Journal of Official Statistics, Vol. 21, No. 2, 2005, pp. 325–341

PRIMA: A New Multiple Imputation Procedure for Binary Variables

Ralf Münnich[1] *and Susanne Rässler*[2]

When investigating unemployment data, one may be interested in estimating the totals of unemployed in subpopulations, e.g., by regional or by contentional differentiation. For the estimation of a population total, typically Horvitz-Thompson type estimates are used. However, often the data are prone to item nonresponse. To achieve valid results for these estimates from a randomization-based perspective, in general variance correction methods are needed.

In this article we discuss imputation issues in large-scale datasets with different scaled variables, laying special emphasis on binary variables. Since fitting a multivariate imputation model can be cumbersome, univariate specifications are proposed which are much easier to perform. The regression-switching or chained equations Gibbs sampler is proposed and possible theoretical shortcomings of this approach are addressed as well as data problems.

A simulation study is done based on the data of the German Microcensus, which is often used to analyze unemployment. Multiple imputation, raking, and calibration techniques are compared for estimating the number of unemployed in different settings. We find that the logistic multiple imputation routine for binary variables, in some settings, may lead to poor point estimates as well as variance estimates. To overcome possible shortcomings of the logistic regression imputation, we derive a multiple imputation-matching algorithm, which turns out to work well.

Key words: Complex surveys; missing data; multiple imputation; logistic regression; Horvitz-Thompson estimator; GREG estimator.

1. Introduction

Typically, in complex surveys we are confronted with variables of different scales with some missing data spread all around. Theoretically, we can apply a multivariate model and derive suitable imputation routines therefrom. Practically, this is not an easy task. The multivariate normal model (see e.g., Schafer 1999b) has become quite popular among statisticians for multiple imputations in multivariate settings. But in many applications, the assumption of a multivariate normal distribution can hardly be justified, for example when binary variables have missing data. Recently, Rubin (2003) has suggested univariate multiple imputation procedures for large-scale datasets. They are successfully used for

[1] Department of Statistics, Econometrics, and Operations Research, Eberhard Karls University of Tübingen, Mohlstraße 36, 72074 Tübingen, Germany. E-mail: ralf.muennich@uni-tuebingen.de

[2] Institute for Employment Research of the Federal Employment Agency, Regensburger Str. 104, 90478 Nürnberg, Germany. E-mail: susanne.raessler@iab.de

Acknowledgments: This study was conducted in connection with the DACSEIS project which is financially supported within the IST programme under FP5 by the European Commission in strong cooperation with Eurostat. For further project details we refer to Münnich and Wiegert (2001) or http://www.dacseis.de. Our special thanks go to Donald B. Rubin for providing always motivating comments on the multiple imputation study.

multiple imputation in the U.S. National Medical Expenditure Survey (NMES) where the dataset to be imputed consists of up to 240 variables of different scales and 22,000 observations. Such routines have been used quite efficiently in the context of *mass imputation*, i.e., imputing a large quantity of data that are missing by design. This is the situation in the so-called data fusion case and the split questionnaire survey designs (see e.g., Rässler 2002). For the pseudouniverses and the simulation study to be performed within the DACSEIS project (cf. http://www.dacseis.de), we therefore suggest such multiple imputation routines as state-of-the-art. The advantages and disadvantages of this approach are described herein, also the necessary pseudocode is provided in S-PLUS/R routines on the electronic *DACSEIS recommended practice manual* (cf. http://rpm.dacseis.de).

It is said that iterative univariate imputations were first implemented by Kennickell (1991) and Kennickell and McManus (1994) (see Schafer and Olsen 1999). Ready to use and available for free via the Internet is a software called MICE, which is a recent implementation of some iterative univariate imputation methods in S-PLUS as well as R (see van Buuren and Oudshoorn 2000). Moreover, there is the free SAS-callable application IVEware, which also provides iterative univariate imputation methods. The intuitively appealing idea behind the iterative univariate imputation procedure is to overcome the problem of suitably proposing and fitting a multivariate model for mixtures of categorical and continuous data by reducing the multivariate imputation task to conventional regression models iteratively completed. In many surveys it may be difficult to propose a sensible joint distribution for all variables of interest. On the other hand there are a variety of procedures available for regression modelling of continuous and categorical univariate response variables such as ordered or unordered logit/probit models (see Greene 2000). Thus any plausible regression model may be specified for predicting each univariate variable that has to be imputed given all the other variables. This approach is known as regression switching, sequential regressions, chained equations, or variable-by-variable Gibbs sampling; see e.g., van Buuren and Oudshoorn (1999). In the chained equations Gibbs sampling approach it is also possible to include only relevant predictor variables, thus reducing the number of parameters.

2. Univariate Multiple Imputation Models for Complex Surveys

2.1. Multiple imputation

The theory and principle of multiple imputation (MI) originates from Rubin (1978). A comprehensive treatment of data augmentation and multiple imputations can be found in Schafer (1997). An introduction to MI is also given by Schafer (1999a), Little and Rubin (2002) and in the context of the DACSEIS project (Rässler 2004). The theoretical motivation for multiple imputation is Bayesian, although the resulting multiple imputation inference is usually also valid from a frequentist viewpoint. Basically, MI requires independent random draws from the posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$ of the missing data Y_{mis} given the observed data Y_{obs}. Since it is often difficult to draw from $f_{Y_{mis}|Y_{obs}}$ directly, a two-step procedure for each of the m draws is useful:

(a) First, we make random draws of the parameters Ξaccording to their observed-data posterior distribution $f_{\Xi|Y_{obs}}$

(b) then, we perform random draws of Y_{mis} according to their conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$

Because

$$f_{Y_{mis}|Y_{obs}}(y_{mis}|y_{obs}) = \int f_{Y_{mis}|Y_{obs},\Xi}(y_{mis}|y_{obs}, \xi) f_{\Xi|Y_{obs}}(\xi|y_{obs}) d\xi \tag{1}$$

holds, with (a) and (b) we achieve imputations of Y_{mis} from their posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$. Due to the data-generating model used, for many models the conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$ is rather straightforward. Often it can be easily formulated for each unit with missing data.

