

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**

(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (v): Successful strategies for implementing new editing and imputation methods

**FAR FROM NORMAL - MULTIPLE IMPUTATION OF MISSING VALUES IN A  
GERMAN ESTABLISHMENT SURVEY**

**Invited paper**

Submitted by the German Institute for Employment Research<sup>1</sup>

**I. INTRODUCTION**

1. For many datasets, especially for non mandatory surveys, missing data are a common problem. Deleting units that are not fully observed, using only the remaining units is a popular, easy to implement approach in this case. This can possibly lead to severe bias if the strong assumption of a missing pattern that is completely at random (MCAR, [Rubin \(1987\)](#)) is not fulfilled. Imputing missing values can help to handle this problem. However, ad hoc methods like, e.g., mean imputation can destroy the correlation between the variables. Furthermore, imputing missing values only once (single imputation) generally doesn't account for the fact that the imputed values are only estimates for the true values. After the imputation process, they are treated like truly observed values leading to an underestimation of the variance in the data and by this to  $p$  values that are too significant.

2. Multiple imputation as proposed by [Rubin \(1978\)](#) overcomes these problems. With multiple imputation, the missing values in a dataset are replaced by  $m > 1$  simulated versions, generated according to a probability distribution for the missing values given the observed data. More precisely, let  $Y_{obs}$  be the observed and  $Y_{mis}$  the missing part of a dataset  $Y$ , with  $Y = (Y_{mis}, Y_{obs})$ , then missing values are drawn from the Bayesian posterior predictive distribution of  $(Y_{mis}|Y_{obs})$ , or an approximation thereof.

3. But even though the general concept of multiple imputation is easy to implement and software packages that will automatically generate multiply imputed datasets from the original data exist for most standard statistical software, application to real datasets often imposes additional challenges that need to be considered and often can not be handled with of-the-shelf imputation programs. Maintaining all skip patterns and logical constraints in the data is difficult and cumbersome. Besides, depending on the variable to be imputed, it might be necessary to define different imputation models for different subsets of the data to increase the quality of the model and to exclude some variables from the imputation model to avoid multicollinearity problems. These model specifications usually

---

<sup>1</sup>Prepared by Jörg Drechsler (joerg.drechsler@iab.de).

can not be included in the standard software. Furthermore the quality of the imputations needs to be monitored, even if the implicit assumption of a missingness pattern that is missing at random (MAR) can not be tested with the observed data. This does not mean, the imputer can not test the quality of his imputations at all. [Abayomi \*et al.\* \(2008\)](#) suggest several ways of evaluating model based imputation procedures.

4. In this paper we suggest some adjustments for the standard multiple imputation routines to handle the real data problems mentioned above and illustrate a successful implementation of multiple imputation for complex surveys by describing the multiple imputation project for a German establishment survey, the IAB Establishment Panel.

5. The remainder of the paper is organized as follows: In Section II we recapitulate the basic concept of multiple imputation. In Section III we briefly discuss the two main approaches for multiple imputation, joint modeling and sequential regression, and discuss their advantages and disadvantages. In Section IV we present some adjustments for standard multiple imputation routines to handle problems that often arise with real data. We don't claim that the ideas presented in this Section are new. They have been suggested in several other papers. The main aim of this Section is to give an overview of potential problems that are likely to arise in real data applications and to provide a summary of possible solutions in one paper to free the potential multiple imputation user from the burden of a complete literature review in hopes of finding a solution to his specific problem. In Section V we describe results from the multiple imputation project at the German Institute for Employment Research (IAB) that heavily relies on the methods described in Section IV: The multiple imputation of missing values in the German IAB Establishment Panel. In this Section we also discuss the methods we used to evaluate the quality of the imputations. The paper concludes with some final remarks.

## II. The concept of multiple imputation

6. Multiple imputation, introduced by [Rubin \(1978\)](#) and discussed in detail in [Rubin \(1987, 2004\)](#), is an approach that retains the advantages of imputation while allowing the uncertainty due to imputation to be directly assessed. With multiple imputation, the missing values in a dataset are replaced by  $m > 1$  simulated versions, generated according to a probability distribution for the true values given the observed data. More precisely, let  $Y_{obs}$  be the observed and  $Y_{mis}$  the missing part of a dataset  $Y$ , with  $Y = (Y_{mis}, Y_{obs})$ , then missing values are drawn from the Bayesian posterior predictive distribution of  $(Y_{mis}|Y_{obs})$ , or an approximation thereof. Typically,  $m$  is small, such as  $m = 5$ . Each of the imputed (and thus completed) datasets is first analyzed by standard methods designed for complete data; the results of the  $m$  analyses are then combined in a completely generic way to produce estimates, confidence intervals, and test statistics that reflect the missing-data uncertainty properly. In this paper, we discuss analysis with scalar parameters only, for multidimensional quantities see [Little and Rubin \(2002\)](#), Section 10.2 To understand the procedure of analyzing multiply imputed datasets, think of an analyst interested in an unknown scalar parameter  $Q$ , where  $Q$  could be, e.g. the mean of a variable, the correlation coefficient between two variables or a regression coefficient in a linear regression.

