Robust Multiple Imputation

Dissertation zur Erlangung des Doktorgrades an der Fakultät
Erziehungswissenschaft, Psychologie und Bewegungswissenschaft,
Fachbereich Psychologie der Universität Hamburg
vorgelegt von R.N. de Jong
Hamburg, 2012

21.11.2012
Prof. Dr. Martin Spieß
Prof. Dr. Stef van Buuren

I shall certainly admit a system as empirical or scientific only if it is capable of being *tested* by experience. These considerations suggest that not the *verifiability* but the *falsifiability* of a system is to be taken as a criterion of demarcation ... *It must be possible for an empirical scientific system to be refuted by experience.*

The Logic of Scientific Discovery (1934)
KARL POPPER

If your experiment needs statistics, you ought to have done a better experiment.

The Mathematical Approach to Biology and Medicine (Norman T.J. Bailey, 1967)
ERNEST RUTHERFORD

Contents

1 Introduction 1

1.1 Methods for handling missing data . . . 2

1.1.1 Complete Case Analysis . . . 2

1.1.2 Maximum Likelihood . . . 2

1.1.3 Weighting . . . 3

1.2 Multiple imputation . . . 3

1.3 Research Goals . . . 6

2 The Analyst 10

2.1 ADGP . . . 11

2.1.1 Population model . . . 11

2.1.2 Discussion of assumptions . . . 12

2.2 Imputed Data Analysis . . . 14

2.3 Missing Data Mechanism . . . 16

2.3.1 Missing at Random and Friends . . . 16

2.3.2 Missings in Multiple Variables . . . 17

2.3.3 Ignorability . . . 18

2.3.4 CCA or MII . . . 18

2.3.5 Discussion of assumptions . . . 21

3 The Imputer 22

3.1 Pooling . . . 22

3.2 Compatibility . . . 24

3.3 Imputation by reverse regression . . . 26

3.3.1 The IDGP . . . 26
3.3.2 Compatibility check . . . 27
3.3.3 Reverse Linear Regression . . . 29
3.3.4 Transformations . . . 30
3.3.5 Discussion of assumptions . . . 31
3.4 Multivariate missing data . . . 32
3.4.1 Joint Modeling . . . 33
3.4.2 Fully Conditional Specification . . . 33
3.5 Conclusion . . . 36

4 Imputation Methods 37

4.1 Global Linear Regression (GLR) and Generalized Global Linear Regression (GGLR) . . . 38
4.2 Predictive Mean Matching . . . 39
4.3 `aregImpute` . . . 41
4.4 Local Linear Regression . . . 43
4.5 GAMLSS . . . 45
4.5.1 The IDGP . . . 45
4.5.2 Implementation . . . 46
4.5.3 Extension to multilevel models . . . 48
4.5.4 Discussion of Assumptions . . . 49

5 Simulation Experiments 50

5.1 The use of Monte Carlo Statistical Simulation . . . 50
5.2 Existing work . . . 52
5.3 Design . . . 52
5.3.1 Simple . . . 52
5.3.2 Multiple . . . 58
5.4 Results . . . 59
5.4.1 Normal . . . 59
5.4.2 Skew-Normal and Uniform . . . 64

5.4.3 Uniform Squared (Beta) 72

5.4.4 Student's T 77

5.4.5 Bernoulli 78

5.4.6 Poisson 81

6 Conclusion & Summary 84

6.1 Research Questions 84

6.2 Recommendations 88

6.2.1 MI Users 88

6.2.2 MI Researchers 88

A Simulation Framework 91

B GAMIMP 110

Acronyms

ADGP Analyst DGP

CCA Complete Case Analysis

DAP Data Analysis Procedure

DGP Data Generating Process

FCS Fully Conditional Specification

GGLR Generalized Global Linear Regression

GLM Generalized Linear Model

GLR Global Linear Regression

IDA Imputed Data Analysis

IDGP Imputer DGP

IM Imputation Method

IMDS Imputed Data Set

IPW Inverse Probability Weighting

JM Joint Modeling

LLR Local Linear Regression

LM Linear Model

MCSS Monte Carlo Statistical Simulation

MDM Missing Data Mechanism

MI Multiple Imputation

MII Multiple Imputation Inference

ML Maximum Likelihood

OLS Ordinary Least Squares

PMM Predictive Mean Matching

Chapter 1

Introduction

Statistical inference is used in empirical studies to draw conclusions from data that are subject to random variation. Statistical inference about population quantities of interest requires assumptions about the process which generated the data, or Data Generating Process (DGP). When empirical studies are affected by missing data, as is often the case, analysts also need to make assumptions about the process that caused the missing data, or Missing Data Mechanism (MDM), either explicitly by extending the Data Generating Process to formalize knowledge about the MDM, or implicitly by omitting such specification[1]. Statistical inference based on incomplete data is only valid when the specified DGP, including the MDM, is sufficiently in concordance with the unknown true DGP. The term valid is used loosely here, and will be defined more precise with respect to several modes of inference later on.

A multitude of methods of varying complexity have been proposed to perform statistical inference when the data set is afflicted by missing values, ranging from simple ad hoc methods to techniques with sophisticated statistical underpinnings. Some commonly used strategies for tackling the missing data problem are briefly discussed in Section 1.1, after which the method of primary focus of this work, multiple imputation, is reviewed in Section 1.2. All techniques are illustrated by applying them to the problem of estimating the average weight of a female population when a considerable number of sampled women refuse to be weighted. Finally, in Section 1.3, a number of problems and open questions with respect to how Multiple Imputation (MI) is currently used are identified; these questions will serve as the point of departure of this work.

[1] Applied researchers are often not fully aware that omitting the specification of a Missing Data Mechanism effectuates a "default" MDM; this default MDM is not necessarily equal to the unknown true MDM.

1.1 Methods for handling missing data

1.1.1 Complete Case Analysis

One simple and widely used technique is Complete Case Analysis (CCA), or listwise deletion. After removal of all units with at least one missing datum, the Data Analysis Procedure (DAP) is performed unmodified as if there are no missing values: the MDM is assumed to be "neutral", and is ignored. It is important to realize that the ability to ignore the MDM, or ignorability, is not solely a property of the unknown true MDM, but also depends on the Data Analysis Procedure; a MDM might be ignorable for a maximum likelihood estimator, but not ignorable for a GEE-estimator (generalized estimating equation, see Zeger & Liang (1986)). Two major advantages of CCA are that it allows the use of standard statistical software which the user is familiar with, and that no additional modeling effort is required: therefore, when applicable, CCA is the method of choice. Conditions for ignorability will be discussed in Section 2.3.3 with respect to the DAP of interest in this work.

With respect to the weighting of females, suppose we correctly assume that the probability of a missing value is the same for all sampled females, and therefore independent of weight itself. It can be trivially shown that this MDM is ignorable with respect to the sample mean, so CCA is our procedure of choice: we calculate the sample mean after removing all units with a missing value.

1.1.2 Maximum Likelihood

A more generally applicable and sophisticated approach is likelihood-based inference with incomplete data (ML). Given all available data, which typically includes the observed part of units with one or more missing values, a likelihood function derived from the specified DGP, which might include a MDM, is maximized over parameters of the DGP. The DGP is typically extended to random variables which were previously conditioned upon or marginalized over; for example, in a regression setting with missing values in a predictor, the full joint likelihood needs to be specified, instead of formulating the conditional likelihood of the response variable given the predictor. Because the specified DGP often contains external information which is specific to the situation at hand, there might be no off-the-shelf Data Analysis Procedure available. Compared to CCA, ML is more general, since it can be used with any MDM, and not just ignorable ones; further, ML uses all the data that is available, since it allows for the use of the observed part of units with one or more missing values.

In our example, the researcher suspect that his data are truncated such that all sampled obese females over 90 kg refuse to be weighted. Therefore, he specifies that the data follow a log-normal distribution truncated from the right at 90 kg, and estimates the parameters of this distribution using ML to obtain an estimate of the population mean. It should be emphasized that the inference is highly sensitive to the assumed truncation point.

1.1.3 Weighting

Another sophisticated and arguably general approach to the missing data problem is Inverse Probability Weighting (IPW), where the data of a unit is weighted according to the inverse of the estimated or known[2] probability of observing that particular unit. In our example, the analyst believes that the probability of non-response increases monotone as a function of weight; therefore, he assigns larger sampling weights to heavy females. IPW is generally applicable; a large class of Data Analysis Procedures allows for the specification of sampling weights.

Just as CCA, IPW can be relatively inefficient compared to other approaches, since the observed part of units with one or more missing values is also discarded (Carpenter & Kenward, 2006). This efficiency problem can be alleviated by using doubly robust estimators, which require both the specification and estimation of (a) a parametric model for the response probabilities and (b) a parametric model for the observed and missing data. Doubly robust estimators have the advantageous property that they are robust against misspecification of either model (a) or (b), but not both: at least one of the models needs to be correct. A disadvantage of IPW and doubly robust estimators is possible instability of the estimator when the probability of observing some units approaches zero, and the corresponding sampling weights approach infinity; however, the estimators recently proposed by Tan (2008) have desirable properties in boundedness even if the inverse probability weights are highly variable.

1.2 Multiple imputation

MI (Rubin, 1987) is envisioned as a statistical mode of inference to draw conclusions from incomplete data sets. It involves generating plausible values, called imputations, for each missing datum in the data set. These imputations are generated by an Imputation Method (IM), and are based on the incomplete data set and assumptions formulated in a separate DGP, called the IDGP. The resulting Imputed

[2] Paradoxically, it is more efficient to use estimated than true probabilities (Robins, 1995).

Data Sets (IMDSs), which are free of missing data, combined with the DGP necessary for analysis, called ADGP from now on, are used to perform an Imputed Data Analysis (IDA) for the parameters of scientific interest. Thus, the IDGP contains assumptions necessary for generating imputations, and the ADGP contains assumptions necessary for performing the Imputed Data Analysis. In the MI framework, an analyst is never confronted with incomplete data, and always specifies the ADGP as if there are no missing data: the MDM is ignored for analysis. Likewise, the IDA equals the Data Analysis Procedure which would be applied to completely observed data. Although the MDM is always ignorable for the Imputed Data Analysis, the IDGP might contain the explicit specification of a hypothesized MDM; more often, the MDM is also ignored for imputation.

In our example, the IDA estimator equals the sample mean, and the ADGP consists of the assumption that the observations are i.i.d., and the assumption that the first moment of the random variable weight exists. In the IDGP, the imputer posits a log-normal distribution for the missing and observed weights, where the median of the missing weights is shifted to the right by 20 kg with respected to the log-normal distribution of the observed weights. After estimating the parameters of the log-normal distribution of the observed weights, imputations for the unobserved weights are then drawn from a log-normal distribution whose median is shifted to the right by 20 kg with respect to the distribution fitted to the observed data. Then, the analyst can consistently estimate the mean weight by computing the sample mean from the IMDS.

Unfortunately, treating observed and imputed data on equal footing generally leads to invalid inference, since the analysis does not take into account the additional uncertainty about the imputed data. MI is designed to solve this deficit; in contrast with single imputation, MI requires the imputation and analysis step to be performed at least two times, after which the resulting IDAs are aggregated or “pooled” to form the final Multiple Imputation Inference (MII) using certain rules; a schematic overview of the procedure is depicted in Figure 1.1 on page 5. Because the multiple imputations are not deterministic predictions, but consist of random draws from the predictive distribution implied by the specified IDGP, the point estimates of the IDAs vary; this variance between the multiple point estimates represents the uncertainty about the imputed data, and is incorporated into the MII by the pooling rules.

MI was originally conceived for inference within a survey context, where a design-based perspective is prevalent. This perspective comprises of a concrete population consisting of a finite number of units with properties represented by fixed variables; randomness is induced by drawing a sample with known sampling probability from

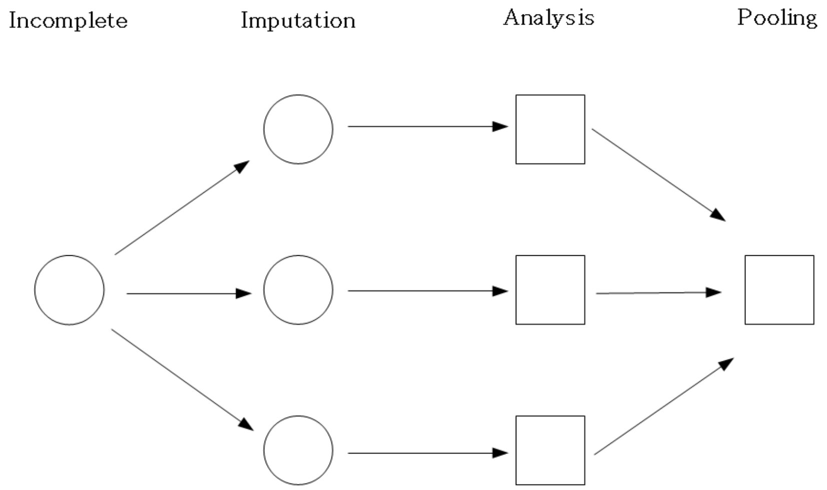

Figure 1.1: Overview of the MI procedure.

this population. However, DAPs used in the social sciences are often based on the linear model, where the sample originates from an infinite hypothetical population, and the properties of the population units are realizations of random variables; assumptions about the unknown properties of these random variables and their relations are formalized in the DGP. Nevertheless, there are no theoretical objections to application of MI outside survey contexts (Rubin, 1987), and in this work MI will be investigated and evaluated from a frequentist model-based perspective.

The principal merit of MI is the separation of the missing data problem from the DAP. As a consequence, the entity which produces the Imputed Data Sets, called imputer, and the entity which analyzes the IMDSs, called analyst, need not be the same. Data analysts who lack the necessary skills and knowledge to correctly handle the missing data problem themselves can thus continue to perform the DAP using the methods and computer programs they are familiar with. One minor caveat is the need to combine the Imputed Data Analyses; fortunately, there also exists software which handles this quite conveniently either by combining the provided collection of IDAs, or encapsulating calculation and pooling of the IDAs given the IMDSs. Previous concerns stemming from the increased computational burden and storage costs associated with imputing and analyzing the data multiple times have been rendered void with the technological advances commonly available nowadays.

MI effectively delegates the task of solving the missing data problem to the imputer, who needs to specify an IDGP. The IDGP contains a specification of the properties of the random variables with missing values, including their relation with other variables, and may contain a hypothesized MDM. Given that the Data Analysis Procedure has favorable frequentist properties if there are no missing data, a necessary condition for validity of the MII is compatibility of the IDGP with the Imputed

Data Analysis. More specifically, the IDGP is compatible when the imputed random variables have the properties necessary for obtaining favorable frequentist properties of certain aggregates of the IDAs. The astute reader will remark that compatibility of an IDGP is defined with respect to an Imputed Data Analysis, and not with respect to an ADGP; an IDGP might not be compatible with every IDA associated with an ADGP (see Nielsen (2003) for examples).

Although the missing data problem is separated from the IDA at the procedural level, it remains tightly coupled with the IDA at the modeling level because compatibility of the IDGP depends on the IDA chosen by the analyst. Incompatibility of the IDGP typically arises when the analyst and imputer insufficiently communicate. It also occurs when the same set of IMDSs is analyzed multiple times by possibly different analysts to obtain unique MIIs, each with an own IDA; this is an intended usage scenario of MI. In this case, the MIIs are valid only when the IDGP is compatible with all IDAs. Absolutely speaking, there exists no regular parametric IDGP which encapsulates the collection of all parametric ADGPs. On the other hand, nonparametric methods suffer from the curse of dimensionality, and typically exhibit poorer finite-sample performance. Creating the ultimate IM which produces valid MII for all ADGPs is therefore likely to be an impossible goal.

It is also possible that the CCA and MII are both valid, but differ in efficiency[3] because the imputer uses less or more information than the analyst; this is a special form of what Meng (1994) calls uncongeniality. When the imputer correctly uses more information than the analyst, the MII is superefficient (Meng, 1994). A superefficient MII is characterized by confidence intervals which are narrower than those of the CCA, but feature at least nominal coverage. On the other hand, when the imputer uses less information than the analyst, MII might be less precise compared to the CCA. Although this special case of uncongeniality (the remaining cases are caused by incompatibility of the IDGP) has gained a lot of attention in the literature and is technically interesting, it does not lead to invalid MII, and is considered to be of secondary importance.

1.3 Research Goals

Summarizing, MI can potentially be used in conjunction with a wide range of existing IDAs[4], while relieving the analyst from the burden of missing values by separating

[3] The efficiency comparisons are made asymptotically, as the number of imputations goes to infinity.

[4] The IDA should be self-efficient (Meng, 1994): the efficiency of the estimator used in the IDA should decreases when there is less data available.

the procedure for handling the missing data from the data analysis, and delegating the task of solving the missing data problem to the imputer. In most applications of MI in the social sciences, the MDM is ignored for imputation, and the role of imputer is played by aforementioned imputation software which automagically [sic] renders the imputations, without using substantial background knowledge about the incomplete data set. Most certainly, these software packages made MI in the eyes of many practitioners an attractive solution to the missing data problem. The practice of applying canned solutions to statistical problems certainly is not new, as apparent in the wide distribution of easy to use statistical software such as SPSS within the academic world; in fact, MI routines have recently been incorporated in SPSS. In itself, delegating calculation of an estimator to a computer program is bona fide; however, users are often unaware of the assumptions accompanying such automated procedures. Analogue to this ritualized use of software, users of MI packages are often enticed to ignore the IDGP underlying the implemented IM. To their defense, the documentation of IMs often does not clearly state the IDGP, and even when the IDGP is specified, deducing the class of compatible IDAs remains difficult. In practice, MII tends to be untransparent compared to ML inference with missing data, because ML inference keeps the missing data problem and the DAP unified. When the IDGP is in fact known, the posited assumptions are rarely completely met, as is the case with most statistical applications; a natural concern is the robustness of MII to violations of the assumptions posited in the IDGP.

Existing research on the robustness of MII has primarily focused on IDAs based on marginal statistics such as the sample mean (Schenker, 1996; Little & An, 2004); perhaps the insistence on marginal statistics stems from the survey context from which MI originates, where most inference procedures are about such population quantities. In the social sciences, the parameters of scientific interest are often the regression coefficients of a Linear Model (LM); these coefficients can also represent treatment effects in simple experimental designs, and are predominantly estimated using the Ordinary Least Squares (OLS) estimator. Both the LM and the OLS estimator will be described in detail in Chapter 2. The main goal of this work is to assess the robustness of MII with respect to the parameters of the linear model using Monte Carlo Statistical Simulation (MCSS); existing research to this end will be briefly reviewed in Section 5.2. Of course, the robustness of MII depends on the used IM; to safeguard scientific reproducibility, transparency, and practical relevancy, IMs under consideration are those implemented and made available for the open-source statistical environment and programming language R (R Development Core Team, 2011). Further, in an attempt to improve upon existing IMs, a new semiparametric IM is proposed in Section 4.5; an implementation of this method is written for the R environment, and a program listing is given in Chapter A[5]. Note that in this work, the MDM is always assumed ignorable for imputation.

Research objective: To assess the robustness of MII — as based on currently available and widely used IMs — with respect to parameters of the LM, and to improve upon existing IMs.

Below, several research questions are listed which serve as a point of departure for reaching the research objective stated above.

Research question 1 When is MI beneficial? (Chapter 2)
After defining the ADGP, DAP, CCA, and IDA, it will be investigated when the MDM is ignorable with respect to the CCA. All possible MDMs are classified, and related to the nomenclature introduced by Rubin (1976). For each class of MDM, a recommendation of either MI or CCA will be given after carefully weighting the advantages and disadvantages of both approaches with respect to validity and efficiency.

Research question 2 When is an IM compatible with the OLS estimator? (Chapter 3)
After defining the MI estimator, the compatibility of an important class of IMs will be investigated with respect to the IDA previously defined in Chapter 2. In particular, it is investigated how to properly impute transformations and interactions of predictor variables.

Research question 3 Which IMs are currently available, and what are their IDGPs? (Chapter 4)
Some widely used and freely available IM will be classified and described. Also, possible incompatibilities between their IDGPs and the OLS estimator are indicated.

Research question 4 How can existing IMs be improved? (Chapter 4)
A new semiparametric robust IM will be proposed, which models the parameters of a specified conditional distribution of the variable to be imputed using smooth functions of predictor variables.

Research question 5 How do IMs perform empirically? (Chapter 5)
After providing an overview of MCSS, extensive simulation studies will be performed which allows for a comparison of the performance (robustness) of some of the IMs listed in Chapter 4, including the proposed one, under a variety of scenarios.

[5] This work is a result of the DFG financed project “Robust and efficient multiple imputation of complex data sets”, which stated more ambitious goals. During the project, it became clear that those goals are hard to obtain; see Chapter 6 for a detailed account.

Finally, all research findings will be summarized and reflected upon in Chapter 6, including retrospective contemplations about the goal of the project, followed methodological approach, and encountered hurdles. Areas of future research are indicated, and recommendations for future researchers and practitioners are given.

Chapter 2

The Analyst

In this chapter, those parts of the Multiple Imputation Inference (MII) procedure are described which principally involve the analyst, starting in 2.1 with a specification and discussion of the Linear Model (LM), which is the ADGP to which this work is constrained. Secondly, the Complete Case Analysis (CCA) and Imputed Data Analysis (IDA) will be defined in 2.2; both consists of the Ordinary Least Squares (OLS) estimator and associated variance estimator. Thirdly, the Missing Data Mechanism (MDM) will be formalized in Section 2.3, along with a taxonomy of MDMs. Also, it will be indicated when a MDM is ignorable for the CCA and MII. Since the analyst always ignores the MDM in a Multiple Imputation (MI) setting, and formulating the MDM is a task of the imputer, discussion of the MDM might seem misplaced in this chapter. However, it is the analyst who must ultimately assess if the acquired Imputed Data Sets are indeed beneficial for the analysis at hand: some MDMs are ignorable for a CCA, but not ignorable for MII, in which case the imputations are better discarded. Therefore, for all identified classes of MDM, a recommendation of either MI or CCA will be given after carefully weighting the advantages and disadvantages of both approaches with respect to validity, and to a lesser extent, efficiency. Key assumptions and limitations of the IDA are discussed in 2.1.2 and 2.3.5.

2.1 ADGP

2.1.1 Population model

In the LM, the response variable y follows the following population regression function:

$$\begin{aligned} y &= \mathrm{E}\left(y|\boldsymbol{x}, s\right) + u \\ &= \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + u \end{aligned} \tag{2.1}$$

where y is the response variable, α is the intercept, $\boldsymbol{x}$ is a $k \times 1$ column vector with predictors, $\boldsymbol{\beta}$ is a $k \times 1$ vector with the associated regression coefficients, u is a latent error variable with $\mathrm{E}\left(u\right) = 0$, and s is a selection indicator with $s = 1$ indicating that $\{\boldsymbol{x}, y\}$ is observed, and $s = 0$ indicating that at least one element of $\boldsymbol{x}$ or y has a missing value. It is assumed the observations are i.i.d. (independent and identically distributed); therefore, the index enumerating the units of the population is suppressed. Notationally, no difference is made between random variables and their realizations. Note that we make explicit the often implicit assumption that the conditional expectation for units with missing values ($s = 0$) equals the conditional expectation for units without missing values ($s = 1$): the MDM is ignored. Moreover, since we are solely interested in features of the distribution of y conditional on $\boldsymbol{x}$, properties of $\boldsymbol{x}$ are omitted from the model specification. The parameter of scientific interest is $\boldsymbol{\beta}$, where each $\boldsymbol{\beta}_i$ represents the expected change in y when the corresponding predictor $\boldsymbol{x}_i$ has increased with one unit.

In the social sciences, measurement form at best interval[1] scales which are constructed by adding an unknown constant c to a ratio scale, as opposed to interval scales such as Celsius for which c is known; therefore, α has typically no structural interpretation, and is scientifically meaningless, although estimating α for the purpose of prediction is legitimate. Even when the outcome measure is a ratio scale such as response time, the marginal mean is seldom of interest.

The assumption that (2.1) is the true model implies that the errors are mean independent of the predictors and the selection indicator:

$$\begin{aligned} \mathrm{E}\left(u|\boldsymbol{x}, s\right) &= \mathrm{E}\left(y - \mathrm{E}\left(y|\boldsymbol{x}, s\right)|\boldsymbol{x}, s\right) \\ &= \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} - \alpha - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} \\ &= 0; \end{aligned} \tag{2.2}$$

[1] The social sciences lack a solid measurement-theoretic foundation; Steven's theory of scales of measurement acts as a surrogate (Michell, 2008). This theory proposes that measurements can be classified into four different types of scales: nominal, ordinal, interval, and ratio.

this is the the key condition sufficient for obtaining consistency of the estimator of $\boldsymbol{\beta}$ as described in Section 2.2.

2.1.2 Discussion of assumptions

The model (2.1) specifies that the linear relationship between $\boldsymbol{x}$ and y is identical for all units in the population. When the relationship between $\boldsymbol{x}$ and y is linear for all population units, but not identical, $\boldsymbol{\beta}$ can only be interpreted as the expected population-averaged change in y for a one unit increase of $\boldsymbol{x}$. No other family of functions has the property that an average of several members from the family is also in the family (Luce, 1997). For example, while the linear relationships between the grades and IQ scores of the n students equal

$$grade_i = \alpha_i + IQ_i\beta_i + u_i \qquad \text{for all } i \in \{1, \dots, n\}, \tag{2.3}$$

the population averaged regression equals as implied by (2.1)

$$grade_i = \bar{\alpha} + IQ_i\bar{\beta} + u_i \qquad \text{for all } i \in \{1, \dots, n\}, \tag{2.4}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are population-averaged regression coefficients. Thus, (2.4) indicates that a one point increase in IQ leads to an expected grade increase of $\bar{\beta} = 2$ units; however, it does not necessarily hold that for pupil i the expected grade increases with $\beta_i = 2$ points[2].

A considerable limitation of the LM is that all parameters are forced to enter the model linearly. It is possible to add transformations of variables to the model; this nevertheless presupposes that the modeler knows the true functional relationship between the predictors and y. Unfortunately, most theories in the social sciences do not explicitly state functional relationships. In fact, most applications of the LM in the social sciences have the primary goal of accumulating evidence in favor of the obvious and trivial fact that $\boldsymbol{\beta} \neq \mathbf{0}$. Further, the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$ is often falsely interpreted as the absence of any type of relationship, while more statistically versed researchers keep “improving” the model by selecting suitable transformations by trial and error, jeopardizing the validity of statistical tests.

A necessary condition for (2.2) to hold is that either all relevant predictors are included in the model, or that the excluded but relevant predictors are mean independent of $\boldsymbol{x}$ and s. This condition is fulfilled in randomized experiments for

[2] In this particular example, the relationship between IQ and grade cannot be causal on the student level since IQ is an attribute; a student is not potentially exposable to all possible levels of IQ (Holland, 1986).

estimation of the treatment effect, since possible confounding effects of excluded predictors are neutralized through the random assignment of units to experimental conditions. However, if in a quasi-experimental research setting there are auxiliary predictors $\boldsymbol{z}$ such that in truth

$$\mathrm{E}\left(y|\boldsymbol{x},\boldsymbol{z},s\right)=\alpha+\boldsymbol{x}\boldsymbol{\beta}+\boldsymbol{z}\boldsymbol{\gamma}, \tag{2.5}$$

but these predictors are excluded from the model, then a necessary condition for (2.2) to hold is that

$$\mathrm{E}\left(\boldsymbol{z}|\boldsymbol{x},s\right)=\mathrm{E}\left(\boldsymbol{z}\right). \tag{2.6}$$

To see this, observe that upon exclusion the effect of the auxiliary predictor $\boldsymbol{z}\boldsymbol{\gamma}$ is “absorbed” by the augmented error term $u^{*}=\boldsymbol{z}\boldsymbol{\gamma}+u$ such that

$$\begin{aligned}\mathrm{E}\left(u^{*}|\boldsymbol{x},s\right)&=\mathrm{E}\left(\boldsymbol{z}|\boldsymbol{x},s\right)\boldsymbol{\gamma}+\mathrm{E}\left(u|\boldsymbol{x},s\right).\\&=\mathrm{E}\left(\boldsymbol{z}\right)\boldsymbol{\gamma}+0\\&=\mathrm{E}\left(u^{*}\right).\end{aligned} \tag{2.7}$$

An often overlooked asymmetry in the LM is that in contrast to y, which includes the random error component u, the predictors $\boldsymbol{x}$ are assumed to be free of measurement error; violation of this assumption may lead to attenuated estimates of $\boldsymbol{\beta}$[3]. Finally, it is implicitly assumed that there is no feedback from the response variable to the predictors; this seems overly simplistic given the dynamic nature of human behavior. For instance, given the relation between the grade on a test y and fear of failure x, an increased fear of failure leads to lower grades, which in turn leads to an increased fear of failure; when an ADGP with a single equation is assumed, estimates of β will in this case be biased towards zero.

