Rässler, Susanne; Schnell, Rainer

Working Paper

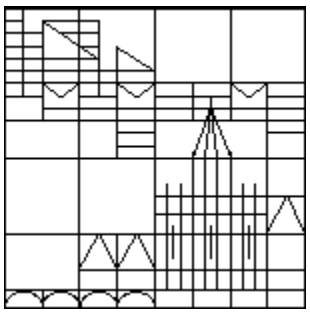
Multiple imputation for unit-nonresponse versus weighting including a comparison with a nonresponse follow-up study

Diskussionspapier, No. 65/2004

Provided in Cooperation with:
Friedrich-Alexander-University Erlangen-Nuremberg, Chair of Statistics and Econometrics

Suggested Citation: Rässler, Susanne; Schnell, Rainer (2004) : Multiple imputation for unit-nonresponse versus weighting including a comparison with a nonresponse follow-up study, Diskussionspapier, No. 65/2004, Friedrich-Alexander-Universität Erlangen-Nürnburg, Lehrstuhl für Statistik und Ökonometrie, Nürnberg

This Version is available at:
https://hdl.handle.net/10419/29622



Multiple Imputation for Unit-Nonresponse versus Weighting including a comparison with a Nonresponse Follow-Up Study

Susanne Rässler, Rainer Schnell
University of Erlangen-Nürnberg, University of Konstanz

09.10.2003

Centre for Quantitative Methods and Survey Research
University of Konstanz

Abstract: The results of a national fear of crime survey are compared with results following the use of different nonresponse correction procedures. We compared naive estimates, weighted estimates, estimates after a thorough nonresponse follow-up and estimates after multiple imputation. A strong similarity between the MI and the follow-up-estimates was found. This suggests, that if the assumptions of MAR hold, carefully selected and collected additional data applied in a MI could yield similar estimates to a nonresponse follow-up at a much lower price and respondent burden.
Keywords: Multiple Imputation, Unit-nonresponse, missing data, complex surveys.

1 Introduction

Unit-Nonresponse rates in Germany are rising much like everywhere else (Schnell 1997). High nonresponse rates are of theoretical and practical importance, because of the need to justify the high survey costs of random samples compared with convenience samples. Since survey results are rarely validated with external data, the possible bias of survey results is usually rated according to the response rate. Of course, this is misleading, since the bias of an estimate depends on the amount of nonresponse and the missing data generating mechanism (MDGP). It is useful to distinguish between three kind of "'missingness"' (Rubin and Little 1987)

- MCAR (missing completely at random),
- MAR (missing at random) and
- MNAR (missing not at random).

The consequences for valid inferences of these types of mechanisms are different and depend on the MDGP and method used to correct results affected by nonresponse. A number of different methods to handle the increasing amount missing data have been developed, for example:

- Procedures based on the available cases only, i.e., only those cases that are completely recorded for the variables of interest
- Weighting procedures such as Horvitz-Thompson type estimators that adjust for nonresponse
- Single imputation like hot-deck-methods
- Multiple imputation (MI) (Rubin 1987)
- Implicitly model-based corrections of parameter estimates such as the expectation-maximization (EM) algorithm

- Explicitly model based corrections of regression estimates of the Heckman-Type (Winship and Mare 1992).

More than one missing data mechanism can be expected in large-scale national surveys. If we consider the MDGPs of such samples as MAR, model based procedures will often be misleading, because the generating mechanism will only be partially modelled. Ad-hoc-procedures like weighting or available cases will fail, because the missing data mechanisms are not MCAR, but only MAR. Furthermore, the variance due to imputation will be underestimated by nearly all procedures even if the MDGPs are MCAR. Therefore, we expect that only the use of multiple imputation procedures will at least approximately result in unbiased estimators and their variance.
Weighting and hot-deck-methods are quite common in official statistics and many public-use-files contain imputed values or weighting factors. These data sets are seldom intended for analytical purposes, and so evaluation of the substitutes are mostly based on population totals. In contrast, EM- and Heckman-procedures are usually intended for analytical surveys with a single purpose model. Interestingly, evaluations of multiple imputation (MI) is usually done for descriptive surveys, which are characterized by item-nonresponse with a large amount of information for every unit. Unit-nonresponse evaluations for MI are quite rare if not a complete novelty. In order to compare different nonresponse correction procedures for unit-nonresponse, we used the data from the DEFECT-project, which included a thorough nonresponse-follow-up.

2 Study design of the DEFECT project

The data presented are part of the DEFECT study on sampling errors and nonsampling errors in complex surveys (Schnell and Kreuter 2000). It is the first nationwide sample to be carried out in Germany with an interpenetrated sampling design. The same questionnaire was used to conduct five independent surveys at 160 sampling points. Four of the five surveys were conducted by professional survey institutes that are highly regarded for their good practices and reliability, while the fifth (the mail survey) was conducted by the DEFECT group itself. This design means that in each of these 160 sampling points, two face-to-face random surveys, a face-to-face survey with quota selection, a CATI random survey, and a mail survey were carried out concurrently (the use of several survey modes and sampling procedures arises because this study was conducted as part of a methodological study with a much larger scope). For each of the three face-to-face surveys, only one interviewer was involved with each sampling point and no interviewer worked in more than one sampling point. This use of this design was intended to permit the statistical separation of interviewer and sampling-point effects, while adhering to the survey institutes' usual procedures. In this paper, we will concentrate on the two face-to-face-surveys. Due to space limitations, only the estimates of one of these surveys will be reported.

