Imputation for hierarchical datasets and responses in intervals

Matthias Speidel

München 2018

Imputation for hierarchical datasets and responses in intervals

Matthias Speidel

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Matthias Speidel
aus Reutlingen

München, den 06.06.2018

Erstgutachter: Prof. Dr. Christian Heumann
Zweitgutachter: Prof. Dr. Stef van Buuren
Drittgutachter: Prof. Dr. Martin Spieß
Tag der Disputation: 23.11.2018

Contents

Acknowledgments

My deepest gratitude goes to ...

- Jörg Drechsler, who introduced me to the academic world and was a great guide through it, and Christian Heumann who took care of me as an external student and guided me through the process of finishing this thesis.
- Stef van Buuren and Martin Spieß for being the external reviewers; Thomas Augustin and Helmut Küchenhoff for being part of the examination committee.
- Jörg Drechsler, Shahab Jolani, Hans Kiesl and Joseph Sakshaug for collaborating on the articles building the basis of this thesis.
- Frauke Kreuter and all colleagues at the Statistical Methods group at the Institute for Employment Research for making the time in the academic world so enjoyable.
- All current and past members of the Department of Statistics at the Ludwig-Maximilians-Universität München, especially Thomas Augustin, for the fruitful discussions at various meetings.
- Shahab Jolani for enabling my research visit at the Maastricht University.
- Malte Schierholz and Julia Plaß for their suggestions for improvements regarding the draft of this thesis.
- Julia Plaß for being a great mental support.
- My family for understanding the restraints of a doctoral candidate.

Zusammenfassung

Viele Datensätze haben nicht die Form, wie es die standard Analyse-Methoden voraussetzen. Variablen können das Problem mit sich bringen, dass einzelne Antworten fehlen, (mutmaßlich) gerundet wurden, oder nur als Intervall vorliegen - auch alle Defizite gleichzeitig sind möglich. In der Literatur wird in diesen Fällen von "groben Daten" gesprochen. Neben der Anpassung der Analyse-Methode, ist ein etabliertes Vorgehen das Ersetzen von groben Daten mit plausiblen Werten, genannt "Imputation". Je nach Datensituation bzw. Analysevorhaben, fallen die Imputationsmethoden allerdings unterschiedlich aus. In den Beiträgen zu dieser kumulativen Dissertation wird die Notwendigkeit, sowie die theoretische und praktische Machbarkeit der adäquaten Imputation von groben Daten in einer Vielzahl von Fällen dargestellt.

Eine Methode zur gleichzeitigen Imputation von fehlenden, gerundeten und Intervall-Beobachtungen in einer Zielvariablen wird in Beitrag 1 vorgestellt. Ein Modell für die latente Rundungstendenz und ein Modell für die Zielvariable werden gemeinsam behandelt. Dabei trägt jede Beobachtung zu einer gemeinsamen Likelihood bei. Basierend auf den Maximum-Likihood-Schätzern werden für die groben Werte plausible Werte aus einer (trunkierten) Verteilung gezogen.

Beitrag 2 behandelt Methoden für den Umgang mit fehlenden Daten in hierarchischen Datensätzen. Zwei häufig diskutierte Methoden werden analytisch verglichen. Es wird gezeigt, warum Imputation basierend auf cluster-spezifischen fixen Effekten zu systematisch höheren Varianzen der zufälligen Effekte führt und welche Kenngrößen den Bias in welche Richtung beeinflussen. Basierend auf einer Simulationsstudie wird der Bias (in den Simulationssettings) quantifiziert. Desweiteren wird motiviert, warum die angenommene Funktion des Ausfallprozesses, innerhalb eines Fehlendmechanismus, das Ausmaß des Bias beeinflussen kann.

Mit Beitrag 3 wird ein R-Paket der Öffentlichkeit frei zugänglich gemacht, mit dem fehlende Beobachtungen in hierarchischen Daten für eine Vielzahl von Variablentypen imputiert werden können. Ebenfalls implementiert ist die Methode zur gleichzeitigen Imputation von fehlenden, gerundeten und Intervalldaten. Teilaspekte dabei sind die technische Verarbeitung von Intervallinformationen in einer Variable, sowie damit verbundene Methoden zur Handhabung. Bezogen auf gerundete Daten wird eine Heuristik zur Erkennung von möglichen Rundungsgraden vorgestellt.

Summary

Many data sets do not have the form required by standard analysis methods. Variables can entail the problem that some responses are missing, (probably) rounded, or only available in an interval - also all deficiencies can be present simultaneously. In literature, such cases are called “coarse data”. Beside the adjustment of the analysis method, an established approach is the replacement of coarse data by plausible values, called “imputation”. Dependent on the data situation, respectively the desired analysis, the imputation methods differ. The contributions of this cumulative dissertation present the necessity, along with the theoretical and practical feasibility of an appropriate imputation of coarse data for a variety of cases.

A method for the simultaneous imputation of missing, rounded and interval-observations in a target variable is presented in contribution 1. A model for the latent rounding tendency and a model for the target variable are treated jointly. In the process each observation contributes to a common likelihood. Based on the maximum likelihood estimates, the coarse data are replaced by plausible values drawn from a (truncated) distribution.

Contribution 2 covers methods for the handling of coarse data in hierarchical data sets. Two frequently discussed methods are compared analytically. It is shown why imputation based on cluster specific fixed effects lead to systematically higher variances of the random effects and which parameters influence the bias in which direction. Via a simulation study, the bias (under the simulation settings) is quantified. Furthermore, a motivation is given, why within a missing mechanism, the presumed function about the missing process can influence the bias.

With contribution 3, an R package is made freely available to the public, allowing the imputation of missing values in hierarchical data for a variety of variable types. Likewise, the method for simultaneously imputing missing, rounded and interval data is implemented. Partial aspects of this are the technical realization of interval information in one variable along with related routines for their handling. With reference to rounded data, a heuristic for the detection of possible rounding degrees is presented.

List of Figures

List of Tables

1 Introduction

Charles Babbage, polymath and inventor of the *Difference Engine*, a mechanical calculator, wrote in his 1864 book that he was asked twice "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answer come out?"(Babbage, 1864, p. 67). Even if he was able to give an answer, Babbage admitted "I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."(Babbage, 1864, p. 67).

This anecdote is instructive in many ways. It shows that the problem of "wrong figures" in computations together with the demand for correct(ed) results existed already more than 150 years ago; and it points out misconceptions about the power of machines and the need to see things from the applicant's perspective, paired with the competence explaining the machine's mechanisms.

1.1 Motivation

Nowadays "wrong figures", or more general, data where not the exact value is observed, but a set of observations containing the true value, called *coarse data* (Heitjan and Rubin, 1991), are omnipresent. When data are collected, the unit of observation might not be able to be contacted/observed at all, or if it is a person, successfully contacted, he/she might refuse to participate entirely or to answer specific questions. Even a given answer is not guaranteed to be a precise, correct answer: they might actually be incorrect or imprecise (cf. Lynn 2008 for an overview of possible factors confounding the data collection).

Within a data set the extent of coarse data can be considerable. The impacts of using coarse data are often overlooked in practice (Sterne et al., 2009), but can be severe. Broadening the knowledge and awareness about the impacts of coarse data and how to cushion negative effects has become a field of research in (applied) statistics and applied sciences.

1.2 Answered and open questions in literature

Statistical literature is providing ideas and practical solutions for different situations in which coarse data appear. An overall consensus in literature is that ignoring the problem of coarse data (including the removal of observations with coarse data) is only viable in very limited settings (Schafer, 1999; White and Carlin, 2010; de Jong et al., 2016; White et al., 2011; van Buuren, 2012). Approaches generally viable are the adjustment of the analysis model and the modification of the data. One, nowadays broadly accepted, approach of data modification is *multiple imputation* (Rubin, 1987), where (presumably) coarse data are replaced by "plausible values". Multiple imputation for selected cases of coarse data is the topic of this thesis.

Literature concludes that different desired analysis models (due to different data situations) require different imputation methods (Meng, 1994). Therefore a major question in this field of research is, what imputation method to use in the given data and analysis situation.

Ample research has been conducted when the desired analysis model is a generalized linear model and when missing values in the target variable of this analysis model shall be imputed (e.g. Rubin 1987; Schafer 1999; White et al. 2011; van Buuren 2012). A first rather open question, not considered in this thesis, is the imputation of covariates from the analysis model (White and Carlin 2010; Grund et al. 2016). Covariates in the analysis model normally need no distributional assumption, but when they are imputed this additional assumption has to be made (Bartlett et al., 2015). Literature has a tendency towards that all variable lacking some observations should be imputed without respect to their position in the analysis model, but some authors found this approach to be problematic in some settings (Grund et al. 2016).

A second question is the imputation of coarse, but not missing data (Heitjan and Rubin, 1990). In literature methods have been proposed to impute rounded data (when for example a large portion of individuals report a value divisible by 1000) (e.g. Wang and Heitjan 2008; van der Laan and Kuijvenhoven 2011; Zinn and Würbach 2016) or responses given as intervals (e.g. Law and Brookmeyer 1992; Dorey et al. 1993; Heeringa 1993; Raghunathan et al. 2001; Royston 2007). For the situation when missing data, rounded data and interval data are present at once, a gap was found in literature.

A third question is the imputation of missing data when the desired analysis model is a generalized linear mixed model, considering a hierarchical structure of the data (e.g. Reiter et al. 2006; van Buuren 2011; Enders et al. 2016; Zhou et al. 2016; Lüdtke et al. 2017). Literature focused on mainly two different imputation approaches for hierarchical data, compared them and evaluated their performances. But all articles where limited in their setting under consideration or relied on simulation studies.

An aspect not to neglect in the science about appropriate imputation methods and in applied research is the availability of suitable imputation methods: if the scientist is not willing or able to implement the imputation method it cannot be used and compared in her/his research (Azur et al., 2011). A lack of imputation routines for coarse data and missing hierarchical data has been found.

1.3 Contributions

The thesis provides three contributions to the issues mentioned in the previous section:

	full reference	*cited as*
Contribution 1	Jörg Drechsler, Hans Kiesl and Matthias Speidel (2015): MI double feature: Multiple Imputation to address nonresponse and rounding errors in income questions. *Austrian Journal of Statistics 44* (2), 59-71.	Drechsler et al. (2015)
Contribution 2	Matthias Speidel, Jörg Drechsler and Joseph Sakshaug (2018): Biases in Multilevel analyses caused by cluster-specific fixed effects imputation. *Behavior Research Methods 50* (5), 1824-1840. First Online: 24 August 2017.	Speidel et al. (2018)
Contribution 3	Matthias Speidel, Jörg Drechsler and Shahab Jolani (2018), hmi: Hierarchical Multiple Imputation. R package version 0.9.13 and Matthias Speidel, Jörg Drechsler and Shahab Jolani (2018). R Package hmi: a Convenient Tool for Hierarchical Multiple Imputation and Beyond. *IAB-Discussion Paper 16/2018*. Manuscript submitted to the *Journal of Statistical Software*.	Speidel et al. (2018a) and Speidel et al. (2018b)

The gap of simultaneous imputation of missing, rounded and interval data was closed by contribution 1 (Drechsler et al., 2015). Chapter 2, gives a general introduction to the different types of coarse data including their occurrence and reasons, followed by current strategies of handling coarse data. Into this literature, the approach from contribution 1 is embedded. Chapter 3 sets up the notation for linear mixed models and the two imputation models that are most often considered in literature, followed by a short summary of the current state of the literature. The main part of this section is covered by contribution 2 (Speidel et al., 2018), which provides more general and analytical results to the research about hierarchical imputation. Chapter 4 covers the implementation of suitable imputation routines in statistical software. Beside the deployment of yet unimplemented routines, contribution 3 (Speidel et al., 2018a,b) provides a new approach on processing interval data. The declarations about the contributions to each publication are given at the beginning of these chapters. Chapter 5 ends the thesis with a short conclusion and an outlook on further fields of study.

2 Imputation for missing, interval and rounded values

In contribution 1 (Drechsler et al., 2015) a method for *simultaneously* imputing missing, interval and rounded data was proposed. The description of this method is embedded into the literature about the occurrence of such coarse data, and the literature on imputing them.

Declaration of contributions to the article: Generally speaking, the ideas in the article about using imputation for un-rounding rounded responses were written by Jörg Drechsler and Hans Kiesl based on their work that later was published in Drechsler and Kiesl (2016), while the parts to impute missing data and values given in intervals were written by the author, based on ideas of Jörg Drechsler and Hans Kiesl. In more detail: the author's contribution was to write the parts about missing data and the attempts to reduce nonresponse by survey agencies and the implications for the Panel Study Labour Market and Social Security (PASS), and parts of the statistical details. For the real data application, the author extended the R code from Jörg Drechsler and Hans Kiesl to be able to consider missing and interval values and conducted the adapted analysis. The results after using imputation for all situations are given by the author, including the evaluation of the Gibbs-samples run during the imputation. Minor contributions to the conclusion and outlook were made by the author.

2.1 Motivation: occurrence of coarse data

Missing data, interval data or rounded data are found in nearly every real data set. An empirical evaluation of the frequency and methodological/psychological reasons for their occurrence are presented. In this thesis *coarse data* encompasses only missing data, interval data and rounded data. Others of the manifold types of measurement errors are out of scope of this thesis. For a handbook on measurement errors see for example Buonaccorsi (2010) or Biemer et al. (2017).

2.1.1 Missing data

Against the ideal condition, in practice it is unlikely that every unit selected into the sample will be observed. In the field of survey methodology Groves and Peytcheva (2008) found in a meta-study response rates between 14 and 72% with an average of 36%. In accordance with this, Kreuter (2013) stated that unit response rates "[...] in the 30 to 40 percent range [are] common to many household surveys in Europe"(Kreuter, 2013, p. 24). Common reasons for missing data on the unit level (*unit nonresponse*) are failure to

locate or to contact the sample unit, refusal or inability of the sample unit to participate, inability of data collector and sample unit to communicate adequately or technical failure (Lynn, 2008, p. 37). The consequences of unit nonresponse can be serious. In addition to the loss of statistical power (Rubin, 1987; de Leeuw et al., 2003), a low unit response rate can, but do not necessarily leads to, nonresponse bias (Rubin, 1987; Groves, 2006).

A pivotal characteristic of data sets is whether respondents and nonrespondents are systematically different (Rubin, 1978; Griffin et al., 2011). This holds for both, missing sampling units and missing data in response variables of a sampling unit (*item nonresponse*). An early article looking at item nonresponse is Ferber (1966). In the survey under examination by Ferber, only 37.5% of the respondents returned a questionnaire without any missing information. Several years later, Denscombe (2009) found item nonresponse rates in a survey to be between 0 and 21.3%. The reasons for item nonresponse are manifold but major issues stated in literature are the sensitivity and complexity of a question (Shoemaker et al., 2002; Yan et al., 2010). The impact of item nonresponse is considerable: without further decisions (like dropping units with item nonresponse or imputing missing values), most desired statistical analyses (like regression) are simple not doable (Raudenbush and Bryk, 2002). Anticipating Section 2.2: restricting the analysis to complete observed units reduces the statistical power and might cause a substantial nonresponse bias (Raudenbush and Bryk, 2002).

2.1.2 Interval data

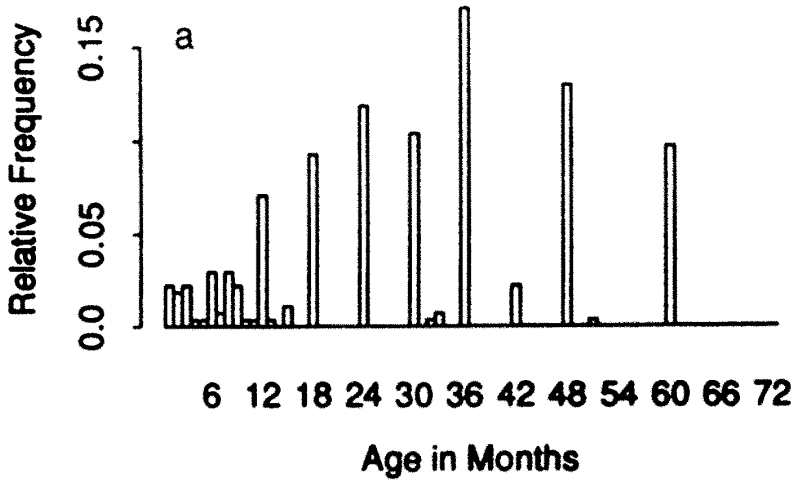
As mentioned in Section 2.1.1, the sensitivity or complexity of a question increases the chances of observing nonresponse. When survey participants are asked to provide their income or other wealth related information *both* issues are present. Some participants are, due to sensitivity concerns not willing to provide an exact answer, while others are not able to remember the exact value and thus report "Don't Know". Both issues lead to the fact that in surveys, income or other wealth related questions are often amongst those with the highest nonresponse rate (cf. e.g. Heeringa 1993, Loosveldt et al. 2002, or Schenker et al. 2006).

To accommodate the issues of those participants in some surveys individuals who initially refused to answer or said "Don't Know" are asked a follow up question to get some range of the true value. This was for example done in the US-American *Survey of Consumer Finances* (SCF) by letting the respondents select a range from a shown *range card* (e.g. \$0 - \$5000, cf. Kennickell 1991), leading to interval responses. Asking initial nonresponders for a range makes it possible to collect at least some information from participants who initially refused to give an answer, but are willing to provide some range (Drechsler et al., 2015) or from participants who reported to "don't know" the precise value, but who can name an interval covering the real value (Kennickell, 1996). In this way, nonresponse-rates are reduced (Juster and Smith, 1997; Drechsler et al., 2015). A consequence of interval data is that most standard analyses, only designed for point precise data, are thus not doable. For that reason, new analysis models, like the Tobit model (Tobin, 1958) had to be developed.

2.1.3 Rounded data

A further, less prevalent behavioral pattern of respondents is the tendency to report a value that is a multiple of a rounding base. For example Heitjan and Rubin (1990) studied the reported age of children. The empirical distribution showed that the majority of responses were multiples of 6 months (see Figure 2.1). The rounding bases considered by the authors in this case were 6 and 12. A similar pattern can be found for example in the German

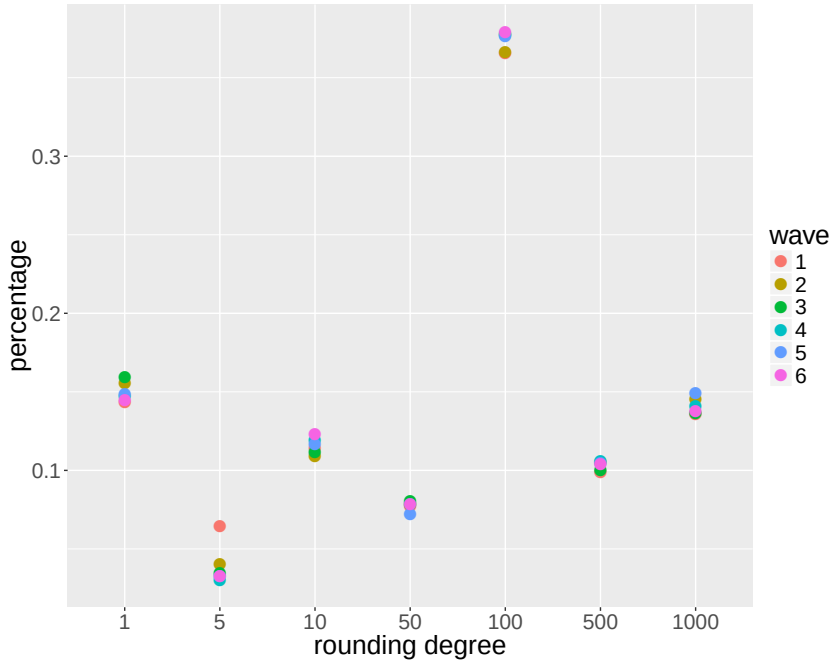
Figure 2.1: Part (a) from Figure 1 in Heitjan and Rubin (1990), p. 306, showing the relative frequencies of reported ages in months of children in Tanzania



Panel Study Labor Market and Social Security (PASS, Trappmann et al. 2010) where the monthly net income of households was surveyed. In the PASS data a noticeable portion of respondents reported a monthly net income divisible by 10, 100, 500 or 1000 (e.g. 3000 or 4500). For the first six waves (from 2006 - 2012) of the PASS data, the percentages of households reporting data with a certain rounding degree are plotted in Figure 2.2. Beside the fact that the percentages are quite stable across the different years, it is noticeable that about 15% of respondents reported a value divisible by 1000 and more than 35% reported a value to be divisible by 100 (but not 1000) which remarkably exceeds the expected numbers under a nonheaped distribution. A psychological explanation for rounding starts with the observation that multiples of 10, like 10, 50, 100 or 1000 are *cognitive reference points* (Rosch, 1975). Secondly, a trait of many people is to only remember the first digit(s) of their income precisely and replace the forgotten digits by zeros (Hanisch, 2005). Both circumstances combined lead to responses which are rounded to multiples of 10, 100, or 1000. The accumulation of individually rounded values, leads to the result of heaped variables.

Rounding in continuous variables can have considerable impacts on the analysis: amongst

Figure 2.2: Rounding degrees in the PASS data over multiple waves



the potentially biased quantities of interest are the moments of a normally distributed variable (Sheppard, 1898), regression coefficients (Augustin and Wolff, 2004) or the rate of people in danger of poorness (60% of the median income, Drechsler and Kiesl 2016) (see Schneeweiss et al. 2010 for a review on this topic).

2.2 General strategies on handling coarse data

The best action regarding coarse data is self-evidently the attempt to avoid observing coarse data. For literature on prevention of coarse data see for example de Leeuw et al. (2003) or Biemer et al. (2017). Still, in most cases the occurrence of coarse data cannot be averted completely, so strategies to handle coarse data had to be elaborated. For unit nonresponse, *nonresponse weighting* is a commonly used strategy (Kreuter et al., 2009). For item nonresponse, which is in the focus of this thesis, very roughly speaking there are two general distinct strategies on handling coarse data:

1. Modifying the analysis models (like adjusted Maximum Likelihood estimation or using the EM-Algorithm Dempster et al. 1977)

2. Modifying the data set (like removing coarse observations or using imputation Rubin 1987)

Both strategies have their pros and cons; a few shall be mentioned here. On the one hand, if available, adjusted Maximum Likelihood (ML) estimates are sometimes seen as a gold

standard (Enders and Bandalos, 2001; Raghunathan, 2004). On the other hand imputation is said to be more practical (Raghunathan, 2004) and to be, in some situations, more robust against model violations than adjusted ML estimates (de Leeuw et al., 2003; Wang and Hall, 2010; el Messlaki et al., 2010) (but not in all cases - see He and Raghunathan 2009). Dropping coarse observations is only valid in very special settings (cf. e.g. Enders and Bandalos 2001; Raghunathan 2004).

This thesis will focus on imputation, a prominent form of data modification.

2.3 Imputation

In an early contribution to the handling of missing data Yates (1933) proposed the idea to replace missing values by their least square prediction $x_i \cdot \widehat{\beta}$, later known as *Yates' Method* (Little and Rubin, 2002). Similar but more general is *Buck's Method* proposed by Buck (1960): for each missing value pattern across p variables, a (multivariate) regression equation is formulated and subsequently used to replace missing values. A milestone for contemporary imputation literature was Rubin (1976), where different types of missing mechanisms have been introduced. Using the notation of Little and Rubin (2002), with small changes, the basics of imputation are the following:

- Let X be a $n \times p$ data set with elements x_{ij} being the observation in row i and column j.
- Let x_i be the i-th row vector of X and X_j the j-th column vector. When X_j is the of special interest (for example when selected as target variable in a regression model), it shall be noted as Y with elements y_i.
- X can be split up into its observed part X_{obs} and its missing part X_{mis} so that $X = (X_{obs}, X_{mis})$.
- Let R be a response indicator with elements r_{ij} being 1 if x_{ij} is observed and zero other wise.
- The missing process can be described by a function $f(R|X, \psi)$ where ψ are unknown parameters governing the missing process.
- Based on the factorization $f(R, X, \psi) = f(R|X, \psi) \cdot f(X, \psi)$, the different missing processes can be defined:
  - Data are called to be *Missing Completely At Random* (MCAR) if $f(R|X, \psi) = f(R|\psi)$ holds for all X, ψ (cf. eq. (1.1) in Little and Rubin 2002)
  - Data are called to be *Missing At Random* MAR if $f(R|X, \psi) = f(R|X_{obs}, \psi)$ holds for all X_{mis}, ψ (cf. eq. (1.2) in Little and Rubin 2002)
  - Data, neither MCAR or MAR are called to be *Missing Not At Random* (MNAR).

- When generalized to coarse data, the definitions are analogue (cf. e.g. Tsiatis 2006, Chapter 7):
  - when the coarsening mechanism is independent of any variables, the data are *Coarsened Completely At Random* (CCAR). The less restrictive *Coarsening At Random* (CAR) requires that the coarsening mechanism can be modeled with the observed variables while *Coarsening Not At Random* is present if the CAR assumption does not hold.
  - X_{obs} and X_{mis} represents the precisely observed and coarsely observed data.
  - R becomes the coarsening indicator.

If MAR or MCAR is given (latter, an assumption that cannot be tested without further information or assumptions, cf. e.g. Wang and Hall 2010), valid inferences for Q, the set of *Complete Data Statistics* (Rubin, 1988) for X and ψ is possible. The full likelihood is: (eq. (6.45) in Little and Rubin 2002)

$$L(Q, \psi | X_{obs}, R) \propto f(X_{obs}, R | Q, \psi) \tag{2.1}$$

From a theoretical point of view, imputation is not necessary to obtain valid inferences for Q. Still, the posterior distribution (eq. (10.8) in Little and Rubin 2002)

$$p(Q|X_{obs}) = \int p(Q|X_{mis}, X_{obs}) \cdot p(X_{mis}|X_{obs}) \, dX_{mis} \tag{2.2}$$

proofed to be useful. This equations shows that it is possible to simulate the posterior distribution of Q. Firstly, coarse values are replaced multiply ($M \geq 2$ times) by "plausible values" $X^{\star}_{mis}$ from the joint posterior distribution $p(X_{mis}|X_{obs})$. Secondly, based on each completed data set, an estimate $\widehat{Q}$ for the Complete Data Statistics from the posterior distribution $p(Q|X_{mis}, X_{obs})$ is calculated. Lastly, the M estimates for $\widehat{Q}$ can be analyzed using Rubin's combining rules (Rubin, 1987). This process, called *Multiple Imputation* (MI), is based on the *(Single) Imputation* proposed by Rubin (1978), where missing values are imputed only once. The expansion from Single Imputation to Multiple Imputation allowed to incorporate the uncertainty of imputation within the analysis. For the rest of the thesis by "imputation", the Multiple Imputation is meant (and not Single Imputation).

Since this starting point, an ongoing question in the literature about imputation is how a "plausible value" should be derived. Imputation methods shown to be suitable in one situation might fail in another.

A very crucial quality of an imputation method, elaborated by Meng (1994), is *congeniality*. Roughly summarized: the imputation model has to be consistent with the analysis model. This means for example that at least every variable used in the analysis model should be included in the imputation routine and that the assumptions in the imputation model are compatible with those in the analysis model.

From a method used to provide edited data to external users, imputation broadened to an "in-house" tool, used by researchers to handle their missing data problem adequately (Barnard and Meng, 1999).

Also from the starting point of imputing missing continuous values in cross sectional surveys, two major directions of development of imputation methods can be found. The first is data structure including adaptations for other variables types (like binary or categorical data) and adaptations for more complex data formats (e.g. panel studies or hierarchical data). The second direction is to use imputation not only for missing data but in general for coarse data. The idea behind imputing coarse data is the same to the idea of imputing missing data: When the data are CCAR or CAR, firstly, imputation parameters can be drawn from their posterior distribution and secondly, coarse values are replaced by plausible values based on a model combining the observed information and the imputation parameters.

2.3.1 Regression based imputation

Imputation methods can be differentiated in multiple ways. Some methods, (like mean imputation) replace coarse values with deterministic values, while other methods are stochastic, drawing (pseudo) random values following a probability distribution. For regression based imputations with more than one coarsened variable, two different approaches are found in literature: *joint modeling* and *sequential regression*. In the former a joint distribution of all variables in the data set is assumed to be present (e.g. a multivariate normal distribution), and from this distribution all coarse values are replaced in one step by a draw from the joint distribution. Beside the fact that in many settings the class of this joint distribution is unclear, in multilevel imputation it cannot handle missing data in random slopes variables (Enders et al., 2016). In sequential regression imputation, this joint distribution is tried to be approximated by repeatedly imputing every coarsened variable in X_{mis}. For each variable a conditional distribution is assumed. So step 1. "drawing plausible values $X^{\star}_{mis}$ from the joint posterior distribution $p(X_{mis}|X_{obs})$" (see previous section) is done by intermediate steps that are repeated multiple times (e.g. 10 times) consecutively:

1. For θ, the vector of imputation parameters, within the imputation model for the current variable, new values $\theta^{\star}$ are drawn from $p(\theta \mid X_{obs})$, their posterior distribution given the precisely observed other covariates - which might include previously imputed values.

2. Draw replacements for the coarse values in the current variable from the predictive distribution of the coarse data given the precisely observed other covariates and the drawn parameters from the previous intermediate step: $p(X_{j,mis} \mid \theta^{\star},\ X_{obs})$.

For example using a linear regression imputation model, in the first step the regression coefficients $\beta^{\star}$ and the residual variance $(\sigma^{\star})^2$ can be drawn from a normal distribution for $\beta^{\star}$ and a χ^2-distribution for $(\sigma^{\star})^2$. In the second step, coarse observations might replaced by random draws from a $N(X_{mis} \cdot \beta^{\star},\ (\sigma^{\star})^2)$ distribution.
A disadvantage of the sequential regression approach is that convergence is only possible under the existence of the joint distribution. However, Liu et al. (2014) showed that the joint distribution will exist under rather general conditions and Zhu and Raghunathan

(2015) found that even in cases of nonexistence, correctly specified conditional distributions will lead to consistent estimates.

2.3.2 Other common approaches for imputation of missing data

In the course of time many imputation methods have evolved, each having its specific strengths and weaknesses. Beside regressions based imputation methods, which are in focus of this thesis, other common imputation methods shall be sketched briefly. In *(un)conditional mean imputation*, missing values in a variable are replaced by the (un)conditional mean of this variable (cf. e.g. Little and Rubin 2002 for a discussion). *Data Augmentation*, developed by Tanner and Wong (1987), is a mixture of the EM-Algorithm and Multiple Imputation: Missing values are replaced by draws from a regression model that depends on the observed and previously imputed values in this variable. In the first step of *Predictive mean matching*, proposed by Little (1988), observations with missing values are matched to complete observations (through a distance measurement). In the second step, from the pool of complete donor observations values are selected and imputed for the missing values.

2.4 Methodology of imputing interval data

Early uses of imputation to replace interval values by precise values was in the context of survival analysis (Taylor et al., 1986; Muñoz et al., 1989; Taylor et al., 1990; Dorey et al., 1993). Such imputation models, that need to be congenial with the survival analysis model, are out of scope of this thesis, focusing on (generalized) linear mixed models. In the latter context, for the 1983 *Survey of Consumer Finances* (SCF) interval data "were later translated into a single value by coders using a set of rules"(Kennickell, 1996, p. 440). Later, in the 1989 SCF, Kennickell (1991) imputed interval data with an imputation model. It followed a Gibbs sampling approach, imputing missing values in the income variable sequentially and repeatedly, by draws from a conditional normal distributions. As both, refusals and interval answers were considered to be missing, the imputation model was solely based on precise observations. Furthermore the values, imputed for the interval responders were not bound by the intervals given, which Kennickell stated to "be an important line of research for this project in the future"(Kennickell, 1991, p. 6). Later, for the 1995 SCF Kennickell (1996) used the bounds of interval data as truncation points of the conditional normal distributions.

In the meantime, Heeringa (1993) imputed interval and missing data in the *Health and Retirement Survey* (HRS) following the *General Location Model*. The model consists of two steps. First a multinomial model is fit to estimate an interval category the nonresponder might fall into (e.g. [5000; 9999]). In the second step, for each interval category a separate imputation model is estimated based on the precise observation falling into this category. These models are then used to impute precise values for missing and interval responses. The authors list several statistical and practical problems (like when multiple categorical

variables are present, sparse cells or skewness within open ended categories).

Bhat (1994) proposed a method for missing values in ordered income categories, where the income distribution and the response probabilities are modeled jointly, based on a selection modeling approach.

Raghunathan et al. (2001) described a general sequential regression approach for interval data with draws from the truncated normal distribution. Schenker et al. (2006) imputed values by draws from the truncated predictive distribution. Royston (2007) implemented an imputation model for interval data for *Stata*. He assumes that the actual values behind the intervals originate from a normal distribution. From the posterior predictive distribution of the distribution's parameters, imputation parameters are drawn. These imputation parameters and the individual bounds from the interval observed, are then used to sample imputation values from a truncated normal distribution.

Formally written: suppose for the independent and identically distributed (iid) true values y_i from a variable Y, an indicator I_i indicates whether the value is observed precisely ($I_i = 0$) or in intervals ($I_i = 1$). For interval observations, instead of the true y_i an interval with lower bound $\underline{y_i}$ and upper bound $\overline{y_i}$ is observed. Missing values are handled as interval data with bounds $\underline{y_i} = -\infty$ and $\overline{y_i} = \infty$. The likelihood for the regression imputation parameters $\omega = \{\beta, \sigma^2\}$, based on Y and covariates X is:

$$L(\beta, \sigma^2|Y, X) = \prod_{i=1}^{n} \left((1 - I_i) \cdot f(y_i, x_i \cdot \beta, \sigma^2) + I_i \cdot \left[F(\overline{y_i}, x_i \cdot \beta, \sigma^2) - F(\underline{y_i}, x_i \cdot \beta, \sigma^2)\right]\right) \quad (2.3)$$

with f being the density and F the cumulative distribution function of a normal distribution with mean $x_i \cdot \beta$ and variance σ^2.

Maximizing this likelihood yields to maximum likelihood estimates $\widehat{\omega}$ for ω. Let $J(\widehat{\omega})$ be the negative inverse of the Hessian matrix of the log-likelihood of $\widehat{\omega}$. These parameters are used to draw actual imputation parameters $\omega^\star = \{\beta^\star, (\sigma^\star)^2\}$ from a $N(\widehat{\omega},\ J(\widehat{\omega}))$ distribution which approximate draws from $f(\omega|Y,\ X)$ with assumed flat priors for all parameters. As a last step, each interval value in Y is replaced by a draw from a truncated normal distribution with mean $x_i \cdot \beta^\star$, variance $(\sigma^\star)^2$ and bounds $\underline{y_i}$ and $\overline{y_i}$.

2.5 Methodology of imputing rounded data

A very early contribution to literature about rounding (sometimes called *binning*) was Sheppard (1898), where moments of a normal distribution under rounding were reviewed. In the next decades, handling of rounding in the data was treated by adapting the analysis model (see e.g. Hanisch 2005 or Schneeweiss et al. 2010 for an extensive review). Later, Heitjan and Rubin (1990) proposed the idea to use imputation to deal with rounded values in reported ages. (Presumable) rounded values are replaced by values from the assumed truncated normal distribution. Schenker et al. (2006) used imputation for missing and interval answers in income questions without considering rounded responses. Similar approaches were later used for various topics (Wang and Heitjan, 2008; van der Laan and Kuijvenhoven, 2011; Zinn and Würbach, 2016).

Here the method from Heitjan and Rubin (1990) is delineated concisely: The target variable Y given some covariates X is assumed to be normally distributed:

$$Y|X \sim N\left(X \cdot \beta, \sigma^2\right) \tag{2.4}$$

For a value y_i, in general k rounding degrees K, which are whole positive numbers, are possible. For example K can be an element of the set of possible rounding degrees $\{1,\ 10,\ 100,\ 1000\}$, with $k = 4$. Which rounding degree respondent i actually "picked" is assumed to be driven by a value g_i of a latent variable G. The more thresholds $\kappa = \{\kappa_1,\ \kappa_2,\ \ldots,\ \kappa_{k-1}\}$ the latent variable g_i exceeds, the higher the rounding degree K of respondent i.

G is assumed to be, conditioned on Y and some other variables V, normally distributed:

$$G|Y, V \sim N\left(\gamma_0 + Y \cdot \gamma_1 + V \cdot \gamma_2,\ \tau^2\right) \tag{2.5}$$

with some regression coefficients $\gamma_0,\ \gamma_1,\ \gamma_2$. Under the assumptions of Y and V being independent for given X and G, and X being independent given V, the distribution for Y and G, given X and V, is a bivariate normal: distribution

$$Y, G|X, V \sim N\left(\begin{pmatrix} X \cdot \beta \\ \gamma_0 + X \cdot \beta \cdot \gamma_1 + V \cdot \gamma_2 \end{pmatrix},\ \begin{pmatrix} \sigma^2 & \gamma_1 \cdot \sigma^2 \\ \gamma_1 \cdot \sigma^2 & \tau^2 + \gamma_1 + \gamma_1^2 \cdot \sigma^2 \end{pmatrix}\right) \tag{2.6}$$

In contribution 1 (Drechsler et al., 2015) G is modeled based on a cumulative logit model with some covariates.

The set of parameters to be estimated is given by $\Omega = (\beta, \sigma^2, \gamma_1, \gamma_2, \kappa_1, \ldots, \kappa_{k-1})$. Note: γ_0 was fixed at 0 and τ^2 at 1 to keep the model identifiable.