In quasi-experimental settings, the maintenance of the liberal and strong assumptions of the LM decreases the credibility of statistical inference. When (2.2) does not hold due to misspecification of the functional form, omission of relevant predictors, measurement error in the predictors, feedback from the response variable to the predictors, or non ignorability of the MDM, any causal interpretation is no longer possible, and the estimates can only be interpreted as a linear projection. However, linear projections typically exhibit lackluster predictive power; if prediction is the primary objective, there are superior alternatives available which are

[3] Although in some applications the relationship between the response quantity and contaminated predictor is of legitimate interest, such as the relationship between the perceived number of alcoholic consumptions x and drunkenness y (although in this example there is probably also a feedback effect).

not hampered by the linearity assumption, and are therefore better at tracking the data. At best, the LM is used as a mathematically convenient approximation to the analysis of variance of simple experimental designs (for more involved designs such as the split-plot or hierarchical designs, LM fails to give the right answer (Gelman, 2005)). Nevertheless, due to its prevalence in applied research, (2.1) is the ADGP under consideration in this work.

2.2 Imputed Data Analysis

In this section we define the IDA the analyst is assumed to use, which consists of an estimator of $\boldsymbol{\beta}$, and a measure of the precision of this estimator. Suppose a random sample of size N is drawn where all $\{\boldsymbol{x}_i, y_i, s_i\}_{i=1}^N$ are independent and identically distributed (i.i.d.) according to some joint distribution such that (2.1) holds. The sample is described by the following random variables:

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} - \bar{\boldsymbol{x}}^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} - \bar{\boldsymbol{x}}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} - \bar{\boldsymbol{x}}^{\mathrm{T}} \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{bmatrix} \qquad \boldsymbol{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix}, \tag{2.8}$$

where $\mathbf{X}$ and $\boldsymbol{y}$ are demeaned. The observations are set in deviations from the sample mean because the intercept α is often of little scientific interest (as explained in 2.1). Further, of the N units in the sample, the number of completely observed units is defined as $n_{\text{obs}} = \sum_{i=1}^N \boldsymbol{s}_i$, and the number of cases with at least one missing value as $n_{\text{mis}} = N - n_{\text{obs}}$.

The estimator of $\boldsymbol{\beta}$ is the OLS-estimator:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \boldsymbol{s}) = \boldsymbol{\Psi}^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{C}\boldsymbol{y}), \tag{2.9}$$

where

$$\begin{aligned} \boldsymbol{\Psi} &= (\mathbf{X}^{\mathrm{T}}\mathbf{C}\mathbf{X}) \\ \mathbf{C} &= \operatorname{diag} \boldsymbol{s}, \end{aligned} \tag{2.10}$$

with $\mathbf{C}$ being the selection matrix. Provided that 2.2 holds, it can be shown that $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is consistent and unbiased (Cameron & Trivedi, 2005).

For estimating the precision of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, the ADGP is extended with the assumption that the errors u in (2.1) are homoscedastic:

$$\mathrm{Var}(u|\boldsymbol{x}, s) = \sigma^2. \tag{2.11}$$

Assumptions (2.2) and (2.11) can be summarized as

$$\mathrm{E}\left(u^l|\boldsymbol{x}, s\right) = \mathrm{E}\left(u^l\right) \qquad l \in \{1, 2\}. \tag{2.12}$$

When (2.12) holds, the limit distribution of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}_N} - \boldsymbol{\beta}) &\xrightarrow{\mathcal{D}} (\mathrm{plim}\, N^{-1}\boldsymbol{\Psi}_N)^{-1}(\mathrm{dlim}\, \frac{1}{\sqrt{N}}\mathbf{X}_N^{\mathrm{T}}\mathbf{C}_N\boldsymbol{u}_N) \\ &= \mathcal{N}(0, (\mathrm{plim}\, N^{-1}\boldsymbol{\Psi}_N)^{-1}\sigma^2),\end{aligned} \tag{2.13}$$

where dlim denotes convergence in distribution.

The variance of the errors σ^2 can be consistently estimated using

$$\widehat{\sigma^2} = (n_{\mathrm{obs}} - k - 1)^{-1}\boldsymbol{y}^{\mathrm{T}}\mathbf{C}(\mathbf{I}_N - \mathbf{C}\mathbf{X}\boldsymbol{\Psi}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{C})\boldsymbol{y}, \tag{2.14}$$

such that the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ can be consistently estimated as

$$\hat{V}_{\mathrm{OLS}}(\mathbf{X}, \boldsymbol{y}, \boldsymbol{s}) = \widehat{\sigma^2}\boldsymbol{\Psi}^{-1}. \tag{2.15}$$

Both the CCA and IDA, and also the Data Analysis Procedure (DAP), are based on the OLS estimator:

$$\left.\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{CCA}}(\mathbf{X}, \boldsymbol{y}, \boldsymbol{s}) &= \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}(\mathbf{X}, \boldsymbol{y}, \boldsymbol{s}) \\ \hat{V}_{\mathrm{CCA}}(\mathbf{X}, \boldsymbol{y}, \boldsymbol{s}) &= \hat{V}_{\mathrm{OLS}}(\mathbf{X}, \boldsymbol{y}, \boldsymbol{s})\end{aligned}\right\} \tag{2.16}$$

$$\left.\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{IDA}}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}}) &= \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}}, \mathbf{1}) \\ \hat{V}_{\mathrm{IDA}}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}}) &= \hat{V}_{\mathrm{OLS}}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}}, \mathbf{1})\end{aligned}\right\} \tag{2.17}$$

where $\{\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}}\}$ is an Imputed Data Set (IMDS). Since there are no missing values left after imputation, $\boldsymbol{s}$ is fixed at $\mathbf{1}$ for the IDA[4]. Both the CCA and IDA are valid when their limit distribution equals (2.13); a sufficient condition for validity is (2.12). Note that the final Multiple Imputation Inference is a function of several IDAs.

[4] As will be discussed in 2.3, it is actually recommended to exclude cases with missing values in $\boldsymbol{y}$ from the IDA.

2.3 Missing Data Mechanism

2.3.1 Missing at Random and Friends

The MDM is formalized by assigning a conditional distribution to $\boldsymbol{s}$. Below, several types of MDMs are defined and illustrated using the problem of estimating the average weight of a female population, which was already discussed in Chapter 1. Suppose an i.i.d. sample of size N is drawn from the target population of females, where the random variable $\boldsymbol{y}$ represents the weights in kg, and $\boldsymbol{s}_i$ indicates if $\boldsymbol{y}_i$ is observed or missing. However, some females refuse to be weighted, and $\boldsymbol{y}$ is consequently afflicted by missing values. We denote the random variables which are observed by $\boldsymbol{y}_{\text{obs}} = (\boldsymbol{y}_i)_{i \in 1,\ldots,N : \boldsymbol{s}_i = 1}$, and the variables whose values are missing by $\boldsymbol{y}_{\text{mis}} = (\boldsymbol{y}_i)_{i \in 1,\ldots,N : \boldsymbol{s}_i = 0}$. The MDMs below are ordered, starting with those which are most restrictive with respect to the DAP.

Missing Not At Random (MNAR)

The selection indicator is dependent on $\boldsymbol{y}_{\text{mis}}$. For example, consider the following MDM, which is a deterministic function of y

$$p(\boldsymbol{s}|\boldsymbol{y}) = \prod_{i=1}^{N} I(y_i < a)^{\boldsymbol{s}_i} I(y_i \geq a)^{1-\boldsymbol{s}_i} \qquad a > 0, \tag{2.18}$$

where $I(\cdot)$ is the indicator function, and our measurement device critically fails for females heavier than a. It is trivial to show that the expected value of the truncated sample does not equal the expected value of the sample; the MDM is therefore not ignorable for any mode of inference.

Missing At Random Locally (MAR-L)

The selection indicator only depends on $\boldsymbol{y}_{\text{obs}}$. It is important to note that this assumption is only made for the realized value of $\boldsymbol{s}$, and is therefore called a local assumption (Jaeger, 2005). It can be shown that, conditional on $\boldsymbol{s}$, this restriction is sufficient for consistency of a maximum likelihood estimate of the mean. However, the MDM is not ignorable for the distribution of the maximum likelihood estimator (Nielsen, 1997).

Missing At Random Globally (MAR-G)

The distribution of the selection indicator only depends on $\boldsymbol{y}_{\text{obs}}$ for all possible realizations of $\boldsymbol{s}$[5]. The global MAR assumption and some technical conditions are sufficient to completely ignore the MDM for a maximum likelihood estimator of the mean of $\boldsymbol{y}$. In our example, the sensitive nature of the study makes

[5] See Nielsen (1997) for a more precise definition.

it unlikely that the probability of a missing weight for a female depends only on the observed weights of other females. Moreover, allowing the missingness of weight for a female to depend on the weight of other females contradicts the i.i.d. assumption about each observation $\{\boldsymbol{y}_i, \boldsymbol{s}_i\}$.

Missing Completely at Random Locally (MCAR-L)

The selection indicator is independent of $\boldsymbol{y}$. It is important to note that this assumption is only made for the realized value of $\boldsymbol{s}$, and is therefore called a local assumption (Jaeger, 2005). Rubin (1976) shows that, when the missing values are MCAR-L, the MDM is ignorable for inference which is conditional on $\boldsymbol{s}$ and based on the sample distribution of $\boldsymbol{y}$. Thus, inference based on for example the sample mean of $\boldsymbol{y}_{\text{obs}}$ is valid given the realized value of $\boldsymbol{s}$, but not necessarily for other realizations of $\boldsymbol{s}$; therefore, this inference is of questionable worth, at least from a frequentist perspective.

Missing Completely at Random Globally (MCAR-G)

The selection indicator is independent of $\boldsymbol{y}$ for all possible realizations of $\boldsymbol{s}$. This implies that the units for which $\boldsymbol{y}$ is observed are a simple random subsample of the original sample. Therefore, this MDM is unconditionally ignorable for all modes of inference. For example, inference based on the sample mean of $\boldsymbol{y}_{\text{obs}}$ is valid for all possible realizations of $\boldsymbol{s}$.

2.3.2 Missings in Multiple Variables

When there are missings in multiple columns of the $N \times k$ matrix $\mathbf{X}$, the selection indicator $\boldsymbol{s}$ is no longer suited, and the following observed data indicator matrix is used instead:

$$\mathbf{R} = \sum_{i=1}^{N} \sum_{j=1}^{k} I_{☠}(\mathbf{X}_{ij}) \mathbf{E}_{ij}, \tag{2.19}$$

where $I_{☠}(a)$ indicates if a is observed, and $\mathbf{E}$ is the $N \times k$ matrix with $\mathbf{E}_{ij} = 1$, and all other elements zero. All observed cases of $\mathbf{X}$ are indicated with $\mathbf{X}_{\text{obs}} = (\mathbf{X}_{ij})_{i,j:\mathbf{R}_{ij}=1}$, and all missing values of $\mathbf{X}$ by $\mathbf{X}_{\text{mis}} = (\mathbf{X}_{ij})_{i,j:\mathbf{R}_{ij}=0}$.

The relationship between the selection indicator and observed data indicator matrix is

$$\boldsymbol{s} = \sum_{i=1}^{N} \boldsymbol{e}_i \prod_{j=1}^{k} \mathbf{R}_{ij},$$

where $\boldsymbol{e}_i$ is the $n_{\text{obs}} \times 1$ vector with a one in the ith component, and zero elsewhere.

For a multivariate MDM involving two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ and an observed data indicator $\mathbf{R} = \begin{bmatrix} r_x & r_y \end{bmatrix}$, the missing data are MAR when for all i

$$p(\mathbf{R}_i|\boldsymbol{x}_i, \boldsymbol{y}_i) \;\;=\;\; \begin{cases} p_1 & \text{if } \mathbf{R}_i = \begin{bmatrix}0 & 0\end{bmatrix} \\ g_{01}(\boldsymbol{y}_i) & \text{if } \mathbf{R}_i = \begin{bmatrix}0 & 1\end{bmatrix} \\ g_{10}(\boldsymbol{x}_i) & \text{if } \mathbf{R}_i = \begin{bmatrix}1 & 0\end{bmatrix} \\ 1 - g_{01}(\boldsymbol{y}_i) - g_{10}(\boldsymbol{x}_i) - p_1 & \text{if } \mathbf{R}_i = \begin{bmatrix}1 & 1\end{bmatrix} \end{cases}, \qquad (2.20)$$

where the probability of missing values does not depend on $\boldsymbol{x}_i$ when $\boldsymbol{x}_i$ is missing, and where the probability of missing values does not depend on $\boldsymbol{y}_i$ when $\boldsymbol{y}_i$ is missing, and p_1 is a constant (Little & Rubin, 2002). The MDM (2.20) is therefore perhaps unrealistic, and becomes more implausible when the number of components of $\mathbf{X}$ that is affected by missing values increases. On the other hand, whereas the univariate MAR conditions discussed in 2.3 allowed the observed data indicator for a unit to depend on the observed values of other units, (2.20) respects the i.i.d. nature of the Data Generating Process (DGP) (2.1).

2.3.3 Ignorability

A sufficient condition for ignorability of the MDM with respect to the CCA is that the selection indicator is conditionally independent of $\boldsymbol{y}$ given $\mathbf{X}$:

$$p(\boldsymbol{s}|\mathbf{X}, \boldsymbol{y}) = \prod_{i=1}^{N} p(\boldsymbol{s}_i|\mathbf{X}_i). \qquad (2.21)$$

When (2.21) holds, the MDM is ignorable for the CCA, since $\boldsymbol{s} \perp \boldsymbol{y}|\mathbf{X} \Rightarrow \boldsymbol{u} \perp \boldsymbol{s}|\mathbf{X}$. The MDM is ignorable for imputation when condition (2.20) holds; a rationale will be given in Section 3.3.1.

2.3.4 CCA or MII

Confusingly, the mapping between the three types of MDMs, namely MCAR, MAR, and MNAR, and validity of the CCA is not injective; therefore, in Table 2.1 on page 19 it is indicated for all possible types of MDMs if the CCA is valid, and when imputation is recommended, with efficiency being of secondary concern. The different classes of MDMs defined in are discussed below:

Class	Missings in	$p(\boldsymbol{s}\vert\cdot)$	Rubin's term	CCA valid	Impute[a]
1	$\boldsymbol{y}$	$\mathbf{X}$	MCAR[b], MAR	Yes	No
2	$\boldsymbol{y}$	$\boldsymbol{y}$	MNAR	No	No
3	$\mathbf{X}$	$\mathbf{X}_{\text{obs}}$	MCAR[b], MAR	Yes	No
4	$\mathbf{X}$	$\mathbf{X}_{\text{mis}}$	MNAR	Yes	No
5	$\mathbf{X}$	$\boldsymbol{y}, \mathbf{X}_{\text{obs}}$	MAR	No	Yes

Table 2.1: Table exhaustively categorizing all classes of MDMs in terms of the nomenclature in Rubin (1976). For each class it is indicated if the CCA is valid with respect to $\boldsymbol{\beta}$, or that an alternative strategy such as imputation is recommended, efficiency considerations aside.

[a] Without using external information, and ignoring efficiency considerations.

[b] When $p(\boldsymbol{s}|\boldsymbol{y}, \mathbf{X}) = p(\boldsymbol{s})$

Class 1

Missing values in $\boldsymbol{y}$ are MAR or MCAR, which implies that the MDM is conditionally independent of $\boldsymbol{y}$ given $\mathbf{X}$, and ignorable with respect to the CCA. In general, imputation is not recommended in this case, since the units with missing $\boldsymbol{y}$ have zero information. To see this, a likelihood-based perspective is adopted, where the conditional likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{y}) = f(\boldsymbol{y}|\mathbf{X}; \boldsymbol{\theta}), \tag{2.22}$$

and $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$. Further, let $\mathbf{X}_{\text{mis}}$ represent the rows (units) of $\mathbf{X}$ for which the corresponding element of $\boldsymbol{y}$ is missing; $\mathbf{X}_{\text{mis}}$ itself is observed! Von Hippel (2007) shows that when the marginal likelihood of the observed data is calculated

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) &= \mathcal{L}(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) \int f(\boldsymbol{y}_{\text{mis}}|\mathbf{X}_{\text{mis}}; \boldsymbol{\theta}) d\boldsymbol{y}_{\text{mis}} \\ &= \mathcal{L}(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) \times 1,\end{aligned}$$

the likelihood of the cases with $\boldsymbol{y}$ missing is 1. Although we already know the answer, which is one, MII estimates this part of the likelihood by approximating the integral using simulation; imputation will thus only result in unnecessary estimation error. Further, there is the risk of imputing using an incompatible IDGP, which may lead to invalid MII.

One exception to the argument above occurs when the imputer uses information external to the analyst to produce "superefficient" MII (Meng, 1994), a subject which was already briefly touched in Section 1.2. The leading example is when the imputer uses fully observed auxiliary variables $\mathbf{Z}$, which are excluded from the ADGP but are nevertheless predictive for $\boldsymbol{y}$, to increase the precision of the imputations. When the IDGP includes auxiliary predictors $\boldsymbol{z}$, and the following conditions are satisfied:

1. (2.6) holds; otherwise, $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is inconsistent,
2. the number of imputations is “large enough”,

then MII is superefficient. Confidence intervals of $\boldsymbol{\beta}$ as produced by the MII will have nominal or larger coverage, but are narrower and thus considered "better" than those of the CCA. However, condition 1 is rather stringent, and 2 may be computationally prohibitive.

Class 2

The selection indicator is dependent on $\boldsymbol{y}$, and therefore the MDM is not ignorable for the CCA. Further, the missing values in $\boldsymbol{y}$ are MNAR, which violates (2.20), and implies that the MDM is also not ignorable for imputation: the IDGP must include a hypothesized MDM. In this work, only MDMs which are ignorable for imputation are considered.

Class 3

Missing values are in $\mathbf{X}$ and the MDM is dependent on $\mathbf{X}_{\text{obs}}$, and the CCA is valid, albeit inefficient. It is true that the loss of information suffered by CCA can become quite large when multiple predictors are afflicted by missing data. However, as we will see in Section 3.2 and Chapter 5, producing imputations which lead to valid MII is a very hard problem, where inadequate solutions easily lead to invalid MII. When the MDM is known to be independent of $\boldsymbol{y}$, analysts are therefore advised to count their blessings, discard any imputed values, and proceed with a CCA. On the other hand, when the analyst is relatively certain that the assumptions of the documented imputation model are met, MII may lead to considerable gains in efficiency.

Class 4

Missing values are in $\mathbf{X}$ and the MDM is dependent on $\mathbf{X}_{\text{mis}}$; the MDM is not ignorable for imputation. However, since $\boldsymbol{s}$ is (conditionally) independent of $\boldsymbol{y}$, the MDM is ignorable for the CCA, which is often overlooked.

Class 5

Missing values are in $\mathbf{X}$ and the MDM is dependent on $\boldsymbol{y}$, which implies that the CCA is invalid. However, since $\boldsymbol{s}$ is (conditionally) independent of $\mathbf{X}_{\text{mis}}$, the MDM is ignorable for imputation; moreover, MI tends to produce "better" estimates than CCA for this class of MDMs, even if the IDGP is slightly misspecified. Therefore, we recommend the use of MI in this case, and will focus on this class of MDMs in this work.

2.3.5 Discussion of assumptions

It should now be clear that given ADGP (2.1), MI without the availability of relevant auxiliary predictors and without specifying a MDM seems only purposeful when $\mathbf{X}$ is afflicted by missing data, which corresponds to Class 5 and Class 3. Imputation does not improve precision if missing values are confined to the outcome, and if there are no relevant auxiliary predictors of y available.

Generating imputations for $\mathbf{X}_{\text{mis}}$ while ignoring the MDM requires the assumption that the missing data in $\mathbf{X}$ are MAR or MCAR. Unfortunately, by its very definition, the validity of the MAR assumption cannot be tested without information external to the data set. The MDM may in fact be MNAR, in which case an Imputation Method (IM) which ignores the MDM will lead to invalid MII. Schafer (1997) argues that including additional predictors $\mathbf{Z}$ in the imputation model for $\mathbf{X}_{\text{mis}}$ can make the MAR assumption more plausible when these variables explain when $\mathbf{X}_{\text{mis}}$ is missing. Moreover, credibility in the MII can be increased by performing a sensitivity analysis; this requires that the imputer produces multiple imputations under a series of hypothesized MDMs, each resulting in a MII. Analysts typically lack the knowledge and skills to interpret the results of such an analysis.

When the selection indicator is independent of $\boldsymbol{y}$, the CCA is consistent and is the easiest way out. Unfortunately, there is no silver bullet; neither CCA nor MII is valid for all identified classes of MDMs, and the choice between the two depends on often unverifiable assumptions about the MDM. Because an ignorable MDM is specified by omitting it from the DGP, care must be taken to make this implicit assumption explicit by means of documentation.

It is important to realize that although MI seeks to separate the missing data problem from the DAP, the responsibility for the MII ultimately lays with the analyst. Because the IMDSs give little clue about the IDGP assumed by the imputer, it is paramount that the analyst procures the documentation describing the IDGP, and assesses if the IDGP is compatible with the IDA; this is also necessary when the IMDSs are generated by the analyst himself using a computer program.

Chapter 3

The Imputer

After defining the MI estimator, the compatibility of an import class of IMs will be investigated with respect to IDA (2.17). In particular, it is investigated how to properly impute transformations and interactions of predictor variables. Finally, strategies for imputing missings in multiple variables will be compared and contrasted in 3.4.

3.1 Pooling

The core of MI theory (Rubin, 1987) consists of the rules necessary to combine the IDAs into a MII. More formally, these rules define an estimator $\hat{\boldsymbol{\beta}}_{\text{MI}}$ and associated variance estimator $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\text{MI}})$, which are functions of the IDAs, and together constitute the MII. Although compatibility of the IDGP with the IDA is defined in Section 3.2 from a frequentist perspective, the MI estimators are justified from a Bayesian perspective, where the Bayesian random counterpart to the parameter $\boldsymbol{\beta}$ is denoted by $\dot{\boldsymbol{\beta}}$. The posterior distribution given only the observed data can be found by averaging the imputed data posterior $f(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R})$ over the multiple imputations as drawn from the posterior predictive distribution of the missing values $f(\mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R})$:

$$\begin{aligned} f(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}) = \iint & f(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R}) \\ & \times f(\mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}) d\mathbf{X}_{\text{mis}} d\boldsymbol{y}_{\text{mis}}. \quad (3.1) \end{aligned}$$

By centering $f(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R})$ at the estimator of $\boldsymbol{\beta}$ in the hypothetical case that all data are observed (which is the estimator used in the DAP):

$$\begin{aligned} \mathrm{E}\left(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R}\right) &= \hat{\boldsymbol{\beta}}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1}) \\ \mathrm{Var}(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R}) &= \hat{V}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1}), \end{aligned} \tag{3.2}$$

the posterior mean marginalized over the missing data equals

$$\begin{aligned} \mathrm{E}\left(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right) &= \mathrm{E}\left(\mathrm{E}\left(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R}\right)|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right) \\ &= \mathrm{E}\left(\hat{\boldsymbol{\beta}}_{\text{OLS}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right), \end{aligned} \tag{3.3}$$

where $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is shorthand for $\hat{\boldsymbol{\beta}}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1})$. The marginal posterior variance equals

$$\begin{aligned} \mathrm{Var}(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}) &= \mathrm{E}\left(\mathrm{Var}(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R})|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right) \\ &\quad + \mathrm{Var}(\mathrm{E}\left(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \mathbf{R}\right)|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}) \\ &= \mathrm{E}\left(\hat{V}_{\text{OLS}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right) + \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}). \end{aligned} \tag{3.4}$$

Equation (3.3) and (3.4) suggest a simple numerical procedure for simulating the posterior mean and variance of $\dot{\boldsymbol{\beta}}$, which was already described in (1.2). First, m independent imputations are drawn from $f(\mathbf{X}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R})$, after which m IMDSs $\{\tilde{\mathbf{X}}^i, \tilde{\boldsymbol{y}}^i\}_{i=1}^m$ are assembled. From each IMDS an IDA is calculated, resulting in m simulations $\{\hat{\boldsymbol{\beta}}^i_{\text{IDA}}, \hat{V}^i_{\text{IDA}}\}_{i=1}^m$ of certain features of the posterior distribution of $\dot{\boldsymbol{\beta}}$ such that for an infinite number of independent imputations, the average of the m IDAs

$$\hat{\boldsymbol{\beta}}_{\text{MI}} = m^{-1}\sum_{i=1}^{m} \hat{\boldsymbol{\beta}}_{\text{IDA}}(\tilde{\mathbf{X}}^i, \tilde{\boldsymbol{y}}^i). \tag{3.5}$$

is a consistent estimator of the (marginalized) posterior mean of $\dot{\boldsymbol{\beta}}$:

$$\operatorname*{plim}_{m\to\infty} m^{-1}\sum_{i=1}^{m} \hat{\boldsymbol{\beta}}_{\text{IDA}}(\tilde{\mathbf{X}}^i, \tilde{\boldsymbol{y}}) = \mathrm{E}\left(\hat{\boldsymbol{\beta}}_{\text{OLS}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right) = \mathrm{E}\left(\dot{\boldsymbol{\beta}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right). \tag{3.6}$$

Further,

$$\boldsymbol{\Omega}_{W,\infty} = \operatorname*{plim}_{m\to\infty} m^{-1}\sum_{i=1}^{m} \hat{V}^i_{\text{IDA}} = \mathrm{E}\left(\hat{V}_{\text{OLS}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R}\right)$$

$$\boldsymbol{\Omega}_{B,\infty} = \operatorname*{plim}_{m\to\infty} m^{-1}\sum_{i=1}^{m} (\hat{\boldsymbol{\beta}}^i_{\text{IDA}} - \hat{\boldsymbol{\beta}}_{\text{MI}})(\hat{\boldsymbol{\beta}}^i_{\text{IDA}} - \hat{\boldsymbol{\beta}}_{\text{MI}})^{\mathrm{T}} = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}|\mathbf{X}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \mathbf{R})$$

are the within-imputation variance and between-imputation variance, such that the posterior variance of $\dot{\boldsymbol{\beta}}$ equals

$$\mathrm{Var}(\boldsymbol{\beta}|\mathbf{X}_{\mathrm{obs}}, \boldsymbol{y}_{\mathrm{obs}}, \mathbf{R}) = \boldsymbol{\Omega}_{W,\infty} + \boldsymbol{\Omega}_{B,\infty}.$$

Rubin (1987) motivates that for a finite number of imputations, the posterior variance can be estimated as the sum of the estimated within-imputation variance and estimated between-imputation variance:

$$\begin{aligned}\hat{V}_{\mathrm{MI}} &= \hat{\boldsymbol{\Omega}}_{W,m} + (1 + \frac{1}{m})\hat{\boldsymbol{\Omega}}_{B,m} \qquad (3.7)\\ &= m^{-1}\sum_{i=1}^{m} \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{i}_{\mathrm{IDA}}) + (1 + \frac{1}{m})(m-1)^{-1}\sum_{i=1}^{m}(\hat{\boldsymbol{\beta}}^{i}_{\mathrm{IDA}} - \hat{\boldsymbol{\beta}}_{\mathrm{MI}})(\hat{\boldsymbol{\beta}}^{i}_{\mathrm{IDA}} - \hat{\boldsymbol{\beta}}_{\mathrm{MI}})^{\mathrm{T}},\end{aligned}$$

where the factor $1 + \frac{1}{m}$ is a correction for the finite number of imputations.

Robins & Wang (2000) propose an alternative estimator which is more efficient, but also more complicated to compute; in applied work, (3.7) is predominantly used. Therefore, only $\hat{V}_{\mathrm{MI}}$ is considered in this work.

3.2 Compatibility

Now that the MII $\{\hat{\boldsymbol{\beta}}_{\mathrm{MI}}, \hat{V}_{\mathrm{MI}}\}$ is defined, it will be investigated which requirements must be fulfilled for an IDGP to be compatible with the IDA defined in (2.17). More formally, an IDGP is compatible when:

1. The MI estimator $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$ is a consistent estimator for $\boldsymbol{\beta}$ as $n \to \infty$, and
2. The MI variance estimator $\hat{V}_{\mathrm{MI}}$ is a consistent estimator for $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{MI}})$.

Note that in contrast with (3.6), which given an asymptotic Bayesian justification for using $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$ as an estimator of the DAP estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ marginalized over the missing data as $m \to \infty$ and with n fixed, the compatibility conditions stated above concern the asymptotic frequentist properties of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$ as an estimator of the true parameter $\boldsymbol{\beta}$ as $n \to \infty$ and with m fixed.

