LIfBi WORKING PAPERS

Irina Hondralis and Elisa Himbert

AN APPLICATION OF MULTIPLE IMPUTATION USING NEPS SC1 DATA— A COMPARISON OF R AND STATA

LIfBi Working Paper No. 78
Bamberg, November 2018

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

An Application of Multiple Imputation using NEPS SC1 Data—a Comparison of R and Stata

Dr. Irina Hondralis[*]
Goethe-University Frankfurt am Main

Elisa Himbert[*]
Leibniz Institute for Educational Trajectories

E-mail address of lead author:

hondralis@soz.uni-frankfurt.de

[*] Both authors contributed equally to this work.

Bibliographic data:

Hondralis, I. & Himbert E. (2018). *An application of multiple imputation using NEPS SC1 data—a comparison of R and Stata* (LIfBi Working Paper No. 78). Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

An Application of Multiple Imputation using NEPS SC1 Data—a Comparison of R and Stata

Abstract

This working paper describes the application of multiple imputation in the context of survey data, where missing data are a common problem. Consequently, multiple imputation methods have become a widely used approach to handle incomplete data. Despite the pervasive use of multiple imputation, more detailed and practically oriented examples and illustrations are limited. Hence, we base our user-friendly application example on data from the National Educational Panel Study SC1 (NEPS). With this working paper, we demonstrate how to apply multiple imputation with R and Stata. We further illustrate an imputation approach on how to conduct data preparation and data management prior to conducting multiple imputation. Both software programs have their advantages: R allows to specify the predictor matrix manually, which supports specific tailoring to individual data problems but can also be quite time-consuming. Both software tools offer a broad range of imputation methods, while R presents a greater variety of imputation methods based on machine learning. However, up to this date, R is less suited for longitudinal imputation, whereas we found longitudinal data preparation and imputing for longitudinal data more straight forward and easier to implement in Stata. These pitfalls should be kept in mind, when deciding on the imputation model and selecting software.

Keywords:

missing data, multiple imputation, mice, R, Stata, FCS, chained equations

1. Introduction

Missing data are a common problem in survey data. Consequently, multiple imputation methods have become a widely used approach to handle incomplete data in recent years (Little & Rubin 2002; van Buuren 2012; Carpenter and Kenward 2013). Despite the pervasive use of multiple imputation, examples and guidelines how to conduct data preparation and data management prior to conducting multiple imputation are limited. Hence, this working paper describes the application of multiple imputation in the context of survey data and places a focus on data management.

In this working paper, we illustrate the pitfalls social researchers are often confronted with and try to outline a typical approach on how to handle missing data by exemplifying it, using data from the National Educational Panel Study (NEPS) Starting Cohort 1 (SC1).[1] The NEPS SC1 panel study is affected, similarly to many other panel studies, by the fact that not all variables were surveyed in all four waves along with the presence of missing values. In our case study, we are interested in estimating the probability of returning to the labor market in second year after childbirth and since the imputation model needs to reflect the analysis strategy, we intend to perform an imputation for wave 2, while we also include information from the previous wave. Before elaborating further on our imputation and data management strategy for the SC1, which is our missing data case study, we provide a short overview of multiple imputation and refer to further literature. We then describe how to conduct multiple imputation using R (version 3.4.0) and Stata (version 14.1), both are commonly used software in the social sciences and compare their advantages and disadvantages. We end with a brief discussion of our imputation strategy and reflect on its performance.

2. Multiple Imputation

Introduced by Donald Rubin (1987) multiple imputation (MI) has meanwhile been integrated into common software and gained vast popularity in applied science. To delve deeper into the mechanism behind multiple imputation, an overview of the basic literature can be found at Little and Rubin (2002), Allison (2002) or Carpenter and Kenward (2013), while the latter present a more user-oriented introduction. To conduct MI computationally with R or Stata, we refer to van Buuren & Groothuis-Oudshoorn (2011)'s R package mice and to Stata's mi suite of commands (StataCorp 2017).

At the beginning, it is necessary to make assumptions on the missing data mechanism in order to impute correctly and draw inferences from the results after the imputation process. However, it is hardly feasible to know exactly which mechanisms have created the missing values, since social research is frequently confronted with partially observed data sets. Although, it is possible to make plausible inference from observed data, definite assumptions can only be drawn from complete data sets or from simulations (Carpenter & Kenward 2013).

[1] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Newborns, doi:10.5157/NEPS:SC1:4.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network

The relevant properties, that have to be taken into account, are MCAR (Missing completely at random), MAR (Missing at random) and MNAR (Missing not at random). MCAR occurs when there is no correlation between missing values and other observed and unobserved variables. When the data is MCAR, it can be assumed that complete case analysis will produce unbiased parameter estimates and valid inference is possible, while suffering to some extend from less statistical power. MAR in contrast, occurs when the missingness of a certain variable is related to other observed variables in the data set, but not to the variable itself with its own unobserved values. For example, in our data, migrants are more likely to have missing information on the variable mother-child interaction than non-migrants (migrant status predicts the missingness of another variable in the data set). The MAR assumption is less restrictive and most imputation approaches such as mice generally start from the MAR assumption. If both MAR and MCAR do not apply, MNAR holds, which means that the probability of missing values is due to unknown variations and that the value of the unobserved variable itself is correlated with missingness (Rubin 1976; Templ & Filzmoser 2008; van Buuren 2012; Carpenter & Kenward 2013). Under MNAR, a different modelling approach has to be taken into consideration, which can be handled by some statistical software (for more information turn to van Buuren & Groothuis-Oudshoorn (2011)). We, however, focus on the MAR assumption, which can also be used when data is MCAR.

In order to apply multiple imputation under MAR, we first examine - graphically and descriptively - our missing data set to understand the missing data mechanism. Second, the imputation model needs to be specified, concerning the selection of variables that require imputation and the choice of variables that are suitable to explain the mechanism for the missing values (the so-called predictor variables). Beyond this, we determine the functional form of the imputation method and select between different methods to accommodate non-linear relationships and categorical variables, before deciding on the number of imputations. Third, we apply Rubin's combining rules to compound multiply imputed estimates and conduct statistical analysis afterwards. Last, we evaluate the quality of our imputation model with different diagnostics to test whether certain steps need to be altered.

3. Application on NEPS SC1

3.1 Data set and Analytic Strategy

To demonstrate multiple imputation, we rely on data from the National Educational Panel Study (NEPS) Starting Cohort 1 (SC1) (Blossfeld, Roßbach & von Maurice 2011). NEPS SC1 includes a register-based probability sample consisting of 3,500 newborns between February and July 2012 (for details see Weinert et al. 2016). The first wave of the survey took place when the children were at least six months and no more than eight months old in order to derive valid and comparable measurements of infant development. The second wave took place at the age of 12-17 months, the third at the age of 25-27 months, and the fourth at the age of 37-39 months. In the fourth wave, still 2,389 children and 2,478 parents took part. From the 1st to the 3rd wave of the survey, mostly female interviewers who had been specially trained for the study, visited the homes of the selected families and conducted video-based observations, for example for the parent-child-interaction. Additionally, one parent (usually the mother) in every family was surveyed by computer-assisted personal interviews. In the 2nd wave, parents were interviewed by telephone and direct measures at the children's homes were only assessed in half of the sample. From the 3rd wave onwards, parents could

choose between a personal or telephone interview. Similarly to other panel studies, this panel data set suffers from typical item non-response problems, drop outs, and that a number of variables are only available in certain waves and not across all waves.