For each individual i, let s_i denote the possibly rounded value which is observed and might differ from the real value y_i. $S = (s_1, \ldots, s_n)$ is the vector of actually observed values. The likelihood function for Ω given S and covariates X and V (assuming independent observations) is

$$\begin{aligned} L(\Omega|S, X, V) &= \prod_{i=1}^{n} f(s_i|x_i, v_i, \Omega) \\ &\propto \prod_{i=1}^{n} \iint_{A(s_i)} f(g, y|x_i, v_i, \Omega) dy dg, \end{aligned} \tag{2.7}$$

where $A(s_i)$ is the set of (g_i, y_i) that are consistent with an observed s_i. This is when the assumed rounding mechanism $K_i \cdot \lfloor y_i/K_i + 1/2 \rfloor$ matches s_i, with g_i and κ leading to the estimated K_i. For example when $s_i = 1500$, a proposed $y_i = 1668$ and proposed rounding degree $K_i = 500$ match. Whereas for the same rounding degree $y_i = 1790$ would not match.

Given these assumptions, the imputation of heaped values is done in three steps:

1. maximizing the likelihood from equation (2.7) leading to maximum likelihood estimates $\widehat{\Omega}$ and $J(\widehat{\Omega})$, the negative inverse of the Hessian matrix of the likelihood with $\widehat{\Omega}$ plugged in.

2. Draw imputation parameters $\Omega^\star \sim N(\widehat{\Psi}, J(\widehat{\Omega}))$.

3. For each rounded observation i, draw proposal values for (g_i, y_i) from the bivariate normal distribution of equation (2.6) using imputation parameters $\Omega^\star$.

4. If the proposal values for (g_i, y_i) fit together with observed s_i, the proposal for y_i is used as an imputation value, otherwise, they new proposal values are drawn.

For the sake of completeness, two other works on imputing variables affected by heaping shall be mentioned. First: Marcus et al. (2013) modeled heaped data following a generalized beta of the second kind function. Second: the function `correctHeaps` from the `R` package `simPop` by Templ et al. (2017) use imputation to unheap data (by either a log-normal, normal or uniform distribution) rounded to multiples of 5 or 10. Both approaches do not use further information from covariates, what limits the application to settings of Coarsening Completely At Random.

2.6 Contribution to literature: the simultaneous imputation of missing, heaped and interval data

2.6.1 Extension of the likelihood: adding responses in intervals

To the best of the author's knowledge, no article covered the simultaneous imputation of missing, heaped and interval data. Contribution 1 (Drechsler et al., 2015) closed this gap. The approach by Heitjan and Rubin (1990) for imputing heaped data was expanded to additionally impute intervals and missing values together with heaped data, based on a common likelihood.

The likelihoods for interval data (eq. (2.3)) and rounded data (eq. (2.7)) are combined. The likelihood to consider all types of coarse data mentioned is:

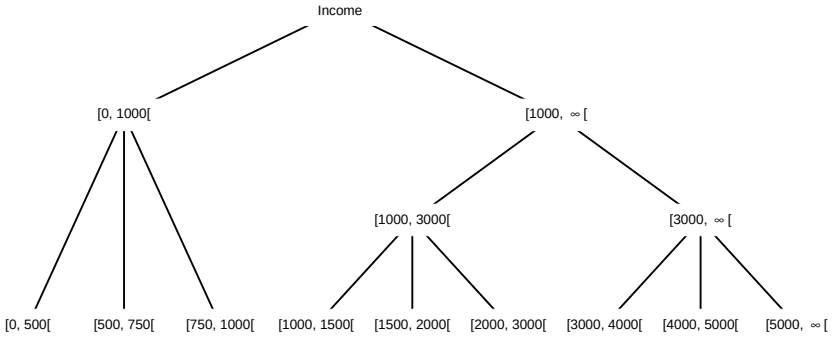
$$L(\Omega|S,X,V) \propto \prod_{i=1}^{n} \Bigg\{ (1-I_i) \cdot \Big[\iint\limits_{A(s_i)} f(g,y|x_i,v_i,\Omega) dy dg \Big] + I_i \cdot \Big[F(\overline{y_i}, x_i \cdot \beta, \sigma^2) - F(\underline{y_i}, x_i \cdot \beta, \sigma^2) \Big] \Bigg\} \tag{2.8}$$

with F being the cumulative normal distribution function and f its density; the indicator function I_i is 1 if observation i is observed in an interval and 0 if not. Missing data are not considered to be rounded, but are seen as an interval observation with bounds $-\infty$ and $+\infty$.

2.6.2 Real data application

In contribution 1 (Drechsler et al., 2015) the approach for simultaneous imputation of missing values, rounded values and interval data was applied to the first six waves of the

Figure 2.3: Copy of Figure 1 from Drechsler et al. (2015), p. 60, showing the questiontree used data collection process for the PASS data



German Panel Study Labor Market and Social Security (PASS) data, starting with the year 2006. Initial nonresponders to the monthly income question have been asked if they are willing to provide an interval of their income starting with "above or below 1000 euros?". If they provide an answer, it is tried to narrow down the intervals a bit. The process is sketched in Figure 2.3. At each step, the respondent can chose to stop, leading to 13 different income intervals. 76.96% of the initial nonresponder provided an interval answer, dropping the nonresponse rate from 4.56% to 1.05%. As already shown in Figure 2.2, a large portion of the respondents reported a net income divisible by multiples of 100 or 1000. As potential rounding degrees governed by G (cf. eq. (2.5)) the values 1, 5, 10, 50, 100, 500 and 1000 were considered. The authors used a set of 10 covariates (for example the household size or the income from savings) for the income imputation model. The complete data statistics Q of interest was the rate of people in risk of poverty, defined as the fraction of observations having an income below 60% of the median income. This rate was calculated for three different situations: 1) complete cases analysis ignoring the coarseness of the data, using all non-missing values as they are 2) using the un-rounding imputation method by Heitjan and Rubin (1990) ignoring interval and missing data, 3) using the method proposed in contribution 1 (Drechsler et al., 2015) addressing rounded, missing and interval data simultaneously. The results showed no clear trend with respect to the location of the poverty rates based on complete data analysis compared to the imputed data sets. In general the differences between complete case analysis and the imputation methods were larger than the differences within the imputation methods. Still some differences between both imputation methods were noticeable, indicating that for the income information the CCAR assumption does not hold and thus they should be treated adequately.

2.6.3 Model evaluations

The article included a detailed demonstration of model evaluations: For the evaluation of the income model each observation y_i with smallest rounding degree 1 (e.g. 1767) was

imputed 1000 times based on an untruncated $N(x_i \cdot \beta^\star,\ (\sigma^\star)^2)$ distribution, with $\beta^\star$ and $(\sigma^\star)^2$ being the imputation parameters based on the ML-estimates from equation (2.8). On the basis of these 1000 artificial observations, which should approximate the posterior distribution, the empirical $\alpha/2$ and $1-\alpha/2$ quantiles (with $\alpha = 0.01,\ 0.05,\ 0.1$) were calculated. For each observation the relative frequency of being within those bounds, the *coverage rate*, was derived. The averages of the coverage rates were reasonable close to the expected numbers, with one exception where only 93.76% of the observations where within the 99% interval.

The rounding model was evaluated by using the individuals' estimates for the rounding tendency and the true income. In each of the 100 imputations, 100 times those estimates are used to generate artificially heaped data, leading to 10000 data sets. In each data set the relative frequencies of rounding to the degrees 1, 5, 10, 50, 100, 500 and 1000 were calculated and lastly averaged. The calculated frequencies based on those re-rounded data were quite close to the frequencies observed in the actual data, with slightly more individuals who have been rounded to the nearest multiple of 100.

2.7 Conjectures for future research

Three conjectures, beyond the content of contribution 1 (Drechsler and Kiesl, 2016), which are worth future verification/falsification, about the biases in poverty rates shall be given.

1. If the median changes due to rounding, this immediately changes the poverty threshold and thus the poverty rate. Withal this effect is minor to the second effect:

2. If the poverty threshold lies close below (resp. above) a common rounding value, it could happen that a noticeable portion of individuals having a precise value below (resp. above) the threshold round to a value above (resp. below) the threshold. This would result in an underestimation (resp. overestimation) of the poverty rate. For example if the threshold is 1800, all individuals between 1500 and 1800 "choosing" a rounding degree 1000 will round to 2000 and thus will be counted as "rich" even if they should be counted as "poor".

3. Both conjectures combined: A poverty threshold slightly below a popular rounding value will lead to an underestimation of the poverty rate; a threshold above a popular rounding value to an overestimation.

3 Biases in multilevel models after fixed effects imputation

After the first direction of generalization of imputation (coarse data; cf. Chapter 2) this chapter covers the second direction of generalization: the more complex data structures, which was the topic of contribution 2 (Speidel et al., 2018). The chapter starts with a brief summary of linear mixed models. Subsequently two frequently discussed methods for hierarchical imputation are described: the cluster specific fixed effects imputation (also called *dummy imputation*) and the multilevel imputation. After a review of literature evaluating both methods, the major contributions of Speidel et al. (2018) are shown. The chapter ends with an outlook.

Declaration of contributions to the article: The majority of the article was written by the author. The following parts were joint work by the author and Jörg Drechsler: the literature overview, the *Multilevel modeling* and *Imputation models* sections, the appendices A, C and F. The real data application and the overall technical implementation in `R` was done by the author with valuable input by Jörg Drechsler and Joseph Sakshaug regarding methodological issues and code. The evaluation and explanation of the variance ratios in the *Simulation* section was sole work by Jörg Drechsler. Joseph Sakshaug provided thorough improvements in grammar and spelling of the whole article. The improvements, requested by the reviewers, were done by the author.

3.1 Motivation

Many data sets have in common that some sort of clustering (e.g. students in classes or repeated measurements for the same individual) is present (cf. e.g. Hox and Roberts 2011). In such cases the analysis has to be executed cautiously: clustered observations tend to be homogeneous (cf. e.g. Osborne and Neupert 2013, p. 188), which is against the independence assumption of classical linear regression models. For such *hierarchical* or *multilevel data*, special analysis models had to be developed.

Early articles on multilevel modeling are Rao (1959) and Elston and Grizzle (1962), mainly focusing on the implications for variance estimates. Analysis methods have been developed to explicitly model the cluster effects (a detailed history of the multilevel analysis literature can be found for example in Raudenbush and Bryk 2002, chapter 1). In this development linear models have been expanded to linear mixed models.

3.2 Linear mixed models

There are various names for linear mixed models (lmm) and several ways of parametrization. Here the notation from contribution 2 (Speidel et al., 2018) is used. Alike for linear models, for linear mixed models a linear relationship between the target variable Y and some covariates X and Z is assumed. Identical to linear models, the effect of X on Y is modeled by a global fixed effect β. The crucial difference are the C clusters, all having a vector of cluster specific effects u_c, with $c = 1, \ldots, C$ being the index for the clusters. These cluster specific effects model the effect of Z on Y. Often Z is a subset of X (cf. e.g. Snijders 2015 or Resche-Rigon and White 2018). By x_{ic} and z_{ic} the i-th row vector of X and Z in cluster c is denoted; with $i = 1, \ldots, n_c$ being the index for the units belonging to cluster c, and n_c being the number of observations in cluster c. The standard form and assumption of multilevel models is:

$$\begin{aligned} y_{ic} =& x_{ic} \cdot \beta + z_{ic} \cdot u_c + \varepsilon_{ic} \\ u_c \sim& N(0,\ \Sigma), \\ \varepsilon_{ic} \sim& N(0,\ \sigma^2) \end{aligned} \tag{3.1}$$

In literature, models with Z being a column of 1s, are called *random intercepts models*. If further (non-constant) variables are present in Z, the models are called *random slope models*. In more general settings, not considered here, more than two levels or cluster specific error variances are possible.

3.3 Methodology of imputing hierarchical data

Even if missing values in the target variable of hierarchical data sets generally lead to unbiased and efficient estimates (cf. e.g. Carpenter and Kenward 2013), there can be reasons, like balancedness or completeness of the data set (van Buuren, 2011), why such missing values shall be imputed. As mentioned in Section 2.3, when it is decided to impute coarse values, the imputation model should be congenial to the analysis model. For the imputation of hierarchical data this means that they should incorporate the hierarchy of the data adequately. Using the notation from contribution 2 (Speidel et al., 2018), two methods, often discussed in literature, are presented: the cluster specific fixed effects imputation and the multilevel imputation.

3.3.1 Imputation of hierarchical data using dummy variables

A very simple way to incorporate the hierarchy from the data into the standard linear regression model is to include cluster specific dummy variables.

For each cluster c (without the reference cluster) a cluster specific dummy variable I_c is defined being 1 for the observations belonging to cluster c and 0 otherwise. Each covariate in Z (the set of covariates which are assumed to have a cluster specific effect) are then

interacted with each of these cluster specific dummy variables. For individual i in cluster c the model equation is:

$$\begin{aligned} y_{ic} =& \tilde{x}_{ic} \cdot \beta + z_{ic} \cdot u_c + \varepsilon_{ic} \\ \varepsilon_{ic} \sim& N(0,\ \sigma^2) \end{aligned} \tag{3.2}$$

with $\tilde{x}_{ic} = x_{ic} \backslash z_{ic}$ (to keep the model identifiable). The essential difference to the multilevel model equation (3.1) is that u_c is assumed to be fixed, and not a realization from a normal distribution. Following the usual presupposition of uninformative priors, the imputation parameter for the residual variance is drawn from an inverse chi-squared distribution; the other imputation parameters (β and u) are drawn from a normal distribution. In the final step, imputed values are drawn from a normal distribution conditional on the imputation parameters and the values of the variables X and Z. All details are described in the section *Cluster specific fixed effects imputation* of contribution 2 (Speidel et al., 2018).

3.3.2 Imputation of hierarchical data using a multilevel model

An imputation model, truly congenial to the linear mixed model equation (3.1), should follow the same assumptions. This mainly means that the random effects u are assumed to follow a $N(0, \Sigma)$ distribution. Generally, a Gibbs sampler is needed in order to get imputation parameters, as their posterior distribution is not available in closed form (cf. e.g. Gelman and Hill 2006). Under the assumption of flat priors, iterative draws from the following conditional distributions (the comprehensive formulas are written in the *Multilevel imputation* section in contribution 2 - Speidel et al. 2018) allow to approximate the posterior distribution:

- $\beta^\star$, the global fixed effects for the imputation model are drawn from a normal posterior distribution.
- $u_c^\star$, the cluster specific random effects are (multivariate) normally distributed.
- $(\sigma^\star)^2$, the residual variance is based on a χ^2 distribution.

In the final imputation step missing values in Y are imputed by drawing:

$$y_{ic} \sim N\left(x_{ic} \cdot \beta^\star + z_{ic} \cdot u_c^\star,\ (\sigma^\star)^2\right). \tag{3.3}$$

3.4 State of literature

Summarizing the literature review from contribution 2 (Speidel et al., 2018), it can be said that the usage of imputation for hierarchical data sets is relatively new, and so is the related field of research. The essentials regarding the imputation model are:

- Ignoring clustering at all, generally leads to biased results (Reiter et al., 2006; Enders et al., 2016; Zhou et al., 2016).

- Including dummies to incorporate the hierarchy tends to be better than ignoring clustering at all (van Buuren, 2011).

- Dummy imputation leads to conservative inferences for the regression coefficients (Reiter et al., 2006; Andridge, 2011).

- Dummy imputation can lead to biased random effect variances (van Buuren, 2011; Drechsler, 2015).

- Multilevel imputation tends to be a good choice (Taljaard et al., 2008; Drechsler, 2015; Zhou et al., 2016; Lüdtke et al., 2017), even if the normal assumption for the random effects is violated (Yucel and Demirtas, 2010).

- In some situations multilevel imputation can face some shortcomings (Grund et al., 2016).

- In some situations ignoring the clustering or dummy imputation can be sufficient (Zhou et al., 2016; Lüdtke et al., 2017).

3.5 Contribution to literature

The articles mentioned in the literature review section 3.4 are subject to two important limitations: They only consider random intercept models (Reiter et al., 2006; Andridge, 2011; van Buuren, 2011; Drechsler, 2015; Enders et al., 2016; Zhou et al., 2016; Taljaard et al., 2008; Lüdtke et al., 2017) or they only rely on simulation studies (Reiter et al., 2006; van Buuren, 2011; Enders et al., 2016; Zhou et al., 2016; Taljaard et al., 2008; Grund et al., 2016).

Thus, two central aspects of contribution 2 (Speidel et al., 2018) to literature was that

1. findings regarding the comparison of cluster specific fixed effects imputation and multilevel imputation were *analytically* generalized.

2. the findings hold for settings with arbitrarily many cluster specific variables. This includes findings about the impact on random slopes, which have been called for in literature by Drechsler (2015); Lüdtke et al. (2017) or Grund et al. (2016).

Regarding the content of contribution 2 (Drechsler et al., 2015) it was analytically derived why the variance of the random effects in the analysis model is positively biased when a cluster specific fixed effects imputation model, instead of a multilevel imputation model, is used. Formulas reveal the relevant parameters governing the bias. A simulation study was conducted to quantify the bias and an application to real data gave further insights about the differences of the imputation methods.

3.5.1 Analytic comparison

The distribution of the random effects itself cannot be derived in closed form. Therefore the conditional covariance matrices (conditioned on all other parameters in the model, denoted by a dot) of the cluster specific effects are compared. For the fixed effects imputation the covariance matrix of all cluster specific effects u_c in cluster c has the form:

$$Var\left(u_c^{fix}|.\right) = \left(1/\sigma^2 \cdot Z_c^{obs'} Z_c^{obs}\right)^{-1}, \tag{3.4}$$

For the multilevel imputation, the form is:

$$Var\left(u_c^{multi}|.\right) = \left(1/\sigma^2 \cdot Z_c^{obs'} Z_c^{obs} + \Sigma^{-1}\right)^{-1}. \tag{3.5}$$

with Z_c^{obs} being cluster c's observations in the random effect covariates matrix Z having no missing value in Y.

For simplicity it can be defined $A := 1/\sigma^2 \cdot Z_c^{obs'} Z_c^{obs}$ and $B := \Sigma^{-1}$ leading to:

$$\begin{aligned} Var\left(u_c^{fix}|.\right) &= (A)^{-1} \\ Var\left(u_c^{multi}|.\right) &= (A+B)^{-1}. \end{aligned} \tag{3.6}$$

Findings from analytic comparison

It was shown that the fixed effects imputation covariance matrix is Loewner larger than the multilevel imputation matrix:

$$Var\left(u_c^{fix}|.\right) = (A)^{-1} >_L (A+B)^{-1} = Var\left(u_c^{multi}|.\right) \tag{3.7}$$

This means that the variance for any cluster specific effect is larger after fixed effects imputation than after multilevel imputation. Given unbiased variance estimates after multilevel imputation this inevitably yields positively biased random effect variances after fixed effects imputation. The finding $Var(u_c^{fix}|.) >_L Var(u_c^{multi}|.)$ is equivalent to following ellipsoid equations inequality:

$$z' \cdot Var(u_c^{multi}|.)^{-1} \cdot z > z' \cdot Var(u_c^{fix}|.)^{-1} \cdot z \tag{3.8}$$

for any vector $z \neq 0$. Figuratively speaking this means that the ellipsoid of the random effects after *cluster specific fixed effects imputation* always fully encloses the *multilevel imputation*-ellipsoid. When the ellipsoids are interpreted as confidence regions, the set of random effects (inspected jointly) after cluster specific fixed effects imputation will vary more in every possible direction than the random effects after multilevel imputation do.

The next analytical finding is the multiplicative difference between both covariance matrices. It also does not provide the size of the bias, but it reveals in which situations, the fixed effects imputation covariance matrix approaches the multilevel covariance matrix.

The multiplicative difference is:

$$Var\left(u_c^{fix}|.\right) = \left(I + \left[Z_c^{obs'} Z_c^{obs}\right]^{-1} \cdot \sigma^2 \cdot \Sigma^{-1}\right) \cdot Var\left(u_c^{multi}|.\right) \tag{3.9}$$

Equation (3.9) shows that higher random effects variances in Σ will decrease the bias, whereas higher residual variances σ^2 will increase it. The third influential component is Z_c^{obs}. With an increasing number of observations, the bias will decrease. An insight, not found in literature so far, is that different missing functions, within the MAR setting, will influence the shape of Z_c^{obs} and thus the amount of bias. A parameter not appearing in the equation, and thus not influential on the bias, is C, the number of available clusters.

3.5.2 Simulation study

To check the theoretical findings and to quantify the bias, a simulation study was conducted. Artificial data were generated under the assumptions of a linear mixed model (see equation (3.1)) with varying cluster sizes, residual variances and missing functions. The missing function generated about 50% MAR item nonresponse (and in one setting MCAR as a special case). The fixed effects and multilevel imputation models included cluster specific intercepts and cluster specific slopes, so did the multilevel analysis model.

For no key figure examined (not shown here), a major problem after multilevel imputation was found. Regarding the fixed effects imputation, the theoretical findings have been confirmed:

- There is no bias in the point estimates of the global fixed effects regression coefficients, but overestimation of the random effects variances.
- Higher residual variances σ^2 increase the bias.
- Larger cluster sizes reduce the bias.
- The missing function indirectly, by shaping Z_c^{obs}, influences the bias.

For the last point, a general explanation is missing, but for the setup of the simulation in contribution 2 it was explained how the main diagonal elements of $Z_c^{obs'} Z_c^{obs}$ changed in relation to the missing pattern. A further finding without theoretical basis was the close relation between random effects variances and the variances of the global fixed effects regression parameters together with the pattern introduced by the missing functions. As an explanatory approach, it was shown how the different global fixed effects covariates can be written as either random intercepts or random slopes variables. Conjectures about the relation between fixed effects and random effects variance are given in the Section 3.6. A noticeable consequence of the overestimation of fixed effect variances is the increased chances of type II errors (false negative).

Regarding the sizes of the observed relative empirical biases, it can be said, dependent on the setting under consideration, that the median of the relative empirical bias of the random intercepts variance was close to 0% at its best, but above 250% at its worst.

3.5.3 Real data application

Based on two variables from starting cohort 3 of the National Educational Panel Study (NEPS, Blossfeld et al. 2011) the impact of different imputation methods was shown on real data. The final subset consisted of 630 students with 29 missing values in the target variable, giving a missing rate of 4.6%. The estimates based on both imputation methods couldn't be compared to the true values as those are obviously unknown, but were in line with the theoretical findings: the random effects variances estimates are noticeable larger after cluster specific fixed effects imputation than after multilevel imputation. The confidence intervals for the global fixed effects largely overlapped with slightly larger intervals for fixed effects imputation. A noticeable larger variance partition coefficient (a generalization of the intra class correlation) was found after cluster specific fixed effects imputation.

3.6 Outlook

Three distinct but viable fields of further research are:

- The impact of the missing function on parameters of interest: Currently missing values are classified whether they are MCAR, MAR or MNAR. But within MAR or MNAR multiple mechanism are possible having different effects on the disparity between Y and Y^{obs}.
- The close relation between fixed effects variance and random effects variance: A heuristic explanation is that a high random effects variance mean that the random effects vary largely around the fixed effect. In this case it is obvious that the uncertainty (loosely speaking: expressed in the variance) about the location of the fixed effect is much higher.
- A further generalization of the finding: higher levels of clustering, cross classified clusters or missing values in the covariate(s) need to be studied. Missing values in the covariate(s) might be a largish issue as Grund et al. (2016) found (via simulation) some results after multilevel imputation to be biased. With the presumed gold standard to be biased, analytic explanations are needed to elucidate this phenomenon.

4 The R package hmi

Contribution 3 (Speidel et al. 2018a and Speidel et al. 2018b) makes the imputation routines, presented in the previous sections, available to the public and proposes new ideas on the implementation of interval data and the detection of rounding degrees in variables with rounded values. This chapter starts with a short motivation and an overview of existing software for imputation of coarse data, before the contribution 3 is described in more detail.

- Declaration about the contributions for the software (Speidel et al., 2018a): Nearly the entire code was developed and written by the author. The basics about imputing rounded income was coded by Jörg Drechsler. Jörg Drechsler gave valuable comments and ideas about issues rising up during the development of the package. Shahab Jolani helped explaining and functionality and properties of mids objects and gave valuable comments for the implementation.
- Declaration about the contributions for the article (Speidel et al., 2018b): The text was mainly written by the author (especially sections 5-7) and Jörg Drechsler (especially sections 1- 4). Jörg Drechsler and Shahab Jolani improved various passages either directly by modifying them or requesting changes to be done by the author. Shahab Jolani thoroughly checked the document for accurate argumentation, and a good, concise structure.

4.1 Motivation

As shown in the previous sections, the kind of coarse data discussed (missing values in hierarchical data sets, rounded and interval data) are a relatively new field in literature and therefore suitable software to deal with them is sparse. As mentioned by Andridge (2011), Azur et al. (2011) or Speidel et al. (2018), software limitations, respectively the sparseness of suitable software, is likely to be a reason why inadequate models are proposed and used instead.

By the development and deployment of the R-Package `hmi`, Speidel et al. (2018a) contributed to the availability of hierarchical and interval imputation in one of the most popular programming languages (TIOBE software BV 2018; Institute of Electrical and Electronics Engineers 2017).

4.2 Existing software for imputation of coarse data

Existing software for the imputation of interval data:
In `R` for the survival analysis setting, the packages `MIICD` (Delord, 2017) and `icenReg`

(Anderson-Bergman, 2017) provide tools to impute interval data. Since they are based on the survival analysis modeling approach, these methods differ systematically from those presented in Section 2.4. To the best of the author's knowledge, such general methods for imputing interval data are only available in `SAS` (cf. Royston 2007) and `IVEware` (Raghunathan et al., 2016).

Existing software for the imputation of rounded data:
A function designed to correct for rounded data can be found in the `R` package `simPop` (Templ et al., 2017). It is only able to impute multiples of 5 or 10 and does not take into account any covariates, making it only usable when CCAR can be assumed. To the best of the author's knowledge, a general implementation of the method by Heitjan and Rubin (1990) is not present yet.

Existing software for the imputation of missing hierarchical data:
Tools providing multilevel imputation methods can be found in `MPlus` (Asparouhov and Muthén, 2010), `REALCOM-IMPUTE` (Carpenter et al., 2011) and the external `SAS` macro `MMI_IMPUTE` (Mistler, 2013). All these tools rely on the joint modeling approach with the drawbacks mentioned in Section 2.3.1 - mainly the inability to impute missing values in covariates.

In `R`, the packages `mice` (van Buuren and Groothuis-Oudshoorn, 2011), `micemd` (Audigier and Resche-Rigon, 2018), `pan` (Schafer, 2016), and `jomo` (Quartagno and Carpenter, 2018) provide multilevel imputation methods, but with some limitations. `mice` and `micemd` provide only multilevel methods for continuous, binary and integer variables, but not for (ordered) categorical or semi-continuous variables. `pan` and `jomo` also rely on the joint modeling approach. Beside the stand alone software `blimp` (Enders et al., 2017) to the best of the author's knowledge, no software provides imputation methods for generalized linear mixed models[1]. A unique feature of `hmi` is the joint implementation of single and multilevel missing data imputation methods regarding many types of variables (including ordered categorical and semi-continuous variables) and a single level method for simultaneously imputing missing, rounded and interval data. The software is described in a technical report (Speidel et al. 2018b); a draft, congruent with this technical report was submitted to the *Journal of Statistical Software*.

4.3 Content of `hmi`

Contribution 3 (Speidel et al. 2018a,b) closes the gaps mentioned in the previous section by providing and describing the `R`-package `hmi`, which is based on the sequential regression approach. Core features of the package are:

- The implementation of imputation methods for missing data in both, single level and multilevel settings for many types of variables.

[1] With generalized linear mixed models, variables Y from the exponential family can be modeled using a linear predictor $l = X \cdot \beta + Z \cdot u$ and a link function f such that $E(Y) = \mu = f^{-1}(l)$.

- The implementation of the method for simultaneous imputation of rounded, interval and missing data from contribution 1 (Drechsler et al., 2015) including its technical side aspects.
- Convergence checks after multilevel imputation.
- Compatibility with `mice`.

Following Donald Rubin's wish for "simply-used appropriate software for creating multiple imputations and analyzing multiply-imputed data"(Rubin, 1996, p. 486), a further feature of `hmi` is a generally easy handling of the software. For example the user can only provide her or his desired analysis model as `model_formula` (and possibly a `family` argument - for example to specify Poisson analysis model) together with the coarse data to `hmi`. The packages figures out the appropriate imputation methods for each variable, tries to build imputed data sets congenial to the given analysis model and by default returns pooled results following the given analysis model. Several input parameters allow detailed control about the imputation process (not shown here - see Section 5.1 in Speidel et al. 2018b). Functions to facilitate the setup are provided as well.

4.3.1 Imputation methods

After the checks on proper inputs, `hmi` extracts from the analysis model (if given) which variable is the target variable, which variables are fixed effects covariates, which are random effect covariates and what variable is the cluster indicator (if a multilevel analysis model was given). If the user specified the imputation methods for the variables this specification is used, otherwise classification heuristics (not shown here - see Section 5.5 in Speidel et al. 2018b) are used to determine a suitable imputation method. Table 4.1 lists the supported types of variables and the corresponding imputation routines. Following the sequential regression approach, the variables are imputed step by step using the previously imputed variables as covariates - until after a given number of *cycles* convergence is expected to be achieved. The final state of all imputed values at this stage is saved as one imputation run. Then the data are reset and a next imputation runs restarts with the first cycle. `mice` (van Buuren and Groothuis-Oudshoorn, 2011) is used for most single level missing data imputation methods. The multilevel missing data imputation routines are own implementations: `MCMCglmm` (Hadfield, 2010) returns parameter draws for generalized linear mixed models which are used as imputation parameters, subsequently used in combination with the observed covariates to draw plausible values for the missing values.

The method the simultaneous imputation of rounded, missing and interval observations presented in contribution 2 (Drechsler et al., 2015) is an own implementation. A relevant side aspect of providing this method is the technical implementation of interval observations. The packages `survival` (Therneau, 2018) and `linLIR` (Wiencierz, 2012) store interval informations in two separate columns. This approach was not feasible for the `hmi` package; instead, `hmi` stores lower and upper bounds in a character separated by a semicolon (e.g. `"1234.56;3000"`) and declares them as class `interval`. Several functions are

Table 4.1: The supported variable types and corresponding imputation models

Variable type	Imputation model (single level setting)	Imputation model (multilevel setting)
Binary	Logit model	Multilevel logit model
Categorical	CART (*classification and regression trees*)	Multilevel multinomial model
Ordered categorical	Ordered logit model	Multilevel ordered logit model
Continuous	Linear model	Multilevel linear model
Semi-continuous	Combination of logit and linear model cf. Rubin 1987	Combination of multilevel logit and multilevel linear model
Count	Poisson model	Multilevel Poisson model
Rounded continuous (including intervals)	Only single level: Method from contribution 1 (Drechsler et al., 2015)	

delivered with the package to generate interval objects, to perform calculations on interval objects, to plot and tabulate them, to split an interval up into its lower and upper bounds and to switch between the formats of `linLIR` and `hmi`. A further novelty related to the imputation of variables affected by heaping, is a function suggesting potential rounding degrees for a given variable Y: divisors (positive integers, dividing a number without rest) appearing in Y twice more often than expected are rounding degree candidates. Starting with the highest, a candidate is considered to be an actual rounding degree if 1. it is a divisor of at least two other divisors in the data and 2. at least 20% of the data are divisible by this candidate; with observations divisible by a larger actual rounding degree not being counted.

4.3.2 Output

`hmi` uses the `mids` (multiply imputed data set) format introduced by the `mice` package. This format includes for example the original data set, the imputed data and the arithmetic means and variances of the imputed variables across the cycles and imputation runs. The usage of this format in `hmi` allows users, familiar with `mice`, to use all functions they know from the `mice` package on objects created by `hmi`. Two elements, not found in objects created by `mice`, but included in objects created by `hmi` are `gibbs` and `pooling`. The former element contains the chains from the Gibbs-samplers from multilevel imputation

methods. It will be described in Section 4.3.3 in more detail. The latter element contains the outcome from pooling analysis results which are run based on the imputed data, the `model_formula` and `family` given to `hmi` by the user. Details about the pooling can be found in Section 4.3.4

4.3.3 Convergence checks

In evaluating the behavior of the imputation runs, convergence in two areas has to be achieved. First: when imputation parameters are based on a Gibbs-sampler, these samples should have converged to actual draws from the joint posterior distribution. For all variables based on a Gibbs-sampler, these Gibbs-samples are stored as element `gibbs` in the output of `hmi`. One possibility to monitor convergence is Geweke's test on stationarity (Geweke, 1992). Roughly speaking, the test compares the mean of the first 10% and the last 50% of a Gibbs-samples chain. After adequate transformations, under the null hypothesis, the difference between these means is asymptotically normally distributed. The test is implemented in the package `coda` (Plummer et al., 2006) and will be automatically run on every chain in `gibbs`, when the function `chaincheck`, delivered with `hmi`, is called.

Secondly: across the different cycles the distribution of the imputed data should differ only randomly and not systematically to justify the assumption that the draws of the missing data are based on the joint posterior distribution. Plotting the means and variances/standard deviations of the imputed data across the different cycles gives an impression about how stable the distribution is. Also a good *mixing* across the different imputation runs (i.e. similar means and variance for all imputation runs) should be seen. A graphical tool, delivered by `mice`, for monitoring is available by calling `plot` on `mids` objects which will plot the chain means and standard deviations. Section 4.4 includes an example of monitoring convergence.

4.3.4 Pooling

When the imputation is completed, and a `model_formula` was given to `hmi`, the function checks whether it is a single or multilevel model and calls the single or multilevel pooling routines from `mice` using `model_formula` (and `familiy` if given). Finally, `hmi` returns this result in the element `pooling`. Currently the pooling routines from `mice` are limited to fixed effects parameter and their variances; pooling of random effect covariance matrix elements are not supported. For pooling of variance parameters or other complete data statistics Q, `hmi` provides the function `hmi_pool` which takes as arguments a `mids` object and a function, e.g. called `analysis_function`, defined by the user. The only input parameter of this function (`analysis_function`) is a (yet not initialized) completed data set. By *completed data set*, a data set is meant where the coarse observations are replaced by the imputed observations. The body of the function defines the desired analyses. This can be a regression analysis or anything else. Finally the complete data statistics are returned in a vector. Internally, `hmi_pool` now passes every of the `M` completed data sets to the given function (`analysis_function`), stores the resulting vectors of complete data

statistics Q estimates and finally averages the estimates for Q over the M resulting vectors. Therefore, the use of `hmi_pool` is limited to complete data statistics Q which sensibly can be averaged.

4.4 Examples of application

For a short illustration for some of the package's functionalities, the hierarchical data set `Gcsemv` regarding 1905 students in 73 schools is used. Beside school and student ID, the data set contains the variables `gender` (0 = boy, 1 = girl), `coursework` (score of a coursework) and `written` (score in a written questionnaire). Harvey Goldstein and the *Centre for Multilevel Modelling* (CMM) from the University of Bristol granted the data to be included into the `hmi` package. A description of the data can be found in Creswell (1991) and Goldstein (2011).

After loading the package and the data, a function from `hmi` is used to create a list of types in order to check `hmi`'s variable classification:

```
library("hmi"); as.data.frame.interval <- hmi:::as.data.frame.interval
data(Gcsemv)
list_of_types_maker(Gcsemv)
$school
[1] "categorical"
$student
[1] "categorical"
$gender
[1] "binary"
$written
[1] "cont"
$coursework
[1] "cont"
```

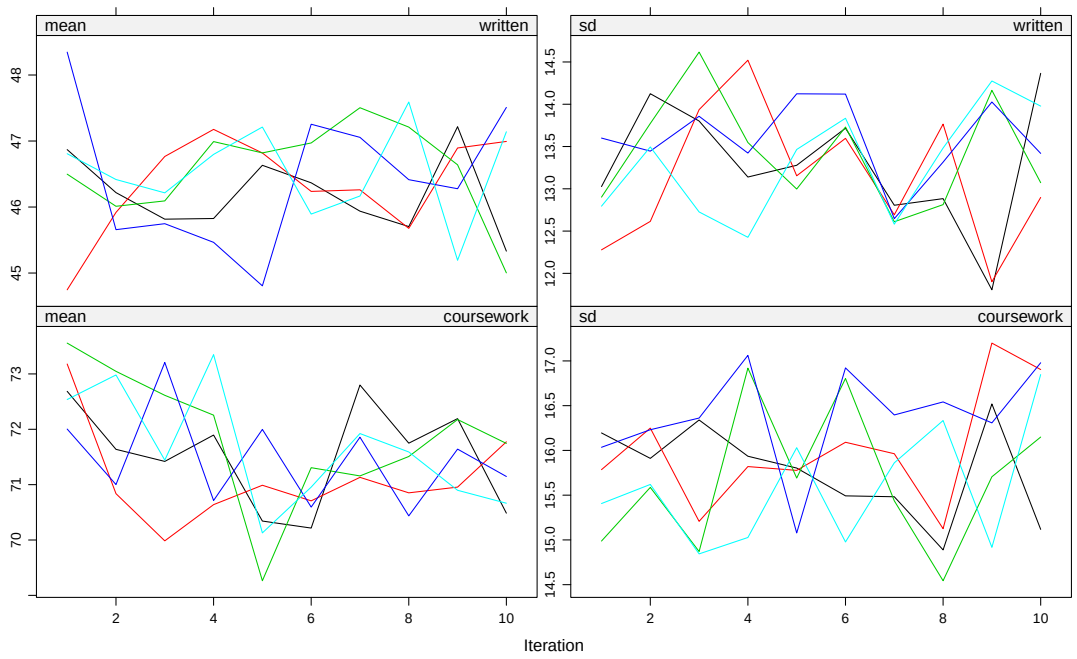
The classification gave correct results, so the wrapper function can be called without specification of the `list_of_types` argument. As `model_formula` a linear mixed model with random intercepts and random slopes for the gender is specified.

```
set.seed(123)
Gcsemv_mids <- hmi(data = Gcsemv, model_formula =
                   written ~ 1 + gender + coursework + (1 + gender|school))
```

When the imputation runs were completed, the performance shall be first examined by checking the Gibbs-sampler chains on convergence:

```
chaincheck(Gcsemv_mids, thin = NULL)
12 out of 695 chains (1.73%) did not pass the convergence test.
For alpha = 0.01, the expected number is 6.95.
```

Figure 4.1: Copy of Figure 1 from Speidel et al. (2018b), p. 38: chain means and standard deviations of imputed example data



The output shows that a larger number than expected did not pass the stationarity test; but the graphical inspection (not reported here) showed no worrisome patterns. As second performance examination, the chain means and standard deviations of the imputed variables across the 10 iterations and 5 imputation runs shall be plotted:

```
library("mice")
plot(Gcsemv_mids, layout = c(2, 2))
```

The resulting Figure 4.1 showed good mixing and no systematical trend.