2.1 The sampling process

The addresses of households to be contacted in these surveys were selected in a multistage procedure. In the first stage, 160 sampling points were randomly selected from a nationwide register of election districts. In the second stage, an address-random procedure was used to select households within the 160 election districts. The actual household selection was carried out by eight members of the research group, who personally visited each of the sampling points. Starting from a randomly selected address in each sampling point, the group members noted the address of every third household along the random-walk route until they had gathered 110 addresses. Of these 110 addresses, the first 64 (16*4) in each point were randomly assigned to the four random surveys mentioned above using a shuffle procedure, which meant that a total of 2,560 addresses were initially sent to the survey institutes. Shortly thereafter, four additional addresses were given to the institutes for each sampling point in order to replace the neutral dropouts they reported in the first round of surveys (for example, respondents who had either died or moved away since the address collection had been conducted). At some sampling points, this number of addresses was not enough to permit a total of at least six completed interviews at all sampling points. In those cases, the institute was supplied with additional addresses for each of the respective sampling points. However, the interviewers attempted to make at least four contact attempts with each of the original addresses before the fall-back sample was used. Each selected address was photographed and the corresponding building was location rated by a member of the research team according to architectural criteria (for example cameras, alarms) and estimated social economic status of the inhabitants. Additionally for each sampling point (PSU) a long list of aggregate statistics from federal data bases such as population density, number of foreigners and number of crime incidents known to the police was compiled. These data are available for each selected address, regardless whether an interview was obtained or not.
The addresses received by the institutes were therefore household addresses, but the goal was an individual sample of the target population. The target population was defined as all German-speaking inhabitants of these households selected who were 18 years or older, one of whom was randomly selected for survey participation. The institutes themselves were responsible for this last sampling stage. In the face-to- face random surveys, the potential respondent was selected by the interviewer using a variant of a Kish grid. For the other two random surveys (CATI and mail), the potential respondent was selected using the last-birthday procedure.

2.2 Survey topic

The survey topic was fear of crime. The topic was chosen for this methodological study for several reasons. First, it was intended to boost participation, since crime is a matter of at least some concern to a wide variety of people. Second, several kinds of questions can be asked about fear of crime (factual, attitudinal, sensitive, etc.), the use of such questions allows a comparison of design effects for different kinds of items. In constructing the questionnaire, we used both the well-established indicators typically used in fear of

crime surveys and a set of items developed in-house (Kreuter 2002). All questions went through a series of pretest phases.

2.3 Fieldwork

Fieldwork was done by 173 and 164 interviewers between October 1999 and February 2000. The actual field procedures were decided by the survey organizations. Despite some differences in details (for example, time scheduling of call-backs), the results of the fieldwork was remarkable similar (see table 1). Each company made an independent attempt to interview survey nonrespondents from the face-to-face-survey by CATI-interviewers. Due to legal constraints and limits of phone number listings, only about 2/3 of the nonrespondents were eligible for the nonresponse study. With respect to the interviews using a full length questionnaire, the companies were equally successful: About 27% of the eligible nonrespondents were successfully interviewed. The interviewer was instructed to use a short nonresponse questionnaire consisting of only two core questions from the study if a full length interview was not possible. With regard to this short NR-interview, the results of the two companies are stunningly different: 3% versus 16% of the nonrespondents could be motivated to provide answers to the core questions. However, the nonresponse studies yielded 399 and 274 additional interviews for initial nonrespondents. Therefore, we could compare the results of statistical nonresponse compensation procedures with the results of increased fieldwork efforts by changing response modes.

Number of Interviewers	173	164
Number of Interviews	1326	1345
Response Rate	41.5	39.26
Eligible for Nonresponse-Study	1422	1061
Number of NR-Interviews	399	274
Response-Rate	28.1	25.8
Number of short NR-Interviews	48	228
Response-Rate	3.4	16.0

Table 1: Details and results of the fieldwork

3 Nonresponse mechanism and analyst's model

Previous work from our group has suggested a strong inverse relationship between the probability of being contacted by an interviewer and the probability of victimization (Schnell 2002). A restriction to easy contactable respondents therefore was expected to yield an overestimation of fear of crime and an underestimation of actual victimization. Therefore we attempted to model the response mechanism with respect to the probability of being contacted. Since we expected a high SES to correlate with a lower fear of

crime and a higher victimization risk we included the SES estimation of the address enumerators in the imputation.
The question asked for fear of crime was the widely used "'standard indicator"' of fear of crime "'Is there any area right around here - that is, within 1 km - where you would be afraid to walk alone at night?"' (Yes/No).
The estimation task in this paper is to fit a logistic regression model for the analysis according to

$$logit(fear) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 for + \beta_4 dens + \beta_5 PKZ + \beta_6 ses \qquad (1)$$

with PKZ = crime rate, ses = 5-point scale of socio-economic status, for = % of foreigners within the sampling point and $dens$ = local population density.
In the DEFECT project additional information has been collected which is either point specific or case or area specific, respectively. Therefore, in this very rare situation, a couple of valuable additional information is available, thus, we are able to treat the unit-nonresponse in this project as a special case of item-nonresponse and to correct for the possible nonresponse bias by multiple imputation. Finally, the following data matrix for the imputation model as well as for the analyst's model resulted:

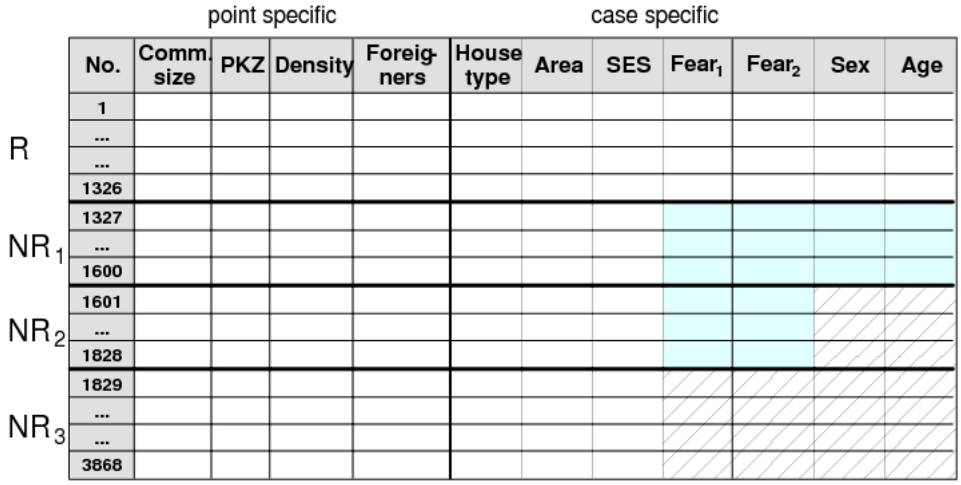


Figure 1: DEFECT unit-nonresponse treated as item-nonresponse for FTF No. 1

Moreover, the nonresponse pattern is monotone which makes multiple imputation relatively easy to perform, see section 5.2.

4 Weighting adjustment

The most common procedures to correct for unit-nonresponse in official statistics and survey research in the social sciences are weighting procedures. Although different techniques are available (from propensity weighting to raking), the most commonly

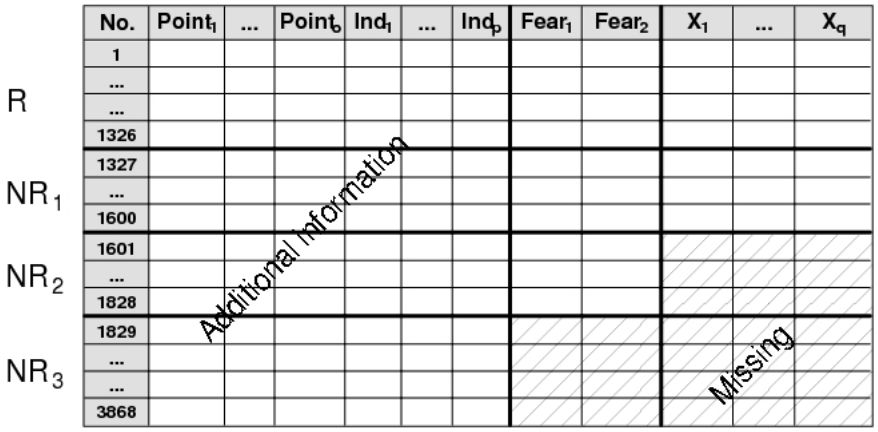


Figure 2: Unit-nonresponse treated as item-nonresponse

used technique is the application of cell weights. Hereby the individual weight w_i of an observation i within a cell j is computed as ratio of the number of observations within a cell n_j multiplied by design weights d_j and the reciprocal sampling fraction (N/n) to the population total of that cell (N_j):

$$w_i = \frac{n_j d_j N/n}{N_j} \tag{2}$$

It is even more interesting to compare the results based on multiple imputed data sets with the results produced by rather complex survey weights, because weighting can be seen as a single conditional mean imputation. To illustrate this, let us discuss for a moment a typical estimate of a population total Y using weights such as the so-called weighting-class estimator. Let the sample be divided into G homogeneous cells or groups with respect to the assumed response generating process. Then, let n_j denote the expected or planed sample size in group or cell j, $j = 1, 2, \ldots, J$, e.g. among young working women, and m_j is the number of respondents in this group. Then, if only sample counts are used in the weighting procedure, the weighting-class estimator according to Oh and Scheuren (1983) is given by

$$\widehat{Y} = \frac{N}{n} \sum_{j=1}^{J} \frac{n_j}{m_j} \sum_{i=1}^{m_j} y_i = \frac{N}{n} \sum_{j=1}^{J} n_j \overline{y}_j^{obs} \tag{3}$$

with n denoting the sample size without nonresponse. It is easily seen, that the weighting-class estimate as given in (3) equates the estimate derived by single conditional mean imputation. [1]

In practice, most often the population totals of the cells are unknown, but the marginals of different weighting variables are known for the population. In this situation, a set

[1] Thus, naive estimates of standard errors and confidence intervals will be biased downwards. The derivation of an unbiased variance estimator is cumbersome, see Djerf (2001).

of weighting vectors has to be estimated, which satisfies all constraints given by the population margins. In most cases, this is done by an iterated proportional fitting algorithm (IPF). In order to follow common practice in Germany, we computed the individual weights with iterative proportional fitting.[1] The IPF was applied for each survey according to two different crosstabulations:

- G1: sex × age × labor force status × state (256 cells)
  based on the microcensus (a survey of a 1% sample of the population, conducted by the federal bureau of statistics)
- G2: occupational status × community size (25 cells)
  based on the "'media analysis"', the largest commercial survey used in Germany.