7. Inferences for this parameter for datasets with no missing values usually are based on a point estimate  $q$ , a variance estimate  $u$ , and a normal or Student's  $t$  reference distribution. For analysis of the imputed datasets, let  $q_i$  and  $u_i$  for  $i = 1, 2, \dots, m$  be the point and variance estimates achieved from each of the  $m$  completed datasets. To get a final estimate over all imputations, these estimates have to be combined using the combining rules first described by [Rubin \(1978\)](#).

8. For the point estimate, the final estimate simply is the average of the  $m$  point estimates  $\bar{q}_m = \frac{1}{m} \sum_{i=1}^m q_i$  with  $i = 1, 2, \dots, m$ . Its variance is estimated by  $T = \bar{u}_m + (1 + m^{-1})b_m$ , where  $\bar{u}_m = \frac{1}{m} \sum_{i=1}^m u_i$  is the "within-imputation" variance,  $b_m = \frac{1}{m-1} \sum_{i=1}^m (q_i - \bar{q}_m)^2$  is the "between-imputation" variance, and the factor  $(1 + m^{-1})$  reflects the fact that only a finite number of completed-data estimates  $q_i, i = 1, 2, \dots, m$  are averaged together to obtain the final point estimate. The quantity  $\hat{\gamma} = (1 + m^{-1})b_m/T$  estimates the fraction of information about  $Q$  that is missing due to nonresponse.

9. Inferences from multiply imputed data are based on  $\bar{q}_m, T$ , and a Student's  $t$  reference distribution. Thus, for example, interval estimates for  $Q$  have the form  $\bar{q}_m \pm t(1 - \alpha/2)\sqrt{T}$ , where  $t(1 - \alpha/2)$  is the  $(1 - \alpha/2)$  quantile of the  $t$  distribution. [Rubin and Schenker \(1986\)](#) provide the approximate value  $\nu_{RS} = (m - 1)\hat{\gamma}^{-2}$  for the degrees of freedom of the  $t$  distribution, under the assumption that with complete data, a normal reference distribution would have been appropriate. [Barnard and Rubin \(1999\)](#) relax the assumption of [Rubin and Schenker \(1986\)](#) to allow for a  $t$  reference distribution with complete data, and suggest the value  $\nu_{BR} = (\nu_{RS}^{-1} + \hat{\nu}_{obs}^{-1})^{-1}$  for the degrees of freedom in the multiple-imputation analysis, where  $\nu_{obs} = (1 - \hat{\gamma})(\nu_{com})(\nu_{com} + 1)/(\nu_{com} + 3)$  and  $\nu_{com}$  denotes the complete-data degrees of freedom.

### III. TWO APPROACHES TO GENERATE IMPUTATIONS FOR MISSING VALUES

10. Over the years, two different methods emerged to generate draws from  $P(Y_{mis}|Y_{obs})$ : joint modeling and fully conditional specification (FCS), often also referred as sequential regression or chained equations. The first assumes that the data follow a specific multivariate distribution, e.g. a multivariate normal distribution. Under this assumption a parametric multivariate density  $P(Y|\theta)$  can be specified with  $\theta$  representing parameters from the assumed underlying distribution. Within the Bayesian framework, this distribution can be used to generate draws from  $(Y_{mis}|Y_{obs})$ . Methods to create multivariate imputations using this approach have been described in detail by [Schafer \(1997\)](#), e.g., for the multivariate normal, the log-linear, and the general location model.

11. FCS on the other hand does not depend on an explicit assumption for the joint distribution of the dataset. Instead, conditional distributions  $P(Y_j|Y_{-j}, \theta_j)$  are specified for each variable separately. Thus imputations are based on univariate distributions allowing for different models for each variable. Missing values in  $Y_j$  can be imputed for example by a linear or a logistic regression of  $Y_j$  on  $Y_{-j}$ , depending on the character of  $Y_j$ , where  $Y_{-j}$  denotes all columns of  $Y$  excluding  $Y_j$ . 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.

12. In general, imputing missing values by joint modeling is faster and the imputation algorithms are simpler to implement. Furthermore, if the underlying joint distribution can be specified correctly, joint modeling will guarantee valid results with the imputed dataset. However, empirical data will seldom follow a standard multivariate distribution, especially if they consist of a mix of numerical and categorical variables. Besides, FCS provides a flexible tool to account for bounds, interactions, skip patterns or constraints between different variables. None of these restrictions that are very common in survey data can be handled by joint modeling. In practice the imputation task is often centralized at the methodological department of the statistical agency and imputation experts will fill in missing values for all the surveys conducted by the agency. Imputed datasets that don't fulfill simple restrictions like non-negativity or convex constraints will never be accepted by subject matter analysts from other departments. So preserving these constraints is a central element of the imputation task.