Regarding the pooling, for brevity, only the pooled results which are delivered by default are shown:

```
summary(Gcsemv_mids$pooling)
              estimate  std.error statistic        df p.value
(Intercept) 21.4285513 1.54661329 13.855145 228.08527       0
gender1     -5.4004356 0.59328192 -9.102647 153.58281       0
coursework   0.4042744 0.01919767 21.058509  64.10292       0
```

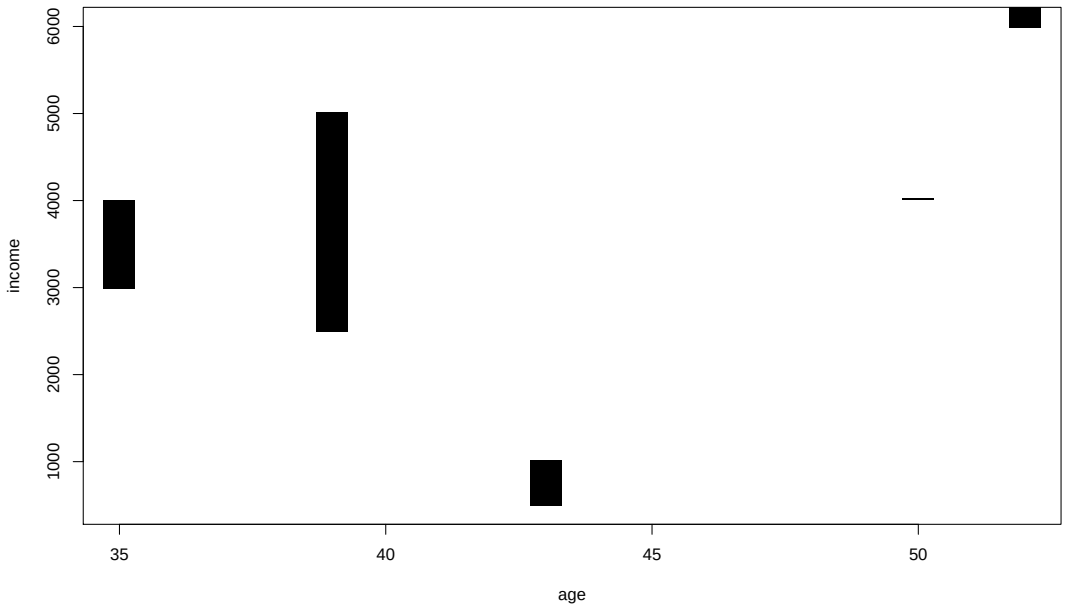
To illustrate some properties and functionalities of `interval` objects developed for this package, five artificial observations with an income as interval variable and a precise age variable are generated.

```
income <- generate_interval(lower = c(3000, 2500, 500, 4017, 6000),
                            upper = c(4000, 5000, 1000, 4017, Inf))
age <- c(35, 39, 43, 50, 52)
df <- data.frame(income, age)
# gives
     income age
1 3000;4000  35
2 2500;5000  39
3  500;1000  43
4 4017;4017  50
5  6000;Inf  52
```

A function for plotting interval objects is implemented in the package. In Figure 4.2 an example is given:

```
hmi:::plot.interval(income ~ age, data = df)
```

Figure 4.2: The income interval variable plotted against the precise age variable



4.5 Outlook

For the future development of `hmi` potential steps and milestones on the methodological side are:

- to further generalize the imputation models by allowing more levels of clustering or heteroscedastic variances
- to include imputation routines for variables on the second level (e.g. the teacher's age or the cluster mean of a covariate)

For a better usability, a major task is the improvement of the run time. A continual task, that already started, since users from universities in Germany and the USA started working with `hmi`, is to collect and incorporate user feedback. Looking at the R Project in general, the promotion (and on the long run, a default implementation) of interval data is overdue.

5 Concluding Remarks and Outlook

In contribution 1 (Drechsler et al., 2015), it was shown how different types of coarse data (missing, rounded and intervals), frequently occurring for example in income variables, can be imputed simultaneously, based on one likelihood. Contribution 2 (Speidel et al., 2018) analytically proofed the superiority of the multilevel imputation over the cluster specific fixed effects imputation with respect to the random effects variance matrices. Contribution 3 (Speidel et al. 2018a and Speidel et al. 2018b) makes both theoretical sound imputation methods for hierarchical data and responses in intervals available to a broad audience. The contributions boosted the reasons to impute missing hierarchical, rounded or interval data in three ways: by rising the awareness of the problem, by presenting solutions and by providing suitable software. Naturally, all contributions have room for improvements: feedback from empirical studies, (further) generalization of the methods involved or improvements in functionality.

Aside these improvements, an interesting field for future research would be an "update2.0" of CCAR, CAR and CNAR.

- For example it could be the case that one variable has a very small influence on the coarsening probabilities in another variable. Even if theoretically the coarsening process changed from CCAR to CAR, practically the impact of methods based on the CCAR assumption would be neglectable. A broad, systematical evaluation of violations of assumptions is missing.

- As shown in contribution 2, within CAR, different missing functions are leading to different results. A framework for different CAR mechanism has to be developed.

- The topic of mixed coarsening mechanism is seldom discussed in literature (exceptions are for example van Buuren et al. 1999 or Gong 2012). Traditionally, the assumptions about the coarsening mechanisms are made for all observations in a variable, but actually it seems plausible that for a first fraction of a variable, some observations are coarsened completely at random, for a second fraction CAR and for a third fraction CNAR is present.

Bibliography

Anderson-Bergman, C. (2017). `icenReg`: Regression models for interval censored data in `R`. *Journal of Statistical Software 81*(12), 1–23.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal 53*(1), 53–74.

Asparouhov, T. and B. Muthén (2010). Multiple imputation with `Mplus`. *MPlus Web Notes*.

Audigier, V. and M. Resche-Rigon (2018). *micemd: Multiple Imputation by Chained Equations with Multilevel Data*. R package version 1.2.0.

Augustin, T. and J. Wolff (2004). A bias analysis of weibull models under heaped data. *Statistical Papers 45*(2), 211–229.

Azur, M. J., E. A. Stuart, C. Frangakis, and P. J. Leaf (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research 20*(1), 40–49.

Babbage, C. (1864). *Passages from the Life of a Philosopher*. London: Longman.

Barnard, J. and X.-L. Meng (1999). Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical Methods in Medical Research 8*(1), 17–36.

Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research 24*(4), 462–487.

Bhat, C. R. (1994). Imputing a continuous income variable from grouped and missing income observations. *Economics Letters 46*(4), 311–319.

Biemer, P. P., E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West (Eds.) (2017). *Total Survey Error in Practice*. New-York: Wiley.

Blossfeld, H.-P., H.-G. Roßbach, and J. von Maurice (Eds.) (2011). *Education as a Lifelong Process - The German National Educational Panel Study (NEPS)*, Volume 14. Zeitschrift für Erziehungswissenschaft.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological) 22*(2), 302–306.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC Press.

Carpenter, J. R., H. Goldstein, and M. G. Kenward (2011). `REALCOM-IMPUTE` software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software 45*(5), 1–14.

Carpenter, J. R. and M. G. Kenward (2013). *Multiple Imputation and its Application*. John Wiley & Sons.

Creswell, M. (1991). A multilevel bivariate model. In R. Prosser, J. Rasbash, and H. Goldstein (Eds.), *Data Analysis with ML3*. London: Institute of Education.

de Jong, R., S. van Buuren, and M. Spiess (2016). Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics - Simulation and Computation 45*(3), 968–985.

de Leeuw, E. D., J. Hox, and M. Huisman (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics 19*(2), 153–176.

Delord, M. (2017). *MIICD: Multiple Imputation for Interval Censored Data*. R package version 2.4.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Denscombe, M. (2009). Item non-response rates: A comparison of online and paper questionnaires. *International Journal of Social Research Methodology 12*(4), 281–291.

Dorey, F. J., R. J. A. Little, and N. Schenker (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine 12*(17), 1589–1603.

Drechsler, J. (2015). Multiple imputation of multilevel missing data – rigor versus simplicity. *Journal of Educational and Behavioral Statistics 40*(1), 69–95.

Drechsler, J. and H. Kiesl (2016). Beat the heap: An imputation strategy for valid inferences from rounded income data. *Journal of Survey Statistics and Methodology 4*(1), 22–42.

Drechsler, J., H. Kiesl, and M. Speidel (2015). MI double feature: Multiple imputation to address nonresponse and rounding errors in income questions. *Austrian Journal of Statistics 44*(2), 59–71.

el Messlaki, F., L. Kuijvenhoven, and M. Moerbeek (2010). Making use of multiple imputation to analyze heaped data. Technical report, Utrecht University.

Elston, R. C. and J. E. Grizzle (1962). Estimation of time-response curves and their confidence bands. *Biometrics 18*(2), 148–159.

Enders, C. K. and D. L. Bandalos (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal 8*(3), 430–457.

Enders, C. K., B. T. Keller, and R. Levy (2017). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*.

Enders, C. K., S. A. Mistler, and B. T. Keller (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods 21*(2), 222–240.

Ferber, R. (1966). Item nonresponse in a consumer survey. *Public Opinion Quarterly 30*(3), 399–415.

Gelman, A. and J. Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Geweke, J. (1992). Evaluating the accuracy of sampling based approaches to calculating posterior moments. In J. B. Bernando, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 169–193. Oxford, UK: Clarendon Press.

Goldstein, H. (2011). *Multilevel Statistical Models* (4 ed.). Chichester (UK): John Wiley & Sons.

Gong, J. (2012). *Modeling longitudinal data with mixed dropout mechanisms using extended pattern mixture model*. Ph. D. thesis, University of Medicine and Dentistry of New Jersey.

Griffin, J. M., A. B. Simon, E. Hulbert, J. Stevenson, J. P. Grill, S. Noorbaloochi, and M. R. Partin (2011). A comparison of small monetary incentives to convert survey nonrespondents: a randomized control trial. *BMC Medical Research Methodology 11*(81), 1–8.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly 70*(5), 646–675.

Groves, R. M. and E. Peytcheva (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *The Public Opinion Quarterly 72*(2), 167–189.

Grund, S., O. Lüdtke, and A. Robitzsch (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. *Behavior Research Methods 48*(2), 640–649.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The `MCMCglmm` `R` package. *Journal of Statistical Software 33*(2), 1–22.

Hanisch, J. U. (2005). Rounded responses to income questions. *Allgemeines Statistisches Archiv 89*(1), 39–48.

He, Y. and T. E. Raghunathan (2009). On the performance of sequential regression multiple imputation methods with non normal error distributions. *Communications in Statistics - Simulation and Computation 38*(4), 856–883.

Heeringa, S. G. (1993). Imputation of item missing data in the health and retirement survey. *Proceedings of the Survey Research Methods Section*, 107–116.

Heitjan, D. F. and D. B. Rubin (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association 85*(410), 304–314.

Heitjan, D. F. and D. B. Rubin (1991). Ignorability and coarse data. *The Annals of Statistics 19*(4), 2244–2253.

Hox, J. and J. K. Roberts (Eds.) (2011). *Handbook of Advanced Multilevel Analysis*. New York: Routledge.

Institute of Electrical and Electronics Engineers (2017, 07). The 2017 top programming languages. `https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages` (retrieved on 2018-06-01).

Juster, F. T. and J. P. Smith (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association 92*(440), 1268–1278.

Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section*, 440–445.

Kennickell, A. B. (1996). Using range techniques with capi in the 1995 survey of consumer finances. *Proceedings of the Survey Research Methods Section*, 440–445.

Kreuter, F. (2013). Facing the nonresponse challenge. *The ANNALS of the American Academy of Political and Social Science 645*(1), 23–35.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T. M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R. M. Groves, and T. E. Raghunathan (2009). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 173*(2), 389–407.

Law, C. G. and R. Brookmeyer (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine 11*(12), 1569–1578.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2 ed.). Hoboken, New Jersey: Wiley-Interscience.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics 6*(3), 287–296.

Liu, J., A. Gelman, J. Hill, Y.-S. Su, and J. Kropko (2014). On the stationary distribution of iterative imputations. *Biometrika 101*(1), 155–173.

Loosveldt, G., J. Pickery, and J. Billiet (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics 18*(4), 545–557.

Lüdtke, O., A. Robitzsch, and S. Grund (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods 22*(1), 141–165.

Lynn, P. (2008). The problem of nonresponse. In E. D. de Leeuw, J. J. Hox, and D. A. Dillman (Eds.), *International Handbook of Survey Methodology*, Chapter 3, pp. 35–55. Lawrence Erlbaum Associates, London: Taylor & Francis.

Marcus, J., R. Siegers, and M. M. Grabka (2013). Preparation of data from the new soep consumption module: Editing, imputation, and smoothing. Data Documentation, DIW 70, Berlin.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science 9*(4), 538–573.

Mistler, S. A. (2013). A `SAS` macro for applying multiple imputation to multilevel data. *Proceedings of the SAS Global Forum*.

Muñoz, A., M.-C. Wang, S. Bass, J. M. G. Taylor, L. A. Kingsley, J. S. Chmiel, B. F. Polk, and The Multicenter AIDS Cohort Study Group (1989). Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type1 (HIV-1) seroconversion in homosexual men. *American Journal of Epidemiology 130*(3), 530–539.

Osborne, J. W. and S. D. Neupert (2013). Multiple imputation of multilevel data. In T. Teo (Ed.), *The Handbook of Advanced Multilevel Analysis*, Chapter 9, pp. 187–198. Rotterdam, The Netherlands: SensePublishers.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). `CODA`: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Quartagno, M. and J. Carpenter (2018). *`jomo`: A Package for Multilevel Joint Modelling Multiple Imputation*. R package version 2.6-1.

Raghunathan, T. E. (2004). What do we do with missing data? some options for analysis of incomplete data. *Annual Review of Public Health 25*(1), 99–117.

Raghunathan, T. E., J. M. Lepkowski, J. van Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*(1), 85–95.

Raghunathan, T. E., P. W. Solenberger, P. A. Berglund, and J. van Hoewyk (2016). *`IVEware`: Imputation and Variance Estimation Software.*

Rao, C. R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika 46*(1/2), 49–58.

Raudenbush, S. W. and A. S. Bryk (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2 ed.). Thousand Oaks: Sage Publications, Inc.

Reiter, J. P., T. E. Raghunathan, and S. K. Kinney (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology 32*(2), 143–150.

Resche-Rigon, M. and I. R. White (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research 27*(6), 1634–1649.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology 7*(4), 532–547.

Royston, P. (2007). Multiple imputation of missing values: Further update of `ice`, with an emphasis on interval censoring. *Stata Journal 7*(4), 445–464.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, 20–34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Hoboken (NJ): John Wiley & Sons.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics*, Volume 3, pp. 395–402. Oxford University Press.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*(434), 473–489.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research 8*(1), 3–15.

Schafer, J. L. (2016). ***`pan`**: Multiple Imputation for Multivariate Panel or Clustered Data.* R package version 1.4.

Schenker, N., T. E. Raghunathan, P.-L. Chiu, D. M. Makuc, G. Zhang, and A. J. Cohen (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association 101*(475), 924–933.

Schneeweiss, H., J. Komlos, and A. S. Ahmad (2010). Symmetric and asymmetric rounding: A review and some new results. *Advances in Statistical Analysis 94*(3), 247–271.

Sheppard, W. F. (1898). On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society s1-29*(1), 353–380.

Shoemaker, P. J., M. Eichholz, and E. A. Skewes (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research 14*(2), 193–201.

Snijders, T. A. B. (2015). The multilevel flavours of multilevel issues for networks. In E. Lazega and T. A. B. Snijders (Eds.), *Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications (Methodos Series)*, pp. 15–46. Springer.

Speidel, M., J. Drechsler, and S. Jolani (2018a). *hmi: Hierarchical Multiple Imputation*. R package version 0.9.13.

Speidel, M., J. Drechsler, and S. Jolani (2018b). R package hmi: A convenient tool for hierarchical multiple imputation and beyond. Technical report, IAB-Discussion Paper 16/2018.

Speidel, M., J. Drechsler, and J. W. Sakshaug (2018). Biases in multilevel analyses caused by cluster-specific fixed-effects imputation. *Behavior Research Methods 50*(5), 1824–1840. Online first 24 August 2017.

Sterne, J. A. C., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ 338*.

Taljaard, M., A. Donner, and N. Klar (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal 50*(3), 329–345.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

Taylor, J. M. G., A. Muñoz, S. M. Bass, A. J. Saah, J. S. Chmiel, and L. A. Kingsley (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statistics in Medicine 9*(5), 505–514.

Taylor, J. M. G., K. Schwartz, and R. Detels (1986). The time from infection with human immunodeficiency virus (HIV) to the onset of AIDS. *The Journal of Infectious Diseases 154*(4), 694–697.

Templ, M., B. Meindl, A. Kowarik, and O. Dupriez (2017). Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software 79*(10), 1–38.

Therneau, T. M. (2018). *A Package for Survival Analysis in S*. version 2.42-3.

TIOBE software BV (2018, 05). TIOBE index for May 2018. https://www.tiobe.com/tiobe-index/ (retrieved on 2018-06-01).

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Trappmann, M., S. Gundert, C. Wenzig, and D. Gebhardt (2010). PASS: a household panel survey for research on unemployment and poverty. *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 130*(4), 609–622.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox and J. K. Roberts (Eds.), *The Handbook of Advanced Multilevel Analysis*, Chapter 10, pp. 173–196. Milton Park, UK: Routledge Academic.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. United States: Taylor & Francis Group.

van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine 18*(6), 681–694.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in R. *Jornal of Statistical Software 45*(3), 1–67.

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

Wang, C. and C. B. Hall (2010). Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine 29*(6), 671–679.

Wang, H. and D. F. Heitjan (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine 27*(19), 3789–3804.

White, I. R. and J. B. Carlin (2010). Bias and efficiency of multiple imputation compared with complete case analysis for missing covariate values. *Statistics in Medicine 29*(28), 2920–2931.

White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine 30*(4), 377–399.

Wiencierz, A. (2012). *linLIR: linear Likelihood-based Imprecise Regression*. R package version 1.1.

Yan, T., R. Curtin, and M. Jans (2010). Trends in income nonresponse over two decades. *Journal of Official Statistics 26*(1), 145–164.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *The Empire Journal of Experimental Agriculture 1*(2), 129–142.

Yucel, R. M. and H. Demirtas (2010). Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics & Data Analysis 54*(3), 790–801.

Zhou, H., M. R. Elliott, and T. E. Raghunathan (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics 32*(1), 231–256.

Zhu, J. and T. E. Raghunathan (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association 110*(511), 1112–1124.

Zinn, S. and A. Würbach (2016). A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics 43*(4), 682–703.

Attached contributions

The contributions 1-3 are attached into this thesis at the following pages:

- Contribution 1: p. 50-62
- Contribution 2: p. 63-79
- Contribution 3: p. 80-136

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

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

Jörg Drechsler
Institute for Employment Research

Hans Kiesl
OTH Regensburg

Matthias Speidel
Institute for Employment Research

Abstract

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

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

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

1. Introduction

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

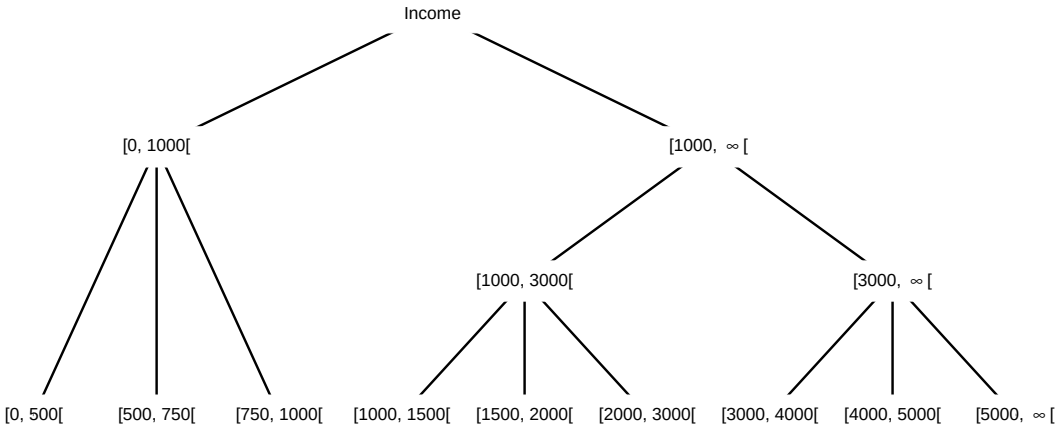


Figure 1: Implied income intervals based on partial income information collected from respondents unwilling to provide their exact income.

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

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

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

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

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

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

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

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

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

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

2. Strategies to adjust for rounding errors

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. Otherwise draw again.

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

3. Extensions for (partial) nonresponse

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

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

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

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

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

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

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

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

Table 2: Covariates included in the income model.

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

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

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

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

4.1. Evaluation of the model assumptions

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

The income model

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

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

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

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

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

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

The rounding behaviour model

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

As expected the percentages differ substantially between the observed income and the unrounded income. Most of the values (80.05%) in the unrounded data (second row in the table) are only divisible by one and the percentages decrease quickly as the rounding base increases (note that we assume that values in the unrounded data are always rounded to the nearest euro). This is different for the observed data (first row). Only 14.94% of the data are only divisible by 1 and 37.34% of the records have a maximum rounding base of 100. The divisibility of the re-rounded data (third row) is reasonably close to the observed data. Again, most records are in the category with a maximum rounding base of 100, although the percentage of records that fall into this category is slightly overestimated (45.79%). This overestimation leads to a slight underestimation of the percentage of records that are only divisible by one (9.67%). For most of the remaining categories the percentages based on the

re-rounded data are fairly close to the percentages based on the observed data: the difference in percentage points is less than 1.2 for the rounding bases 5, 10, and 500. The percentage of records with maximum rounding bases of 50 and 1,000 differ somewhat more between the observed and the re-rounded data (1.75 and 4.12 percentage points respectively). Overall the results indicate a reasonable fit of the rounding behaviour model.

4.2. Results

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

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

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

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

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

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

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

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

5. Conclusions and Outlook

Obtaining reliable income information from surveys is notoriously difficult. Income is considered sensitive information and survey respondents often find it difficult to remember their exact income. In this paper we suggested a strategy to address two common potential sources of bias: nonresponse and rounding. Our multiple imputation approach tackles both problems simultaneously and provides a simple tool to incorporate interval information when making inference based on the collected data. The application to the PASS survey showed that adjusting for these two factors can have a direct impact on politically important measures such

as the poverty rate. We found that rounding has a higher impact on the results than nonresponse at least for our study. The changes in the poverty rates that we found in our empirical evaluation are modest although an increase of the poverty rate by 1.4% as observed for wave 3 of the PASS survey would likely cause some political discussions. We believe that the main reason for the relatively small changes lies in the robustness of the poverty measure which is based on the median of the income distribution. It would be an interesting area of future research to evaluate the impacts on less robust measures such as the income quintile share ratio (see, for example, Eurostat (2014b)) which computes the ratio of the 80% and the 20% quantile of the income distribution as a measure of income inequality.

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Affiliation:

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

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
April 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-09-30
Accepted: 2014-11-21

Behav Res
DOI 10.3758/s13428-017-0951-1

Biases in multilevel analyses caused by cluster-specific fixed-effects imputation

Matthias Speidel[1] · **Jörg Drechsler**[1] · **Joseph W. Sakshaug**[1,2]

Abstract When datasets are affected by nonresponse, imputation of the missing values is a viable solution. However, most imputation routines implemented in commonly used statistical software packages do not accommodate multilevel models that are popular in education research and other settings involving clustering of units. A common strategy to take the hierarchical structure of the data into account is to include cluster-specific fixed effects in the imputation model. Still, this ad hoc approach has never been compared analytically to the congenial multilevel imputation in a random slopes setting. In this paper, we evaluate the impact of the cluster-specific fixed-effects imputation model on multilevel inference. We show analytically that the cluster-specific fixed-effects imputation strategy will generally bias inferences obtained from random coefficient models. The bias of random-effects variances and global fixed-effects confidence intervals depends on the cluster size, the relation of within- and between-cluster variance, and the missing data mechanism. We illustrate the negative implications of cluster-specific fixed-effects imputation using simulation studies and an application based on data from the National Educational Panel Study (NEPS) in Germany.

Keywords Hierarchical multiple imputation · Cluster-specific fixed-effects imputation approach · Multilevel imputation approach · Linear mixed model

✉ Matthias Speidel
matthias.speidel@iab.de

[1] Institute for Employment Research, Regensburger Strasse 104, 90478 Nuremberg, Germany

[2] The University of Manchester, Oxford Road, M13 9PL Manchester, UK

Introduction

Missing values are a common problem in survey data, which can lead to bias if the nonresponse is not properly taken into account by the analyst. A widely accepted strategy to deal with this problem is imputation, which is based on the idea that missing values are replaced with plausible values to produce a completed dataset on which standard analysis models can be applied by the analyst with no, or a less severe, nonresponse bias.

A procedure to take the uncertainty from imputation directly into account is *multiple imputation* (MI). With MI, values are not imputed just once, but $M \geq 2$ times. This leads to M datasets that need to be analyzed, each with the same method leading to M estimates of the parameters of interest and their standard errors. The final inference is obtained by using simple procedures to combine the different results (Rubin's combining rules, Rubin 1987). For applications of (multiple) imputation in educational research see, for example, the overview by Peugh and Enders (2004).

From a theoretical perspective, it is essential that the imputation model is congenial to the model used by the analyst to ensure unbiased results based on the imputed data. Broadly speaking, congeniality means that the model specifications of the imputation model and the analysis model are compatible, i.e., they should be based on the same modeling assumptions (see Meng 1994 and Kenward and Carpenter 2007 for more details). For example, if the analyst is interested in explaining the performance of students in a competence test and uses socio-economic status as one of the predictors, but this predictor is not used when imputing missing values in the competence test, the imputation model and the analysis model would be uncongenial. Therefore, an imputation method should always be developed keeping in

Published online: 24 August 2017

mind the assumed analysis model to be carried out on the imputed data.

These considerations also hold for hierarchical datasets. These are datasets in which individual measurements are grouped; for example, students observed within the same class or repeated measurements on the same individual. Such hierarchical datasets might be analyzed using multilevel models (see Goldstein 1987 or O'Connell and McCoach 2008 and the short review provided in the section "Multilevel modeling" of this paper). Thus, to ensure congeniality, multilevel models should also be used at the imputation stage. However, most of the statistical software packages that are commonly used for imputation such as SAS, SPSS, or Stata, do not provide imputation methods explicitly designed for hierarchical data. To our knowledge, the only tools that allow for multilevel imputation models are the external SAS macro `MMI_IMPUTE` developed by Mistler (2013), some multiple imputation routines in MPlus (Asparouhov and Muthén, 2010), the standalone software `REALCOM-IMPUTE` (Carpenter et al., 2011), which also offers interfaces for MLwiN and Stata, and the R packages `mice` (van Buuren et al., 2015), `pan` (Schafer, 2016), and `jomo` (Quartagno & Carpenter, 2016).

However, `mice` is limited to two levels of hierarchy and continuous dependent variables while all other imputation routines rely on the restrictive joint modeling approach. Joint modeling, which assumes a joint density for all variables with missing data, is especially problematic if the model of interest is a random slopes model, since unlike the sequential regression approach implemented in `mice`, the joint modeling approach cannot deal with missing data in the slope variables (Enders et al., 2016, see also Drechsler 2011 for a general discussion of the pros and cons of the joint modeling approach).

Due to the sparseness of suitable software, using cluster-specific fixed-effects imputation has been recommended in the literature (Diaz-Ordaz et al., 2016; Graham, 2009). This approach is carried out by including dummy variables, representing the cluster membership of the observations, into the data (see section "Cluster-specific fixed-effects imputation"). This imputation strategy is also endorsed on the FAQ website for the multiple imputation module in Stata (StataCorp, 2011). Since the cluster-specific fixed-effects approach is easy to implement using standard imputation software, it has been used for the imputation of missing values in hierarchical datasets (see for example, Brown et al., 2009; Clark et al., 2010; Zhou et al., 2016). Research about imputation in hierarchical settings has only been undertaken in recent years with the earliest papers on this topic focusing only on the impacts on global fixed-effects (the regression coefficients) inferences (Reiter et al., 2006; Taljaard et al., 2008; Andridge, 2011). In educational research, it is often the random effects themselves (or derivatives, such as the intra-class correlation) that are of particular interest when measuring the school effect (Lenkeit, 2012; Nye et al., 2004; McCaffrey et al., 2004b). The impacts on random effects were addressed in later papers but the authors either only focused on random intercept models (Drechsler, 2015; Lüdtke et al., 2017; Zhou et al., 2016), or the evaluations were limited to running simulation studies without analytical derivations to identify which factors influence the bias observed in the simulation studies (van Buuren, 2011; Enders et al., 2016; Grund et al., 2016). In random intercept models, it is assumed that within a cluster, the average intercept deviates from the global intercept by a cluster-specific random value. For example, this could mean that in a class the students score on average four points higher on a math test than the average population of students. This is in contrast to a random coefficients model where the effect of a covariate, x on y, randomly deviates from the global effect; for example, if the performance x in a previous test has a higher effect in a class than on average.

To our knowledge, the impact on random effects if fixed-effects models with cluster-specific slopes are used for imputation has not yet been studied analytically, despite the demand for such research (Drechsler, 2015; Lüdtke et al., 2017; Grund et al., 2016). Our paper closes this research gap by comparing cluster-specific fixed-effects imputation and multilevel imputation and generalizing the evaluations to all types of random coefficient models. We derive analytically why the variance of the random effects in the analysis model is positively biased when a cluster-specific fixed-effects imputation model, instead of a multilevel imputation model, is used. Further, we find that beyond the three factors governing this bias that were already identified in Drechsler (2015) (for the special case of random intercept models), the bias also depends on the mean and variance of the observed data (which are governed by the missing data mechanism). We present support for these findings using simulation studies and a real data application.

The remainder of the article is organized as follows: Section "Related research" summarizes the findings from previous studies, highlights their limitations, and describes our contributions to fill these research gaps. Section "Multilevel modeling" summarizes the ideas behind multilevel modeling and introduces the relevant notation. The different imputation methods are described in section "Imputation models". The following section compares the different imputation strategies analytically and derives which factors influence the bias in random effects-based inferences. The theoretical findings are confirmed using extensive simulations in the "Simulation study" section. In the "Real data application" section, we compare the results of the imputation methods on educational research data. Finally, in the "Conclusion" section we provide a summary of our findings with some practical guidance and provide an

outlook for further research on the topic of hierarchical data imputation.

Related research

As mentioned previously, research about imputation in hierarchical settings is relatively sparse. Reiter et al. (2006) illustrated that ignoring clusters in the imputation process can lead to biased analysis results for clustered sampling designs. They also illustrated that including cluster-specific fixed intercepts for each cluster in the imputation model will lead to conservative inferences for the global fixed effects in the analysis model, increasing the chances of type II errors. In substantive research, this could mean that some covariates are found to have no significant effect on the target variable, while in reality there is one, which would have been found if a proper imputation would have been conducted. Taljaard et al. (2008) compared several imputation routines in a cluster randomized trial setting (clustered randomized trials are typically analyzed using multilevel models but sometimes imputed based on a cluster-specific fixed-effects approach). They found that simple imputation routines (such as cluster mean imputation) can be a suitable choice, but is inferior in performance compared to a congenial (multilevel) imputation. Andridge (2011) also focused on cluster randomized trials. She showed analytically and empirically that the MI variance estimator for the global fixed effects will be conservative if cluster-specific fixed-effects imputation models are used. All three papers leave two kinds of research gaps. First, they limited their evaluations to random intercept models; and second, they all dealt with situations in which the random effects are only nuisance parameters. Thus, none of them evaluated the impacts of cluster-specific fixed-effects imputation on random effects inferences. However, as illustrated in the Introduction, these inferences are often of major interest in education research.

The first paper that also evaluated the impacts on random effects inferences is van Buuren (2011). In a simulation study, the author evaluated the consequences of ignoring the hierarchical structure completely or incorporating dummy variables for the clusters in a random intercept model. He found that ignoring the hierarchy in the data causes biases in random effects inferences and even biases the global fixed effects if missing values occur in the explanatory variables. A further finding was that incorporating dummies for the clusters in the imputation model improves the inferences for the global fixed effects but the estimated variances of the random effects can still be biased. Still, this work was limited to random intercepts and did not explain the results analytically.

Recently, several theoretical articles, comparing imputation methods in a multilevel setting, have appeared. Drechsler (2015) theoretically explained the bias found in the simulations of van Buuren (2011) and illustrated that the bias depends on the cluster size, the missing data rate, and the intra-class correlation (ICC), which, in random intercepts models, is the proportion of variance between clusters relative to the total variance. Like van Buuren (2011), he only focused on random intercept models. Lüdtke et al. (2017) again only focused on random intercept models. They compared a single level imputation (which ignores the clustering of the data), a cluster-specific fixed-effects imputation (incorporating cluster-specific intercepts), and a multilevel imputation with respect to the bias in the intra-class correlation. They derived their results analytically and included a simulation study. Generally, they favored the multilevel imputation, but in some settings the single level imputation performed acceptable as well. The dummy imputation could be appropriate when the clusters and ICC are large and when the focus is on the regression coefficients only. The first paper to also consider random slopes was published by Grund et al. (2016). The authors evaluated the performance of two multilevel imputation strategies and listwise deletion under various settings. They found that the multilevel imputation methods worked well, as long as the missing data only occur in the dependent variable. If missings occur in the covariates, then random effects variances would be biased, an issue we will discuss later. The authors did not consider the dummy variable approach as an alternative to the multilevel imputation model.

Enders et al. (2016) mainly compared joint modeling (imputing all variables in one step) and sequential regression (imputing the variables step by step) in a setting of random intercepts and random slopes. Besides these imputation techniques, they also evaluated the performance of single-level imputation and including dummy variables for cluster-specific intercepts, but not cluster-specific slopes. They found that joint modeling and sequential regression produced similar results in random intercepts models. Joint modeling performed better when contextual effects (cluster means, etc.) were incorporated into the analysis model, while the sequential regression approach performed best in random slopes settings. The poor performance of the dummy variable and joint modeling approach in the random slopes context is not surprising since, except for the sequential regression approach, the authors only considered models that ignore the cluster-specific slopes. Finally, Zhou et al. (2016) proposed an approach to impute a binary variable for rare events in a multilevel setting. The idea is to generate synthetic populations and then to draw plausible values for the missing values from the posterior predictive distribution based on these populations. Via simulation based on a random intercept model, they compared their approach with a single-level imputation, an imputation model with intercept dummies for strata and clusters, and a random intercepts

imputation model. Results indicated poor coverage rates for single-level imputation. The fixed- and random- effects imputation models and their approach worked mostly well with some shortcomings, and random slopes were not considered.

To summarize, while all these articles cover imputation strategies for hierarchical data, they are subject to three important limitations: They only consider random intercept models (Reiter et al., 2006; Andridge, 2011; van Buuren, 2011; Drechsler, 2015; Enders et al., 2016; Zhou et al., 2016; Taljaard et al., 2008; Lüdtke et al., 2017), they only rely on simulation studies to evaluate the impact of different imputation approaches (Reiter et al., 2006; van Buuren, 2011; Enders et al., 2016; Zhou et al., 2016; Taljaard et al., 2008), or they do not evaluate the cluster-specific fixed-effects imputation approach as an alternative to the multilevel imputation model (Grund et al., 2016). Our contribution to the literature is that we analytically generalize the findings regarding the cluster-specific fixed-effects imputation compared to the multilevel imputation model by considering a setting with (arbitrarily many) cluster-specific variable dummies. We also show which factors govern the potential bias from cluster-specific fixed-effects imputation.