For analytical tractability, compatibility will only be investigated in the case of a single predictor affected by missing values, which is denoted by $\mathbf{X}_{(1)}$ (the first column of $\mathbf{X}$); the completely observed predictors are denoted by $\mathbf{X}_{(-1)}$ (all columns of $\mathbf{X}$ except the first one). Further, it is assumed that the MDM is ignorable for imputation, which is satisfied when the missing values in $\mathbf{X}_{(1)}$ are MAR such that

$$f(\boldsymbol{s}|\mathbf{X}_{(1)}, \mathbf{W}) = \prod_{i=1}^{N} f(\boldsymbol{s}_i|\mathbf{W}_i), \qquad (3.8)$$

where $\mathbf{W} = \begin{bmatrix} \boldsymbol{y} & \mathbf{X}_{(-1)} \end{bmatrix}$; in Section 3.3.1 it is shown that this MDM is ignorable for imputation.

Naturally, consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$ requires consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{IDA}}$. It is useful to partition (2.17) in terms of cases which had a missing value in $\mathbf{X}_{(1)}$, and cases which were already fully observed:

$$\hat{\boldsymbol{\beta}}_{\mathrm{IDA}} = \left(\mathbf{X}_{\mathrm{obs}}^{\mathrm{T}}\mathbf{X}_{\mathrm{obs}} + \tilde{\mathbf{X}}_{\mathrm{mis}}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathrm{mis}}\right)^{-1}\left(\mathbf{X}_{\mathrm{obs}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{obs}} + \tilde{\mathbf{X}}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{mis}}\right), \tag{3.9}$$

where $\tilde{\mathbf{X}}_{\mathrm{mis}} = \begin{bmatrix} \tilde{\mathbf{X}}_{\mathrm{mis},1} & \mathbf{X}_{\mathrm{mis},-1} \end{bmatrix}$, and the index denoting the imputation number is suppressed. Note that $\tilde{\mathbf{X}}$ and $\tilde{\boldsymbol{y}}$ are demeaned as in (2.8) after imputation, not before. In the hypothetical case that the missing data are in fact observed, $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ can be written similar to (3.9):

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\mathbf{X}_{\mathrm{obs}}^{\mathrm{T}}\mathbf{X}_{\mathrm{obs}} + \mathbf{X}_{\mathrm{mis}}^{\mathrm{T}}\mathbf{X}_{\mathrm{mis}}\right)^{-1}\left(\mathbf{X}_{\mathrm{obs}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{obs}} + \mathbf{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{mis}}\right). \tag{3.10}$$

When the IM produces imputations such that

$$\begin{aligned} \operatorname{plim} N^{-1}\tilde{\mathbf{X}}_{\mathrm{mis}}^{\mathrm{T}}\tilde{\mathbf{X}}_{\mathrm{mis}} &= \operatorname{plim} N^{-1}\mathbf{X}_{\mathrm{mis}}^{\mathrm{T}}\mathbf{X}_{\mathrm{mis}} \\ \operatorname{plim} N^{-1}\tilde{\mathbf{X}}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{mis}} &= \operatorname{plim} N^{-1}\mathbf{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{y}_{\mathrm{mis}} \end{aligned} \tag{3.11}$$

it follows from (3.9) and (3.10) that

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_{\mathrm{IDA}}(\tilde{\mathbf{X}}, \boldsymbol{y}) = \operatorname{plim} \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1}) = \boldsymbol{\beta}. \tag{3.12}$$

Note that to demean $\tilde{\mathbf{X}}_{(1)}$ correctly, it is also necessary that

$$\operatorname{plim} N^{-1}\tilde{\mathbf{X}}_{\mathrm{mis}} = \operatorname{plim} N^{-1}\mathbf{X}_{\mathrm{mis}}.$$

From (3.11) it becomes clear that a sufficient condition for consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$ is that the first two asymptotic sample moments of the imputed variable match the first two asymptotic sample moments of the variable with missings, and the asymptotic sample covariance between the imputed variable and the other variables $\mathbf{X}_{(-1)}$ and $\boldsymbol{y}$ should match the asymptotic sample covariance between the variable with missings and the other variables. Thus, it is not necessary that the distribution of the imputed variable conditional on the other variables fully matches the corresponding conditional distribution of the variable with missings. This also implies that with respect to consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$, there are no statistical objections to imputing “unrealistic” values, a point which will be discussed in more detail in Section 3.3.4. Further, all variables referenced in the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ should be included as predictors in the IDGP. Although it might be believed that incorporating $\boldsymbol{y}$ in the imputation model will artificially strengthen the relationship between $\mathbf{X}_{(1)}$ and $\boldsymbol{y}$, (3.11) shows that actually the converse is true: omitting $\boldsymbol{y}$ will attenuate the linear relationship between $\mathbf{X}_{(1)}$ and $\boldsymbol{y}$.

3.3 Imputation by reverse regression

Since (2.1) is a regression model, a natural idea is to impute $\mathbf{X}_{\text{mis},1}$ using a "reverse regression" IDGP where $\mathbf{X}_{(1)}$ and $\boldsymbol{y}$ swap roles. Indeed, the majority of IMs under consideration are special cases of this model, which will be described in detail in Section 3.3.1. Further, it will be attempted to verify if imputations produced by this model satisfy conditions (3.11), which is a necessary condition for compatibility as defined in Section 3.2. In 3.3.4, it will be investigated how to impute transformations of predictor variables. Finally, assumptions necessary for the application of IMs based on this model will be discussed in Section 3.3.5.

3.3.1 The IDGP

The variables with missing values follow the following population regression function:

$$\boldsymbol{x}_1 = \mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right) + \sigma(\boldsymbol{w})v, \tag{3.13}$$

where (3.13) pertains to a single observation, $\mathrm{E}\left(v|\boldsymbol{w}\right) = 0$, and $\sigma^2(\boldsymbol{w})$ is the scedastic function. In contrast with the error term in (2.1), the variance of $\boldsymbol{x}_1$ is not assumed to be homoscedastic for reasons which will be made clear in Section 3.3.3. A special case of (3.13) is the reverse linear regression imputation model with a linear conditional expectation and constant scedastic function.

The conditional expectation $\mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right)$ and conditional variance $\mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{w})$ are estimated using the observed data $\{\mathbf{X}_{\text{obs},1}, \mathbf{W}_{\text{obs}}\}$, which is valid when (3.8) holds such that $v \perp \boldsymbol{s}|\boldsymbol{w}$. Let $\hat{\mu}(\cdot)$ and $\hat{\sigma}^2(\cdot)$ be estimates of $\mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right)$ and $\sigma^2(\cdot)$, respectively. Imputations for $\boldsymbol{x}_1$ are then generated as

$$\tilde{\boldsymbol{x}}_1 = \hat{\mu}(\boldsymbol{w}) + \hat{\sigma}(\boldsymbol{w})\tilde{v}, \tag{3.14}$$

where $\tilde{v}$ follows some distribution with $\mathrm{E}\left(\tilde{\nu}\right) = 0$ and $\mathrm{Var}(\tilde{\nu}) = 1$. Unfortunately, the ADGP as defined in Section 2.1 does not contain assumptions about the conditional distributions of $\boldsymbol{x}_1$; out of convenience, $\tilde{v}$ is often taken to be distributed standard normal. Algorithm 3.1 gives a condensed description of the necessary steps required for generating imputation according to (3.13).

Algorithm 3.1 Imputation by reverse regression

```
1. Estimate the parameters of $\hat{\mu}(\cdot)$ and $\hat{\sigma}^2(\cdot)$ using the observed cases:
   $\hat{\boldsymbol{\eta}}_\mu = \hat{\boldsymbol{\eta}}_\mu(\mathbf{X}_{1,\text{obs}}, \mathbf{W}_{\text{obs}})$ and $\hat{\boldsymbol{\eta}}_{\sigma^2} = \hat{\boldsymbol{\eta}}_{\sigma^2}(\mathbf{X}_{1,\text{obs}}, \mathbf{W}_{\text{obs}})$
2. Simulate $\left(\dot{\boldsymbol{\eta}}_\mu, \dot{\boldsymbol{\eta}}_{\sigma^2}\right)$ from the (approximated) posterior distribution of $\left(\hat{\boldsymbol{\eta}}_\mu, \hat{\boldsymbol{\eta}}_{\sigma^2}\right)$
3. Predict the conditional mean and conditional variance for the units with missing values:
   $\hat{\boldsymbol{m}} = \hat{\boldsymbol{\mu}}(\mathbf{W}_{\text{mis}}|\dot{\boldsymbol{\eta}}_\mu)$ and $\hat{\boldsymbol{s}} = \hat{\boldsymbol{\sigma}}^2(\mathbf{W}_{\text{mis}}|\dot{\boldsymbol{\eta}}_{\sigma^2})$
4. Draw $n_{\text{mis}}$ imputations $\tilde{\mathbf{X}}_{\text{mis},1} \sim \mathcal{N}(\hat{\boldsymbol{m}}, \operatorname{diag}\hat{\boldsymbol{s}})$
5. Repeat steps 2 through 4 $m$ independent times, where $m$ is the number of desired imputations
```

3.3.2 Compatibility check

It will now be verified if imputations generated according to model (3.13) satisfy condition (3.11). Note that the imputations $\tilde{\mathbf{X}}_{(1)}$ as generated from (3.14) are not independently distributed because they are dependent on the observed cases through the estimates $\hat{\mu}(\cdot)$ and $\hat{\sigma}^2(\cdot)$. However, given the observed cases and selection indicator $\boldsymbol{s}$, the imputations are i.i.d, which allows for analysis of the conditional asymptotic behavior of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{IDA}} - \boldsymbol{\beta})$. Let $b := x_1$ and $\tilde{b} := \tilde{x}_1$. Because the imputations depend on the estimators $\hat{\mu}(\cdot)$ and $\hat{\sigma}^2(\cdot)$ which change with sample size, the imputed values are entries in the following triangular array as $N \to \infty$:

$$\begin{array}{lll}
\tilde{b}_{1,1} = \hat{\mu}_1(\mathbf{W}_1) + \hat{\sigma}_1(\mathbf{W}_1)\tilde{\boldsymbol{v}}_1 & & \\
\tilde{b}_{2,1} = \hat{\mu}_2(\mathbf{W}_1) + \hat{\sigma}_2(\mathbf{W}_1)\tilde{\boldsymbol{v}}_1 & \tilde{b}_{2,2} & \\
\qquad\cdots & \cdots & \qquad\cdots \\
\tilde{b}_{N,1} = \hat{\mu}_N(\mathbf{W}_1) + \hat{\sigma}_N(\mathbf{W}_1)\tilde{\boldsymbol{v}}_1 & \cdots & \tilde{b}_{N,N} = \hat{\mu}_N(\mathbf{W}_N) + \hat{\sigma}_N(\mathbf{W}_N)\tilde{\boldsymbol{v}}_N
\end{array},$$

where the entries are row-wise i.i.d. conditional on $OD = \{\mathbf{X}_{1,\text{obs}}, \mathbf{W}_{\text{obs}}, \boldsymbol{s}\}$. Further, the following assumptions are made:

Assumption 1 The estimators $\hat{\mu}(\cdot)$ and $\hat{\sigma}(\cdot)$ are consistent, such that for all $i \in \text{mis}$

$$\operatorname{plim} \tilde{b}_{N,i} = \mathrm{E}\left(\boldsymbol{x}_1|\mathbf{W}_i\right) + \sigma(\mathbf{W}_i)\tilde{\boldsymbol{v}}_i. \tag{3.15}$$

Assumption 2 $\tilde{b}$ is bounded.

Assumption 3 $\frac{n_{\text{obs}}}{n_{\text{mis}}} = \lambda$ as $n \to \infty$, where $0 < \lambda < \infty$.

Using the first two assumptions, the dominated convergence theorem gives:

$$\begin{aligned} \lim \mathrm{E}\left(\tilde{b}^2_{N,i}\right) &= \mathrm{E}\left(\operatorname{plim} \tilde{b}^2_{N,i}\right) \\ &= \mathrm{E}\left([\mathrm{E}\left(\boldsymbol{x}_1|\mathbf{W}_i\right) + \sigma(\mathbf{W}_i)\tilde{\boldsymbol{v}}_i]^2\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(\boldsymbol{x}_1|\mathbf{W}_i\right)\right)^2 + \mathrm{Var}(\mathrm{E}\left(\boldsymbol{x}_1|\mathbf{W}_i\right)) + \mathrm{E}\left(\sigma^2(\mathbf{W}_i)\right) \\ &= \mathrm{E}\left(b_i^2\right) \end{aligned} \tag{3.16}$$

for all $i \in$ mis. Further,

$$\begin{aligned} \lim \mathrm{E}\left(\tilde{b}_{N,i}\mathbf{W}_i\right) &= \mathrm{E}\left(\operatorname{plim} \tilde{b}_{N,i}\mathbf{W}_i\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(\boldsymbol{x}_1|\mathbf{W}_i\right)\mathbf{W}_i\right) \\ &= \mathrm{E}\left(b_i\mathbf{W}_i\right) \end{aligned} \tag{3.17}$$

for all $i \in$ mis. Applying a weak law of large numbers for triangular arrays using (3.17) and (3.16) gives

$$\begin{aligned} \operatorname{plim} N^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathrm{mis}}\tilde{\mathbf{X}}_{\mathrm{mis}} &= (1+\lambda)^{-1}\mathrm{E}\left(\mathbf{X}^{\mathrm{T}}_{\mathrm{mis}}\mathbf{X}_{\mathrm{mis}}\right) \\ \operatorname{plim} N^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathrm{mis}}\boldsymbol{y}_{\mathrm{mis}} &= (1+\lambda)^{-1}\mathrm{E}\left(\mathbf{X}^{\mathrm{T}}_{\mathrm{mis}}\boldsymbol{y}_{\mathrm{mis}}\right), \end{aligned} \tag{3.18}$$

which shows that imputations generated according to model (3.3.1) lead to consistent estimation of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$ conditional on OD.

The conditional asymptotic behavior is only of secondary importance, and thus a device is needed to infer the unconditional statistical properties of the MII from the derived conditional asymptotic behavior. An often employed technique in the literature (see for example Aerts et al. (2002)) is to derive conditional on OD the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\mathrm{IDA}}$ centered around a consistent estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OD}} = h(OD)$, which is solely based on the observed cases, such that

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{IDA}} - \hat{\boldsymbol{\beta}}_{\mathrm{OD}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_1) \qquad \text{given } OD \text{ for almost every } OD. \tag{3.19}$$

When unconditionally

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{OD} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_2),$$

lemma 1 of either Schenker & Welsh (1988) or Nielsen (2003) states that

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{IDA}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_1 + \Sigma_2), \tag{3.20}$$

unconditionally. There are several possible choices for the “centering estimator”. Letting $\hat{\boldsymbol{\beta}}_{\mathrm{OD}} = \hat{\boldsymbol{\beta}}_{\mathrm{CCA}}$ restricts the compatibility of (3.13) to those situations in Table 2.1 on page 19 where the complete case analysis is consistent and the sole motivation for MI is increased efficiency. However, under the posited MDM (3.8), $\hat{\boldsymbol{\beta}}_{\text{CCA}}$ is inconsistent, and the only consistent OLS estimator is $\hat{\boldsymbol{\beta}}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1})$, which is based in part on unobserved missing cases. Unfortunately, (3.20) does not hold when $\hat{\boldsymbol{\beta}}_{\text{OD}} = \hat{\boldsymbol{\beta}}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1})$, because conditioning on OD does not reduce $\hat{\boldsymbol{\beta}}_{\text{OLS}}(\mathbf{X}, \boldsymbol{y}, \mathbf{1})$ to a non-random constant.

The correct application of (3.20) requires that the centering estimator reduces to a constant when conditioning on the observed data. If the probabilities of observing a unit are known, $\boldsymbol{\beta}$ can be consistently estimated using an Inverse Probability Weighting estimator based on OD; such an estimator can then be successfully used as centering estimator for $\hat{\boldsymbol{\beta}}_{\text{IDA}}$. When the response probabilities are unknown but (3.8) holds, the corresponding sampling weights can be obtained by estimating a specified model $p(s|\boldsymbol{w})$; analogue to (3.19), $\hat{\boldsymbol{\beta}}_{\text{IDA}}$ can then be centered around the Inverse Probability Weighting estimator using the estimated sampling weights while conditioning on $\{\mathbf{W}, \mathbf{X}_{\text{obs},1}\}$, after which (3.20) can be applied. However, this implies that data analysts cannot consistently estimate the variance of $\hat{\boldsymbol{\beta}}_{\text{IDA}}$ by (2.17), which is the estimator commonly used in the Data Analysis Procedure when there are no missing data; instead, estimating the variance involves estimation of the model $p(s|\boldsymbol{w})$ by for example logistic regression. Since the solution described above does not allow for the use of IDA (2.17), and since no alternative solutions could be found, the analysis ends here.

3.3.3 Reverse Linear Regression

When the observations are independent and distributed multivariate normal then $\operatorname{E}(\boldsymbol{x}_1|\boldsymbol{w})$ is linear and $\operatorname{Var}(\boldsymbol{x}_1|\boldsymbol{w})$ is homoscedastic such that

$$\begin{aligned} \operatorname{E}(\boldsymbol{x}_1|\boldsymbol{w}) &= \tilde{\alpha} + \boldsymbol{w}\tilde{\boldsymbol{\beta}} \\ \operatorname{Var}(\boldsymbol{x}_1|\boldsymbol{w}) &= \tilde{\sigma}^2. \end{aligned} \tag{3.21}$$

The linearity and homoscedasticity of both the direct and reverse regression only hold when $\boldsymbol{x}_1$ and y are distributed bivariate normal given $\boldsymbol{x}_{-1}$ (Spanos, 1995). We illustrate the consequences of non-linearity with an example where $k = 1$. Further, we take $\boldsymbol{x} \sim \mathcal{SN}(\alpha)$, where $\mathcal{SN}(\alpha)$ is the Skew-Normal distribution (Azzalini, 2005) with skewness parameter α:

$$f(\boldsymbol{x}; \alpha) = 2\phi(x)\Phi(\alpha x), \qquad x \in \mathbb{R} \tag{3.22}$$

and y follows (2.1). See Figure 5.1 on page 55 for a plot of the Skew-Normal density. Note that when $\alpha = 0$, (3.22) reduces to the standard normal density. If

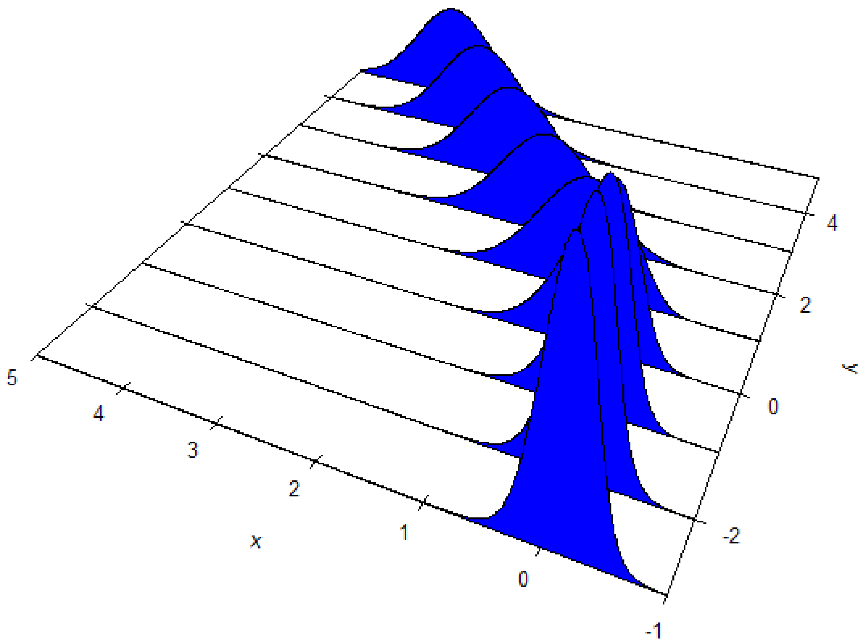

Figure 3.1: Conditional density plot of $x|y$ with $\alpha = 5$, $\sigma^2 = \mathrm{Var}(\boldsymbol{x}) = 1 - \frac{2\alpha^2}{\pi(\alpha^2+1)}$, such that the coefficient of determination $R^2 = \frac{\mathrm{Var}(\boldsymbol{x})}{\mathrm{Var}(\boldsymbol{x})+\sigma^2} = 0.5$.

we take $u \sim \mathcal{N}(0, \sigma^2)$, the conditional distribution Figure 3.1 on page 30 of x given y is depicted in (3.1); the non-linearity of the expectation violates the linearity assumption of (3.21), and may lead to invalid MII.

In the special case that the missing data are MCAR such that $f(\boldsymbol{x}, y|s = 1) = f(\boldsymbol{x}, y|s = 0)$, the fully observed cases $\{\mathbf{X}_{\mathrm{obs}}, \boldsymbol{y}_{\mathrm{obs}}\}$ can be considered a simple random sub-sample from $\{\mathbf{X}, \boldsymbol{y}\}$. Therefore, instead of condition (3.11),

$$\begin{aligned} \operatorname{plim} n_{\mathrm{mis}}^{-1} \tilde{\mathbf{X}}_{\mathrm{mis}}^{\mathrm{T}} \tilde{\mathbf{X}}_{\mathrm{mis}} &= \operatorname{plim} n_{\mathrm{obs}}^{-1} \mathbf{X}_{\mathrm{obs}}^{\mathrm{T}} \mathbf{X}_{\mathrm{obs}} \\ \operatorname{plim} n_{\mathrm{mis}}^{-1} \tilde{\mathbf{X}}_{\mathrm{mis}}^{\mathrm{T}} \boldsymbol{y}_{\mathrm{mis}} &= \operatorname{plim} n_{\mathrm{obs}}^{-1} \mathbf{X}_{\mathrm{obs}}^{\mathrm{T}} \boldsymbol{y}_{\mathrm{obs}} \end{aligned} \tag{3.23}$$

is sufficient for consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$. This suggests that the linear imputation model or any method which generates imputations according to a consistent estimate of the covariance matrix of the observed data is valid if the missing data are MCAR, even if the reverse regression $\mathrm{E}\left(\mathbf{X}|\mathbf{W}\right)$ is non-linear; this will be demonstrated empirically in 5.4.3. However, if the MDM equals (3.8), then (3.23) is insufficient, and condition (3.11) must hold instead.

3.3.4 Transformations

As mentioned in 2.1.2, one can model non-linear relationships between variables using the LM by transforming either or both the response and predictor variables. For example, in economics, often the log of income is modeled. It is also possible that both the original and transformed predictor variable enter model (2.1). When $\mathbf{X}_1$ is afflicted by missing values and simultaneously transformed, there are two alternative imputation strategies: first impute and then transform, and first transform and then impute (von Hippel, 2009). First imputing the original variable and then transforming the original using the imputed original variable seems an attractive strategy, as it preserves the relationship between the original variable and transformed variable; there is no chance for any inconsistencies. Analogue to (3.17), $\lim \mathrm{E}\left(g(\tilde{b}_{N,i})\mathbf{W}_i\right)$ should equal $\mathrm{E}\left(g(b_{N,i})\mathbf{W}_i\right)$, where $g(\cdot)$ is the transformation function. However, when first imputing and then transforming,

$$\begin{aligned}
\lim \mathrm{E}\left(g(\tilde{b}_{N,i})^{\mathrm{T}}\mathbf{W}_i\right) &= \mathrm{E}\left(\operatorname{plim} g(\tilde{b}_{N,i})^{\mathrm{T}}\mathbf{W}_i\right) \\
&= \mathrm{E}\left(g\left(\mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right) + \sigma(\mathbf{W}_i)\tilde{\boldsymbol{\nu}}_i\right)^{\mathrm{T}}\mathbf{W}_i\right) \\
&\neq \mathrm{E}\left(\mathrm{E}\left(g(\boldsymbol{x}_1)|\boldsymbol{w}\right)\mathbf{W}_i\right) \qquad \text{in general} \\
&= \mathrm{E}\left(g(b_{N,i})\mathbf{W}_i\right);
\end{aligned} \tag{3.24}$$

thus, the first impute and then transform strategy is fundamentally flawed.

The first impute and then transform technique is often called passive imputation because the values of the transformed variable are not actively imputed, but instead calculated from the imputed original variable. Passive imputation is implemented in several software packages because some imputers desire to assess the quality of imputations using a plausibility criterion (Kuchler & Spiess, 2009); a plausibility criterion defines the set of imputed values which are deemed realistic, and prohibits inconsistencies between the original variable and transformed variable. The statistically relevant criterion, however, is the consistency of $\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$, and the requirements for consistency of the MI inference are postulated in (3.11); any other criterion is irrelevant to the validity of the inference. Thus, transformed variables afflicted by missing values should be treated as any other variable with missing values, and should be imputed separately from the original variables, at least when using a reverse regression imputation model. In the case of (3.13), this involves estimating $\mathrm{E}\left(g(\boldsymbol{x}_1)|\boldsymbol{w}\right)$ and $\mathrm{Var}(g(\boldsymbol{x}_1)|\boldsymbol{w})$. These remarks also apply to transformations involving multiple predictors, for example interaction effects.

3.3.5 Discussion of assumptions

Imputation by reverse regression defines a functional relationship between $\boldsymbol{x}_1$ and $\boldsymbol{w}$ via the conditional expectation operator. However, this might fail to capture the essential relation between predictor and response. For example, in the following

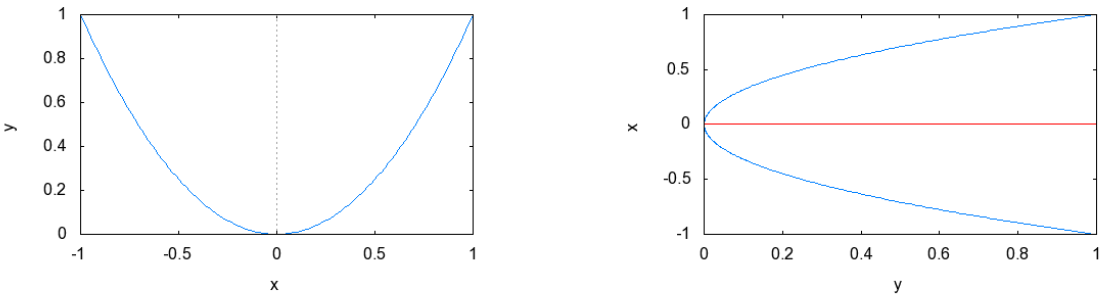

Figure 3.2: Plot of the direct and reverse regression of $y = x^2$

population regression function the predictor $\boldsymbol{x}_1$ enters the model non-linearly:

$$y = \alpha + g(\boldsymbol{x}_1)\beta + \boldsymbol{x}_{-1}^{\mathrm{T}}\gamma + u,$$

where $g(\cdot)$ is not one-to-one such as $g(\boldsymbol{x}) = \boldsymbol{x}^2$ with $\boldsymbol{x}_1 \in \mathbb{R}$. Then, the reverse relation (3.13) of $\boldsymbol{x}_1$ given $\boldsymbol{w}$ is multivalued, in which case the conditional expectation might blur important features of the data. Figure 3.2 on page 32 illustrates the case $g(\boldsymbol{x}) = \boldsymbol{x}^2$, where imputations of $\boldsymbol{x}_1$ generated according to (3.13) will be concentrated around the conditional expectation, which is depicted by the red horizontal line. Without knowledge of which of the two branches in Figure 3.2 on page 32 the observations with missing values belong, imputation will attenuate the relationship between $g(\boldsymbol{x}_1)$ and y; this seems to be a fundamental problem of the reverse regression approach.

A possible solution to the above problem, and one that incidentally also allows for the impute-transform strategy, is to use a IMs which estimate the full conditional distribution of $\boldsymbol{x}_1$, and not just the conditional mean and variance. If the IM produces imputations which weakly converge to the variable with missing values such that $(\tilde{\mathbf{X}}_{1,\text{mis}}|\mathbf{W}_{\text{mis}}) \xrightarrow{\mathcal{D}} (\mathbf{X}_{1,\text{mis}}|\mathbf{W}_{\text{mis}})$, and $g : \mathbb{R} \rightarrow \mathbb{R}$ is continuous, then $(g(\tilde{\mathbf{X}}_{1,\text{mis}})|\mathbf{W}_{\text{mis}}) \xrightarrow{\mathcal{D}} (g(\mathbf{X}_{1,\text{mis}})|\mathbf{W}_{\text{mis}})$ also.

3.4 Multivariate missing data

Imputation by reverse regression as discussed in Section 3.3 is a class of IM capable of imputing missing values in a single predictor variable; in practice, multiple predictor variables are afflicted by missing values[1]. There are two main approaches available for dealing with missings in multiple variables: Joint Modeling (JM) and Fully Conditional Specification (FCS).

[1] As discussed in Section 2.3.4, discarding observations with a missing value in the response variable is a viable strategy.

