	0.80	67.4
Naïve imputation	18.02	$2.20*10^{-5}$	$3.19*10^{-5}$	1.45	92.5
Improved imputation	18.52	$2.34*10^{-5}$	$3.02*10^{-5}$	1.29	97.6

4.4 Results

The results are summarized in Table 2. The four rows of the table present the results under different analysis scenarios. The first row provides the results if the true income were available in the sample. The second row contains the results, if only the rounded income were available but the analyst treated it as if it would be the true income. The third and fourth row provide the results if the analyst used the naïve or the improved imputation approach respectively. The first column contains the average point estimates across the 1,000 simulation runs. Given that the poverty rate in the population is 18.46%, we find that the estimate of the poverty rate is unbiased only if the true income is available or if the improved imputation method is used. If no adjustments are made the poverty rate based on the rounded income is overestimated by 0.74 percentage points. The naïve imputation underestimates the true poverty rate by more than 0.4 percentage points. The fourth column contains the ratio of the average of the estimated variances (using the multiple imputation combining rules for row three and four) over the true variance of the estimated poverty rates across the 1,000 simulation runs. If the variance estimate is unbiased this ratio should be one. Only the variance estimate based on the true income is unbiased. The variance is underestimated for the rounded data while both imputation methods overestimate the true variance. The overestimation is a direct result of the fact that the estimated poverty rate is a function of a sample quantile (the median) and it is well known that the MI variance estimate tends to be conservative if sample quantiles are used to estimate the quantiles in the population (see for example Meng (1994) for a discussion of related issues). In fact, we verified in a small simulation that the variance estimates would be unbiased for other parameters of interest such as the average income (results not shown for brevity). It is also interesting to note that the additional information used in the imputation model regarding the parametric distribution of the income which is not used in the analysis model leads to superefficiency as defined by Rubin (1996), i.e., the true variance after imputation (reported in column 2) is smaller than the variance based on the original data. In any case, the MI variance estimate will always ensure confidence validity, i.e., the actual coverage will never be smaller than the nominal coverage, which is not true if the rounded income is used directly. The last column reports the percentage of times the 95% confidence interval for the estimated poverty rate contains the true poverty rate. The confidence interval of the rounded income without adjustments clearly has less than nominal coverage due to bias and underestimation of the true variance, the other coverage rates are close to the nominal coverage rate with a small undercoverage for the naïve imputation method and a small overcoverage for the improved imputation method which is a direct result of the conserva-

tive variance estimate. Given these results it is obvious that (except for the results based on the true income which would not be available in practice) only the improved imputation method provides unbiased point estimates and a confidence interval with at least the nominal coverage rate.

5 Application to the Panel Study "Labour Market and Social Security"

In this section we apply the imputation approach to the German panel study "Labour Market and Social Security". The panel study, 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 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).

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. As discussed in the introduction (see Table 1) the exact reported income seems to be subject to rounding. More than 80% of the reported income values are divisible by 10 and more than 60% are divisible by 100. To obtain estimates for the true income for this dataset according to the procedure described in Section 3.2, we need to set up the models for the rounding behaviour and the income distribution. For the rounding behaviour we stick with the simple model already used for the simulation study, i.e., we assume that the tendency to round only depends on the true income. This model could certainly be extended in practice but the model evaluations discussed in the following section indicate a good model fit for this simple model. Given 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 (since all income values are reported as integers, we actually assume that all respondents round at least to the nearest 1). To model the true income, we assume a log-normal distribution for income conditional

Table 3: 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

on a set of covariates X. Details about the covariates included in the model are contained in Table 3.

In the model sparsely populated categories among the X variables containing less than 5% of the records are always collapsed. We also drop households that claimed to receive unemployment benefits while at the same time having a monthly household income of more than 5,000 euros. Such a high income is unrealistic for unemployment benefits recipients under the German Social Security System. Furthermore, we exclude households with income below (above) the 0.5% (99.5%) percentile of the income distribution before maximizing the likelihood since these records were identified as influential outliers which caused problems during the maximization. However, these records are still included when the poverty measures are calculated. Since only records in the very tails of the income distribution are affected, the fact that the imputation model to "unround" the reported income might be misspecified for those records has no impact on the computed poverty measure. Finally, we standardize each variable in the dataset to avoid multicollinearity issues and problems due to the large differences in the range of the variables in X when estimating the parameters of the model. This also means that we impute the standardized income. The final income is obtained by backtransforming the imputed value to its original scale.