This weighting scheme is a bit more detailed than the one usually used. Given the fact, that a weighting scheme could be interpreted as a conditional mean imputation, this detailed weighting scheme seem to be a plausible model for mean imputation of fear of crime: All weighting variables have clear theoretical links and strong empirical correlations with fear of crime and victimization. Therefore, we feel that no weighting scheme for this survey topic using currently available data would perform better than this.

5 Multiple imputation

5.1 The basic principle

The theory and principle of multiple imputation (MI) originates from Rubin (1978).[2] 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 $f_{U_{mis}|U_{obs}}$ of the missing data given the observed data. Since it is often difficult to draw from $f_{U_{mis}|U_{obs}}$ directly, a two-step procedure for each of the m draws is useful:

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

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

Because

$$f_{U_{mis}|U_{obs}}(u_{mis}|u_{obs}) = \int f_{U_{mis}|U_{obs},\Xi}(u_{mis}|u_{obs},\xi) f_{\Xi|U_{obs}}(\xi|u_{obs}) d\xi \qquad (4)$$

[1] We used a proprietary IPF-implementation of one of the largest social research companies in Germany (Infratest) called "'Gemsoq"'. We like to thank the staff of Infratest for their permission to use the program for this data.

[2] MI is extensively described by Rubin(1987), though this book is hard to read. An excellent and comprehensive treatment of data augmentation and multiple imputation can be found in Schafer (1997). An introduction to MI is also provided by Schafer (1999a) or Little and Rubin (2002).

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

In contrast, the corresponding observed-data posteriors $f_{\Xi|U_{obs}}$ usually are 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.[1] In MCMC the desired distributions $f_{U_{mis}|U_{obs}}$ and $f_{\Xi|U_{obs}}$ are achieved as stationary distributions of Markov chains which are based on the easier to compute complete-data distributions.

To proceed further, let θ denote a scalar quantity of interest that is to be estimated, such as a mean, variance, or correlation coefficient. Notice that now θ can be completely different from the data model used before to create the imputations. In the remainder of this section, the quantity θ to be estimated from the multiply imputed data set, has to be distinguished from the parameter ξ used in the model for imputation. Although θ (analysis) could be an explicit function of ξ (imputation), one of the strengths of the multiple imputation approach is that this need not be the case. In fact, θ (analysis) could even be the parameter of the imputation model, then the imputation and analysis model are the same and are said to be congenial (Meng, 1995). However, multiple imputation is designed for situations where the analyst and the imputer are different, thus, the analyst's model could be quite different from the imputer's model. As long as the two models are not overly incompatible or the fraction of missing information is not high, inferences based on the multiply imputed data should still be approximately valid. Moreover, if the analyst's model is a sub-model of the imputer's model, i.e., the imputer uses a larger set of covariates than the analysts and the covariates are good predictors of the missing values, then MI inference is superior to the best inference possible using only the variables in the analyst's model. This property is called superefficiency by Rubin (1996). On the other hand, if the imputer ignores some important correlates of variables with missing data, but these variables are used in the analyst's model, then the results will be biased. Consider, for example, expenditure surveys and the situation of imputing income without using expenditure. This refers to an imputation being done under the hypotheses of zero correlation between income and expenditure. Unfortunately, this is not the case, thus, results will be biased.[2] Moreover, the imputer's model also allows the researcher to use in-house variables such as additional information from interviewers (residential area, neighborhood, house size, number of cars or garages etc.) which are typically not available to the analyst but may show some correlation with the missing

[1] These are extensively discussed by Schafer (1997).

[2] Rubin (1987) and Schafer (1997, Chapter 4) and their references therein discuss the distinction between θ (analysis) and ξ (imputation) more fully.

variables. The DEFECT project was designed to collect such background information.
As described earlier, $U = (U_{obs}, U_{mis})$ denotes the random variables concerning the data with observed and missing parts, and $\widehat{\theta} = \widehat{\theta}(U)$ denotes the statistic that would be used to estimate θ if the data were complete. Furthermore, let $\widehat{var}(\widehat{\theta}) = \widehat{var}(\widehat{\theta}(U))$ be the variance estimate of $\widehat{\theta}(U)$ based on the complete data set.
The MI principle assumes that $\widehat{\theta}$ and $\widehat{var}(\widehat{\theta})$ can be regarded as an approximate complete-data posterior mean and variance for θ, with

$$\widehat{\theta} \approx E(\Theta | u_{obs}, u_{mis})$$

and

$$\widehat{var}(\widehat{\theta}) \approx var(\Theta | u_{obs}, u_{mis})$$

based on a suitable complete-data model and prior; see also Schafer (1997), p. 108. Moreover, we should also assume that with complete data, tests and interval estimates which are based on the normal approximation

