Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets

Jaap P. L. Brand

Acknowledgment

The research described in this thesis was funded by TNO Prevention and Health in Leiden, The Netherlands. This support is gratefully acknowledged.

Brand J.P.L.
Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.
Thesis Erasmus University Rotterdam – with summary in dutch.
ISBN: 90-74479-08-1

Printed by: Print Partners Ispkamp, Enschede

Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets

Ontwikkeling, implementatie en evaluatie van multiple imputatiestrategieën voor de statistische analyse van incomplete gegevensbestanden

Proefschrift

Ter verkrijging van de graad van doctor
Aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Professor dr. P.W.C. Akkermans M.A.
en volgens het besluit van het college van promoties

De openbare verdediging zal plaatsvinden op
donderdag 8 april 1999 om 16.00 uur

door

Jacob Pieter Laurens Brand

geboren te Zaandam

Promotiecommissie

Promotor:	Prof. dr. E.S. Gelsema
Co-promotor:	Dr. S. van Buuren
Overige leden:	Prof. dr. ir. J.D.F. Habbema
	Prof. dr. J. Hox
	Prof. dr. Th. Stijnen

To my parents

CONTENTS

Chapter 1

General introduction

The incompleteness of data sets is a pervasive problem in statistical data analysis. Missing data can have many causes: respondents may be unwilling to answer some questions (item nonresponse) or refuse to participate in a survey (unit nonresponse), transcription errors, dropout in follow-up studies and clinical trials, and joining of two not entirely matching data sets.

Problems associated with missing data are:

1. The sample may not be representative when there is systematic nonresponse. For instance, evidence exists that in sample surveys, low-income or high-income individuals are more likely not to fill in their incomes than middle-income individuals [1], so that the resulting sample may overrepresent the middle incomes;

2. Loss in statistical power. An incomplete data set contains less information about the parameters of interest than the hypothetical complete data set. Consequently, with incomplete data conclusions are less powerful, i.e., standard errors are larger, confidence intervals are wider and p-values are less significant, than in case of complete data;

3. Efficient statistical analysis of incomplete data can be more complicated. When for instance for logistic regression the covariates are incompletely observed, logistic regression cannot be applied directly to the entirely observed data set.

When confronted with incomplete data, researchers usually opt for ad hoc approaches from which listwise deletion and imputation are most popular. In listwise deletion, cases which are

not completely observed are discarded and with imputation each missing data entry is replaced by an estimate of it. In both approaches, the resulting data set is analyzed with the desired statistical method for complete data. An advantage of both approaches is their simplicity and the possibility of applying existing statistical software for complete data. Although both approaches are reasonable in case of a small fraction of incomplete cases, they have serious disadvantages when this fraction is larger. In listwise deletion, the statistical analysis may be biased when complete cases differ systematically from incomplete ones. Moreover, listwise deletion is inefficient in the sense that it may lead to a large potential waste of data; a data reduction of 50% or more is no exception (see also chapter 7).

Imputations can be generated in various ways. A conceptually simple imputation method is mean imputation, in which each missing data entry is replaced by the mean of the observed values of the corresponding variable in the data set. The disadvantages of mean imputation are, that it results in an underestimation of variances and a distortion of relationships between variables, since in this method missing data entries for the same variable of a data set are replaced by the same value. This latter disadvantage is especially serious for multivariate statistical analysis. A more advanced and better imputation method is hot-deck imputation [2], in which for an imputation variable y, each missing data entry y_{mis} is replaced by an observed value of y with a set of observed covariates similar to the observed covariates of y_{mis}. A disadvantage of any imputation method is, that standard errors are underestimated, confidence intervals are too narrow, and p-values are too significant, suggesting a higher precision than in fact can be concluded from the observed data. This is due to the fact that the extra uncertainty due to missing data is not reflected, since the imputed values are treated as if they were fixed, observed values.

Multiple imputation, as proposed by Rubin [1], is the best approach known at this time. With multiple imputation, for each missing data entry of an incomplete data set m likely values based on a statistical model are filled in (imputed). When the statistical model describes the data adequately and the imputations are generated from the predictive distribution of the missing data Y_{mis} given the observed data Y_{obs}, the difference between m imputed values for each missing data entry will reflect the extra uncertainty due to missing data. From the resulting multiply imputed data set containing for each missing data entry m imputed values,

m completed data sets are obtained such that the i-th completed data set is the incomplete data set imputed by the i-th imputations. These m completed data sets are separately analyzed by the desired statistical method for complete data and the m intermediate completed data results are pooled into one final result according to explicit procedures. For each missing data entry, the m imputations can be efficiently stored by linking them as string to the corresponding empty cell, which is facilitated by modern data base techniques such as used in data warehousing. Generally, $m = 5$ imputations are sufficient for valid statistical analysis.

Advantages of multiple imputation are:

- A better statistical validity than can be obtained with ad hoc approaches;
- Multiple imputation is statistically efficient, since the entire observed data set is used in the statistical analysis. Efficiency can be interpreted as the degree to which all information about the parameter of interest available in the data set is used. In clinical trials it is compulsory to only use certificated statistical software and in the future multiple imputation may prove to be the only certified method for statistical analysis of incomplete data;
- Use of commercial statistical software packages for complete data is possible, which is the same advantage as of ad hoc approaches. Such packages are often certified and reliable since they have been extensively tested;
- Multiple imputation saves costs, since for the same statistical power, multiple imputation requires a smaller sample size than listwise deletion;
- Once imputations have been generated by an expert, researchers can use them for their own statistical analyses.

Despite its advantages, multiple imputation has been applied on a small scale only. This is caused by the fact that multiple imputation is laborious, requires expertise, and that so far no standard multiple imputation software is available. This latter observation is especially important, since due to the general availability of powerful computers, statistical analysis is frequently carried out by the applied researcher using standard software, rather than by a

statistician. To make multiple imputation available to a larger group of users, it would be helpful if the technique is implemented in a way that is transparent to end users.

This study describes the development of an interactive system embedding multiple imputation, called the missing data engine, in the client-server based HERMES (HEalth care and Research MEdiating System) Medical Workstation environment [3-5] developed at the Department of Medical Informatics of the Erasmus University Rotterdam, The Netherlands. The main goal of HERMES is the integration of applications and data bases in order to offer clinical users a transparent access to existing data base systems and applications. The possibility to encapsulate existing statistical software packages as autonomous entities and its client-server architecture makes HERMES an attractive environment for the implementation of multiple imputation [6-8].