5.1 Evaluation of the Model Assumptions

Since the proposed rounding adjustment strategy is purely model based, an evaluation of the model assumptions is essential. However, a direct evaluation of the two models is difficult since both dependent variables – the true income and the rounding behaviour – are not observed. Thus, we rely on posterior predictive simulations (Gelman et al., 2004: Chap. 6) to check whether our model assumptions are reasonable.

Table 4: 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	98.69	94.87	98.03	98.21	96.28	97.94
95.00	95.86	92.96	94.15	94.43	93.75	95.14
90.00	93.11	90.27	90.66	90.06	89.95	90.78

5.1.1 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 (4). 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 not divisible by 5). 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 4. The empirical coverages are always very close to the nominal coverages indicating a good fit for the income model.

5.1.2 The Rounding Behaviour Model

To evaluate the quality of the rounding behaviour model, we repeatedly re-round the imputed (unrounded) income variable 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 5 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

Table 5: 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	5	10	50	100	500	1,000
Observed income (%)	3.51	12.73	8.04	37.34	10.11	13.37
Unrounded income (%)	10.03	8.28	1.15	1.06	0.13	0.27
Re-rounded income (%)	2.64	13.33	9.85	46.64	8.62	9.59

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 limited to wave 6 from the PASS data for readability. Similar results were obtained for the other waves. As expected the percentages differ substantially between the observed income and the unrounded income. Most of the values (70.07%) in the unrounded data (second row in the table) are not divisible by any of the numbers 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 14.90% of the data are not divisible by any of the given numbers 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 (46.64%). This overestimation leads to a slight underestimation of the percentage of records that are not divisible (9.59%) by any of the numbers. For the remaining categories the percentages based on the re-rounded data are fairly close to the percentages based on the observed data indicating a good fit of the rounding behaviour model.

5.2 Results

We evaluate the rounding effects on the poverty rate for all six waves of the PASS survey available so far. 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). Since the main aim of the paper is to illustrate the effects of rounding, observations with missing data in any of the variables included in the model are deleted for simplicity. Incorporating the rounding procedure into a sequential regression multivariate imputation (SRMI, Raghunathan et al. (2001)) procedure would be straightforward. The parameters found by maximizing the likelihood in (4) can also be used to impute missing income values.

Table 6 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 (2013)). The first column contains the results based on the original rounded data without any adjustments. The second column contains the results for the multiply imputed true income based on $m = 25$ imputations. The 95% confidence intervals

Table 6: Estimated poverty rates from the PASS survey (with 95% confidence intervals reported in brackets).

wave	original data	corrected data
wave 1	17.31	16.35
	(15.79;18.83)	(15.14;17.55)
wave 2	16.91	16.98
	(15.76;18.05)	(15.69;18.27)
wave 3	14.27	15.40
	(12.22;16.33)	(13.91,16.90)
wave 4	14.89	14.61
	(13.64;16.15)	(13.40;15.81)
wave 5	16.34	15.75
	(14.80;17.88)	(14.41;17.10)
wave 6	15.95	16.27
	(14.42;17.48)	(14.81;17.72)

reported in brackets are based on bootstrap variance estimates.

The rounding effects vary from year to year. In some years the estimated poverty rate only differs slightly (< 0.5 percentage points) between the original data and the corrected data (waves 2, 4, and 6). In other years the poverty rate is considerably smaller after the correction (waves 1 and 5). Finally, the poverty rate can also be substantially larger after the correction (wave 3). These different effects of the rounding are not surprising. Whether the poverty rate is smaller or larger in the rounded data depends to a large extent on whether the true median income is close to one of the spikes in the rounded data. In these cases the median computed from the rounded data will likely be equal to the income value at the spike, i.e., it will overestimate the true median if the true median is below the spike and it will underestimate if the true median is above the spike. If the median is estimated too large, the poverty threshold (60% of the median income) is also too large and in general the poverty rate will be overestimated (unless there is a counter balancing effect of the rounding for the low-income group, i.e., the estimated poverty threshold is slightly below another spike in the rounded data). Similarly, if the median is estimated too low, the poverty threshold will be underestimated and the poverty rate will be too small (again ignoring any effects at the threshold). In any case it is obvious that rounding can have a strong impact on measures such as the poverty rate and ignoring this effect will generally give misleading results.