13. Overall, joint modeling will be preferable, if only a limited number of variables need to be imputed, no restrictions have to be maintained, and the joint distribution can be approximated reasonably with a standard multivariate distribution. For more complex imputation tasks only fully conditional specification will enable the imputer to preserve constraints inherent in the data. In this case, convergence of the Gibbs sampler should be carefully monitored. A simple way to monitor problems with the iterative imputation procedure is to store the mean of every imputed variable for every iteration of the Gibbs Sampler. A plot of the imputed means over the iterations can indicate if there is only the expected random variation between the iterations or if there is a trend between the iterations indicating problems with the model. Of course no observable trend over the iterations does not guarantee convergence since the monitored estimates can stay stable for hundreds of iterations before drifting off to infinity. Nevertheless, this is a straightforward method to identify flawed imputation models. More complex methods to monitor convergence are discussed in [Arnold \*et al.\* \(1999\)](#) and [Gelman \*et al.\* \(2004\)](#).

#### IV. REAL DATA PROBLEMS AND POSSIBLE WAYS TO HANDLE THEM

14. The basic concept of multiple imputation is straightforward to apply and multiple imputation software like IVEware in SAS ([Raghunathan \*et al.\*, 2002](#)), `mice` ([Van Buuren and Oudshoorn, 2000](#)) and `mi` ([Su \*et al.\*, 2009](#)) in R, `ice` in Stata ([Royston, 2005](#)) (for FCS), and the stand alone packages NORM, CAT, MIX, and PAN ([Schafer, 1997](#))(for joint modeling) further reduce the modeling burden for the imputer. However, simply applying standard imputation procedures to real data can lead to biased or inconsistent imputations. Several additional aspects have to be considered in practice, when imputing real data. Unfortunately most of the standard software with the positive exception of the new `mi` package in R can only handle some of these aspects:

##### A. Imputation of skewed continuous variables

15. One problem that especially arises when modeling business data is that most of the continuous variables like turnover or number of employees are heavily skewed. To control for this skewness, we suggest to transform each continuous variable by taking the cubic root before the imputation. We prefer the cubic root transformation over the log transformation that is often used in the economic literature to model skewed variables like turnover, because the cubic root transformation is less sensitive to deviations between the imputed and the original values in the right tail of the distribution. Since the slope of the exponential function increases exponentially whereas the slope of  $f(x) = x^3$  increases only quadratically, a small deviation in the right tail of the imputed transformed variable has more severe consequences after backtransformation for the log transformed variable than for the variable transformed by taking the cubic root.

##### B. Imputation of semi-continuous variables

16. Another problem with modeling continuous variables that often arises in surveys is the fact that many of these variables in fact are semi-continuous, i.e. they have a spike at one point of the distribution, but the remaining distribution can be seen as a continuous variable. For most variables, this spike will occur at zero. To give an example, in our dataset the establishments are asked how many of their employees obtained a collage degree. Most of the small establishments do not require such high skilled workers. In this case, we suggest to adopt the two step imputation approach suggested by [Raghunathan \*et al.\* \(2001\)](#): In the first step we impute whether the missing value is zero or not. For that, missing values are imputed using a logit model with outcome 1 for all units with a positive value

for that variable. In the second step a standard linear model is applied only to the units with observed positive values to predict the actual value for the units with a predicted positive outcome in step one. All values for units with outcome zero in step one are set to zero.

### C. Imputation under non-negativity constraints

17. Many survey variables can never be negative in reality. This has to be considered during the imputation process. A simple way to achieve this goal is to redraw from the imputation model for those units with imputed values that are negative until all values fulfill the non-negativity constraint. In practice, usually an upper bound  $z$  has to be defined for the number of redraws for one unit since it is possible that the probability to draw a positive value for this unit from the defined model is very low. The value for this unit is set to zero, if  $z$  draws from the model never produced a positive value. However, there is a caveat with this approach. Redrawing from the model for negative values is equivalent to drawing from a truncated distribution. If the truncation point is not at the very far end of the distribution, even simple descriptive analyses like the mean of the imputed variable will significantly differ from the hypothetical true value. For this reason, this approach can only be applied, if the probability to draw negative values from the model is very low and we only want to prevent that some very unlikely unrealistic values are imputed. If the fraction of units that would have to be corrected with this approach is too high, the model is miss-specified and needs to be revised. Usually it is helpful to define different models for different subgroups of the data. To overcome the problem of generating too many negative values, a separate model for the units with small values should be defined. Another possible strategy is implemented in the software package `mi` in R: Take the square root of each strictly positive variable before imputation and backtransform after the imputation. This will guarantee that no negative imputed values can occur.