$$(\widehat{\theta} - \theta)/\sqrt{\widehat{var}(\widehat{\theta})} \sim N(0, 1) \tag{5}$$

should work well. Notice that the usual maximum-likelihood estimates and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions.
Suppose now that the data are missing and we make $m > 1$ independent simulated imputations $(U_{obs}, U_{mis}^{(1)})$, $(U_{obs}, U_{mis}^{(2)})$, $\ldots$, $(U_{obs}, U_{mis}^{(m)})$ enabling us to calculate the imputed data estimate $\widehat{\theta}^{(t)} = \widehat{\theta}(U_{obs}, U_{mis}^{(t)})$ along with its estimated variance $\widehat{var}(\widehat{\theta}^{(t)}) = \widehat{var}(\widehat{\theta}(U_{obs}, U_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. Figure 3 illustrates the multiple imputation principle. From these m imputed data sets the multiple imputation estimates are computed.

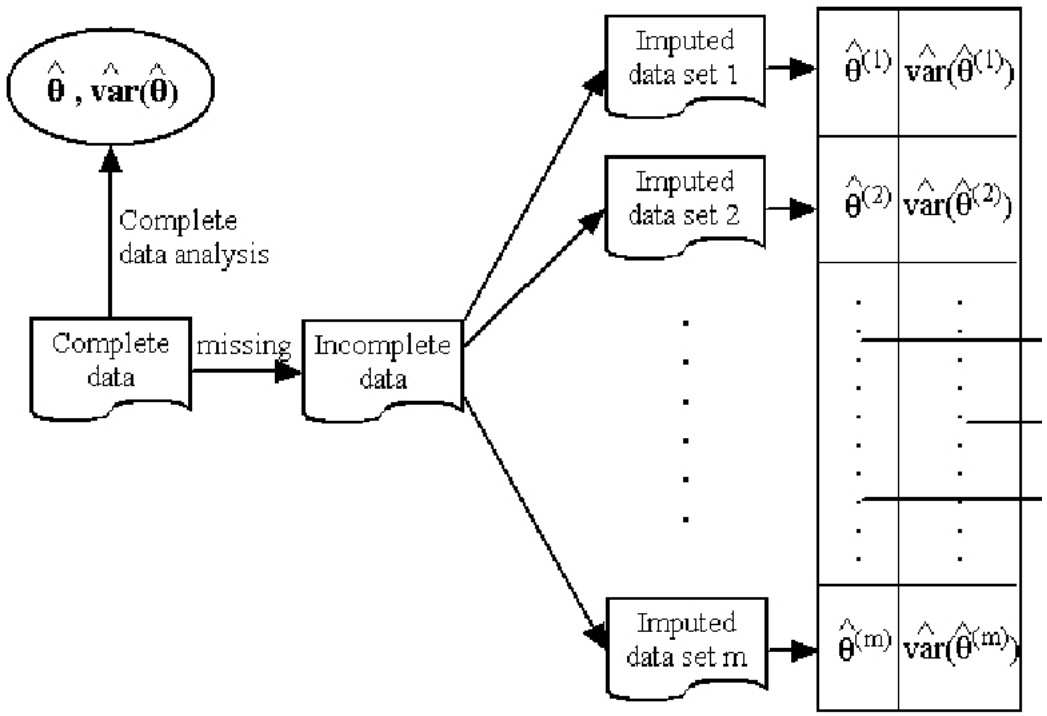


Figure 3: The multiple imputation principle

The MI point estimate for θ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\widehat{\theta}^{(t)}. \tag{6}$$

To obtain a standard error $\sqrt{\widehat{var}(\widehat{\theta}_{MI})}$ for the MI estimate $\widehat{\theta}_{MI}$ we first calculate the "between-imputation" variance

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2, \tag{7}$$

and then the "within-imputation" variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\widehat{\theta}^{(t)})\,. \tag{8}$$

Finally, the estimated total variance is defined by

$$\begin{aligned}\widehat{var}(\widehat{\theta}_{MI}) &= T = \widehat{var}(\widehat{\theta})_{within} + (1+\frac{1}{m})\widehat{var}(\widehat{\theta})_{between} \\ &= W + \frac{m+1}{m}B. \end{aligned} \tag{9}$$

Notice that the term $((m+1)/m)B$ enlarges the total variance estimate T compared to the usual analysis of variance with $T = B + W$; $(m+1)/m$ is an adjustment for finite m. An estimate of the fraction of missing information γ about θ due to nonresponse is given by

$$\widehat{\gamma} = \frac{(1+1/m)B}{T}. \tag{10}$$

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

$$(\widehat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v \quad \text{and} \quad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \tag{11}$$

with the degrees of freedom[1]