6 Conclusion

Obtaining valid information on the income distribution from survey data is a difficult task. There is ample discussion in the literature of potential biases from high nonresponse rates for income questions that are coupled with a missing data mechanism that is definitely not missing completely at random. While most researchers try to adjust their analyses to

account for the nonresponse another phenomenon is widely ignored: The potential bias that might arise because survey respondents hardly ever report their exact income, and instead provide only a rounded estimate. In this paper we have illustrated the potential negative effects this rounding can have on important measures such as the poverty rate. We proposed an imputation procedure that generates estimates of the true income given the reported rounded income and showed that unbiased estimates can be obtained from the imputed data by simply using the standard multiple imputation combining rules proposed by Rubin (1978). The major advantage of the approach lies in the fact that the analyst no longer needs to worry how to best adjust his or her analysis for the rounding effects. Standard analysis procedures on the imputed data will give valid inferences. This is especially important since most inferences – not only the poverty rate – will be biased if based on the rounded data. Although it would probably be possible to develop adjustments for each estimate, the imputation approach works as an omnibus tool since the imputed income values can be treated as the true income and no further adjustments are necessary. Of course this only holds if the models for the imputation are correctly specified. As with any imputation method, a misspecified imputation model will always lead to biased results. Thus, careful model evaluations akin to the evaluations presented in Section 5 should always be conducted. Still, it is important to keep in mind that the common practice of ignoring the rounding completely will be guaranteed to give biased results.

It should 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 or her 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 data collected based on this approach.

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. The authors find (looking only at individuals with a total family income below $52,500) that in the National Health Interview Survey (NHIS) which uses the screener variable approach, 35.6% of the individuals reported an income divisible by $5,000 and 20.9% reported an income divisible by $10,000. In the CPS (ACS) those numbers reduced to 11.0% (16.2%) and 6.2% (9.5%) 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 individual components, for two reasons. First, we believe that most of the survey respondents will not have an income in all the categories that are usually listed. 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% 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 this rounding behaviour will already negatively affect the inferences obtained from the rounded data. Thus, while the effect of rounding on official 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.

References

Bollinger, C.R.; Hirsch, B.T. (2013): Is Earnings Nonresponse Ignorable? In: The Review of Economics and Statistics, Vol. 95, p. 407–416.

Clementi, F.; Gallegati, M. (2005): Pareto's law of income distribution: Evidence for Germany, the United Kingdom, and the United States. In: Chatterjee, A.; Yarlagadda, S.; Chakrabarti, B. (Eds.) Econophysics of wealth distributions, Milan: Springer, p. 3–14.

Czajka, J.L.; Denmead, G. (2008): Income Data for Policy Analysis: A Comparative Assessment of Eight Surveys. Final report to the U.S. Department of Health and Human Services submitted by Mathematica Policy Research, Inc.

Dempster, A.P.; Rubin, D.B. (1983): Rounding error in regression: the appropriateness of Sheppard's correction. In: Journal of the Royal Statistical Society B, Vol. 45, p. 51–59.

Eurostat (2013): Glossary: Equivalised disposable income - Statistics Explained (2013/6/2). http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Equivalised_disposable_income>.

Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. (2004): Bayesian Data Analysis. London: Chapman and Hall, second ed..

Heitjan, D.F. (1994): Ignorability in General Incomplete-Data Models. In: Biometrika, Vol. 81, p. 701–708.

Heitjan, D.F. (1989): Inference from grouped continuous data: a review. In: Statistical Science, Vol. 4, p. 164—-179.

Heitjan, D.F.; Rubin, D.B. (1991): Ignorability and Coarse Data. In: The Annals of Statistics, Vol. 19, p. 2244–2253.

Heitjan, D.F.; Rubin, D.B. (1990): Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping. In: Journal of the American Statistical Association, Vol. 85, p. 304–314.

Liu, T.Q.; Zhang, B.X.; Hu, G.R.; Bai, Z.D. (2010): Revisit of Sheppard corrections in linear regression. In: Science China Mathematics, Vol. 53, p. 1435–1451.

Manski, C. F.; Molinari, F. (2010): Rounding Probabilistic Expectations in Surveys. In: Journal of Business & Economic Statistics, Vol. 28, p. 219–231.

Meng, Xiao-Li (1994): Multiple-imputation Inferences With Uncongenial Sources of Input (Disc: P558-573). In: Statistical Science, Vol. 9, p. 538–558.