### D. Imputation under linear constraints

18. In many surveys the outcome of one variable by definition has to be equal to or above the outcome of another variable. For example, the total number of employees always has to be at least as high as the number of part-time employees. When imputing missing values in this situation, [Schenker \*et al.\* \(2006\)](#) suggest the following approach: Variables that define a subgroup of another variable are always expressed as a proportion, i.e. all values for the subgroup variable are divided by the total before the imputation and thus are bounded between zero and one. A logit transformation of the variables guarantees that the variables will have values in the full range  $] - \infty, \infty[$  again. Missing values for these transformed variables can be imputed with a standard imputation approach based on linear regressions. After the imputation all values are transformed back to get proportions again and finally all values are multiplied with the totals to get back the absolute values. To avoid problems on the bounds of the proportions, we suggest setting proportions greater than 0.999999 to 0.999999 before the logit transformation and to use the two step imputation approach described in Section IV.B to determine zero values.

### E. Skip patterns

19. Skip patterns, e.g. a battery of questions are only asked if they are applicable, are very common in surveys. Although it is obvious that they are necessary and can significantly reduce the response burden for the survey participant, they are a nightmare for anybody involved in data editing and imputation or statistical disclosure control. Especially, if the skip patterns are hierarchical, it is very difficult to guarantee that imputed values are consistent with the filtering rules. With fully conditional specification, it is straightforward to generate imputed datasets that are consistent with all

these rules. The two step approach described in Section IV.B can be applied to decide if the filtered questions are applicable. Values are imputed only for the units selected in step one. Nevertheless, correctly implementing all filtering rules is a labor intensive task that can be more cumbersome than defining good imputation models. Furthermore, the filtering can lead to variables that are answered by only a small fraction of the respondents and it can be difficult to develop good models based on a small number of observations.

## V. MULTIPLE IMPUTATION OF MISSING VALUES IN THE IAB ESTABLISHMENT PANEL

20. In this Section we describe results from the multiple imputation project at the German Institute for Employment Research: The imputation of missing values in the wave 2007 of the IAB Establishment Panel.

### A. The dataset

21. The IAB Establishment Panel <sup>2</sup> is based on the German employment register aggregated via the establishment number as of 30 June of each year. The basis of the register, the German Social Security Data (GSSD) is the integrated notification procedure for the health, pension and unemployment insurances, which was introduced in January 1973. This procedure requires employers to notify the social security agencies about all employees covered by social security. As by definition the German Social Security Data only includes employees covered by social security - civil servants and unpaid family workers for example are not included - approx. 80% of the German workforce are represented. However, the degree of coverage varies considerably across the occupations and the industries.

22. Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 17 classes for the industry <sup>3</sup>. These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in Western Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually - since 1996 with over 4,700 establishments in Eastern Germany in addition. In the wave 2007 more than 15,000 establishments participated in the survey. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy.

### B. The imputation task

23. Most of the 284 variables included in the wave 2007 of the panel are subject to nonresponse. Only 26 variables are fully observed. However, missing rates vary considerably between variables and are modest for most variables. 65.8% of the variables have missing rates below 1%, 20.4% of the

---

<sup>2</sup>The approach and structure of the establishment panel are described for example by Fischer *et al.* (2008) and Kölling (2000).

<sup>3</sup>From 2000-2003 20 industry classes were used, before 2000 16 classes were used.

variables have missing rates between 1% and 2%, 15.1% rates between 2% and 5% and only 12 variables have missing rates above 5%. The five variables with missing rates above 10% are *subsidies for investment and material expenses* (13.6%), *payroll* (14.4%), *intermediate inputs as proportion of turnover* (17.4%), *turnover in the last fiscal year* (18.6%), and *number of workers who left the establishment due to restructuring measures* (37.5%). Obviously, the variables with the highest missing rates contain information that is either difficult to provide like *number of workers who left the establishment due to restructuring measures* or considered sensitive like *turnover in the last fiscal year*. The variable *number of workers who left the establishment due to restructuring measures* is only applicable to 626 establishments in the dataset, who declared they had restructuring measures in the last year. Of these 626 only 391 establishments provided information on the number of workers that left the establishment due to these measures. Clearly, it is often difficult to tailor exactly which workers left as a result of the measures and which left for other reasons. This might be the reason for the high missing rates. The low number of observed values are also problematic for the modeling task, so this variables should be used with caution in the imputed dataset.