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2 \tag{12}$$

From (11) we can see that the multiple imputation interval estimate is expected to produce a larger interval than an estimate based only on one single imputation (SI). The multiple imputation interval estimates are widened to account for the missing data uncertainty and simulation error; see Schafer (1999). Notice that confidence intervals under MI can be shorter than confidence intervals based only on the complete or available cases (AC). This is especially true if the imputed sample is substantially larger than the complete case sample. Therefore, typically, the following comparisons hold for most surveys and most estimates of standard errors:

$$\text{s.e.(SI)} < \text{s.e.(truth)} < \text{s.e.(MI)} < \text{s.e.(AC)}.$$

5.2 Multiple imputation for large scale surveys with binary variables

The multivariate normal model, as it is provided by Schafer (1999b) with NORM, has become quite popular among statisticians for multiple imputation in multivariate settings. But in many applications, the assumption of a multivariate distribution can hardly be justified, for example with binary variables. Recently, Rubin (2003) suggests univariate multiple imputation procedures for large-scale data sets. They are successfully used for multiple imputation in the U.S. National Medical Expenditure Survey (NMES), where the data set to be imputed consists of up to 240 variables of different scale types and 22.000 observations. Such routines have been used quite efficiently in the context of "mass imputation", i.e., imputing a high amount of data that are typically missing by design. This is the situation in the so-called data fusion case and the split questionnaire

[1] The degrees of freedom are based on a Satterthwaite approximation, see Rubin and Schenker (1986) or Rubin (1987), pp. 76-77. For small data sets an improved expression for the degrees of freedom is given by Barnard and Rubin (1999). They relax the assumption of a normal reference distribution of (5) for the complete-data interval estimates and tests to allow a t distribution, and they derive the corresponding degrees of freedom for the MI inference to replace the formula (12) given here. Moreover, additional methods are available for combining vector estimates and covariance matrices, p-values, and Likelihood-ratio statistics (see Little and Rubin, 2002).

survey designs, see, e.g., Rässler (2002).
Iterative univariate imputations were first implemented by Kennickell (1991) and Kennickell and McManus (1994); see Schafer and Olsen (1999).[1] 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 is a variety of procedures available for regression modeling of continuous and categorical univariate response variables such as ordered or unordered logit/probit models (see Greene 2000). Thus any plausible regression model $Y|X = x, \Theta = \theta$ may be specified for predicting each univariate variable Y_{mis} that has to be imputed given all the other variables. This approach is also known as regression switching, chained equations, or variable-by-variable Gibbs sampling.[2]
To illustrate the principle of the regression-switching let us assume the simple case with 3 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."

This procedure allows great flexibility due to the possible conditional specifications. Each specification is simply a univariate regression. It should be noted 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 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,

[1] Ready to use and freely available via the Internet, software MICE is a recent implementation of some iterative univariate imputation methods in S-PLUS as well as R; see van Buuren and Oudshoorn (2000). There is also the free SAS-callable application IVEware, which also provides iterative univariate imputation methods.

[2] See van Buuren and Oudshoorn (1999). In the variable-by-variable Gibbs sampling approach it is also possible to include only relevant predictor variables, thus reducing the number of parameters.

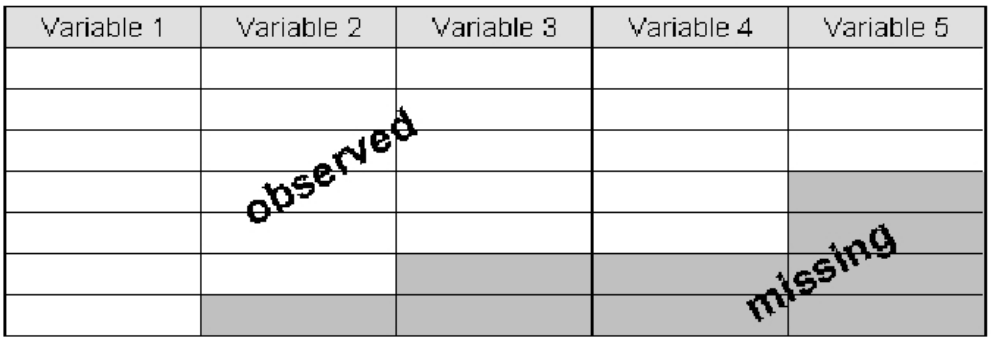


Figure 4: Monotone pattern of missingness

even if there is no such joint distribution for the data, the Markov chain Monte Carlo Method (MCMC) can be implemented, and each conditional specification may give a good empirical fit to the data[1]
If, on the other hand, the missingness has a monotone pattern, as is the case in our study (see figure 4), then 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 fitted is conditional only on the variables to the left. The resulting univariate models are automatically distributionally compatible. Finally, for this test application we applied the algorithm provided by MICE and generated $m = 5$ imputations.

6 Results

At first glance, the results based on respondents suggests a strong effect of sex, age and SES on fear of crime (see table 2, model R). Cell weighting (model RW) result in very similar estimates, with the exception that the effect of SES get insignificant. If we consider initial respondents and respondents to the NR-follow-up (model RNR), the effect of SES is even stronger than for the initial respondents alone. If initial and follow-up-respondents were weighted (model RNRW), the estimates are very close to the unweighted estimates (model RNR). So, weighting initial respondents seem to reduce the effect of SES, all other effects are not effected by weighting alone.
A comparison of this results with the estimates of MI (see table 3, model RMI) shows that MI of initial respondents lead to similar estimates of sex and age as RNR or RNRW. The use of the follow-up data with MI (model RNRMI) further increase the estimates of sex, age and SES. If the additional data of the core questions were used (model RNRNR2MI), the crime rate, which was insignificant in all other models, shows a significant effect:

[1] For details, see Rubin (2003), van Buuren and Oudshorn (2000), and Brand (1999).