3.4.1 Joint Modeling

JM involves specifying a joint distribution for all variables in the data set, including the response indicator:

$$g(\mathbf{X}, \boldsymbol{y}, \mathbf{R}; \boldsymbol{\eta}, \boldsymbol{\phi}) = f(\mathbf{X}, \boldsymbol{y}; \boldsymbol{\eta}) p(\mathbf{R}|\mathbf{X}, \boldsymbol{y}; \boldsymbol{\phi}), \tag{3.25}$$

where $\boldsymbol{\eta}$ are the parameters of the imputation model, and $\boldsymbol{\phi}$ are the parameters of the posited missing data mechanism. Since the focus of this work is on ignorable missing data mechanisms, (3.25) reduces to $f(\mathbf{X}, \boldsymbol{y}; \boldsymbol{\eta})$. Provided that the DAP and IDA employ maximum likelihood estimators derived from the specified joint distribution (3.25), Nielsen (2003) proofs consistency of the MII and shows that the associated variance estimator (3.7) is weakly unbiased. Schafer (1997) and Little & Rubin (2002) developed IMs based on the multivariate normal or general location model which are capable of imputing missing values in multiple variables for general missing data patterns using the Expectation Maximization algorithm or Gibbs sampler as described in Algorithm 3.2. The presentation used in Algorithm 3.2 is taken from Liu et al. (2012) to illuminate the relation to an alternative approach discussed next.

Algorithm 3.2 Gibbs sampler JM

```
Step 1     (a) Draw $\boldsymbol{\eta}$ from $f(\boldsymbol{\eta}|\boldsymbol{y}, \mathbf{X}_{1,\text{obs}}, \mathbf{X}_{-1})$
           (b) Draw $\tilde{\mathbf{X}}_{1,\text{mis}}$ from $g(\mathbf{X}_{1,\text{mis}}|\boldsymbol{y}, \mathbf{X}_{1,\text{obs}}, \mathbf{X}_{-1}, \boldsymbol{\eta})$
               ⋮
Step k     (a) Draw $\boldsymbol{\eta}$ from $f(\boldsymbol{\eta}|\boldsymbol{y}, \mathbf{X}_{k,\text{obs}}, \mathbf{X}_{-k})$
           (b) Draw $\tilde{\mathbf{X}}_{k,\text{mis}}$ from $g(\mathbf{X}_{k,\text{mis}}|\boldsymbol{y}, \mathbf{X}_{k,\text{obs}}, \mathbf{X}_{-k}, \boldsymbol{\eta})$
Step k + 1 (a) Draw $\boldsymbol{\eta}$ from $f(\boldsymbol{\eta}|\boldsymbol{y}_{\text{obs}}, \mathbf{X})$
           (b) Draw $\tilde{\boldsymbol{y}}_{\text{mis}}$ from $g(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \mathbf{X}, \boldsymbol{\eta})$
```

3.4.2 Fully Conditional Specification

An alternative modeling approach for specifying the IDGP which has become increasingly popular in recent years is MI by sequential regression models (Raghunathan et al., 2001), which is also known under the name Fully Conditional Specification (?). This approach was conceived primarily due to a lack of joint models (distributions) when facing missing values in a mix of categorical, continuous, and count data. With the FCS approach, the imputer does not specify a joint distribution (3.25); rather, for each variable with missing data a distribution conditional on other variables in the data set is specified:

$$f_i(\mathbf{X}_i|\boldsymbol{y}, \mathbf{X}_{-i}; \boldsymbol{\eta}_i, \boldsymbol{\phi}_i). \tag{3.26}$$

FCS implementations typically employ the Markov Chain described in Algorithm 3.3. This involves visiting each variable with missing values, and drawing imputations from the posterior predictive distribution corresponding to the specified conditional model. Note that all other variables besides the currently visited are already rendered complete due to previous iterations, or assigned starting imputations. These starting values are often draws from the marginal distribution of the variable with missing values. Tables 3.1 through 3.3 on the following page illustrate the algorithm iterating through two variables with missings.

Algorithm 3.3 Markov Chain FCS

```
Step 1      (a) Draw $\boldsymbol{\eta}_1$ from $f(\boldsymbol{\eta}_1|\boldsymbol{y}, \mathbf{X}_{1,\text{obs}}, \mathbf{X}_{-1})$
            (b) Draw $\tilde{\mathbf{X}}_{1,\text{mis}}$ from $f_1(\mathbf{X}_{1,\text{mis}}|\boldsymbol{y}, \mathbf{X}_{1,\text{obs}}, \mathbf{X}_{-1}, \boldsymbol{\eta}_1)$
Step k      (a) Draw $\boldsymbol{\eta}_k$ from $f(\boldsymbol{\eta}_k|\boldsymbol{y}, \mathbf{X}_{k,\text{obs}}, \mathbf{X}_{-k})$
            (b) Draw $\tilde{\mathbf{X}}_{k,\text{mis}}$ from $f_k(\mathbf{X}_{k,\text{mis}}|\boldsymbol{y}, \mathbf{X}_{k,\text{obs}}, \mathbf{X}_{-k}, \boldsymbol{\eta}_k)$
Step k + 1  (a) Draw $\boldsymbol{\eta}_{k+1}$ from $f(\boldsymbol{\eta}_i|\boldsymbol{y}_{\text{obs}}, \mathbf{X})$
            (b) Draw $\tilde{\boldsymbol{y}}_{\text{mis}}$ from $f_{k+1}(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \mathbf{X}, \boldsymbol{\eta}_{k+1})$
```

The main difference between Algorithm 3.2 and Algorithm 3.3 is that in each step the JM approach updates all parameters of the joint imputation model (3.25), while the FCS only updates the set of imputation parameters associated with the conditional distribution (3.26) of the current step. The FCS framework splits a possibly high dimensional imputation model into multiple one-dimensional problems, which allows for the application of a wealth of existing univariate statistical models. For example, a reverse regression model conditional on the other variables could be specified for each predictor with missing values. Additionally, frameworks implementing FCS can be easily extended with custom IMs, such as the methods based on reverse regression.

Unfortunately, the great flexibility of the FCS approach allows for the possibility that there does not exists a joint distribution (3.25) such that $f_j(\mathbf{X}_j|\mathbf{X}_{-j}, \boldsymbol{y}, \boldsymbol{\eta}_j) = g(\mathbf{X}_j|\mathbf{X}_{-j}, \boldsymbol{y}, \boldsymbol{\eta})$ for all j and $f_{k+1}(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\eta}_{k+1}) = g(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\eta})$ (Liu et al., 2012). The specified conditional models are compatible when the implied joint distribution does exist, and incompatibility when it does not. An example of a compatible model is when all conditional distributions are normal with the conditional mean linearly

y	x
.36	☠
☠	.98
☠	.12
.81	☠
⋮	⋮

Table 3.1: Initial state; general missing data pattern in two variables.

$\Downarrow$	
y	x
.36	.55
☠	.98
☠	.12
.81	.18
⋮	⋮

Table 3.2: Start of first iteration, first variable is being imputed. Missing values in the second variable have been replaced with starting values, which are often drawn from the corresponding marginal distribution.

	$\Downarrow$
y	x
.36	☠
.87	.98
.86	.12
.81	☠
⋮	⋮

Table 3.3: First variable is imputed using the specified model conditional on the second variable; second variable is about to be imputed.

$\Downarrow$	
y	x
.36	.35
☠	.98
☠	.12
.81	.47
⋮	⋮

Table 3.4: Second variable is imputed using the specified model conditional on the first variable; the algorithm keeps iterating until a stopping criterion is fulfilled.

dependent on all other variables, which implies a multivariate normal distribution. Incompatible models may arise easily due to the great flexibility when specifying the conditional models, especially when these vary in complexity and richness. Up until recently, the limited simulation study of van Buuren et al. (2006) provided the only insight into the consequences of incompatibility. However, recently work of Liu et al. (2012) attempted to fill the theoretical void and provides an analysis of the characteristics of the Markov Chain when the conditional models are compatible and when they are incompatible. Liu et al. (2012) also prove consistency of the Multiple Imputation Inference for a special class of incompatible conditional distributions; however, consistency of the variance estimator could not be proven.

3.5 Conclusion

IMs based on the reverse regression IDGP defined in 3.13 are ubiquitous. However, in the highly simplified case of missing values in a single predictor variable and a MDM which is MAR, obtaining compatibility of the reverse regression IDGP without making distributional assumptions requires the consistent estimation of a model for the selection indicator conditional on the observed data; thus, data analysts cannot use the OLS variance estimator described in 2.2. From a frequentist perspective, no theoretical justification can be given for the application of existing and new IMs based on this IDGP for the purpose of Multiple Imputation Inference under a Missing Data Mechanism belonging to Class 5 (see Table 2.1 on page 19) while using the OLS estimator as Imputed Data Analysis. Further, the passive imputation method for imputing transformed variables is found to be flawed, which confirms the analysis presented by von Hippel (2009). Finally, a fundamental problem of the reverse regression method 3.13 arises when the reverse relation is multivalued.

Chapter 4

Imputation Methods

In this chapter both existing and experimental IMs are described; most of these methods are compared empirically in the simulation studies in Chapter 5. Experimental IMs are the nonparametric Local Linear Regression (LLR) and GAMLSS Imputation Methods described in Section 4.4 and Section 4.5, respectively. A fairly complete list of FCS frameworks which are available as add-on package to the `R` software environment for statistical computing is given in table 4.1; implementations of the IMs Global Linear Regression, Predictive Mean Matching, and `aregImpute` are provided there. The proposed GAMLSS method is made available as a plug-in to `mice` with a code listing and documentation given in appendix B.

Table 4.1 lists contributed add-on packages of FCS implementations together with version information. The IMs implemented in the listed software packages, and the two newly proposed IMs, are all in one way or another based on the reverse regression IDGP discussed in 3.3. While the GLR IM equals the reverse linear regression method already described in 3.3.3, the GGLR method as presented in (4.1) is based on the Generalized Linear Model (GLM), and allows for the imputation of non-continuous data. The Predictive Mean Matching (PMM) discussed in 4.2 can be seen as a type of random k-nearest-neighbor method, with the distance between the linear predictors of the reverse linear regression as (pseudo)metric. In 4.4 the LLR is proposed, which seeks to relax the linearity restriction of the GLR method. The method described in 4.3 is a combination between PMM and LLR, which also draws imputations from the k-nearest-neighbors, but uses a more involved metric. Finally, the newly proposed GAMLSS IM models both the mean and dispersion parameters of a specified distribution using generalized additive models.

`R` Package	Tested Version
`mice` (van Buuren & Groothuis-Oudshoorn, 2010)	`2.10`
`mi` (Gelman et al., 2010)	`0.09-14`
`Hmisc` (Harrell, 2010)	`3.9-0`
`BaBooN` (Meinfelder, 2011)	`0.1-6`

Table 4.1: Exhaustive list of FCS implementations in `R`, along with tested version number.

4.1 Global Linear Regression (GLR) and Generalized Global Linear Regression (GGLR)

The GLR is the most basic member of the imputation by reverse regression class of methods, and is already discussed in Section 3.3.3. It is designated "Global" to distinguish it from Local Linear Regression. Although elementary, this method is implemented in all imputation software, and together with 4.2 remains one of the most widely used method for the imputation of continuous data. Apart from availability, another advantage is the relative numerical robustness of the IM, especially when the algorithm employs regularization techniques such as ridge regression to alleviate possible problems with multicollinearity.

A generalization of the GLR is the GGLR, which is based on a flexible generalization of the LM called the GLM (McCullagh & Nelder, 1989), where the response variable $\boldsymbol{x}_1$ is assumed to be generated from a (conditional) distribution in the exponential family. The conditional expectation and variance of $\boldsymbol{x}_1$ are related to the linear predictor through the inverse of a link function $g(\cdot)$:

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right) &= g^{-1}(\tilde{\alpha} + \boldsymbol{w}\tilde{\boldsymbol{\beta}}) \\ \mathrm{Var}(\boldsymbol{x}_1|\boldsymbol{w}) &= v(g^{-1}(\tilde{\alpha} + \boldsymbol{w}\tilde{\boldsymbol{\beta}})), \end{aligned} \tag{4.1}$$

where $v(\cdot)$ is the scedastic function mapping the predicted mean to the conditional variance; its form follows from the specified distribution and link function.

The GGLR requires specification of a conditional distribution for $\boldsymbol{x}_1$ and link function $g(\cdot)$, where the choice of link function is somewhat arbitrary. The specification of $f(\boldsymbol{x}_1|\boldsymbol{w})$ is typically based on the observed range of values of $\boldsymbol{x}_1$, where it is implicitly assumed that the marginal distribution $f(\boldsymbol{x}_1)$ belongs to the same family as the conditional distributions $f(\boldsymbol{x}_1|\boldsymbol{w})$; however, this is only true in some special cases. One example is when the data are distributed multivariate normal, in which case $g(x) = x$ and $v(x) = \sigma^2$, which corresponds to the GLR method for imputing continuous data. Another case is

$$\begin{aligned} x &\sim \text{Bernoulli}(p) \\ y|x &\sim \mathcal{N}(\alpha + x\boldsymbol{\beta}, \sigma^2), \end{aligned}$$

which implies that $x|y \sim \text{Bernoulli}(\frac{e^{\tilde{\alpha}+x\tilde{\boldsymbol{\beta}}}}{1+e^{\tilde{\alpha}+x\tilde{\boldsymbol{\beta}}}})$ (Efron, 1975). However, suppose $x \sim \text{Poisson}(\lambda)$. Then,

$$\begin{aligned} \mathrm{E}\left(x\right) &= \mathrm{Var}(x) \\ &= \mathrm{Var}(\mathrm{E}\left(x|y\right)) + \mathrm{E}\left(\mathrm{Var}(x|y)\right). \end{aligned} \tag{4.2}$$

`mi` provides an IM based on a GLM with a conditional Poisson distribution for x, and Kleinke et al. (2012) propose a GGLR with a Negative Binomial distribution, which is frequently used for modeling over-dispersed count data. However, specifying a GGLR with $x|y \sim \text{Poisson}(e^{\tilde{\alpha}+x\tilde{\boldsymbol{\beta}}})$ or an over-dispersed distribution implies that $\mathrm{Var}(x|y) \geqq \mathrm{E}\left(x|y\right)$ for all y, which in turn implies that

$$\mathrm{E}\left(\mathrm{Var}(x|y)\right) \geqq \mathrm{E}\left(\mathrm{E}\left(x|y\right)\right) = \mathrm{E}\left(x\right), \tag{4.3}$$

which contradicts (4.2) in general: ironically, the imputation model for x should allow for under-dispersion instead of over-dispersion, such that $\mathrm{Var}(x|y) < \mathrm{E}\left(x|y\right)$ for all y .

Although elementary, GGLRs are implemented in almost all imputation software, and together with PMM remain one of the most widely used IMs. A disadvantage of the method is the restrictions on the functional form of the conditional mean and variance of $\boldsymbol{x}_1$, which may lead to inconsistent estimation of the conditional expectation and variance, and ultimately to invalid MII.

4.2 Predictive Mean Matching

PMM was first proposed in the seminal book of Rubin (1987) and in Little (1988). A comparison with the GLR method when estimating the marginal mean and marginal distribution of a variable with missing values was undertaken in a simulation study by Schenker & Welsh (1988). None of the articles mentioned above derived the large-sample properties of the method, and only Schenker & Welsh (1988) tested the method empirically, although with respect to marginal statistics. Despite lack of theoretical and empirical support, the method is currently adopted as the standard method in the widely used `mice` package for MII with respect to $\boldsymbol{\beta}$.

PMM can be seen as a type of random k-nearest-neighbor method. Given a metric $d : \mathbb{R}^{2k} \rightarrow \mathbb{R}$ and a query point for which $\boldsymbol{x}_1$ is missing, the p nearest neighbors of the query point are sought to obtain a set of p donor values from which an imputation is randomly drawn. What differentiates PMM from nearest neighbor methods is the metric used, which is defined in terms of the linear predictor of the reverse linear regression:

$$d_{PMM}(\boldsymbol{a}, \boldsymbol{b}) = |\boldsymbol{a}\dot{\boldsymbol{\beta}} - \boldsymbol{b}\dot{\boldsymbol{\beta}}| = |(\boldsymbol{a} - \boldsymbol{b})\dot{\boldsymbol{\beta}}|, \tag{4.4}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are realizations of $\boldsymbol{w}$, and $\dot{\boldsymbol{\beta}}$ are (approximated) draws from the posterior distribution of the parameters of the reverse linear regression. Since matching is done using the linear predictor and imputed values are "live" or observed, the method can also be used for the imputation of non-continuous data without the need for iterative maximum likelihood fitting. Algorithm 4.1 describes the method more formally.

Assuming (a) the distance function d is topologically equivalent to Euclidean distance, and (b) the size of the donor pool increases with the sample size as $p(n) = \sqrt{n}$, Dahl (2007) shows that

$$(\tilde{\mathbf{X}}_{1,\text{mis}}|\mathbf{W}_{\text{mis}}) \xrightarrow{\mathcal{D}} (\mathbf{X}_{1,\text{mis}}|\mathbf{W}_{\text{mis}}). \tag{4.5}$$

Further, an upper bound of the correlation between an imputed value $\tilde{\mathbf{X}}_{1,i}$ with $i \in \text{mis}$, and any measurable function $f(\cdot)$ of the observed cases $(\mathbf{W}, \mathbf{X}_{1,\text{obs}})$ is given as

$$|\text{Cor}\left(\tilde{\mathbf{X}}_{1,i}, f(\mathbf{W}, \mathbf{X}_{1,\text{obs}})\right)| \leq n^{-\frac{1}{4}}. \tag{4.6}$$

Thus, when condition (a) and condition (b) are fulfilled, nearest neighbor imputation methods produce imputations which are (asymptotically) independent over observations, and have the correct conditional distribution. However, all implementations of PMM listed in Table 4.1 on page 38 set $p(n) = c$, where c is typically 5 or 10, and thus violate condition (b) because the number of nearest neighbors does not increase with sample size. Further, the use of metric (4.4) and $\mathbf{W}$ having multiple columns violates condition (a), since then $d(a, b)$ can be zero even if $a \neq b$. Thus, the asymptotic properties described above are no longer guaranteed, and it is unknown if current implementations of PMM produces imputations which are asymptotically independent over observations, and have the correct conditional distribution.

Problems may occur when regions of the sample space are sparsely populated, possibly due to the MDM. For example, in Figure 4.1 on page 41 the area between $-.1 < y < .3$ has a lot of missing values in x. Because of the low number of observed values of x, the same donors are considered for each missing value, which might result in underestimation of $\hat{V}_{\text{MI}}$. Further, PMM is unable to extrapolate cor-

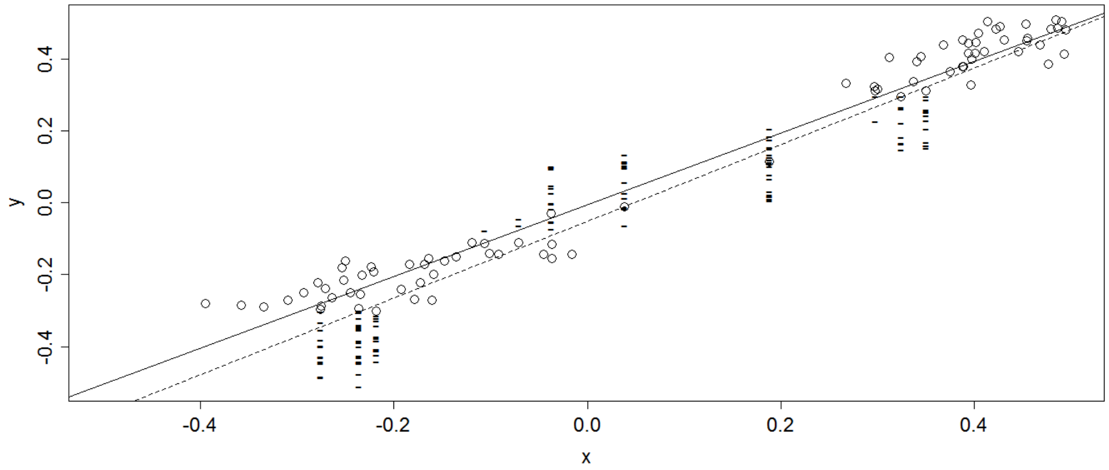

Figure 4.1: PMM with a sparsely populated region between $-.1 < y < .3$, and truncation at $y < -.3$. The circles are observed values, and the horizontal bars imputed values. The black line depicts the regression slope based on the observed and missing values, and the dashed line is based on the observed and imputed values.

rectly from the observed values to the truncated region $y < -.3$, leading to a biased estimate of the regression slope. Although often heralded for imputing "realistic" values, the resulting inability to properly inter- and extrapolate can be a serious weakness of the method, especially when the MDM is selective.

The PMM implementation of the `mice` package version draws from the three closest donors. However, it features the following unusual distance function,

$$d_{MICE}(\boldsymbol{a}, \boldsymbol{b}) = |\boldsymbol{a}\dot{\boldsymbol{\beta}} - \boldsymbol{b}\tilde{\boldsymbol{\beta}}|,$$

where $\tilde{\boldsymbol{\beta}}$ are the posterior means of the parameters of the reverse regression model, and $\dot{\boldsymbol{\beta}}$ are draws from the corresponding posterior distribution. According to the CHANGELOG file of the `mice` package, this distance function is supposed to "add between imputation variability in the case of a single predictor"; no theoretical justification is given in the package documentation.

4.3 `aregImpute`

The function `aregImpute` in the package `Hmisc` is another readily available alternative for end users. This IM has not been published: there are no large-sample results available, and the method has not been evaluated using simulations. The only source of information, apart from the program code itself, is the (rather sparse) documen-

Algorithm 4.1 PMM (taken in part from Dahl (2007))

```
1. Fit model (3.21) to $(\boldsymbol{x}_1, \mathbf{W})$
2. Draw $\dot{\alpha}$ , $\dot{\boldsymbol{\beta}}$, and $\dot{\sigma}^2$ from their respective (approximated) posterior distributions
3. For each $q \in$ mis
    (a) For each $i \in$ obs
        $d_i = d_{PMM}(\mathbf{W}_q, \mathbf{W}_i)$
    (b) Let $I \subset$ obs so that $|I|{=}k$ and $(i \in I, j \notin I) \Rightarrow d_i \leq d_j$
    (c) Let $l$ be a random element of $I$
    (d) Let the imputed value be equal to $\mathbf{X}_{1,l}$
4. Repeat steps 2 through 3 $m$ independent times, where $m$ is the number of desired imputations
```

tation contained in Harrell (2010). Therefore, we cannot describe the method in much detail, and the package remains somewhat of a "black box".

First, the algorithm finds those transformations of the predictors $f_j(\mathbf{W}_{(j)})$ which lead to optimal prediction of a linear transformation of $\mathbf{X}_{(1)}$ in the following additive model:

$$\tilde{c} + \mathbf{X}_{(1)}\tilde{d} = \tilde{\alpha} + \sum_{j=1}^{J} f_j(\mathbf{W}_{(j)})\tilde{\boldsymbol{\beta}}_j + \boldsymbol{\nu}, \tag{4.7}$$

where the $f_j(\cdot)$ are restricted cubic spline basis functions with a user specified fixed number of knots. After estimation of (4.7), a variant of PMM using weighted probability sampling of donor values is used to generate imputations, where the weights are inversely proportional to the following distance function:

$$d_{areg}(\boldsymbol{a}, \boldsymbol{b}) = \sum_{j=1}^{J} |(f_j(\boldsymbol{a}_j) - (f_j(\boldsymbol{b}_j))\tilde{\boldsymbol{\beta}}_j|, \tag{4.8}$$

and where $\boldsymbol{a}$ and $\boldsymbol{b}$ are realizations of $\boldsymbol{w}$. Te method uses the simple non-parametric bootstrap to approximate draws from the Bayesian posterior distribution of the parameters of the imputation model. Since the final imputed values are produced using PMM, `aregImpute` can also be used for the imputation of non-continuous data.

4.4 Local Linear Regression

As described in 3.3.5, $\mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right)$ is non-linear when $\boldsymbol{x}_1$ and y are not distributed bivariate normal given $\boldsymbol{x}_{-1}$. The GGLR method imposes restrictions to the functional form of the conditional mean and variance of $\boldsymbol{x}_1$, and may therefore fail to consistently estimate (3.13). Further, nearest neighbor methods such as PMM may run into problems when regions of the sample space are sparsely populated. To be robust against possible non-linearities, a non-parametric technique such as local linear regression can be used to estimate $\mathrm{E}\left(\boldsymbol{x}_1|\boldsymbol{w}\right)$. The IM we propose in this section uses a form of locally weighted learning, where for each missing datum a query is put forward. Each query is answered by fitting a local model to the data points near the query point. The local model used is a linear regression, where the data are inversely weighted according to their distance to the query point. This local linear estimator is minimax efficient and is one of the best known approaches for boundary correction (Li & Racine, 2004).

It is a common belief in the local learning literature that the final performance of the local model is most sensitive to the bandwidth and to the distance metric used (Atkeson et al., 1997). As a starting point, we use a global distance metric. However, the bandwidth is selected locally on a query-by-query basis. This allows for a better adaptation to the local characteristics of the data. After the structural and parametrical identification of the local model, the query is answered by drawing an imputation from the approximated posterior predictive distribution of the local linear model.

Given a query point $\mathbf{W}_q$, the parameters $\tilde{\alpha}$ and $\tilde{\boldsymbol{\beta}}$ of a local linear approximation of $f(\cdot)$ in a neighborhood of $\mathbf{W}_q$ can be obtained by solving the local polynomial regression

$$\sum_{i \in \text{obs}} \left\{ \left(\mathbf{X}_{1,i} - (\tilde{\alpha} + \mathbf{W}_i \tilde{\boldsymbol{\beta}}) \right)^2 K\left(\frac{d(\mathbf{W}_i, \mathbf{W}_q)}{h} \right) \right\}, \tag{4.9}$$

where $K(\cdot)$ is the weight function, and $d(\mathbf{W}_i, \mathbf{W}_q)$ is the distance function in predictor space from the query point $\mathbf{W}_q$ to the i^{th} data point $\mathbf{W}_i$ (Bontempi et al., 2000). In the literature, no substantial empirical differences have been found with regard to the choice of weight function (Atkeson et al., 1997). However, when the *uniform weighting kernel*

$$K\left(\frac{d(\mathbf{W}_i, \mathbf{W}_q)}{h} \right) = \begin{cases} 1 & \text{if } d(\mathbf{W}_i, \mathbf{W}_q) \leq h \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

is adopted, the optimization of the bandwidth h can be conveniently reduced to the optimization of the number of neighbors p to which a unitary weight is assigned in the local regression evaluation. The distance between the query point $\mathbf{W}_q$ and an input $\mathbf{W}_i$ is computed using the weighted Euclidean distance

$$d(\mathbf{W}_i, \mathbf{W}_q) = \sqrt{(\mathbf{W}_i - \mathbf{W}_q)\mathbf{M}(\mathbf{W}_i - \mathbf{W}_q)^{\mathrm{T}}}, \tag{4.11}$$

where $\mathbf{M}$ is determined by the global relative linear influence of the predictors. More specifically, $\mathbf{M}$ is a diagonal matrix with

$$\mathbf{M}_{jj} = \sqrt{\frac{\hat{\gamma}_j^2}{\sum_{i=1}^{k} \hat{\gamma}_i^2}}, \tag{4.12}$$

where $\hat{\gamma}_j$ is the j^{th} of k standardized regression coefficients as estimated by least-squares on the whole training set, excluding the intercept.

To find the local model with the optimum number of neighbors p, $p_{max} - p_{min}$ models are fitted, where p_{min} and p_{max} are tuning parameters and control the maximum and minimum number of neighbors to use, respectively. We set $p_{min} = 3k$ and $p_{max} = n_{\mathrm{obs}}$. For each fitted model, the mean squared error

$$\mathrm{MSE}(p) \approx \frac{1}{p} \sum_{i \in S} \left(e_i^{cv}(p)^2 \right) \tag{4.13}$$

is evaluated, where S is the set containing the p nearest neighbors of $\mathbf{W}_q$, and $\mathbf{e}^{cv}(p)$ contains the leave-one-out (l-o-o) errors associated with the model with p number of neighbors:

$$e_i^{cv}(p) = \mathbf{X}_1 - \mathbf{W}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_{-i}(p) - \tilde{\alpha}, \tag{4.14}$$

where $\tilde{\boldsymbol{\beta}}_{-i}(p)$ are the parameter estimates of the local linear model without using the i^{th} datum. Cross-validation by means of the l-o-o errors is used as a model selection criterion, because it allows for significant computational shortcuts during model fitting and validation using the Prediction Sum of Squares (PRESS) statistic (Allen, 1974). The model $\hat{p}$ with the smallest MSE(p) is then selected to answer the query, where a final prediction is obtained as follows:

$$\tilde{\mathbf{X}}_{1,q}(\mathbf{W}_q) = \tilde{\alpha} + \mathbf{W}_q^{\mathrm{T}} \tilde{\boldsymbol{\beta}}(\hat{p}) \tag{4.15}$$

The mean squared error of the selected model is used as an estimate of the local error variance:

$$\hat{\sigma}^2(\mathbf{W}_q) = \mathrm{MSE}(\hat{p}) \tag{4.16}$$

The implementation of the IM is largely based on the `R` package `lazy` (Birattari & Bontempi, 2003).