Multilevel modeling

With hierarchical data, each individual belongs to one of J clusters. Assuming that individuals within the same cluster are relatively homogeneous, it makes sense to extend the standard linear regression model to account for this. For example, school classes can be homogeneous if the school district lies in an area with many pupils from a specific socio-economic group. As the literature has identified, the socio-economic background tends to be influential on many educational issues (American Psychological Association, 2017), and analyses about the students' educational abilities have to take this homogeneity into account. The multilevel model (or *linear mixed model* as it is often referred to in statistics) is an extension of the linear model and has been a common analysis model for hierarchical data for many years (see, for example, Hedeker and Gibbons 1997 or Verbeke and Molenberghs 2009). A multilevel model incorporates cluster level *random effects* in addition to the global *fixed effects* to take the data hierarchy into account. The general multilevel model is given by:

$$\begin{aligned} y_{ij} &= X_{ij}\beta + Z_{ij}\gamma_j + \varepsilon_{ij}, \\ \gamma_j &\sim N(0, \Sigma), \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \end{aligned} \tag{1}$$

where y_{ij} is the value of the target variable Y for individual $i = 1, \ldots, n_j$ in cluster $j = 1, \ldots, J$, with n_j being the size of cluster j. X_{ij} is a $(1 \times P)$ vector containing the variables for which a constant effect across all clusters is assumed (generally this will include a 1-column for the intercept). β is the $(P \times 1)$ vector containing the global fixed effects. Z_{ij} is a $(1 \times K)$ vector containing the variables for which it is assumed that the effects vary between the clusters. Often Z is a subset of X, meaning that a variable can either have only a global fixed effect or both a global fixed effect and random effects, but will never be modeled as having random effects only. γ_j is a $(K \times 1)$ vector containing the cluster-specific random effect(s) for cluster j. They allow the effect(s) of Z to vary between the clusters and are assumed to follow a multivariate normal distribution with zero-mean and covariance matrix Σ. This modeling strategy implicitly assumes that the observed clusters represent a random selection from a larger population of clusters. The assumption is met if 1,000 schools in the U.S. are sampled from the existing 100,000+ schools, but when characteristics are measured on all 50 U.S. states, including random effects for the states, it is not appropriate, as the states are the basic population and not a sample from it. For later use, we define γ to be the $J \times K$ matrix containing all random effects $\gamma = (\gamma_1', \ldots, \gamma_J')'$. Finally, ε_{ij} is the error term and σ_ε^2 its variance.

To give an example in which situation the multilevel modeling approach could be used in educational research, consider the following model that analyzes the relationship between the score in a math test in year 1 and in year 3 of schooling:

$$\begin{aligned} scoreyear3_{ij} &= \beta_0 + scoreyear1_{ij} \cdot \beta_1 \\ &\quad + \gamma_{0j} + scoreyear1_{ij} \cdot \gamma_{1j} + \varepsilon_{ij} \end{aligned} \tag{2}$$

This modeling strategy would imply that there is a global average score β_0 (say 10) that students have in year 3 if their score in year 1 was 0. For each additional point scored in year 1, the expected score in year 3 increases by β_1 (say 0.8) points, on average. Now, for each cluster, these effects are assumed to vary randomly around the global effects. For example, it could be the case that in school 27 the expected average score is higher (say 11.5, implying $\gamma_{0,27} = 1.5$) but the effect of the test in year 1 is lower (say 0.6, implying that $\gamma_{1,27} = -0.2$).

Imputation models

Imputation methods based on the multiple imputation approach generally consist of two steps: First, a set of model parameters is drawn from their posterior distributions given the data. In the second step, missing values are replaced by repeated draws from the specified distribution given the parameters drawn from step one. This section describes these two steps for the two imputation models

to be compared: the *cluster-specific fixed-effects imputation* and the *multilevel imputation* model.

Cluster-specific fixed-effects imputation

The easiest way to extend the standard linear (multiple) imputation procedures to account for the hierarchy in the data is to incorporate individual fixed effects for each cluster. In this case, the parametric model is given by:

$$y_{ij} = X_{ij}\beta + Z_{ij}\gamma_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \tag{3}$$

The only (yet crucial) difference to Eq. 1 is that γ_j is no longer assumed to be a realization from a normal distribution, but rather assumed to be fixed. In practice, this implies that a dummy variable for each cluster is included in the model and each variable in Z is interacted with this dummy. Let $I_j = I(y_{ij} \in cluster_j)$ be the indicator function that equals 1 if y_{ij} belongs to cluster j, and equals zero otherwise. The model to be estimated is given by:

$$\begin{aligned} y_{ij} &= \sum_{p=1}^{P-T} X_{ijp}\beta_p + \sum_{j=1}^{J-1}\sum_{k=1}^{K} Z_{ijk} I_j \gamma_{jk} + \varepsilon_{ij}, \\ \varepsilon_{ij} &\sim N\left(0, \sigma_\varepsilon^2\right), \end{aligned} \tag{4}$$

where $p = 1, \ldots, P$ is the index for the P variables in X and $k = 1, \ldots, K$ is the index for the K variables contained in Z. Without loss of generality, we assume that X is sorted so that those T variables in X, that are also included in Z, are included in the last T columns of X. These variables need to be dropped (in addition to the reference categories for the dummy variables) to keep the model identified.

Since this is a standard linear regression model, with the usual assumption of uninformative priors (Bartlett et al., 2015), the draws for the first step of the imputation come from the following posterior distributions:

$$\begin{aligned} \tilde{\sigma}_\varepsilon^2 &\sim \chi^{-2}\left(n^{obs} - d, \left[n^{obs} - d\right] \cdot \hat{\sigma}_\varepsilon^2\right), \\ \tilde{\delta} &\sim N\left(\hat{\delta}, \left[V^{obs\prime} V^{obs}\right]^{-1} \cdot \tilde{\sigma}_\varepsilon^2\right), \end{aligned} \tag{5}$$

where χ^{-2} is an inverse Chi-squared distribution, n^{obs} is the number of individuals over all clusters for which the outcome Y is observed, and $d = (J-1) \cdot K + P - T$ is the number of coefficients that need to be estimated. $\delta = \{\gamma_{11}, \ldots, \gamma_{(J-1)K}, \beta_1, \ldots, \beta_{P-T}\}$ is the collection of parameters to be estimated and $V = \{Z_1 I_1, \ldots, Z_1 I_{J-1}, Z_2 I_1, \ldots, Z_K I_{J-1}, X_1, \ldots, X_{P-T}\}$ is the matrix of explanatory variables. V^{obs} is the subset of V containing those observations for which the outcome Y is observed. Note that we assume that all explanatory variables are fully observed or that missing values in these variables have been imputed in previous steps, as missing values in an explanatory variable can cause biases in some parameter estimates (Grund et al., 2016). Lastly, $\hat{\sigma_\varepsilon}^2$ and $\hat{\delta}$ are the ordinary least squares estimates for σ_ε^2 and δ.

In the second step, missing values are imputed by randomly drawing values from

$$Y^{imp} \sim N(V^{imp}\tilde{\delta}, \tilde{\sigma}_\varepsilon^2), \tag{6}$$

where Y^{imp} and V^{imp} denote the subset of Y and V for which Y is missing.

Multilevel imputation

Since the posterior distribution of the parameters of the multilevel model is not available in closed form, a Gibbs sampler is required for the first step of the imputation (see for example Gelman and Hill (2006) for details). Assuming uninformative priors, draws from the following conditional models need to be iterated until convergence (for readability we use |. to indicate conditioning on all other parameters and the data at each step of the Gibbs sampler):

The global fixed effects for the imputation model are drawn from the normal posterior distribution:

$$\begin{aligned} &\tilde{\beta}|. \sim N(\beta^\star, \Sigma^\star) \quad \text{with} \\ &\beta^\star = \left(X^{obs\prime} X^{obs}\right)^{-1} X^{obs\prime}\left(y^{obs} - Z^{obs}\tilde{\gamma}\right) \\ &\Sigma^\star = \tilde{\sigma_\varepsilon}^2 \cdot \left(X^{obs\prime} X^{obs}\right)^{-1} \end{aligned} \tag{7}$$

The residual variance is based on the posterior χ^2 distribution with $n^{obs} - 1$ degrees of freedom

$$\tilde{\sigma_\varepsilon}|. \sim \sqrt{\sum_{j=1}^{J}\sum_{i=1}^{n_j^{obs}} \frac{\left(y_{ij}^{obs} - X_{ij}^{obs}\tilde{\beta} - Z_{ij}^{obs}\tilde{\gamma}_j\right)^2}{\chi^2_{n^{obs}-1}(n^{obs} - 1)}} \tag{8}$$

The variance of the random effects is drawn from the posterior Wishart distribution with $J + K$ degrees of freedom

$$\begin{aligned} &\tilde{\Sigma}|. \sim Wishart(\Sigma^\star)^{-1}_{J+K} \quad \text{with} \\ &\Sigma^\star = \left(\tilde{\gamma}'\tilde{\gamma} + S_p\right)^{-1}, \end{aligned} \tag{9}$$

where $S_p = K \cdot \hat{\Sigma}^{obs}$ is the prior for the random effects variance and $\hat{\Sigma}^{obs}$ is the estimated random effects variance based on the observed data.

The cluster-specific random effects are (multivariate) normally distributed

$$\begin{aligned} &\tilde{\gamma}_j|. \sim N(\gamma_j^\star, \Omega_j) \\ &\gamma_j^\star = \left(Z_j^{obs\prime} Z_j^{obs} + \tilde{\sigma_\varepsilon}^2 \cdot \tilde{\Sigma}^{-1}\right)^{-1} Z_j^{obs}\left(y_j^{obs} - X_j^{obs}\tilde{\beta}\right) \\ &\Omega_j = \tilde{\sigma}_\varepsilon^2 \cdot \left(Z_j^{obs\prime} Z_j^{obs} + \tilde{\sigma}_\varepsilon^2 \cdot \tilde{\Sigma}^{-1}\right)^{-1} \end{aligned} \tag{10}$$

In the second step, missing values in Y are imputed by drawing from:

$$y_{ij} \sim N\left(X_{ij}\tilde{\beta} + Z\tilde{\gamma}_j, \tilde{\sigma}_\varepsilon^2\right). \tag{11}$$

Theoretical juxtaposition of the two imputation models

Both imputation models have an important common feature: they allow one to incorporate cluster-specific effects. The main difference is that cluster-specific fixed-effects imputation assumes that the cluster effects are fixed quantities, whereas in multilevel models it is assumed that the cluster effects are random deviations from the global effect and these deviations follow a known distribution.

Including many dummy variables for the cluster-specific fixed-effects imputation can result in a large amount of parameters to be estimated (cf. Enders et al., 2016). On the other hand, one drawback of the multilevel imputation is its computational complexity resulting in relatively long run times and the task to monitor convergence of the imputation runs. It is well known (see, for example, Wooldrige 2010) that both models provide consistent estimates of the global fixed effects in a multilevel analysis model. However, as illustrated by Reiter et al. (2006) and Andridge (2011), the estimated variances of these global fixed-effects estimates will be biased after a cluster-specific fixed-effects imputation.

Furthermore, because the cluster-specific effects are modeled differently within the imputation, we also expect that the inferences of the random effects will be affected in the analysis model. Since the variance components are often of major interest in multilevel modeling, we will focus on the impact on the estimated covariance matrix of the random effects.

Directly quantifying the impact is difficult since the distribution of the random effects cannot be obtained in closed form. Thus, we follow the approach of Drechsler (2015) and compare the covariance matrix of the cluster-specific effects conditioning on all other parameters in the model. Since for the cluster-specific fixed-effects approach the conditional cluster-specific effects in one cluster are independent of the other clusters this conditional covariance matrix can be computed based solely on the information from the cluster. For cluster j, the matrix is given by (see Appendix A for details):

$$\begin{aligned} Var\left(\gamma_j^{fix}|\beta, V^{obs}\right) &= \sigma_\varepsilon^2 \cdot \left(Z_j^{obs'} Z_j^{obs}\right)^{-1} \\ &= \left(1/\sigma_\varepsilon^2 \cdot Z_j^{obs'} Z_j^{obs}\right)^{-1}, \end{aligned} \tag{12}$$

where $\gamma_j^{fix} = \{\gamma_{1j}, \ldots, \gamma_{Kj}\}'$ is the collection of cluster-specific fixed effects, $\beta = \{\beta_1, \ldots, \beta_{P-T}\}$ is the collection of global fixed effects, V^{obs} is the observed data, and Z_j^{obs} is the subset of records in Z_j for which Y is observed, where Z_j contains those variables in cluster j for which cluster-specific effects are assumed (in the example above Z_j is a matrix with a column of 1s for the intercept and the score of students in year 1 from class j). As noted above, the same conditional covariance matrix for the multilevel model is given by (Goldstein 2011 p. 69)

$$\begin{aligned} Var\left(\gamma_j^{multi}|\ldots, V^{obs}\right) &= \sigma_\varepsilon^2 \cdot \left(Z_j^{obs'} Z_j^{obs} + \sigma_\varepsilon^2 \cdot \Sigma^{-1}\right)^{-1} \\ &= \left(1/\sigma_\varepsilon^2 \cdot Z_j^{obs'} Z_j^{obs} + \Sigma^{-1}\right)^{-1}. \end{aligned} \tag{13}$$

The analytic comparison of Eqs. 12 and 13 is the main part of this section and key to this article. In the appendix, these equations are compared in detail regarding their Loewner-ordering (a mathematical concept to compare matrices), their additive and multiplicative difference, and their representations as ellipsoids (a multidimensional generalization of two-dimensional ellipses). Here we want to limit ourselves to the major findings. The first major finding (see Appendix B for details and proofs): $Var(\gamma^{fix}|.)$ is Loewner larger than $Var(\gamma^{multi}|.)$ and therefore the variances of the estimated random effects are always larger for the cluster-specific fixed-effects imputation. Assuming a correctly specified analysis model, this implies that after cluster-specific fixed-effects imputation, the estimated variances on the second level of the multilevel analysis model will always have a positive bias. The second major finding (see Appendix C): The multiplicative difference between the two variances (14) allows one to draw many conclusions regarding the causes of bias induced by the fixed-effects imputation:

$$\begin{aligned} Var\left(\gamma_j^{fix}|.\right) = &\left(I + \left[Z_j^{obs'} Z_j^{obs}\right]^{-1} \cdot \sigma_\varepsilon^2 \cdot \Sigma^{-1}\right) \\ &\cdot Var\left(\gamma_j^{multi}|.\right) \end{aligned} \tag{14}$$

On the one hand, the difference depends on the ratio of the two variance components σ_ε^2 and Σ. Higher random effects variances in Σ will decrease the bias, whereas higher residual variances σ_ε^2 will increase it. Intuitively this makes sense. If the residual variance σ_ε^2 (i.e., the variance on the individual level) is small relative to the cluster level variance Σ, this implies that all the variation is between the clusters and thus the multilevel model coincides with the cluster-specific fixed-effects model. Both imputation models will lead to similar results in this case. However, if the individual level variance is large relative to the cluster level variance, results based on a cluster-specific fixed-effects analysis model will differ from the results obtained from a multilevel analysis model and we would expect to see a similar effect if cluster-specific fixed-effects and multilevel models are used at the imputation stage.

Besides the ratio of the two variance components, the difference also depends on the matrix of explanatory variables in cluster j. Under rather general conditions, the difference decreases with increasing cluster size since the main diagonal elements of $\left(Z_j^{obs'} Z_j^{obs}\right)^{-1}$ decrease as n_j increases (see Appendix D). Again, this is plausible, since the shrinkage effect of the multilevel model generally decreases with increasing cluster size and thus the differences between the two models also decreases with the size of the cluster. An implication that is easily overseen is that the difference will implicitly also depend on the missing data mechanism since Z_j^{obs} only contains those records for which Y is observed. If, for example, the missingness in Y is positively correlated with Z, i.e., the probability for Y to be missing is higher for larger Z, the matrix $Z_j^{obs'} Z_j^{obs}$ will look different than if the missingness is negatively correlated with Z. We will address this issue in the next section. We also note that Eq. 14 reveals that the bias does not depend on the number of available clusters since the number of clusters J does not appear in the equation.

The third major finding: The ellipsoid of the random effects after *cluster-specific fixed-effects imputation* always fully encloses the *multilevel imputation*-ellipsoid (see Appendix E). One interpretation is that the confidence region for the joint distribution of the conditional parameters for $\gamma_j^{fix}|.$ fully encloses the confidence region for $\gamma_j^{multi}|.$ for any significance level α. This allows us to make a more general statement compared to the first finding: the set of random effects (inspected jointly) will vary more in every possible direction (regardless of their covariance) after cluster-specific fixed-effects imputation. Thus, we would generally overestimate the variability on the second level of our multilevel model. This directly implies that the "classical" intra-class correlation ICC $= \sigma_0^2/(\sigma_0^2 + \sigma_\varepsilon^2)$, with σ_0^2 being the variance on the second level, will be positively biased in a random intercepts setting (the fraction increases as σ_0^2 increases while σ_ε^2 remains constant).

Simulation study

To evaluate whether the identified differences between the two models also lead to substantial bias in the inferences obtained from the imputed dataset, we run extensive simulation studies in R (R Core Team, 2016). The simulations (repeated 1000 times) consist of four steps:

1. Data generation
2. Inducement of nonresponse
3. Multiple ($M = 50$) imputation based on both the cluster-specific fixed-effects and multilevel imputation models described above
4. Running a multilevel analysis model on the imputed dataset

In the following, we will describe each step in detail.

Data generation

To limit the number of parameters that need to be evaluated, we assume the model of interest has, besides the random intercepts, just *one* random slope variable. We do not expect any further insights from the inclusion of further random coefficients.

For the simulation, we assume that the analysis model is correctly specified, i.e., the analysis model matches the data generating process. Of course, this assumption is often not met in practice; however, it is moot to discuss potential biases from imputation if the analysis model would already be biased in the absence of any missing data. For simplicity, we only include two explanatory global fixed-effects variables—W_1 varying at the individual level (e.g., the test score in year 1) and W_2 varying at the cluster level (e.g., the teachers age)—in our random coefficients analysis model. These two variables were generated according to the following models:

$$\begin{aligned} W_1 &\sim N(1, 2 \cdot I_n) \\ W_2 &\sim N(3, 1.5 \cdot I_J), \end{aligned} \tag{15}$$

where I_n and I_J denote the identity matrices (a matrix with 1s on the main diagonal and 0s elsewhere) of dimension n and J, where n is the number of individuals and J is the number of clusters. In other words, we have n independent draws from a normal distribution with mean 1 and standard deviation 2 and J independent draws from a normal distribution with mean 3 and standard deviation 1.5. Our random coefficient model is given as:

$$Y = X\beta + Z\gamma + \varepsilon, \quad \text{with} \quad \varepsilon \sim N(0, I_n \cdot \sigma_\varepsilon^2), \tag{16}$$

where $X = \{1, W_1, W_2^*, W_1 \cdot W_2^*\}$, $Z = \{1, W_1\}$, and W_2^* is the cluster level variable W_2 "blown-up" to have the same length as the other variables by repeating each entry j n_j times, where n_j is the cluster size for cluster j and $j = \{1, \ldots, J\}$. So the model has an intercept, two fixed-effects covariates, and their interaction as global fixed-effects variables in the model. Besides the fixed-effects variables, the model contains a random intercept and a random slope variable. The values of the global fixed effects are set to $\beta = \{2, 1, 1.5, -0.3\}$ and the random effects are generated as

$$\begin{aligned} \gamma &\sim N(0, \Sigma) \quad \text{with} \\ \Sigma &= \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.7 & -0.3 \\ -0.3 & 0.8 \end{pmatrix} \end{aligned} \tag{17}$$

We keep the cluster sizes equal for all clusters, but alter them across different simulation settings between 15, 25, and 50. The number of clusters is fixed at 30 and is not altered further as the number of clusters does not affect the bias (see

previous section). We run the simulation for different values of the residual variance σ_ε (1.0, 1.5, and 2.0), allowing us to examine the bias under different intra-class correlations. Furthermore, because the missing data mechanism (described in detail below) influences the bias, the simulation results are presented under five different models for the nonresponse.

The nonresponse model

Step two in the simulation design is the inducement of missing values. In our simulation, the missingness is limited to Y, and the missingness mechanism is modeled based on a logistic function of W_1. Since we identified the missingness mechanism as influential for the amount of bias, we need a model that allows for some flexibility regarding the influence of W_1 on the probability of Y to be missing. We decided to use the following model:

$$P\left(Y_{ij} = \text{NA}|\tilde{W}_{1ij}, s\right) = MR \cdot (1-s) + 2 \cdot MR \cdot s \cdot logit\left(\tilde{W}_{1ij}\right)^{-1} \quad (18)$$

where MR denotes the desired missing rate, which we fix at 0.5. $\tilde{W}_{1ij}$ is the standardized version of W_{1ij}, i.e., $\tilde{W}_{1ij} = (W_{1ij} - \bar{W}_1)/\sqrt{var(W_{1ij})}$. The parameter s governs the influence of W_1 on the probability of Y to be missing. Figure 1 illustrates the missing data probability functions for different settings of s. Using this model has several implications:

- To obtain a valid probability model, the range of s needs to be bounded by $\{max(-1, [1-1/MR]), min(1, [1-MR]/MR)\}$. As we set $MR = 0.5$, s is bounded by $\{-1, 1\}$.
- $s = 0$ implies Missing Completely At Random (MCAR, see Rubin 1976).
- $s > 0$ ($s < 0$) implies a positive (negative) correlation between x and the probability to be missing and thus Missing At Random (MAR, see Rubin 1976).
- Larger values of $|s|$ imply a stronger influence of W_1 on the probability of Y to be missing.
- If $\tilde{W}_1$ is symmetrically distributed around 0, the expected missing rate over all records in a dataset is equal to MR.
- Records with $\tilde{W}_1$ values close to 0 will be missing with a probability equal to MR.
- The record with the smallest (resp. largest) possible W_1 value will have a probability for Y to be missing close to $(1-s) \cdot MR$ (resp. $(1+s) \cdot MR$).

In our simulations, we alter s within $\{-1, 0.5, 0, 0.5, 1\}$ to evaluate the impact of the missing data mechanism. Whenever less than six observed records remain in one of the clusters, the missing data generation is repeated for this cluster to ensure numerical stability.

Parameters of interest

As discussed above, we assume that the analysis model of interest is a random slopes model that is congenial to the data generating process. Point estimates of the global regression parameters β should not be biased by a cluster-specific fixed-effects imputation procedure, so we do not focus on them. Instead, we look at the variances of the global fixed-effects and the random-effects variances σ_0^2 and σ_1^2, often reported in educational research to evaluate how much of the total variance in the outcome variable is explained by the cluster level units. Both imputation methods are programmed using own code following the description in the section "Imputation models". The functions for the multilevel imputation will be incorporated in the R package `hmi` by Speidel et al. (2017) in the future. All parameter estimations for the multilevel analysis model are computed using the function `lmer` from the R-package `lme4` by Bates et al. (2016).

Results of the simulation study

We discuss the impacts on the random effects first before describing the implications for the variances of the fixed effects. We only present results for the cluster-specific fixed-effects imputation. Results for the original data (before values were deleted) and for imputation based on the multilevel model did not show any significant bias and we omit them for brevity. In order to make the differences in the estimations $\hat{\theta}_{run}$, $run = 1, \ldots, 1000$ for $\theta = \{\sigma_0^2, \sigma_1^2\}$ easily comparable, we look at the *empirical relative bias*:

$$\frac{\hat{\theta}_{run} - \theta}{|\theta|}, \theta \neq 0. \quad (19)$$

If, for example, the true value is 0.7 and the estimate 0.71, the empirical relative bias is $(0.71 - 0.7)/|0.7| \approx 0.014$, which is an overestimation of 1.4%. An unbiased method has an empirical relative bias of 0. As the simulations work empirically, not even the estimates on the original data will yield an empirical relative bias of exactly 0. Therefore, a small relative empirical bias is tolerable. As a rough guideline, we refer to Grund et al. (2016). They consider relative biases of ±5% for global fixed effects and ±30% for variance parameters to be noteworthy.

Implications for the random effects

Figures 2 and 3 show the relative empirical biases for the random intercepts and random slopes variances for all combinations of the cluster size, residual variance, and missing

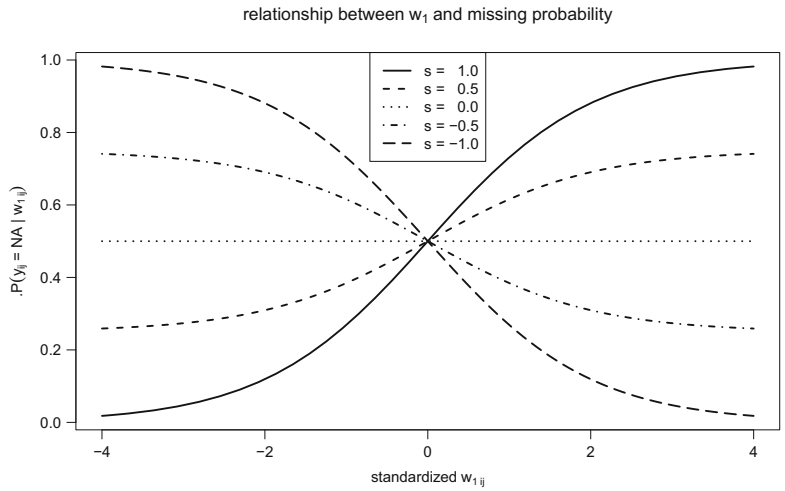


Fig. 1 Illustration of missing data probabilities for different settings of s (=relationship between w_1 and missing data probability) from strong positive ($s = 1.0$) over missing completely at random (MCAR; $s = 0.0$) to strong negative ($s = -1.0$)

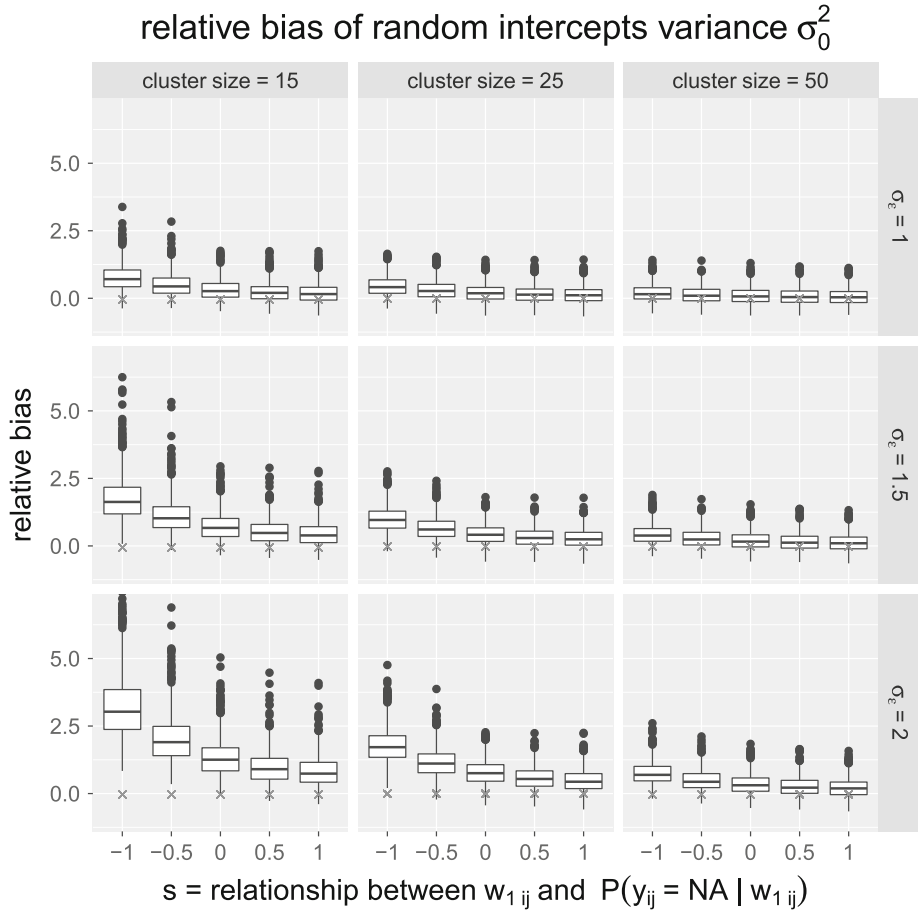


Fig. 2 Relative bias for the estimated variance of the random intercept. The cross marks the median empirical bias of the estimates on the original data as a reference. 10 points (out of 45k) larger than 7 are not shown for readability of the figure

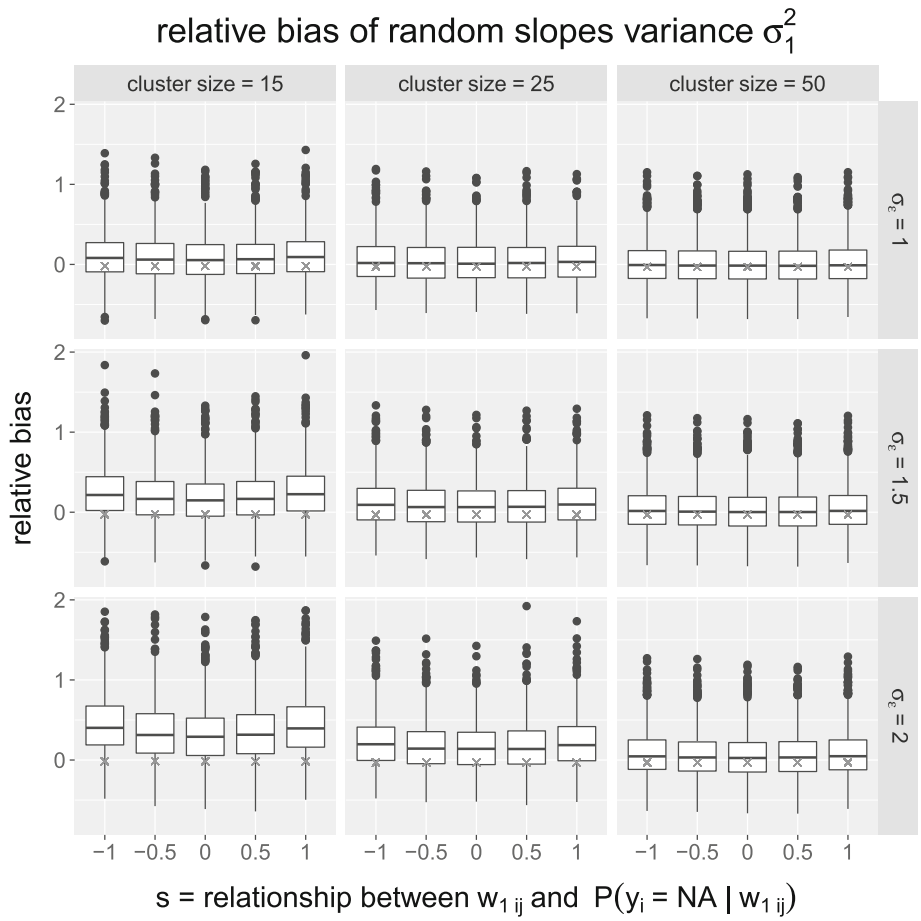


Fig. 3 Relative bias for the estimated variance of the random slope. The cross marks the median empirical bias of the estimates on the original data as a reference. 3 points (out of 45k) larger than 2 are not shown for readability of the figure

data mechanism. A boxplot centered around 0 indicates empirical unbiasedness.

In most settings, the random effects variances are overestimated. In some settings they are (practically) unbiased, but never underestimated. The amount of bias decreases with increasing cluster size, but increases with increasing residual variance. These results are in line with our derivations in the previous section. The bias for the random intercepts is generally larger than the bias for the random slopes (the median relative bias of the random intercept is almost three for $s = -1$, $\sigma_\varepsilon = 2$, and cluster size equal to 15, whereas the median relative bias of the random slopes never exceeds 0.5). We also see that the bias depends on the missing data mechanism. We find a decreasing bias with increasing s for the intercepts and a U-shaped effect for the slopes. It is difficult to explain the process behind these results in general because the bias is governed by distributional properties of Z_{obs}, the random effect variables of those individuals with an observed target variable value (see Eq. 14). We provide some explanations for the observed relationship between the missing data mechanism and the bias for our specific setup in Appendix F.

Implications for the global fixed effects

We do not expect to see any bias in the point estimates of the global fixed effects since both the cluster-specific fixed-effects imputation model and the multilevel imputation model provide unbiased point estimates of the true population parameters. This was confirmed in our simulation study (results not shown for brevity). However, as Reiter et al. (2006) and Andridge (2011) point out, the variances of the global fixed effects should be overestimated if the cluster-specific fixed-effects imputation approach is used. Our simulation study also confirmed this finding. Figure 4 contains variance ratios for all global fixed effects for all

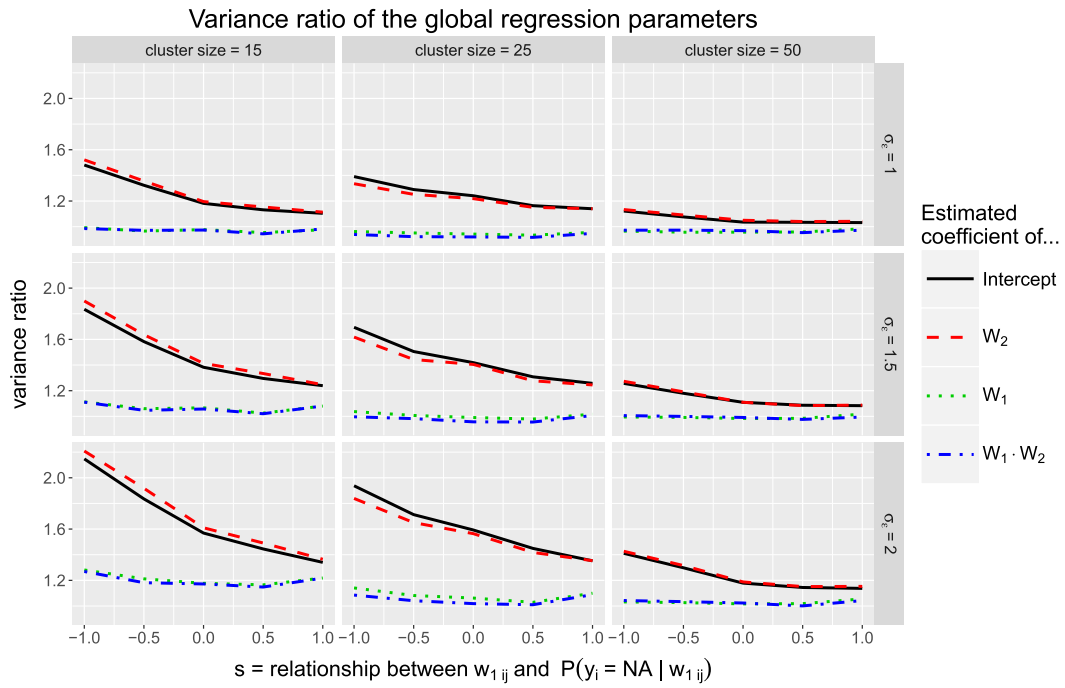


Fig. 4 The variance ratio of the global regression parameters (=median of estimated variances of $\hat{\beta}$ divided by the empirical variance of $\hat{\beta}$)

simulation setups. The variance ratios are computed by dividing the median estimated variance by the true variance of the point estimates across the 1,000 simulation runs. Most of the ratios are greater than 1 indicating that the variance is generally overestimated leading to conservative point estimates and an increased chance of type II errors. The few cases in which the variance ratios are less than 1 seem to be artifacts, since the variance ratios for the original data before deletion (not reported) are even smaller in these cases, indicating a general bias in the analysis procedure. However, beyond confirming results previously discussed in the literature, the figure also illustrates that there is a close relationship between the biases in the random effects variances and the biases in the global fixed-effects variances. As with the random effects, the biases in the variance ratios decrease with increasing cluster size and increase with increasing residual variance. As discussed above, these results are to be expected as the multilevel imputation and cluster-specific fixed-effects imputation become more similar with increasing cluster size and decreasing residual variance. The effect of the nonresponse mechanism s on the bias needs some further explanations. Note that the negative relationship between s and the bias for the regression coefficients of the intercept and W_2 follows the relationship found for the random intercept variances, whereas the U-shaped relationship for the regression coefficients of W_1 and the interaction between W_1 and W_2 mimics the relationship found for the random slopes variances. This can be explained if we note that we can express the random slopes model in Eq. (16) in a different way:

$$\begin{aligned} Y_{ij} &= \alpha_j + \beta_j W_{1ij} + \varepsilon_{ij}, \quad \text{with} \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \\ \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim N\left(\begin{bmatrix} \gamma_0^\alpha + \gamma_1^\alpha W_{2j} \\ \gamma_0^\beta + \gamma_1^\beta W_{2j} \end{bmatrix}, \Sigma \right). \end{aligned} \tag{20}$$

Rewriting the model like this is helpful because it illustrates that there is a relationship between the random intercepts α_j and the coefficients γ_0^α and γ_1^α, and likewise a relationship between the random slopes β_j and γ_0^β and γ_1^β. Relating this notation to the notation in Eq. 16, γ_0^α and γ_1^α are the regression coefficients for the intercept and W_2, while γ_0^β and γ_1^β are the coefficients for W_1 and the interaction between W_1 and W_2. This explains why the biases for the variance of the coefficients of the intercept and W_2 follow a similar pattern as the bias of the random intercept variance. Likewise, we better understand the relationship between the biases for the variance of the coefficients of W_1 and the interaction term, and the bias of the random slopes variance. To our knowledge, this topic has not been addressed in the literature so far and determining the exact relationship between the two effects analytically would be an interesting topic for future research.