Multiple imputation, especially when implemented in a missing data engine, can be useful in the following settings:

- **Clinical and epidemiological research:** With an easier access to routinely collected medical data which may be distributed throughout the entire hospital, and an easier use of existing statistical software for analyzing this data, multiple imputation is a powerful additional option. Medical data sets which are retrieved from multiple sources are often incomplete due to errors occurring during data entry or due to joining of two or more possibly not entirely matching data sets. Clinical and epidemiological researchers can use a missing data engine to apply multiple imputation prior to the statistical analysis of a retrieved incomplete data set;

- **Health Surveys:** A health survey is a study in which health aspects of a population are investigated. Well known is the Third National Health and Nutrition Examination Survey (NHANESIII) in which the health and nutritional status of the U.S. population is assessed;

- **Postmarketing surveillance (PMS):** In postmarketing surveillance (PMS), side effects of drugs, often not detected during clinical trials, are reported by GPs, pharmacists, and other health care providers, and stored into a data base. The resulting data set can be statistically analyzed in order to generate and test hypotheses about previously unrecorded

adverse drug reactions. Data sets resulting from PMS often suffer from incompleteness [9], so that a missing data engine can be useful for PMS;

Other settings in which a missing data engine may be useful are:

- **Statistical software companies:** It is to be expected that existing statistical packages will be equipped with a multiple imputation front end;
- **World Wide Web:** It is possible to provide a missing data engine with a login service from the World Wide Web. After logging in, a user can then use it for the imputation of an incomplete data set, or for the statistical analysis of a multiply imputed data set.

For the missing data engine as developed here, a general imputation strategy has been implemented in which for many types of data sets appropriate imputations can be generated. New in this approach is that for each imputation variable (an incomplete variable for which imputations are to be generated), a separate imputation model can be specified, including the specification of a set of predictor variables and an imputation method (linear regression, logistic regression, etc.). The possibility to select predictor variables for each imputation variable makes this approach especially useful for imputation and analysis of large data sets with many variables. The developed imputation methods are validated by means of a simulation study and the implementation of these methods in the missing data engine has been tested by comparing the results with those of a simulation program written on a different platform. The missing data engine has been used in practice in a study conducted at TNO Prevention and Health in Leiden, The Netherlands.

1.1 Thesis outline

In **chapter 2**, the concept behind the three basic types of missing data mechanisms MCAR (Missing Complete At Random), MAR (Missing At Random), and MNAR (Missing Not At Random) is illustrated by means of simple numerical examples. A data set consisting of measurements of blood pressures used as illustration material in this chapter will also be used elsewhere in this thesis.

An overview of existing approaches for the statistical analysis of incomplete data sets is given in **chapter 3**. It is argued that multiple imputation is the best approach currently available. Multiple imputation is described in detail in **chapter 4**. This chapter describes our approach to the generation of imputations for multivariate incomplete data sets and proposes a strategy for the selection of its parameters. This chapter also contains a non-technical explanation of the conditions for proper imputation [1] to be used as a validation criterion for imputation methods.

The validation by means of a simulation study of some of the imputation methods developed in chapter 4 is described in **chapter 5**. **Chapter 6** describes the implementation of multiple imputation in the HERMES medical workstation and **chapter 7** describes the application of the missing data engine to a large study conducted at the TNO Prevention and Health in Leiden.

Bibliography

[1] Rubin DB, Multiple imputation for nonresponse in surveys. Wiley New York, 1987

[2] Ford BL, An overview of hot-deck procedures. In Madow WG, Olkin I, and Rubin DB (Eds.), Incomplete Data in Sample Surveys, Vol. 2, Theory and Biblographies. Academic Press, New York, p 185-207, 1983

[3] Van Mulligen EM, An Architecture for an Integrated Medical Workstation: Its Realization and Evaluation, PhD thesis, Department of Medical Informatics, Erasmus University Rotterdam, The Netherlands, 1993

[4] Van Mulligen EM, Timmers T, Van Bemmel JH, A New Architecture for Integration of Heterogeneous Software Components. Methods of Information in Medicine. 1993: 32-292-301

[5] Van Mulligen EM, Timmers T, Brand JPL, Cornet R, Van den Heuvel F, Kalshoven K, Van Bemmel JH, HERMES a health care workstation integration architecture. International Journal of Bio-Medical Computing, 1994:24:267-275

[6] Van Buuren S, Van Mulligen EM, Brand JPL, Routine multiple imputation in statistical databases. In JC Frenc and H Hinterberger (Eds), Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management, Charlottesville, Virginia, September 28-30, 1994 (pp. 74-78). Los Alamitos IEEE Computer Society Press.

[7] Brand JPL, Van Buuren S, Van Mulligen EM, Timmers T, Gelsema ES, Multiple imputation as a missing data machine. In Ozbolt, J.G. (Ed), Proceedings of the eighteenth annual sym-

posium on Computer Apllications in Medical Care (SCAMC) (pp. 303-307), Philadelphia: Hanley & Belfus, Inc., 1994

[8] Van Buuren S, Van Mulligen EM, Brand JPL, Omgaan met ontbrekende gegevens in statistische databases: Multiple imputatie in HERMES. Kwantitatieve Methoden, 1995 nummer 50, 5-25

[9] Zeelenberg C, Modeling of Health Care for Information Systems Development. In Van Bemmel JH, Musen MA (Eds.), Handbook of Medical Informatics (pp. 357-374), Springer, 1997

Chapter 2

An illustration of MCAR, MAR and MNAR missing data mechanisms by examples

2.1 Introduction

Assumptions about the occurrence of missing data are usually formulated as a missing data mechanism. Essential for the definition of such a mechanism are assumptions about the existence of a hypothetical complete data set which is only partly observed, due to a possibly unknown process. This process is described as a stochastic mechanism, possibly related to the values in the hypothetical complete data set, and determines which data entries are observed and which are not. A missing data entry is defined as one which is not observed, but could have been observed. If, for instance, a physician fails to record the numerical size of a gastric ulcer of a particular patient, this size is a missing data entry. If, however, this size has not been observed since the patient has no ulcer at all, this entry will be called idle rather than missing. Generating imputations for idle data entries makes no sense.