In our empirical example, we estimate the probability for those mothers who have not returned in the first year of returning to work in the second year after childbirth (18-27 months). We define our dependent variable as a return to employment whenever the respondent states either being part-time employed or full-time. We construct the variable return to employment from the variable that measure the mother's current employment status in every survey wave. If no information is available for this variable, which is true for 25% of cases, we rely on information from the employment calendar that respondents only filled out in wave 2 and 4, allowing us to fill in values for 10% of the cases. For the remaining missing information, we filled in gaps when the respondent's employment status remained constant in the wave prior and after our target wave, allowing us to fill in values for 2.5% of cases, leaving us with 7.5 % of missing values on the dependent variable.

Our analysis sample includes all mothers who gave birth after observing child development for the first time in the survey, while we exclude male respondents, even though they answered to being the main care-taker, as well as teenage pregnancies below the age of 18 years. Due to fact that different theoretical mechanisms for returning to the labor market in the second year after childbirth can be expected for these two groups. We observe all mothers in our analysis sample until we observe a positive outcome, this is a return to the labor market in second survey wave. Alternatively, observations are right censored, either when a mother gave birth to another child or when she remained out of employment beyond wave 2.

3.2 Variables for the analysis model

Relying on survey data from the questionnaire answered by the mother, we control for the mothers' age, migration background, educational attainment, and mental problems. These are available in all four survey waves. Mental problems are measured by constructing an index based on how depressed the mother felt in the last four weeks and how unsatisfied and drained she felt with her mother role (for more details refer to Linberg, Freund & Mann 2017). We also control for child features, such as birth order and health (available in wave 2-3), as well as household factors, such as household composition and household income. Aspects of the children's non-cognitive development were measured indirectly by asking their mothers about different aspects of child temperament. For our case study, we concentrate on negative affectivity (e.g. how quickly the child gets angry when not getting what it wants) (available in all four waves). We z-standardized the dimension of the child's temperament. The parent-child interaction is directly observed in wave 1 and 2 via video recording and assessed on a rating system (for more information refer to Sommer & Mann 2015).

The dependent variable is a binary variable where 1 describes the mother's return to the labor market after childbirth at the time of the second survey. We used logistic regression models to assess the relationship between the mother's return and children's temperament, while adjusting for potentially confounding factors (mothers' age, migration background, educational attainment, mental problems, the child's birth order and health, number of

children in the household, mother-child interaction, home-learning environment, and household income).[2]

3.3 Assessment of Missing data

In the second wave of SC1, 968 of 2762 (35 %) individuals had data available for a complete case analysis, including all variables in the regression analysis. 7.5 percent of the panel participants had missing values on the outcome variable return, while 0.4 percent of explanatory variable, negative affectivity, was missing. Fully observed variables are: migration background, child's sex, mother's age at childbirth and second or higher order birth. We found most missing values on the variable mother-child interaction with 59.27% missing values. This is due to the fact that only 200 hours of video recording was prepared for scientific analysis (Sommer & Mann 2015), as video recording was not carried out or was interrupted, when the child was asleep or felt unwell. This was more often the case for migrants and less educated families, indicating that missing values are likely not missing by design.[3] For scientific analysis only complete videos were prepared (Mann & Sommer 2015). Table 1 shows the co-occurrence of missing values across variables (column 2-6). When comparing the summary statistics of variables in Table 1 to summary statistics based on a complete case analysis (not shown), consisting of 968 individuals, slight differences become visible: Mothers with completely observed data are slightly older (32.62 vs. 32.60 year of age), more often have a tertiary degree (44% vs. 38%), less often have a migration background (9% vs. 10%), more often have less children in household (0.73 vs. 0.78 children,) and more often live in households with higher household incomes (8.00 vs 7.92 EUR (log)). Assessing the missing data in our case study suggest that our analysis can benefit from MI, since we have a substantial amount of missing data and restricting the sample to a complete case analysis can bias results, as there are differences on key variables between the sample based on complete cases and those with missing information.

[2] Please note that this analysis is simplified and neglects several underlying mechanisms concerning parenting as well as child development and temperament. We use this simplified analysis for illustrative purposes.

[3] However, we can only assume that missing information is not missing by design, as the reasons for missing video recordings are not documented in the data.

Table 1

Summary Statistics with observed data and after imputation with Stata and R

	Model 1 Observed					Model 2 Imputation with Stata		Model 3 Imputation with R	
	Mean	Sd	Min	Max	% miss	Mean	Sd	Mean	Sd
Child Characteristics									
Negative Affectivity (centered)	0.24	1.10	-4	2	0.40	0.43	1.10	0.24	1.10
Medical Problems	0.07	0.26	0.00	1.00	0.33	0.07		0.07	
Mother Characteristics									
Mother-Child Interaction	3.40	0.55	1.00	4.80	59.27	3.37	0.56	3.36	0.56
Mother's age at birth of child[1]	32.60	4.99	18.00	54.00	0.00	32.60	4.99	32.60	4.99
No vocational degree (Ref.)[1]	0.16	0.37	0.00	1.00	0.18	0.16		0.16	
Vocational degree[1]	0.46	0.50	0.00	1.00	0.18	0.46		0.46	
Tertiary degree[1]	0.38	0.49	0.00	1.00	0.18	0.38		0.38	
German (Ref.)									
Migration background[1]	0.20	0.40	0.00	1.00	0.00	0.20		0.20	
Psychological well-being	-0.03	0.71	-0.99	2.64	0.65	-0.03	0.71	-0.03	0.71
Joint learning activity infrequently) (Ref.)									
Joint learning activity daily	0.73	0.44	0.00	1.00	0.18	0.73		0.73	
Household Characteristics									
Number of children	0.78	0.95	0.00	7.00	0.11	0.78	0.95	0.78	0.95
First child in household (Ref.)									
Second child or beyond in household[1]	0.54	0.50	0.00	1.00	0.00	0.54		0.54	
Log household income (in €)	7.97	0.54	5.19	11.51	8.07	7.97		7.97	

Source: SOEP v31 linked with regional data on the county-level (2006-2014).
[1] Auxiliary variables not imputed with Stata.

Testing whether your data is MNAR or MAR (further explanation in Chapter 2) is in reality often not possible, as complete information on missingness is required to clarify which mechanism has caused the missing values (Carpenter & Kenward 2013). Instead, visual assessment of missing values is recommended to display the missing data pattern to further gain information about the mechanism (van Buuren 2012).

The VIM package in R with the functions `matrixplot()` and `aggregateplot()` are well suited to inspect missing values (Templ & Filzmoser 2009). The matrixplot in Figure 1 is interactive and it is of great use to explore the missing data mechanisms (see Zhang 2015).The graph in Figure 1 shows, for example, how the missingness of the mother-child interaction, as well as for all other depicted variables, depends on the migrant status. When the respondent has a migration background (black), a lot of missing values are prevalent (pink). The same effect is visible for the variable education. For lower educated mothers (white and brighter shades of grey) the mother-child interaction tends to be more often unobserved (pink) in comparison to higher educated mothers (darker shades of grey and black). Graphically detecting such relationships between missing and observed variables makes it more realistic to assume MAR,

as missing values depend on available information in the data set. Figure 2, depicting the so-called aggregateplot, consists of two visualizations of missingness. The left half of Figure 2 shows the variables containing the most missing values. The graph on the right, refers to the co-occurrence of missing values across several variables and also indicates how many missing values are present in total. For example, Figure 2 shows that most missing values are present for the mother-child interaction, while all other values are observed. The second most common case is that all variables are observed (white box in second bottom line on the right).