In contrast, the corresponding observed-data posteriors $f_{\Xi|Y_{obs}}$ are usually difficult to derive for those units with missing data, especially when the data have a multivariate structure and different missing data patterns. The observed-data posteriors are often not standard distributions from which random numbers can easily be generated. However, simpler methods have been developed to enable multiple imputation based on Markov chain Monte Carlo (MCMC) techniques. In MCMC the desired distributions $f_{Y_{mis}|Y_{obs}}$ and $f_{\Xi|Y_{obs}}$ are achieved as stationary distributions of Markov chains which are based on the easier to compute complete-data distributions.

Typically, $m = 5$ values are imputed for each missing datum according to some distributional assumptions. We conduct, say, $m > 1$ independent simulated imputations $(Y_{obs}, Y_{mis}^{(1)})$, $(Y_{obs}, Y_{mis}^{(2)})$, $\ldots$,$(Y_{obs}, Y_{mis}^{(m)})$. Then standard complete-case analysis can be performed for each of the m imputed datasets, enabling us to calculate the imputed data estimate $\hat{\theta}^{(t)} = \hat{\theta}(Y_{obs}, Y_{mis}^{(t)})$ along with its estimated variance $\hat{V}(\hat{\theta}^{(t)}) = \hat{V}(\hat{\theta}(Y_{obs}, Y_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. Finally the complete-case estimates are combined according to the MI rule, which means that the MI point estimate for θ is simply the average

$$\hat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}^{(t)} \tag{2}$$

To obtain a standard error $\sqrt{\hat{V}(\hat{\theta}_{MI})}$ for the MI estimate $\hat{\theta}_{MI}$, we first calculate the *between-imputation* variance

$$\hat{V}_b(\hat{\theta}) = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2 \tag{3}$$

and then the *within-imputation* variance

$$\hat{V}_w(\hat{\theta}) = \frac{1}{m} \sum_{t=1}^{m} \hat{V}(\hat{\theta}^{(t)}) \tag{4}$$

Finally, the estimated total variance is defined by

$$\hat{V}(\hat{\theta}_{MI}) = \hat{V}_w(\hat{\theta}) + (1 + \frac{1}{m}) \hat{V}_b(\hat{\theta}) \tag{5}$$

For large sample sizes, tests and two-sided $(1 - \alpha)100\%$ interval estimates can be based on Student's t-distribution

$$(\hat{\theta}_{MI} - \theta)/\sqrt{\hat{V}(\hat{\theta}_{MI})} \sim t_v \quad \text{and} \quad \hat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{\hat{V}(\hat{\theta}_{MI})} \tag{6}$$

with the degrees of freedom

$$v = (m-1)\left(1 + \frac{\hat{V}_w(\hat{\theta})}{(1+m^{-1})\hat{V}_b(\hat{\theta})}\right)^2 \tag{7}$$

MI is in general applicable when the complete-data estimates are asymptotically normal (as ML estimates are) or t-distributed; see e.g., Rubin and Schenker (1986), Rubin (1987), Barnard and Rubin (1999), or Little and Rubin (2002).

From (6) we can see that the multiple imputation interval estimates are widened to reflect the uncertainty due to imputation and the model choice. If this in done *correctly* the MI method is said to be proper in Rubin's sense. From a frequentist perspective this means that randomization valid inference will be drawn from the multiply imputed datasets; for further discussion see Brand (1999) or Rässler (2004). To show analytically whether an MI method is proper is a difficult task and often only simulation studies can help. In our setting, the logistic regression imputation is obviously not proper. Therefore we propose another routine, which is based on the proper linear regression imputation method.

2.2. *Regression-switching*

To illustrate the principle of the regression switching let us assume the simple case with three variables A, B, and C, each with missing data. Then Rubin (2003) proposes:

- Begin by arbitrarily filling in all missing B and C values.
- Then, fit a model of $A|B, C$ using those units where A is observed, and impute the missing A values.
- Next, toss the imputed B values, and fit a model of $B|A, C$ using those units where B is observed, and impute the missing B values.
- Next, toss the imputed C values, and fit a model of $C|A, B$ using units where C is observed, and impute the missing C values.
- Iterate.

Since this describes a Markov chain Monte Carlo (MCMC) procedure, in general, the changed equations are cycled through until convergence of the chains can be expected. This procedure allows great flexibility due to the different conditional specifications. Each specification simply is a univariate regression. It has to be mentioned that there are some theoretical shortcomings, because it is possible to generate incompatible distributions via implicit contradictions in the specified conditional specifications. The practical implications of this phenomenon in iterative univariate imputation are still quite unknown (see Schafer and Olsen 1999). A *real* Gibbs sampler starts with an existing but intractable joint distribution for the variables of interest, iteratively generating random variables from easier to operate full conditional distributions derived from its joint distribution. In the context of iterative univariate imputations the conditional distributions are specified in the hope that these conditional distributions will define a suitable joint model. However, even if there is no such joint distribution for the data, the MCMC method can be implemented,

and each conditional specification may be a good empirical fit to the data (see Rubin 2003, van Buuren and Oudshorn 2000, and Brand 1999).