Unfortunately, the package `lazy` does not support Bayesian inference. Therefore, it is impossible to obtain multiple imputations by drawing from the posterior predictive distribution. To nevertheless incorporate the added variance due to non-response into the MII, the posterior predictive distribution of the missing values (3.1) is approximated by the bootstrap predictive distribution (Harris, 1989) as described in 4.5.2.

Although experimenting with LLR helped to explore the possibilities of using nonparametric techniques as imputation models, the method has been discarded in favor of the GAMLSS method described in 4.5, because GAMLSS generalizes more easily to non-continuous data, and offers improved estimation of the conditional variance of $\boldsymbol{x}_1$.

4.5 GAMLSS

In the literature, several methods are available which jointly model the conditional expectation and conditional variance, and iteratively estimate both using nonparametric techniques. For example, Yu & Jones (2004) propose a local linear regression method with estimators based on the local normal likelihood. Rigby & Stasinopoulos (1996) propose a similar idea using semi-parametric additive models based on the penalized normal likelihood. Both approaches involve first fitting the conditional mean using local linear regression or a smoother while holding the conditional variance fixed, and then fitting the conditional variance using local linear regression or smoother while holding the conditional mean fixed. Rigby & Stasinopoulos (2005) propose the GAMLSS model, which allows for relaxation of the normality assumption and the specification of arbitrary families of conditional distributions for $\boldsymbol{x}_1$, even ones outside of the exponential family.

4.5.1 The IDGP

In the IDGP of the GAMLSS IM, at least the mean and dispersion parameters of a specified distribution $\mathcal{D}$ are modeled using additive terms:

$$\begin{aligned} g_1(\boldsymbol{\mu}) =& \tilde{\alpha}_{(1)} + \sum_{j=1}^{k} h_{1j}(\mathbf{W}_{(j)}) \\ g_2(\boldsymbol{\sigma}) =& \tilde{\alpha}_{(2)} + \sum_{j=1}^{k} h_{2j}(\mathbf{W}_{(j)}), \end{aligned} \tag{4.17}$$

where $g_i(\cdot)$ are monotonic link functions which relate the parameters $\boldsymbol{\eta}$ of the conditional distribution $\mathcal{D}$ to the predictor variables $\boldsymbol{w}$, and h_{ij} are smoother terms (Rigby & Stasinopoulos, 1996, 2005). The distribution, which we denote by $\mathcal{D}$, defaults to normal for continuous data, but alternatives can be chosen from a broad range of alternatives. This enables users in combination with a suitable link function to restrict the drawn imputations to a certain range by specifying for example a truncated normal distribution, and allows for easy generalization of the method to discrete and count data. An improvement over the predecessor method described in 4.4 is that the conditional variance is also estimated using an additive model, which allows for better adaptation to local heteroscedasticity. The downside of the necessity of specifying $\mathcal{D}$ is of course that misspecification can lead to invalid MII; this will be illustrated in 5.4.6. This problem is aggravated by the fact that there are often no theoretical considerations for choosing a distribution, since theory is typically focused on the ADGP, and not on aspects of the scientifically uninteresting IDGP.

If besides location and scale $\mathcal{D}$ has up to two shape parameters $\{\nu, \tau\}$ and the sample size is relatively large, we can extend (4.17) by modeling these parameters additively:

$$\begin{aligned} g_3(\boldsymbol{\nu}) &= \tilde{\alpha}_{(3)} + \sum_{j=1}^{k} h_{3j}(\mathbf{W}_{(j)}) \\ g_4(\boldsymbol{\tau}) &= \tilde{\alpha}_{(4)} + \sum_{j=1}^{k} h_{4j}(\mathbf{W}_{(j)}) \end{aligned} \tag{4.18}$$

Since this extended model portrays the conditional distribution $f(\boldsymbol{x}_1|\boldsymbol{w})$ more accurately, the resulting imputations may be of higher quality compared to those whose IDGP solely consists of (4.17).

4.5.2 Implementation

The `R` implementation of the IM, whose code listing is given in Chapter A, uses the `gamlss` package (Rigby & Stasinopoulos, 2005) to fit model (4.17). Rigby & Stasinopoulos (2005) provide a description of the algorithms used by this package; however, no large sample properties are derived. Implemented smoothing terms h_{ij} include cubic smoothing splines, penalized splines, and local regression. In principle, any smoother can be used; however, penalized B-splines Eilers & Marx (1996) proved to be computationally the most stable. More specifically, the smoother used in the simulation studies in Chapter 5 consists of a penalized B-spline with 20 knots, a piecewise polynomial of the second degree with a second order penalty, and au-

Algorithm 4.2 GAMLSS imputation

```
1. Fit Model (4.17), possibly extended to (4.18), using the observed data
   $\{\mathbf{X}_{1,\text{obs}}, \mathbf{W}_{\text{obs}}\}$.

2. Resample $\mathbf{X}_{1,\text{obs}}$ as follows:

   $\mathbf{X}^*_{1,\text{obs}} \sim \mathcal{D}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\tau}})$

   Define a bootstrap sample $B := \left\{\mathbf{X}^*_{1,\text{obs}}, \mathbf{W}_{\text{obs}}\right\}$

3. Refit model (4.17) or (4.17) and (4.18) using $B$. Draw $n_{\text{mis}}$ imputations for
   $\mathbf{X}_{1,\text{mis}}$ as follows:
   $\tilde{\mathbf{X}}_{1,\text{mis}} \sim \mathcal{D}(\dot{\boldsymbol{\mu}}, \dot{\boldsymbol{\sigma}}, \dot{\boldsymbol{\nu}}, \dot{\boldsymbol{\tau}})$

4. Repeat step (2) and (3) $m$ independent times, where $m$ is the number of
   imputations.
```

tomatic selection of the smoothing parameter using the Local Maximum Likelihood criterion. For high amounts of smoothing, the fit of this smoother approaches linearity.

Unfortunately, the package `gamlss` does not support Bayesian inference. Therefore, it is impossible to obtain multiple imputations by drawing from the posterior predictive distribution. To nevertheless incorporate the added variance due to nonresponse into the MII, the posterior predictive distribution of the missing values (3.1) is approximated by the bootstrap predictive distribution (Harris, 1989):

$$\begin{aligned} f^*(\mathbf{X}_{\text{mis},1}|\mathbf{X}_{\text{obs},1}, \mathbf{W}) = \int & f(\mathbf{X}_{\text{mis},1}|\tilde{\boldsymbol{\eta}}, \mathbf{W}_{\text{mis}}) \\ & \times f(\tilde{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}(\mathbf{X}_{\text{obs},1}, \mathbf{W}_{\text{obs}}))d\tilde{\boldsymbol{\eta}}, \end{aligned} \tag{4.19}$$

where $\tilde{\boldsymbol{\eta}}$ denote the possible values of the imputation model parameters, $\hat{\boldsymbol{\eta}}(\mathbf{X}_{\text{obs},1}, \mathbf{W}_{\text{obs}})$ is the estimator of said parameters, and $f(\tilde{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}(\mathbf{X}_{\text{obs},1}, \mathbf{W}_{\text{obs}})$ is the sampling distribution of the imputation parameters evaluated at the estimated values of the parameters. $f(\tilde{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}(\mathbf{X}_{1,\text{obs}}, \mathbf{W}_{\text{obs}})$ is simulated by performing the parametric bootstrap, and acts as a surrogate for the posterior distribution of the parameters of the imputation model. A full description of the algorithm is given in Algorithm 4.2. An advantages compared to a fully Bayesian approach is that no prior information – which is typically lacking – needs to be specified.

Even though the implementation of penalized smoothing splines in the package `gamlss` is considered to be the most stable, it has been observed that the in some cases Algorithm 4.2 may fail to converge. This is frequently traced to the algorithm which selects the smoothing parameter. The implementation of the IM catches such an occurrence, and then falls back to a cubic smoothing spline[1] with a fixed smoothing parameter consisting of one additional degrees of freedom on top of the linear term, which indicates a very large amount of smoothing. Even with these measures in place, GAMLSS may fail to converge in some scenarios, especially for low sample sizes, as will become apparent in the simulation studies.

A difference between `aregImpute` and the proposed IM is that the former fixes the number of knots of the transformations f_j to a default fixed value, while GAMLSS optimizes the smoothing parameter of h_{ij} using cross-validation; after all, the performance of a smoother is extremely sensitive to the appropriateness of the chosen smoothing parameter. Also, `aregImpute` draws by default imputations from the observed values using PMM, while `GAMLSS` samples imputations from $\mathcal{D}$ using the estimated parameter values.

4.5.3 Extension to multilevel models

Many research designs in the social sciences yield data that have a hierarchical, nested or clustered structure. Examples include pupils within classes, children within families, occasions within an individual, experiments within batches, tests within laboratories, and so on. Classic statistical techniques, including the variance estimator (2.15), fail to take into account that observations from the same cluster are likely to be dependent on each other, and are therefore not suited to analyze such data. Multilevel models have been developed as an alternative, and are often used in the social sciences for the analysis of data with these complex patterns of variability. An basic multilevel model is the following extensions of model (2.1):

$$y = \alpha + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + u + \boldsymbol{b}_l, \tag{4.20}$$

where l is a fixed index denoting to which of the L groups (classes, families, individuals) the observation belongs, and

$$\boldsymbol{b} \sim \mathcal{N}(0, \tau^2\mathbf{I}_L) \quad \text{and} \quad u \sim \mathcal{N}(0, \sigma^2). \tag{4.21}$$

Equations (4.20) and (4.21) together are called the random intercept model, with $\boldsymbol{b}_l$ being the random effects; this model can be estimated by for example Maximum Likelihood. Note that (4.21) extends (2.2) to $\mathrm{E}\,(u|\boldsymbol{x}, \boldsymbol{b}_l, s) = 0$, and also implies that $\mathrm{E}\,(\boldsymbol{b}_l|\boldsymbol{x}, s) = 0$; in economics, researchers are hesitant to make the latter assumption, and prefer to use the fixed effects model, which will not be discussed here.

[1] The function `cs` in the `GAMLSS` package

IDGPs which ignore the hierarchical structure of the data may lead to invalid Multiple Imputation Inference; therefore, Schafer & Yucel (2002) present an IDGP based on a generalization of model (4.20) within the Joint Modeling framework. De Jong (2006) implemented an IM for the imputation of hierarchical dichotomous data within the Fully Conditional Specification framework. IDGP (4.17) can easily be extended to a random intercept model by adding an additive term for the random effects $\boldsymbol{b}$; the `GAMLSS` package already supports this.

All multilevel IDGPs mentioned above assume that the random effects follow a normal distribution.Yucel & Demirtas (2010) investigated how robust the Imputation Method proposed in Schafer & Yucel (2002) is to violations of this distributional assumption. When the rate of missingness is relatively high and the true distribution of the random effects deviates from normality, they found that the validity of the MII is adversely affected, especially for parameters such as τ^2 which describe the distribution of the random-effects. One possible solution would be to develop IMs with less restrictive distributional assumptions about the random effects.

4.5.4 Discussion of Assumptions

When the IDGP contains many predictors, a problem non-parametric smoothers face is the 'curse of dimensionality', where the volume of predictor space grows so fast that the available data becomes sparse; this generally leads to an explosion of the variance of the non-parametric estimator, and computational problems. One strategy to cope with the curse is to force predictors to enter the model additively as in (4.17). Although the additivity assumption allows for the incorporation of a moderate number of predictors in the imputation model, it cannot capture the effects of potential interactions between the predictors of the imputations model. When interaction terms need to be included in the model, they should be explicitly specified; even if the direct regression is additive, it is generally unknown if the functional relationship relating the predictors of the IM to the parameters of the specified conditional distribution is also additive. Another possible limitation is that the functions $h_{ij}(\cdot)$ to be estimated in (4.17) and (4.18) should be sufficiently smooth; functions with pronounced discontinuities might lead to incompatible imputations.

Finally, estimating arbitrary smooth functions using flexible non-parametric estimators requires more data than required for a linear regression, and the GAMLSS IM might not be appropriate for small samples.

Chapter 5

Simulation Experiments

In this chapter, the simulation experiments are described which were conducted to investigate the performance of the Imputation Methods (IMs) listed in Chapter 4. Firstly, an overview of Monte Carlo Statistical Simulation (MCSS) will be given in 5.1, along with a discussion of its limitations; in particular, MCSS will be compared and contrasted with theoretical large-sample results, and it will be shown how the two methods complement each other. A brief overview of existing work will be given in 5.2. Then, the used experimental design will be described in detail in 5.3. Finally, the results are presented and interpreted in 5.4; conclusions are deferred to Chapter 6.

5.1 The use of Monte Carlo Statistical Simulation

In the planning phase of the research project, simulation studies were decided to be the main research instrument for gaining insight into the performance of existing and the proposed IMs. Considering the important role this technique played in the project and in light of the experiences gained, it seems prudent to describe how Monte-Carlo experiments are used in this work and that of contemporary research. Note that this section contains subjective statements and interpretations.

In statistical research, the advent of cheap and potent personal computers has catapulted the field of computational statistics to new highs both in research and teaching. Capitalizing on increased computational power, researchers in this field often employ MCSS. These simulations experiments involve the repeated sampling from a statistical model or Data Generating Process (DGP), and applying statistical algorithms to the generated data sets. MCSSs can complement theoretical large-sample results by providing insight into the finite sample properties of statistical methods; analytical finite-sample properties are only available for a limited number of estimators. After the large sample properties of a proposed inference technique has been rigorously investigated, MCSSs offer some reassurance that the method not only works "on paper", but is also of practical worth. Also, the performance of a number of estimators can be compared to each other. Finally, note that it is entirely possible for an estimator to have good large-sample properties but poor finite-sample performance, and vice versa.

Exposing large-sample properties of an estimator is a key element of proper statistical research; for instance, consistency is usually considered a minimum requirement for an inference procedure (Serfling, 2002). However, with the surge of computational statistics, it has become more common to propose statistical techniques and methods without an accompanying analysis of their large-sample properties. This phenomenon was especially prevalent in Multiple Imputation (MI) research, where the complexity of the problem has hindered rigorous analysis; for example, as described in 4.2, there are no asymptotic results supporting application of the Predictive Mean Matching (PMM) method for multiple imputation inference about $\boldsymbol{\beta}$. With respect to deriving the statistical properties of the Fully Conditional Specification (FCS) framework, it seems that the development of the mathematical tools necessary for obtaining analytical results strays behind the recent advancements in computing technology. Statistical research has put less emphasis on theoretical results, and frequently employs MCSSs to show the alleged superiority of new statistical contraptions.

Decreased emphasis on theory is arguably also affecting research in the social sciences; for instance, the body of knowledge of Psychology is fragmented, with studies often lacking embedment in a substantive overarching theoretical framework. With the number of publications becoming the new academic currency, and the rise of the publish-or-perish phenomenon (De Rond, 2005), researches are pressured to publish frequently and fast, or forfeit their career. Developing interesting and creative theories is too time-consuming, and reviewers might reject novel ideas when they run counter to the majority view. Instead, a risk adverting strategy seems to work the system, chase statistical significance, and churn out formulaic and relatively risk free papers.

When setting up a MCSS, a scenario needs to be specified, which is defined in terms of the values of the parameters which index the family of DGPs as described in (2.1). Possible parameters are the distribution of the predictors, distribution of the missing data indicator, sample size, and the true value of $\boldsymbol{\beta}$. While the limiting behavior obtained by asymptotic analysis typically generalizes to all possible scenarios of this family – provided that the sample size is large enough[1] – results of a simulation

[1] It is typically unknown at which sample size the asymptotic behavior "kicks in"; further, this

study can strictly speaking only be extended to all possible sample originating from a given scenario. Naturally, it is impossible to enumerate or simulate all possible scenarios. Therefore, one can never prove on the basis of simulation studies that an inference procedure performs favorable in general. On the contrary, MCSSs might give misleading results when the performance of an estimator fluctuated strongly with a change in simulation scenario.

Because it is unclear how to assign a measure to the simulation space, the average performance of a method over all possible scenarios remains unknown. Researchers are unable to randomly sample scenarios, and instead deliberately choose which scenarios to simulate. Therefore, evaluating estimators using MCSSs fails to provide an objective frame of reference. Since favorable simulation results are more likely to be published, it is enticing for researchers to present the method in the most favorable conditions; counter-examples where a proposed method fails are not sought often enough. With these cautions in mind, the IMs described in Chapter 4 are now compared using simulations.

5.2 Existing work

Parallel to the initial phase of the project, He & Raghunathan (2009) was published, which investigates within the FCS framework the sensitivity of the Global Linear Regression (GLR) and PMM IMs to different conditional distributions of the variables to be imputed. In a setting with three variables and missing data which are Missing Completely at Random (MCAR), they demonstrate that with respect to the estimation of regression coefficients, currently used MI procedures can in fact give worse performance than Complete Case Analysis (Complete Case Analysis (CCA)) under seemingly innocuous deviations from standard (multivariate normality) simulation conditions.

5.3 Design

5.3.1 Simple

First we will empirically investigate the performance of all the IMs described in Chapter 4 in the context of a simple linear regression model following (2.1) with a single predictor variable x. The true parameter values are $\alpha = 0$ and $\beta = 1$; since as discussed in Chapter 2 the intercept is of no scientific interest, only the regression

may depend on other simulation parameters.

	Imputed Data Analysis (IDA)	IM
y	Response	Predictor
x	Predictor	Response

Table 5.1: Role of the variables in the Completed Data Inference (CDI) and Imputation Model (IM) when there are missing values in x.

slope β is reported in the results. Thus, there is a single variable with missing values, and the imputation model is limited to a single fully observed predictor variable y. It is again important to realize that, just as in the theoretical discussion in 3.3, the predictor x and response variable y swap roles in the imputation model, as depicted in 5.1; experience has it that this crucial but confusing point is often misunderstood.

The general nature of MI typically means it is applied to large data sets. This is also in aligned with current research practices in the social sciences, which often feature data sets with few cases and a large number of variables. One could therefore argue that this simulation study is severely limited and of little practical relevance. However, constraining the scope of the experiments to a single variable with missing values allows for the isolation of defects in the IMs from possible confounding issues stemming from the FCS framework in which the IMs are ultimately embedded. A further constraint is the limitation to a single predictor variable in the imputation model. However, obtaining acceptable performance in this basic setting is not a trivial task, and acceptable performance is a prerequisite for more involved scenarios. Further, this basic scenario simplifies experimenting with the distribution of the predictor x and other parameters of the simulation, because computational cost is less high than with multiple predictors. Finally, two simulation studies with multiple predictor variables are presented in 5.3.2.

The three simulation parameters which are structurally varied are the distribution of x, coefficient of determination R^2, and sample size N. All studies have 1000 replications and $m = 10$ imputations, and a normal distribution for the errors of the complete data model (2.1). A very important factor is the distribution of the predictor x. The following continuous densities are considered:

$$\begin{aligned}
\text{Normal:} \quad f(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \\
\text{Skew-Normal:} \quad f(x) &= 2\phi(x)\Phi(\lambda x), \\
\text{Uniform:} \quad f(x) &= \begin{cases} 1 & \text{if } 0 \leqq x \leqq 1 \\ 0 & \text{otherwise} \end{cases} \\
\text{Squared Uniform (Beta):} \quad f(x) &= \frac{x^{-\frac{1}{2}}(1-x)}{B(\frac{1}{2},1)} \\
\text{Student T:} \quad f(x) &= \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}(1+\frac{x^2}{v})^{-\frac{v+1}{2}},
\end{aligned} \tag{5.1}$$

where $\lambda = 5$ is the shape parameter, and $v = 3$ is the degrees of freedom. Figure 5.1 on page 55 features a plot of the continuous distributions used. Further, the following non-continuous distributions are considered

$$\begin{aligned}
\text{Poisson:} \quad f(x) &= \frac{\kappa^x}{x!} e^{-\kappa} \\
\text{Bernoulli:} \quad f(x) &= p^x(1-p)^{1-x},
\end{aligned} \tag{5.2}$$

with $\kappa = 2$. The study is performed for all combinations of the distributions listed above and the factor levels $R^2 \in \{.25, .50, .75\}$ and $N \in \{200, 500, 1000\}$. Note that the square of a standard uniform variable, denoted by Squared Uniform above, equals the Beta distribution with parameters $\alpha = \frac{1}{2}$ and $\beta = 1$.

For all studies, the following Missing Data Mechanism (MDM) is imposed:

$$p(s|y) = \begin{cases} (\varphi_1)^{1-s}(1-\varphi_1)^s & \text{if } y < \tilde{y} \\ (\varphi_2)^{1-s}(1-\varphi_2)^s & \text{if } y \geqq \tilde{y}, \end{cases} \tag{5.3}$$

where $\tilde{y}$ is the sample median. Setting $\varphi_1 = .1, \varphi_2 = .7$ results in 40% missing data in x, where the chance of a missing datum in x is .1 when the corresponding value of y is smaller than the sample median of y, and .7 when it is larger than the median; note that this MDM corresponds with Class 5 in Table 2.1 on page 19. While holding the MDM constant at (5.3), the coefficient of determination determines the extent to which the missing values are MAR, with R^2 approaching 0 implying the missing data are in fact MCAR and evenly spread, and a high value of R^2 giving rise to a strongly systematic MDM with the potential of thinning out select regions of the sample space. To replicate parts of the study of He & Raghunathan (2009), two simulation studies are conducted where 40% of the values of $\boldsymbol{x}_1$ are MCAR, and $f(x)$ is the normal or the Beta distribution; all other simulation parameters are equal to those in the MAR experiments.

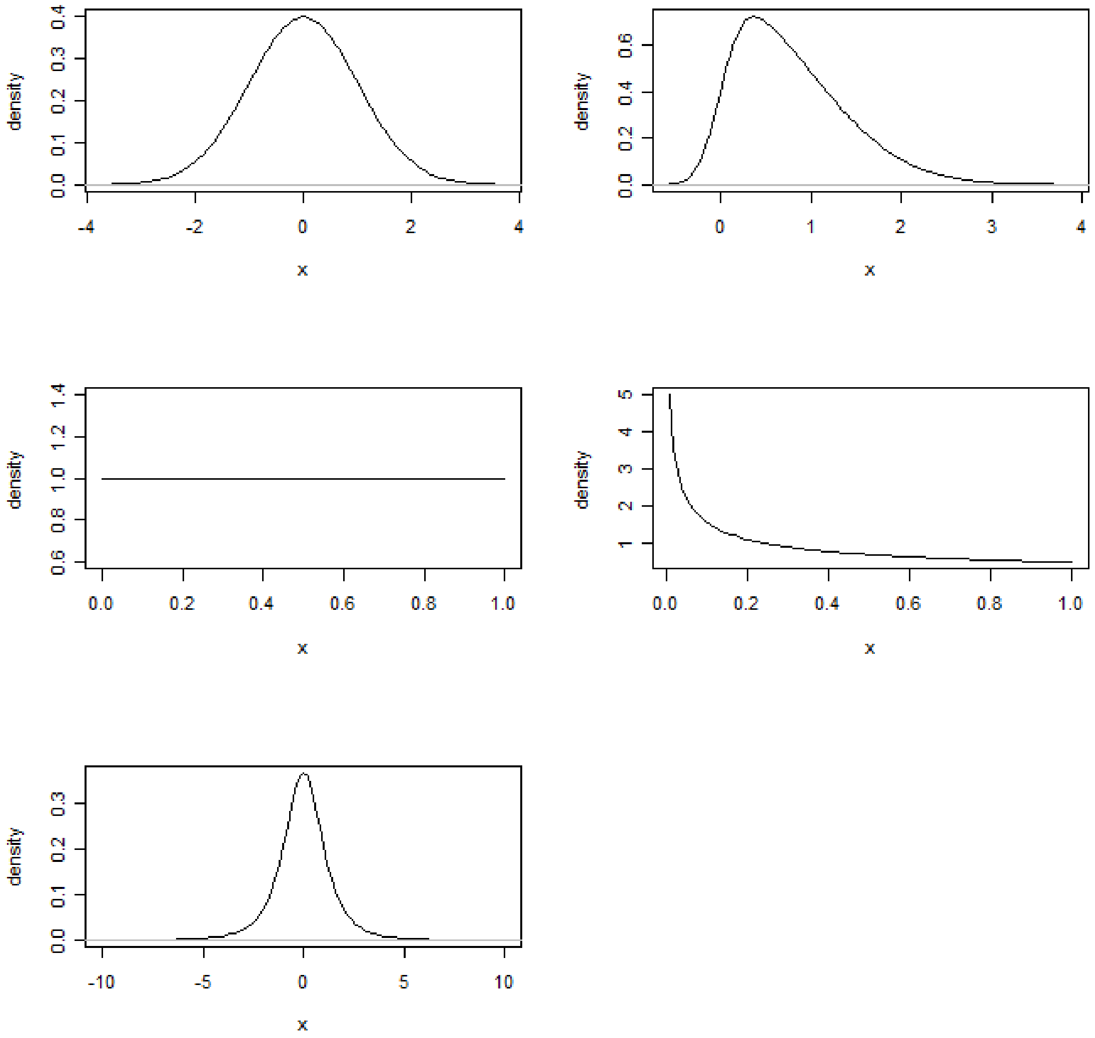

Figure 5.1: Plots of the densities in (5.1), from left to right: Standard normal, Skew-Normal with shape parameter $\lambda = 5$, Uniform, Squared Uniform, and T with $\upsilon = 3$ degrees of freedom.

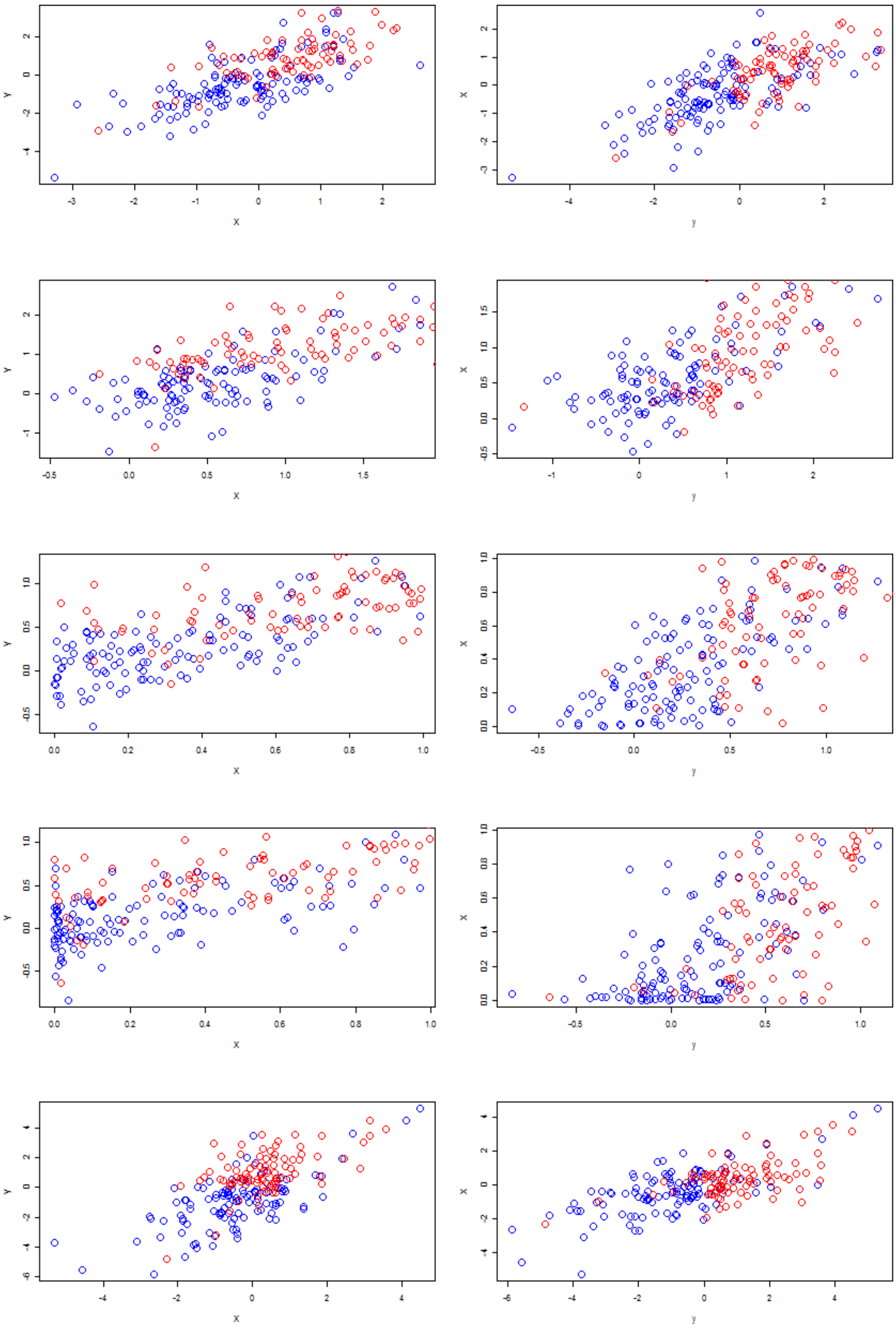

Figure 5.2: Scatter plots of both the direct and reverse regression when the covariate is distributed, from top to bottom, Standard Normal, Skew-Normal with shape parameter $\lambda = 5$, Uniform, Squared Uniform, and T with $\upsilon = 3$ degrees of freedom. The red dots are missing values, and the blue dots are observed.