Table 1 Estimates of the variance parameters for National Educational Panel Study data

parameter	available case analysis	fixed-effects imp.	multilevel imp.
σ_0^2	0.0228	0.0310	0.0232
σ_1^2	0.0869	0.1167	0.0844
σ_{01}	-0.0441	-0.0596	-0.0449
σ_ε^2	0.1056	0.1058	0.1057

Real data application

In this section, we evaluate whether our theoretical and simulation-based findings are relevant in an applied setting. An appropriate field of research where random effect variances are of particular interest is the evaluation of teacher effectiveness (see e.g., Lenkeit 2012, Nye et al., 2004, or more generally McCaffrey et al., 2004b). Thus, we use data from the Starting Cohort 3 of the National Educational Panel Study (Blossfeld et al., 2011). The NEPS, run by the Leibniz Institute for Educational Trajectories, is an extensive study in Germany that aims to measure the reasons and impacts of educational decisions over the entire life course. To achieve this goal, surveys are conducted in a multi cohort sequence design in which six different cohorts are followed for several years. The cohorts are selected to cover the entire life span starting with an infant cohort, a kindergarten cohort, a cohort of pupils in elementary school, etc. The final cohort is an adult cohort that represents adults aged 23 to 64 by the time of the first interview. The six starting cohorts were recruited between 2009 and 2012 containing more than 60,000 target persons. 5th grade students comprise the Starting Cohort 3. Many performance related items have been administered to these students, including two math competence tests in years 2010 and 2012. Because the aim of this real data application is not to draw conclusions from sophisticated educational analyses, but to show the impact of the imputation methods, we consider a simple model for students' achievements in math competence tests:

$$math7_{ij} = \beta_0 + \gamma_{0j} + math5_{ij} \cdot (\beta_1 + \gamma_{1j}) + \varepsilon_{ij}, \quad (21)$$

where variable $math5$ is the 5th grade test score and $math7$ the test score in grade 7. We conditioned on those students who had the same teacher in both years and only missing values in $math7$. This resulted in $n = 630$ students overall and 29 students having missing values in $math7$ ($\rightarrow$ missing rate of 4.6%). We multiply (M = 50) imputed $math7$ using both methods (the cluster-specific fixed-effects imputation and the multilevel imputation) and estimated the multilevel model from Eq. 21 with `lmer`. To isolate the teacher effect, several additional control variables would normally be included on both levels of the model in practice (see, for example, McCaffrey et al., 2004a). Since including more variables will not provide additional insights regarding the implications of the two imputation strategies, we keep the model simple for illustrative purposes.

Because we only focus on those observations for which missingness is limited to the dependent variable, we can use available case analysis as a benchmark, since available case methods will provide unbiased estimates in this case (assuming the model is correctly specified). The estimations of the random effects variances in Table 1 are in line with the findings from the theoretical section and the simulation study: after the cluster-specific fixed-effects imputation the variance estimates are recognizably higher while the estimates based

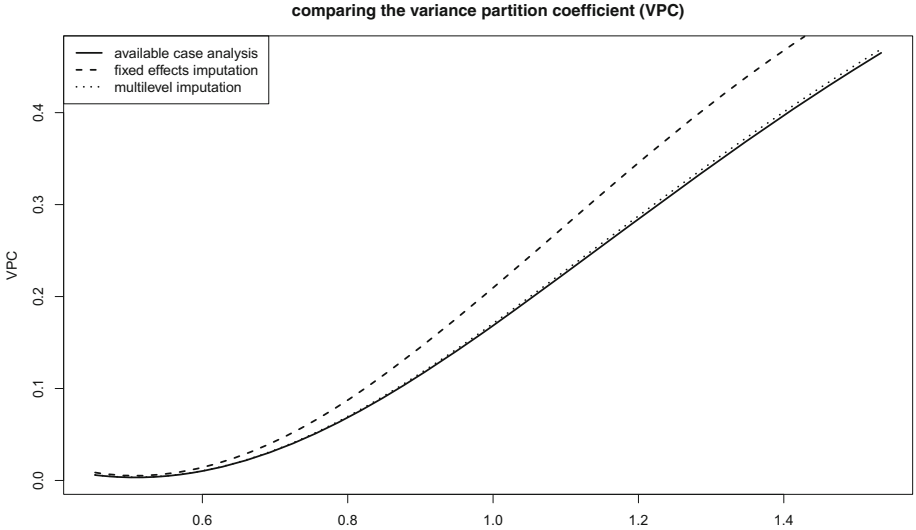


Fig. 5 The variance partition coefficient (VPC) for the range of $x = math5$ in the NEPS data

Table 2 Point estimates and confidence interval (CI) properties of the global fixed effects for National Educational Panel Study data

parameter	point estimate	0.025%-quantile	0.975%-quantile	CI-length
β_0^{aca}	0.3830	0.3123	0.4537	0.1414
β_0^{fix}	0.3808	0.3027	0.4589	0.1562
β_0^{multi}	0.3823	0.3105	0.4541	0.1437
β_1^{aca}	0.2914	0.1563	0.4266	0.2703
β_1^{fix}	0.2967	0.1473	0.4461	0.2988
β_1^{multi}	0.2924	0.1550	0.4298	0.2748

on multilevel imputation are very close to the benchmark values based on available case analysis. The overestimation after cluster-specific fixed-effects imputation is substantial considering that less than 5% of the data were imputed. As mentioned above, a commonly computed measure for the impact of the clustering on the total variance is the intra-class correlation (ICC). For models with more than only random intercepts, the "classical" ICC $= \sigma_0^2/(\sigma_0^2 + \sigma_\varepsilon^2)$ is no longer sufficient to summarize the contribution of the clusters to the total variance. Goldstein et al. (2002) proposed the *variance partition coefficient* (VPC). The VPC is a function of the predictor variable x and the variance components and shows the 'importance' of the clusters for different values of x:

$$vpc = \frac{\sigma_0^2 + 2 \cdot \sigma_{01} \cdot x + \sigma_1^2 \cdot x^2}{\sigma_0^2 + 2 \cdot \sigma_{01} \cdot x + \sigma_1^2 \cdot x^2 + \sigma_\varepsilon} \tag{22}$$

Figure 5 shows the VPCs based on the available cases, the data after the cluster-specific fixed-effects imputation and after the multilevel imputation. One can see that the clusters would be viewed as being more 'important' under cluster-specific fixed-effects imputation than under multilevel imputation.

Results for the global fixed effects are presented in Table 2. The point estimates are almost identical for all methods. Considering the uncertainty of the estimates as expressed by the 95% confidence intervals, the difference between the inferences based on the three different analysis strategies is small, since the confidence intervals overlap to a large extent. The last column of the table shows that the confidence intervals for the global fixed effects are larger for the cluster-specific fixed-effects imputation, which is also in line with theoretical expectations.

Conclusion

Contributing to the discussion about suitable imputation methods for hierarchical data, we present theoretical and empirical evidence to the supposition that the cluster-specific fixed-effects imputation is highly likely to bias variance parameter estimates in a multilevel analysis model. A simulation showed that the bias can be severe. Even though the simulation study was limited to random intercepts and random slopes, the theory holds for any number of random effects variables (starting from only random intercepts models and ending with models where each variable is treated as random). Therefore, we generally advise using multilevel imputation models. Even if there are only random intercepts, including a random slope variable should do no harm ("Consider all coefficients as potentially varying", Gelman and Hill 2006, p. 549).

The high variance in the cluster-specific effects under a cluster-specific fixed-effects imputation also negatively affects the coverage rates of the global fixed effects and increases the probability of false conclusions. In a real data application, we showed that even for a small missing data rate (less than 5%) the results can substantially differ. This is a further reason to use multilevel models for imputation.

A shortcoming of the multilevel imputation is the relatively high runtime. While the cluster-specific fixed-effects imputation took between 0.15 and 4 s (median 0.28) in our simulation settings, the multilevel imputation took between 1.5 and 5.2 min (median 2.2). This can be a severe drawback if many variables need to be imputed in a dataset. A second technical shortcoming of the multilevel imputation is the need to monitor convergence, which is not needed for the cluster-specific fixed-effects imputation as all posterior distributions can be obtained in closed form.

As Eq. 14 showed, both imputation methods will produce similar (if not identical) results in three conditions: large cluster sizes, large differences between the clusters, and small residual variances. We cannot provide general thresholds for these parameters to be "large" or "small," but if the researcher sees one (or better yet, as our simulation showed: more) of these conditions met, s/he might consider using a cluster-specific fixed-effects imputation instead of a multilevel imputation model for convenience. This is especially relevant if there are many variables to impute and not enough time to conduct multilevel imputations.

In this paper, we limited our analysis to missing values in the target variable as a starting point. Still, it would be worthwhile to investigate the impacts of missing values in the covariates. Grund et al. (2016) conducted a simulation study for this scenario and found some results after multilevel imputation to be biased. With the presumed gold standard to be biased, analytic explanations are needed to elucidate this phenomenon. Further research on higher level models, comprising more than two levels, or cross-classified clusters would also be beneficial.

Acknowledgements This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Priority Programme "Education as a Lifelong Process" [SPP 1646] - DR 831/2-2. This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3 - 5th Grade, doi:10.5157/NEPS:SC3:3.1.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Appendix

To simplify all of the following equations, we define $A := 1/\sigma_\varepsilon^2 \cdot Z_j^{obs'} Z_j^{obs}$ and $B := \Sigma^{-1}$. This allows us to write:

$$\begin{aligned} Var\left(\gamma_j^{fix}|.\right) &= (A)^{-1} \\ Var\left(\gamma_j^{multi}|.\right) &= (A+B)^{-1}. \end{aligned} \tag{23}$$

Appendix A

Following Drechsler (2015) we write $\phi' = (\gamma', \beta')$ for the set of cluster-specific and global effects in Eq. 4. $R = (Z_{ind}, X)$ is the $n \times ([J-1]\cdot K + P - T)$ matrix of regression variables, with Z_{ind} being the $n \times ([J-1]\cdot K)$ matrix of intercept and slope dummies. Generally, for linear models it holds that $\phi|\mu, R, \sigma_\varepsilon^2 \sim N(\mu, \Sigma = \sigma_\varepsilon^2 \cdot [R'R]^{-1})$. Let $\mu' = (\mu_1', \mu_2')$ be partitioned so that μ_1 contains the expected values for γ and μ_2 contains the expected values for β. We partition Σ in a similar way so that Σ_{11} consists of a $(J-1)\cdot K \times (J-1)\cdot K$ matrix containing the covariance matrix of the cluster-specific effects. Likewise, Σ_{22} is the $(P-T)\times(P-T)$ dimensional matrix containing the covariance matrix of the global effects. With this partitioning we have

$$\begin{pmatrix} \gamma^{fix}|\mu, R, \sigma_\varepsilon^2 \\ \beta^{fix}|\mu, R, \sigma_\varepsilon^2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \sigma_\varepsilon^2 \cdot \begin{bmatrix} (R'R)_{11} & (R'R)_{12} \\ (R'R)_{21} & (R'R)_{22} \end{bmatrix}^{-1} \right) \tag{24}$$

Since the joint distribution of the effects is multivariate normal it holds that the conditional variance of the cluster-specific effects given the global effects is

$$Var(\gamma^{fix}|\mu, R, \sigma_\varepsilon^2, \beta^{fix}) = \Sigma_{11} - \Sigma_{12}[\Sigma_{22}]^{-1}\Sigma_{21} \tag{25}$$

In order to simplify this equation we use Harville (1997, Corollary 8.5.12 p. 100) which shows that for a nonsingular matrix $1/\sigma_\varepsilon^2 \begin{bmatrix} (R'R)_{11} & (R'R)_{12} \\ (R'R)_{21} & (R'R)_{22} \end{bmatrix}$ and its inverse $\sigma_\varepsilon^2 \cdot \begin{bmatrix} (R'R)_{11} & (R'R)_{12} \\ (R'R)_{21} & (R'R)_{22} \end{bmatrix}^{-1} = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, partitioned in the same way, it holds that $\Sigma_{11} = \sigma_\varepsilon^2 \cdot [(R'R)_{11}]^{-1} + \Sigma_{12}[\Sigma_{22}]^{-1}\Sigma_{21}$. Replacing this expression of Σ_{11} in Eq. 25 yields $Var(\gamma^{fix}|V^{obs}, \beta^{fix}) = \sigma_\varepsilon^2 \cdot [(R'R)_{11}]^{-1}$.

Multiplication rules for block matrices yield $\sigma_\varepsilon^2 \cdot \{([Z_{ind}, X]'[Z_{ind}, X])_{11}\}^{-1} = \sigma_\varepsilon^2 \cdot \{Z_{ind}'Z_{ind}\}^{-1}$. For a cluster j this means that $Var(\gamma_j^{fix}|V^{obs}\beta^{fix}) = \sigma_\varepsilon^2 \cdot (Z_j'Z_j)^{-1}$.

Appendix B

Here we show that the conditional variance of each of the cluster-specific fixed effects in the cluster-specific fixed-effects imputation model is larger than the conditional variance of the corresponding random effect in the multilevel imputation model, i.e., $diag[Var(\gamma^{fix}|.)] > diag[Var(\gamma^{multi}|.)]$. To start our proof, we look at the additive disparity of $Var(\gamma^{fix}|.)$ and $Var(\gamma^{multi}|.)$.

$$\begin{aligned} Var(\gamma_j^{multi}|.) + \Delta &= Var(\gamma_j^{fix}|.) \Leftrightarrow \\ \Delta &= Var(\gamma_j^{fix}|.) - Var(\gamma_j^{multi}|.) \Leftrightarrow \\ \Delta &= A^{-1} - (A+B)^{-1} \end{aligned} \tag{26}$$

To show that $diag[Var(\gamma^{fix}|.)] > diag[Var(\gamma^{multi}|.)]$, we need to show that Δ is positive definite since the main diagonal elements of positive definite matrices are always positive (Harville 1997, corollary 14.2.13 p. 214).

According to the definition of the Loewner order a Hermitian matrix M_2 is Loewner larger than a Hermitian matrix M_1 ($M_2 >_L M_1$) if the difference $M_2 - M_1$ is positive definite. It also holds that if $M_2 >_L M_1$ then $M_1^{-1} >_L M_2^{-1}$ (see e.g. Siotani 1967 eq. 3 p. 246 or Horn and Johnson 1990 theorem 7.7.4 p. 471). So to show that Δ is positive definite we need to show that $A^{-1} >_L (A+B)^{-1}$ or equivalently that $A+B >_L A$. The last statement is trivially fulfilled because the difference $A + B - A = B$ is positive definite (because B is the inverse of the positive covariance matrix Σ and the inverse of a positive definite matrix is also positive definite Harville 1997, corollary 14.2.11 p. 214).

Appendix C

To further understand which factors influence the difference between $Var(\gamma^{fix}|.)$ and $Var(\gamma^{multi}|.)$ it is informative to identify the multiplicative factor H for which it holds that:

$$H \cdot Var\left(\gamma_j^{multi}|.\right) = Var\left(\gamma_j^{fix}|.\right). \tag{27}$$

Simple matrix manipulation reveals that

$$\begin{aligned} H(A+B)^{-1} &= (A)^{-1} \Leftrightarrow \\ H(A+B)^{-1}\cdot(A+B) &= (A)^{-1}(A+B) \\ H &= (A)^{-1}A + (A)^{-1}B \\ H &= I + (A)^{-1}B \\ H &= I + \left(1/\sigma_\varepsilon^2 \cdot Z_j^{obs'} Z_j^{obs}\right)^{-1}\Sigma^{-1} \end{aligned} \tag{28}$$

Appendix D

If Z is a $n_0 \times p$ data matrix then observing additional $n_1 \times p$ data Z_{new} results in the $(n_0 + n_1) \times p$ data matrix $Z^\star = \begin{pmatrix} Z \\ Z_{new} \end{pmatrix}$. According to the multiplication rules for block matrices (see e.g., Harville 1997 section 2.2) it holds that

$$Z^{\star'}Z^\star = \begin{pmatrix} Z' & Z'_{new} \end{pmatrix}\begin{pmatrix} Z \\ Z_{new} \end{pmatrix} = Z'Z + Z'_{new}Z_{new}. \tag{29}$$

Given the results in Appendix B it follows that $\left(Z^{\star'}Z^\star\right)^{-1} <_L \left(Z'Z\right)^{-1}$. This shows that $\left(Z_j^{obs'}Z_j^{obs}\right)^{-1}$ decreases as n_j^{obs} increases.

Appendix E

The $(1-\alpha)$-confidence region for a $p \times p$ covariance matrix can be represented as a p-dimensional ellipsoid (see e.g., Press et al., 2007 or Scheffé 1999). An ellipsoid for a matrix Σ is the set of points $z \neq 0$ that fulfill the equation $z' \cdot \Sigma^{-1} \cdot z = c$ with c being a constant scalar. The value of c can be used to define the $(1-\alpha)$-confidence ellipsoid. Therefore we give c a subscript $\delta := 1-\alpha$ and write c_δ. If c_δ increases the ellipsoid becomes larger (covers more area/volume). So a larger c_δ means a higher level of certainty. We will show that for any value of z the ellipsoid equation for $Var(\gamma^{multi}|.)$ will give a higher critical value c_δ than for $Var(\gamma^{fix}|.)$, i.e., $z' \cdot Var(\gamma^{multi}|.)^{-1} \cdot z = c_\delta^{multi} > c_\delta^{fix} = z' \cdot Var(\gamma^{fix}|.)^{-1} \cdot z \Leftrightarrow \delta^{multi} > \delta^{fix}$. To make this point clearer let us assume $\delta^{multi} = 0.99 > \delta^{fix} = 0.90$. This would imply that only 1% of the data drawn based on $Var(\gamma^{multi}|.)$, but 10% of the data drawn based on $Var(\gamma^{fix}|.)$, are expected to exceed the point z. So for any value of δ the *fix*-ellipsoid fully encloses the *multi*-ellipsoid. And if an ellipsoid E_2, representing matrix M_2, fully encloses another ellipsoid E_1, representing matrix M_1, one can say that values drawn from M_2 vary more in every possible direction than values drawn from M_1. The proof for our case is simple. It uses the result from Appendix B: $Var(\gamma^{multi}|.)^{-1} >_L Var(\gamma^{fix}|.)^{-1}$ which is equivalent to $z' \cdot Var(\gamma^{multi}|.)^{-1} \cdot z > z' \cdot Var(\gamma^{fix}|.)^{-1} \cdot z$ for any $z \neq 0$.

Appendix F

Here we want to provide explanations why the random intercepts variance bias is negatively correlated with s (= relationship of W_1 and $P(y = NA|W_1)$) and the random slopes bias shows a U-shaped pattern for our specific data setting. As mentioned in the analytical section, the properties of Z_j^{obs} (the observations of Z_j for which Y is observed after nonresponse was generated) influence the $(Z_j^{obs'}Z_j^{obs})^{-1}$ part of the multiplicative difference between $Var(\gamma^{multi}|.)$ and $Var(\gamma^{fix}|.)$ (see also Equations 27 and 28 in Appendix C).

In our random slopes setting $Z = (1, W_1)$. This implies that $\left(Z_j^{obs'}Z_j^{obs}\right)^{-1}$ becomes

$$\begin{aligned} &\left(n_{obs\ j}\cdot\sum_{i=1}^{n_{obs\ j}}\left[W_{obs\ 1ij}^2\right] - \left[\sum_{i=1}^{n_{obs\ j}} W_{obs\ 1ij}\right]^2\right)^{-1} \\ &\cdot\begin{pmatrix} \sum_{i=1}^{n_{obs\ j}}\left[W_{obs\ 1ij}^2\right] & -\sum_{i=1}^{n_{obs\ j}} W_{obs\ 1ij} \\ -\sum_{i=1}^{n_{obs\ j}} W_{obs\ 1ij} & n_{obs\ j} \end{pmatrix} \end{aligned} \tag{30}$$

The main diagonal elements in the second term of the product together with the determinant (first term of the product) in Eq. 30 govern the bias in the random intercepts and random slopes. We can simplify (30) further by noting that the determinant can be rewritten as (for readability we will drop the indices i, j, and 1 from here):

$$n_{obs}\cdot\sum\left(W_{obs}^2\right) - \left(\sum W_{obs}\right)^2 = n_{obs}^2 \cdot var(W_{obs}), \tag{31}$$

where $var(W_{obs}) = 1/n_{obs}\cdot\sum\left(W_{obs} - \bar{W}_{obs}\right)^2$. The proof is straightforward if we notice that this empirical variance can be rewritten as

$$\begin{aligned} var(W_{obs}) &= 1/n_{obs}\cdot\left\{\sum\left(W_{obs}^2\right) - 1/n_{obs}\cdot\left(\sum W_{obs}\right)^2\right\} \\ &= 1/n_{obs}\cdot\sum\left(W_{obs}^2\right) - 1/n_{obs}^2\cdot\left(\sum W_{obs}\right)^2 \\ \Rightarrow n_{obs}^2\cdot\ var(W_{obs}) &= n_{obs}\cdot\sum\left(W_{obs}^2\right) - \left(\sum W_{obs}\right)^2 \end{aligned} \tag{32}$$

After this simplification, it can be seen that the main diagonal elements of $(Z_j^{obs'}Z_j^{obs})^{-1}$ are $\sum\left(W_{obs}^2\right)/\left(n_{obs}^2\cdot var[W_{obs}]\right)$ and $(n_{obs}\cdot var[W_{obs}])^{-1}$.

We will start with the second component, the contribution to the random slopes bias. When the expected n_{obs} remains constant (as in our setting), the bias solely depends on $var(W_{obs})$. Note that as $|s|$ increases, it will be more likely that those Y-values with W_1 values in the tails of the distribution of W_1 will be deleted (see Fig. 1). For symmetric distributions such as the normal distribution that we use in our setting this implies that $var(W_{obs})$ will

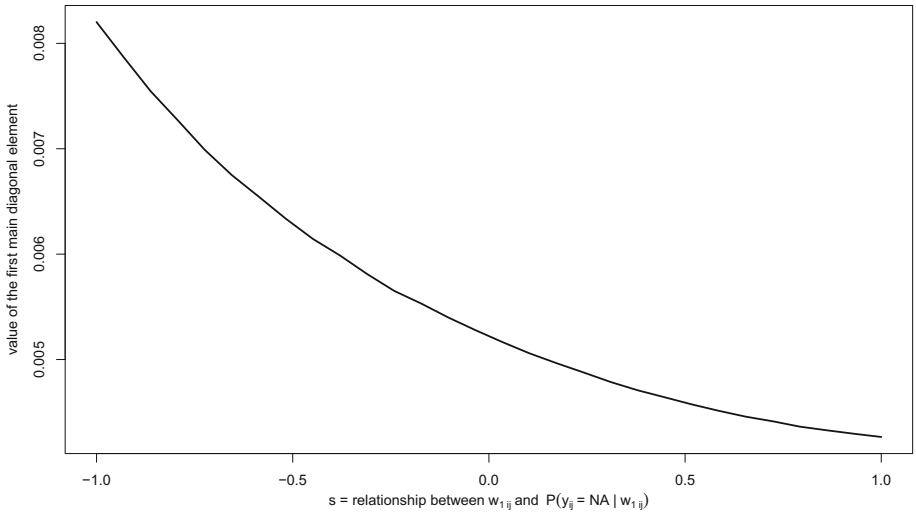


Fig. 6 Simulation of how $\sum \left(W_{obs}^2\right) / \left(n_{obs}^2 \cdot var[W_{obs}]\right)$ changes as s takes different values

decrease and $var(W_{obs})^{-1}$ will increase and thus the bias in the variances of the random slopes will increase. This explains the U-shaped pattern seen in Fig. 3. Regarding the contribution to the random intercepts bias, we find that s has a negative relationship to the bias for our data setting. But whether $\sum \left(W_{obs}^2\right) / \left(n_{obs}^2 \cdot var[W_{obs}]\right)$ is large or small highly depends on W_{obs}. We simulated this ratio for our data setup for various values of s. Figure 6 displays this ratio. The negative relationship between s and $\sum \left(W_{obs}^2\right) / \left(n_{obs}^2 \cdot var[W_{obs}]\right)$ is in line with our empirical findings regarding the bias for the variance of the random intercepts (see Fig. 2). However, unlike the U-shaped relationship between s and the bias for the variance of the slopes that should hold for any random slopes model with a symmetric distribution of the slope variable, we emphasize that the relationship between s and the variance of the random intercepts is specific to this data setting and might well be reversed in other data settings.

References

American Psychological Association (2017). Education and socioeconomic status. http://www.apa.org/pi/ses/resources/publications/education.aspx, accessed: 2017-06-22.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, *53*(1), 53–74. https://doi.org/10.1002/bimj.201000140.

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with mplus. MPlus Web Notes https://www.statmodel.com/download/Imputations7.pdf.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., & Green, P. (2016). lme4: Linear Mixed-Effects Models using 'Eigen' and S4. http://cran.r-project.org/web/packages/lme4/index.html, r package version 1.1-12.

Blossfeld, H. P., Roßbach, H.G., & Von Maurice, J. (2011). Education as a Lifelong Process - The German National Educational Panel Study (NEPS), vol. 14, Zeitschrift für Erziehungswissenschaft. http://link.springer.com/journal/11618/14/2/suppl/page/1.

Brown, E., Graham, J., Hawkins, J., Arthur, M., Baldwin, M., Oesterle, S., Briney, J., Catalano, R., & Abbott, R. (2009). Design and analysis of the community youth development study longitudinal cohort sample. *Evaluation Review*, *33*, 311–324.

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, *45*(5), 1–14. https://doi.org/10.18637/jss.v045.i05, http://www.jstatsoft.org/v45/i05.

Clark NM, Shah S, Dodge JA, Thomas LJ, Andridge RR, & Little RJ (2010). An evaluation of asthma interventions for preteen students. *Journal of School Health*, *80*(2), 80–87. https://doi.org/10.1111/j.1746-1561.2009.00469.x

Diaz-Ordaz, K., Kenward, M. G., Gomes, M., & Grieve, R. (2016). Multiple imputation methods for bivariate outcomes in cluster randomised trials. *Statistics in Medicine*, *35*(20), 3482–3496. https://doi.org/10.1002/sim.6935.

Drechsler, J. (2011). Multiple imputation in practice—a case study using a complex German establishment survey. *Advances in Statistical Analysis*, *95*(1), 1–26.

Drechsler, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, *40*(1), 69–95. https://doi.org/10.3102/1076998614563393.

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–240. https://doi.org/10.1037/met0000063.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Goldstein, H. (1987). Multilevel Models in Educational and Social Research. Charles Griffin Book, Charles Griffin & Company, https://books.google.de/books?id=UKzrAAAAMAAJ.

Goldstein, H. (2011). Multilevel Statistical Models, 4th edn., Wiley, Chichester, UK.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231. http://www.tandfonline.com/doi/abs/10.1207/S15328031US0104_02.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. Behavior Research Methods, pp 640–649, https://doi.org/10.3758/s13428-015-0590-3.

Harville, D. A. (1997). Matrix algebra from a statistician's perspective. Springer.

Hedeker, D., & Gibbons, R. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psych Methods*, *2*(1), 64–78. https://doi.org/10.1037/1082-989X.2.1.64.

Horn, R. A., & Johnson, C. R. (1990). Matrix Analysis, reprint edn. Cambridge University Press, http://amazon.com/o/ASIN/0521386322/.

Kenward, M. G., & Carpenter, J. (2007). Multiple Imputation: Current Perspectives. *Statistical Methods in Medical Research*, *16*(3), 199–218. https://doi.org/10.1177/0962280206075304.

Lenkeit, J. (2012). How effective are educational systems? A value-added approach to measure trends in pirls. *Journal for Educational Research Online*, *4*(2), 143–173. http://www.j-e-r-o.com/index.php/jero/article/view/317/157.

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165.

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton LS (2004a). Evaluating Value-Added Models for Teacher Accountability, 0th edn. RAND Corporation, http://amazon.com/o/ASIN/0833035428/.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004b). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101. https://doi.org/10.3102/10769986029001067.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–573. https://doi.org/10.1214/ss/1177010269, http://projecteuclid.org/euclid.ss/1177010269.

Mistler, S. A. (2013). A SAS Macro for Applying Multiple Imputation to Multilevel Data. Proceedings of the SAS Global Forum https://support.sas.com/resources/papers/proceedings13/438-2013.pdf.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects?. *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237, http://epa.sagepub.com/content/26/3/237.short?rss=1&ssource=mfc.

O'Connell, A. A., & McCoach, D. B. (2008). Multilevel modeling of educational data. Information Age Publishing.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3rd edn. Cambridge University Press. http://prefixwww2.units.it/ipl/students_area/imm2/files/Numerical_Recipes.pdf.

Quartagno, M., & Carpenter, J. (2016). jomo: A package for Multilevel Joint Modelling Multiple Imputation. http://CRAN.R-project.org/package=jomo, r package version 2.1-2.

R. Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, *32*(2), 143–149.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3581, http://biomet.oxfordjournals.org/content/63/3/581.short.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470316696.

Schafer, J. L. (2016). pan: Multiple Imputation for Multivariate Panel or Clustered Data. http://cran.r-project.org/web/packages/pan/index.html, r package version 1.4.

Scheffé, H. (1999). The analysis of variance, 1st edn. Wiley-Interscience,.

Siotani, M. (1967). Some applications of Loewner's ordering on symmetric matrices. *Annals of the Institute of Statistical Mathematics*, *19*(2), 245–259. http://www.ism.ac.jp/editsec/aism/pdf/019_2_0245.pdf.

Speidel, M., Drechsler, J., & Jolani, S. (2017). hmi: Hierarchical Multiple Imputation. prefixhttps://CRAN.R-project.org/package=hmi, r package version 0.7.4.

StataCorp (2011). Accounting for clustering with mi impute. http://www.stata.com/support/faqs/statistics/clustering-and-mi-impute/ (retrieved on 09.11.2016).

Taljaard, M., Donner, A., & Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biometrical Journal*, *50*(3), 329–345. https://doi.org/10.1002/bimj.200710423.

van Buuren, S. (2011). *Multiple imputation of multilevel data*, (pp. 173–196). Milton Park, UK: Routledge Academic. chapter 10.

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). mice: Multivariate imputation by chained equations. https://cran.r-project.org/web/packages/mice/index.html, r package version 2.25.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. New-York: Springer.

Wooldrige, J. (2010). *Econometric analysis of cross section and panel data*, 2nd edn. Cambridge, MA: The MIT Press.

Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics*, *32*(1), 231–256. https://doi.org/10.1515/JOS-2016-0011

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

IAB-Discussion Paper 16/2018

Articles on labour market issues

R Package hmi: A Convenient Tool for Hierarchical Multiple Imputation and Beyond

Matthias Speidel
Jörg Drechsler
Shahab Jolani

ISSN 2195-2663

R Package hmi: A Convenient Tool for Hierarchical Multiple Imputation and Beyond

Matthias Speidel (IAB)
Jörg Drechsler (IAB)
Shahab Jolani (Maastricht University)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

Abstract

Applications of multiple imputation have long outgrown the traditional context of dealing with item nonresponse in cross-sectional data sets. Nowadays multiple imputation is also applied to impute missing values in hierarchical data sets, address confidentiality concerns, combine data from different sources, or correct measurement errors in surveys. However, software developments did not keep up with these recent extensions. Most imputation software can only deal with item nonresponse in cross-sectional settings and extensions for hierarchical data – if available at all – are typically limited in scope. Furthermore, to our knowledge no software is currently available for dealing with measurement error using multiple imputation approaches.

The `R` package `hmi` tries to close some of these gaps. It offers multiple imputation routines in hierarchical settings for many variable types (for example, nominal, ordinal, or continuous variables). It also provides imputation routines for interval data and handles a common measurement error problem in survey data: Biased inferences due to implicit rounding of the reported values. The user-friendly setup which only requires the data and optionally the specification of the analysis model of interest makes the package especially attractive for users less familiar with the peculiarities of multiple imputation. The compatibility with the popular `mice` package ensures that the rich set of analysis and diagnostic tools and post-imputation commands available in `mice` can be used easily once the data have been imputed.

Zusammenfassung

Anwendungen von Multipler Imputation sind längst über den klassischen Kontext der Behandlung von fehlenden Beobachtungen in Querschnittsstudien heraus gewachsen. Heutzutage wird Multiple Imputation auch verwendet um fehlenden Werten in hierarchischen Datensätzen zu imputieren, um Vertraulichkeits-Interessen zu begegnen, um Datensätze aus verschiedenen Quellen zu kombinieren oder um Messfehler aus Erhebungen zu korrigieren. Die meiste Imputationssoftware kann allerdings nur mit fehlenden Beobachtungen in Querschnittsdaten umgehen und Erweiterungen für hierarchische Daten - sofern überhaupt vorhanden - sind typischerweise in ihrem Umfang begrenzt. Unserem Kenntnisstand nach, ist aktuell keine Software für den Umgang mit Messfehlern, basierend auf Multiplen Imputationsmethoden, vorhanden. Das R-Packet `hmi` versucht einige dieser Lücken zu schließen. Es bietet Multiple Imputationsroutinen in hierarchischen Settings für viele Variablentypen (zum Beispiel nominal, ordinal oder stetige Variablen). Zudem stellt es Imputationsmethoden für Intervalldaten bereit und behandelt ein übliches Messfehlerproblem in Befragungsdaten: Verzerrungen aufgrund impliziten Rundens der berichteten Werte. Der nutzerfreundliche Aufbau, der nur die Daten und optional eine Spezifizierung des Analysemodels benötigt, macht das Paket besonders attraktiv für Nutzer die weniger vertraut mit den Besonderheiten von Multipler Imputation sind. Die Kompatibilität mit dem populären Paket `mice` stellt sicher, dass der reichhaltige Satz an Analyse- und Diagnosewerkzeugen,

und Befehlen für das Imputationsergebnis aus `mice`, einfach angewandt werden kann, sobald die Daten imputiert wurden.

JEL classification: C38; C83

Keywords: hierarchical data, multiple imputation, multilevel models, measurement error, heaping, R

1 Introduction

Forty years after Donald Rubin's seminal paper (Rubin, 1978) which introduced the concept of multiple imputation, the approach has been shown to be useful in many contexts going far beyond the classical item nonresponse in cross sectional surveys for which it was originally proposed (Reiter/Raghunathan, 2007). Today, multiple imputation is used to deal with nonresponse in hierarchical data sets (Carpenter/Kenward, 2013: chap. 9), address confidentiality concerns by disseminating synthetic data instead of the original data (Drechsler, 2011), concatenate files from different data sources (Rubin, 1986; Rässler, 2003; Reiter, 2012), address measurement error in self-reported health information (Schenker/Raghunathan/Bondarenko, 2010), handle changes in the coding of variables in longitudinal studies (Clogg et al., 1991; Schenker, 2003), or impute plausible values for coarse data (Taylor/Schwartz/Detels, 1986; Heitjan/Rubin, 1990; Raghunathan et al., 2001). As discussed in Heitjan/Rubin (1991) coarse data are data for which the true values are not observed in a precise way. This includes missing data as a special case, but also rounding, grouping, censoring and interval data. Examples of applications of multiple imputation for coarse data include Gartner/Rässler (2005); Jenkins et al. (2011); Drechsler/Kiesl/Speidel (2015).

While classical imputation methodology as discussed for example in Rubin (1987) or van Buuren (2012) is sufficient for some of these applications, adjusted methodology is required for others. However, although all major statistical software such as `SPSS`, `Stata`, `SAS`, or `R` offer multiple imputation routines today, the available methodology is typically limited to the classical methodology for cross-sectional surveys. Some software also provides methods for dealing with hierarchical data structures, but as we will illustrate in Section 2.6, current implementations are limited in scope. With the exception of the recently implemented software package `synthpop` (Nowok/Raab/Dibben, 2016) which was specifically developed for generating synthetic data sets for disclosure protection, no software exists to our knowledge for applications such as the coarse data problem discussed above, which require modifications of the traditional multiple imputation framework.

The `R` package `hmi` closes some of the gaps of currently available software by offering four important contributions:

1. It offers imputation routines for hierarchical data using multilevel (mixed-effects) models for all variable types based on the sequential regression approach, which unlike the joint modeling approach can also handle item nonresponse if random slope models need to be estimated (see Section 2.4 for details)
2. It provides routines for dealing with rounding in reported values based on the methodology proposed in Heitjan/Rubin (1991).
3. It offers routines for imputing plausible values if it is only known (for some of the observations) that the exact value lies in certain intervals, for example if the data are censored. Currently, such imputation routines are only available in `Stata`.
4. It allows to deal with item nonresponse, interval information and rounding within the same variable simultaneously following the approach described in Drechsler/Kiesl/Speidel (2015).

The package also offers imputation tools for "classical" missing data problems by calling imputation routines available in the popular multiple imputation package `mice` (van Buuren/Groothuis-Oudshoorn, 2011). Since the objects generated using `hmi` are structured similar to objects generated using `mice` (both are `mids` objects), the rich set of analysis and diagnostic tools and post-imputation commands available in `mice` can be used easily once the data have been imputed. Furthermore, the package provides imputation routines for semi-continuous variables, that is, variables which have a spike at one value (typically zero), but can be considered continuous otherwise. These imputation routines are available in several software packages, but are not offered in `mice`.

To facilitate the usage of the package for less experienced users, the selection of suitable imputation models is highly automated, that is, the user only needs to provide the data. The package will identify the most appropriate imputation models for each variable with missing values using decision rules described in Section 5 of this paper. Additionally, the user can specify the substantive model he or she wants to run on the imputed data set. In this case `hmi` will use the same set of predictors and the same functional form as the substantive model for all imputation models in an effort to make the congeniality assumption more plausible. As discussed in Meng (1994), congeniality between the imputation model and the substantive model is important to avoid biased inferences based on the imputed data. We illustrate in Section 2.3 that specifying the substantive model is especially important if multilevel models will be fitted at the analysis stage since this will ensure that the hierarchical structure of the data will also be taken into account at the imputation stage. The package is available at `https://cran.r-project.org/package=hmi`.

The remainder of the paper discusses the main contributions of the package and provides detailed illustrations on how the package can be used. Specifically, Sections 2 to 4 address multiple imputation for hierarchical, interval and rounded data. Each section starts by illustrating the inferential problems caused by the various data deficiencies followed by a brief review of the required multiple imputation methodology for addressing the said problem. Limitations of currently available software and our contributions are also discussed. Section 5 describes the `hmi` package in detail: all mandatory and optional arguments, the internal checks, the handling of the model formula, the types of supported variables, and the implemented convergence checks will be presented. In Section 6 we provide real data applications to illustrate the implementation of the different features of the package. We end with a conclusion.