Missing data mechanisms can be divided into the three basic classes Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR):

- **Missing Completely At Random (MCAR):** The most stringent class of missing

data mechanisms is the class denoted as Missing Completely At Random (MCAR). A missing data mechanism is called MCAR, if the probability of each entry to be missing is independent of the values in the hypothetical complete data set. This is equivalent to the assumption that for each variable with missing data, the observed values constitute a random sub-sample of the original complete data set of this variable. An implication of this equivalence is that MCAR is a necessary condition for the validity of complete-case analysis in which only completely observed cases are analyzed. The MCAR assumption, however, is generally too stringent to be realistic in many practical situations, but it may hold when it is plausible that the missing data is solely due to random failures such as transcription errors;

- **Missing At Random (MAR):** If the probability of an entry to be missing possibly depends on observed data but is independent of unobserved data, the underlying missing data mechanism is called Missing At Random (MAR). If there is no dependency on observed nor on unobserved data, one has MCAR as a special case of MAR. The definition of MAR provides a minimal condition on which valid statistical analysis can be performed without modelling the underlying missing data mechanism. Under the assumption of MAR, all information about the missing data, necessary for performing valid statistical analysis, is contained in the observed data, but structured in a way that complicates the analysis. To perform valid statistical analysis, it is necessary to take all observed data into account. Complete case analysis is generally not valid under MAR. Under the MAR assumption it can be detected whether or not the underlying missing data mechanism is MCAR. To establish the stronger condition of MCAR under the MAR assumption, several statistical tests exist [1,2,3];
- **Missing Not At Random (MNAR.):** A missing data mechanism is called Missing Not At Random (MNAR), if the probability of an entry to be missing depends on unobserved data. In this case, unobserved values can be either the unknown value of the missing data entry itself or other unobserved values. While a MCAR missing data mechanism is a special case of MAR, the two classes of MAR and MNAR are disjunct and constitute a partitioning of all possible missing data mechanisms. I.e.,

a missing data mechanism can be either MAR or MNAR but not both. In the case of MNAR, valid statistical analyses cannot be performed without modelling the underlying missing data mechanism. In fact, from the observed data alone, it cannot be detected whether the missing data mechanism is MAR or MNAR. Additional information must be brought to bear. This can be, for instance, an extra random sample among non-respondents, or assumptions about the distribution of the hypothetical complete data. If, on the basis of prior knowledge, it is expected that the complete sample is symmetrically distributed and if an asymmetrical distribution is observed, it may be concluded that the underlying missing data mechanism is MNAR. If it is unknown whether the original complete data is symmetrically distributed, an observed asymmetrical distribution gives no definite answer whether the missing data mechanism is MAR or MNAR.

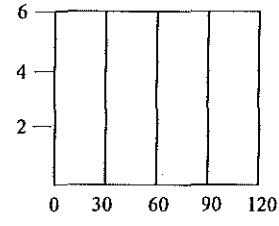
Generally, very little is known about the cause of missing data, so that modelling assumptions are often hard to verify. If the missing data is under control, i.e., if the occurrence of missing data is considered in the study design, the missing data mechanism can typically be assumed to be MAR. Otherwise, the MAR assumption is often not tenable. When little is known about the missing data mechanism, a suitable approach is sensitivity analysis. With sensitivity analysis, the robustness against violation of the MAR assumption is investigated by hypothesizing various plausible MNAR missing data mechanisms and by verifying if incorporation of these mechanisms leads to conclusions different from those expected under MAR.

The distinction between MCAR and MAR is confusing since the term MAR is somewhat in contradiction with its definition. Although the term MAR suggests a random missing data mechanism, it is not random in the sense that the occurrence of missing data may depend on observed values. The MAR assumption is merely a minimal condition for performing valid statistical analysis without having to model the missing data mechanism. It is not easy to imagine what MAR actually means and to understand the real distinction between MAR and MNAR. The main purpose of this chapter is to clarify the definitions of MCAR, MAR as given in [4] and MNAR as given in [5] by means of examples. A brief and simple numerical example is first given. In the second example, the effects of MCAR, MAR and MNAR on descriptive statistics, such as means and standard deviations, are examined for a data set containing

Travelling time in minutes

Complete	Incomplete
43	*
15	15
98	98
12	12
113	113
78	78
9	*
58	58
68	68
100	*
88	88
51	*
29	29
81	81
10	10
105	*
71	*
49	49
51	51
117	117
22	*
115	115
67	*
41	41

Frequency distribution complete:



Frequency distribution incomplete:

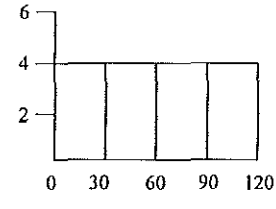


Figure 2-1: An example of a MCAR missing data mechanism.

measurements of blood pressures. This data set is artificially made incomplete by a MCAR, MAR and an MNAR missing data mechanism. Example 2 will also be used elsewhere in this thesis. The formal notation used for the definition of a missing data mechanism is given in section 2.2. The counter-intuitive nature of the MAR definition is illustrated in section 3.

Example 1 *The Figures 2.1, 2.2 and 2.3 are artificial numerical examples of the concepts of MCAR, MAR and MNAR, respectively.*

In the column 'Complete' of Figure 2.1, travelling times in minutes of 24 employees are given. Of these 24 travelling times, only 16 are observed as given in the column 'Incomplete', where the missing entries are marked with an asterisk.. E.g., the first and seventh entries are missing and their true values are 43 and 9. For all 24 entries representing the complete data

Income

Complete	Incomplete
3211	3211
2018	*
9196	9196
10111	*
6318	*
5115	*
4117	4117
2915	2915
9999	*
2762	2762
7134	*
7165	7165
4123	*
2368	*
5965	5965
4127	4127
10976	10976
7432	*
8154	*
3061	*
2511	2511
2163	2163
9162	*
3167	3167

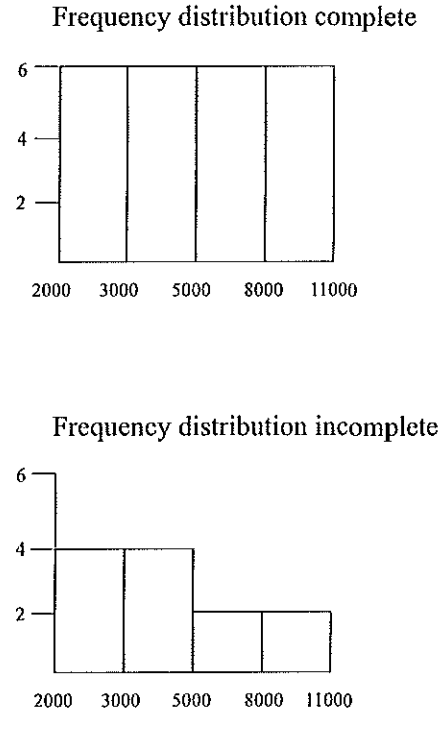


Figure 2-2: An example of an MNAR missing data mechanism.

and for the 16 observed entries representing the incomplete data, frequency distributions for the categories 0-30, 31-60, 61-90 and 91-120, are made. From these two graphs, it appears that both the complete- and the incomplete data are uniformly distributed over the 4 categories with 6 and 4 entries per category, respectively. The missing data entries show independence with respect to the values of the travelling times, so that the missing data in this example may have been generated by a MCAR missing data mechanism.