2.3. MI algorithm for missing continuous variables

To impute missing data for a continuous variable Y, such as income or expenditure in household budget surveys, we propose to apply a simple linear regression model. As prior distributions the usual uninformative or flat priors are used. The continuous variable (income or expenditure data, whichever has data missing) may be transformed by its logarithm before performing the imputations. Note that after imputation the values have to be transformed back. To assure that only values are imputed that lie within a certain range, also upper and lower bounds can be given. After performing the final imputation step for the missing Y values, each row of the imputed dataset is examined to see whether any of the imputed values is out of range. In such cases these values are redrawn until the constraints are satisfied. According to Schafer (1997, p. 204), this procedure leads to approximately proper multiple imputations under a truncated normal model.

The basic algorithm according to Rubin (1987, p. 166) is as follows.

- Assume the underlying data model of a linear regression
$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$
- Assume that y has n_{mis} missing data, variables X are fully observed or already imputed. y_{obs} and X_{obs} refer to the jointly observed part, X_{mis} to the missing part y_{mis}.
- Let $\hat{\beta}$ and $\hat{\sigma}^2 = (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})/(n_{obs} - p)$ be the least squared estimates from the observed data.
- Multiple imputation procedure for $j = 1, 2, \ldots m$:
  1. Draw $\sigma^2 | X \sim (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})\chi^{-2}_{n_{obs}-p}$
  2. Draw a vector of p variables from $\beta | \sigma^2, X \sim N(\hat{\beta}, \sigma^2 (X'_{obs} X_{obs})^{-1})$
  3. Draw $Y_{mis} | \beta, \sigma^2, X \sim N(X_{mis}\beta, \sigma^2)$ independently for every missing value $i = 1, 2, \ldots, n_{mis}$

For the independent variables X all available auxiliary variables from the universes may be taken. If both income and expenditure have missing values, then the regression switching can be applied. Since income and expenditure are typically highly correlated, it seems adequate to incorporate both variables in one imputation model.

2.4. MI algorithm for missing binary variables

For a binary target variable with missing values, we first base the imputations on a logistic regression model. For labor force surveys or the microcensus and its data, the employment variable has to be either recoded to zero and one (e.g., to 1 if $y =$ unemployed and to 0 otherwise) or split into dummy variables (e.g., to employment (yes/no) and unemployment (yes/no)), leaving the third category for nonlabor force or simply the rest. Then the following algorithm is proposed by Rubin (1987, p. 169).

- Assume the underlying data model of a logistic regression
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = X\beta, \quad p = P(Y = 1|X)$$

- Assume that y has n_{mis} missing data, variables X are fully observed or already imputed. y_{obs} and X_{obs} refer to the jointly observed part, X_{mis} to the missing part y_{mis}.
- Let $\hat{\beta}$ be the iterative least squares estimates from the observed data (or any approximate ML estimate) and $\hat{V}(\hat{\beta})$ its estimated covariance matrix (e.g., from the inverse Fisher information matrix $I(\hat{\beta})^{-1}$).
- Apply the large sample normal approximation for $j = 1, 2, \ldots, m$:
  1. Draw a vector of p variables from $\beta|X \sim N(\hat{\beta}, \hat{V}(\hat{\beta}))$
  2. For every $i \in mis$ calculate $p_i = 1/(1 + \exp(-X_i'\beta))$
  3. Draw n_{mis} independent uniform (0,1) random numbers u_i for $i = 1, 2, \ldots, n_{mis}$ and if $u_i > p_i$ impute $Y_i = 0$, otherwise impute $Y_i = 1$.

For the independent variables X again all available information may be taken. If more than one (dummy) variable (after recoding) has missing data, then we may impute the most populous category first versus the rest. If zero is imputed, then we impute the next category versus the rest, and so on. If one is imputed once, all remaining categories are set to zero.

2.5. *Refined MI algorithm for binary variables*

First experiences with the logistic regression approach show that in a couple of cases the large sample approximation for the estimated variance $\hat{V}(\hat{\beta})$ of the logistic regression does not work well. According to Rubin and Schenker (1987) this is due to the diverging shapes of the true and the approximated likelihood, and a stabilizing prior may be used instead of the standard noninformative prior. Alternatively, the computer-intensive importance sampling-resampling algorithm may be used instead of the large sample approximation. To get a computationally easy and fast-working solution for better imputations we propose a new algorithm based on the simple linear MI routine of Section 3. First imputations according to this linear model are created for all units missing or not, although the dependent Y variable is a binary one. Then, for each unit with missing Y values, from the units with observed Y values the unit with the nearest imputed Y_{imp} is searched and its Y_{obs} value is imputed. This procedure basically follows the *predictive mean matching* approach by Rubin (1986) and Little (1988) though the matching is not done on the mean but on the imputed values. We will call this approach *predictive imputation matching* (PRIMA), which was earlier denoted by MI linbin throughout the DACSEIS simulation study (cf. http://rpm.dacseis.de). Whether this approach leads to proper imputations has to be left to future research. However, the results are quite encouraging, as will be shown later in this article.

3. Typical Problems That May Occur with Regression-switching

3.1. *Incorporating the sampling design*

In the multiple imputation model, stratification can be incorporated by including strata indicators as covariates. Clustering may be incorporated by multilevel models that include random cluster effects (see Little and Rubin 2002, p. 90, or Schafer and Yucel 2002). On the other hand, these effects can be controlled by a design-based complete-data inference.

3.2. Collinearity

It should be noted that the problem of collinearity might occur when the coefficients of the linear regression model are to be estimated from the observed data. If some covariates of X show only very little variability it may happen that the remaining part in X_{obs}, which belongs to the observed part of Y, has one or more variables with constant values. To avoid such cases, it is useful to perform plausibility checks to exclude such variables from the regression model.

3.3. Discrete data

For continuous discrete data we propose proceeding with the linear regression model and the PRIMA or directly using the predictive mean matching approach as already mentioned above.

Both procedures have the great advantage that only values, which are really observed, can be imputed and the imputation is even more robust against misspecification of the linear model (see Rässler et al. 2002).