### C. Imputation models

24. Since the dataset contains a mixture of categorical variables and continuous variables with skewed distributions and a variety of often hierarchical skip patterns and logical constraints, it is impossible to apply the joint modeling approach. We apply the fully conditional specification approach, iteratively imputing one variable at a time, conditioning on the other variables available in the dataset. For the imputation we basically rely on three different imputation models. The linear model for the continuous variables, the logit model for binary variables and the multinomial logit for categorical variables with more than two categories. Multiple imputation procedures for these models are described in [Raghunathan \*et al.\* \(2001\)](#). In general, all variables that don't contain any structural missings are used as predictors in the imputation models in hopes of reducing problems from uncongeniality ([Meng, 1994](#)). In the multinomial logit model for the categorical variables the number of explanatory variables is limited to 30 variables found by stepwise regression to speed up the imputation process. To improve the quality of the imputation we define several separate models for the variables with high missing rates like turnover or payroll. Independent models are fit for Eastern and Western Germany and for different establishment size classes. All continuous variables are subject to non negativity constraints and the outcome of many variables is further restricted by linear constraints. To complicate the imputation process most variables have huge spikes at zero and as mentioned before the filtering rules are often hierarchical. We therefore have to rely on a mixture of the adjustments presented in [Section IV](#). Since the package `mi` was not available at the beginning of this project and other standard packages could not deal with all these problems or did not allow detailed model specification, we use own coding in R for the imputation routines to generate  $m = 5$  datasets.

### D. Checking the quality of the imputations

25. It is difficult to check the quality of the imputations for missing values, since information about the missing values usually by definition is not available and the assumption that the response mechanism is ignorable ([Rubin, 1987](#)), necessary for obtaining valid imputations if the response mechanism is not modeled directly, can not be tested from the observed data. A response mechanism is considered ignorable, if, given that the sampling mechanism is ignorable, the response probability only depends on the observed information.<sup>4</sup> If these conditions are fulfilled, the missing data is said to

---

<sup>4</sup>The additional requirement that the sampling mechanism is also ignorable ([Rubin, 1987](#)), i.e. the sampling probability only depends on observed data, is usually fulfilled in scientific surveys. The stratified sampling design of the IAB establishment panel also satisfies this requirement.

be *missing at random (MAR)* and imputation models only need to be based on the observed information. As a special case, the missing data is said to be *missing completely at random (MCAR)*, if the response mechanism does not depend on the data (observed or unobserved), which implies that the distribution of the observed data and the distribution of the missing data are identical. If the above requirements are not fulfilled, the missing data is said to be *missing not at random (MNAR)* and the response mechanism needs to be modeled explicitly. [Little and Rubin \(2002\)](#) provide examples for non-ignorable missing-data models.

26. As noted before, it is not possible to check, if the missing data is *MAR* or *MCAR* with the observed data. But even if the *MAR* assumption can not be tested, this does not mean, the imputer can not test the quality of his or her imputations at all. [Abayomi et al. \(2008\)](#) suggest several ways of evaluating model based imputation procedures. Basically their ideas can be divided in two categories: On the one hand, the imputed data can be checked for reasonability. Simple distributional and outlier checks can be evaluated by subject matter experts for each variable to avoid implausible imputed values like a turnover of \$ 10 million for a small establishment in the social sector. On the other hand, since imputations usually are model based, the fit of these models can and indeed should be tested. [Abayomi et al. \(2008\)](#) label the former as *external* diagnostic techniques, since the imputations are evaluated using outside knowledge and the latter *internal* diagnostic techniques, since they evaluate the modeling based on model fit without the need of external information.

27. To automate the external diagnostics to some extend, [Abayomi et al. \(2008\)](#) suggest to use the Kolmogorov Smirnov test to flag any imputations for which the distribution of the imputed values significantly differs from the distribution of the observed values. Of course a significant difference in the distributions does not necessarily indicate problems with the imputation. Indeed, if the missing data mechanism is *MAR*, we would expect the two distributions to differ. The test is only intended to decrease the number of variables that need to be checked manually, implicitly assuming that no significant difference between the original and the imputed data indicates no problem with the imputation model.

28. However, we are skeptical about this automated selection method, since the test is sensitive to the sample size, so the chance of rejecting the null hypothesis will be lower for variables with lower missing rates and variables that are answered only by a subset of the respondents. Furthermore it is unclear what significance level to choose and as noted above rejection of the null hypothesis does not necessarily indicate an imputation problem, but even more important not rejecting the null hypothesis is not a guarantee that we found a good imputation model which is implicitly assumed by this procedure.

29. To check for possible flaws in the imputations, we plot the distributions for the original and imputed values for every variable and check if any notable differences between these distributions can be justified by differences in the distributions of the covariates. [Figure 1](#) displays the distributions for two representative variables. Original values are presented in black imputed values in red. The left variable (*payroll*) represents a candidate that we did not investigate further, since the distributions almost match exactly. The right variable (*number of participants in further education (NB.PFE)*) is an example for a variable for which we tried to understand the difference between the distribution of the observed values and the distribution of the imputed values before accepting the imputation model.

30. Obviously, most of the imputed values for the variable *NB.PFE* are significantly larger than the observed values for this variable. To understand this difference, we examine the dependence between the missing rate and the establishment size. In [Table 1](#) we present the percentage of missing units in 10 establishment size classes defined by quantiles and the mean of *NB.PFE* within these quantiles.