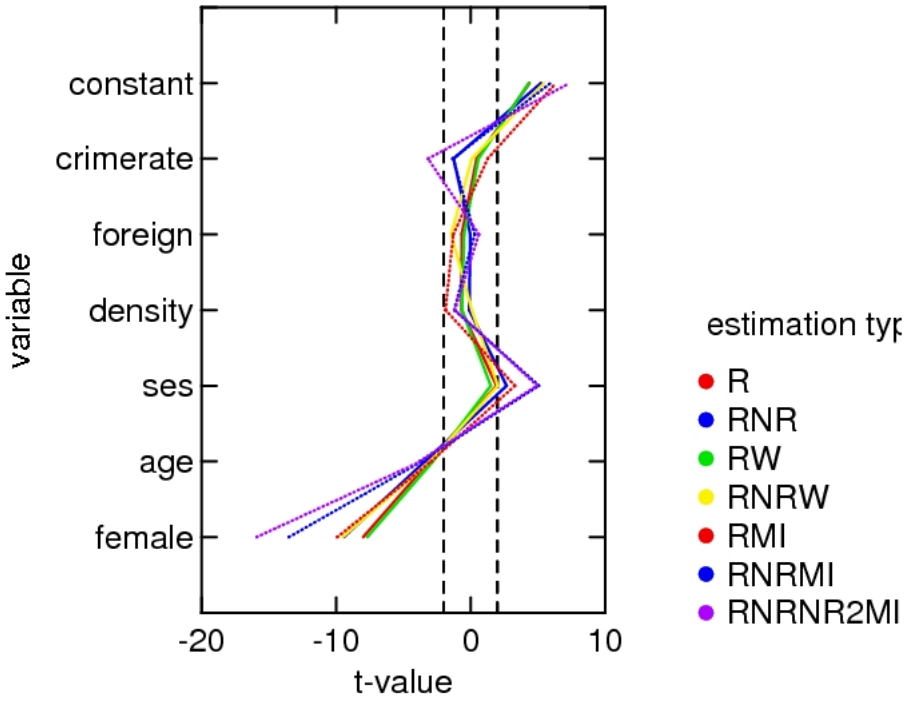


Figure 5: Results overview

Rising crime rates result in higher fear of crime, even controlling for sex, SES, age, population density and number of foreigners. Additionally, the constant of this model is the highest of all models. Therefore, we feel confident in the conclusion, that the MI-estimates are closest to the unknown true values.

7 Conclusions and outlook

The results encourage the use of multiple imputation for unit-nonresponse in survey practice. It seems remarkable, that - in accordance to theory and in strong contrast to survey practice - that we found no differences between naive and weighted estimates based only on the respondents. According to our expectations, the hard to contact subpopulations seem to be different from easy accessible persons, so a NR-follow-up yield different estimates of three variables (sex, SES, const). An additional weighting of the NR-follow-up does not improve the estimates, the results are nearly identical with the naive estimates based on the nonresponse follow-up. Most interestingly, the estimated effects after MI point in the expected direction, even more, the more information is used in creating multiple imputations. Finally, although the core questions are answered only by a few people, the information gained resulted in estimates, which seem to be sociologically plausible.

Moreover, we have argued that weighting can typically be seen as a sort of single imputation and therefore tends to give biased estimates of the sampling variance. In contrast, given MAR we could expect correct sampling variances and bias reduction by using multiple imputation, especially with multivariate data analysis techniques. The chained equations are a very flexible imputation tool allowing to exploit all valuable information and to display uncertainty due to nonresponse and imputation.

Unit-nonresponse is the result of a complex interaction of interviewer behavior and respondent behavior. So there is not one MDGP, but a mixture of MDGPs. If we model only one MDGP, the results may be heavily biased. This may be the explanation why sometimes increasing response rates led to more bias: By increasing the amount of hard-to-contact persons in the sample, we are modeling only one MDGP. Therefore, although non-response follow-ups are useful, one should not rely on their results alone. By using all available information, MI may be suited to handle situations where nonresponse is due to an unknown mixture of MDGPs - as long as the MAR-assumption holds. But fortunately, with the exception of some medical surveys, missing not at random seem to be rare in the social sciences.

Based on this result, we are tempted to recommend the use of auxiliary variables like estimated SES, the use of core questions for refusals and multiple imputation as a practical procedure to handle nonresponse under MAR. Of course, this recommendation should be tested with other data sets.

References

Barnard, J., 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.

Casella, G., George, E.I. (1992), Explaining the Gibbs Sampler, The American Statistician, 46, 167-174.

Djerf, K. (2001), Properties of Some Estimators Under Unit Nonresponse, Research Report, Statistics Finland, Helsinki.

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

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

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

Kreuter, F. (2002), Kriminalitätsfurcht: Messung und methodische Probleme, Opladen, Leske+Budrich, (in German).

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

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

Meng, X.L. (1995), Multiple-Imputation Inferences with Uncongenial Source of Input (with discussion), Statistical Science, 10, 538-573.