In the column 'Complete' of Figure 2.2, the incomes in guilders per month of 24 employees are given. Of these incomes, 12 are observed as given in the column 'Incomplete'. Frequency distributions of the complete- and incomplete data are made over the categories 2000-3000,

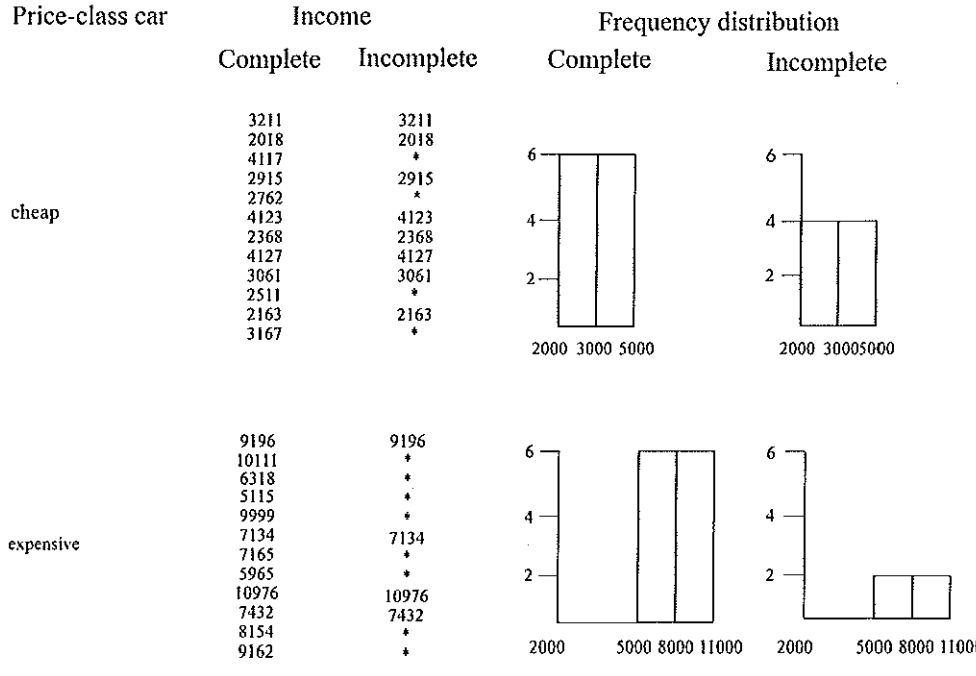


Figure 2-3: An example of a MAR missing data mechanism.

3000-5000, 5000-8000 and 8000-11000. Contrary to the situation in Figure 2.1., the frequency distributions of the complete- and incomplete data are different. While the two lower income categories contain 2 missing data entries per category, the two higher income categories have 4 missing data entries. The occurrence of missing data may be related to the value of income, so that the missing data mechanism here may be MNAR.

In Figure 2.3, the observed and unobserved incomes are given as in Figure 2.2, but an additional variable containing information about the price class of the cars of the employees is also given. The variable 'Price-class car' is observed for all 24 employees and is subdivided into the categories 'cheap' and 'expensive'. The incomes are distributed over these two price classes and for each price class, frequency distributions of the complete and incomplete data have been made. The incomes of employees with a cheap car range from 2000 to 5000 and those of employees with expensive cars from 5000 to 11000. The number of missing incomes

of employees with cheap cars is 2 for both two lower income categories. For employees with expensive cars, this number is 4 for both higher income categories. Thus, within each category of 'Price-class car', the missing data mechanism is MCAR. In this example, the occurrence of missing data depends only on 'Price-class car', so that the underlying missing data mechanism is MAR. The last two numerical examples demonstrate that a missing data mechanism which is initially MNAR may change into MAR by adding a completely observed relevant predictor variable for the missing data entries. Although not a general rule, in practice an MNAR missing data mechanism can often be made closer to MAR by including relevant predictor variables for the incomplete variables.

2.2 Notation for the definition of missing data mechanisms

A data set y may be represented by an $(n \times k)$ rectangular data matrix, where n is the number of cases and k is the number of variables. Let y_{ij} be the observation on the $j-th$ variable in the $i-th$ case of the data matrix y. If some data entries are missing, the data set y can be decomposed into (y_{obs}, y_{mis}, R), where y_{obs} contains the observed values of y and y_{mis} contains the unknown values of the missing data entries of y, and where R is defined by:

$$R_{ij} = \begin{cases} 0 & \text{if } y_{ij} \text{ is not observed} \\ 1 & \text{if } y_{ij} \text{ is observed} \end{cases} \tag{2.1}$$

R is called the response indicator matrix. Thus, R defines the positions of the observed and of the missing data in the data matrix. The structures y_{obs} and y_{mis}, representing the observed and the missing values must be defined in such way that, together with the response indicator R, the complete data matrix y can be reconstructed. One way to define y_{obs} and y_{mis} is illustrated in Figure 2.4.

In Figure 2.4, the data matrix 'incomplete' is the observed part of the data matrix 'complete'. Which entries of 'complete' are observed in 'incomplete' can be read from the response indicator R. For instance, $R_{1,1} = 1$ and $R_{1,2} = 0$ indicate that entry $y_{1,1}$ of 'complete' has been observed and entry $y_{1,2}$ of 'complete' is missing. The row-vectors y_{obs} and y_{mis} are constructed by reading through R from left to right, starting in the upper left corner, and placing the data

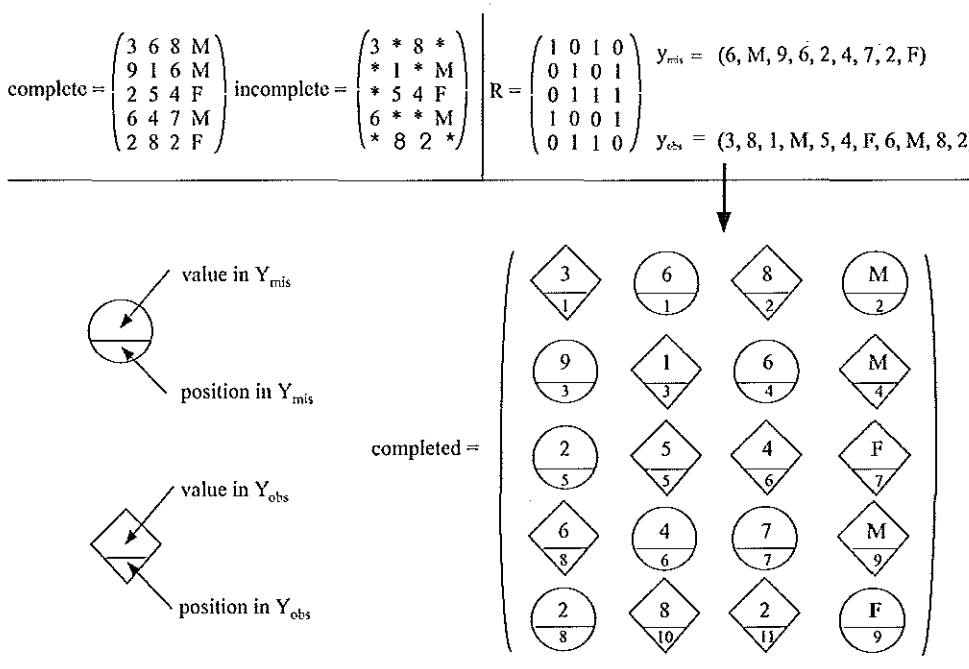


Figure 2-4: One possible definition of y_{obs}, y_{mis} and R.