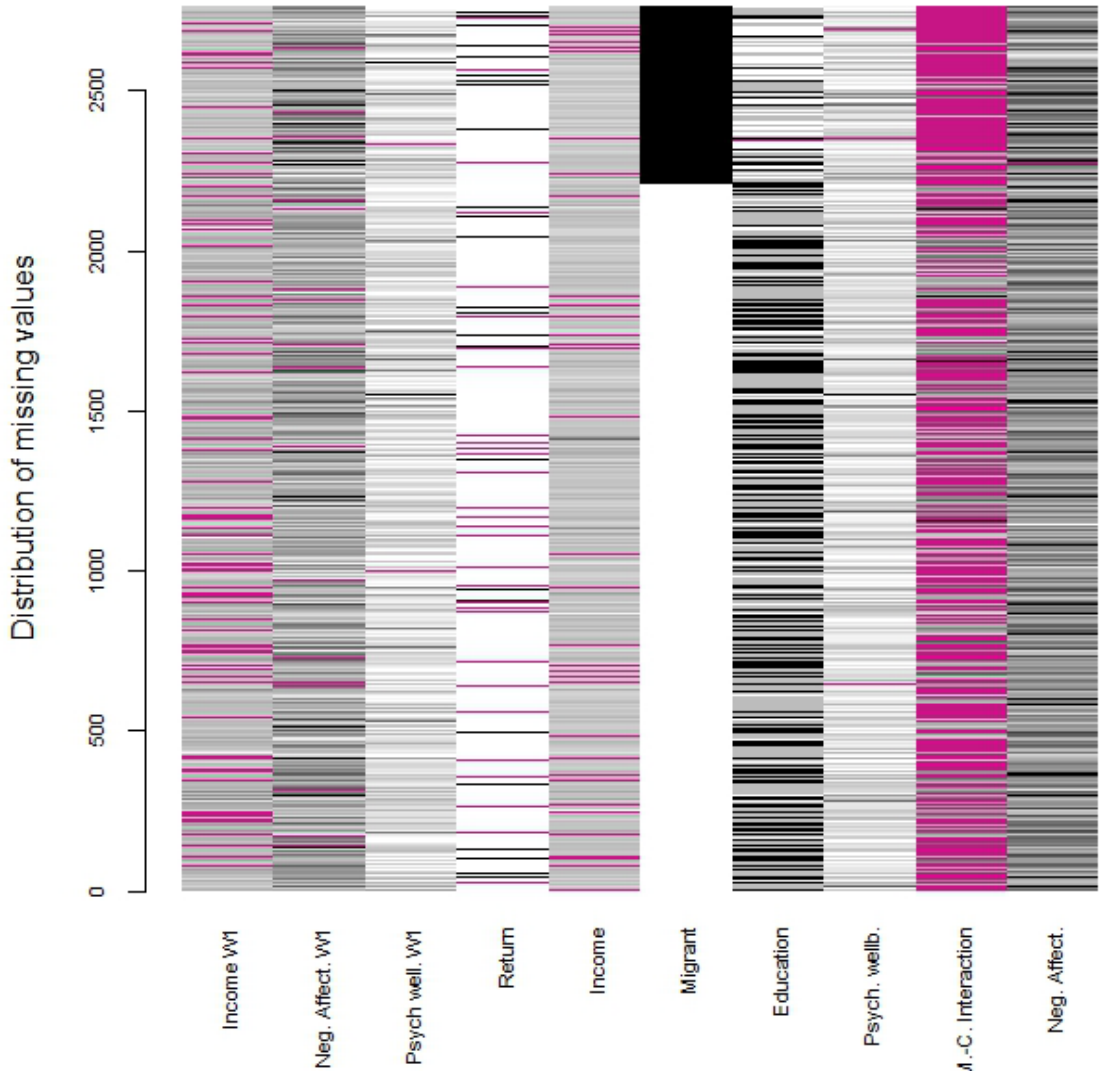


Figure 1. Matrixplot in R, NEPS SC1, Wave2 (SC1:4.0.0)[4]

[4] For reasons of displaying the graph properly, only variables with a high amount of missing values were selected.