3.4. Monotone missingness versus incompatibility

The regression-switching approach can suffer from the theoretical limitation of possibly generating incompatible distributions via implicit contradictions on their conditional specification (Rubin 2003). This may be of importance if the rate of missingness is high and there are a lot of different conditional specifications. If the missingness has a monotone pattern (for illustration see Fig. 1) we may impute the variable from left to right always regressing the variable to be imputed on all other variables on the left. This procedure is continued until all missing values have been imputed.

Monotone patterns of missingness have the advantage that they can be handled noniteratively because each imputation model being fit is conditioning only on the variables to the left. The resulting univariate models are automatically distributionally compatible. When the missing data are not monotone but *approximately* monotone, Rubin (2003) proposes first imputing those values, which disturb the monotone missingness pattern regressing on all other variables (thus, possibly specifying an incompatible Gibbs sampler) and then impute the monotone missing values from left to right.

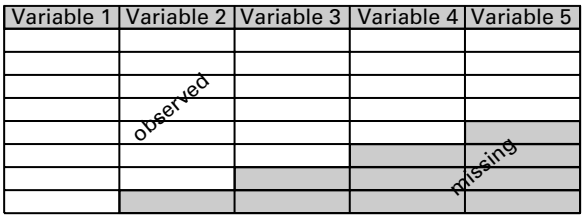

Fig. 1. Monotone pattern of missingness

4. The Multiple Imputation Study of the German Microcensus

4.1. Estimating unemployment in the German Microcensus

The German Microcensus (GMC) is a 1% sample survey of persons and households that is conducted on a yearly basis (cf. *Statistisches Bundesamt 1999*). The main purpose is to analyse the structure of the population, the labor market, including the labor participation, and the housing situation. As in Germany no separate Labor Force Survey exists, the relevant unemployment figures, which refer to the ILO definition of unemployment are drawn from the GMC.

The sampling design is stratified cluster sampling, where the strata are the 214 regional classes within the federal states and finally the five house size classes. Within these 1,070 strata, clusters are built of approximately 15–20 persons. From every 100 of these clusters, one is drawn randomly. Further details on the GMC can be found in Heidenreich (1994), Meyer (1994), Münnich (2001), or Quatember (2002).

In this study, the number of unemployed in Germany is of interest. Unfortunately, in Germany two definitions of unemployment exist. One refers to the labor participation similar to the Eurostat definition as unemployment type U_1. Data for this variable are gained from the variable EF504 in the GMC. Unemployment type U_2 which refers to job-seeking people who are registered at the Federal Employment Agency (*Bundesagentur für Arbeit*) in Nürnberg and is therefore different from U_1.

Subsequently, we will use as the estimation variable U_1, which is denoted by Y. U_2 will be used as an appropriate auxiliary variable, denoted by X. The first-order inclusion probabilities in the GMC are all $\pi_i = 1/100$ for $i = 1, \ldots, N$.

4.2. Estimators of interest

The estimators of interest are based on the classical Horvitz-Thompson estimator for a total τ

$$\hat{\tau}_{HT} = \sum_{i=1}^{n} d_i \cdot y_i \tag{8}$$

with observed values y_i and weights i. In general, the d_i are the design weights $1/\pi_i$ where π_i denotes the first-order inclusion probability (cf. Särndal et al. 1992, or Lohr 1999).

Then, the well-known GREG estimator can be expressed as a linear weighted sum of the observed values

$$\hat{\tau}_{GREG} = \sum_{i=1}^{n} w_i \cdot y_i \tag{9}$$

where $w_i = d_i \cdot g_i$ are the GREG weights with

$$g_i = 1 + \left(\sum_{k=1}^{N} x_k - \sum_{k=1}^{n} d_k \cdot x_k \right)' \left(\sum_{k=1}^{n} d_k \cdot x_k \cdot x_k' \right)^{-1} \cdot x_i \tag{10}$$

and the individual auxiliary information vector x_i. However, the w_i can also be derived from calibration and raking approaches which are discussed in D'Arrigo and Skinner (2004).

As the corresponding variance estimator, the jackknife linearized variance which considers the calibrated weights was chosen. In the case of stratified random sampling with H strata one obtains the following (D'Arrigo and Skinner, 2004, p. 8):

$$\hat{V}(\hat{\tau}_{GREG}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} \left(w_{hj} \cdot e_{hj} - \frac{1}{n_h} \sum_{k=1}^{n_h} w_{hk} \cdot e_{hk} \right)^2 \tag{11}$$

$e_{hj} = y_{hj} - x_{hj} \cdot \hat{B}$ denote the estimated residuals obtained by using the least squares estimate $\hat{B}$ of the multiple regression coefficient. This will be used later in the simulation study of the German Microcensus.

In the case of nonresponse, the design weights and sample sizes have to be updated according to the actual nonresponse.

The combination of point estimator (9) and variance estimator (11) was studied in detail in D'Arrigo and Skinner (2004) and turned out to be a recommended method in the presence of nonresponse and was therefore used as a benchmark in many of the DACSEIS simulations (cf. DACSEIS recommended practice manual).

The second estimator using weights for compensating for nonresponse is defined according to Lundström and Särndal (2002, Sections 6.3 and 6.4). The variance estimator is given by

$$\hat{V}(\hat{\tau}) = \hat{V}_{SAM}(\hat{\tau}) + \hat{V}_{NR}(\hat{\tau}) \tag{12}$$

where

$$\hat{V}_{SAM}(\hat{\tau}) = \sum_{i=1}^{r} \sum_{j=1}^{r} (d_i d_j - d_{ij}) (g_i v_{si} e_i) (g_j v_{sj} e_j) - \sum_{i=1}^{r} d_i (d_i - 1) v_{si} (v_{si} - 1) (g_i e_i)^2 \tag{13}$$

$$\hat{V}_{NR}(\hat{\tau}) = \sum_{i=1}^{r} d_i^2 v_{si} (v_{si} - 1) e_i^2 \tag{14}$$

The weights v_{si} are drawn from

$$v_{si} = 1 + \left(\sum_{k=1}^{n} d_k x_k - \sum_{k=1}^{r} d_k \cdot x_k \right)' \left(\sum_{k=1}^{r} d_k \cdot x_k \cdot x_k' \right)^{-1} \cdot x_i \tag{15}$$

with r respondents in the sample, the g_k are the same as before and the e_k are gained from the least squares estimate of the regression coefficient on the set of respondents.