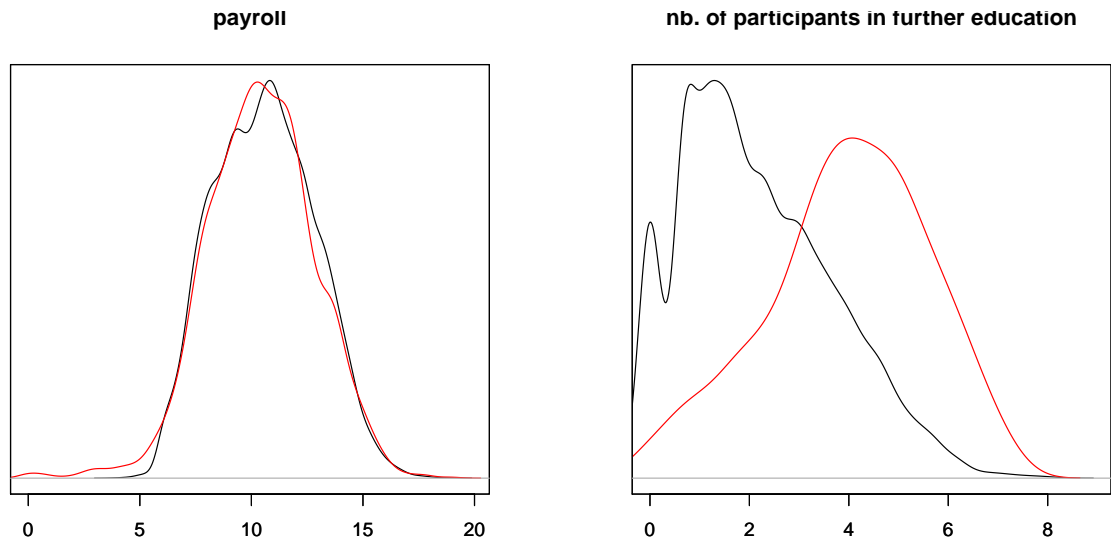


FIGURE 1. Observed (black) and imputed (red) data for payroll and number of participants in further education

The missing rates are low up to the sixth establishment size class. Beyond that point the missing rates increase significantly with every class. The average number of further education participants increases steadily with every establishment size class with largest increases in the second half of the table. With these results in mind, it is not surprising that the imputed values for that variable are often larger than the observed values.

TABLE 1. missing rates and means per quantile for *NB.PRE*.

est. size quantile	missing rate in %	mean(NB.PFE) per quantile
1	0.09	1.61
2	0.00	2.49
3	0.57	3.02
4	0.36	4.48
5	0.44	6.09
6	0.37	9.53
7	0.85	15.48
8	1.16	26.44
9	3.18	56.39
10	6.66	194.09

31. We inspected several continuous variables by comparing the distributions of the observed and imputed values in our dataset and did not find any differences in the distributions that could not be explained by the missingness pattern. However, these comparisons are only meaningful, if enough observations are imputed. Otherwise the distributions between observed data and imputed data might look completely different, only because using kernel density estimation to produce a smooth distribution graph is not appropriate in this context. For this reason we restricted the density comparisons to variables with more than 200 imputed values above zero. For the remaining variables we plotted

histograms to check for differences between the observed and imputed values and to detect univariate outliers in the imputed data.

32. We also investigated if any weighted imputed value for any variable lay above the maximum weighted observed value for that variable. Again, this would not necessarily be problematic, but we did not want to produce any unrealistic influential outliers. However, we did not find any weighted imputed value that was higher than the maximum of its weighted observed counterpart.

33. For the internal diagnostics, we used three graphics to evaluate the model fit: A Normal Q-Q plot, a plot of the residuals from the regression against the fitted values and a binned residual plot (Gelman and Hill, 2006). The Normal Q-Q plot indicates if the assumption of a normal distribution for the residuals is justified by plotting the theoretical quantiles of a normal distribution against the empirical quantiles of the residuals. The residual plot visualizes any unwanted correlation between the fitted values and the residuals. The binned residual plot plots the average fitted value against the average residual within predefined bins. This is especially helpful for categorical variables since the output of a simple residual plot is difficult to interpret if the outcome is discrete. We therefore only focused on the binned residual plots for the discrete variables.

34. Figure 2 again provides an example of one model (one of the models for the variable turnover) that we did not inspect any further and one model (for the variable *number of participants in further education with college degree (NB.PFE.COL)*, for which we checked the model for necessary adjustments.