The results of the simulation study are reported in tables which all share common elements. The "Method" column identifies the mode of inference, where "COM" stands for the complete data analysis; this is the analysis on the complete data, before any cases are deleted, and should be taken as the golden standard for first moment accuracy of the estimator. The complete data analysis should not be confused with the Complete Case Analysis denoted by CCA, which represents the analysis on the completely observed cases. All other entries are multiple imputation inferences using the indicated IM. Of the methods listed in Table 4.1 on page 38, `BaBooN` terminated with an error message. The author of the package was contacted on April 14, 2011 with a detailed report of the error message and the circumstances under which it occurred. On January 17, 2012, the date on which the simulation studies were finalized, the bug was not resolved; therefore, `BaBooN` is dropped from this and all further simulation studies. Further, the output of the package `mi` is suppressed in most scenarios since it performs very similar to `mice`; this comes as no surprise, since the package is a reimplementation of `mice` with some additional post-imputation diagnostic capabilities strapped on.

All IMs are assessed on four criteria:

- Number of simulations which failed due to computational problems, indicated by the column headed by the skull symbol (☠). Although a couple of failed simulations are tolerated, a large number of failures (☠ > 10) is considered problematic and indicative of structural weaknesses in the implementation of the algorithm. Note however that nonparametric techniques often need more observations to function properly.

- Bias, as can be calculated by comparing the third column headed by $\hat{\beta}$ with the true parameter value of 1. First-moment accuracy is a primary requirement of IMs, and is indicative of its ability to "track the available data". The aggregated parameter estimate is calculated as the average of the 1000 simulations of the parameter estimates; note that the parameter estimates in the case of MI inference are aggregates of the IDAs.

- Coverage, as indicated by the fifth column denoted by $COV(\hat{\beta})$. This column gives the proportion of replications where the true parameter lays inside the 95% confidence interval as produced by the Multiple Imputation Inference (MII); under-coverage occurs when coverages are lower than the nominal level of 95%; the performance of a method can be regarded to be poor if its coverage drops below 90% and hence leads to substantially increased Type-I error rates. On the other hand, high coverage rates (approximately $> 97\%$), or over-coverage, may be the result of the overestimation of variances. Correspondingly, the estimates tend to be conservative and thus lead to increased Type-II error rates.

- Efficiency, as indicated by the fourth column denoted by $\hat{SD}(\hat{\beta})$. Although bias and coverage are considered primary requirements, IMs which are unbiased and have nominal coverage might be considered unattractive when they sport large standard errors. The standard errors are calculated by taking the average of the square roots of the estimated variances of the estimator.

	Method	☠	$\hat{\beta}$	$\hat{SE}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.25$	COM	0	0.996	0.123	0.949
	CCA	0	0.865	0.152	0.833
	GLR	0	0.980	0.155	0.949
	PMM	0	0.974	0.154	0.896
	AREGIMPUTE	0	0.974	0.141	0.882
	GAMLSS	2	0.999	0.175	0.950

Table 5.3: Example results table

5.3.2 Multiple

Two simulation experiments were conducted with four jointly independent predictor variables, each having unit variance and associated regression coefficients $\boldsymbol{\beta} = \left[1, \sqrt{2/3}, \sqrt{2/3}, \sqrt{2/3}\right]$, and unit residual variance. Note that also in these experiments, only $\boldsymbol{x}_1$ is afflicted by missing values. Further, R^2 is fixed at .75; the regression coefficients, residual variance, and covariance matrix of the predictor are chosen such that the amount of variance explained by $\boldsymbol{x}_1$ equals $R^2_{\boldsymbol{x}_1} = .50$, which is canonical in the sense that it is the middle value of the set of coefficients of determination $\{.25, .50, .75\}$ in the simple experiments with only $\boldsymbol{x}_1$.

For the studies with multiple predictors, the following MDM is imposed:

$$p(s|lp) = \begin{cases} (\varphi_1)^{1-s}(1-\varphi_1)^s & \text{if } lp < \tilde{lp} \\ (\varphi_2)^{1-s}(1-\varphi_2)^s & \text{if } lp \geqq \tilde{lp}, \end{cases} \tag{5.4}$$

where

$$lp = -.4\boldsymbol{x}_2 - .4\boldsymbol{x}_3 - .4\boldsymbol{x}_4 + .50y \tag{5.5}$$

is an approximation of the reverse linear predictor, and $\tilde{lp}$ is the sample median of lp. Again, setting $\varphi_1 = .1$ and $\varphi_2 = .7$ results in 40% missing data in $\boldsymbol{x}_1$.

5.4 Results

5.4.1 Normal

The first simulation scenario features a normal distribution for the predictor variable x, which implies that x and y are distributed bivariate normal; this is a standard simulation condition for assessing the performance of IMs. Results of the simulation study are presented in Table 5.4 on page 61. Since the missing data are MAR, the CCA is biased, which leads to under-coverage. The under-coverage of CCA seems unaffected by the coefficient of determination, but becomes worse with increasing sample size. The fact that there are better alternatives available than MI in this simple scenario, such as Maximum Likelihood (ML) with incomplete data, does not detract from the requirement that all IMs perform adequate.

Since x and y are distributed bivariate normal, both the direct and reverse regression are linear, and the GLR method is expected to be perfectly adequate; in fact, when the missing values are MAR, the use of GLR is only warranted when the observed data are distributed according to a multivariate normal distribution. This is confirmed in the simulation results, where the GLR is virtually unbiased and has nominal coverage. As is to be expected, the aggregated standard errors are larger than those of the golden standard set by the complete data analysis. This loss of precision is due to the missing values; MI does not make up data.

Given the linearity of the reverse regression, PMM is more flexible than needed in this scenario, and may potentially suffer from the theoretical issues described in 4.2. However, the MDM does not truncate the sample space, although Figure 5.2 on page 56 shows thinning of the sample space for large values of y. Surprisingly, the method suffers from mild to moderate under-coverage, with coverage rates ranging between .892 and .922. With respect to bias, PMM performs roughly equal to GLR, which means very limited empirical bias. On the other hand, the standard errors are slightly smaller than those of the GLR model, which is counter-intuitive since PMM is more flexible and uses less information external to the data than the GLR method; more specifically, it does not use the information that the errors are normally distributed. The unsatisfactory performance of PMM did not arise in the simulation studies of He & Raghunathan (2009), probably because they simulated a MCAR MDM which does not attrite the sample space as selectively as MDM (5.3). Indeed, if the missing values in x are MCAR instead of MAR, and all other scenario parameters are held constant, the coverages rates of PMM as presented in Table 5.5 on page 62 are acceptable and range from .921 to .940.

The performance of `aregImpute` is comparable or slightly worse than that of PMM, with coverages ranging between the .866-.916 interval. A possible explanation is that aregImpute as described in Algorithm 4.1 also performs a predictive mean matching step, and thereby suffers from the same problem as the mice implementation of the PMM algorithm. Although no reference to the poor performance of PMM and aregImpute has been found in the literature, the aregImpute documentation nevertheless states that:

> "When match="closest", predictive mean matching does not work well when fewer than 3 variables are used to predict the target variable, because many of the multiple imputations for an observation will be identical. In the extreme case of one right-hand-side variable and assuming that only monotonic transformations of left and right-side variables are allowed, every bootstrap resample will give predicted values of the target variable that are monotonically related to predicted values from every other bootstrap resample. This causes predictive mean matching to always match on the same donor observation."

This excerpt suggests that the problematic performance of PMM and aregImpute is related to the number of predictors used in the imputation model; if this is the case, then PMM and aregImpute should perform better in the simulation studies with four predictors described in Table 5.6 on page 65. When the IDGP features multiple predictors and with a $R^2_{\boldsymbol{x}_1} = .50$, the coverages of PMM for $\boldsymbol{\beta}_1$ range from .892 to .904, and the coverages of aregImpute range from .909 to .924; although the coverages are slightly better than in the single predictor study, they are still clearly below the nominal level. Another argument against the explanation cited above is that aregImpute was configured not to impute the value of the closest donor, but to perform weighted multinomial probability sampling of the donor values. Moreover, the PMM implementation in mice imputes a value from the nearest five neighbors, each having an equal probability of being selected. Thus, both methods are prevented from matching the same donor observation for every multiple imputation of a missing datum. As the results in Table 5.5 on page 62 show, aregImpute and PMM performs better when the missing values in x are MCAR; this offers support for the explanation given in Section 4.2.

GAMLSS is expected to give unbiased results, albeit with some loss of efficiency compared to GLR. The conditional distribution $\mathcal{D}$ is specified to be normal. Looking at the results in Table 5.4 on page 61 and Table 5.5 on page 62, bias is comparable to that of GLR and thus negligible, although the standard errors are moderately larger than those of GLR; this is the price to pay for the greater flexibility of the model. However, for larger sample sizes, the difference in efficiency diminishes. GAMLSS unfortunately fails to converge in a total of three cases for the lowest sample size condition; however, in the larger sample conditions no problems arise.

Lastly, the performance of GAMLSS in the study with multiple predictors (see Table 5.6 on page 65) is comparable to the study with a single predictor; this indicates that if the additivity assumption is correct, GAMLSS successfully circumvents the curse of dimensionality, at least for a moderate amount of predictor variables.

Table 5.4: Normal distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.25$	COM	0	0.996	0.123	0.949
	CCA	0	0.865	0.152	0.833
	GLR	0	0.980	0.155	0.949
	PMM	0	0.974	0.154	0.896
	AREGIMPUTE	0	0.974	0.141	0.882
	GAMLSS	2	0.999	0.175	0.950
$n = 200 \quad r^2 = 0.50$	COM	0	1.002	0.071	0.951
	CCA	0	0.912	0.092	0.839
	GLR	0	0.995	0.085	0.955
	PMM	0	1.001	0.082	0.905
	AREGIMPUTE	0	0.988	0.080	0.866
	GAMLSS	1	1.008	0.099	0.944
$n = 200 \quad r^2 = 0.75$	COM	0	1.001	0.041	0.964
	CCA	0	0.956	0.056	0.867
	GLR	0	1.001	0.051	0.950
	PMM	0	1.011	0.049	0.892
	AREGIMPUTE	0	0.994	0.048	0.868
	GAMLSS	0	1.006	0.062	0.948
$n = 500 \quad r^2 = 0.25$	COM	0	1.002	0.078	0.940
	CCA	0	0.874	0.096	0.717
	GLR	0	0.993	0.097	0.943
	PMM	0	0.992	0.094	0.903
	AREGIMPUTE	0	0.996	0.088	0.899
	GAMLSS	0	1.003	0.106	0.944
$n = 500 \quad r^2 = 0.50$	COM	0	1.001	0.045	0.941
	CCA	0	0.912	0.058	0.649
	GLR	0	0.998	0.053	0.953
	PMM	0	1.000	0.050	0.912
	AREGIMPUTE	0	0.995	0.050	0.904
	GAMLSS	0	1.005	0.059	0.939

Table 5.4: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 500 \quad r^2 = 0.75$	COM	0	1.000	0.026	0.952
	CCA	0	0.954	0.035	0.732
	GLR	0	0.999	0.032	0.957
	PMM	0	1.005	0.031	0.922
	AREGIMPUTE	0	0.999	0.030	0.881
	GAMLSS	0	1.002	0.035	0.953
$n = 1000 \quad r^2 = 0.25$	COM	0	0.999	0.055	0.947
	CCA	0	0.871	0.068	0.525
	GLR	0	0.994	0.068	0.956
	PMM	0	0.995	0.066	0.910
	AREGIMPUTE	0	0.993	0.062	0.896
	GAMLSS	0	1.000	0.072	0.950
$n = 1000 \quad r^2 = 0.50$	COM	0	1.000	0.032	0.949
	CCA	0	0.913	0.041	0.431
	GLR	0	0.998	0.037	0.941
	PMM	0	1.000	0.035	0.904
	AREGIMPUTE	0	0.998	0.035	0.916
	GAMLSS	0	1.002	0.040	0.948
$n = 1000 \quad r^2 = 0.75$	COM	0	1.000	0.018	0.952
	CCA	0	0.955	0.025	0.544
	GLR	0	1.000	0.023	0.944
	PMM	0	1.003	0.022	0.915
	AREGIMPUTE	0	1.000	0.021	0.901
	GAMLSS	0	1.002	0.024	0.944

Table 5.5: Normal distribution, MCAR

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.250$	COM	0	1.001	0.123	0.942
	CCA	0	1.002	0.159	0.939
	GLR	0	0.996	0.149	0.938
	PMM	0	0.996	0.144	0.921
	AREGIMPUTE	0	0.991	0.139	0.916
	GAMLSS	0	1.011	0.159	0.938

Table 5.5: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.500$	COM	0	1.001	0.071	0.948
	CCA	0	1.002	0.093	0.944
	GLR	0	0.999	0.084	0.957
	PMM	0	1.002	0.080	0.940
	AREGIMPUTE	0	0.997	0.080	0.923
	GAMLSS	0	1.009	0.091	0.951
$n = 200 \quad r^2 = 0.750$	COM	0	1.000	0.041	0.949
	CCA	0	1.002	0.053	0.944
	GLR	0	1.001	0.050	0.938
	PMM	0	1.009	0.048	0.922
	AREGIMPUTE	0	1.001	0.047	0.918
	GAMLSS	0	1.005	0.055	0.942
$n = 500 \quad r^2 = 0.250$	COM	0	0.999	0.077	0.944
	CCA	0	1.001	0.100	0.943
	GLR	0	0.997	0.094	0.942
	PMM	0	0.997	0.090	0.939
	AREGIMPUTE	0	0.995	0.087	0.923
	GAMLSS	0	1.005	0.095	0.947
$n = 500 \quad r^2 = 0.500$	COM	0	1.001	0.045	0.959
	CCA	0	1.000	0.058	0.958
	GLR	0	0.999	0.053	0.956
	PMM	0	1.001	0.050	0.933
	AREGIMPUTE	0	0.998	0.050	0.940
	GAMLSS	0	1.004	0.053	0.951
$n = 500 \quad r^2 = 0.750$	COM	0	0.999	0.026	0.948
	CCA	0	1.000	0.034	0.951
	GLR	0	0.999	0.031	0.950
	PMM	0	1.002	0.030	0.938
	AREGIMPUTE	0	1.000	0.030	0.929
	GAMLSS	0	1.002	0.032	0.950
$n = 1000 \quad r^2 = 0.250$	COM	0	1.002	0.055	0.957
	CCA	0	1.002	0.071	0.956
	GLR	0	1.002	0.066	0.954
	PMM	0	1.002	0.064	0.939
	AREGIMPUTE	0	0.998	0.062	0.933
	GAMLSS	0	1.006	0.066	0.939

Table 5.5: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 1000 \quad r^2 = 0.500$	COM	0	1.001	0.032	0.948
	CCA	0	1.001	0.041	0.945
	GLR	0	1.000	0.037	0.944
	PMM	0	1.001	0.035	0.926
	AREGIMPUTE	0	0.999	0.035	0.931
	GAMLSS	0	1.003	0.037	0.950
$n = 1000 \quad r^2 = 0.750$	COM	0	1.000	0.018	0.938
	CCA	0	1.000	0.024	0.938
	GLR	0	1.000	0.022	0.947
	PMM	0	1.001	0.021	0.937
	AREGIMPUTE	0	1.001	0.021	0.924
	GAMLSS	0	1.001	0.022	0.938

5.4.2 Skew-Normal and Uniform

The second and third simulation study feature a marginal skew-normal and uniform distribution for the predictor variable x, respectively. Judging from Figure 3.1 on page 30, which features a non-linear conditional expectation of x given y and heteroscedastic conditional variance, the GLR method is expected to fail. Indeed, as the results in Table 5.7 on page 66 indicate, the GLR method breaks down with coverages ranging between 0.472 and 0.916. The under-coverage seems primarily due to substantial empirical biases ranging from .051 to .075, which are comparable to those of the CCA. Although PMM and `aregImpute` have negligible bias, their coverage rates are equal to those of the normal study, and remain poor. The performance of PMM and `aregImpute` continues to be substandard irrespective of the conditional distribution of x, and will not be addressed in the discussion of the remaining studies with a single predictor.

For the GAMLSS approach, the conditional distribution $\mathcal{D}$ of x is specified as normal. Since u is simulated from a normal distribution, the assumption that $\mathcal{D}$ is distributed normal corresponds with the assumption that x and y are distributed multivariate normal. However, this implies that the conditional expectation of x given y is linear, which is not true. Despite this logical inconsistency, Figure 3.1 on page 30 shows that the conditional distributions given y are roughly symmetrical in form, and (3.11) implies that only the first two moments need to be correct for consistent MII. Moreover, the Skew-Normal distribution has full support, so there is no need to

	Method	☠	Estimates				Standard Errors				Coverages			
			$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{SE}(\hat{\beta}_1)$	$\hat{SE}(\hat{\beta}_2)$	$\hat{SE}(\hat{\beta}_3)$	$\hat{SE}(\hat{\beta}_4)$	$COV(\hat{\beta}_1)$	$COV(\hat{\beta}_2)$	$COV(\hat{\beta}_3)$	$COV(\hat{\beta}_4)$
n= 200	COM	0	1.001	0.812	0.818	0.814	0.072	0.072	0.072	0.072	0.945	0.953	0.967	0.956
	CCA	0	0.912	0.811	0.818	0.810	0.093	0.090	0.089	0.089	0.844	0.942	0.946	0.952
	GLR	0	0.995	0.813	0.818	0.815	0.086	0.085	0.084	0.084	0.949	0.951	0.940	0.959
	PMM	0	1.005	0.813	0.817	0.814	0.082	0.084	0.084	0.084	0.892	0.947	0.945	0.955
	AREGIMPUTE	0	0.971	0.814	0.817	0.815	0.091	0.086	0.086	0.086	0.924	0.948	0.946	0.961
	GAMLSS	0	1.013	0.813	0.817	0.814	0.112	0.084	0.084	0.084	0.941	0.946	0.940	0.951
n= 500	COM	0	1.001	0.814	0.818	0.817	0.045	0.045	0.045	0.045	0.945	0.937	0.951	0.942
	CCA	0	0.914	0.814	0.816	0.817	0.058	0.055	0.055	0.055	0.672	0.930	0.951	0.958
	GLR	0	0.998	0.814	0.817	0.818	0.053	0.053	0.052	0.052	0.944	0.943	0.948	0.947
	PMM	0	1.002	0.814	0.817	0.818	0.050	0.052	0.052	0.052	0.892	0.943	0.949	0.950
	AREGIMPUTE	0	0.989	0.815	0.817	0.818	0.052	0.053	0.053	0.053	0.909	0.946	0.946	0.950
	GAMLSS	0	1.012	0.814	0.817	0.818	0.061	0.052	0.052	0.052	0.935	0.945	0.935	0.946
n= 1000	COM	0	1.000	0.816	0.817	0.817	0.032	0.032	0.032	0.032	0.958	0.955	0.960	0.953
	CCA	0	0.911	0.814	0.814	0.814	0.041	0.039	0.039	0.039	0.412	0.953	0.950	0.964
	GLR	0	0.999	0.816	0.815	0.816	0.037	0.037	0.037	0.037	0.951	0.947	0.953	0.947
	PMM	0	1.001	0.815	0.815	0.816	0.035	0.037	0.037	0.037	0.904	0.952	0.951	0.942
	AREGIMPUTE	0	0.994	0.816	0.815	0.816	0.036	0.038	0.038	0.038	0.921	0.944	0.950	0.949
	GAMLSS	0	1.007	0.815	0.816	0.816	0.040	0.037	0.037	0.037	0.946	0.950	0.948	0.946

Table 5.6: Multivariate normal distribution

restrict the imputed values via the conditional distribution; all in all, the normal distribution does not seem an unreasonable choice.

Table 5.7: Skew-Normal distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.250$	COM	0	1.004	0.123	0.954
	CCA	0	0.925	0.163	0.892
	GLR	0	1.061	0.170	0.916
	PMM	0	0.984	0.158	0.901
	AREGIMPUTE	0	0.968	0.143	0.868
	GAMLSS	3	0.974	0.201	0.956
$n = 200 \quad r^2 = 0.500$	COM	0	1.000	0.071	0.948
	CCA	0	0.951	0.099	0.896
	GLR	0	1.067	0.092	0.861
	PMM	0	1.003	0.087	0.866
	AREGIMPUTE	0	0.984	0.083	0.863
	GAMLSS	5	1.004	0.119	0.940
$n = 200 \quad r^2 = 0.750$	COM	0	0.999	0.041	0.953
	CCA	0	0.973	0.060	0.911
	GLR	0	1.051	0.055	0.834
	PMM	0	1.020	0.054	0.870
	AREGIMPUTE	0	0.993	0.051	0.841
	GAMLSS	11	1.019	0.068	0.916
$n = 500 \quad r^2 = 0.250$	COM	0	1.000	0.078	0.950
	CCA	0	0.922	0.103	0.851
	GLR	0	1.065	0.105	0.885
	PMM	0	0.991	0.094	0.907
	AREGIMPUTE	0	0.988	0.088	0.881
	GAMLSS	0	0.971	0.122	0.960
$n = 500 \quad r^2 = 0.500$	COM	0	1.000	0.045	0.946
	CCA	0	0.958	0.062	0.873
	GLR	0	1.075	0.057	0.745
	PMM	0	1.008	0.052	0.884
	AREGIMPUTE	0	0.994	0.051	0.873
	GAMLSS	3	1.012	0.068	0.934

Table 5.7: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 500 \quad r^2 = 0.750$	COM	0	1.001	0.026	0.959
	CCA	0	0.977	0.037	0.906
	GLR	0	1.053	0.034	0.694
	PMM	0	1.012	0.033	0.884
	AREGIMPUTE	0	0.999	0.031	0.862
	GAMLSS	9	1.013	0.040	0.928
$n = 1000 \quad r^2 = 0.250$	COM	0	1.000	0.055	0.943
	CCA	0	0.925	0.072	0.796
	GLR	0	1.068	0.074	0.828
	PMM	0	0.998	0.066	0.917
	AREGIMPUTE	0	0.990	0.062	0.882
	GAMLSS	0	0.981	0.085	0.952
$n = 1000 \quad r^2 = 0.500$	COM	0	1.000	0.032	0.949
	CCA	0	0.955	0.044	0.804
	GLR	0	1.074	0.040	0.550
	PMM	0	1.002	0.037	0.885
	AREGIMPUTE	0	0.997	0.036	0.887
	GAMLSS	0	1.006	0.045	0.941
$n = 1000 \quad r^2 = 0.750$	COM	0	1.000	0.018	0.951
	CCA	0	0.977	0.026	0.853
	GLR	0	1.051	0.024	0.472
	PMM	0	1.005	0.024	0.893
	AREGIMPUTE	0	1.001	0.022	0.883
	GAMLSS	0	1.007	0.027	0.950

Since the conditional mean is not restricted to be a linear function of the predictors, and the conditional variance is not restricted to be constant, the GAMLSS approach is expected to offer robust performance in the skew-normal scenario. Generally speaking, these expectations are fulfilled: only the case with $n = 200$ and $r^2 = 0.75$ features slight undercoverage. Since GAMLSS is the only method with adequate coverages, any comparison of confidence interval lengths is meaningless. More troubling are the 11 failed simulations where the `GAMLSS` algorithm failed to converge, which is deemed unacceptable. For $n = 500$ and $r^2 = 0.75$ the coverage is good, but the number of failures is again high. Part of the problem is the automatic method for selecting the bandwidth of the penalized B-spline, as indicated by this except from the documentation of the `GAMLSS` package:

> "Note that the local (or performance iterations) methods can occasionally make the convergence of gamlss less stable compared to models where the degrees of freedom are fixed."

However, as described in Section 4.5, when the penalized B-spline fails to fit, the imputation algorithm falls back to a cubic smoothing spline with a fixed smoothing parameter; even with this fallback mechanism in place, the implementation of the `GAMLSS` estimator suffers from general numerical stability problems, not unlike other user contributed packages (see Chapter 6).

A simulation experiment has also been conducted with $\boldsymbol{x}_1$ distributed skew-normal with a standardized third cumulant of .85, and $\boldsymbol{x}_{-1}$ distributed standard normal, with all predictors jointly independent. The results of the study are presented in Table 5.8 on page 71; GAMLSS performs comparable to the study with a single predictor.

In the case of x having a standard uniform distribution, it may be desirable to restrict the imputed values to lay between zero and one; this can be accomplished by letting $\mathcal{D}$ be the Beta distribution. The more flexible Generalized Beta distribution (Rigby & Stasinopoulos, 2005) is also tested, which is endowed with two additional shape parameters. Finally, we test the normal distribution, denoted by GAMLSS (normal) in the table, even though this choice of $\mathcal{D}$ leads to the imputation of potentially unrealistic values.