2 Multiple imputation for hierarchical data sets

Hierarchical data sets are data sets in which individual records are nested within groups. Typical examples include students in the same class or repeated measures of the same individual. In such settings, the assumption of independent observations, needed for the classical linear regression model, does not hold since records belonging to the same group tend to be more homogeneous than records belonging to different groups. To account for these cluster effects, multilevel models (also referred to as random effects or mixed effects models depending on the field of study) are often employed. In the following, we provide a

brief summary of the methodology behind multilevel modeling starting with multilevel linear models for continuous variables. Then, we discuss extensions to multilevel generalized linear models for any variable type from the exponential family. A more detailed introduction can be found in any textbook on multilevel modeling, for example in Raudenbush/Bryk (2002). The brief overview will form the basis for our discussion of appropriate imputation strategies for hierarchical data and details about their implementation and available software in Sections 2.3 to 2.7.

2.1 Multilevel linear models

Paraphrasing from Speidel/Drechsler/Sakshaug (2017), multilevel linear models assume a linear relationship between the continuous target variable Y and some covariates X and Z. The effect of X on Y is governed by some global fixed effects β; the effect of Z on Y by some cluster specific random effects γ. Often Z is a subset of X, meaning that variables that are assumed to have a random effect are also included as fixed effect variables in the model.

The standard multilevel model has the form

$$\begin{aligned} y_{ij} =& x_{ij}\beta + z_{ij}\gamma_j + \varepsilon_{ij}, \\ \gamma_j \sim& N(0,\ \Sigma), \\ \varepsilon_{ij} \sim& N(0,\ \sigma^2), \end{aligned} \tag{1}$$

with $j = 1, \ldots, J$ being the index for the clusters, $i = 1, \ldots, n_j$ being the index for the units belonging to cluster j, and n_j being the number of observations in cluster j. The parameter β contains the global fixed effects, similar to the regression coefficients in classical linear regression models. The parameters γ_j are the cluster specific random effects, which are assumed to follow a normal distribution with zero mean vector and variance matrix Σ. These random effects and the normality assumption for them is a key difference to the classical linear regression model. The parameter ε_{ij} is the error term which is normally distributed with zero mean and variance σ^2, which is constant for all clusters.

Multilevel linear models can be generalized to more than two levels and residual variances being heteroscedastic across the clusters. Since `hmi` can only handle two levels of hierarchy and homoscedastic residuals at the moment, we do not cover these extensions here. The interested reader is referred to Raudenbush/Bryk (2002) or Snijders/Bosker (2011) for more details on these topics.

2.2 Multilevel generalized linear models

The step from multilevel linear models to multilevel generalized linear models (mglm) is analogous to the step from classical linear models to generalized linear models (glm). Both enable model estimation for variables from the exponential family using a linear predictor l and a link function f such that $E(Y) = \mu = f^{-1}(l)$. The major difference between mglm and glm is that the linear predictor in mglm also has random effect variables Z with

regression coefficients $\gamma = \{\gamma_1, \ldots, \gamma_J\}$ leading to $l = X\beta + Z\gamma + \varepsilon$. These random effects and their covariance matrix Σ also have to be considered when estimating the model.

The link function is defined according to the type of variable that is modeled. For example for continuous variables the identity link is used and for count data the log-link. In general no closed form solution for the parameter estimates exist, so Markov Chain Monte Carlo (MCMC) methods or other iterative procedures are required for estimation (Gelman/Hill, 2006; Hadfield, 2010).

2.3 Dealing with missing values in hierarchical data

Hierarchical data are not spared from nonresponse and multiple imputation can be a convenient strategy to address this problem. Several researchers have shown that ignoring the hierarchical structure at the imputation stage will lead to biased inferences when analyzing the data (Reiter/Raghunathan/Kinney, 2006; van Buuren, 2011; Enders/Mistler/Keller, 2016; Zhou/Elliott/Raghunathan, 2016; Lüdtke/Robitzsch/Grund, 2017). Furthermore, accounting for the clustering by adding indicator variables for the clusters (fixed effects modeling) will still introduce bias if the analysis is based on a multilevel model (Taljaard/Donner/Klar, 2008; Andridge, 2011; Drechsler, 2015; Speidel/Drechsler/Sakshaug, 2017). To avoid this bias due to uncongeniality between the imputation and the analysis model, all manuscripts suggest using multilevel models also at the imputation stage.

2.4 Multiple imputation using multilevel models

With multiple imputation missing values are imputed multiply ($M \geq 2$ times) to be able to take the uncertainty from imputation into account. The imputed values are random draws from the distribution of the missing data given the observed data. Let $D = \{D_{obs},\ D_{mis}\}$ denote the data D separated into an observed part (D_{obs}) and a missing part (D_{mis}) and let θ contain the parameters which govern the distribution of D. To obtain approximate draws from $f(D_{mis} \mid D_{obs})$ multiple imputation repeatedly applies the following two steps:

1. Draw a new set of parameters $\theta^\star$ from their posterior distribution given the observed data: $f(\theta \mid D_{obs})$.
2. Draw replacements for the missing values from the predictive distribution of the missing data given the observed data and the drawn parameters from the previous step: $f(D_{mis} \mid \theta^\star,\ D_{obs})$.

Valid inferences based on the imputed data can be obtained using the generic inferential procedures first described in Rubin (1978). For further details regarding the general properties of multiple imputation we refer to any textbook on multiple imputation, for example Rubin (1987); van Buuren (2012); Carpenter/Kenward (2013).

As pointed out above, if the model to be estimated on the imputed data is a multilevel model, a similar model specification should be used at the imputation stage to ensure unbiased results. Thus, for continuous variables the imputation model should follow the

model specification given in Equation (1) and the two generic multiple imputation steps described above consist of the following two steps:

1. Draw a new set of parameters $\theta^\star = \{\beta^\star,\ \gamma^\star,\ \Sigma^\star,\ (\sigma^\star)^2\}$ from their posterior distribution.

2. Generate imputed values by drawing from

$$y_{ij}^{imp} = x_{ij}^{imp}\beta^\star + z_{ij}^{imp}\gamma_j^\star + \varepsilon_{ij}{}^\star \qquad \varepsilon_{ij}^\star \sim N(0,\ (\sigma^\star)^2), \tag{2}$$

where the superscript *imp* identifies all records for which Y is imputed. Unlike in the classical linear regression case, no closed form solutions exist for the posterior distribution of the parameters. Thus, Markov Chain Monte Carlo methods or other approximations (Jolani, 2018) are generally required to update the parameters. We refrain from providing the details of the iterative procedure here for brevity. The interested reader is referred to Goldstein (2011) for a detailed description of Gibbs sampling methods for hierarchical data and to Carpenter/Kenward (2013: chap. 9) and Drechsler (2015) for applications in the missing data context.

2.5 Joint modeling vs. sequential regression for multilevel multiple imputation

Two general strategies exist for imputing missing values if more than one variable is affected by nonresponse: joint modeling and sequential regression. The *joint modeling* approach specifies a joint distribution for all variables with missing data (potentially conditioning on fully observed variables) and draws imputed values based on this distribution. For example, if all variables to be imputed are continuous, a multivariate normal distribution is typically specified for those variables affected by nonresponse. A major drawback of the joint modeling approach in the multilevel context is that it cannot be used if missingness also occurs in the random slope variable(s) (Carpenter/Kenward, 2013; Enders/Mistler/Keller, 2016). Furthermore, the specification of a joint distribution can be difficult, if different variable types need to be modeled.

The *sequential regression* approach (also known as *chained equations* or *fully conditional specification*) does not require modeling the joint distribution directly. Instead, conditional distributions are specified for each variable to be imputed. The variables are imputed sequentially, conditioning on the other variables in the data set. However, some of the predictors in the imputation model might themselves contain imputed values. Thus, the model estimates will change if these imputed values are updated. To account for this, the procedure of sequentially imputing each variable has to be repeated several times, until the draws from the conditional distribution converge to draws from the implicitly specified joint distribution (see Raghunathan et al. (2001) for further details on the sequential regression approach).

A downside of the approach is that convergence is only guaranteed if this joint distribution exists. However, Liu et al. (2014) and Zhu/Raghunathan (2015) show that the joint distribution will exist under rather general conditions and even if this is not the case, inferences

based on the imputed data will still be consistent as long as the conditional distributions are correctly specified.

2.6 Existing imputation routines for hierarchical data and their limitations

To our knowledge the only `R` (R Core Team, 2016) packages allowing hierarchical multiple imputation are `jomo` (Quartagno/Carpenter, 2018), `mice` (van Buuren/Groothuis-Oudshoorn, 2011), `micemd` (Audigier/Resche-Rigon, 2018) and `pan` (Schafer, 2016). Currently, `mice` is limited to continuous variables for hierarchical settings and cannot impute other variable types using a multilevel model. `micemd` also provides multilevel imputation functions for binary and integer variables, but not for categorical variables with more than two categories. A downside of `jomo` and `pan` is the fact that they rely on the joint modeling approach, with the drawbacks mentioned in the previous section.

Imputation routines based on multilevel models have also been developed for other statistical software packages: For `SAS` the external macro `MMI_IMPUTE` (Mistler, 2013) can be used. `Mplus` (Asparouhov/Muthén 2010) and the stand alone software `REALCOM-IMPUTE` also offer some multilevel multiple imputation routines. All of these imputation routines also use the joint modeling approach. To our knowledge, the only other software allowing multilevel imputation based on the more flexible sequential regression approach is the recently released standalone software `blimp` (Enders/Keller/Levy, 2017).

2.7 Our contribution for the imputation of hierarchical data

As mentioned in the introduction, `hmi` is designed to provide multilevel imputation routines for many relevant variable types, including semi-continuous variables based on the flexible sequential regression approach. Furthermore, it also offers single level models for all types of variables, for situation where a multilevel model is not applicable.

If an analysis model is specified, the package will automatically use the same set of predictors and the same functional form as the substantive model for all imputation models to avoid introducing bias in the analysis, because relationships which are important to the analyst are not reflected in the imputation models. If no analysis model is given, all variables are imputed using single level models by default. However, if desired, the user can manually specify which imputation models should be used for each variable.

For single level imputation, the package relies on the imputation routines implemented in `mice`. Own code is used for all multilevel imputation routines. The draws from the posterior distribution of the parameters of the multilevel models are obtained using MCMC methods implemented in the `MCMCglmm` package (Hadfield, 2010).

If multilevel imputations are employed, the package also stores the model parameters at each iteration of the MCMC chains, to enable the users to monitor the convergence of the chains. The users can either extract this information to run their own convergence diagnostics or they can rely on the checks implemented in the package. Per default the package runs Geweke's stationarity test (Geweke 1992) on each chain, plots those chains

that failed the test and provides some summary information on the number of chains which failed the test (see Section 5.8 for details).

3 Multiple imputation for interval data

Interval data (sometimes called *bracketed response*) comprise all data for which an interval covering the true value is given instead of the exact value. According to this definition both, grouped and censored data can be treated as interval data. With *grouped data*, a set of precise observations is grouped into a single response group. For example in cancer research the number of positive lymph nodes might only be collected in categories 0, 1-3, 4-9 and 10+ (Royston, 2007) or age might only be reported in five year intervals for confidentiality reasons. Grouped data can also arise if surveys aim to maximize response rates for sensitive or difficult questions. For example, in the *Survey of Consumer Finances* (SCF) range cards are shown to respondents who refuse to provide information regarding their exact income, asking them to pick one of the ranges depicted (e.g. 0-5,000 $\$$) or to pick a category following a decision tree (Kennickell, 1991). A similar procedure is implemented in the *National Health Interview Survey* (NHIS), where initial nonresponders to the question regarding the yearly income are asked whether their income is above or below $20,000$ USD and in a next step a range card with 44 income categories is shown (Schenker et al., 2006). The German Panel Study *Labour Market and Social Security* (PASS) also asks initial nonresponders consecutive questions about intervals covering the true income (Trappmann et al., 2010). These approaches help to collect at least some information for respondents initially refusing to provide an answer (Drechsler/Kiesl/Speidel, 2015) or selecting “don’t know” for the exact income question (Kennickell, 1996).

Censoring refers to the situation in which values above (or below) a given threshold are not observed. The only information available is that the true value must be above (or below) the known threshold. Censoring from the left typically arises in situations in which technical equipment will not detect the measure of interest if its concentration is below a certain limit. For example, in the study presented in Pilcher et al. (2007), the number of HI viruses in human blood is only measurable once it is above a given threshold of detection. Censoring from the right often occurs in public use files, in which top coding is applied to reduce the risk of re-identification. This is for example done in the US-American Current Population Survey (CPS) (Larrimore et al., 2008). An example of right censoring in biology is the time to seed germination as the time it takes for a seed to germinate can be longer than the duration of the study (Scott/Jones, 1990).

3.1 Analyzing interval data

Obtaining valid inferences if only interval information is available for (parts of) some of the variables can be complex. The most common strategy is to adjust the likelihood accordingly. For example, in linear regression models, the well known tobit model (Tobin, 1958) can be used to account for censoring in the dependent variable. This approach can easily be extended to other forms of interval data but iterative procedures are typically required

to find the maximum likelihood estimates in this case. Since most software packages do not offer routines for dealing with interval data beyond the tobit model, some applied researchers rely on naive approaches for analyzing the interval data: A common approach is to ignore the interval information completely, using only those observations for which exact information is observed. This approach is always inefficient, since available information is not used. It can also introduce bias, if those units that only provide interval information differ form those units which provide exact information. In fact, Heeringa/Little/Raghunathan (1997) showed that the tendency to only report intervals for income increases with income. Thus, results solely based on the exact reports are likely to be biased.

To simplify the analysis for applied researchers, imputation approaches can be used to generate plausible values given the interval information. This offers the advantage that the analysts no longer need to find appropriate ways for incorporating the interval information. They can rely on standard analysis models using the plausible values for inference. However, just like in the standard nonresponse context, care needs to be taken to ensure that valid inferences can be obtained from the imputed data.

For example, a naïve imputation approach which is sometimes applied in practice uses the midpoint or the upper bound of each reported interval as the imputed value (Law/Brookmeyer, 1992; Dorey/Little/Schenker, 1993). The data are then analyzed treating the imputed values as the true exact values. These approaches are valid only in very limited settings since they will generally underestimate the variance in the imputed data (Law/Brookmeyer, 1992; Kim/Xue, 2002).

To fully account for the uncertainty resulting from the fact that only intervals instead of exact values are observed initially, multiple imputation approaches are required which generate imputations by drawing from the conditional distribution of the exact values given the interval information (and additional information from other variables available in the data set).

Imputation approaches have been used for several data sets to facilitate the analysis for the user. For example, since 1995 the *Survey of Consumer Finances* generates imputed income values by drawing from truncated normal distributions using the bounds of the reported intervals as truncation points.

An application of the joint modeling approach for imputation of interval data is discussed in Heeringa (1993). The author imputed interval and missing data in the *Health and Retirement Survey* (HRS) using the *general location model*. One major disadvantage of the general location model is that the multivariate normal distribution needs to be estimated for each cell of the table spanned by crossclassifying all categorical variables. Thus, the approach can only be used if the number of categorical variables is very limited to ensure a sufficient number of observations for estimating the normal distribution within each cell. A second problem can be sparse cells in the interval variable, making the imputation model unreliable. The author noticed this problem especially for the largest income category which typically included only few, very wealthy individuals. The true income distribution in this category also might be very skewed, violating the normality assumption.

For settings with ordered income categories affected by item nonresponse, Bhat (1994) proposed an imputation method modeling the income distribution and the response probabilities jointly using a selection modeling approach.

Raghunathan et al. (2001) describes a general sequential regression approach for interval data. Plausible values are generated by drawing from truncated normal distributions. The parameters for the model are estimated using those observations for which an exact value is available. New parameters for the truncated normal model are drawn using sampling/importance resampling (SIR, Rubin 1988). This approach is also implemented in the multiple imputation software `IVEware` (Raghunathan et al., 2016). The software was also used to impute plausible values for interval answers in the *National Health Interview Survey* (NHIS), (Schenker et al., 2006).

Royston (2007) implemented an imputation model for interval data for `Stata`. He extended the approach of Raghunathan et al. (2001) by also using the information from the respondents that only provided an interval when estimating the parameters of the imputation model. To obtain parameter estimates the joint likelihood of the income of the exact reporters and the income of the interval reporters is maximized under the implicit assumption that the conditional distribution of the true income given the covariates in the model is the same for both groups. Instead of using SIR, draws from the posterior distribution of the parameters are only approximated by drawing from a multivariate normal distribution centered around the maximum likelihood estimates of the parameters. Compared to the approach of Raghunathan et al. (2001) this strategy offers the advantage that it uses all available information and that it can also be used if only interval information is available.

A similar approach was later used by Drechsler/Kiesl/Speidel (2015) for simultaneous imputation of interval, rounded, and missing data. For interval data without rounding, the approach simplifies to the method described by Royston (2007) and is separately implemented in `hmi`.

Several (multiple) imputation approaches have also been proposed for the special case of survival data (Taylor/Schwartz/Detels, 1986; Muñoz et al., 1989; Taylor et al., 1990; Dorey/Little/Schenker, 1993). In survival analysis censoring is a common problem since for those units that entered a certain state of interest (for example unemployment) previous to the start of the study or are still in that state at the time the study is terminated, the true time of entry or exit is unknown. Imputation routines for survival data differ systematically from the imputation routines for interval data in other data sets since survival models need to be used for imputation to ensure congeniality between the imputation and the analysis model. Multiple imputation routines for this special type of data are implemented in the `R` package `icenReg` (Anderson-Bergman, 2017). Imputations in `icenReg` can be based on proportional hazards, proportional odds or accelerated failure time models. Since `icenReg` already provides a convenient tool for dealing with survival data, we did not implement these routines in `hmi` and we limit the description of the imputation methodology in the next section to applications outside the survival analysis context. The interested reader is referred to Grover/Gupta (2015) or Anderson-Bergman (2017) for details regarding imputation routines for survival data.

3.2 Methodology of multiple imputation for interval data

Let $y = \{y_1, \ldots, y_n\}$ be the realizations of the variable of interest– possibly transformed to fulfill the normality assumption of linear regression models – for which only interval information is available for some or all of the n observations in the data. Let $x = \{x_1, \ldots, x_n\}$ be the realizations of any other variables X available in the data set which might help to predict the values of y. We assume that

$$Y|X \sim N(X\beta, \sigma^2) \tag{3}$$

If exact values would be observed for all records, the likelihood of the model parameters would be

$$L(\beta, \sigma^2|y, x) = \prod_{i=1}^{n} f\left(y_i, \mu_i = x_i'\beta, \sigma^2\right) \tag{4}$$

with f being the density of a normal distribution.

If only interval information is available for some of the respondents, we need to introduce some additional notation. Let I_i be an indicator function that equals zero if exact information is available and equals one if only interval information is available for individual i (the interval information includes missing data as a special case with interval bounds $-\infty$ and $+\infty$). Let $\underline{y_i}$ and $\overline{y_i}$ be the lower and upper bound of the interval for unit i. The extended likelihood that also takes the interval information into account is given by

$$L(\beta, \sigma^2|y, x) = \prod_{i=1}^{n} \left((1 - I_i) f(y_i, x_i'\beta, \sigma^2) + I_i \left[F(\overline{y_i}, x_i'\beta, \sigma^2) - F(\underline{y_i}, x_i'\beta, \sigma^2)\right]\right), \tag{5}$$

with F being the cumulative distribution function of the normal distribution. Maximizing this likelihood will provide estimates for the parameters $\theta = \{\beta, \sigma^2\}$. To approximate a draw from the posterior distribution of $f(\theta|y,\ x)$ under the assumption of flat priors for all parameters, we can draw from

$$\theta^\star \sim MVN(\hat{\theta},\ I(\hat{\theta})), \tag{6}$$

where $\hat{\theta}$ contains the maximum likelihood estimates of θ, and $I(\hat{\theta})$ is the negative inverse of the Hessian matrix of the log-likelihood with $\hat{\theta}$ plugged in.

Plausible values for interval respondents can be imputed by drawing from a truncated normal distribution $N_t(\mu, \sigma^2)$ with $\mu = x'\beta^\star$, $\sigma^2 = (\sigma^\star)^2$, where $\beta^\star$ and $(\sigma^\star)^2$ are the parameters drawn form the approximate posterior distribution as described above. The truncation points are given by the bounds of the reported interval. Imputations for those respondents that refused to provide any information are obtained by drawing from a normal distribution with parameters $\mu = x'\beta^\star$ and $\sigma^2 = (\sigma^\star)^2$.

3.3 Our contribution for the imputation of interval data

To our knowledge, imputation routines for interval data following the procedures described above are currently only available in `Stata`. For the special case of survival data imputation routines following a completely different methodology are available in the `R` package `icenReg` by Anderson-Bergman (2017). The `hmi` package is the first `R` package to offer general imputation routines for interval data beyond the survival data context. The package also provides a new solution for storing information on lower and upper bounds of the interval information in *one* variable together with a set of functions for handling interval data.

The idea is to store the bounds in a character variable separated by a semicolon. Such an interval object can be generated using `generate_interval` or split into its lower and upper bounds by `split_interval`. See Section 5.5 for details and Section 6.2 for examples.

4 Multiple imputation for data affected by heaping

Another form of coarse data are data for which the reported values are implicitly rounded. The rounding can either be identical for all individuals (for example if individuals round off their age), or subject to different rounding degrees. Many individuals rounding to the same value lead to heaps in the empirical distribution of the data. Therefore, this form of rounding with unknown rounding degrees is often referred to as heaping in the literature. It typically occurs, if the respondent is unwilling or unable to provide an exact value and instead reports a value which is a multiple of some common rounding base to implicitly express his or her uncertainty regarding the estimate. In many cases, multiples of 10, 100, or 1,000 are used. In other situations, the respondent uses a higher level of aggregation (such as years instead of months or weeks instead of days) for the estimate. For example, Heitjan/Rubin (1990) studied reported ages for young children in Tanzania and noted several heaps at certain values, such as 6 or 12 months. Huttenlocher/Hedges/Bradburn (1990) found heaps at multiples of seven for questions which asked how many days ago an event took place. Wang/Heitjan (2008) identified several heaps at multiples of 20 in questions regarding cigarette consumption because the common pack of cigarettes contains 20 cigarettes.

Table 1 taken from Drechsler/Kiesl/Speidel (2015) illustrates the problem using reported monthly household income in the German panel study *Labour Market and Social Security* (PASS) (Trappmann et al., 2010) for the year 2008/2009. The table provides the percentage of the reported monthly income values that are divisible by a given round number. It seems that most respondents tend to round their income. More than 60 percent of the reported values are divisible by 100 and less than 16 percent of the values are not divisible by 5. Czajka/Denmead (2008) report similar problems for the *American Community Survey* and the *Current Population Survey*.

The major problem with heaping is that inferences will be biased if the reported values are treated as face value (Hanisch, 2005). For example, Drechsler/Kiesl (2016) illustrate that

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

Source: PASS data, own calculations

important policy measures such as the poverty rate can be substantially biased if heaping in the reported income is not taken into account.

4.1 Analyzing rounded data

Starting with Sheppard (1898) several methods have been proposed to account for rounding at the analysis stage (see for example Hanisch 2005 or Schneeweiss/Komlos/Ahmad 2010 for a review). However, most of the rounding literature assumes symmetric rounding intervals that can be derived directly from the reported value. For example, if distance is reported in kilometers, it is assumed that the true distance must be in the interval *reported distance* $\pm$ *500 meters*. However, this does not generally hold for heaping. As illustrated below, the rounding interval can not be inferred directly with data affected by heaping.

Instead of accounting for the rounding at the analysis stage multiple imputation methodology can be used to account for the rounding at the data processing stage. A multiple imputation strategy to obtain plausible values for the true values based on the reported values accounting for the uncertainty from rounding was first proposed by Heitjan/Rubin (1990) for age data affected by heaping. Related approaches were later used for self-reported cigarette counts (Wang/Heitjan, 2008), rounded unemployment durations (van der Laan/Kuijvenhoven, 2011) and self-reported income (Drechsler/Kiesl/Speidel, 2015; Drechsler/Kiesl, 2016; Zinn/Würbach, 2016).

4.2 Methodology of multiple imputation for data affected by heaping

There is an important difference between interval observations treated in Section 3 and rounded observations: With interval observations the interval in which the true value must lie is known. This is not the case for rounded observations. For example, if the reported income is 1,800, we do not know whether this is the exact true value, or if the true value has been rounded to the closest 5, 10, 50, or 100. To account for this uncertainty, we also need to model the rounding process.

The methodology presented in this section is based on the ideas first discussed in Heitjan/Rubin (1990). We summarize the main ideas of the approach here borrowing heavily from Drechsler/Kiesl (2016). For further details we refer to Heitjan/Rubin (1990) or Drechsler/Kiesl (2016).

To be able to account for the heaping in a variable, two models need to be specified: one model for the variable of interest and one model for the rounding behavior. Let Y be the variable of interest. Similar to Section 3 we assume that the conditional distribution of Y

given some covariates X is given as

$$Y|X \sim N(X\beta, \sigma^2) \tag{7}$$

To model the rounding behavior, an ordered probit model can be specified, i.e., a normally distributed latent variable G is assumed which may (linearly) depend on Y and some covariates Z (where some or all components of Z might be in X and vice versa):

$$G|Y, Z \sim N(\gamma_0 + Y\gamma_1 + Z\gamma_2, \tau^2). \tag{8}$$

The thresholds of the ordered probit model separate the different degrees of rounding. For example, if the assumed possible degrees of rounding are 1, 10, 50, and 100, an ordered probit model with four categories would be estimated.

Based on these model assumptions, the joint distribution of Y and G can be specified. The set of parameters to be estimated is given by $\Psi = (\beta, \sigma^2, \gamma_1, \gamma_2, k_1, ..., k_{p-1})$, where $k_1, ..., k_{p-1}$ denote the thresholds of the probit model assuming p possible degrees of rounding (note that γ_0 is fixed at 0 and τ^2 at 1 to make the ordered probit model identifiable). For each individual i, $i = 1, \ldots, n$, with n being the sample size, let s_i denote the rounded value which is observed instead of the true y_i, and $s = (s_1, \ldots, s_n)$. The likelihood function for Ψ given s_i and covariates x_i, z_i (assuming independent observations) may then be written as

$$\begin{aligned} L(\Psi|s, x, z) &= \prod_i f(s_i|x_i, z_i, \Psi) \\ &\propto \prod_i \iint_{A(s_i)} f(g, y|x_i, z_i, \Psi) dy dg, \end{aligned} \tag{9}$$

where $A(s_i)$ is the set of (g, y) that are consistent with an observed s_i. The parameter vector Ψ can be estimated by maximizing $L(\Psi|s, x, z)$ using numerical methods.

To generate imputations of Y, the first imputation step (drawing a new set of parameters from their joint posterior distribution) can again be approximated by drawing from

$$\Psi^\star \sim MVN(\hat{\Psi}, I(\hat{\Psi})), \tag{10}$$

where $\hat{\Psi}$ contains the maximum likelihood estimates of Ψ, and $I(\hat{\Psi})$ is the negative inverse of the Hessian matrix of the log-likelihood with $\hat{\Psi}$ plugged in.

For the second imputation step (generating imputed values for Y) a simple rejection sampling approach is implemented:

1. Draw candidate values for (y_i^{imp}, g_i) from a truncated bivariate normal distribution using parameters from $\Psi^\star$, where the truncation points are given by the maximal possible degree of rounding given the observed value s_i (for example, for an observed income value 850 with possible degrees of rounding 1, 10, 50, 100, and 1,000, y_i is bounded by 825 and 875 and g_i has to be in $]-\infty, k_3^\star[$).

2. Accept the drawn values for y_i as imputation value if they are consistent with the observed rounded value, i.e., when rounding the drawn value for y_i according to the drawn rounding indicator g_i gives the observed value s_i.

3. Otherwise draw again.

4.3 Our contribution for the imputation of data affected by heaping

The `R` package `simPop` (Templ et al., 2017) provides a function for generating plausible values if heaps only occur at multiples of 5 or 10. However, no other rounding degrees can be considered and no covariates can be incorporated into imputation model. `hmi` provides a more general imputation routine for variables affected by heaping following the methodology presented above. With `hmi` flexible degrees of rounding can be specified and covariates can be incorporated in both, the model for the rounding process and the imputation model. The package will declare variables to be affected by heaping if certain criteria are met, but it is also possible for the user to manually decide, which variables are affected. For details how to register variables accordingly see Section 5.1 and the *Rounded continuous variables* paragraph in Section 5.5.

It is also possible to use `hmi` for dealing with situations in which missing observations, interval observations and rounded observations occur simultaneously. This will typically be the case for surveys asking for income or other sensitive questions. Since nonresponse to the income question tends to be high, it is common practice to ask respondents whether their income lies in certain intervals if they are unwilling or unable to provide exact income values. In this situation three potential outcomes are possible: the respondent remains unwilling to provide any information at all and thus the income value is missing. Alternatively, the respondent might not provide an exact value but might be willing to indicate an interval in which his or her income lies. Finally, the respondent might report a supposedly exact value, which considering Table 1 will still be a rounded estimate of the true income in many cases. To deal with such a situation the likelihood function in Equation (9) needs to be extended to also account for the interval information:

$$L(\Psi|s,x,z) \propto \prod_{i=1}^{n} \Big\{ (1-I_i) \iint\limits_{A(s_i)} f(g,y|x_i,z_i,\Psi) dy dg + I_i \Big[F(\overline{y}_i, \mu_i = x_i'\beta, \sigma^2) - F(\underline{y}_i, \mu_i = x_i'\beta, \sigma^2) \Big] \Big\} \tag{11}$$

Imputed values for the interval data can be obtained by drawing from a truncated distribution as described in Section 3. See Drechsler/Kiesl/Speidel (2015) for an application and for further details regarding the imputation procedure. To our knowledge, `hmi` is the only imputation routine which is able to simultaneously impute rounded, missing and interval observations.

5 Software

The main function of the package hmi is the wrapper function called hmi. It performs all input checks, data preparations, calls of different imputation functions depending on the type of variable to be imputed and generates the output. In the simplest case the user just passes her or his data to hmi. In this case all variables with missing values are imputed based on a single level imputation model including all other variables in data as predictors. Under this scenario, the package works similar to other multiple imputation packages in R such as mice or mi (Su et al., 2011). The full flexibility of the package is unleashed, if the user additionally passes her or his (multilevel) analysis model to hmi and/or makes further specifications.

5.1 Input

These are the arguments which can be specified with hmi:

- data: The (partially observed/rounded) data set specified as a data.frame. Data in the matrix format are converted into a data.frame. For multilevel imputation the data have to be in the *long format*, meaning that observations belonging to the same cluster have to be stacked in rows and a cluster indicator needs to be available. Data in the *wide format* have to be converted to the *long format* using for example the packages reshape2 (Wickham, 2007) or tidyr (Wickham/Henry, 2018).
- model_formula: This argument requires a formula representing the desired analysis model which should be run once the data have been imputed. If model_formula is specified, hmi will try to set up imputation models which are in line with this model. In the multilevel case model_formula is used to identify fixed effects and random effects covariates and the cluster indicator. See Section 5.3 for details.
- family: A family object supported by glm (resp. glmer). This argument is not needed in the imputation process, it only facilitate the automated pooling (see Section 5.9) when the dependent variable in model_formula is not continuous. For example, for count data the appropriate call would be family = "poisson". Setting the family argument will ensure that the correct model is used when hmi calculates the appropriate multiple imputation inferences for the specified analysis model.
- additional_variables: With this argument the user can specify variables (separated by +, e.g. "x8 + x9") which should be included in the imputation models beyond those variables already included in the analysis model as specified in model_formula. Instead of using additional_variables the user might extend the model_formula and run a reduced analysis model with hmi_pool (or use the analysis tools provided by mice).
- list_of_types: If users are not satisfied with the automatic classification of the variable types by hmi (see Section 5.5), they can specify a list containing their own classifications. For example a user might want to treat a variable as continuous while

it was identified as count data (imputations would then be based on a linear regression model in this case instead of the Poisson model which is the default for count data). The explicit specifications in `list_of_types` are binding for `hmi` and overrule all other implicit specifications in any other attribute. For example, only missing values will be imputed in a variable specified to be continuous even if rounding degrees and/or a rounding formula are specified for this variable. To change this, the variable would need to be explicitly specified as rounded continuous in `list_of_types`. The list contains elements, named like the variables. Each element is a character of one keyword (e.g. `list_of_types = list(x1 = "cont", x2 = "categorical")`) to denote the imputation routine that shall be used for this variable. See Section 5.5 for all supported keywords and Section 5.6 for more explanations about the pre-definition of the variable types and Section 6.1 for an real data example.

- `M`: The number of imputed data sets that should be generated. The default value is 5.
- `maxit`: Similar to `mice`, `maxit` defines the number of cycles of the sequential regression imputation procedure that should be run before one imputed data set is stored (see also Section 2.5). The default value is 10, unless only one variable needs to be imputed. In this case the number of iterations is set to 1 as no updating of other variables is required.
- `nitt`: An integer defining the number of iterations that should be used for the Gibbs sampler whenever a variable is imputed using multilevel models based on the MCMC routines implemented in the package `MCMCglmm` (Section 2.4). Higher values imply a higher chance of convergence, but also increase the runtime of the imputation process. Convergence can be checked after imputation using the function `chaincheck` (see Section 5.8 for details). By default 22,000 iterations are run.
- `burnin`: An integer defining the number of MCMC draws of the `MCMCglmm` routines to be discarded as burn in. Higher values increase the chance of drawing values from a chain that has converged, but `burnin` has to be strictly lower than `nitt`. Furthermore a sufficient number of draws (say 1,000) should remain after discarding the burn in order to be able to effectively test convergence of the chain after the imputation run. The default value is 2,000.
- `pvalue`: By default `hmi` tries to include all variables as predictors in the imputation model. This can lead to unstable parameter estimates if the number of predictors is large. As a consequence imputations can vary erratically generating implausible imputed values way outside the observed range of values. A strategy to limit this problem is to exclude insignificant variables from the imputation model via a variable selection procedure (this strategy is also implemented in the multiple imputation software `IVEware`). If specified, the package `hmi` uses a backward selection procedure to identify the final imputation model: In the first step a (multilevel generalized) linear model is estimated using all variables as predictors. In the next step a new regression model is estimated such that the variable with the highest p value above `pvalue` is removed. This is repeated until each variable included in the model have a p value smaller or equal to `pvalue` or until only one variable remains in the model. Excluding

insignificant variables stabilizes the imputation process in most situations, but will typically bias the (conditional) correlation between imputed and excluded variables towards zero in any analysis performed on the imputed data. Therefore we advise to use this option conservatively, that is, we recommend generating imputations using the default value (i.e., `pvalue = 1`, which means no variables are removed). Lower values – say, 0.5 or 0.2 – can be specified, if the imputations based on the default setting show unacceptably large variances. We also note that variables are automatically removed if their effect cannot be estimated, that is, if the estimated coefficient is `NA`.

- `mn`: Estimating cluster specific parameters based on very few observations can lead to unstable estimates. As an ad hoc approach the user can specify a minimum number (`mn`) of observations a cluster should contain. The smallest clusters with less then `mn` observations will then be collapsed with the second smallest cluster until all clusters have at least `mn` observations. As this approach violates the assumption of independent normally distributed cluster effects and the individual effects of the collapsed clusters will no longer be reflected in the imputed data, this approach should be used with caution. The default value is 1, leading to no collapsing.

- `k`: Categorical variables with many categories can lead to unstable estimates since a large number of dummy variables needs to be included in the imputation model and some categories might be sparsely populated. To avoid such problems, `k` gives the maximum number of categories a categorical variables is allowed to have when used as covariate in an imputation model. Variables with more than `k` categories will be excluded from all imputation models. By default the number is ∞, leading to no removal. A less restrictive solution to avoid unstable estimates is to prevent the inclusion of insignificant dummy variables in the imputation model by setting an appropriate values for `pvalue`. In some situation it could be acceptable to classify ordinal variables with many categories as continuous in `list_of_types`.

- `spike`: This argument accepts a single numeric value or a list for which the names of the list entries match the names of semi-continuous variables (variables which have a spike at one value of the distribution but can be considered continuous otherwise). By setting `spike` to be an integer, the user can specify at which value the spike(s) might be found in the variable(s). In many cases, a spike will be found at zero, for example if a household survey asks for the taxes payed or a business survey asks for the number of employees hired in the previous year. However, there could be situations in which a spike occurs at a different value. For example, responses regarding the monthly net income will typically have a spike at the social security transfer level. In cases of different spikes for different variables, the parameter `spike` should be a list. For example, if `x2` has a spike at 0 and `x7` has a spike at 416 (which is minimum amount of social security payments in Germany), the attribute would need to be specified as `spike = list(x2 = 0, x7 = 416)`. The function `list_of_spikes_maker` can be used to generate such a list with suggested spikes (returning the mode for all variables for which more than 10 percent of the values are equal to the mode). This list can be adopted according to the needs of the user and then passed to `hmi` via the `spike` attribute. If `spike` contains a list, the names in the list implicitly define

which variables should be treated as semi-continuous, that is, there is no need to additionally register the variables as `semicont` in `list_of_types`. However, if a different variable type is explicitly provided in `list_of_types` for this variable, the variable is treated according to this type since explicit specifications in `list_of_types` dominate any implicit specifications through any of the other attributes. The *Semi-continuous variables* paragraph in Section 5.5 describes the heuristic used to decide whether a variable should be treated as semi-continuous if neither a numeric value nor a list is specified. It also provides details how semi-continuous variables are imputed.