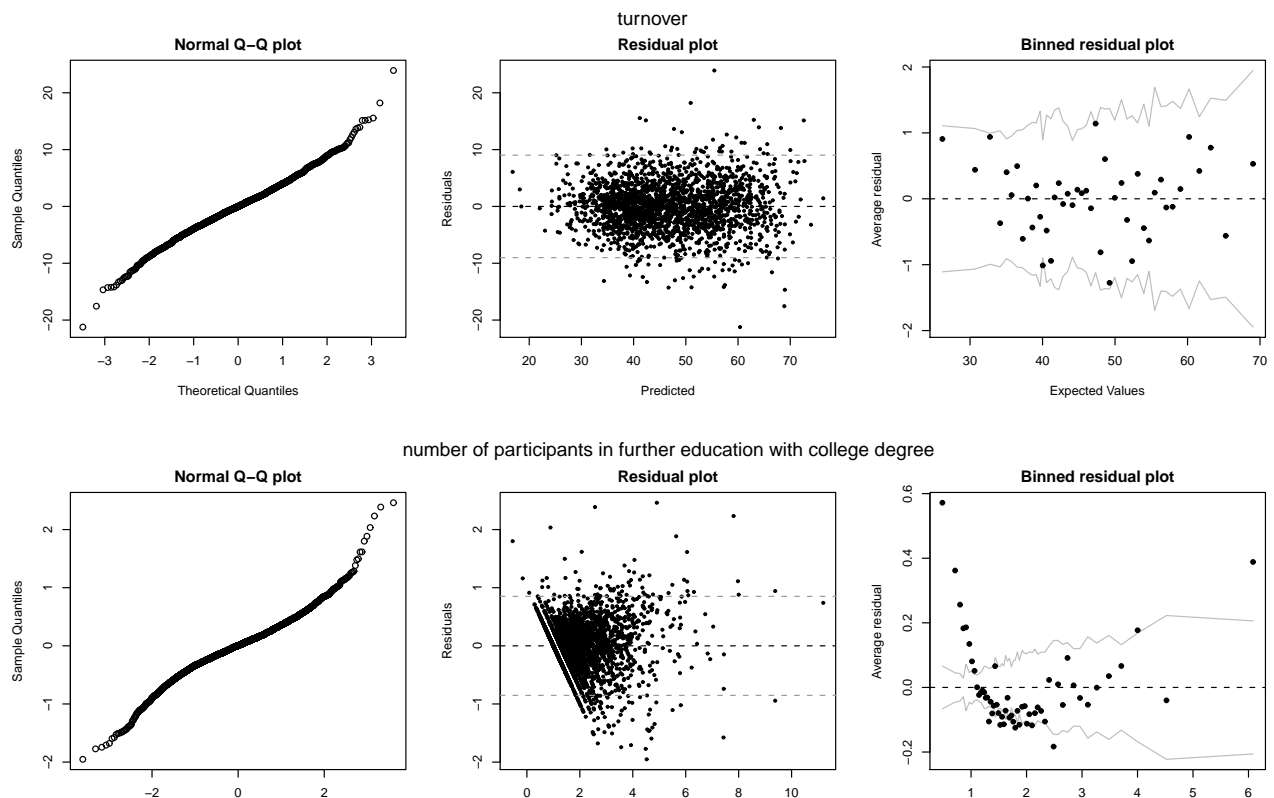


FIGURE 2. model checks for turnover and number of participants in further education with college degree

35. For both variables the assumption that the residuals are more or less normal distributed seems to be justified. For the variable turnover, the two residual plots further confirm the quality of the model. Only a small amount of residuals fall outside of the grey dotted 95% error lines for the residual plot and none of the averaged residuals falls outside the grey 95% error lines for the binned residuals. This is different for *NB.PFE.COL*. Although still most of the points are inside the 95% error lines, we see a clear relationship between the fitted values and the residuals for the small values and the binned residuals for these small values all fall outside the error lines. However, this phenomenon can be explained if we inspect the variable further. Most establishments don't have any participants in further training with college degree and we fitted the model only to the 3426 units that reported to have at least one participant. 648 of these units report that they had only 1 participant, leading to a spike at 1 in the original data. Since we simply fit a linear model to the observed data, the almost vertical line in the residual plot is not surprising. It contains all the residuals for all the units with only 1 participant in the original data. The binned residual plot indicates that the small fitted values sometimes severely underestimate the original values. The reason for this again is the fact that the original data is truncated at 1 whereas the fitted values are predictions from a standard linear model that would even allow negative fitted values, since we computed the fitted values before the adjustments for non-negativity described in Section IV.C. The consequence is a light overestimation for the larger fitted values.

36. We found similar patterns in some other variables that had huge spikes at 1. We could have tried to model the data with a truncated distribution or we could have applied the semi-continuous approach described in Section IV.B to model the spike at 1 separately, but since we expect that the non-negativity adjustments reduce this effect and we found that the fitted values are almost normally distributed around 1 for the units with 1 participant, we decided to avoid making the already complex modeling task even more difficult.