entry corresponding to the $i-th$ '1' in R into the $i-th$ component of y_{obs} and placing the data entry corresponding to the $i-th$ '0' in R into the $i-th$ component of y_{mis}. E.g., the first and second '1' in R correspond to the two data entries $y_{1,1}$ and $y_{1,3}$, so that the first and second component in y_{obs} are 3 and 8. Similarly, the first and second '0' in R correspond to the two data entries $y_{1,2}$ and $y_{1,4}$, so that the first two components in y_{mis} are 6 and 'M'. The reconstruction of the complete matrix y from y_{mis}, y_{obs} and R is shown by the matrix 'completed'. In this matrix the diamonds and the circles correspond to the observed and the missing data entries, respectively. The upper parts of the diamonds and circles contain the corresponding values of y_{obs} and y_{mis}, while the lower parts contain the corresponding component numbers in y_{obs} and y_{mis}.

Any missing data mechanism can be defined by specifying the conditional distribution $P(R|y)$ of the response indicator given the complete data. Formal definitions of MCAR and

MAR can now be given. A missing data mechanism is MCAR if it can be specified as:

$$P(R|y) = P(R), \tag{2.2}$$

and MAR if it can be specified as:

$$P(R|y) = P(R|y_{obs}). \tag{2.3}$$

In Eq. 2.2, the response indicator R is independent of the complete data y, in Eq. 2.3, the response indicator R depends only on the observed data y_{obs}. Any missing data mechanism with R depending on the values of y_{mis} is an MNAR missing data mechanism.

In Figure 2.1, the fractions of missing data are equal to $\frac{1}{3}$ for each category. In this case the underlying missing data mechanism can be described by

$$P(R_{i1} = 0|y_{i1}) = P(R_{i1} = 0) = \frac{1}{3} \tag{2.4}$$

This missing data mechanism is MCAR since R_i is independent of any value of y. The missing data mechanism as specified in Eq. 2.4 is only a possible missing data mechanism in this case, since from a complete and an incomplete data set the underlying missing data mechanism can only be estimated but not determined. The fractions of missing data for the incomes in Figure 2.2 are $\frac{1}{3}$ for incomes lower than 5000 and $\frac{2}{3}$ for incomes higher than or equal to 5000. The underlying missing data mechanism can be described by

$$P(R_{i1} = 0|y_{i1}) = \begin{cases} \frac{1}{3} & \text{if} \quad y_{i1} < 5000 \\ \frac{2}{3} & \text{if} \quad y_{i1} \geq 5000 \end{cases} \tag{2.5}$$

This missing data mechanism is MNAR, since R_{i1} depends on the value of y_{i1} which is not observed if $R_{i1} = 0$. To describe a missing data mechanism for the situation in Figure 2.3, let y_{i1} and y_{i2} be the income and price class of the car of the $i - th$ employee. The price classes

variable	*N*	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
rs	*1034*	*138.2*	*24.0*	*80*	*240*
rd	*1034*	*79.5*	*13.2*	*40*	*133*
ps	*1034*	*147.7*	*30.6*	*80*	*265*
pd	*1034*	*77.2*	*15.7*	*40*	*147*

Table 2.1: Descriptive statistics for the population.

y_{i2} are completely observed. A missing data mechanism for Figure 2.3 can be described by

$$P(R_{i1}=0|y_{i1},y_{i2}) = P(R_{i1}=0|y_{i2}) = \begin{cases} \frac{1}{3} & \text{if } y_{i2} = \text{'cheap'} \\ \frac{2}{3} & \text{if } y_{i2} = \text{'expensive'} \end{cases} \tag{2.6}$$

The missing data mechanism in Eq. 2.6 is MAR since R_{i1} depends on the value of y_{i2} only and this variable is observed for all cases.

Example 2 *In this example, more extended MCAR, MAR and MNAR missing data mechanisms are described. Their impact on several statistics is analyzed by successively generating three incomplete data sets, one for each of these missing data mechanisms. All three cases start with the same complete data set, in which missing entries are created using Monte Carlo techniques. The complete data set is generated as a random sample of 394 cases without replacement, from a larger data set containing blood pressure measurements of 1034 persons. This larger data set is treated as the population and the sample generated from this population is regarded as the complete data.*

The population is part of a Ph.D.-study in which the safety, side effects and feasibility of the drug Dobutamine as a method to detect coronary artery disease is investigated [6-13]. Dobutamine is an alternative for the usual exercise stress test for patients who are not able to do this test for medical reasons. The population consists of systolic and diastolic blood pressures measured during rest and during a state of maximal effort simulated by the drug Dobutamine. Descriptive statistics for the population of 1034 persons are given in Table 2.1. All values are in mm Hg. Table 2.2 contains the correlation matrix for the population.

In the names of the variables, 'r' refers to the state of rest, 'p' refers to the state of maximal (peak) effort, 's' stand for systolic blood pressure, and 'd' stands for diastolic blood pressure. The

	rs	*rd*	*ps*	*pd*
rs	*1.00*	*0.64*	*0.55*	*0.40*
rd	*0.64*	*1.00*	*0.41*	*0.55*
ps	*0.55*	*0.41*	*1.00*	*0.69*
pd	*0.40*	*0.55*	*0.69*	*1.00*

Table 2.2: Correlation matrix for the population.

variable	*N*	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
rs	*394*	*136.7*	*21.3*	*85*	*200*
rd	*394*	*78.8*	*12.5*	*40*	*111*
ps	*394*	*146.4*	*30.2*	*80*	*265*
pd	*394*	*76.7*	*16.3*	*40*	*147*

Table 2.3: Descriptive statistics for the complete data set (the sample).

correlation between the variables ps and pd is the largest (0.69). The second largest correlation (0.64) is the correlation between the variables rs and rd. The two smallest correlations are those between the variables rs and pd (0.40) and between the variables rd and ps (0.41). This is understandable since these two pairs of variables differ both in physical state and in kind of blood pressure.

Descriptive statistics and the correlation matrix for the sample are given in Tables 2.3 and 2.4, respectively. Figure 2.5 contains a scatterplot matrix for the sample. In the diagonal cells of this matrix, the names of the variables and their corresponding minimum and maximum values are displayed. Each off-diagonal cell contains a scatter plot of the variable in its corresponding row versus the variable in its corresponding column.