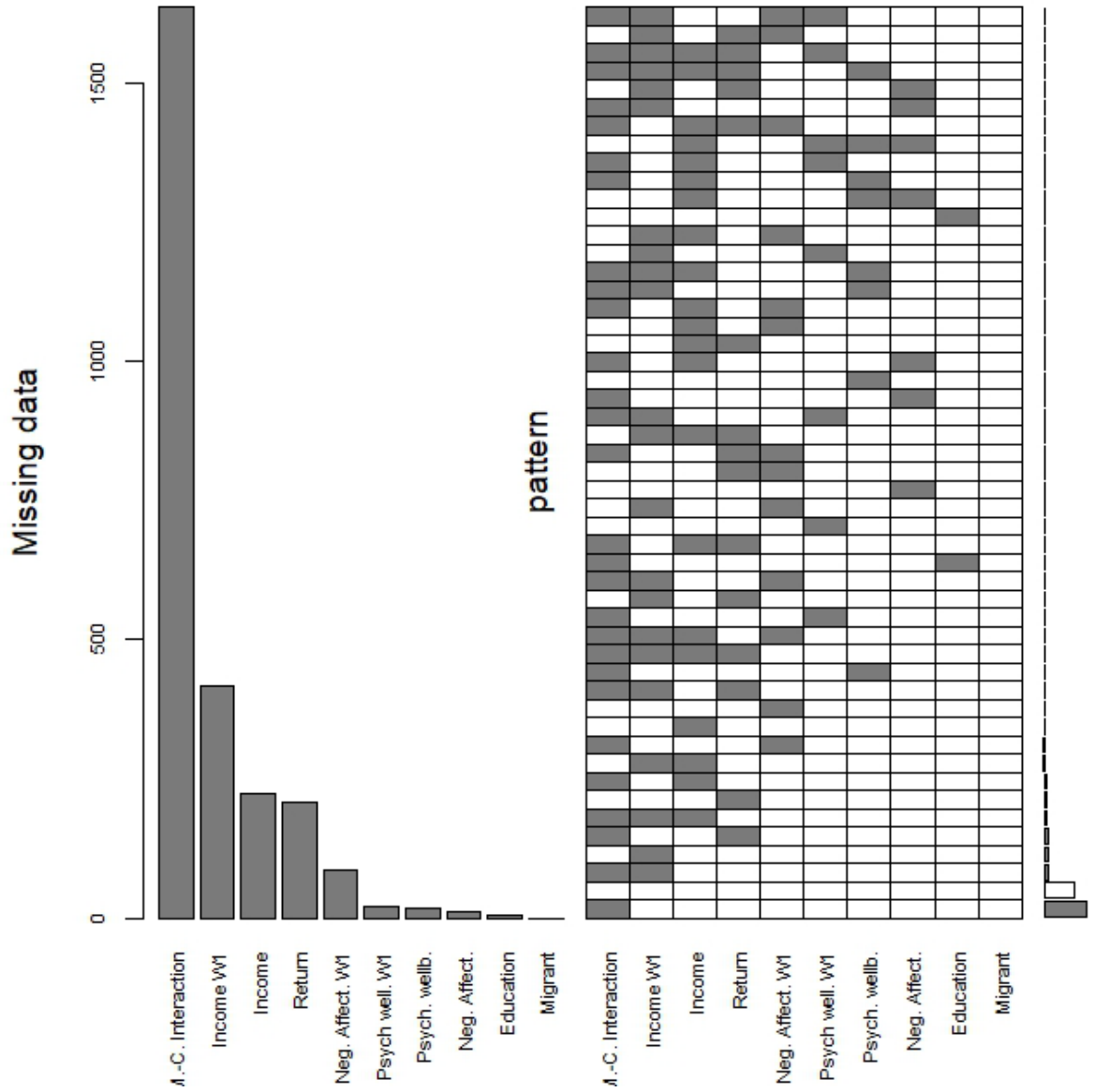


Figure 2. Aggregateplot in R, NEPS SC1, Wave2 (SC1:4.0.0)[5]

Similarly, in Stata, `misstable` can assist in displaying missing data, such as the frequence of missing values (`misstable summarize`) and the missing pattern in the data (`misstable pattern`).

Based on the comparison of summary statistics for key variables between the full sample and a complete case analysis, as well as different plots of missingness, it becomes noticeable that the missingness of certain variables are related to other observed variables, which provide tentative evidence for the MAR assumption. Nevertheless, it is possible that other unobserved factors have an impact on the missingness.

[5] For purposes of displaying the graphical depiction, only variables with a high amount of missing values were selected.

4. Proposed imputation model

Multivariate imputation with chained equations (mice) is considered an adequate choice when a joint data distribution for the missing data cannot be specified, which is often the case for multivariate data sets. Unlike joint modeling techniques, developed by Schafer (1997), where a joint multivariate distribution describes the data, mice uses a fully conditional specification (FCS). It examines each variable and specifies a multivariate imputation model for each incomplete variable, thereby iterating over a set of conditional densities (van Buuren & Groothuis-Oudshoorn 2011).

The R package mice written by Stef van Buuren generates multiple imputation with FCS. It provides tools for analyzing imputed data and it pools analysis results based on Rubin's combining rules (Rubin 1987). In Stata, the mice method is implemented in the chained method (`mi impute chained`) and in order to impute multiple variables sequentially it uses a Gibbs-like algorithm for the FCS (StataCorp 2017: 140). Likewise, Stata uses Rubin's combining rules to conduct estimations from multiple imputed data (StataCorp 2017: 47). In the following, we describe how we proceed to set up our MI model.

4.1 Data Preparation

In our case study, we are interested in the probability of having returned to the labor market in the second year after childbirth, as opposed to remaining out of the labor market. Therefore, we perform a cross-sectional imputation based on wave 2 and disregard the other three waves, in contrast to a multivariate imputation based on a longitudinal data structure. We also decided to follow this procedure as most imputation packages in R, are written for analysis of individuals nested within groups, whereas packages that allow to impute for a data structure where individuals are nested within individuals are less commonly available. However, Stata's mi suite of commands allows to impute values for a data structure where individuals are nested within individuals, but for comparability to the multiple imputation in R and since results do not significantly differ between imputing for a cross-sectional data structure and a longitudinal data structure in Stata[6], we decided to conduct a cross-sectional imputation.

For our SC1 data set, consisting of a total of 13 subfiles, five subfiles (pParent, xDirectMeasures, spEmp, MethodsCATI, MethodsCAPI) are taken into account in order to perform the imputation. We start off with pParent, where we create extended missing codes from NEPS missing codes with the Stata `nepsmiss` command. For the imputation in Stata, all missing values that should later be imputed need to be filled in with ".". For the employment episode of the employment module (spEmp), we select harmonized spells, when available and concentrate on the employment episodes survey in wave 2. Based on the person number and wave, we merge information from the other subfiles to pParent. In total, we restrict the SC1

[6] The procedure for multiple imputation with a longitudinal data structure in Stata is similar to the procedure with a cross-sectional data structure: Before the multivariate imputation, we set the longitudinal data, stored in a long format, to a wide format. After deleting all design based missing variables, that were not survey in certain waves, we declared the storage style as `mi set flong,` instead of `mi set wide.` Step 2-5 (see Appendix for the provided Stata code) do not change. Finally, the data needs to be reshaped from the wide format to the long format with the `mi reshape long` command.

data set to the relevant control variables, the dependent variable returning to employment and potential predictor variables for the imputation model (see 4.2.1). Afterwards, we transfer the Stata data set to R. To transfer the Stata data sets to R[7], we use the function `read.dta13()` from the readstata13 package (Garbuszus et al. 2018) with the options `missing.type=TRUE, nonint.factors=TRUE, generate.factors=TRUE`. We highly recommend checking whether R correctly assigned the appropriate scale of measurement to all variables. For example, in our case study, R loaded our indicator for the child's temperament as a factor variable, whereas we decide to treat it as a semi-continuous variable. Aligning with von Hippel (2009), who recommended to "first transform, then impute" (p. 266), we construct an indicator (range from 0-6) consisting of several items to assess the child's temperament and then z-standardized the indicator for the child's temperament prior to the imputation. To check this strategy, we also transformed the variable after the conducting the imputation, but we found that it made no difference.

4.2 Imputation Model

There are several requirements that need to be fulfilled when setting up an imputation model. Accordingly, van Buuren & Groothuis-Oudshoorn (2011) recommend not to run imputations based on the default configurations, but rather to adapt each step to the individual data set, since "real problems need tailoring" (p.16). Foremost, the imputation model should preserve the relations in the data and account for the process that generated the missing values. To produce consistent and realistic imputed values, several decisions are necessary. This implies deciding which variables to include, deciding on the functional form of the model, choosing a suited imputation method depending on variables' scale of measurement as well as non-linear relations in the data. Setting up an appropriate imputation model is crucial to produce valid estimates and standard errors, as well as to obtain unbiased inference (Nguyen, Carlin & Lee 2017).

4.2.1 Variable selection for the imputation model

The imputation model should include at least all variables in the analytical model to preserve the relationships between variables of interest (von Hippel 2007) and should further incorporate other auxiliary variables. A guideline is to include as many relevant variables as possible, including interaction terms and higher order terms (Collins, Schafer & Kam 2001; Hippel 2009; Marshall, Altman, Holder & Royston 2009; Graham 2012). Among others, Collins, Schafer & Kam (2001) recommend using auxiliary variables that are highly correlated with the incomplete variables. Based on this, we chose to include the following auxiliary variables in our imputation model: living in East/West of Germany, child's sex, child's birthweight (being a time-constant variable and a measurement is only available for wave 1, we inserted the values from wave 1 to wave 2), and the number of words the child is able to speak (answered by the mother in wave 2). Socio-demographic characteristics of the mother, such as age, education and work status are available in every wave and were included predictor variables. Since the SC1 is a panel study, we can make use of repeated measurements of time-varying variables as auxiliary variables in our imputation model. Hence, we included information from wave 1 on the child's temperament, household income, mother's psychological well-being, home-learning environment, and the number of children, assuming that these variables can

[7] The SC1 data is available in Stata and SPSS format.

improve the prediction of the missing values in wave 2. In total, we selected 13 predictor variables for the imputation of wave 2. Due to the fact, that we also have single mothers in the data set, we were only able to select variables that contain information for single mothers and mothers with a partner. For example, we were unable to include the partner's education and employment status as no information is available for single mothers.