4.3. Simulation results

The data used are the same, which are used within the DACSEIS simulation study. For a detailed description of the universe dataset generation mechanism we refer to Münnich and Schürle (2003). The simulation set-ups are best viewed in Münnich and Magg (2004).

The universe of Saarland (SAL) consists of 1,089,381 individuals. Within the Monte-Carlo study, 10,000 samples were evaluated for point estimation and 1,000 for variance estimation. This choice was based on some very computer-intensive resampling-based

variance estimation procedures within the entire DACSEIS study while providing an adequate benchmark of the variance $V(\hat{\theta})$.

The nonresponse mechanism was introduced into the estimation variable as item nonresponse. In the case of the German Microcensus, this seems appropriate since the unemployment variable refers to the register variable. In practice, the assignment of units between the two unemployment variables is still problematic for legal reasons. However, within the simulation study, this link could be achieved. The influence of poor assignments of units was studied in Wiegert and Münnich (2004).

The nonresponse rate is 25% in all simulations below. Other nonresponse rates and further estimators can be taken from the DACSEIS recommended practice manual (cf. http://rpm.dacseis.de).

Four simulation tasks are described as follows:

Task 1: The target variable is the total U_1 of unemployed in Saarland. The auxiliary variable for estimation and calibration is the number of job-seeking, U_2. The universe is referred to as `sub0`. The number of unemployed is $\theta_2 = 38,713$ and the number of individuals is $N_1 = 1,089,381$.

Task 2: The simulation of Task 1 is performed on the house-size classes 1–3 in regional stratum 1 of Saarland. The number of unemployed is now $\theta_2 = 12,222$ and, the number of individuals is $N_2 = 312,137$. The universe is referred to as `sub5`.

Task 3: The simulation of Task 1 is now performed on a subset of Saarland, the house-size classes 3 in all three regional strata. Here, the number of unemployed is $\theta_3 = 3,784$ and the number of individuals is $N = 52,635$. The universe is referred to as `sub1`.

Task 4: The target variable is the unemployed U_1 in the age class $[45; 65)$ in the entire population of Saarland (`sub0`; $N = 1,089,381$). In this case, only $\theta_4 = 15,274$ individuals are unemployed. Again, the number of job seeking, U_2, was used as auxiliary variable for the estimation and imputation process.

The tasks were chosen according to the size of the universe, the proportion of unemployed in the areas of interest, and the different stratifications (`sub1` and `sub5`).

The estimators of interest are (in parenthesis the numbers when only four are shown):

1 (1): GREG estimator with nonresponse correction and jackknife linearized variance estimator
2 (2): Calibration estimator according to Lundström and Särndal
3 (–): Horvitz-Thompson estimator with logit MI (cf. Section 2.4)
4 (3): Horvitz-Thompson estimator with PRIMA (cf. Section 2.5)
5 (–): GREG estimator with logit MI
6 (4): GREG estimator with PRIMA

For the Horvitz-Thompson and the GREG estimator in connection with MI, the g-weights-residual variance estimator was applied. Within the simulation study, hardly any differences occur for the point estimators beyond the recommended $m = 5$ imputed datasets. For variance estimation, there was some increase in efficiency when using $m = 15$ but little further increase was observed for $m = 30$. Since the computation effort was rather small on these datasets for the MI routines, we chose to take $m = 30$ imputations within this simulation.

The dark lines in the graphs indicate the true value whereas the light lines indicate the average of the simulated distribution of the estimator of interest. The measures graphs contain the MSE values (left) and the coverage rates (right) for the estimators. The upper bar indicates the true relative root MSE of the estimator, the middle bar the simulated relative standard error, and finally the lower bar the average of the estimated relative standard errors from the samples within the simulation. The coverage rates considered were 90% (upper bar) and 95% (lower bar).

With regard to task one, in Fig. 2 the performance of the six point and variance estimators is shown.

All estimates yield comparable results. The GREG calibration with nonresponse correction, however, is slightly preferable to the other estimators. The Horvitz-Thompson estimator is certainly inferior to the other estimators. The MSE-based measures and coverage rates indicate rather small differences (cf. Fig. 3). However, in this case, the MI logistic regression seems to have a small advantage as compared to the MI PRIMA routine with respect to the coverage rates. This tends to be more a problem of the GREG estimator in comparison with the standard Horvitz-Thompson.

In Fig. 4 we see that turning to the subpopulation while still using the number of unemployed as the target variable, the estimators using logit imputation become unstable. The variance estimator itself seems a little more stable in comparison with the point estimator, which is due to the different numbers of simulation runs. However, the estimators still seem to be unbiased.

In order to be able to compare the other estimators with each other, one can see in Fig. 5 that the GREG-based estimators, including the Lundström and Särndal formulae, show little difference whereas the gain in efficiency compared to the Horvitz-Thompson is very large, which results from the highly correlated auxiliary variable.