The sample size of the complete data set is 394. The differences in mean, standard error, and in the correlation matrix between the sample and the population are due to sampling error. In Figure 2.5, the clouds of points seem to fan out for the larger values. This is most clearly seen in the scatterplot between the variables ps and pd. Due to this heteroscedasticity the data are

	rs	*rd*	*ps*	*pd*
rs	*1.00*	*0.61*	*0.44*	*0.34*
rd	*0.61*	*1.00*	*0.38*	*0.52*
ps	*0.44*	*0.38*	*1.00*	*0.70*
pd	*0.34*	*0.52*	*0.70*	*1.00*

Table 2.4: Correlation matrix for the sample.

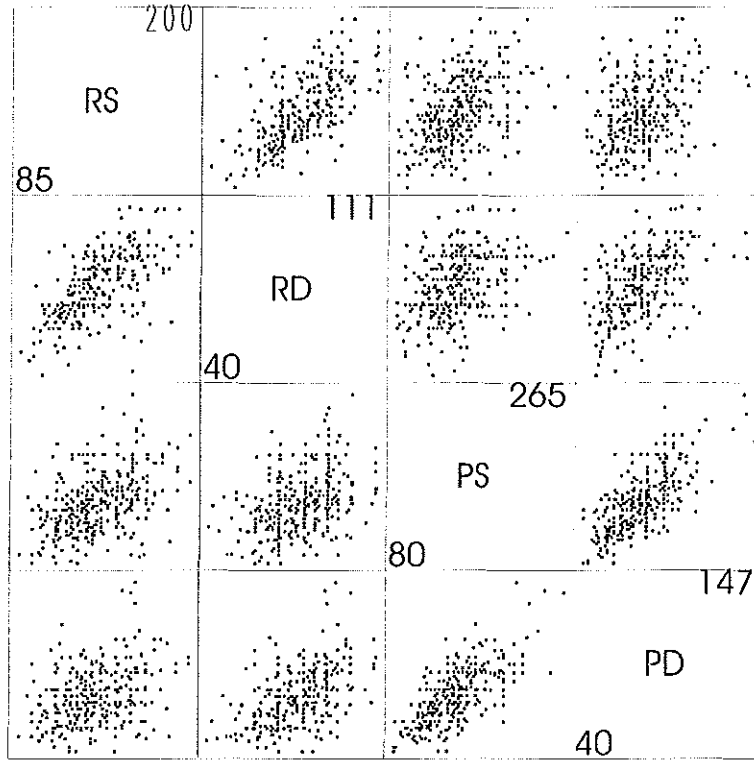


Figure 2-5: Scatterplot matrix of the blood pressures in the sample.

not expected to fit well to a multivariate normal model. Nevertheless, the deviations from this model will be regarded as not too serious.

The three missing data mechanisms to be explored below, are MCAR, MAR and MNAR in relation to the variable ps. With MCAR, the occurrence of missing data in ps is independent of the values of ps and of the other variables. The MAR mechanism as implemented here, generates missing data in ps, the occurrence of which solely depends on the variable pd. By the strong correlation between the variables ps and pd, the occurrence of missing data in ps under MAR depends indirectly on the values of ps itself. The occurrence of missing data generated in ps by the MNAR mechanism implemented here, solely depends on the value of ps itself. All

rs	*rd*	*ps*	*pd*	*% among incomplete cases*
0	*0*	*0*	*1*	*35*
0	*1*	*0*	*1*	*25*
1	*0*	*0*	*1*	*25*
1	*1*	*0*	*1*	*15*

Table 2.5: The four missing data patterns for all three missing data mechanisms.

three missing data mechanisms are designed to generate an average of 45.5% incomplete cases such that Table 2.5 holds true.

In this Table, a '0' indicates that the corresponding variable is not observed and a '1' indicates that it is observed in the corresponding missing data pattern. It is seen that in all four missing data patterns the variable ps is missing and the variable pd is always observed. The percentages in the rightmost column are the expected percentages with which the corresponding patterns appear among all incomplete cases. For instance, the three missing data mechanisms generate missing data in such a way that on the average in 35% of the incomplete cases the variable pd is the only observed variable and in 15% of the incomplete cases ps is the only missing variable.

The MCAR missing data mechanism generates incomplete data by distributing the incomplete cases randomly over all sample cases according to the missing data patterns in table 2.5. The MAR missing data mechanism is designed to satisfy

$$\frac{P(R_{ps} = 0 \mid pd < med(pd))}{P(R_{ps} = 0 \mid pd \geq med(pd))} = 3. \tag{2.7}$$

In Eq. 2.7, R_{ps} is the response indicator for ps and med(pd) is the median of pd. According to Eq. 2.7, the expected fraction of missing data in ps among cases with a value of pd smaller than its median is 3 times larger than the expected fraction of missing data among cases with a value of pd equal to or larger than its median. This missing data mechanism is indeed MAR, since pd is completely observed. However, due to the strong correlation between the variables ps and pd, a larger fraction of missing data is expected in the lower values of ps than in the higher values.

The MNAR missing data mechanism is specified by

$$\frac{P(R_{ps}=0 \mid ps < med(ps))}{P(R_{ps}=0 \mid ps \geq med(ps))} = 3. \tag{2.8}$$

Eq. 2.8 is of the same type as Eq. 2.7, but now the occurrence of missing data in the variable ps depends directly on the value of the variable ps itself.

The algorithm is described in chapter 5 of this thesis. In the Tables 2.6 through 2.8, simple descriptive statistics for the three incomplete data sets are given. The correlation matrices for these data sets are given in the Tables 2.9 through 2.11.

variable	*nobs*	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
rs	*290*	*137.5*	*21.3*	*85*	*200*
rd	*302*	*79.3*	*12.3*	*47*	*111*
ps	*231*	*145.0*	*28.9*	*80*	*227*
pd	*394*	*76.6*	*16.0*	*40*	*147*

Table 2.6: Descriptive statistics of the incomplete data under MCAR.

variable	*nobs*	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
rs	*294*	*140.6*	*22.7*	*85*	*200*
rd	*295*	*81.4*	*12.5*	*47*	*111*
ps	*221*	*153.7*	*31.8*	*80*	*265*
pd	*394*	*82.9*	*15.9*	*40*	*147*

Table 2.7: Descriptive statistics of the incomplete data under MAR.

variable	*nobs*	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*
rs	*279*	*139.4*	*21.6*	*91*	*200*
rd	*285*	*80.6*	*12.3*	*47*	*111*
ps	*220*	*157.0*	*29.8*	*88*	*265*
pd	*394*	*80.80*	*16.6*	*49*	*147*

Table 2.8: Descriptive statistics of the incomplete data under MNAR.