Preston, I. (1995): Sampling Distributions of Relative Poverty Statistics. In: Journal of the Royal Statistical Society. Series C, Vol. 44, p. 91–99.

Raghunathan, T. E.; Lepkowski, J. M.; van Hoewyk, J.; Solenberger, P. (2001): A multivariate technique for multiply imputing missing values using a series of regression models. In: Survey Methodology, Vol. 27, p. 85–96.

Rubin, D. B. (1996): Multiple Imputation After 18+ Years. In: Journal of the American Statistical Association, Vol. 91, p. 473–489.

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

Schenker, Nathaniel; Raghunathan, Trivellore E.; Chiu, Pei Lu; Makuc, Diane M.; Zhang, Guangyu; Cohen, Alan J. (2006): Multiple Imputation of Missing Income Data in the National Health Interview Survey. In: Journal of the American Statistical Association, Vol. 101, p. 924–933.

Schneeweiss, H.; Komlos, J.; Ahmad, A.S. (2010): Symmetric and asymmetric rounding: a review and some new results. In: Advances in Statistical Analysis, Vol. 94, p. 247–271.

Sheppard, W.F. (1898): On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. In: Proceedings of the London Mathematical Society, Vol. 29, p. 353–380.

Trappmann, M.; Gundert, S.; Wenzig, C.; Gebhardt, D. (2010): PASS: a household panel survey for research on unemployment and poverty. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 130, p. 609–622.

U.S. Census Bureau (2013): Current-Population Survey (CPS) – Imputation of Unreported Data Items. http://www.census.gov/cps/methodology/unreported.html (retrieved on 19.12.2013).

U.S. Census Bureau (2009): Design and Methodology – American Community Survey. U.S. Government Printing Office, Washington, D.C.

Van der Laan, J.; Kuijvenhoven, L. (2011): Imputation of rounded data. Statistics Netherlands Discussion Paper no. 201108.

Recently published

No.	Author(s)	Title	Date
8/2013	Alm, B. Engel, D. Weyh, A.	Einkommenseffekte von Betriebswechslern: Neue Befunde für Ostdeutschland	6/13
9/2013	Pauser, J.	Capital mobility, imperfect labour markets, and the provision of public goods	8/13
10/2013	Bauer, A.	Mismatch unemployment: Evidence from Germany 2000-2010	8/13
11/2013	Werner, D.	New insights into the development of regional unemployment disparities	8/13
12/2013	Eggs, J.	Unemployment benefit II, unemployment and health	9/13
13/2013	Vallizadeh, E. Muysken, J. Ziesemer, Th.	Migration, unemployment, and skill downgrading: A specific-factors approach	9/13
14/2013	Weber, E. Zika, G.	Labour market forecasting: Is disaggregation useful?	9/13
15/2013	Brenzel, H. Gartner, H. Schnabel, C.	Wage posting or wage bargaining? Evidence from the employers' side	9/13
16/2013	Dengler, K.	The effectiveness of sequences of One-Euro Jobs	10/13
17/2013	Hutter, Ch. Weber, E.	Constructing a new leading indicator for unemployment from a survey among German employment agencies	10/13
18/2013	Schwengler, B.	Einfluss der europäischen Regionalpolitik auf die deutsche Regionalförderung	10/13
19/2013	Bossler, M.	Recruiting abroad: the role of foreign affinity and labour market scarcity	11/13
20/2013	Forlani, E. Lodigiani, E. Mendolicchio, C.	The impact of low-skilled immigration on female labour supply	11/13
21/2013	Singer, Ch. Toomet, O.-S.	On government-subsidized training programs for older workers	12/13
22/2013	Bauer, A. Kruppe, Th.	Policy Styles: Zur Genese des Politikstilkonzepts und dessen Einbindung in Evaluationsstudien	12/13
1/2014	Hawranek, F. Schanne, N.	Your very private job agency	1/14

As per: 2014-01-14

For a full list, consult the IAB website
http://www.iab.de/de/publikationen/discussionpaper.aspx

Imprint

IAB-Discussion Paper 2/2014

Editorial address
Institute for Employment Research
of the Federal Employment Agency
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Regina Stoll, Jutta Palm-Nowak

Technical completion
Jutta Palm-Nowak

Website
http://www.iab.de

Download of this Discussion Paper
http://doku.iab.de/discussionpapers/2014/dp0214.pdf

ISSN 2195-2663

For further inquiries contact the author:

Jörg Drechsler
E-mail joerg.drechsler@iab.de