Based on the recommendation in the MI literature to preserve all relationships between variables of interest (e.g. von Hippel 2007; White Royston 2009), we decided to adopt the strategy suggested by von Hippel (2007) to include the dependent variable in the imputation model and to use multiple imputation, then deletion (MID) for the dependent variable. The study of Little (1992: 1227) suggests that "cases with Y missing provide minor […] information for the regression of interest". Von Hippel supports this idea to some extent but modifies this approach in a way that Y can still be imputed. This implies that we use return to employment in the imputation model, but after the imputation we delete all imputed values for returning to employment, which allows us to model the relationship between the dependent variable and the independent variables. Therefore, a fully imputed outcome variable is used as a predictor for other variables, but its imputed values are then set to be missing again. This corresponds to Littles finding that "cases with missing Y are useful for imputation, but not for analysis" (Hippel 2017: 84).

Under the condition that auxiliary variables are defined as such in Stata, a distinctive difference between Stata and R is that in R auxiliary variables predict the missingness of other variables but are also imputed at the same time. In contrast, Stata does not impute values for auxiliary variables when they are defined as such and instead uses them only to predict missingness in other model variables. In comparison to Stata, R ends up with imputed values for all variables, including all auxiliary variables, whereas Stata may not fill all missing values, when there are missing values on the auxiliary variables.

4.2.2 Predictor Matrix

It is an R specific feature that the user is able to manually specify the predictor matrix, meaning the user can choose which predictors to use to predict each incomplete variable. Mice offers with `quickpred()` an easy, time-saving way to set up a predictor matrix, but for our case study, we felt it was less suited, as it selected too few variables as predictors as well as unsuited variables, such as the person identification number. In order to specify predictors manually, we imputed the data set only once with `mice()` and modified the `imp$predictorMatrix`, a matrix with 0/1 data, according to the suggestions of van Buuren (2012). As we end up with relatively small data sets with 24 variables, van Buuren (2012) recommends conditioning on as many variables, apart from some exceptions, as this makes the MAR assumption more plausible (see also Collins, Schafer & Kam 2001). To set up the predictor matrix, we based our variable selection on theoretical assumptions and binary t-tests: We considered variables as non-predictor variables, which have been set to zero, when they had too many missing values and when they provided no useful information to the others, for example the person identification number. We also excluded those variables with high multicollinearity and where no reasonable causal inference was found, such as child characteristics should not explain missingness in mother's education. When the predictor matrix is ill specified and suffers, among other things, from linear dependencies and multicollinearities, mice automatically sorts out those variables, but produces a warning. The

warning object can be accessed via `imp$loggedEvents`. To prevent automatic removals, the user can set the listed variables to zero in the predictor matrix (van Buuren 2012). For our case study, we found it is quite a laborious task to specify the whole predictor matrix manually, especially with a large number of variables. We therefore recommend, to take the default object from the empty mice imputation and modify it according to the data needs.

4.2.3 Imputation Method

In this step, several decisions need to be made on, for instance, the functional form of the imputation model, on how to impute variables of different scale of measurement and non-normal distributed variables. In order to specify a multivariate imputation model for each incomplete variable, mice provides an adequate choice with FCS (van Buuren & Groothuis-Oudshoorn 2011). In R, within the mice package, similarly to the specification of the predictor matrix, the default setting can be called with the `imp$method`. In the default setting, R already proposes certain methods depending on the scale of measurement of the variables. If required, the user can alter imputation methods for each variable. In R, the following methods are included in the mice package: `pmm, norm, mean, logreg, polyreg, cart, sample` (van Buuren & Groothuis-Oudshoorn 2011). Additionally, several other methods that are based on machine learning or that are suitable for 2-level analysis are also provided. In Stata, mice is implemented in the `mi impute chained` command and the user can choose between the following methods depending on the scale of measurement: `regress, pmm, truncreg, intreg, logit, ologit, mlogit, poisson,` and `nbreb` (StataCorp 2017: 141). In our example, we use predictive mean matching (pmm) for the measurement of the child's non-cognitive development and temperament, the mother-child interaction, household income, number of children, and psychological well-being, since pmm only draws from already existing values in the data. For comparability, we set the number of donors to 5, since then a random draw from 5 cases with closest predicted values is guaranteed, which does not apply in the default Stata setting with knn(1) (Morris, White & Royston: 2014). For binary variables, we chose univariate logistic regression (`logreg` in R and `logit` in Stata), for ordered factor variables with more levels we use ordered logistic regressions (`polr` in R and `ologit` in Stata, also known as proportional odds model), and multi-categorical variables we impute with multinomial logistic regression (`polyreg` in R and `mlogit` in Stata). Another possibility to impute income is the cart method in R, as suggested by Aßmann et al. (2017). Due to comparability between R and Stata, we focused only on methods offered in both packages.

4.2.4 Visit sequence

The visit sequence sets the imputation order. In R the default setting is set to imputing data from left to right, instead of reordering the data set according to the missing data rate. Re-ordering can, however, be achieved through the `visitSequence` argument, called `vis=c()`. We chose to re-order the data in a way that the variables most dependent on others, as well as containing many missing values were set to the right. This corresponds to the default imputation order of mice, which also imputes from left to right, which means that information of previously imputed variables is used for imputing subsequent variables. In Stata, by default variables are imputed according to the missing data rate, meaning from the most observed to the least observed. The user can specify imputation order with the `orderasis` request. When the algorithm passes through the individual variables frequently, which can be specified through the iteration number, the visit sequence is negligible (van Buuren 2012; StataCorp

2017). Nevertheless, we found the setting of the visit sequence valuable to control for marginal order effects.

4.2.5 Number of iterations and imputed data sets

Lastly, we set the number of iterations and number of imputed data sets. To ensure convergence, we inspected the trace plots generated by `plot()` in R. We visually detect that 40 iterations are sufficient for our case (Figure 4). In R, we set the number of iterations to 40 (`maxit=40`). In Stata, the default are 10 iterations and the number of iterations can be altered through `burnin(40)`. For the number of imputed data sets, a general rule with respect to the Fraction of Missing Information (FMI) is to take the variable with the highest rate of FMI and multiply it with 100 (White, Royston & Wood 2001: 388). In our case, the variable with the highest missingness is the mother-child interaction variable with 0.84 FMI. Based on this general rule and due to computational capacities, we round the number of imputations to the highest decile 90 (100*0.84). In R, the number of imputations are set through the command `m=90` and in Stata through the option `add(90)`.