- `rounding_degrees`: If the user wants to generate plausible values for variables affected by heaping following the methodology described in Section 4, she or he can specify the rounding degrees which should be included in the model. The argument can either be a single numeric vector or a list for which the names of the list entries match the names of the variables affected by heaping. In this case each element of list contains a numeric vector specifying the various rounding degrees. For example if the age of children is reported in months, heaps might occur at multiples of 1, 6, or 12 while the monthly income might be rounded to multiples of 1, 10, 100, or 1,000. If plausible values should be generated for both variables, the user would need to specify `rounding_degrees = list(age = c(1, 6, 12), income = c(1, 10, 100, 1000))`. The function `list_of_rounding_degrees_maker` generates such a list with individually suggested rounding degrees for each variable found to be affected by heaping. This list can be adapted by the user according to his or her needs. See the *Rounded continuous variables* paragraph in Section 5.5 for details regarding when a variable is considered to be heaped and what rounding degrees are used in which scenarios. In Section 6.3 a data example on imputing variables affected by heaping is given.

- `rounding_formula`: For heaped continuous variables the user can specify a formula for the rounding process, that is he or she can specify, which predictor variables should be included in Equation (8). The standard `formula` notation should be used but no dependent variable needs to be specified. To give an example, the formula specification could be `~y + x2 + x15`, where `y` represents the variable affected by rounding and `x2` and `x15` are two other variables from the data set. Again, the argument can either be a formula or a list with element names identical to the names of the heaped variables. In the letter case each list element must contain a formula for the rounding process. The function `list_of_rounding_formulas_maker` generates such a list. This list can be adapted by the user according to his or her needs. The default formula is `~.`, meaning that all variables are included as main effects in the model for rounding. We note that maximizing the likelihood in Equation (9) is tricky since the boundaries of the integrals also need to be estimated. If the rounding model is too complex or if too many rounding degrees are specified, the iterative procedure for maximizing the likelihood might not converge. The function `hmi` will issue a warning whenever the optimizer did not converge or when the Hessian matrix of the maximum likelihood procedure cannot be inverted (which is typically a strong indication of numerical problems of the estimation procedure). In such cases, we generally

recommend to either drop predictors from the `rounding_formula` or discard some of the specified `rounding_degrees`.

- `pool_with_mice`: As long as `pool_with_mice` is set to be `TRUE`, which is the default, `hmi` internally uses the functions from `mice` to obtain the final results for the analysis model specified in `model_formula`. The results are returned as an additional attribute called `pooling` within the output object. Note that the output object generated by `hmi` differs from the output generated by `mice` in this case. This can be avoided if `pool_with_mice` is set to `FALSE`. Currently, the synergy of `hmi` and `mice` supports the automatic calculation of the final inferences for (generalized multilevel) linear analysis models. The default pooling of (ordered) categorical variables, is not supported, except for categorical variables in the single level case. A more flexible, but somewhat inconvenient function for pooling is `hmi_pool`, which is delivered with this `hmi` package (see Section 5.7 for details).

5.2 Checks and preparations

The package `hmi` runs several initial checks before starting with the actual imputation:

- All inputs are checked to ensure correct formating (e.g., `data` must be set up as a `data.frame`, many other attributes must either contain a list or a vector of numeric values, etc). See `?hmi` or the previous section for details on the attribute specifications.
- If any of the variables included in `data` has more than 90 percent missing values, the program asks the user whether he or she wants to keep this variable or to quit the program to adjust the `data` accordingly.
  - Variables which are completely missing will cause a warning; they do not contain any information and will not be imputed.
  - Observations with missing values for all variables will also cause a warning for the same reasons.
- Variables included in `model_formula` which are not in `data` will cause an error. Note that `hmi` currently only supports two levels of hierarchy in the multilevel imputation models. Thus, only one cluster ID can be specified in `model_formula`.
- If a multilevel model is specified in `model_formula` but less than three clusters are found, the user is asked to run a single level imputation or to process the data in a different manner.
- If a multilevel model is specified in `model_formula` and the cluster variable contains missing values, the user is asked whether those should be removed (recommended), categorically imputed (not recommended) or the imputation process should be canceled.
- If `nm` is specified, clusters with less than `nm` observations are collapsed (see Section 5.1 for details).

The following additional preparing steps are taken for each imputation model during the imputation process:

- If more than one constant variable is included in the imputation model, only one is kept to avoid multicollinearity. For the same reason one variable is dropped from multilevel imputation models of unordered categorical variables, whenever two predictor variables are highly correlated ($\rho > 0.99$).
- If a value for `k` is specified, categorical variables with more than `k` categories are removed from the current imputation model (see Section 5.1 for details).
- If a value for `pvalue` is specified, variables with p values larger than `pvalue` are removed from the current imputation model in an iterative procedure (see Section 5.1 for details).
- During the first imputation cycle, interval variables are treated as factors whenever they appear as covariates in one of the imputation models, until they have been imputed themselves: Once they have been imputed, the plausible values are used as predictors instead of the interval information. If there are many unique intervals in an interval variable, the user may consider setting a limit for the maximum number of allowed factors using the attributes `k`.

5.3 `model_formula`

In the single level case, the model specified in `model_formula` has to follow standard `formula` conventions for `lm` in `R` (see `?formula`). For multilevel models the notation used by `lmer` (`lme4` package by Bates et al. 2015) must be used. The notation for multilevel models as implemented in `lme4` closely follows the notation for single level models with the main difference that random effect variables are added in parentheses. The cluster identifier is also included within the parentheses separated from the random effect variables(s) by a vertical bar. To illustrate, a possible model specification might be `y ~ x1 + x2 + x3 * x4 + (1 + x2|ID)`. In this model an intercept, four main effects and one interaction are specified as fixed effects. The intercept and `x2` also have random effects. The variable `ID` contains the cluster identifier.

If interactions are specified in `model_formula`, they are also used as predictors in the imputation models of all other variables in an effort to achieve congeniality. Note that the package currently does not follow the sophisticated approach suggested by Carpenter/ Goldstein/Kenward (2011) for dealing with interactions in the analysis model, instead it uses passive imputation meaning that after each iteration the interaction term is updated by multiplying the current imputed versions of the main effects (cf. e.g. Seaman/Bartlett/White 2012).

5.4 Imputation cycles

In the first cycle of the sequential regression imputation routine, the variables are sorted and imputed by increasing number of missing observations following the approach of

Raghunathan et al. (2001). In this cycle only those variables with no missing values or variables that have been previously imputed are used as predictors in the imputation model. If all variables have missing values, the variable with the lowest missing rate will be imputed by taking random samples from the observed cases of this variable. In all other imputation cycles, all variables are included as main effects in the imputation model, unless `pvalue` is specified. If `model_formula` is specified, the imputation model follows this model as closely as possible. This implies that the imputation and analysis model coincide when the dependent variable in the analysis model needs to be imputed. If, on the other hand, a covariate in the analysis model needs imputation, this variable takes the place of the dependent variable in the imputation model and the actual dependent variable in the analysis model becomes an independent variable in the imputation model. For example, if the analysis model is `y ~ 1 + x1 + x2 + (1 + x1|ID)` and the covariate `x1` needs imputation, the imputation model becomes `x1 ~ 1 + y + x2 + (1 + y|ID)`.

Depending on the situation the imputation model can either be a single or multilevel model. If `model_formula` contains a single level model, or when no analysis model is specified, the imputation model always will be a single level model. However, specifying a multilevel model in `model_formula` generally implies that a multilevel model will also be used for all imputation models. In the first cycle it can happen that the random effect covariate(s) have missing values. In such cases single level models are estimated until the random effect covariates(s) have been imputed. If the cluster ID has missing values, we recommend to remove the missing cases from the data set. In case the user opt against this, the missing cases are imputed using a single level model for categorical variables.

The number of cycles is defined by `maxit` unless only one variable contains missing values. In this situation, imputed values will be drawn from the correct distribution in the very first iteration (because all predictor variables are fully observed), and thus the number of iterations can be set to 1. The default number of imputation cycles, for situations with more than one missing variable, is 10. For a more cautious approach the user might set `maxit` to a larger value. After `maxit` cycles, the imputed values are stored, building a completed (imputed) data set. Then the process starts again, until `M` (default value: 5) imputed data sets have been generated.

5.5 The different supported types of variables

Different variable types (continuous, binary, etc.) require different imputation routines. For example, for binary variables it is not desirable in most cases to get imputed values different from `0` or `1`. And factor variables with levels `"A"`, `"B"` and `"C"` need an imputation routine different from the routines for binary and continuous variables.

The package `hmi` distinguishes nine different types of variables. The following section describes the internal strategies to assign a type to each variable and how the imputation model works for that type. Users not satisfied with these default choices can specify the types of variables in advance by setting up a `list_of_types`. Section 5.6 explains how this is done.

5.5.1 Binary variables (keyword "binary")

Variables are considered to be binary if there are only two unique values in the observed data. This includes for example 0 and 1 or "m" and "f". This default classification might fail for small data sets or if a third possible category is unobserved. For example, in a small health survey it could happen that non of the respondents reported to have had two (or more) Bypass surgeries. So here a count variable would falsely be classified as binary. (Multilevel) logistic regression models are used to impute binary variables.

5.5.2 Continuous variables (keyword "cont")

Any numeric vector that is not one of the other types is considered to be continuous. Imputation models are based on (multilevel) linear regression models described in Section 2.4.

5.5.3 Semi-continuous variables (keyword "semicont")

If a variable is not defined explicitly (via list_of_types) or implicitly (via an entry for this variable in spike), a variable is identified as semi-continuous by hmi if more than 10 percent of the observations share the same value (this value is then called *spike*), but the remainder of the observations can be considered continuous. To which spike the variable is tested, depends on the specifications in spike: if it is explicitly or implicitly defined, the value in spike is used (i.e. the numeric values of spike or the list element in spike for this variable - dependent on how spike is specified). If those elements are empty, hmi uses the mode (most frequent observation) of the variable, irrespectively of the 10 percent threshold. This threshold is only relevant if the variables are not explicitly or implicitly specified as semi-continuous. In theses cases, the mode, or spike if it is a numeric value, is used to check whether the 10 percent threshold is exceeded or not.

The approach for imputing semi-continuous variables implemented in hmi follows the ideas presented in Rubin (1987) and Raghunathan et al. (2001). The variable is imputed in two steps. In the first step a temporary indicator variable is generated that equals 0 if the observed value is equal to the spike and 1 otherwise. Missing values in this indicator variable are then imputed using (multilevel) logit models. In the second step, missing observations with an imputed value of 1 for the temporary indicator variable are imputed based on a (multilevel) linear regression imputation model, using only those observed cases of the semi-continuous variable that are not equal to the spike. The missing observations with an imputed value of 0 for the temporary indicator variable are replaced by the value of the spike.

5.5.4 Interval variables (keyword "interval")

Variables where some observations contain only interval information (e.g. $[2000; 3000]$) are called interval variables. The technical implementation requires a specification for interval data. To our knowledge there is no general technical standard for handling interval data in R. The packages survival (Therneau, 2018) and linLIR (Wiencierz, 2012) provide functionalities to handle interval data. Both packages generate auxiliary objects in which the information for the lower and upper bound are stored separately. We did not follow this approach for our package since it would require an inconvenient workflow to link both interval bounds (for all interval variables) appropriately. Instead we define a new class interval for interval variables. Technically each observation in such an interval variable is coded as "l;u" with l and u denoting the lower and upper bound of the interval. Both bounds can either be numerical values, NA, -Inf or Inf. Two examples would be "1234.56;3000" and "-1234.56;Inf".

We also implemented functions to run basic calculations on interval data (+, -, *, /, %%, exp, log, ^, sqrt, floor, ceiling, and round), to generate interval data based on one (as.interval) or two vectors (generate_interval), or to split interval data into their lower and upper bounds (split_interval). How to use these functions is illustrated in Section 6.2.

For interval variables, the imputation routine described in Section 3 is used. As mentioned in Section 5.2, interval variables are treated as factor variables during the first imputation cycle - until the variable itself has been imputed. Once plausible values have been generated for this variable, these imputed values will be used instead of the interval information in the following cycles whenever the (former) interval variable is used as a predictor in one of the other imputation models.

5.5.5 Rounded continuous variables (keyword "roundedcont")

Whether a variable is treated as "rounded continuous", (i.e., when the variable is affected by heaping), depends on the information contained in the attributes list_of_types, rounding_degrees and rounding_formula.

- list_of_types is always binding. If there is an entry in list_of_types for the variable, it will be imputed using imputation routines appropriate for the specified type irrespective of the information provided in any of the other attributes. Thus, if the variable is registered as roundedcont in list_of_types, it will be treated as affected by heaping irrespective whether potential degrees of rounding are specified in rounding_degrees or not. Vice versa, if the variable is registered to be of any other type, its missing values will be imputed using imputation methods appropriate for this variable type, but the heaping in this variable will be ignored even if rounding degrees are specified for this variable.
- If no explicit method is specified for the variable in list_of_types, hmi checks whether

rounding_degrees or a rounding_formula are specified for it, implying that the variable should be treated as rounded continuous.

- If no explicit or implicit classification is found, hmi classifies the variable internally. The classification tests for rounding degrees $1,\ 10,\ 100,\ 1,000$ or, if given, the general vector in rounding_degrees. A variable is classified as "rounded continuous" if more than 50 percent of the values in this variable are divisible by the specified rounding degrees (ignoring rounding to the nearest integer).

Variables classified to be rounded continuous (including variables having heaps, missing values and intervals at the same time) are imputed following the methodology described in Section 4. Which rounding degrees are used for generating plausible values depends on the provided specifications:

- For variables explicitly or implicitly specified to be rounded continuous, the information provided in rounding_degrees is decisive. If rounding_degrees contains a vector, the values of this vector are used for all variables specified to be affected by heaping. If it contains a list and this list has an element for the variable under consideration, the rounding degrees specified in this list element are used. If the list element or rounding_degrees is NULL, the heuristic explained in Appendix A.1 is used for suggesting rounding degrees.
- For variables classified by hmi as rounded continuous, the rounding degrees 1, 10, 100, 1000 or, if given, the general vector in rounding_degrees is used.

5.5.6 Count variables (keyword "count")

Except for variables which are identified to be semi-continuous all variables containing no more than 20 different integers are treated as count data per default. Variables with more than 20 integers are considered to be continuous to avoid treating continuous variables for which only integers are reported in the data (such as income data) as count data. The user can override these rules by simply specifying a variable with more than 20 different integers to be count or a variable with less than 20 integers to be cont in the list_of_types.

Imputations are generated based on a Poisson model for this variable type. MCMCglmm is used to obtain the required draws of the model parameters from their respective posterior distributions for both, single and multilevel models.

5.5.7 Categorical variables (keyword "categorical")

Unordered factor variables (or variables with more than two categories - if they are not one of the previous types) are considered to be categorical variables.

To impute these variables in a single level setting hmi uses the cart approach implemented in mice. The approach constructs a classification tree based on the observed data and

then samples imputed values from suitable leaves of this tree for individuals for which the variable is missing.

In the multilevel setting, we use the `categorical` specification in `MCMCglmm` to obtain draws of the model parameters from their posterior distribution based on a multilevel multinominal regression model. Imputations for the missing values are generated using own routines implemented in `hmi`.

5.5.8 Ordered categorical variables (keyword `"ordered_categorical"`)

If a factor variable is ordered, `hmi` treats it as `"ordered_categorical"`. Missing values in this variable are imputed based on an ordered logistic (for single level models) or ordered probit regression (for multilevel models). For single level models `mice` is used to generate the imputations. For multilevel models `MCMCglmm` is used to obtain the required draws of the model parameters from their posterior distribution and imputations are generated using own routines implemented in `hmi`.

5.5.9 Intercept variable (keyword `"intercept"`)

A variable for which all observed records share the same value is considered a constant variable and thus registered as an intercept variable. Missing values in this variable are replaced by the value observed for the other records.

If the user defines a `model_formula` containing an intercept variable (even if it is only implicit like in `y ~x1 + x2`) and there is no intercept variable in the data set, `hmi` temporarily includes such a variable for the imputation process. This can be suppressed by using `y ~0 + x1 + x2` or `y ~-1 + x1 + x2`. Vice versa, as mentioned in Section 5.2, if `model_formula` contains constant variables in addition to the intercept, these variables are automatically removed from the imputation model to keep the model identified.

5.6 Pre-definition of the variable types

The package `hmi` tries to make an educated guess, which imputation model is most suitable for which variable. Still, we encourage users to explicitly specify which imputation model should be used for each variable or at least to check whether the imputation models suggested by the package are reasonable. Imputation models for each variable can be specified using `list_of_types`. This attribute expects a `list` in which each element of it has the name of a variable in the data frame. The named element has to contain a single character string denoting the type of the variable (the keywords from the previous section). The user can pass her or his `data` to the function `list_of_types_maker` to see which imputation model would be suggested by `hmi` for which variable. Calling this function can also be useful to obtain an object which already contains a list with entries for all variables in the data set. This object can then be modified as required. Examples for generating and modifying this list are shown in Section 6.1.

We emphasize again that the specifications provided in `list_of_types` will dominate any other specifications. For example, if the attribute `rounding_degrees` contains specific degrees of rounding for variable `x11`, but this variable is specified as continuous in `list_of_types`, the variable will be treated like any other continuous variables, meaning that only the missing values in this variable will be imputed based on a (multilevel) linear regression model. No adjustments will be performed to deal with the heaps in the data.

5.7 Output of `hmi`

The package is build to allow a seamless integration into `mice`. Most importantly, the output generated by `hmi` can be treated like a multiply imputed data set generated with `mice`, that is, all the tools available in `mice` for analyzing and modifying the imputed data sets can be applied directly. The technical details regarding the structure of the `hmi` output are described here, practical examples are shown in the *Monitoring convergence* and *Analyzing the imputed data* paragraphs of Section 6.1.

Similar to `mice`, `hmi` returns a so called `mids`-object (multiply imputed data sets). These objects contain the original data set, the imputed values, the chain means and variance of the imputed values, and several additional elements (see van Buuren/Groothuis-Oudshoorn 2011). The fact that `hmi` returns a `mids`-object enables users familiar with `mice` to use functions designed for `mice`-outputs without switching barriers. For example, running the generic `plot()`-function on a `mids`-object calls the function `plot.mids` showing the means and standard deviations of the imputed values for all variables over the different imputations and cycles, regardless whether the `mids`-object came from `mice` or `hmi`. Another example is the `complete`-function delivered by `mice` which returns the imputed data set.

The function `hmi` returns two additional elements within the `mids`-object which are not available from `mice`: `gibbs` and `pooling`. The former allows checking the convergence of the gibbs-sampler chains generated by `MCMCglmm` (a convenient tool for checking convergence is available through the function `chaincheck`, see Section 5.8 for details). The later gives the pooled results (that is the final inferences based on the combining rules for multiply imputed data) from passing the `model_formula` to the pooling functions from `mice` (see Section 5.9 for details).

5.8 Convergence checks

For every imputed variable, the function `plot.mids` (delivered by `mice`) shows the mean and standard deviation of the imputed values across the `maxit` iterations and `M` imputation cycles. See Figure 1 in Section 6.1 as an example. This tool helps to evaluate whether draws based on the sequential regression approach converged to draws from the underlying joint distribution of the missing data given the observed data (see van Buuren/Groothuis-Oudshoorn 2011 for more details on this convergence measure).

If multilevel models are used for imputation (or if a Poisson model is used in general) additional convergence tests are necessary since the posterior draws of the model parameters

are obtained using a Gibbs sampler in these cases. Thus, we need to ensure that the Gibbs sampler actually converged before the parameters were drawn. Detailed information about all the MCMC chains from all models is available through the element `gibbs`. This is a multidimensional list. The first dimension distinguishes the different imputation runs. The elements in this layer are therefore called `"imputation1"`, `"imputation2"`, ..., `"imputation[M]"`. The second layer is for the cycles with names `"cycle1"`, ..., `"cylce[maxit]"`. The next layer is for the variable that has been imputed. For example, an element named `"x1"` stands for the imputation of `"x1"`. The last layer distinguishes between `"Sol"` and `"VCV"`. The names are adopted from `MCMCglmm` where the elements `"Sol"` and `"VCV"` in the output represent the point estimates (of the fixed effects and cluster specific effects) and the variance parameter estimates (the elements of the random effects covariance matrix and the residual variance), receptively. `hmi` only exports the fixed effects point estimates from `"Sol"` due to workspace considerations: `MCMCglmm` estimates `nitt` cluster specific effects for every random effects variable in every cluster. This would imply that if the user wants to run `nitt = 5000` iterations for a random intercepts and slopes model with only one fixed effects variable on a data set with 60 clusters, the dimension of the resulting matrix would already be $5000 \times (2+2 \cdot 60)$. If such a matrix would be saved for two variables and the imputation procedure is based on `maxit = 10` iterations and `M = 20` imputations, the final output would already contain $20 \times 10 \times 2 \times 5000 \times (2+2 \cdot 60) \approx 2$ million elements. Thus, to keep the size of the generated output manageable even if several variables are imputed based on multilevel models and/or the number of clusters is large, convergence can only be monitored for the fixed effects and the variance components.

To facilitate the convergence evaluations, the user can apply the function `chaincheck` to the output provided by `hmi`. The function implements the stationarity test proposed by Geweke (1992) and plots the results. The null hypothesis of the stationary test is that the expected values behind the means $\bar{x}_A$ and $\bar{x}_B$ of the first 10 percent and last 50 percent of the chain (after discarding the burn in) are equal. The test statistic for this test is $T = (\bar{x}_A - \bar{x}_B)/\sqrt{\hat{\sigma}(\bar{x}_A)^2 + \hat{\sigma}(\bar{x}_B)^2}$, where $\hat{\sigma}(\bar{x}_A)^2$ and $\hat{\sigma}(\bar{x}_B)^2$ are the estimated variances of the arithmetic means of the first 10 percent and last 50 percent of the chain after discarding the burn in. T asymptotically follows a standard normal distribution. So if $|T|$ exceeds the $1 - \alpha/2$ quantile of the standard normal distribution, the null hypothesis can be rejected. The test is implemented in the function `geweke.diag` from the `R` package `coda` (Plummer et al., 2006) and `chaincheck` calls this function. Beyond the `mids`-object generated by `hmi` the user can also pass the desired significance level `alpha` for the test statistic and the desired `burnin` (expressed as a percentage of the total length of the chain) to the `chaincheck` function. By default (`plot = TRUE`), `chaincheck` will plot all chains for which the null hypothesis was rejected. Each plot contains the information which parameter and which variable, in which cycle and imputation is depicted. Furthermore, the test statistic T is shown. Note that no adjustments are made for the multiple testing problem and thus a certain number of tests will show significant results ("chain did not converge") by chance (Type I error). For example in a setting with `maxit = 5`, `M = 5`, two variables to impute and an imputation model with 2 fixed effects and two random effects variables and a significance level of `alpha = 0.01`, the number of expected false positives is $5 \cdot 5 \cdot 2 \cdot (2+4+1) \cdot \alpha = 3.5$. The function `chaincheck` will print the actual and expected number of

failed test. Note that the test is only meant to highlight potentially convergence problems. The provided plots can then be used to decide, whether the identified chains really indicate problems of the Gibbs sampler.

For large numbers of chains and thus larger numbers of expected false positives, it might be more convenient not to plot the chains failing the convergence test. This can be done by setting the function parameter `plot = FALSE`. We note that users are free to use their own convergence diagnostics since results from all the chains are available in the `gibbs` attribute of the `mids`-object generated by `hmi`.

High autocorrelation can increase the number of false positives. The parameter `thinning` allows to increase the thinning of the Chains to reduce auto correlation (the default value is `1`). As a rule of thumb, the number of values in the chain should not fall below 1000. By setting `thinning = NULL`, the number of remaining values is set to be approximately 1000. Note that setting a value for `thinning` will not affect the imputation procedures. The parameters will only affect which chain values are used when computing Geweke's test.

If the Gibbs sampler apparently did not converge, (currently) a new call of `hmi` has to be initiated with an increased number of iterations for the Gibbs-sampler (parameter `nitt`).

5.9 Pooling

The functions `with` and `pool` from `mice` are flexible tools for analyzing and pooling multiply imputed data sets. `hmi` uses these functions to obtain the final results for the analysis model specified by `model_formula` and `family`. The results can be accessed in the `mids`-object through its element `pooling`. Currently, `mice` only pools global fixed effects of multilevel regression model. In some situations, other parameters such as the variance components from the different levels of the hierarchical model might be relevant for the user. Therefore `hmi` delivers the function `hmi_pool` as a flexible alternative to the functionality available in `mice`. The function needs two inputs:

1. the multiply imputed data set (the `mids` object created with `hmi` or `mice`) and
2. a predefined analysis function which takes a completed data set as input, and returns a vector with the desired complete data statistics (e.g. the regression coefficients or random effects variance estimates).

`hmi_pool` calculates the parameters defined in the analysis function on each of the completed data sets in the `mids`-object and averages them, that is `hmi_pool` will only provide point estimates but not their associated estimated variances. The pooling is only valid when averaging is reasonable. For example it would be invalid to pool factor loadings from factor analysis where the signs of loadings have no meaning (comparable to whether `"m"` or `"f"` is the reference category in a regression model). Examples how to use `hmi_pool` are given in the *Analyzing the imputed data* paragraph in Section 6.1 and on the help page `?hmi_pool`.

6 Application examples

To illustrate the generation of plausible values for multilevel data, interval data and variables affected by heaping three step-by-step examples from three real data sets are given.

6.1 Multilevel data

To illustrate the main functionality of the package `hmi`, we use the data set Gcsemv containing information on the General Certificate of Secondary Education (GSCE) in the UK. The data set, which was collected in 1989 and contains 1905 students in 73 schools, is one of the data sets used in Goldstein (2011). It is freely available on the website of the *Centre for Multilevel Modelling* (CMM) at the University of Bristol under the following URL `http://www.bristol.ac.uk/cmm/media/team/hg/msm-3rd-ed/gcsemv.xls`. It is also included in the package `hmi` to allow users to replicate the examples given in this section. We thank Harvey Goldstein and the CMM for allowing us to incorporate the data into the `hmi` package. The variables contained in the data set are described in Table 2. A more detailed description of the data can be found in Creswell (1991).

Table 2: Variables included in the Gcsemv data.

variable	description
`school`	School ID
`student`	Student ID within this school[1]
`gender`	Gender (0 = boy, 1 = girl),
`written`	(Numeric) score in a written questionnaire
`coursework`	(Numeric) score for a coursework

Source:
`http://www.bristol.ac.uk/cmm/learning/mmsoftware/data-rev.html#gcsemv`

6.1.1 Before starting imputation

If the package has not been installed previously, the very first step is to install the `hmi` package via `install.packages("hmi")`. Once the package has been installed it can be attached to the current session, and the `Gcsemv` data can be loaded. The code for these two steps is:

```
library("hmi")
data(Gcsemv)
```

A short summary of the data shows (among other information) that the data set has 202 missing values in the written exam covariate and 180 missing values in the coursework covariate. Thus, the missing rate in those variables is 10.6 percent and 9.4 percent respectively. There are no rows with missing values in both variables, so the number of incomplete observations in total is 382 or 20.0 percent.

[1] the student ID is not unique since students in different schools can have the same ID

```
summary(Gcsemv)
     school        student     gender      written          coursework
 68137  : 104   77     :  14   0: 777   Min.   : 0.625   Min.   :  9.259
 68411  :  84   83     :  14   1:1128   1st Qu.:37.500   1st Qu.: 62.963
 68107  :  79   53     :  13            Median :46.875   Median : 75.926
 68809  :  73   66     :  13            Mean   :46.798   Mean   : 73.435
 22520  :  65   27     :  12            3rd Qu.:55.625   3rd Qu.: 86.111
 60457  :  54   110    :  12            Max.   :90.000   Max.   :100.000
 (Other):1446   (Other):1827            NA's   :202      NA's   :180
```

A list containing the suggested variable types for each variable in the data set can be obtained by:

```
list_of_types_maker(Gcsemv)
$school
[1] "categorical"

$student
[1] "categorical"

$gender
[1] "binary"

$written
[1] "cont"

$coursework
[1] "cont"
```

If the user is not satisfied with the suggested types, he or she might save the list, modify it, and pass the modified list to `hmi`. For example, if `coursework` contained the average grade of every student and the user prefers to treat that variable as ordered categorical, he or she can type:

```
modified_list <- list_of_types_maker(Gcsemv)
modified_list$coursework <- "ordered_categorical"
```

The modified list would then be passed to `hmi` by setting the attribute `list_of_types = modified_list`.

6.1.2 Running the imputation

The next (optional) step is to set up the `model_formula`, i.e. the final model of interest which should be estimated based on the multiply imputed data (see Section 5.3). In the

example given below, interest lies in the influence of gender and performance in previous coursework on the written exam. The intercept and the effect of gender are allowed to vary across the schools. They are added as random effects in the `model_formula`.

```
model_formula <- written ~ 1 + gender + coursework + (1 + gender|school)
```

Now the data and `model_formula` can be passed to the wrapper function `hmi`. The results are saved in an object called `dat_imputed`. Note that for full reproducibility a seed for the pseudo-random number generator is specified. Since no value is specified for the number of imputations, the default number of `M=5` imputed data sets will be generated. `hmi` will provide a progress bar during the imputation process.

```
set.seed(123)
dat_imputed <- hmi(data = Gcsemv, model_formula = model_formula)
Imputation progress:
0%   20%  40%  60%  80%  100%
|----|----|----|----|
```

6.1.3 Monitoring convergence

Before running any analysis models on the newly generated `mids`-object, it is always a good idea to check the convergence of all imputation routines. Some examples of how to do this based on the output generated by `hmi` are presented in this section.

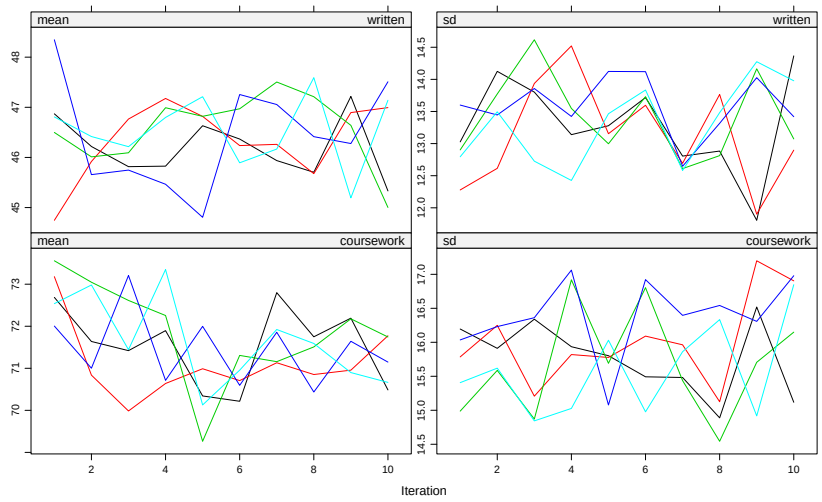
Diagnostic plots regarding the convergence of the sequential regression procedure can be obtained for example by `plot(dat_imputed)`. The command will plot the arithmetic mean and standard deviation of the imputed values for each imputed variable across the `maxit` cycles separately for each of the `M` imputations. In the given example calling the `plot` command will produce graphs for the variables `"written"` and `"coursework"` since these are the only two variables which have been imputed previously. Each graph contains five different lines for each of the `M = 5` imputations. Each line consists of ten points for each of the `maxit = 10` iterations.

```
plot(dat_imputed, layout = c(2, 2)))
```

Convergence (potentially after some burnin iterations) can be assumed if the following two points are fulfilled:

1. There is no inherent trend in any of the lines.
2. The lines from the different imputations mix well, i.e. there is sufficient overlap between the different lines.

Figure 1: Mean (left) and standard deviation (right) for the imputed variables in the Gcsemv data across 10 iterations for 5 imputations.



Source: Gcsemv data, own calculations

Examining the plots in Figure 1, both requirements seem to be met.

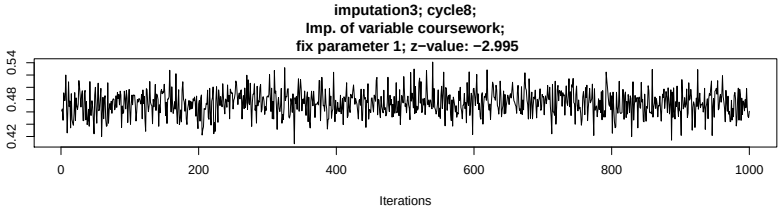
Given that the model specified in `model_formula` is a hierarchical model, multilevel models have also been used as imputation models. Since these models can only be estimated using MCMC methods, formal checks regarding the convergence of these models are also required. The function `chaincheck` runs convergence tests using the Geweke statistic for each chain of the MCMC method and plots traceplots for all those parameters for which the test indicates a failure of convergence (see Section 5.8 for details on the test). The function also provides the information how often the null hypothesis is rejected and compares this number to the expected number of false rejections due to type I error.

```
chaincheck(dat_imputed, thin = NULL)
12 out of 695 chains (1.73%) did not pass the convergence test.
For alpha = 0.01 the expected number is 6.95.
```

For the given example the traceplots for the fixed effects in the models which did not pass the stationarity test show no problematic pattern (one traceplot is shown in Figure 2 the others are omitted for brevity). But the plots for the variance parameters show signs of autocorrelation (one chain is shown in Figure 3). For highly autocorrelated chains it is more likely that the mean of the first 10 percent of the chain differs from the mean of the last 50 percent of the chain and thus the null hypothesis of the Geweke test (which basically assumes equivalence of the two means) is rejected. Note however, that autocorrelation would only be a problem, if multiple draws from the same chain would be used. Since only one value from a chain is used for each imputation in `hmi`, autocorrelation within a

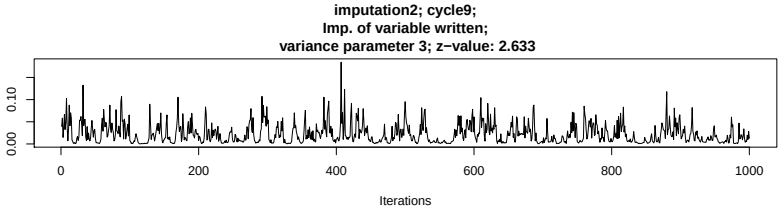
chain is generally irrelevant for `hmi`. Therefore it can be concluded, that for the package's purposes, all parameters in all imputation models show good convergence properties.

Figure 2: Traceplot of one fixed effects parameter which formally did not pass the stationarity test.



Source: Gcsemv data, own calculations

Figure 3: Traceplot of a variance parameter showing signs of high autocorrelation.



Source: Gcsemv data, own calculations

6.1.4 Analyzing the imputed data

In this section different possibilities for obtaining valid inferences based on the imputed data are shown. In general, valid inferences can be obtained by analyzing each completed data set separately and combining the results according to Rubin's combining rules (Rubin, 1987).

The package `mice` offers the functions `with` and `pool` to obtain final inferences based on the imputed data sets for a broad class of analyses. These functions can also be used with objects generated by `hmi` since they only require a `mids`-object as input. We refer to van Buuren/Groothuis-Oudshoorn (2011) for more details how to use these functions. Note that `hmi` also calls these functions internally if a model is specified in `model_formula` and `pool_with_mice = TRUE` (which is the default). The regression results are directly available through the element `pooling` from the `mids` object. This element is not available in `mids` objects generated by `mice`; it is a special feature of `hmi`. It will not be included if `pool_with_mice = FALSE`.

```
summary(dat_imputed$pooling)
                   est         se         t        df     Pr(>|t|)
(Intercept) 21.4285513 1.54661329 13.855145 228.08527 0.000000e+00
```

```
gender1     -5.4004356 0.59328192 -9.102647 153.58281 4.440892e-16
coursework   0.4042744 0.01919767 21.058509  64.10292 0.000000e+00
                lo 95     hi 95 nmis       fmi    lambda
(Intercept) 18.3810747 24.476028   NA 0.1306011 0.1230109
gender1     -6.5724822 -4.228389   NA 0.1642805 0.1534680
coursework   0.3659238  0.442625  180 0.2666766 0.2441485
```

However, `pool` can only be used with estimation commands that return a list of coefficients and their variance matrix. Thus, for example, no information is returned regarding the variance components on the different levels if `pool` is used to provide the results of a multilevel analysis. However, the estimated variances on the different levels can be of interest in some applications. For this reason `hmi` offers the option to pass an analysis function setup by the user to the function `hmi_pool` which will run the specified analyses on each imputed data set and return the final point estimates but not their variances. Thus, this function can be used in situations in which the variance of the point estimates cannot be estimated (or is not of interest to the analyst), but averaging the point estimates from the different data sets is still a valid approach.

In the following example, the user is interested in the global fixed effects and the elements of the random effects covariance matrix of the multilevel model from the running example. To obtain the final results, she or he would first need to specify the analysis function:

```
#The input of the function is a complete data set
#(which will be provided by hmi_pool later).

analysis_function <- function(complete_data){

  # Generate an empty list for storing the results of interest
  parameters_of_interest <- list()

  # Specify the analysis model of interest
  my_model <- lmer(written ~ 1 + gender + coursework + (1 + gender | school),
              data = complete_data)

  # Specify, which parameters from the model should be returned.
  # The fixed effects:
  parameters_of_interest[[1]] <- fixef(my_model)
  # The covariance matrix of the random effects:
  parameters_of_interest[[2]] <- VarCorr(my_model)[[1]][ , ]

  # Turn the list into a vector to simplify labeling:
  ret <- unlist(parameters_of_interest)

  # Optionally: label the output:
  names(ret) <- c("intercept", "gender", "coursework",
```

```
                    "sigma0", "sigma01", "sigma10", "sigma1")

  #Return the results.
  return(ret)
```

This function can then be passed to `hmi_pool` to obtain the final point estimates for the specified parameters. As the `analysis_function` in this example calls `lmer` from the `lme4` package, this packages has to be loaded in advance.

```
library("lme4")
hmi_pool(mids = dat_imputed, analysis_function = analysis_function)

 intercept     gender coursework     sigma0    sigma01    sigma10     sigma1
21.4285513 -5.4004356  0.4042744 42.3474332 -2.7057949 -2.7057949  3.1604561
```

The final results for the global fixed effects are identical to the results obtained with `mice`, but the output now also contains the final point estimates of the covariance matrix of the random effects.