The measures in Fig. 6 yield the same results. However, we observe two things that could be seen in many simulations. First, once an estimator starts to become inadequate,

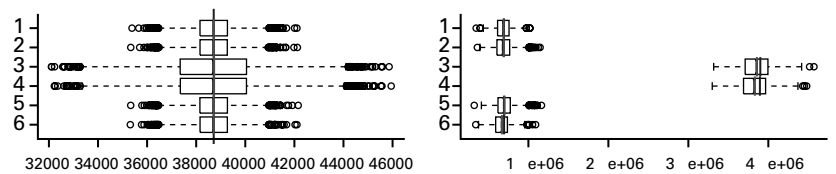

Fig. 2. Comparison of the six point (left) and variance (right) estimators for the total number of unemployed in Saarland (Task 1)

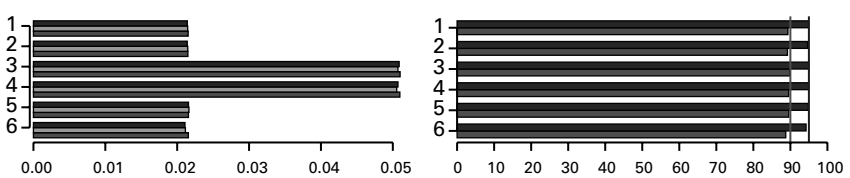

Fig. 3. MSE-based measures (left) and coverage rates (right) of the estimators for the total number of unemployed in Saarland (Task 1)

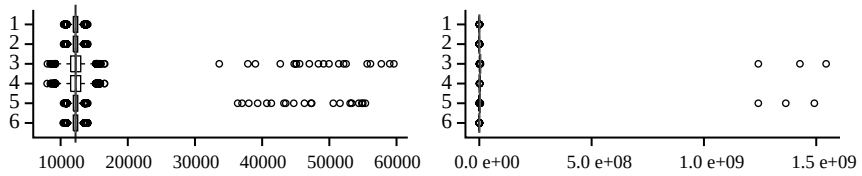

Fig. 4. Comparison of the six-point (left) and variance (right) estimators for the total number of unemployed in Saarland, subregion according to Task 2

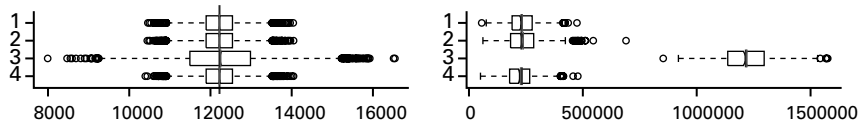

Fig. 5. Comparison of the four-point (left) and variance (right) estimators for the total number of unemployed in Saarland, subregion according to Task 2

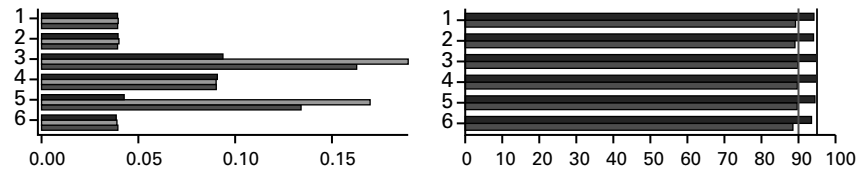

Fig. 6. MSE-based measures (left) and coverage rates (right) of the estimators for the total number of unemployed in Saarland (Task 2)

the corresponding estimated relative standard error tends to become inefficient (all estimators are asymptotically unbiased). Second, the Horvitz-Thompson estimator generally yields slightly more stable confidence coverage rates but with–sometimes–a much larger variance.

Fig. 7 shows the scatterplots of the variance estimators compared with each other. The diagonal line indicates the equal line, and the horizontal and vertical lines the *true* variances. All GREG-based estimators are highly correlated with each other and close to independent of the Horvitz-Thompson estimator.

Turning to Task 3, we find that the more homogeneous strata on the one hand and the smaller size of the area of interest on the other do not produce the extreme outliers, but tend to overestimate the true values in many more cases, which results in biased estimates (cf. Fig. 8). Here, the estimators of interest suffer from the type of cluster sampling used in connection with the classical asymptotical problems of discrete-valued distributions.

Figure 9 again shows the efficiency of the refined MI routine for binary data. In this case, only small differences between the GREG with linearized variance estimator, the Lundström-Särndal and the GREG estimator under multiple imputation in the refined version occur. The coverage rates suffer from a slight underestimation of the true variance (cf. Fig. 10).

Finally, in Fig. 11 again problems arise using the logit imputation for very small proportions of the target variable. In this case $\theta = 15,274$ in the universe, which gives a true proportion of 1.4% unemployed people aged from 45 to 64 years in Saarland. These