The means for the variable ps in the MCAR, MAR and MNAR incomplete data sets are 145.0, 153.7 and 157.0, respectively. The complete data mean for ps is 146.4. At a first glance, the difference between the mean of the MCAR incomplete data set and the complete data mean (146.4-145.0) seems plausible under the MCAR assumption. The deviations of the means of the MAR and MNAR incomplete data sets from the complete data mean seem too large to be explained by the MCAR assumption. As can be expected on the basis of Eq. 2.7 and Eq. 2.8, these means are shifted upwards. The MCAR assumption can be tested by verifying whether the incomplete observations of ps are a random sub-sample from the complete observations of ps. This can be done by constructing from the incomplete data set a 95% confidence interval for the complete data mean under this assumption and to verify whether the complete data mean is included in this interval. Under the assumption that the number of observed values n_{obs} is large enough, a 95% confidence interval is given by:

$$\overline{y}_{n_{obs}} \pm 1.96 SE(\overline{y}_{n_{obs}}) \text{ with,} \tag{2.9}$$

$$SE(\overline{y}_{n_{obs}}) = \left(\sqrt{\frac{1}{n_{obs}} \left(\frac{n - n_{obs}}{n} \right)} \right) S \tag{2.10}$$

In Eq. 2.9, $\overline{y}_{n_{obs}}$ is the incomplete data mean, $SE(\overline{y}_{n_{obs}})$ the standard error of the incomplete data mean, and 1.96 the 0.975 quantile of the standard normal distribution. In Eq. 2.10, S is the standard deviation of the observed ps, and n is the sample size of the complete data. The factor $\left(\frac{n-n_{obs}}{n}\right)$ in the standard error $SE(\overline{y}_{n_{obs}})$ is a correction factor for the finite size n of the complete data from which the incomplete data is drawn as a subsample. For the variable ps in the MCAR data set, the following values apply: $\overline{y}_{nobs} = 145.0$, $n_{obs} = 231$, $S = 28.9$, $n = 394$. The standard error for the incomplete data mean is therefore $SE = \sqrt{1/231 - 1/394} * 28.91 = 1.223$, so that (142.6,147.4) is a confidence interval for the complete data mean. The complete data mean given by 146.40 is included in this interval, so that the results in Table 2.6 are compatible with the MCAR assumption. In the same way it may be verified that the observed values of ps in the other two cases are incompatible with the MCAR assumption.

	rs	*rd*	*ps*	*pd*
rs	*1.00*	*0.62*	*0.42*	*0.33*
rd	*0.62*	*1.00*	*0.32*	*0.49*
ps	*0.42*	*0.32*	*1.00*	*0.67*
pd	*0.33*	*0.49*	*0.67*	*1.00*

Table 2.9: Correlation matrix for the incomplete data under MCAR.

	rs	*rd*	*ps*	*pd*
rs	*1.00*	*0.67*	*0.42*	*0.32*
rd	*0.67*	*1.00*	*0.38*	*0.51*
ps	*0.42*	*0.38*	*1.00*	*0.71*
pd	*0.32*	*0.51*	*0.71*	*1.00*

Table 2.10: Correlation matrix for the incomplete data under MAR.

	rs	*rd*	*ps*	*pd*
rs	*1.00*	*0.61*	*0.39*	*0.28*
rd	*0.61*	*1.00*	*0.34*	*0.48*
ps	*0.39*	*0.34*	*1.00*	*0.67*
pd	*0.28*	*0.48*	*0.67*	*1.00*

Table 2.11: Correlation matrix for the incomplete data under MNAR.

Comparing the correlation matrices of the incomplete data sets (Tables 2.9 through 2.11) with that of the complete data (Table 2.4), no systematic differences are observed. Contrary to the mean, the correlations seem to be robust under all missing data mechanisms. A closer look at the probability density distributions of ps under the three missing data mechanisms will now be given. This can be done by kernel estimators of the following shape:

$$\hat{f}_{nh}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right). \tag{2.11}$$

In Eq. 2.11, $y_1, \ldots, y_n$ is a random sample of size n. The function $K(.)$ can be any density

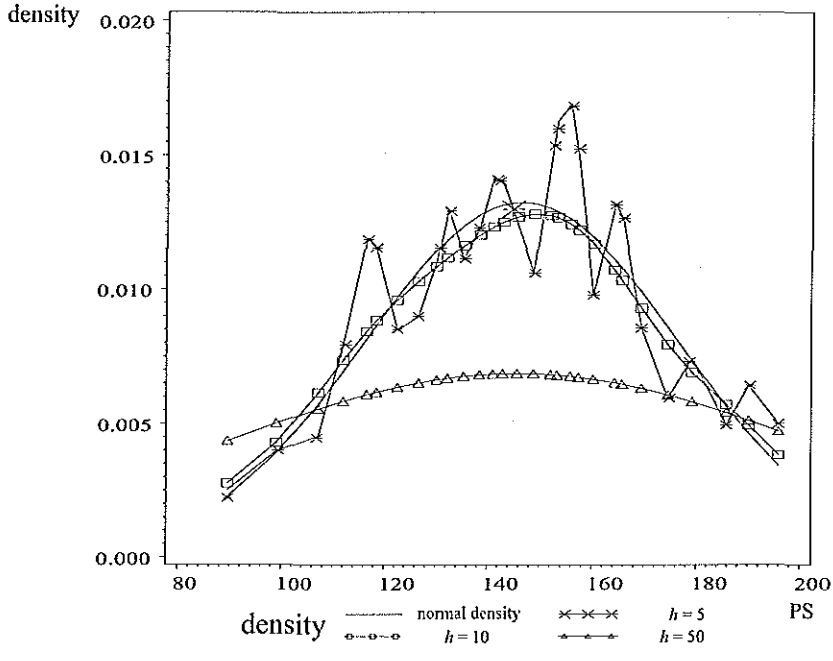


Figure 2-6:

function and is called the kernel. Here the standard normal density, given by $K(x) = \frac{1}{2\pi}e^{-x^2/2}$ *is chosen as the kernel. The quantity h is the so-called bandwidth, which, in principle, is a free parameter. In Figure 2.6, the dependence of the estimated density distribution on the parameter h is illustrated.*

In Figure 2.6, three sample estimates of a univariate normal density distribution are displayed. The sample was generated so as to have the same mean and variance as the variable ps in the complete data studied here. The theoretical density distribution is displayed in the graph without symbols. Density distributions estimated using kernels with bandwidths of 5, 10 and 50 units are shown. It is seen that the kernel estimate with bandwidth 50 is too flat and that the kernel estimate with bandwidth 5 is too unstable. The kernel estimate with bandwidth 10 approximates the theoretical distribution satisfactorily.