Oh, J.L, Scheuren, F. (1983), Weighting Adjustment for Unit Nonresponse, in: Incomplete Data in Sample Surveys. W.G. Madow, I. Olkin, D.B. Rubin (eds.), 2,143-184. Academic Press, New York.

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

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

Robert, C.P., Casella, G. (1999), Monte Carlo Statistical Methods. Springer, New York.

Rubin, D.B. (1978), Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, Proceedings of the Survey Research Methods Sections of the American Statistical Association, 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, Wiley, New York.

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

Rubin, D.B., 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, Chapman and Hall, London.

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

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

Schnell, R. (1997), Nonresponse in Bevölkerungsumfragen, Opladen, Leske+Budrich, (in German).

Schnell, R., Kreuter, F. (2000), Das DEFECT-Projekt: Sampling-Errors und Nonsampling-Errors in komplexen Bevölkerungsstichproben, ZUMA-Nachrichten, 47, 89-101, (in German).

Schnell, R. (2002), Antworten auf Nonresponse. Lecture given at XXXVII Meeting of the German Market Research Society, http://www.uni-konstanz.de/Fuf/Verwiss/Schnell/Antworten_auf_Nonresponse.pdf, (in German).

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

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

Winship, C., Mare, R.D. (1992), Models for sample selection bias, Annual Review of Sociology, 18, 327-350.

Appendix

- Table 2: Results based on respondents
- Table 3: Results based on the $m = 5$ imputed data sets

Respondents R: $N_{obs} = 877$				
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	P-VALUE
sex	-1.1400	0.1428	-7.98	0.000
age	-0.0129	0.0042	-3.06	0.002
ses	0.1805	0.0937	1.93	0.054
density	-0.0000	0.0000	-0.67	0.503
foreign	-0.0079	0.0122	-0.65	0.516
crime rate	0.0000	0.0000	0.46	0.644
constant	1.8914	0.4314	4.38	0.000
Respondents + nonresponse follow-up RNR: $N_{obs} = 1085$				
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	P-VALUE
sex	-1.2190	0.1293	-9.42	0.000
age	-0.0131	0.0039	-3.39	0.001
ses	0.2282	0.0855	2.67	0.008
density	-0.0000	0.0000	-0.11	0.912
foreign	-0.0003	0.0111	-0.02	0.981
crime rate	-0.0000	0.0000	-1.27	0.203
constant	2.0127	0.3850	5.23	0.000
Respondents + weights: RW: $N_{obs} = 892$				
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	P-VALUE
sex	-1.0809	0.1408	-7.68	0.000
age	-0.0112	0.0040	-2.81	0.005
ses	0.1362	0.0910	1.50	0.134
density	-0.0000	0.0000	-0.64	0.525
foreign	-0.0058	0.0119	-0.49	0.628
crime rate	0.0000	0.0000	0.58	0.561
constant	1.7811	0.4113	4.33	0.000
Respondents + nonresponse follow-up + weights: $RNRW$: $N_{obs} = 1096$				
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	P-VALUE
sex	-1.2218	0.1284	-9.51	0.000
age	-0.0110	0.0036	-3.01	0.003
ses	0.1698	0.0843	2.02	0.044
density	0.0000	0.0000	0.09	0.931
foreign	0.0010	0.0110	0.09	0.930
crime rate	-0.0000	0.0000	-1.43	0.152
constant	2.0378	0.3739	5.45	0.000

Table 2: Results based on respondents

MI respondents: RMI: $N_{obs} = 3868$					
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	DF	P-VALUE
sex	-1.1636	0.1174	-9.91	9	0.0000
age	-0.0089	0.0030	-2.93	14	0.0109
ses	0.2847	0.0869	3.28	8	0.0112
density	-0.0001	0.0000	-1.87	19	0.0768
foreign	-0.0114	0.0089	-1.28	14	0.2207
crime rate	0.0000	0.0000	1.26	18	0.2251
constant	1.5683	0.2472	6.34	57	0.0000
MI respondents + nonresponse follow up : $RNRMI$: $N_{obs} = 3868$					
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	DF	P-VALUE
sex	-1.2520	0.0927	-13.50	19	0.0000
age	-0.0088	0.0024	-3.68	71	0.0004
ses	0.3342	0.0658	5.08	19	0.0001
density	-0.0000	0.0000	-1.22	19	0.2390
foreign	0.0030	0.0096	0.31	11	0.7617
crime rate	-0.0000	0.0000	-1.33	6	0.2328
constant	1.6242	0.2736	5.94	24	0.0000
MI respondents + nonresponse + short follow up: $RNRNR2MI$: $N_{obs} = 3868$					
VARIABLE	ESTIMATE	STD.ERR.	T-RATIO	DF	P-VALUE
sex	-1.3064	0.0821	-15.91	48	0.0000
age	-0.0092	0.0024	-3.82	65	0.0003
ses	0.2855	0.0576	4.95	49	0.0000
density	-0.0000	0.0000	-1.17	33	0.2520
foreign	0.0066	0.0106	0.62	8	0.5529
crime rate	-0.0000	0.0000	-3.17	15	0.0064
constant	1.9419	0.2645	7.34	32	0.0000

Table 3: Results based on the $m = 5$ imputed data sets