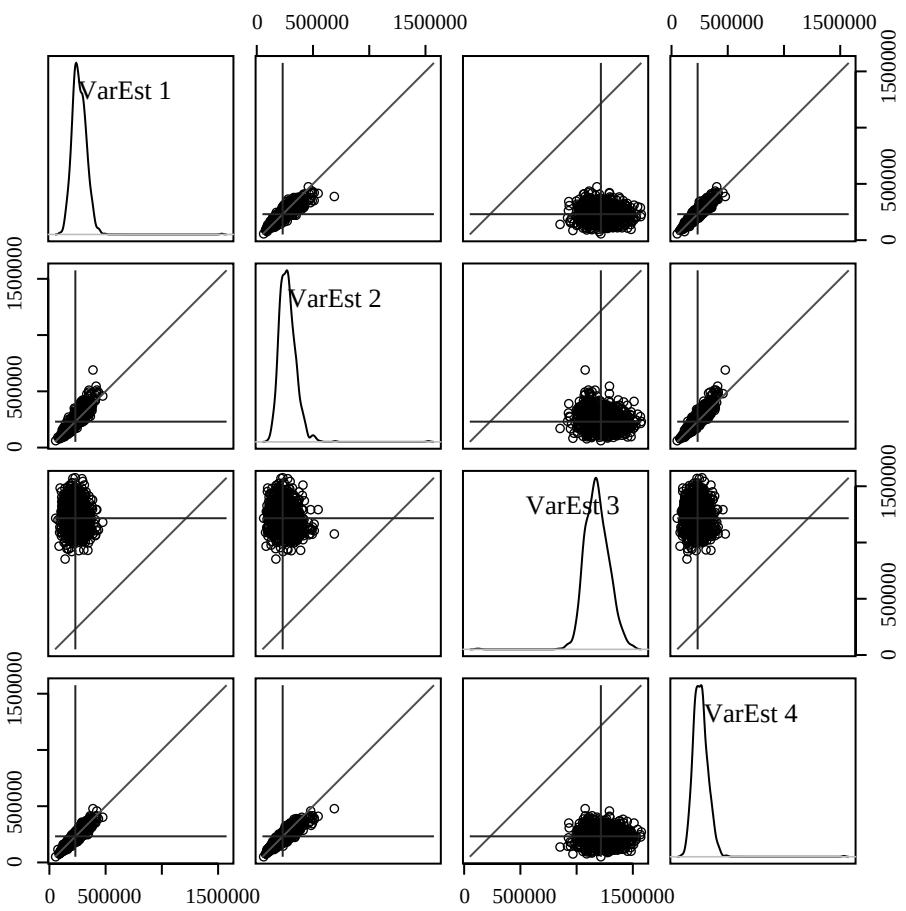

Fig. 7. Scatter plot of the four variance estimators for the total number of unemployed in Saarland, subregion according to Task 2

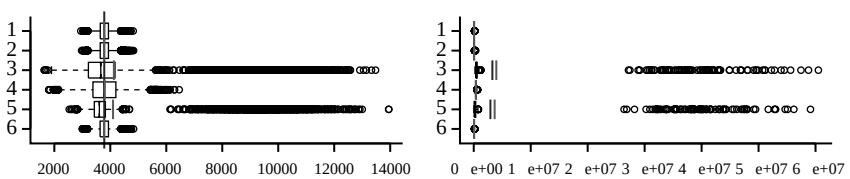

Fig. 8. Comparison of the six point (left) and variance (right) estimators for the total number of unemployed in Saarland, subregion according to Task 3

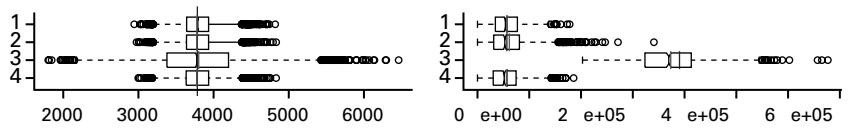

Fig. 9. Comparison of the four point (left) and variance (right) estimators for the total number of unemployed in Saarland, subregion according to Task 3

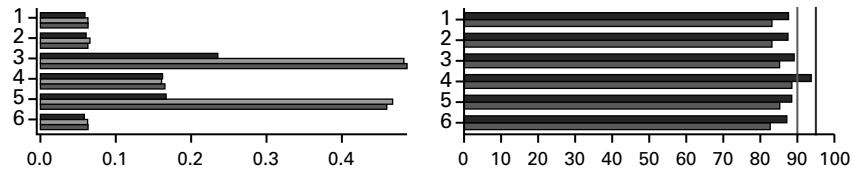

Fig. 10. MSE-based measures (left) and coverage rates (right) of the estimators for the total number of unemployed in Saarland (Task 3)

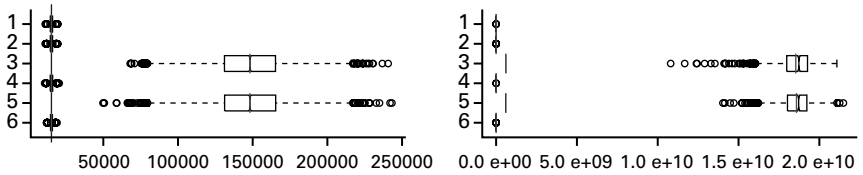

Fig. 11. Comparison of the six point (left) and variance (right) estimators for the total number of unemployed in Saarland in the age class [45; 65)*, according to Task 4*

figures seem to be typical for the logistic regression imputation and may yield – when applied as imputation routine – erroneous point estimates as well as useless variance estimates. In these cases, the maximum likelihood estimation does not seem to converge properly which results in a considerably larger number of unemployed (cf. Section 2.5).

The graphs based on the measures again show proper values for the other four point and variance estimation results (cf. Fig. 12). They also show the insufficiency of the logit based multiple imputation ending up in a severe increase of the MSE and especially the estimated relative standard error.

Further simulation results have shown that an increase of variables used as X in the imputation generally caused further troublesome effects on the logistic regression imputation. So far, in the simulation study no case has been observed where the PRIMA routine has turned out to be problematic. However, the tendency of a small underestimation of the true variance was observed in some cases using several imputation variables in stratification with homogeneous strata.

5. Summary and Outlook

The preceding study has shown some peculiarities of the logistic regression imputation for binary data. The most severe cases obviously may lead to wrong answers in the analysis. The *predictive imputation matching* turned out to be very insensitive with regard to these

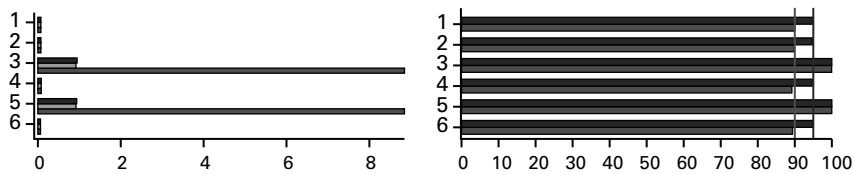

Fig. 12. MSE-based measures (left) and coverage rates (right) of the estimators for the total number of unemployed in Saarland in the age class [45; 65) *(Task 4)*

troublesome data constellations, although we have not yet proved this routine to be proper in Rubin's sense. However, the actual simulations have shown encouraging results.