5. Results based on MI for wave 2

Having now 90 data sets, which are not fully identical, but differ marginally, estimates from each imputed data set must be combined. Although, it may seem tempting to average over the imputed data, this would result into biased estimates. That is why van Buuren (2012) suggests analyzing each of imputed data sets separately which can be realized with the `with()` argument in R and the `mi xeq numlist` in Stata. Rubin (1987) developed a set of rules, the so called “Rubin’s combining rules” (Rubin 1987) to combine standard errors (SE) and coefficients, as the SE and coefficient differ slightly across all imputed data sets. Variation occurs due to the uncertainty created by the missing values and therefore it is necessary to control for the within and between imputation variability. Rubin’s combining rules are incorporated in R through the mice function `pool()`and the `mi estimate` command in Stata.

With our binary outcome variable return to employment in the second year after childbirth, we employed a logistic regression fitted by maximum likelihood for the second survey wave. Our results show that mothers have a statistically significant higher probability of returning to employment in the second year after childbirth, when they live in households with higher household incomes, have a better mother-child interaction and have a migration background. When comparing the regression results before and after the imputation, it becomes obvious that the coefficients are smaller in the logistic regression after having conducted the imputation. Generally, the results are comparable to the complete case analysis, just that the number of observations increased from 968 to 2424 in Stata and 2553 in R, allowing us to use a much larger share of the representatively sampled SC1. The reason for different sample sizes in R and Stata is due to missing values in the auxiliary variables that are also fully imputed in R, but not in Stata. When comparing the results from R and Stata, we noticed mostly similar coefficients, standard errors and significances except for a few variables. In this case, we found the most striking differences for mother-child interaction. A possible explanation is that R imputes all variables iteratively.

Table 2

Probability of returning in the 2. Year after childbirth

	Model 1 Observed (complete case analysis)		**Model 2 Imputation with Stata**		**Model 3 Imputation with R**	
	Coefficient	*SE*	*Coefficient*	*SE*	*Coefficient*	*SE*
Child Characteristics						
Negative Affectivity (centered)	0.19	(0.26)	0.09	(0.08)	0.08	(0.08)
Medical Problems	1.32	(0.69)	0.47+	(0.29)	0.54+	(0.29)
Mother Characteristics						
Mother-Child Interaction	1.28*	(0.59)	0.76*	(0.30)	0.86*	(0.42)
Mother's age at birth of child[1]	0.01	(0.06)	-0.01	(0.02)	-0.01	(0.02)
No vocational degree (Ref.)[1]						
Vocational degree[1]	0.71	(1.18)	0.31	(0.28)	0.30	(0.29)
Tertiary degree[1]	-0.72	(1.30)	-0.07	(0.32)	-0.12	(0.33)
German (Ref.)						
Migration background[1]	0.88	(0.71)	0.67**	(0.22)	0.65***	(0.22)
Psychological well-being	-0.16	(0.41)	0.07	(0.12)	0.05	(0.12)
Home learning environment (joint learning activity infrequently) (Ref.)						
Home learning environment (joint learning activity daily)	-0.99	(0.56)	-0.23	(0.20)	-0.23	(0.21)
Household Characteristics						
Number of children	0.18	(0.42)	-0.03	(0.15)	-0.04	(0.16)
First child in household (Ref.)						
Second child or beyond in household[1]	-0.54	(0.79)	-0.29	(0.28)	-0.27	(0.28)
Log household income (in €)	0.975*	(0.49)	0.24	(0.19)	0.23+	(0.19)
Constant	-16.64***	(4.24)	-7.05***	(1.61)	-7.30***	(1.78)
Observations	968		2424		2553	

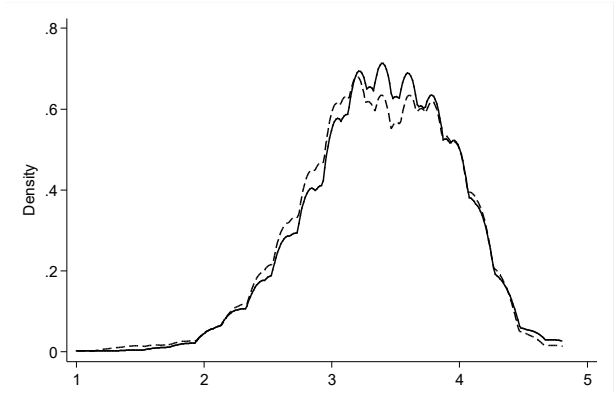
Standard errors in parentheses.
Source: NEPS SC1 (SC1:4.0.0).
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

6. Diagnostics and Performance Checks

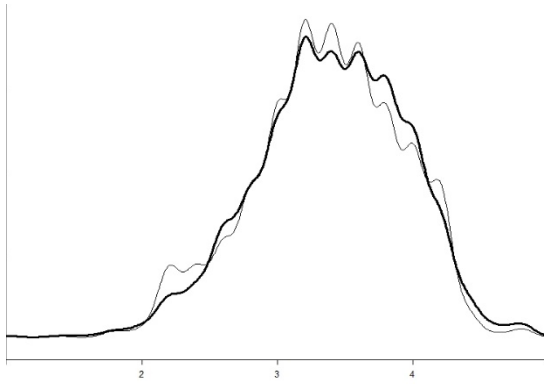
After having conducted the data imputation, it is recommended the imputation model to be evaluated, and also re-check whether the imputations based on the MAR assumption are plausible. As soon as any abnormalities are identified, the imputation has to be re-run (van Buuren & Groothuis-Oudshoorn 2011: 42). Although diagnostics for imputations are still being established in research (Nguyen, Carlin & Lee 2017), it is recommended that the distribution of the imputed and observed data be compared through graphical diagnostics and to check the model fit of the imputation model to the observed data (StataCorp 2017).

To check whether the imputed and observed data have similar distribution, we first compare summary statistics in Table 1. The last four columns contain the information for the imputed data. For R and Stata, the mean and standard deviations are similar.

Moving beyond summaries, graphical representations of the distribution of the observed and imputed data are of great use, especially density plots and histograms. In Stata, the `midiagplots` are helpful in producing graphical diagnostics, such as a density plot (Marchenko & Eddings 2012). In R, the mice package is restricted to few diagnostic tools, such as density plots with the command `densityplot()` and convergence plots with the command `plot()` but more diagnostics can be conducted by using the package mi with diagnostics (Su, Gelman, Hill & Yajima 2011). The figure below, shows a kernel density plot for the 90th imputation of the mother-child interaction and demonstrates that the distribution closely matches the observed data. Although the mother-child interaction suffered from a significant amount of missing values and it does not follow a normal distribution, the `pmm` imputation method, that only uses existing values for imputation, worked relatively well. Diagnostic checks can help to evaluate the imputation performance after having conducted the imputation, such as it can help to identify whether the procedure has to be repeated or whether certain steps in the data imputation need to be altered. In our data set, we experience, when using a different imputation method, such as regress to impute the mother-child interaction, the imputation model performed poorly.



Stata, solid line= observed, dashed line= imputed



R, black line = observed, grey line= imputed, Bandwidth= 0.75

Figure 3. Variable Mother-child interaction

To check the model fit of the imputation model to the observed data, diagnostics for standard regression analysis can be used to examine differences in the model fit across the multiple imputed data sets. For a detailed description how to produce these graphs, turn to Marchenko and Eddings (2011), Nguyen et al. (2017) and Würbach et al. (2016). The latter also provides syntax examples with R and Stata.