Table 5.9: Uniform distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.25$	COM	0	1.002	0.123	0.937
	CCA	0	0.871	0.152	0.855
	GLR	0	0.985	0.155	0.951
	PMM	0	0.992	0.154	0.894
	AREGIMPUTE	0	0.977	0.139	0.880
	GAMLSS (Normal)	12	1.002	0.165	0.950
	GAMLSS (Beta)	0	0.989	0.162	0.936
	GAMLSS (Gen. Beta)	2	1.009	0.164	0.930

Table 5.9: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.50$	COM	0	1.001	0.071	0.960
	CCA	0	0.911	0.092	0.838
	GLR	0	0.997	0.085	0.966
	PMM	0	1.002	0.078	0.931
	AREGIMPUTE	0	0.993	0.077	0.880
	GAMLSS (Normal)	29	1.014	0.094	0.961
	GAMLSS (Beta)	3	1.000	0.088	0.947
	GAMLSS (Gen. Beta)	1	1.001	0.088	0.961
$n = 200 \quad r^2 = 0.75$	COM	0	1.000	0.041	0.944
	CCA	0	0.957	0.056	0.872
	GLR	0	1.006	0.052	0.949
	PMM	0	1.003	0.045	0.903
	AREGIMPUTE	0	0.999	0.045	0.877
	GAMLSS (Normal)	18	1.017	0.054	0.944
	GAMLSS (Beta)	3	1.000	0.049	0.948
	GAMLSS (Gen. Beta)	0	0.997	0.050	0.943
$n = 500 \quad r^2 = 0.25$	COM	0	0.998	0.078	0.955
	CCA	0	0.867	0.096	0.705
	GLR	0	0.988	0.097	0.946
	PMM	0	0.992	0.095	0.913
	AREGIMPUTE	0	0.998	0.087	0.905
	GAMLSS (Normal)	3	1.001	0.101	0.956
	GAMLSS (Beta)	0	0.979	0.101	0.941
	GAMLSS (Gen. Beta)	1	0.995	0.101	0.943
$n = 500 \quad r^2 = 0.50$	COM	0	0.999	0.045	0.947
	CCA	0	0.909	0.058	0.651
	GLR	0	0.997	0.053	0.949
	PMM	0	0.998	0.049	0.900
	AREGIMPUTE	0	0.997	0.049	0.899
	GAMLSS (Normal)	6	1.012	0.057	0.951
	GAMLSS (Beta)	0	0.991	0.054	0.953
	GAMLSS (Gen. Beta)	0	0.993	0.054	0.948

Table 5.9: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 500 \quad r^2 = 0.75$	COM	0	0.997	0.026	0.944
	CCA	0	0.954	0.036	0.740
	GLR	0	1.005	0.033	0.953
	PMM	0	0.998	0.029	0.917
	AREGIMPUTE	0	1.000	0.028	0.910
	GAMLSS (Normal)	8	1.009	0.033	0.958
	GAMLSS (Beta)	1	0.997	0.030	0.954
	GAMLSS (Gen. Beta)	3	0.992	0.031	0.933
$n = 1000 \quad r^2 = 0.25$	COM	0	1.002	0.055	0.959
	CCA	0	0.874	0.068	0.534
	GLR	0	0.996	0.068	0.950
	PMM	0	1.001	0.067	0.920
	AREGIMPUTE	0	0.995	0.062	0.902
	GAMLSS (Normal)	2	1.001	0.071	0.950
	GAMLSS (Beta)	0	0.981	0.071	0.943
	GAMLSS (Gen. Beta)	0	0.998	0.070	0.947
$n = 1000 \quad r^2 = 0.50$	COM	0	0.999	0.032	0.948
	CCA	0	0.909	0.041	0.416
	GLR	0	0.998	0.038	0.956
	PMM	0	0.999	0.035	0.900
	AREGIMPUTE	0	0.997	0.034	0.910
	GAMLSS (Normal)	2	1.009	0.040	0.954
	GAMLSS (Beta)	0	0.993	0.038	0.955
	GAMLSS (Gen. Beta)	0	0.992	0.038	0.943
$n = 1000 \quad r^2 = 0.75$	COM	0	1.000	0.018	0.945
	CCA	0	0.957	0.025	0.582
	GLR	0	1.008	0.023	0.936
	PMM	0	1.001	0.020	0.893
	AREGIMPUTE	0	1.002	0.020	0.922
	GAMLSS (Normal)	1	1.006	0.023	0.943
	GAMLSS (Beta)	0	0.994	0.022	0.932
	GAMLSS (Gen. Beta)	0	0.993	0.022	0.928

The results of the GAMLSS IM in Table 5.9 on page 68 when x has the uniform distribution indicate that imputing under the Normal, Beta, and Generalized Beta distribution gives comparable and adequate results. The Beta distribution seems to capture the features of the conditional distribution quite well, and the additional

	Method	☠	Estimates				Standard Errors				Coverages			
			$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{SE}(\hat{\beta}_1)$	$\hat{SE}(\hat{\beta}_2)$	$\hat{SE}(\hat{\beta}_3)$	$\hat{SE}(\hat{\beta}_4)$	$COV(\hat{\beta}_1)$	$COV(\hat{\beta}_2)$	$COV(\hat{\beta}_3)$	$COV(\hat{\beta}_4)$
n= 200	COM	0	1.000	0.816	0.820	0.816	0.071	0.071	0.071	0.071	0.949	0.956	0.950	0.953
	CCA	0	0.951	0.817	0.819	0.815	0.099	0.087	0.088	0.088	0.907	0.940	0.952	0.956
	GLR	0	1.068	0.817	0.820	0.816	0.092	0.084	0.084	0.084	0.882	0.944	0.963	0.949
	PMM	0	1.011	0.817	0.821	0.817	0.086	0.085	0.086	0.086	0.906	0.941	0.961	0.937
	AREGIMPUTE	0	0.969	0.804	0.808	0.806	0.096	0.088	0.088	0.089	0.919	0.944	0.962	0.946
	GAMLSS	2	0.984	0.810	0.813	0.810	0.145	0.088	0.088	0.089	0.955	0.954	0.958	0.954
n= 500	COM	0	0.998	0.816	0.816	0.817	0.045	0.045	0.045	0.045	0.943	0.937	0.963	0.942
	CCA	0	0.954	0.812	0.813	0.816	0.063	0.055	0.055	0.055	0.871	0.937	0.959	0.957
	GLR	0	1.074	0.815	0.813	0.818	0.057	0.053	0.053	0.053	0.761	0.934	0.966	0.947
	PMM	0	1.005	0.815	0.814	0.819	0.053	0.054	0.054	0.054	0.885	0.938	0.967	0.953
	AREGIMPUTE	0	0.987	0.801	0.800	0.804	0.055	0.055	0.055	0.055	0.898	0.929	0.953	0.945
	GAMLSS	0	0.983	0.807	0.806	0.810	0.080	0.055	0.055	0.056	0.952	0.939	0.967	0.954
n= 1000	COM	0	0.999	0.817	0.816	0.815	0.032	0.032	0.032	0.032	0.948	0.939	0.945	0.945
	CCA	0	0.956	0.816	0.812	0.813	0.044	0.039	0.039	0.039	0.808	0.948	0.947	0.944
	GLR	0	1.073	0.816	0.814	0.814	0.040	0.037	0.037	0.037	0.564	0.946	0.943	0.940
	PMM	0	1.003	0.817	0.814	0.814	0.037	0.038	0.038	0.038	0.900	0.951	0.944	0.945
	AREGIMPUTE	0	0.991	0.802	0.799	0.799	0.038	0.039	0.039	0.039	0.894	0.933	0.936	0.932
	GAMLSS	0	0.984	0.807	0.805	0.805	0.052	0.039	0.039	0.039	0.938	0.950	0.946	0.942

Table 5.8: Multivariate Skew-Normal distribution

flexibility of the Generalized Beta Distribution provided by the two additional parameters seems superfluous. Imputation under the normal model also gives negligible empirical bias, demonstrating that, apart from the first two moments, imputations for x do not need to match the exact shape of the conditional distribution $f(x|y)$. Finally, the good performance of imputations under the Beta distribution demonstrates that the goals of generating plausible imputations and consistent imputations as discussed in Section 3.3.4 are not necessarily incompatible with each other.

To address the stability issues of GAMLSS, some parameters of the optimization algorithm used by the gamlss function have been tweaked; for the remaining simulations, this "enhanced" version is used.

5.4.3 Uniform Squared (Beta)

This simulation study can be interpreted as a assessment of the transform, then impute strategy as described in Section 3.3.4 and von Hippel (2009), where the predictor x is created by taking the square of the original predictor variable z, and where z is standard uniformly distributed. As the scatter plot of the reverse regression in Figure 5.2 on page 56 shows, the conditional expectation of x given y deviates significantly from linearity; moreover, there are outliers visible caused by the skewness of the conditional distribution of x given y for large values of y. Arguably, this scenario features a conditional distribution whose features are difficult to estimate, with the attrition of the MDM further exacerbating the situation. Also note that if the support of z would extend to negative values, the reverse regression approach would break down as discussed in 3.3.5. Given that the values of x lay in the $(0, 1)$ interval, one might want to impute only realistic values; therefore, $\mathcal{D}$ is chooses to be a generalized beta distribution as in 5.4.2.

Table 5.10: Beta distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.250$	COM	0	0.990	0.123	0.942
	CCA	0	0.907	0.162	0.876
	GLR	0	1.039	0.167	0.925
	PMM	0	0.964	0.155	0.902
	AREGIMPUTE	0	0.960	0.139	0.854
	GAMLSS[2] (Gen. Beta)	3	0.976	0.168	0.960
	GAMLSS[2] (Normal)	0	0.972	0.190	0.959

[2] Enhanced

Table 5.10: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.500$	CCA	0	0.968	0.100	0.922
	GLR	0	1.085	0.093	0.824
	PMM	0	0.998	0.081	0.895
	AREGIMPUTE	0	0.985	0.079	0.858
	GAMLSS^2 (Gen. Beta)	3	0.990	0.095	0.947
	GAMLSS^2 (Normal)	0	1.017	0.109	0.934
$n = 200 \quad r^2 = 0.750$	COM	0	0.997	0.041	0.943
	CCA	0	0.992	0.060	0.932
	GLR	0	1.069	0.056	0.779
	PMM	0	1.003	0.049	0.889
	AREGIMPUTE	0	0.993	0.047	0.845
	GAMLSS^2 (Gen. Beta)	0	0.983	0.057	0.934
	GAMLSS^2 (Normal)	0	1.025	0.060	0.944
$n = 500 \quad r^2 = 0.250$	COM	0	1.001	0.078	0.947
	CCA	0	0.922	0.103	0.874
	GLR	0	1.059	0.105	0.882
	PMM	0	0.993	0.096	0.914
	AREGIMPUTE	0	0.990	0.087	0.876
	GAMLSS^2 (Gen. Beta)	1	0.984	0.102	0.951
	GAMLSS^2 (Normal)	0	0.980	0.117	0.948
$n = 500 \quad r^2 = 0.500$	COM	0	1.000	0.045	0.944
	CCA	0	0.971	0.063	0.900
	GLR	0	1.089	0.058	0.661
	PMM	0	1.001	0.050	0.900
	AREGIMPUTE	0	0.997	0.050	0.904
	GAMLSS^2 (Gen. Beta)	2	0.987	0.057	0.943
	GAMLSS^2 (Normal)	0	1.021	0.065	0.933
$n = 500 \quad r^2 = 0.750$	COM	0	0.999	0.026	0.936
	CCA	0	0.992	0.038	0.934
	GLR	0	1.070	0.035	0.504
	PMM	0	1.002	0.031	0.891
	AREGIMPUTE	0	0.999	0.029	0.866
	GAMLSS^2 (Gen. Beta)	0	0.980	0.035	0.906
	GAMLSS^2 (Normal)	0	1.021	0.037	0.938

Table 5.10: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 1000 \quad r^2 = 0.250$	COM	0	0.998	0.055	0.956
	CCA	0	0.918	0.072	0.785
	GLR	0	1.059	0.073	0.866
	PMM	0	0.994	0.067	0.919
	AREGIMPUTE	0	0.990	0.061	0.905
	GAMLSS^2 (Gen. Beta)	1	0.977	0.072	0.946
	GAMLSS^2 (Normal)	0	0.993	0.080	0.940
$n = 1000 \quad r^2 = 0.500$	COM	0	1.001	0.032	0.960
	CCA	0	0.974	0.044	0.897
	GLR	0	1.091	0.040	0.403
	PMM	0	1.001	0.035	0.896
	AREGIMPUTE	0	0.999	0.035	0.909
	GAMLSS^2 (Gen. Beta)	0	0.986	0.040	0.935
	GAMLSS^2 (Normal)	0	1.018	0.045	0.940
$n = 1000 \quad r^2 = 0.750$	COM	0	1.000	0.018	0.947
	CCA	0	0.994	0.027	0.938
	GLR	0	1.071	0.025	0.174
	PMM	0	1.000	0.022	0.905
	AREGIMPUTE	0	1.000	0.021	0.897
	GAMLSS^2 (Gen. Beta)	0	0.980	0.024	0.860
	GAMLSS^2 (Normal)	0	1.011	0.026	0.948

Unfortunately, GAMLSS with a generalized Beta distribution breaks down for high values of the coefficient of determination. When $r^2 = .75$ there is moderate undercoverage, which becomes worse for larger sample sizes. Apparently, for smaller sample sizes, the larger standard errors camouflage the empirical bias. The results might indicate that for this scenario, the IDGP (4.17) is estimated inconsistently. Performance might improve with a varying bandwidth; unfortunately, this feature has not been implemented yet in the `GAMLSS` package. Despite producing potential unrealistic values in the form of negative imputations, GAMLSS with a normal distribution is on target, with the empirical bias being quite acceptable. Therefore, based on these simulation results, it is recommended to specify a conditional normal distribution for continuous variables with missing data when using the GAMLSS imputation method, even if this means that the resulting imputations are unrealistic.

To empirically support the claim that a imputation by reverse linear regression is consistent when the MDM is MCAR, we repeat the previous simulation study under a MCAR mechanism. As predicted, the GLR method performs adequately. However, contrary to what von Hippel (2009) suggests, it is inadvisable to impute x using linear reverse regression for the more general MAR mechanism, as indicated by the results in tables 5.7, 5.9, and 5.10. For larger sample sizes, the empirical bias of GAMLSS with a Normal distribution dominates the bias of GAMLSS with a Generalized Beta distribution.

Table 5.11: Beta distribution - missing values are MCAR

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.250$	COM	0	0.996	0.123	0.954
	CCA	0	0.994	0.159	0.954
	GLR	0	0.987	0.150	0.954
	PMM	0	0.989	0.144	0.940
	AREGIMPUTE	0	0.984	0.138	0.921
	GAMLSS (Gen. Beta)	0	0.994	0.149	0.951
	GAMLSS (Normal)	0	1.012	0.154	0.938
$n = 200 \quad r^2 = 0.500$	COM	0	1.001	0.071	0.945
	CCA	0	1.000	0.092	0.948
	GLR	0	0.997	0.084	0.949
	PMM	0	1.002	0.077	0.936
	AREGIMPUTE	0	0.998	0.077	0.928
	GAMLSS (Gen. Beta)	0	0.994	0.083	0.954
	GAMLSS (Normal)	0	1.021	0.087	0.949
$n = 200 \quad r^2 = 0.750$	COM	0	1.001	0.041	0.951
	CCA	0	1.000	0.053	0.954
	GLR	0	0.999	0.050	0.958
	PMM	0	1.003	0.045	0.938
	AREGIMPUTE	0	1.000	0.045	0.946
	GAMLSS (Gen. Beta)	0	0.989	0.048	0.952
	GAMLSS (Normal)	0	1.020	0.049	0.929
$n = 500 \quad r^2 = 0.250$	COM	0	1.004	0.078	0.946
	CCA	0	1.007	0.100	0.951
	GLR	0	1.004	0.094	0.949
	PMM	0	1.003	0.090	0.944
	AREGIMPUTE	0	1.000	0.087	0.936
	GAMLSS (Gen. Beta)	0	0.999	0.093	0.958
	GAMLSS (Normal)	0	1.004	0.094	0.944

Table 5.11: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 500 \quad r^2 = 0.500$	COM	0	1.000	0.045	0.950
	CCA	0	1.003	0.058	0.945
	GLR	0	1.000	0.052	0.952
	PMM	0	1.002	0.049	0.936
	AREGIMPUTE	0	0.999	0.049	0.941
	GAMLSS (Gen. Beta)	0	0.989	0.051	0.942
	GAMLSS (Normal)	0	1.011	0.053	0.954
$n = 500 \quad r^2 = 0.750$	COM	0	1.000	0.026	0.941
	CCA	0	0.998	0.033	0.938
	GLR	0	0.999	0.031	0.942
	PMM	0	1.000	0.028	0.925
	AREGIMPUTE	0	0.999	0.028	0.931
	GAMLSS (Gen. Beta)	0	0.985	0.030	0.913
	GAMLSS (Normal)	0	1.011	0.031	0.941
$n = 1000 \quad r^2 = 0.250$	COM	0	1.000	0.055	0.954
	CCA	0	1.000	0.071	0.945
	GLR	0	1.000	0.066	0.951
	PMM	0	1.000	0.063	0.932
	AREGIMPUTE	0	0.996	0.061	0.932
	GAMLSS (Gen. Beta)	0	0.991	0.065	0.944
	GAMLSS (Normal)	0	1.001	0.066	0.949
$n = 1000 \quad r^2 = 0.500$	COM	0	0.999	0.032	0.953
	CCA	0	0.997	0.041	0.956
	GLR	0	0.997	0.037	0.958
	PMM	0	0.999	0.034	0.944
	AREGIMPUTE	0	0.996	0.034	0.940
	GAMLSS (Gen. Beta)	0	0.983	0.036	0.923
	GAMLSS (Normal)	0	1.007	0.038	0.952
$n = 1000 \quad r^2 = 0.750$	COM	0	0.999	0.018	0.951
	CCA	0	0.998	0.024	0.955
	GLR	0	0.998	0.022	0.959
	PMM	0	0.999	0.020	0.947
	AREGIMPUTE	0	0.998	0.020	0.945
	GAMLSS (Gen. Beta)	0	0.983	0.021	0.895
	GAMLSS (Normal)	0	1.006	0.022	0.956

5.4.4 Student's T

The fourth simulation study feature a marginal T distribution with three degrees of freedom for the predictor variable x. For the GAMLSS method, $\mathcal{D}$ is specified to be normal. In the simulation study of He & Raghunathan (2009), all tested IMs broke down when the distribution of $\boldsymbol{x}$ was strongly heavy tailed. Unfortunately, as the results in Table 5.12 on page 77 indicate, this is also true for the GAMLSS method, which features biases which are systematically bigger than the GLR method, and coverages rates ranging between .893 and .943. The results of this study suggest that the GAMLSS method, despite its flexibility, is unsuitable for imputation when x has a heavy tailed distributions. While of all methods the coverage rates of GAMLSS are closest to normal, this seems largely due to inflated standard errors.

Table 5.12: T distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.250$	COM	0	1.001	0.129	0.947
	CCA	0	0.892	0.161	0.892
	GLR	0	1.012	0.166	0.933
	PMM	0	0.992	0.168	0.935
	AREGIMPUTE	0	0.962	0.168	0.914
	GAMLSS	0	1.032	0.206	0.939
$n = 200 \quad r^2 = 0.500$	COM	0	0.997	0.075	0.951
	CCA	0	0.917	0.096	0.838
	GLR	0	1.002	0.090	0.904
	PMM	0	1.008	0.096	0.908
	AREGIMPUTE	0	0.976	0.097	0.903
	GAMLSS	2	1.024	0.127	0.924
$n = 200 \quad r^2 = 0.750$	COM	0	1.000	0.044	0.950
	CCA	0	0.950	0.058	0.866
	GLR	0	0.999	0.054	0.888
	PMM	0	1.032	0.060	0.885
	AREGIMPUTE	0	0.986	0.060	0.896
	GAMLSS	1	1.017	0.085	0.911
$n = 500 \quad r^2 = 0.250$	COM	0	0.998	0.081	0.961
	CCA	0	0.889	0.100	0.798
	GLR	0	1.011	0.101	0.932
	PMM	0	0.992	0.102	0.921
	AREGIMPUTE	0	0.970	0.102	0.911
	GAMLSS	1	1.016	0.134	0.943

Table 5.12: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 500 \quad r^2 = 0.500$	COM	0	1.002	0.047	0.954
	CCA	0	0.921	0.059	0.717
	GLR	0	1.006	0.054	0.891
	PMM	0	1.014	0.058	0.902
	AREGIMPUTE	0	0.992	0.057	0.925
	GAMLSS	1	1.020	0.085	0.913
$n = 500 \quad r^2 = 0.750$	COM	0	1.001	0.027	0.952
	CCA	0	0.955	0.035	0.745
	GLR	0	0.999	0.033	0.850
	PMM	0	1.030	0.036	0.824
	AREGIMPUTE	0	0.999	0.035	0.875
	GAMLSS	3	1.013	0.055	0.904
$n = 1000 \quad r^2 = 0.250$	COM	0	1.000	0.056	0.950
	CCA	0	0.893	0.069	0.649
	GLR	0	1.013	0.070	0.912
	PMM	0	1.000	0.069	0.878
	AREGIMPUTE	0	0.984	0.070	0.891
	GAMLSS	8	1.013	0.097	0.940
$n = 1000 \quad r^2 = 0.500$	COM	0	1.000	0.032	0.946
	CCA	0	0.926	0.041	0.548
	GLR	0	1.007	0.037	0.848
	PMM	0	1.016	0.040	0.854
	AREGIMPUTE	0	0.996	0.039	0.888
	GAMLSS	4	1.010	0.068	0.893
$n = 1000 \quad r^2 = 0.750$	COM	0	0.999	0.019	0.945
	CCA	0	0.955	0.025	0.546
	GLR	0	0.995	0.023	0.812
	PMM	0	1.023	0.025	0.797
	AREGIMPUTE	0	0.998	0.024	0.890
	GAMLSS	10	1.003	0.042	0.923

5.4.5 Bernoulli

For the simulation study where x is binary with a Bernoulli distribution, the coefficient of determination is not manipulated, but held fixed at .25; rather, the proportion p of "successes" as defined in (5.2) is manipulated, with $p \in \{.1, .5, .9\}$.

Instead of GLR, which is ill suited for binary data, the Generalized Global Linear Regression (GGLR) implementation of `mice` is tested, which features a Generalized Linear Model (GLM) with logit link. GGLR is expected to perform good, since the true conditional distribution $f(x|y)$ is the Bernoulli distribution with the logistic function linking the linear predictor to the parameter p (see Efron, 1975). As noted in Section 4.2, PMM can be used without modification for the imputation of binary data.

The results of this simulation study are summarized in Table 5.13 on page 79, and confirm that GGLR has good performance. The performance of GAMLSS is acceptable except when $p = .1$ for small sample sizes, as apparent by the high number of failed simulations, bias, and under coverage. This is because the relative small number of "successes", which is further reduced by the MDM since it creates more missing values when y is large, can lead to the well known problem of perfect separation where $x = 1$ always when $y > c$. The `mice` implementation safeguards against this by placing a mildly informative prior on the regression coefficients, which shrinks the estimates and preventing them from going "off the chart"; this fix can also be incorporated into the GAMLSS imputation method by augmenting the data set with data points which represent a sufficiently informative prior. This phenomenon illustrates the computational problems that need to be dealt with due to the demand of executing statistical methods without user supervision.

Table 5.13: Bernoulli distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad p = 0.1$	COM	0	1.002	0.124	0.950
	CCA	0	1.119	0.193	0.847
	GGLR	0	0.955	0.159	0.974
	PMM	0	0.968	0.157	0.911
	AREGIMPUTE	0	0.969	0.142	0.872
	GAMLSS^2	159	0.913	0.192	0.906
$n = 200 \quad p = 0.5$	COM	0	1.001	0.071	0.939
	CCA	0	0.907	0.094	0.827
	GGLR	0	0.993	0.083	0.957
	PMM	0	1.000	0.078	0.935
	AREGIMPUTE	0	0.990	0.075	0.916
	GAMLSS^2	0	0.999	0.085	0.952

Table 5.13: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad p = 0.9$	COM	0	0.996	0.125	0.937
	CCA	0	0.811	0.128	0.694
	GGLR	0	0.965	0.148	0.957
	PMM	0	1.000	0.134	0.933
	AREGIMPUTE	0	0.941	0.160	0.956
	GAMLSS^2	0	0.973	0.166	0.953
$n = 500 \quad p = 0.1$	COM	0	1.004	0.078	0.953
	CCA	0	1.118	0.124	0.785
	GGLR	0	0.984	0.099	0.964
	PMM	0	0.992	0.094	0.887
	AREGIMPUTE	0	0.993	0.089	0.882
	GAMLSS^2	0	0.964	0.129	0.963
$n = 500 \quad p = 0.5$	COM	0	0.998	0.045	0.956
	CCA	0	0.903	0.059	0.626
	GGLR	0	0.995	0.051	0.948
	PMM	0	0.998	0.049	0.930
	AREGIMPUTE	0	0.994	0.047	0.916
	GAMLSS^2	0	0.996	0.053	0.946
$n = 500 \quad p = 0.9$	COM	0	0.998	0.078	0.935
	CCA	0	0.810	0.080	0.359
	GGLR	0	0.987	0.089	0.944
	PMM	0	1.001	0.084	0.922
	AREGIMPUTE	0	0.944	0.101	0.934
	GAMLSS^2	0	0.991	0.097	0.946
$n = 1000 \quad p = 0.1$	COM	0	0.997	0.055	0.955
	CCA	0	1.117	0.087	0.679
	GGLR	0	0.990	0.069	0.955
	PMM	0	0.996	0.065	0.896
	AREGIMPUTE	0	0.997	0.063	0.876
	GAMLSS^2	0	0.992	0.076	0.954
$n = 1000 \quad p = 0.5$	COM	0	1.002	0.032	0.953
	CCA	0	0.909	0.042	0.416
	GGLR	0	1.000	0.036	0.959
	PMM	0	1.001	0.035	0.931
	AREGIMPUTE	0	0.999	0.034	0.932
	GAMLSS^2	0	1.000	0.037	0.960

Table 5.13: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 1000 \quad p = 0.9$	COM	0	1.000	0.055	0.954
	CCA	0	0.813	0.056	0.080
	GGLR	0	0.996	0.062	0.961
	PMM	0	1.001	0.060	0.935
	AREGIMPUTE	0	0.945	0.071	0.909
	GAMLSS^2	0	0.995	0.070	0.964

5.4.6 Poisson

The Poisson simulation study features the `mi` package, which offers an implementation of the GGLR method where the conditional distribution of $\boldsymbol{x}$ is specified to be a Poisson distribution. `mice` provides a GGLR where $\boldsymbol{x}$ has a conditional categorical distribution; this IDGP is also known as polytomous regression. In 4.1 it was shown that the true conditional distribution of $f(x|y)$ is an under-dispersed count distribution; `mi` is therefore expected to fail. Because `GAMLSS` only implements a Poisson distribution and overdispersed count distributions, $\mathcal{D}$ is choosen to be Normal, which results in the imputation of “unrealistic” values.

The results in Table 5.14 on page 81 show horrendous coverages for the POIS method as implemented in `mi`, which confirms the analysis in 4.1. Also, imputing count data as unordered categorical data can lead to invalid MII, as the bad results for the POLY method as implemented in `mice` show. In contrast, the `GAMLSS` method seems to offers good performance, although the empirical bias for the case when $r^2 = .25$ is somewhat disquieting.

Table 5.14: Poisson distribution

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.250$	COM	0	0.998	0.123	0.962
	CCA	0	0.907	0.162	0.896
	POLY	0	0.547	0.181	0.218
	PMM	0	0.976	0.157	0.915
	AREGIMPUTE	0	0.966	0.143	0.855
	POIS	0	0.860	0.146	0.933
	GAMLSS	0	0.974	0.200	0.973

Table 5.14: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 200 \quad r^2 = 0.500$	COM	0	0.999	0.071	0.951
	CCA	0	0.942	0.098	0.890
	POLY	0	0.568	0.131	0.023
	PMM	0	1.007	0.085	0.875
	AREGIMPUTE	0	0.989	0.083	0.845
	POIS	0	0.749	0.093	0.182
	GAMLSS	0	1.007	0.115	0.942
$n = 200 \quad r^2 = 0.750$	COM	0	1.000	0.041	0.935
	CCA	0	0.965	0.059	0.885
	POLY	0	0.583	0.108	0.002
	PMM	0	1.020	0.053	0.872
	AREGIMPUTE	0	0.991	0.051	0.838
	POIS	0	0.629	0.081	0.003
	GAMLSS	0	1.015	0.068	0.924
$n = 500 \quad r^2 = 0.250$	COM	0	0.999	0.078	0.948
	CCA	0	0.916	0.101	0.853
	POLY	0	0.551	0.112	0.003
	PMM	0	0.993	0.095	0.912
	AREGIMPUTE	0	0.991	0.088	0.886
	POIS	0	0.877	0.089	0.781
	GAMLSS	0	0.978	0.119	0.955
$n = 500 \quad r^2 = 0.500$	COM	0	1.000	0.045	0.962
	CCA	0	0.945	0.061	0.839
	POLY	0	0.568	0.082	0.000
	PMM	0	1.006	0.052	0.889
	AREGIMPUTE	0	0.997	0.051	0.882
	POIS	0	0.754	0.059	0.001
	GAMLSS	0	1.007	0.065	0.952
$n = 500 \quad r^2 = 0.750$	COM	0	1.001	0.026	0.949
	CCA	0	0.969	0.037	0.852
	POLY	0	0.580	0.068	0.000
	PMM	0	1.011	0.033	0.886
	AREGIMPUTE	0	1.000	0.031	0.875
	POIS	0	0.634	0.053	0.000
	GAMLSS	0	1.009	0.038	0.940

Table 5.14: Continuation of table on previous page

	Method	☠	$\hat{\beta}$	$\hat{SD}(\hat{\beta})$	$COV(\hat{\beta})$
$n = 1000 \quad r^2 = 0.250$	COM	0	0.999	0.055	0.948
	CCA	0	0.911	0.072	0.757
	POLY	0	0.546	0.080	0.000
	PMM	0	0.995	0.066	0.887
	AREGIMPUTE	0	0.992	0.062	0.880
	POIS	0	0.877	0.063	0.530
	GAMLSS	0	0.980	0.082	0.956
$n = 1000 \quad r^2 = 0.500$	COM	0	1.000	0.032	0.952
	CCA	0	0.944	0.043	0.744
	POLY	0	0.566	0.057	0.000
	PMM	0	1.004	0.036	0.889
	AREGIMPUTE	0	1.000	0.036	0.904
	POIS	0	0.757	0.041	0.000
	GAMLSS	0	1.005	0.044	0.953
$n = 1000 \quad r^2 = 0.750$	COM	0	1.001	0.018	0.952
	CCA	0	0.969	0.026	0.776
	POLY	0	0.581	0.048	0.000
	PMM	0	1.006	0.023	0.892
	AREGIMPUTE	0	1.002	0.022	0.873
	POIS	0	0.635	0.038	0.000
	GAMLSS	0	1.005	0.026	0.944

Chapter 6

Conclusion & Summary

In 6.1, the five research questions stated in Chapter 1 will be answered. Secondly, 6.2.1 will give recommendations for applied researchers who choose to use MI as a solution to their missing data problem. Finally, suggestions and practical recommendations are given on how to proceed with MI research in 6.2.2.

6.1 Research Questions

Research question 1 When is MI beneficial?
In Chapter 2 it was shown that when the Data Analysis Procedure (DAP) equals the Ordinary Least Squares (OLS) estimator, MI without the specification of external information (in the form of auxiliary predictor variables) is only applicable and purposeful when predictors are afflicted by missing data; imputed values for the response variable should generally be discarded to avoid unnecessary loss of efficiency. Imputing while ignoring the missing data mechanism requires the assumption that the missing data in the predictors are at least MAR, that is, the probability of a missing datum in a predictor should be conditionally independent of the datum itself given the response variable and other predictors. Unfortunately, by its very definition, the validity of the MAR assumption cannot be tested without information external to the data set. Therefore, the robustness of the multiple imputation inference to possible hypothesized violations of the MAR assumption should be investigated by performing a sensitivity analysis. Because an ignorable missing data mechanism is specified by omitting it from the DGP, care must be taken to make this assumption explicit by means of documentation.
Although MI seeks to separate the missing data problem from the DAP, the responsibility for the MII ultimately lays with the analyst. Because the imputed data sets give little information about the IDGP assumed by the imputer, it is paramount

that the analyst procures the documentation describing the IDGP, and assesses if the IDGP is compatible with the IDA; this is also necessary when the imputations are generated by the analyst himself using a computer program.