Hence, we conclude that the regression-switching approach seems to be quite promising in large datasets and also for large quantities of missing values. Even in the context of mass imputation, i.e., split questionnaire survey designs and data fusion, we find good frequentist properties. In the U.S. the regression-switching multiple imputation approach is basically applied in the NHANES (a split project) and NMES. The basic routines are already implemented in MICE (SPLUS and R version) and IVEware, Raghunathan's SAS callable application.

6. References

Barnard, J. and Rubin, D.B. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. Biometrika, 86, 948–955.

Brand, J.P.L. (1999). Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. Thesis, Erasmus University Rotterdam. Print Partners Ispkamp, Enschede. The Netherlands.

D'Arrigo, J. and Skinner, C. (2004). Variance Estimation for Estimators Subject to Raking Adjustment. DACSEIS report, deliverable 8.1.

Greene, W. (2000). Econometric Analysis. (4th ed.). Upper Saddle River NJ.

Heidenreich, H.-J. (1994). Hochrechnung im Mikrozensus ab 1990. In S. Gabler, J. Hoffmeyer-Zlotnik, and D. Krebs (eds). Gewichtung in der Umfragepraxis. Opladen: Westdt. Verlag. [In German]

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 1-10.

Kennickell, A.B. and McManus, D.A. (1994). Multiple Imputation of the 1983 and 1989 Waves of the SCF. Proceedings of the American Statistical Association, Section on Survey Research Methods. 523-528.

Little, R.J.A. (1988). Missing-Data Adjustments in Large Surveys. Journal of Business and Economic Statistics, 6, 287–296.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley.

Lohr, S. (1999). Sampling: Design and Analysis. Pacific Grove: Duxbury Press.

Lundström, S. and Särndal, C.-E. (2002). Estimation in the Presence of Nonresponse and Frame Imperfections. Statistics Sweden.

Meyer, K. (1994). Zum Auswahlplan des Mikrozensus ab 1990. In S.Gabler, J.Hoffmeyer-Zlotnik, and D.Krebs (eds). Gewichtung in der Umfragepraxis. Opladen: Westdt. Verlag. [In German]

Münnich, R. (2001). Data Quality of Complex Surveys within the New European Information Society (DACSEIS). NTTS and ETK 2001 Conference Proceedings, EUROSTAT.

Münnich, R. and Magg, K., (2004). The DACSEIS Monte-Carlo Simulation Study. In: Münnich (ed.). Variance Estimation in Complex Surveys. DACSEIS delivarable D1.2 Chapter 1, DACSEIS report.

Münnich, R. and Schürle, J. (2003). On the Simulation of Complex Universes in the Case of Applying the German Microcensus, DASCEIS Research Paper Series 4. http://w210.ub.uni-tuebingen.de/dbt/volltexte/2003/979

Münnich, R. and Wiegert, R. (2001). The DACSEIS Project. DASCEIS Research Paper Series 1. http://w210.ub.uni-tuebingen.de/dbt/volltexte/2001/428/

Quatember, A. (2002). A Comparison of the Five Labour Force Surveys of the DACSEIS Project From a Sampling Theory Point of View. DACSEIS Research Paper Series, 3. http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/547.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Lecture Notes in Statistics, 168 New York, Springer.

Rässler, S. (2004). The Impact of Multiple Imputation for DACSEIS. DACSEIS Research Papers Series 5. http://w210.ub.uni-tuebingen.de/dbt/volltexte/2004/1135.

Rässler, S., Koller, F., and Mäenpää, C. (2002). A Split Questionnaire Survey Design Applied to German Media and Consumer Surveys. Proceedings of the International Conference on Improving Surveys, ICIS, Copenhagen.

Rubin, D.B. (1978). Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. Proceedings of the American Statistical Association, Section on Survey Research Methods, 20–40.

Rubin, D.B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics, 4, 87–95.

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

Rubin, D.B. (2003). Nested Multiple Imputation of NMES via Partially Incompatible MCMC. Statistica Neerlandica, 57, 3–18.

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.

Rubin, D.B. and Schenker, N. (1987). Logit-based Interval Estimation for Binomial Data Using the Jeffreys Prior. Sociological Methodology, 131–144.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Schafer, J.L. (1999a). Multiple Imputation: a Primer. Statistical Methods in Medical Research, 8, 3–15.

Schafer, J.L. (1999b). Multiple Imputation under a Normal Model, Version 2. Software for Windows 95/98/NT, http://www.stat.psu.edu/ ~ jls/misoftwa.html.

Schafer, J.L. and Olsen, M.K. (1999). Modeling and Imputation of Semicontinuous Survey Variables. Technical Report No. 00-39, The Pennsylvania State University.

Schafer, J.L. and Yucel, R.M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. Journal of Computational and Graphical Statistics, 11, 437–457.

Statistisches Bundesamt (StBA; ed.), (1999). Fachserie 1: Bevölkerung und Erwerbstätigkeit. Reihe 4.1.1: Stand und Entwicklung der Erwerbstätigkeit. [In German]

Van Buuren, S. and Oudshoorn, C.G.M. (2000). Multivariate Imputation by Chained Equations. TNO Report PG/VGZ/00.038, Leiden.

Van Buuren, S. and Oudshoorn, K. (1999). Flexible Multivariate Imputation by MICE. TNO Report PG/VGZ/99.054, Leiden.

Wiegert, R. and Münnich, R. (2004). German Register Data for Regression Estimation in Survey Sampling – A Study on the German Microcensus Respecting for Data Protection. Jahrbücher für Nationalökonomie und Statistik, 224(1/2), 247 – 259.

Received April 2005
Revised June 2005