37. Missing rates are substantially lower for the categorical variables. Only 59 out of the close to 200 categorical variables in the dataset have missing rates above 1% and we limited our evaluation to these variables. We compared the response rates in each category for the observed and the imputed values and flagged a variable for closer inspection, if the response rate in one imputed category differed more than 20% from the response rate in the observed category. We further limited our search to categories that contained at least 25 units, since small changes in smaller categories would lead to significant changes in the relative differences for these categories. All 15 variables that were flagged by this procedure had a missing rate below 5% and the differences between the imputed and original response rates could be explained by the missingness pattern for all of them. We select one variable here to illustrate the significant differences between observed and imputed values that can arise from a missingness pattern that is definitely not missing completely at random. The variable under consideration asks for the expectations about the investment in 2007 compared to 2006. Table 2 provides some summary statistics for this variable. We find a significant difference, if we simply compare the observed response rates for each category (column 1) with the imputed response rates for each category (column 2). But if we consider that the missing rate is only 0.2% for this variable, if the unit reported any investments in 2006 but soars to 10.2% if the unit reported that it had no investments in 2006, it is not surprising that the response rates for the imputed values are influenced by the expectations for those units that had no investments in 2006 (column 4) even though only 11.7% of the participants reported no investments in 2006. These response rates differ completely from the response rates for units that reported investments in 2006. Thus the percentage of establishment that expect an increase in investments is significantly larger in the imputed data than it is in the original data.

38. We also examined the binned residual plots for the 59 categorical variables with missing rates above 1%. All plots indicated a good model fit. Graphics are omitted for brevity.

TABLE 2. Expectations for the investments in 2007 compared to 2006 (response rates in % for each category)

category	observed data	imputed data	obs. units with investment 2006	obs. units without investment 2006
will stay the same	36.57	29.20	41.33	0.59
increase expected	38.79	57.66	30.74	99.41
decrease expected	20.33	9.49	23.05	0.00
don't know yet	4.31	3.65	4.88	0.00

39. To check for possible problems with the iterative imputation procedure, we stored the mean for several continuous variables after every imputation round. We did not find any inherent trend for the imputed means for any of the variables. Of course, this is no guarantee for convergence. A possible strategy to measure the convergence of the algorithm that we did not implement in our imputation routines is discussed in [Su et al. \(2009\)](#).

## VI. CONCLUDING REMARKS

40. More than 30 years after its initial proposal by [Rubin \(1978\)](#), multiple imputation has been widely accepted as the best practice to deal with item nonresponse in surveys. Nevertheless, the implementation of good imputation routines is a difficult and cumbersome task and an unsupervised imputation using standard imputation software can harm the results more than simple complete case analysis. In this paper we discussed several problems that often arise when imputing real datasets, severely complicating the imputation process. We summarized several ideas to handle these problems and the detailed discussion of the imputation of missing values in the German IAB Establishment Panel illustrates that generating multiply imputed datasets with high data quality for complex surveys is not an impossible task. However, there are still several ways for improvements and open research questions. We found that defining several independent models for each variable can lead to significant improvements in the imputations. To reduce the runtime of the procedure and to keep the programming burden low, we applied this strategy only to a few variables with very high missing rates. Using this strategy for all variables whenever possible would further improve the quality of the imputations. A definitive shortcoming in our implementation is the limitation to only 30 explanatory variables in the multinomial variables that was necessary due to multicollinearity problems and to speed up the imputation process. This might lead to uncongeniality problems ([Meng, 1994](#)) if the analyst uses explanatory variables in his or her model that were not included in the imputation model. A possible strategy to overcome these limitations would be to use CART models for the categorical variables. [Reiter \(2005\)](#) suggests this strategy in the context of multiple imputation for statistical disclosure control. Further research is needed to investigate under which circumstances this approach is also applicable for multiple imputation for nonresponse.

## References

- Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, C* **57**, 273–291.
- Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models*. Springer.

- Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika* **86**, 948–955.
- Fischer, G., Janik, F., Müller, D., and Schmucker, A. (2008). The iab establishment panel - from sample to survey to projection. Tech. rep., FDZ-Methodenreport, No. 1 (2008).
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis: Second Edition*. London: Chapman & Hall.
- Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Kölling, A. (2000). The iab-establishment panel. *Journal of Applied Social Science Studies* **120**, 291–300.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. New York: John Wiley & Sons.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science* **9**, 538–558.
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.
- Raghunathan, T. E., Solenberger, P., and van Hoewyk, J. (2002). Iweware: Imputation and variance estimation software. Available at: <http://www.isr.umich.edu/src/smp/ive/>.
- Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.
- Royston, P. (2005). Multiple imputation of missing values: Update of ice. *The Stata Journal* **5**, 527–536.
- Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Section on Survey Research Methods*, 20–34. American Statistical Association.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician* **58**, 298–302.
- Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**, 924–933.
- Su, Y., Gelman, A., Hill, J., and Yajima, M. (2009). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software* forthcoming.
- Van Buuren, S. and Oudshoorn, C. (2000). Mice v1.0 user’s manual. report pg/vgz/00.038. Tech. rep., TNO Prevention and Health, Leiden.