6.2 Interval data

To illustrate the usage of the provided functions for `interval`-objects and the imputation of interval data, `hmi` includes three versions of a subset of the 2015-2016 Income File of the National Health and Nutrition Examination Survey (NHANES) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS), 2015-2016). The data set `nhanes_sub` (accessible by typing `data(nhanes_sub)` once the package is loaded) contains the data in their original format (compared to the version available on the NCHS website the data have been slightly modified, for example by coding some variables as factors or collapsing several nonresponse categories into a single category). In the data set `nhanes_mod` some variables have been changed to the internal interval variable format, which is required if plausible values should be imputed for these variables. Finally, `nhanes_imp` contains a multiply imputed data set in which missing and interval information has been replaced with plausible values following the methodology outlined in Sections 2 and 3. These data sets are included for illustrative purposes so that users of the package can compare different versions of the data sets to get a better understanding of how this imputation function works. Table 3 lists the variables present in the `nhanes` data sets.

As an illustrative example, the required steps to prepare the variable `ind310` for generating plausible values, that is, the transformation of the categorical variable from `nhanes_sub` to the interval variable in `nhanes_mod`, are presented here (the interval variable for `ind235` was generated in a similar fashion). Separate lower and upper bounds are defined for each observation (based on the description of `https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/INQ_I.htm`); subsequently they are merged to an interval object by the function `generate_interval`

Table 3: Variables included in the nhanes data sets.

variable	description
inq020	Income from wages/salaries (1 = Yes, 2 = No)
inq012	Income from self employment (1 = Yes, 2 = No)
inq030	Income from Social Security or Railroad Retirement (1 = Yes, 2 = No)
inq060	Income from other disability pension (1 = Yes, 2 = No)
inq080	Income from retirement/survivor pension (1 = Yes, 2 = No)
inq090	Income from Supplemental Security Income (1 = Yes, 2 = No)
inq132	Income from state/county cash assistance (1 = Yes, 2 = No)
inq140	Income from interest/dividends or rental (1 = Yes, 2 = No)
inq150	Income from other sources (1 = Yes, 2 = No)
ind235	Monthly family income (13 categories/an interval object)
ind310	Total savings/cash assets for the family (8 categories/an interval object)
inq320	How do you get to the grocery store? (10 categories)

Source: https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/INQ_I.htm

```
#generate an empty vector of length n
data(nhanes_sub)
low <- array(dim = nrow(nhanes_sub))
up <- array(dim = nrow(nhanes_sub))

#fill in the lower bounds depending on the reported savings category
low[nhanes_sub$ind310 == 1] <- 0
low[nhanes_sub$ind310 == 2] <- 3001
low[nhanes_sub$ind310 == 3] <- 5001
low[nhanes_sub$ind310 == 4] <- 10001
low[nhanes_sub$ind310 == 5] <- 15001
low[nhanes_sub$ind310 == 6] <- 0
low[nhanes_sub$ind310 == 7] <- 20001
low[nhanes_sub$ind310 == 8] <- 0

#fill in the upper bounds depending on the reported savings category
up[nhanes_sub$ind310 == 1] <- 3000
up[nhanes_sub$ind310 == 2] <- 5000
up[nhanes_sub$ind310 == 3] <- 10000
up[nhanes_sub$ind310 == 4] <- 15000
up[nhanes_sub$ind310 == 5] <- 20000
up[nhanes_sub$ind310 == 6] <- 20000
up[nhanes_sub$ind310 == 7] <- Inf
up[nhanes_sub$ind310 == 8] <- Inf

#generate the interval variable
ind310interval <- generate_interval(low, up)

#inspect the first few entries in the generated object
head(ind310interval)
"20001;Inf" "3001;5000" "0;3000" "3001;5000"  "0;3000" "3001;5000"
```

Once the variables are registered as `interval` variables, the data set can be passed to the `hmi` wrapper function. `hmi` will automatically generate plausible values for all variables registered as `interval` variables. For the imputation of the missing and interval data in `nhanes_mod`, we increased the number of iterations to 50, as diagnostic plots showed that the sequential regression procedure did not converge after the default number of 10 iterations.

```
set.seed(123)
nhanes_imp <- hmi(nhanes_mod, maxit = 50)
```

6.2.1 Some useful functions for interval data

The package `hmi` also includes some useful functions to analyze and manipulate interval data. This section provides a short summary of some of the functions available.

`table.interval`: Variables stored in interval format are interpreted as a vector of characters or a factor by most R functions including the `table` command. Without `table.interval`, `table` would order the intervals alphabetically, which can be arbitrary. The function `table.interval` offers improved sorting options. By default, it orders the intervals first by the value of their lower bound and if they are equal, by the value of the upper bound. If the attribute `sort` is set to `"mostprecise_increasing"`, the intervals are first ordered by their length (from small to large) and if the lengths are equal, by the value of the lower bound (from small to large). Using the `table` command on an interval variable will automatically invoke `table.interval` if `hmi` is loaded.

```
table(nhanes_mod$ind310)
```

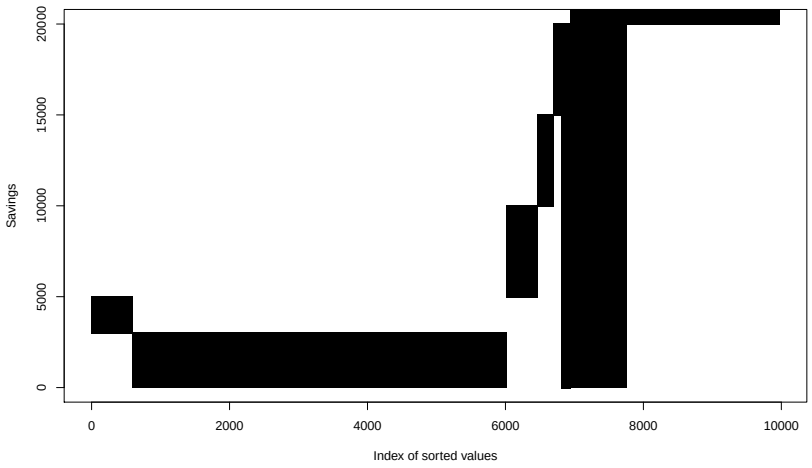
```
     0;3000      0;20000        0;Inf   3001;5000  5001;10000 10001;15000
       5426          128          814          588          450          237
15001;20000   20001;Inf
        110         2218
```

`plot.interval`: To inspect interval variables graphically, the generic plotting function `plot` can be used, which will call `plot.interval`. For example, Figure 4 containing the results for the savings variable from `nhanes_mod` is generated using the following code:

```
plot(nhanes_mod$ind310, ylab = "Savings", sort = "mostprecise_increasing")
```

The figure shows the interval values for `ind310` sorted first by the interval lengths and then by the lower bound. A second option is `sort = "lowerbound_increasing"` sorting the intervals first by the lower bound and then by the upper bound. If no argument is specified for `sort`, the intervals are sorted by their appearance in the data. For each observation the

Figure 4: An interval-data scatter plot.



Source: nhanes data, own depiction

plot draws a line from its lower to its upper bound (plus a small margin to make very small intervals and point precise observations visible). As the lines for observations sharing the same interval are grouped together, they form an area. Thus, the width of the area is an indicator for the relative frequency of this interval. Note that in the example the upper bound for the highest savings category and for the nonrespondents is ∞ which cannot be plotted. Therefore the upper limit of the y-axis by default is the highest finite bound observed (plus a small margin). The axis bounds can be manually altered by the parameters `xlim` and `ylim`.

`center.interval`: This function simply returns a numeric vector containing the midpoint of the reported interval for each observation (for example 1,500 if the interval is `"0;3000"`). Intervals including `Inf` or `-Inf` will return `Inf` or `-Inf`, unless the interval is `"-Inf;Inf"` or the parameter `inf2NA` was set to be `TRUE`. In those cases `NA` will be returned for these intervals. This function can potentially be useful for some descriptive statistics, but we caution the user that treating the midpoint of the reported interval as if it were the originally reported value is rarely a good idea.

```
midpoints <- center.interval(nhanes_mod$ind310)
table(midpoints)
x
   1500  4000.5  7500.5   10000 12500.5 17500.5     Inf
   5426     588     450     128    2371     110    3032
```

`idf2interval` and `interval2idf`: Interval variables are also accepted in some other `R` packages. For example, the package `linLIR` by Wiencierz (2012) provides methods for

regression models with interval variables. However when using this package, the data containing the interval information need to be coded as `idf` (*imprecise data frame*). To ensure that users can switch easily between `idf` and `interval` objects, we implemented `idf2interval` and `interval2idf` which convey an object from one format to the other. Technically, `idf` objects can contain multiple interval variables, so when transforming an `idf` object to fit to the `interval` setting, the (multiple) interval variables from `idf` are stored as variables in a `data.frame`.

```
idf <- interval2idf(nhanes_mod$ind310)
intervaldf <- idf2interval(idf)
```

`split_interval`: This function is basically the inverse function of `generate_interval`. It returns a two column matrix containing the lower bound for each reported interval in the first column and the upper bound in the second column:

```
bounds <- split_interval(nhanes_mod$ind310)
head(bounds)
```

```
      [,1] [,2]
[1,] 20001  Inf
[2,] 3001  5000
[3,]    0  3000
[4,] 3001  5000
[5,]    0  3000
[6,] 3001  5000
```

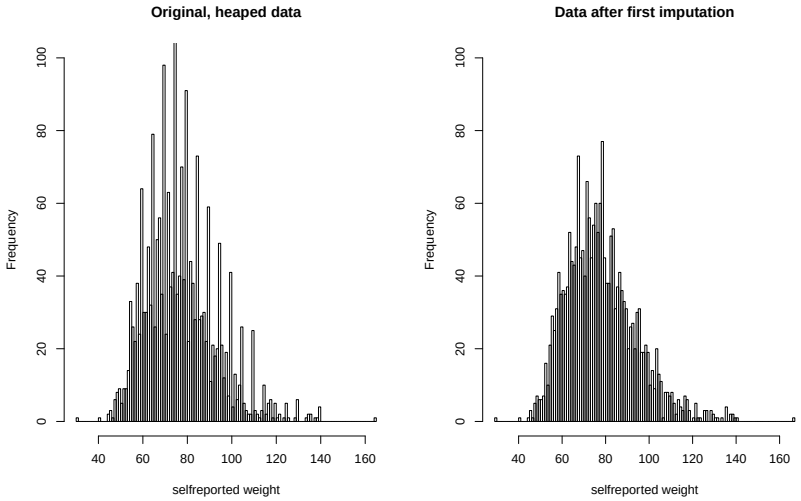
Finally, we note that basic arithmetics (`+`, `-`, `*`, `/`, `%%`) and transformations (`log`, `exp`, `^`, `sqrt`, `round`, `floor`, `ceiling`) can be applied to interval data (for example to change the currency for the reported values):

```
log_savings_in_euro <- log(nhanes_mod$ind310 * 0.8)
```

6.3 Variables affected by heaping

To briefly illustrate how to generate plausible values for a variable affected by heaping, we use the `selfreport` data from the `mice` package. The data set contains 2060 records and 15 variables, merged from multiple Dutch data sets. The left panel of Figure 5 shows a histogram of the self reported weight (variable `wr` in the data set). Heaps at multiples of 5 and 10 are clearly visible and thus, it seems plausible to assume that many respondents round their true weight to the closest 5 or 10 kilograms. Counting the number of records that are divisible by 5 and 10 reveals that almost 40 percent of the records are divisible by 5 and approximately 20 percent of the reported values are divisible by 10:

Figure 5: Selfreported weight from the selfreport data as originally observed (left) and after generating plausible values accounting for potential rounding of the reported values.



Source: selfreport data, own depictions and calculations

```
library("mice")
data(selfreport)
sum(selfreport$wr %%  5 == 0)/nrow(selfreport)
0.3800971

sum(selfreport$wr %% 10 == 0)/nrow(selfreport)
0.1941748
```

Note that these fractions are slightly below the thresholds setup in the heuristic for suggesting rounding degrees as implemented in `list_of_rounding_degrees_maker`. The heuristic would identify 5 as a rounding degree if 40 percent of the data would be divisible by this value and register 10 as a rounding degree if 20 percent of the data are divisible by this value (see Appendix A.1 for details). For this reason, explicit rounding degrees must be provided in this example when calling `hmi`. For the purpose of a short runtime, only two variables are used for imputation in this illustration: the self reported weight (`wr`) and the self reported height (`hr`):

```
set.seed(123)
selfreport_imputed <- hmi(selfreport[, c("hr", "wr")],
                          rounding_degrees = list(wr = c(1, 5, 10)))
```

By default, every variable in the data set is included in the model for the rounding behavior, that is, into the model specified in Equation (8). The model can be adjusted using `rounding_formula`. For example, if only the weight variable (and the intercept) should be

used in the rounding behavior model, this could be achieved by setting `rounding_formula = ~wr`. The right panel of Figure 5 shows the histogram after imputation. The heaps in the data have disappeared.

7 Conclusion

With `hmi` we provide comprehensive, but easy to handle tools for multiple imputation for hierarchical data sets. The package supports imputation methods for all common types of variables. Furthermore, imputation tools for interval and heaped variables are provided. Several internal features of the package ensure that sensible default settings are selected automatically. Thus, even inexperienced users will find the package convenient to use since all they need to provide is their data and potentially the analysis model they want to run on the imputed data. The final results (according to the given analysis model) will also be returned by default. Still, the package offers great flexibility since almost all settings can be defined manually if desired. Multiple imputation point estimates for analyses not supported in `mice` can also be obtained using an additional function provided with the package.

Currently, `hmi` still has some limitations which we hope to address in future releases of the package: Most importantly, the package does not provide any tools for imputing variables from the second level of the hierarchical model, that is, variables which are constant within clusters. A convenient tool for imputing such variables is available in `mice`. Furthermore, the multilevel imputation models are currently limited to two levels of hierarchy and homoscedastic error terms. Finally, ensuring that all Gibbs samplers of the multilevel imputation models have converged is currently left to the user. In future versions of the package, we hope to implement some routines that will automatically ensure that all chains run long enough to ensure convergence.

References

Anderson-Bergman, Clifford (2017): icenReg: Regression Models for Interval Censored Data in R. In: Journal of Statistical Software, Vol. 81, No. 12, p. 1–23.

Andridge, Rebecca R. (2011): Quantifying the Impact of Fixed Effects Modeling of Clusters in Multiple Imputation for Cluster Randomized Trials. In: Biometrical Journal, Vol. 53, No. 1, p. 53–74.

Asparouhov, Tihomir; Muthén, Bengt (2010): Multiple Imputation with Mplus. In: MPlus Web Notes, URL https://www.statmodel.com/download/Imputations7.pdf.

Audigier, Vincent; Resche-Rigon, Matthieu (2018): micemd: Multiple Imputation by Chained Equations with Multilevel Data. URL https://CRAN.R-project.org/package=micemd, r package version 1.2.0.

Bates, Douglas; Mächler, Martin; Bolker, Ben; Walker, Steve (2015): Fitting Linear Mixed-Effects Models Using lme4. In: Journal of Statistical Software, Vol. 67, No. 1, p. 1–48.

Bhat, Chandra R. (1994): Imputing a Continuous Income Variable From Grouped and Missing Income Observations. In: Economics Letters, Vol. 46, No. 4, p. 311–319.

Carpenter, James R.; Goldstein, Harvey; Kenward, Michael G. (2011): REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. In: Journal of Statistical Software, Vol. 45, No. 5, p. 1–14.

Carpenter, James R.; Kenward, Michael G. (2013): Multiple Imputation and its Application. John Wiley & Sons.

Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) (2015-2016): National Health and Nutrition Examination Survey Data. URL https://wwwn.cdc.gov/nchs/nhanes/.

Clogg, Clifford C.; Rubin, Donald B.; Schenker, Nathaniel; Schultz, Bradley; Weidman, Lynn (1991): Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression. In: Journal of the American Statistical Association, Vol. 86, No. 413, p. 68–78.

Creswell, Michael (1991): A Multilevel Bivariate Model. In: Prosser, Robert; Rasbash, Jon; Goldstein, Harvey (Eds.) Data Analysis with ML3, London: Institute of Education.

Czajka, John L.; Denmead, Gabrielle (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., U.S. Department of Health and Human Services, URL https://aspe.hhs.gov/system/files/pdf/75721/report.pdf.

Dorey, Frederick J.; Little, Roderick J. A.; Schenker, Nathaniel (1993): Multiple Imputation for Threshold-Crossing Data With Interval Censoring. In: Statistics in Medicine, Vol. 12, No. 17, p. 1589–1603.

Drechsler, Jörg (2015): Multiple Imputation of Multilevel Missing Data – Rigor Versus Simplicity. In: Journal of Educational and Behavioral Statistics, Vol. 40, No. 1, p. 69–95.

Drechsler, Jörg (2011): Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, Vol. 201. Springer Science & Business Media.

Drechsler, Jörg; Kiesl, Hans (2016): Beat the Heap: An Imputation Strategy for Valid Inferences from Rounded Income Data. In: Journal of Survey Statistics and Methodology, Vol. 4, No. 1, p. 22–42.

Drechsler, Jörg; Kiesl, Hans; Speidel, Matthias (2015): MI Double Feature: Multiple Imputation to Address Nonresponse and Rounding Errors in Income Questions. In: Austrian Journal of Statistics, Vol. 44, p. 59–71.

Enders, Craig Kyle; Keller, Brian T.; Levy, Roy (2017): A Fully Conditional Specification Approach to Multilevel Imputation of Categorical and Continuous Variables. In: Psychological Methods.

Enders, Craig Kyle; Mistler, Stephen Andrew; Keller, Brian T. (2016): Multilevel Multiple Imputation: A Review and Evaluation of Joint Modeling and Chained Equations Imputation. In: Psychological Methods, Vol. 21, No. 2, p. 222–240.

Gartner, Hermann; Rässler, Susanne (2005): Analyzing the Changing Gender Wage Gap based on Multiply Inputed Right Censored Wages. Tech. Rep., IAB-Discussion Paper 05/2005, URL http://doku.iab.de/discussionpapers/2005/dp0505.pdf.

Gelman, Andrew; Hill, Jennifer (2006): Data analysis Using Regression and Multilevel/Hierarchical Models. Cambridge university press.

Geweke, John (1992): Evaluating the Accuracy of Sampling Based Approaches to Calculating Posterior Moments. In: Bernando, J. B.; Berger, J. O.; Dawid, A. P.; Smith, Adrian F. M. (Eds.) Bayesian Statistics 4, Oxford, UK: Clarendon Press, p. 169–193.

Goldstein, Harvey (2011): Multilevel Statistical Models. Chichester (UK): Wiley, 4 ed..

Grover, Gurprit; Gupta, Vinay K. (2015): Multiple Imputation of Censored Survival Data in the Presence of Missing Covariates Using Restricted Mean Survival Time. In: Journal of Applied Statistics, Vol. 42, No. 4, p. 817–827.

Hadfield, Jarrod D. (2010): MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. In: Journal of Statistical Software, Vol. 33, No. 2, p. 1–22.

Hanisch, Jens U. (2005): Rounded Responses to Income Questions. In: Allgemeines Statistisches Archiv, Vol. 89, No. 1, p. 39–48.

Heeringa, Steven G. (1993): Imputation of Item Missing Data in the Health and Retirement Survey. URL http://ww2.amstat.org/sections/srms/Proceedings/papers/1993_013.pdf.

Heeringa, Steven G.; Little, Roderick J. A.; Raghunathan, Trivellore E. (1997): Imputation of Multivariate Data on Household Net Worth. URL http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1997_020.pdf.

Heitjan, Daniel F.; Rubin, Donald B. (1991): Ignorability and Coarse Data. In: The Annals of Statistics, Vol. 19, No. 4, p. 2244–2253.

Heitjan, Daniel F.; Rubin, Donald B. (1990): Inference from Coarse Data via Multiple Imputation with Application to Age Heaping. In: Journal of the American Statistical Association, Vol. 85, No. 410, p. 304–314, URL http://www.jstor.org/stable/2289765.

Huttenlocher, Janellen; Hedges, Larry V.; Bradburn, Norman M. (1990): Reports of Elapsed Time: Bounding and Rounding Processes in Estimation. In: Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol. 16, No. 2, p. 196–213.

Jenkins, Stephen P.; Burkhauser, Richard V.; Feng, Shuaizhang; Larrimore, Jeff (2011): Measuring Inequality Using Censored Data: A Multiple-Imputation Approach to Estimation and Inference. In: Journal of the Royal Statistical Society A, Vol. 174, No. 1, p. 63–81.

Jolani, Shahab (2018): Hierarchical Imputation of Systematically and Sporadically Missing Data: An Approximate Bayesian Approach Using Chained Equations. In: Biometrical Journal, Vol. 60, No. 2, p. 333–351.

Kennickell, Arthur B. (1996): Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances. URL http://ww2.amstat.org/sections/srms/Proceedings/papers/1996_073.pdf.

Kennickell, Arthur B. (1991): Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation. URL http://ww2.amstat.org/sections/srms/Proceedings/papers/1991_001.pdf.

Kim, Mimi Y.; Xue, Xiaonan (2002): The Analysis of Multivariate Interval-Censored Survival Data. In: Statistics in Medicine, Vol. 21, No. 23, p. 3715–3726.

Larrimore, Jeff; Burkhauser, Richard V.; Feng, Shuaizhang; Zayatz, Laura (2008): Consistent Cell Means for Topcoded Incomes in the Public Use March CPS (1976-2007). In: Journal of Economic and Social Measurement, Vol. 33, No. 2,3, p. 89–128.

Law, C. Gordon; Brookmeyer, Ron (1992): Effects of Mid-Point Imputation on the Analysis of Doubly Censored Data. In: Statistics in Medicine, Vol. 11, No. 12, p. 1569–1578.

Liu, Jingchen; Gelman, Andrew; Hill, Jennifer; Su, Yu-Sung; Kropko, Jonathan (2014): On the stationary distribution of iterative imputations. In: Biometrika, Vol. 101, No. 1, p. 155–173.

Lüdtke, Oliver; Robitzsch, Alexander; Grund, Simon (2017): Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies. In: Psychological Methods, Vol. 22, No. 1, p. 141–165.

Meng, Xiao-Li (1994): Multiple-Imputation Inferences with Uncongenial Sources of Input. In: Statistical Science, Vol. 9, No. 4, p. 538–573.

Mistler, Stephen A. (2013): A SAS Macro for Applying Multiple Imputation to Multilevel Data. In: Proceedings of the SAS Global Forum, URL https://support.sas.com/resources/papers/proceedings13/438-2013.pdf.

Muñoz, Alvaro; Wang, Mei-Cheng; Bass, Sue; Taylor, Jeremy M. G.; Kingsley, Lawrence A.; Chmiel, Joan S.; Polk, B. Frank; The Multicenter AIDS Cohort Study Group (1989): Acquired Immunodeficiency Syndrome (AIDS)-free Time After Human Immunodeficiency Virus Type1 (HIV-1) Seroconversion in Homosexual Men. In: American Journal of Epidemiology, Vol. 130, No. 3, p. 530–539.

Nowok, Beata; Raab, Gillian M.; Dibben, Chris (2016): synthpop: Bespoke Creation of Synthetic Data in R. In: Journal of Statistical Software, Vol. 74, No. 11, p. 1–26.

Pilcher, Christopher D.; Joaki, George; Hoffman, Irving F.; Martinson, Francis E.A.; Mapanje, Clement; Stewart, Paul W.; Powers, Kimberly A.; Galvin, Shannon; Chilongozi, David; Gama, Syze; Price, Matthew A.; Fiscus, Susan A.; Cohen, Myron S. (2007): Amplified Transmission of HIV-1: Comparison of HIV-1 Concentrations in Semen and Blood During Acute and Chronic Infection. In: AIDS, Vol. 21, No. 13, p. 1723–1730.

Plummer, Martyn; Best, Nicky; Cowles, Kate; Vines, Karen (2006): CODA: Convergence Diagnosis and Output Analysis for MCMC. In: R News, Vol. 6, No. 1, p. 7–11, URL https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf.

Quartagno, Matteo; Carpenter, James (2018): jomo: A package for Multilevel Joint Modelling Multiple Imputation. URL http://CRAN.R-project.org/package=jomo, r package version 2.6-1.

R Core Team (2016): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Raghunathan, Trivellore E.; Lepkowski, James M.; van Hoewyk, John; Solenberger, Peter (2001): A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models. In: Survey Methodology, Vol. 27, p. 85–96.

Raghunathan, Trivellore E.; Solenberger, Peter W.; Berglund, Patricia A.; van Hoewyk, John (2016): IVEware: Imputation and Variance Estimation Software. URL http://www.src.isr.umich.edu/wp-content/uploads/IVEware-Version-0.3-User-Guide-linked.pdf.

Rässler, Susanne (2003): A Non-Iterative Bayesian Approach to Statistical Matching. In: Statistica Neerlandica, Vol. 57, No. 1, p. 58–74.

Raudenbush, Stephen W.; Bryk, Anthony S. (2002): Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks: Sage Publications, Inc, 2 ed..

Reiter, Jerome P. (2012): Bayesian Finite Population Imputation for Data Fusion. In: Statistica Sinica, Vol. 22, No. 2, p. 795–811, URL http://www.jstor.org/stable/24310035.

Reiter, Jerome P.; Raghunathan, Trivellore E. (2007): The Multiple Adaptations of Multiple Imputation. In: Journal of the American Statistical Association, Vol. 102, No. 480, p. 1462–1471.

Reiter, Jerome P.; Raghunathan, Trivellore E.; Kinney, Satkartar K. (2006): The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data. In: Survey Methodology, Vol. 32, No. 2, p. 143–150.

Royston, Patrick (2007): Multiple Imputation of Missing Values: Further Update of ice, with an Emphasis on Interval Censoring. In: Stata Journal, Vol. 7, No. 4, p. 445–464, URL http://www.stata-journal.com/article.html?article=st0067_3.

Rubin, Donald B. (1988): Using the SIR Algorithm to Simulate Posterior Distributions. In: José Miguel Bernardo, Dennis Victor Lindley, Morris Herman DeGroot; Smith, Adrian Frederick Melhuish (Eds.) Bayesian Statistics, Vol. 3, Oxford University Press, p. 395–402.

Rubin, Donald B. (1987): Multiple Imputation for Nonresponse in Surveys. Hoboken (NJ): John Wiley & Sons.

Rubin, Donald B. (1986): Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. In: Journal of Business & Economic Statistics, Vol. 4, No. 1, p. 87–94, URL http://www.jstor.org/stable/1391390.

Rubin, Donald B. (1978): Multiple Imputations in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse. URL http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1978_004.pdf.

Schafer, Joseph L. (2016): pan: Multiple Imputation for Multivariate Panel or Clustered Data. URL https://CRAN.R-project.org/package=pan, r package version 1.4.

Schenker, Nathaniel (2003): Assessing Variability Due To Race Bridging: Application to Census Counts and Vital Rates for the Year 2000. In: Journal of the American Statistical Association, Vol. 98, No. 464, p. 818–828.

Schenker, Nathaniel; Raghunathan, Trivellore E.; Bondarenko, Irina (2010): Improving on Analyses of Self-Reported Data in a Large-Scale Health Survey by Using Information from an Examination-Based Survey. In: Statistics in Medicine, Vol. 29, No. 5, p. 533–545.

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, No. 475, p. 924–933.

Schneeweiss, Hans; Komlos, John; Ahmad, Amar S. (2010): Symmetric and Asymmetric Rounding: A Review and Some New Results. In: Advances in Statistical Analysis, Vol. 94, No. 3, p. 247–271.

Scott, S. Jeffrey; Jones, Richard A. (1990): Generation Means Analysis of Right-Censored Response-Time Traits: Low Temperature Seed Germination in Tomato. In: Euphytica, Vol. 48, No. 3, p. 239–244.

Seaman, Shaun Robert; Bartlett, Jonathan W.; White, Ian R. (2012): Multiple Imputation of Missing Covariates with Non-Linear Effects and Interactions: An Evaluation of Statistical Methods. In: BMC Medical Research Methodology, Vol. 12, No. 46, p. 1–13.

Sheppard, William Fleetwood (1898): On the Calculation of the Most Probable Values of Frequency-Constants, for Data arranged according to Equidistant Division of a Scale. In: Proceedings of the London Mathematical Society, Vol. s1-29, No. 1, p. 353–380.

Snijders, Tom Augustinus Benedictus; Bosker, Roel (2011): Multilevel Analysis - An Introduction to Basic and Advanced Multilevel Modeling. London: SAGE, 2nd revised edition. ed..

Speidel, Matthias; Drechsler, Jörg; Sakshaug, Joseph W. (2017): Biases in Multilevel Analyses Caused by Cluster-Specific Fixed-Effects Imputation. In: Behavior Research Methods.

Su, Yu-Sung; Gelman, Andrew; Hill, Jennifer; Yajima, Masanao (2011): Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. In: Journal of Statistical Software, Articles, Vol. 45, No. 2, p. 1–31.

Taljaard, Monica; Donner, Allan; Klar, Neil (2008): Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. In: Biometrical Journal, Vol. 50, No. 3, p. 329–345.

Taylor, Jeremy M. G.; Muñoz, Alvaro; Bass, Sue M.; Saah, Alfred J.; Chmiel, Joan S.; Kingsley, Lawrence A. (1990): Estimating the Distribution of Times from HIV Seroconversion to AIDS Using Multiple Imputation. In: Statistics in Medicine, Vol. 9, No. 5, p. 505–514.

Taylor, Jeremy M. G.; Schwartz, Kendra; Detels, Roger (1986): The Time from Infection with Human Immunodeficiency Virus (HIV) to the Onset of AIDS. In: The Journal of Infectious Diseases, Vol. 154, No. 4, p. 694–697, URL http://www.jstor.org/stable/30104955.

Templ, Matthias; Meindl, Bernhard; Kowarik, Alexander; Dupriez, Olivier (2017): Simulation of Synthetic Complex Data: The R Package simPop. In: Journal of Statistical Software, Vol. 79, No. 10, p. 1–38.

Therneau, Terry M. (2018): A Package for Survival Analysis in S. URL https://CRAN.R-project.org/package=survival, version 2.42-3.

Tobin, James (1958): Estimation of Relationships for Limited Dependent Variables. In: Econometrica, Vol. 26, No. 1, p. 24–36.

Trappmann, Mark; Gundert, Stefanie; Wenzig, Claudia; Gebhardt, Daniel (2010): PASS: a Household Panel Survey for Research on Unemployment and Poverty. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 130, No. 4, p. 609–622.

van Buuren, Stef (2012): Flexible Imputation of Missing Data. United States: Taylor & Francis Group.

van Buuren, Stef (2011): Multiple Imputation of Multilevel Data. In: Hox, Joop J.; Roberts, J. Kyle (Eds.) The Handbook of Advanced Multilevel Analysis, chap. 10, Milton Park, UK: Routledge Academic, p. 173–196.

van Buuren, Stef; Groothuis-Oudshoorn, Karin (2011): mice: Multivariate Imputation by Chained Equations in R. In: Jornal of Statistical Software, Vol. 45, No. 3, p. 1–67.

van der Laan, Jan; Kuijvenhoven, Léander (2011): Imputation of Rounded Data. Statistics Netherlands Discussion Paper no. 201108, Statistics Netherlands, URL `https://www.cbs.nl/-/media/imported/documents/2011/08/2011-x10-08.pdf`.

Wang, Hao; Heitjan, Daniel F. (2008): Modeling Heaping in Self-Reported Cigarette Counts. In: Statistics in Medicine, Vol. 27, No. 19, p. 3789–3804.

Wickham, Hadley (2007): Reshaping Data with the `reshape` Package. In: Journal of Statistical Software, Vol. 21, No. 12, p. 1–20, URL `http://www.jstatsoft.org/v21/i12/`.

Wickham, Hadley; Henry, Lionel (2018): `tidyr`: Easily Tidy Data with 'spread()' and 'gather()' Functions. URL `https://CRAN.R-project.org/package=tidyr`, r package version 0.8.0.

Wiencierz, Andrea (2012): `linLIR`: linear Likelihood-based Imprecise Regression. URL `https://CRAN.R-project.org/package=linLIR`, r package version 1.1.

Zhou, Hanzhi; Elliott, Michael R.; Raghunathan, Trivellore E. (2016): Synthetic Multiple-Imputation Procedure for Multistage Complex Samples. In: Journal of Official Statistics, Vol. 32, No. 1, p. 231–256.

Zhu, Jian; Raghunathan, Trivellore E. (2015): Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. In: Journal of the American Statistical Association, Vol. 110, No. 511, p. 1112–1124.

Zinn, Sabine; Würbach, Ariane (2016): A statistical Approach to Address the Problem of Heaping in Self-Reported Income Data. In: Journal of Applied Statistics, Vol. 43, No. 4, p. 682–703.

A Appendix

A.1 Suggestion for rounding degrees

If the user registers a variable as potentially being affected by heaping (by setting the variable type to `roundedcont`) but does not provide `rounding_degrees` for this variable, `hmi` tries to make an educated guess, regarding the possible degrees of rounding which should be used when modeling the heaping. The following heuristic is used to suggest the rounding degrees:

1. For a given continuous variable all possible rounding degrees (*factors* or *divisors* in mathematical terms), are derived for each observation. To give an example, the factors of 10 are $1, 2, 5, 10$. We will call $1, 2, 5$ *subfactors* of 10.
2. For each possible factor identified in step 1, the number of observations divisible by this factor is tabulated.
3. A rough estimate (based on the assumption of a discrete uniform distribution between 0 and ∞) for the expected number of observations being divisible by a factor s is n/s,

where n is the number of records in the data set. For example, the expected number of observations being divisible by $s = 5$ for a data set containing 10,000 records is $n/s = 2000$. If the observed number of individuals being divisible by factor s is at least twice the expected number, s is a "candidate rounding degree".

4. Starting with the highest candidate rounding degree, each candidate has to fulfill two conditions to be stored as an actual rounding degree.

   - At least 20 percent of the data have to be divisible by this candidate; observations which are also divisible by larger rounding degrees which has been previously identified to be an actual rounding degree are not considered. The removal of these records ensures that the currently considered candidate actually contributes to the heaping. For example when 40 percent of the data are divisible by 100, at least 40 percent of the data have to be divisible by 50. By requesting that at least 60 percent of the data are divisible by 50 (if 100 has been identified previously as an actual rounding degree) it is ensured that the fact that a large proportion of the data is divisible by 50 is not only a spurious effect because many observations are rounded to the closest 100.
   - The considered candidate must be a subfactor of at least two other factors found in the data. This prevents that a rounding degree only "explains itself". For example $4,000$ would not be considered to be an actual rounding degree if 27 percent of the individuals reported a value of $4,000$, but no one reported $8,000$ or $12,000$ etc. This condition ensures that lower (and thus more general) rounding degrees such as 1000 are favored.

Recently published

No.	Author(s)	Title	Date
2/2018	Knörr, M. Weber, E.	Labor markets and labor mobility in the French-German border region	1/18
3/2018	Teichert, C. Niebuhr, A. Otto, A. Rossen, A.	Graduate migration in Germany – new evidence from an event history analysis	2/18
4/2018	Osiander, C. Stephan, G.	Unter welchen Bedingungen würden sich Beschäftigte weiterbilden?	2/18
5/2018	Schropp, H.	Ressourcenorientierte Förderung von jungen Menschen im Übergangsmaßnahmen	2/18
6/2018	Schäffler, J. Moritz, M.	German FDI in the Czech Republic – Employment effects in the home country	2/18
7/2018	Fuchs, J. Weber, B.	Fachkräftemangel: Inländische Personalreserven als Altrnative zur Zuwanderung	2/18
8/2018	Wapler, R. Wolf, K. Wolff, J.	Do active labour market policies for welfare recipients in Germany raise their regional outflow into work?	3/18
9/2018	Wanger, S. Zapf, I.	For better or worse? How more flexibility in working time arrangements and fatherhood affect men's working hours in Germany	3/18
10/2018	Warning, A. Weber, E.	Digitalisation, hiring and personnel policy: evidence from a representative business survey	3/18
11/2018	Stepanok, I.	FDI and Unemployment, a Growth Perspective	3/18
12/2018	Knize, V.	Migrant women labor-force participation in Germany	4/18
13/2018	Schierholz, M.; Brenner, L. Cohausz, L.; Damminger, L.; Fast, L.; Hörig, A.; Huber, A.; Ludwig, T.; Petry, A.; Tschischka, L.	Eine Hilfsklassifikation mit Tätigkeitsbeschreibungen für Zwecke der Berufskodierung	5/18
14/2018	Janser,M.	The greening of job in Germany	5/18
15/2018	Dettmann, E. Weyh, A. Titze, M.	Heterogeneous effects of investment grants - Evidence from a new measurement approach	5/18

As per: 2018-05-22

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

Imprint

IAB-Discussion Paper 16/2018
5 June 2018

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

Editorial staff
Ricardo Martinez Moya, Jutta Palm-Nowak

Technical completion
Renate Martin

Website
http://www.iab.de

Download of this Discussion Paper
http://doku.iab.de/discussionpapers/2018/dp1618.pdf

ISSN 2195-2663

For further inquiries contact the authors:

Matthias Speidel
Phone +49.911.179.7956
E-mail Matthias.Speidel@iab.de

Jörg Drechsler
Phone +49.911.179.4021
E-mail Joerg.Drechsler@iab.de

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.2011, §8, Abs. 2 Pkt. .5.)

Hiermit versichere ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 04.12.2018	
Ort, Datum	Matthias Speidel