Among other diagnostic checks, we found the convergence plot helpful in inspecting, whether the imputation across different variables reached convergence. In R, this can be generated by the function `plot()`. Figure 4 contains trace plots for the variables mother-child-interaction (upper two graphs) and child negative affectivity (two graphs at the bottom). On the left of the graph the convergence of the mean values and on the right the convergence of the

standard deviations are plotted. It further shows that with an increasing number of iteration rates (increasing from left to right up to 40 iterations), the trace lines of the plotted variables reach "healthy convergence" (van Buuren & Groothuis-Oudshoorn 2011: 40), which means that they fluctuate to a lower extent around the mean and standard deviation.

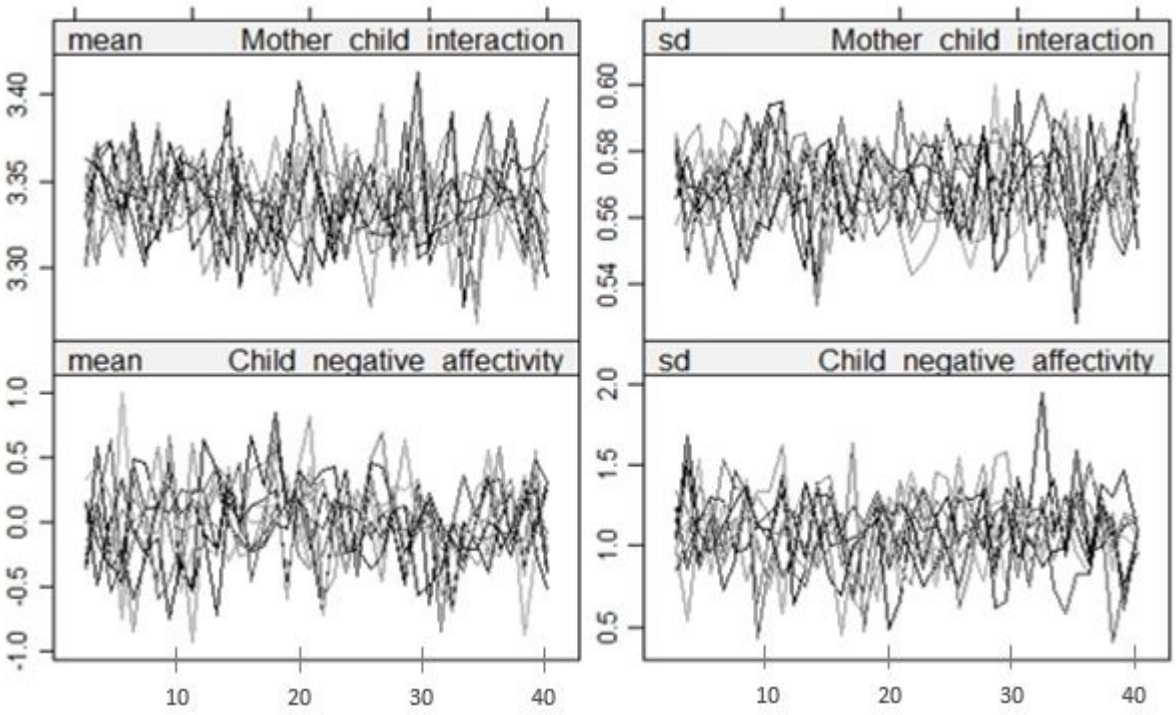


Figure 4. Convergence plot in R for 10 imputed data sets

7. Comparison and Conclusion

In this working paper, our aim was to outline how to handle missing data. To illustrate this, we used data from the National Educational Panel Study (NEPS) Starting Cohort 1 (SC1). We described how to prepare the data before conducting multiple imputation in R or Stata and further, how default settings of `mice()` and `mi impute chained` can be customized. We also briefly summarized the research and pointed the reader towards ongoing research in the field.

We end this working paper with a brief comparison of our experience with conducting the data imputation with R and Stata. We found the data preparation and the execution of the imputation more straightforward and easier to implement in Stata. This might be due to more prevalent literature and user-friendly manuals, which can be easily accessed via an open source. For our case study, our analysis strategy required a cross-sectional analysis of the starting cohort SC1. As the imputation strategy should reflect the analysis strategy, we decided to perform the imputation in a cross-sectional manner. We also included variable measurements of the previous wave into the imputation model. This strategy was also convenient, as to our knowledge, literature on multiple imputation for individuals nested within individual across waves is underrepresented in R. We found constructing the visit sequence, which sets the imputation order, more straightforward in Stata, where the default is that the variables are imputed according to the missing data rate (but manual specification

is possible). Whereas in R, the default setting is set to imputing data from left to right, regardless of the missing data rate. Determining the imputation order of the variables, through the visit sequence, can be of assistance as it can avoid order effects.

In contrast to Stata, R allows the user to manually specify the predictor matrix. Although, we found this to be quite time-consuming it may be of great value in certain cases. While both software tools offer a broad range of imputation methods, R presents a greater variety of imputation methods based on machine learning, such as the classification and regression tree algorithm (cart). This can help to customize the imputation even further to the requirements of the data. Another advantage in R, is that all variables are imputed regardless whether they are auxiliary variables or incomplete variables. In Stata, only the incomplete variables are imputed, and no values are imputed for auxiliary variables. When auxiliary variables possess missing values, it is possible that not all missing values are imputed.

Although performing diagnostics checks are of great importance to evaluate the imputation model, very few have been made available in R and Stata. To date, R offers the largest selection of imputation diagnostics, where most are available in the VIM package and in the package mi with diagnostics (Su, Gelman, Hill & Yajima 2011). In the mice package, we found it quite tedious to change default settings for customized illustration purposes. Stata has the `midiagplots` command for graphical diagnostics and it is possible to perform tailored checks by writing syntax (refer to Marchenko and Eddings (2011), Nguyen et al. (2017) and Würbach et al (2016)).

We conclude with a word of caution. Missing values are imputed simultaneously with chained equations. The imputation model only includes the relationship between specified variables and not between those excluded from the model. Therefore, we recommend to individually specify imputation models depending on the research question and model variables. We hope that with this working paper, we provide an illustration of an imputation approach on how to handle missing values in survey data.

References

Allison, P. D. (2002). *Missing data*. Thousand Oaks. CA: SAGE

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F. & Bela, A. (2017). Nonparametric multiple imputation for questionnaires with individual skip patterns and constrains: The case of income imputation in the National Educational Panel Study. Sociological Methods & Research. 46(4). 864-897

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (2011). Education as a lifelong process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*. Sonderheft 14.

Carpenter, J. & Kenward, M. G. (2013). *Multiple imputation and its application*. Statistics in Practice. Chichester: John Wiley Sons.

Collins, L. M., Schafer, J.L., & Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods. 6(4)*, 330-351.

Garbuszus, J. M. et al. (2018). Package „readstata13“. Retrieved 03.07.2018, from https://github.com/sjewo/readstata13.

Graham, J. W. (2012). *Missing data: analysis and design*. New York: Springer

Linberg, A., Freund, J.-D. & Mann, D. (2017). Bedingungen sensitiver Mutter-Kind-Interaktionen. In: H. Wadepohl, K. Mackowiak; K. Fröhlich-Gildhoff & D. Weltzien (Hrsg.): Interaktionsgestaltung in Familie und Kindertagesbetreuung, 27-52. Wiesbaden: VS-Verlag.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association. 87(420)*, 1227-1237.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. 2nd Edition. Chichester: John Wiley Sons.