Research question 2 When is an IM compatible with the OLS estimator?
In Chapter 3, consistency of the MI estimator ($\hat{\boldsymbol{\beta}}_{\mathrm{MI}}$) was defined as a necessary condition for compatibility of the IDGP with the IDA. A sufficient condition for consistency of the IDA estimator is that the first two (asymptotic sample) moments of the imputed variable match the first two moments of the variable with missings, and the covariance between the imputed variable and the other variables in the DGP should match the covariance between the variable with missings and the other variables. This condition also applies to transformations of the variable with missing data, and so it is generally incorrect to first impute and then transform variable with missings; instead, transformations of variables with missing data should be imputed separately from the original variable. Finally, there are no statistical objections to imputing "unrealistic" values; plausibility criteria are irrelevant to the validity of the Multiple Imputation Inference.
A natural approach to imputing missing data is the reverse regression IDGP. In the case of missings in predictor variables and a MDM which is MAR, obtaining compatibility of the reverse regression IDGP without making distributional assumptions requires the consistent estimation of a model for the selection indicator conditional on the observed data. A fundamental problem of the reverse regression method arises when the reverse relation is multivalued, in which case the conditional expectation might blur important features of the data. For the imputation of missing values in multiple variables, IMs based on reverse regression can be embedded into the FCS framework, although the statistical properties of this framework are far from fully explored.

Research question 3 Which IMs are currently available, and what are their IDGPs?
The majority of IM implemented in popular FCS implementations are either based on the imputation by reverse regression IDGP, or a variant of PMM. PMM can be seen as a type of random k-nearest-neighbor method with a distance function based on the linear predictor of the reverse linear regression. Since imputed values are "live" or observed, the method can also be used for the imputation of non-continuous data. It is unknown if current implementations of PMM produces imputations which are asymptotically independent over observations, and have the correct conditional distribution. Because PMM can only impute observed values, problems may occur when certain regions of the sample space are sparsely populated.

Research question 4 How can existing IMs be improved?
IMs based on the imputation by reverse regression IDGP, such as the GLR and

GGLR, are parametric regression models and pose restrictions on the functional form of the conditional mean and variance of the variable with missing values. These restrictions may lead to inconsistent estimation of the parameters of the imputation model, and ultimately to invalid MII; therefore, it is expected that IMs which jointly estimate the conditional expectation and conditional variance using non-parametric techniques offer better performance. The proposed GAMLSS method models parameters of a specified distribution $\mathcal{D}$ using additive smoother terms, which in combination with a suitable link allows for easy generalization of the method to discrete and count data. Imputations are drawn from the bootstrap predictive distribution, which is an approximation to the posterior predictive distribution; tuning parameters of the method are selected using cross-validation. To cope with the curse of dimensionality, GAMLSS forces predictors to enter the model additively, and will not capture the effects of potential interactions between the predictors of the imputations model.

Research question 5 How do IMs perform empirically?
Simulation studies have been performed where the ADGP consists of a linear regression model with missing values in a single predictor variable, and a strongly systematic MAR mechanism. Experimental conditions include the marginal distribution of the predictor with missing values, the coefficient of determination, and sample size. Although the PMM and `aregImpute` IMs are virtually unbiased, they suffer from mild to moderate under-coverage in all conducted experiments, including the experiment where all variables are jointly normal distributed. The GLR method performs excellent when the variables are jointly normal distributed, but breaks down when the distribution of the predictor deviates from normality, and the reverse regression becomes non-linear; performance is the worst when the coefficient of determination is high. Although the GGLR method performs well when the true family of conditional distributions of the variable with missing values matches the family of its marginal distribution, as is the case in the experiment with binary data, the method breaks down when this is not the case, such as when the predictor is marginally Poisson distributed. Summarizing, the GLR and GGLR are highly sensitive to misspecification of the imputation model, at least in the conducted simulation experiments.
In contrast, the GAMLSS method features better coverage than currently available methods, although the results are not entirely convincing. More specifically, the experiment where the variable with missings is marginally t-distributed suggests that the method copes poorly with heavy-tailed distributions. Further, the implementation of the method suffers from computational problems, which makes it less suitable for unsupervised application. The method has been tested with up

to four predictors, and performance remains stable. GAMLSS performs best when the conditional distribution is specified to be normal, even if the true conditional distribution is discrete.

The majority view of the multiple imputation community seems to be that the imputation problem has essentially been solved in the context of simple linear regression models. It would therefore seem logical to move forward and expand on the foundation provided by the fully conditional specification framework, and develop marginal imputation methods which are suitable for the imputation of complex data sets. However, one of the major findings of this work is that imputation methods frequently used in practice such as parametric regression models and predictive mean matching exhibit poor performance under seemingly innocuous deviations from “standard” (multivariate normality) simulation conditions; in fact, predictive mean matching also suffered from under-coverage in the experiment where all variables were distributed jointly normal. Thus, it seems premature to consider imputation methods for complex hierarchical models when multiple imputations for a simple data analysis model remains problematic.

Data sets can also be complex because they feature a lot of variables. Because manual specification of the imputation model can become tiresome, an automatic predictor selection problem would be greatly beneficial to the usability of software for imputation. However, implementing such a mechanism would add another layer of complexity to a method which is already quite brittle; although imputation methods based on simple statistical models are generally robust and numerically stable, the simulation experiments show that the estimation of state of the art non-parametric models without user supervision can be problematic. Therefore, I am skeptical that efforts aimed at obtaining robust imputation with automatic variable selection will be fruitful. Moreover, since simulation experiments involving non-parametric imputation methods are already very computationally demanding, rigorous validation of automatic variable selection algorithms using simulation studies might become computationally unfeasible.

Non-parametric imputation methods may not be applicable in small samples; for example, in psychology, sample sizes below 100 are the rule rather than the exception. Without introducing external information by detailed model specification (transformations of predictor variables), imputation methods based on parametric linear regression models are not flexible enough, as shown with theoretical arguments and demonstrated empirically; on the other hand, the flexibility of non-parametric methods comes at the expense of higher sample size requirements.

6.2 Recommendations

6.2.1 MI Users

Based on the presented simulation experiments, and the lack of a solid theoretical foundation, users are advised to refrain from the mechanical application of the imputation methods discussed in this work, including the proposed, experimental methods. Although some imputation software such as `mi` and `mice` offer diagnostic plots of the fit of the imputation model and the generated imputations, it remains to be seen if these experimental diagnostic tools truly facilitate visual assessment of the degree in which imputations are compatible with the data analysis, especially since the compatibility and the plausibility criteria do not necessarily overlap; nevertheless, the end user or data analyst is at all times responsible for the multiple imputation inference! On a positive note, the proposed GAMLSS method does offers better performance in the investigated simulation conditions compared to existing alternatives; a full listing of the implementation, together with bindings for `mice`, is given in Chapter A and Chapter B.

6.2.2 MI Researchers

The missing data problem, and MI in particular, is arguably one of the most challenging and involved sub-field of statistics, certainly when viewed from a mathematical perspective. The complexity of the problem makes it difficult to obtain analytical results. Further, large sample results say little about real world performance. Therefore, during the planning phase of the research project, simulation studies were decided to be the main research instrument for gaining insight into the performance of existing and the proposed imputation method. Since conducting simulation studies requires only a modest amount of knowledge about mathematical statistics, restricting the research method to simulation studies made it feasible to conduct the project by a research group primarily vested in applied statistics.

That simulation studies are no substitute for mathematical proofs was known from the start of the project; however, it became clear that the performance of imputation methods fluctuated strongly with changes in simulation scenario. For example, the first year of the research project saw a lot of experimentation with the local linear regression imputation method; due to insufficient computing power, the coefficient of determination of the data analysis model was not manipulated systematically, but fixed at a relatively high value. In this high signal-to-noise ratio, the method performed excellent; however, when increased computational power allowed for more extensive simulation studies, it was discovered that the method unexpectedly performed much worse for low signal to noise ratios due to overfitting. Further, in previous research, the suitability of imputation methods based on parametric regression models and predictive mean matching was also demonstrated using simulation studies; however, this work shows that those methods tend to break down relatively quickly when the distribution of the variable with missing data deviates from normality.

Although simulation studies are a valuable tool to investigate the finite sample properties of statistical methods, they can be misleading when considered in isolation. Further, since it is impossible to enumerate and simulate all possible scenarios, they tend to give inconclusive results. Therefore, I am convinced that the successful study of multiple imputation requires that the body of evidence for new imputation methods and existing methods currently in use is bolstered with analytical large sample results. Due to the complex nature of the missing data problem, I recommend that candidate researchers in this field posses a master or equivalent degree in Mathematics and/or a PhD in Mathematical Statistics, or are closely supervised by staff with these qualifications; nevertheless, even a researcher with these qualifications will need a a considerable amount of time to become familiar with the subject of missing data.

It is important to realize that simulation experiments involving non-parametric imputation methods require a lot of computational power; these demands should be met at the start of the project. The enlistment of computer scientists has proven to be helpful in engineering a system which distributes simulation studies over several low-cost personal computers. Further, simulation studies concerned with investigating the performance of imputation methods are typically very complex due to the necessity of varying the simulation conditions, managing the complete, incomplete, and imputed data sets, and the pooling of the imputed data analyses; the potential for errors is very high. Therefore, it is highly recommended to review and test any simulation code independently by multiple team members who have a firm understanding of the scenarios to be simulated. Because the simulation studies performed in this work have not been verified by an independent party, they should be replicated using the detailed description in Chapter 5, and the code listing in Chapter A.

Also, the problems associated with developing and testing imputation methods are further exacerbated by a lack of large sample results. Because it is unknown what to expect (in a statistical sense), it is difficult to identify the source of poor performance of an imputation method. On the one hand, there is always the possibility of a programming error; on the other hand, there may be conceptual problems with the imputation method. Analytical results would enable differentiating between these two sources of error, at least for sufficiently large enough samples.

Lastly, the R Project for Statistical Computing has proven to be very useful for the rapid development and testing of imputation methods. Nevertheless, the experienced numerical instability of some of the user contributed packages referenced in this work, and an informal source code review lead me to the conclusion that statistical researchers and data analysts should treat every user contributed package as untested and experimental; apparently, not every good statistician is a good programmer.

Appendix A

Simulation Framework

Below, a source code listing is given of the simulation program used to conduct the experiments described in Chapter 5. Listing 1 contains the main simulation function imp.sim, whose parameters are mostly callback functions which are invoked inside the simulation loop:

- dgp(n) is a function which generate the complete data set, where n is the sample size. The function should return a data frame which contains the complete data set. For example, the function dgp.additive in Listing 2 can be used to simulate data from (2.1).
- .mdm(Data) is a function which implements the Missing Data Mechanism, where Data is a data frame which contains the complete data set. The function should return an observed data indicator matrix (2.19). For example, the function univariate.mdm.mar.threshold in Listing 16 return a function which implements (5.3).
- da(Data) is a function which perform the Data Analysis Procedure, Complete Case Analysis, and Imputed Data Analysis, where Data is a data frame which contains a complete, incomplete, or imputed data set. The function may return anything. For example, the function da.lm in Listing 1 is a memory-efficient implementation of (2.16) and (2.17).
- The imp.methods is a list of imputation functions. Valid imputation functions are of type imp.meth(incomplete, R, m, da), where
  - incomplete is a data frame containing the incomplete data set
  - R is the observed data indicator matrix
  - m is the number of imputations

- da is the IDA; for efficiency reasons, imputed data sets are not retained; instead, imputation methods should themselves perform the IDAs on the imputed data sets and return a list with the results of m IDAs. For example, the create.mice.method in Listing 8 returns an imputation function which wraps an arbitrary marginal imputation method from the mice package.

Imputation Method which draw imputations from a bootstrap predictive posterior distribution (4.19) can be constructed using the function create.pbootstrap.method in Listing 12, which takes a function fit (A,B), where

- A contains a training data frame which should be used to fit the parameters of the imputation model
- B contains a data frame with observations for which predictions should be generated
- The result should be a function pred() which generates predictions (with a random component) for B using the training data in A

The GAMLSS method proposed in 4.5 is an example of a pbootstrap.method, and is implemented in Listing 7.

imp.sim returns a list where each element represents one simulation, and each element is a nested list consisting of the following elements:

- **$**"Before␣Deletion␣Analysis"
- **$**"Incomplete␣Data␣Analysis"
- **$**"Imputed␣Data␣Analysis", which contains a list where each elements contains the result of the corresponding call to the imputation functions in imp. methods.

In addition, the results as described above are saved to a file with path **save.path**.

Statistics can be generated from the output of the imp.sim function using the analysis function in Listing 5.

Listing 1: impsim.r

```
require(mice)
require(roxygen)

imp.sim <- function(
    N = 10,                                             # Number of replications
    n = 100,                                            # Sample size
    m = 5,                                              # Number of imputations
    save.path = "results.dat",                          # Path where to safe results
    dgp = Curry(                                        # Data generating process
        dgp.additive,
        covariates = function(n) NULL,
        predict = function(n) 0,
        error = rnorm
    ),
    .mdm = Curry(                                       # Missing data mechanism
        mdm,
        R.y = runif(n) > .5
    ),
    da = function(Data) {                               # Data analysis
        lm(y~1, Data, model=FALSE)
    },
    imp.methods = list(
        mice.mean = create.mice.method(                 # List of Imputation Functions
            mice.method = mice.impute.mean,
            formula = y ~ 1
        )
    ),
    progress = function(i, N) {
        cat("Simulating␣dataset␣", i, "/", N, "\n", sep="")
        flush.console()
    } )
{   # imp.sim function body
    results <- vector(mode="list", length=N)

    for (i in 1:N) {    # Main simulation loop
        progress(i, N)

        Data <- dgp(n)
        before.deletion.analysis <- tryCatch(
            da(Data),
            error = function(e) { NA }
        )

        R <- .mdm(Data)
        is.na(Data[!R]) <- TRUE
        incomplete.analysis <- tryCatch(
            da(Data),
            error = function(e) { NA }
        )

        imputed.analysis <- tryCatch(
            lapply(
                imp.methods,
                do.call,
                list(
                    incomplete.data = Data,
                    R = R,
                    m = m,
                    da = da
                )
            ),
            error = function(e) { NA }
        )

        results[[i]] <- list(
            "Before␣Deletion␣Analysis" = before.deletion.analysis,
            "Incomplete␣Data␣Analysis" = incomplete.analysis,
            "Imputed␣Data␣Analysis" = imputed.analysis
        )
    }

    save(results, file = save.path)
    invisible(results)
}
```

Listing 2: dgp_additive.r

```
dgp.additive <- function(n, covariates, predictor, error) {
    X <- covariates(n)
    y <- predictor(X) + error(n)

    data.frame(X=X, y=y)
}
```

Listing 3: mdm.r

```
# Formula-centric specification of mdm turned out horrible -- function may be scrapped

mdm <- function(data.before.deletion, ...) {
    R <- matrix(
        TRUE,
        NROW(data.before.deletion), NCOL(data.before.deletion),
        dimnames = list(
            rownames(data.before.deletion),
            paste("R.", colnames(data.before.deletion), sep = "")
        )
    )

    e <- eval(substitute(list(...)), data.before.deletion)
    match.index <- match(names(e), colnames(R), 0)
    R[,match.index] <- unlist(e[match.index != 0])

    return(R)
}
```

Listing 4: da_lm.r

```
da.lm <- function(D, formula) {
    fit <- lm(formula, data.frame(D), model = FALSE)

    result <- sapply(
        c(coefficients=coef, vcov=vcov, df.residual = df.residual), do.call, list(object=fit), USE.
            NAMES = TRUE)

    class(result) <- "lm.lite"
    return(result)
}

vcov.lm.lite <- function(object) {
    return(object$vcov)
}

confint.lm.lite <- function(object, parm, level = 0.95, ...) {
    t.confint(object, parm, level)
}


# default implementation of coef and df.residual suffices
```

Listing 5: analysis.r

```
require(mitools)

analysis <- function(
    simulations,
    statistic = coef,
    imp.aggregator = MIcombine,
    aggregator = mean)
{
    imp.statistic <- Compose(imp.aggregator, statistic)

    statistics.list <- lapply(
        simulations,
        function(sim) {
            cbind(
                statistic(sim$"Before␣Deletion␣Analysis"),
                statistic(sim$"Incomplete␣Data␣Analysis"),
                sapply(sim$"Imputed␣Data␣Analysis", imp.statistic)
            )
        }
    )

    statistics.array <- array(
        unlist(statistics.list),
        dim = c(
                dim(statistics.list[[1]]),
                length(statistics.list)
            )
        )

        apply(statistics.array, c(1, 2), aggregator)
}
```

Listing 6: analysis_coverage.r

```
analysis.coverage <- function(x, true.values) {
        as.numeric(
            (true.values > confint(x)[,1]) & (true.values < confint(x)[,2])
        )
}
```

Listing 7: imp_gamlss.r

```
imp.gamlss.fit <- function(
        Data,
        New.Data,
        formula,
        formula.planb = NULL,
        family=NO()
)
{
        fit <- tryCatch(
        gamlss(
            formula,
            formula,
            formula,
            formula,
            family = family,
            data = Data,
            control = gamlss.control(trace = TRUE),
            method = RS(10)
        ),
        error = function(e) {
            gamlss(
                formula.planb,
                formula.planb,
                family = family,
                data = Data,
                control = gamlss.control(trace = TRUE),
                method = RS(10)
            )
        }
    )

    predictions <- predictAll(fit, New.Data, type="response", data = Data)

    function(model) {
        r <- paste(
                "r",
                fit$family[1],
                sep=""
        )

        predictions$y <- NULL
        do.call(
            r,
            args=c(
                predictions,
                n=length(predictions$mu)
            )
        )
    }
}
```

Listing 8: imp_method_mice.r

```
imp.method.mice <- function(incomplete.data, R, mice.method, formula) {
    formula.no.intercept <- update(formula, ~ . - 1)

    MF <- model.frame(
        formula.no.intercept,
        data = incomplete.data,
        na.action = na.pass
    )
    y <- model.response(MF)
    ry <- !is.na(y)
    x <- model.matrix(formula.no.intercept, MF)

    function(...) {
        mice.method(y, ry, x)
    }
}

create.mice.method <- function(mice.method, formula) {
    Curry(
        imp.method.replicate,
        imp.method = Curry(imp.method.mice, mice.method = mice.method, formula = formula)
    )
}
```

Listing 9: imp_method_aregimpute.r

```
imp.method.aregImpute <- function(incomplete.data, R, m, da, formula) {
    results <- aregImpute(formula, incomplete.data, n.impute = m)

    lapply(
        1:m,
        function(imp) {
            incomplete.data[!R] <- unlist(
                lapply(
                    results$imputed,
                    function(x) {
                        if (!is.null(x))
                            x[,imp]
                    }
                )
            )

            da(incomplete.data)
        }
    )
}
```

Listing 10: imp_method_mice_poly.r

```
imp.method.mice.polr <- function(incomplete.data, R, m, da) {
    incomplete.data[] <- lapply(
        incomplete.data,
        function(x) {
            if (any(is.na(x)))
                factor(x)
            else
                x
        }
    )
    mids <- mice(incomplete.data, m, method = "polr", maxit = 1)

    lapply(
        1:m,
        function(imp) {
            completed.data <- complete(mids, imp)
            completed.data[] <- lapply(completed.data, unclass)
            da(completed.data)
        }
    )
}
```

Listing 11: imp_method_mi.r

```
imp.method.mi <- function(incomplete.data, R, mi.method, formula) {
    function(...) {
        imp <- mi.method(formula, incomplete.data)
        imp@random
    }
}

create.mi.method <- function(mi.method, formula) {
    Curry(
        imp.method.replicate,
        imp.method = Curry(imp.method.mi, mi.method = mi.method, formula = formula)
    )
}
```

Listing 12: imp_method_pbootstrap.r

```
imp.method.pbootstrap <- function(incomplete.data, R, fit) {
    # Imputation using the bootstrap predictive distribution

    select.available.cases <- apply(R, 1, all)
    select.incomplete.variables <- !(apply(R, 2, all))
    available.cases <- subset(incomplete.data, select.available.cases)
    master.predict <- fit(available.cases, available.cases)

    function(...) {
        bootstrap.sample <- available.cases
        bootstrap.sample[,select.incomplete.variables] <- master.predict()

        bootstrap.predict <- fit(
            bootstrap.sample,
            subset(incomplete.data, !select.available.cases)
        )

        bootstrap.predict()
    }
}

```

```
create.pbootstrap.method <- function(fit) {
    Curry(
        imp.method.replicate,
        imp.method = Curry(
            imp.method.pbootstrap,
            fit = fit
        )
    )
}
```

Listing 13: imp_method_replicate.r

```
imp.method.replicate <- function(incomplete.data, R, m, imp.method, da) {
    f <- imp.method(incomplete.data, R)

    lapply(
        1:m,
        function(imp) {
            incomplete.data[!R] <- f(imp)
            da(incomplete.data)
        }
    )
}
```

Listing 14: tconfint.r

```
t.confint <- function(object, parm, level, ...) {
    crit<-qt((1-level)/2, df.residual(object))
    se <- sqrt(diag(vcov(object)))

    cbind(
        coef(object) + crit * se,
        coef(object) - crit * se
    )
}
```

Listing 15: mitools/confint.r

```
df.residual.MIresult <- function(object, ...) {
    object$df
}

confint.MIresult <- function(object, parm, level = 0.95, ...) {
    t.confint(object, parm, level)
}
```

Listing 16: setup.r

```
univariate.dgp <- function(sigma2.epsilon = 1 / R2 - 1, R2 = .5, density = runif) {
    Curry(
        dgp.additive,
        covariates = density,
        predictor = identity,
        error = Curry(
            rnorm,
            sd = sqrt(sigma2.epsilon)
        )
    )
}

univariate.da <- Curry(da.lm, formula = y ~ X)

univariate.mdm.mar.threshold <- function(D, location = median, p = c(.7, .1)) {
    mdm(
        D,
        R.X = bquote(runif(length(X)) > .(p)[as.numeric(y < .(location)(y)) + 1])
    )
}

univariate.mdm.mcar <- function(p = .4) {
    Curry(
        mdm,
        R.X = bquote(runif(length(X)) > .(p))
    )
}

univariate.coverage <- Curry(
    analysis.coverage,
    true.values = c(0, 1)
```

```
)


imp.formula = X ~ y

univariate.mice.pmm.impute <- create.mice.method(mice.impute.pmm, imp.formula)

univariate.mice.norm.impute <- create.mice.method(mice.impute.norm, imp.formula)

univariate.gamlss.impute <- create.pbootstrap.method(
    Curry(
        imp.gamlss.fit,
        family=NO(),
        formula = X~pb(y, control = pb.control(degree = 2, order = 2)),
        formula.planb = X~cs(y, df=1)
    )
)
```

Appendix B

GAMIMP

The function create.mice.gamlss(**formula**, **family**) in Listing B returns a wrapper function for the GAMLSS IM which is compatible with mice:

- **formula** specifies a formula compatible with the gamlss function; see the `gamlss` package documentation for details. Note that the relevant variables should also be selected in the mice predictor matrix.
- **family** specifies a distribution compatible with the gamlss function; see the `gamlss` package documentation for details.

Example:

```
# Data <- data.frame(y, x)
mice.impute.gamlss.y <- create.mice.gamlss(y~x, BI())
mice.impute.gamlss.x <- create.mice.gamlss(x~pb(y), NO())
mice(Data, method = c("gamlss.y", "gamlss.x"))
```

```
create.mice.gamlss <- function(formula, family = NO() ) {
        im <- Curry(
                imp.method.pbootstrap,
                fit = Curry(
                        imp.gamlss.fit,
                        formula = formula,
                        family = family
                )
        )

        wrapper <- function(y, ry, x, ...) {
                im(
                        data.frame(y, x),
                        matrix(
                                c(ry, rep(TRUE, ncol(x) * nrow(x))),
                                nrow(x)
                        )
                ) ()
        }
}
```

Bibliography

Aerts, M., Claeskens, G., Hens, N., & Molenberghs, G. (2002). Local multiple imputation. *Biometrika*, 89(2), 375–388.

Allen, D. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1), 125–127.

Atkeson, C., Moore, A., & Schaal, S. (1997). Locally weighted learning. *Artificial intelligence review*, 11(1), 1173.

Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families. *Scandinavian Journal of Statistics*, 32(2), 159–188.

Birattari, M. & Bontempi, G. (2003). *lazy: Lazy Learning for Local Regression.* R package version 1.2-14.

Bontempi, G., Birattari, M., & Bersini, H. (2000). A model selection approach for local learning. *AI Communications*, 13(1), 4147.

Cameron, A. & Trivedi, P. (2005). *Microeconometrics: Methods And Applications.* Cambridge University Press.

Carpenter, J. R. & Kenward, M. G. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. R. Statist. Soc. A*, 169(3), 571–584.

Dahl, F. (2007). Convergence of random kk-nearest-neighbour imputation. *Computational Statistics & Data Analysis*, 51(12), 5913–5917.

de Jong, R. (2006). Multiple imputation of hierarchical dichotomous data by fully conditional specification. Unpublished manuscript.

De Rond, M. (2005). Publish or Perish: Bane or Boon of Academic Life? *Journal of Management Inquiry*, 14(4), 321–329.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352), 892–898.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science*, 11(2), 89–121.

Gelman, A. (2005). Analysis of variance-Why it is more important than ever. *The Annals of Statistics*, 33(1), 1–53.

Gelman, A., Hill, J., Su, Y.-S., Yajima, M., & Pittau, M. G. (2010). *mi: Missing Data Imputation and Model Checking.* R package version 0.09-11.

Harrell, F. E. (2010). *Hmisc: Harrell Miscellaneous.* R package version 3.8-3.

Harris, I. (1989). Predictive fit for natural exponential families. *Biometrika*, 76(4), 675–684.

He, Y. & Raghunathan, T. (2009). On the Performance of Sequential Regression Multiple Imputation Methods with Non Normal Error Distributions. *Communications in Statistics - Simulation and Computation*, 38(4), 856–883.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.

Jaeger, M. (2005). Ignorability in statistical and probabilistic inference. *Journal of Artificial Intelligence Research*, 24(1), 889–917.

Kleinke, K., de Jong, R., Spiess, M., & Reinecke, J. (2012). Multiple imputation of incomplete ordinary and overdispersed count data. Unpublished manuscript.

Kuchler, C. & Spiess, M. (2009). The data quality concept of accuracy in the context of publicly shared data sets. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 3(1), 67–80.

Li, Q. & Racine, J. S. R. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2), 485–512.

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

Little, R. & An, H. (2004). Robust Likelihood-Based Analysis of Multivariate Data. *Statistica Sinica*, 14, 949–968.

Little, R. & Rubin, D. (2002). *Statistical analysis with missing data.* Wiley.

Liu, J., Gelman, A., Hill, J., & Su, Y.-S. (2012). On the stationary distribution of iterative imputations. Unpublished manuscript.

Luce, R. D. (1997). Several Unresolved Conceptual Problems of Mathematical Psychology. *Journal of Mathematical Psychology*, 87, 79–87.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Taylor and Francis.

Meinfelder, F. (2011). *BaBooN: The Bayesian Bootstrap Predictive Mean Matching Package - Multiple and single imputation for discrete data*. R package version 0.0-4.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, 6(1), 7–24.

Nielsen, S. (1997). Inference and missing data: Asymptotic results. *Scandinavian journal of statistics*, 24(2), 261–274.

Nielsen, S. (2003). Proper and improper multiple imputation. *International Statistical Review*, 71(3), 593–627.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raghunathan, T., Lepkowski, J. M., Hoewyk, J. V., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. 27(1), 85–95.

Rigby, R. A. & Stasinopoulos, D. M. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6(1), 57–65.

Rigby, R. a. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3), 507–554.

Robins, J. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429), 106–121.

Robins, J. & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1), 113–124.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

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

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* Taylor & Francis.

Schafer, J. & Yucel, R. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, (August 2012), 37–41.

Schenker, N. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4), 425–446.

Schenker, N. & Welsh, A. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, 16(4), 1550–1566.

Serfling, R. (2002). *Approximation theorems of mathematical statistics.* Wiley.

Spanos, A. (1995). On normality and the linear regression model. *Econometric Reviews*, 14(2), 195–203.

Tan, Z. (2008). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 94(2), 1–22.

van Buuren, S., Brand, J., Groothuis-Oudshoorn, C., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.

van Buuren, S. & Groothuis-Oudshoorn, K. (2010). *MICE: Multivariate Imputation by Chained Equations in R.* R package version 2.10.

von Hippel, P. T. (2007). Regression with missing ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37(1), 83–117.

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

Yu, K. & Jones, M. C. (2004). Likelihood-Based Local Linear Estimation of the Conditional Variance Function. *Journal of the American Statistical Association*, 99(465), 139–144.

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

Zeger, S. L. & Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42(1), 121–130.