Marshall, A., Altmann, D. G., Holder, R. L. & Royston, P. (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines*. BMC Medical Research Methodology*. 9(57*)*, 1-8.

Marchenko, Y. V. & Eddings, W. (2012). A note on how to perform multiple-imputation diagnostics in Stata. *The Stata Journal. 12(3),* 353-367.

Morris, T. P., White, I. R. & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*. *14.* 75-87

Nguyen, C. D., Carlin, J. B. & Lee, K. J. (2017). Model checking in multiple imputation. An overview and case study. *Emerging Themes in Epidemiology. 14(8),* 1-12.

Rubin, D. B. (1976). Inference and missing data. *Biometrika. 63(3)*, 581-590.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley Sons.

Sommer, A. & Mann, D. (2015). Qualität elterlichen Interaktionsverhaltens. Erfassung von Interaktionen mithilfe der Eltern-Kind-Interaktions Einschätzskala im Nationalen Bildungspanel. *NEPS Working Paper. 56*. Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

StataCorp. 2017. Stata: Release 15. Statistical Software. College Station, TX: StataCorp LLC.

Su, Y.-S., Gelman, A., Hill & J., Masanao (2011) Multiple imputation with diagnostics. Opening windows into the black box. *Journal of Statistical Software. 45(2),* 1-31.

Templ, M. & Alfons, A. (2009). VIM: Visualization and imputation of missing values. Retrieved 03.07.2018, from http://cran.r-project.org/package=VIM. R package version 1.3

Templ, M. & Filzmoser, P. (2008). Visualization of missing values using the R-package VIM. Institut für Statistik und Wahrscheinlichkeitstheorie. Forschungsbericht CS-2008-1, 1-13.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: CRC Press.

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). Mice. Multivariate imputation by chained equations in R. *Journal of Statistical Software. 45(3),* 1-67.

Von Hippel, P. T. (2007). Regression with missing Y's. An improved strategy for analyzing multiply imputed data. Sociological Methodology. 37, 83-117.

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology. 39(1),* 265-291.

Weinert, S., Linberg, A., Attig, M., Freund, J.-D. & Linberg, T. (2016). Analyzing early child development, influential conditions, and future impacts: Prospects of a German

newborn cohort study. *International Journal of Child Care and Education Policy*. 10, 1-20.

White, I., Royston, P. & Wood, A. M. (2011). Multiple imputation using chained equations. Issues and Guidance for Practice. *Statistics in Medicine. 30(4),* 377-399.

Würbach, A., Hammon, A., Geissler, F. & Goßmann, S. (2014). Data documentation: imputed data file of starting cohort. Ed. by NEPS-Methods Group. Bamberg.

Zhang, Z. (2015). Missing data exploration. Highlighting graphical presentation of missing pattern. *Annals of Translational Medicine. 3(22)*, 1-17.

Appendix

Code snippets:

R

```
* 1. Employ empty imputation
ini2 <- mice(W2dat, maxit=0)
tail(ini2$loggedEvents, 10)

* 2. Specify predictor matrix
n2 <- ncol(W2dat)
predmat2 <- ini2$predictorMatrix
predmat2[,c(1,8)]<- rep(0,n2) #ID,Wave
predmat2["Education", "Child_sex"] <- 0 #Child_sex cannot explain education
…

* 3. Set the imputation method
meth2 <- ini2$method
meth2[2] <- "polr"
…

* 4. Impute
Imp_W2 <- mice(
 data = W2dat,
 m = 90,
 maxit = 40,
 predictorMatrix = predmat2,
 seed=2403,
 method=meth2
)
Long_W2 <- complete(Imp_W2, "long", include="F") #all 90 data sets stacked, do not
use for analysis
colSums(is.na(Long_W2)) #check, if all missing values were imputed

* 5. MID Strategy: delete
Yimp_pos <- which(is.na(W2dat$Return))
length(unique(Yimp_pos)) #check
Imp_W2$imp$Return[Yimp_pos,] <- NA
```

* 6. Analyze

```
modelFit2 <- with(Imp_W2, glm(Return~ Child_negative_affectivity +
Child_medical_problems + Mother_child_interaction + Education +
Age_at_childs_birth + Migrant_background + Log_HH_income +
Second_or_higher_order_birth + Psychological_wellbeing + Number_of_children +
Home_learning_environment , family = binomial("logit")))

round(summary(pool(modelFit2)),2)

modelFit2$analyses
```

Stata

* 1. Declare the storage style

```
mi set wide
```

* 2. Register variables

```
mi register imputed c_negative_temp_z c_medicalprob_w2 hh_income_tv_log
nr_children_tv joint_activity_daily psy_belastung_tv c_interaction_m_w2 reentry

mi register regular edu_prebirth migrant age_prebirth second_child Ost
c_repeat_words c_birthweight c_sex c_negative_temp_z_w1 hh_income_tv_log_w1
psy_belastung_tv_w1 joint_activity_daily_w1 nr_children_tv_w1
```

* 3. Set seed for reproduction

```
set seed 2403
```

* 4. Impute

```
mi impute chained (pmm, knn(5)) c_negative_temp_z c_interaction_m_w2 ///

      (pmm, knn(5)) hh_income_tv_log psy_belastung_tv nr_children_tv ///

      (logit, augment) c_medicalprob_w2 joint_activity_daily ///

      (logit, augment) reentry /// MID strategy

      = edu_prebirth migrant age_prebirth second_child Ost c_repeat_words ///
      c_birthweight c_sex c_negative_temp_z_w1 hh_income_tv_log_w1 ///
      psy_belastung_tv_w1 joint_activity_daily_w1 nr_children_tv_w1, add(90)
      burnin(40)
```

* 5. MID strategy: delete

```
foreach var of varlist _1_reentry - _90_reentry {

replace `var'=. if mis_reentry==1

}
```

* 6. Analyze

```
mi estimate, post: logit  reentry  c_negative_temp_z  c_medicalprob_w2
joint_activity_daily c_interaction_m_w2 i.edu_prebirth   age_prebirth  migrant
hh_income_tv_log    nr_children_tv  second_child    psy_belastung_tv
```

Working Papers of the Leibniz Institute for Educational Trajectories (LIfBi)
at the University of Bamberg

The LIfBi Working Papers series publishes articles, expert reports, and findings related to data collected and studies conducted at the Leibniz Institute for Educational Trajectories—first and foremost, the National Educational Panel Study (NEPS) in Germany.

LIfBi Working Papers are edited by the LIfBi Board of Directors and the Heads of the LIfBi Departments. The series started in 2011 under the name "NEPS Working Papers" and was renamed in 2017 to broaden the range of studies which may be published here.

Papers appear in this series as work in progress and may also appear elsewhere. They often present preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the LIfBi management or the NEPS Consortium.

The LIfBi Working Papers are available at www.lifbi.de (see section "Institute > Publications"). LIfBi Working Papers based on NEPS data are also available at www.neps-data.de (see section "Data Center > Publications").

Contact:
Leibniz Institute for Educational Trajectories
Wilhelmsplatz 3
96047 Bamberg
Germany
contact@lifbi.de