In Figure 2.7, the graph without symbols represents a kernel estimate of the density distribu-

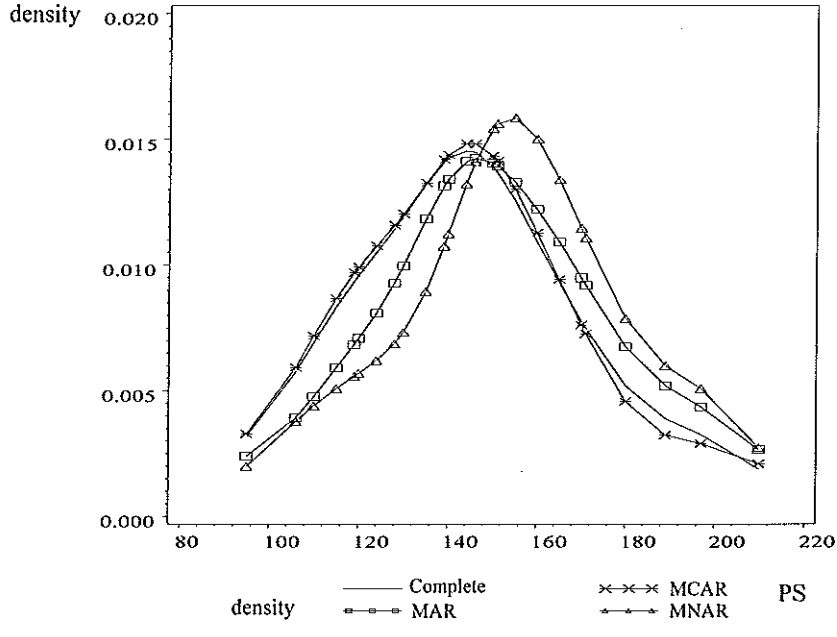


Figure 2-7: Kernal estimates ($h = 10$) of the probability distribution of ps for the complete data set and for the incomplete data sets generated under MCAR, MAR and MNAR.

tion of ps as estimated from the complete data. The kernel estimates of the density distributions of ps in the incomplete samples generated under the MCAR, MAR and MNAR missing data mechanisms are represented by the graphs with stars, squares and triangles, respectively. The graphs for the complete data and for the incomplete data under MCAR show no systematic differences. The observed differences between these two graphs reflect random variation. The estimated density distributions of the incomplete samples generated under the MAR and MNAR missing data mechanisms are incompatible with the complete sample estimate.

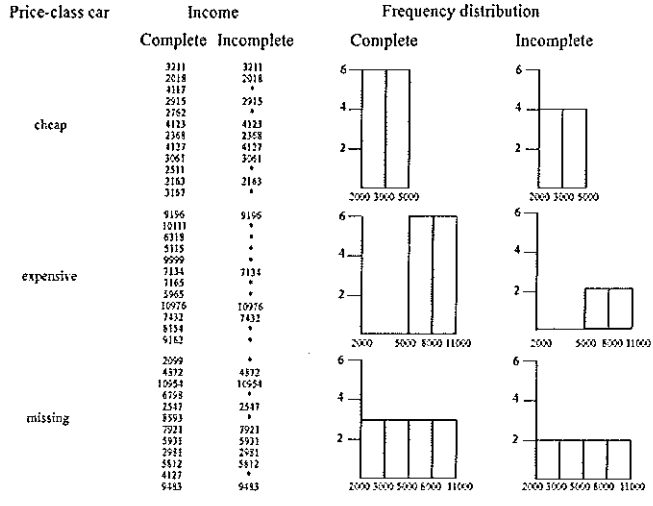


Figure 2-8: An example of a MAR missing data mechanism in which the probability of a missing data entry depends on incomplete covariates.

2.3 A numerical example of a counter-intuitive MAR missing data mechanism.

In the MAR missing data mechanisms in the two previous examples, the probability of occurrence of a missing entry in a variable depends on fully observed covariates. It is possible to construct MAR missing data mechanisms in which this probability depends on covariates which are not fully observed. However, such a missing data mechanism is generally not realistic, since it requires that the probability of an entry in a variable y to be missing could depend on another variable x when and only when x is observed [14]. An example is given in Figure 2.8.

In Figure 2.8, the example of incomes and price classes of cars of Figure 2.3 is extended by 12 extra values of incomes for which the price class of the car is not observed. The first 24 cases for which the price class is observed are the same as in Figure 2.3. For a MAR missing data mechanism it is required that the occurrence of missing data is independent of unobserved values. Consequently, for all employees for whom the price class of the car is missing, the missing incomes are a random sub-sample among the incomes of these employees. In Figure

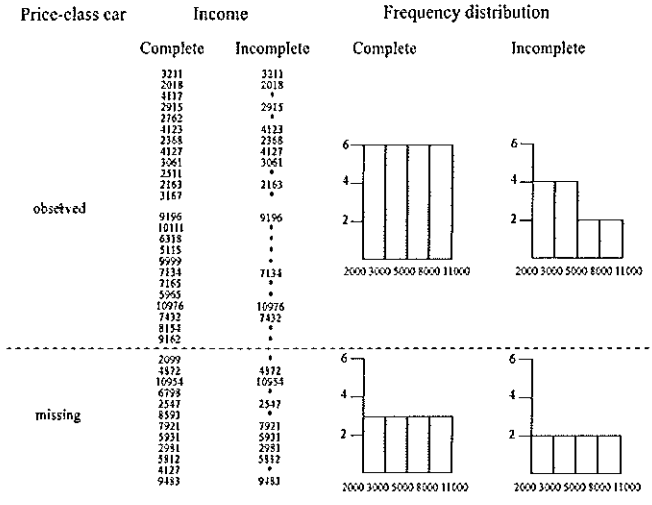


Figure 2-9: Frequency distributions for the complete and incomplete incomes of the employees with observed and missing values for the class of the car.

2.8, this is illustrated by the two frequency distributions where in each income category one income is missing.

In Figure 2.9, the frequency distributions of the complete and incomplete incomes are separately given for the employees with an observed and a missing value for the price class of the car. It appears that among employees with an observed value for the price class, the occurrence of missing data highly depends on the income, whereas among employees for which this price class is not observed, the missing data is independent of income. This is not a very realistic model.

Bibliography

[1] Diggle PJ, Testing for Random Dropouts in Repeated Measurement Data. Biometrics, Vol. 45, 1989:1255-1258

[2] Little RJA, A Test of Missing Completely At Random for Multivariate Data With Missing Values. Journal of the American Statistical Association, Vol. 83, No. 404, 1988

[3] Simonoff JS, Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression. Technometrics, Vol. 30, No. 2, 1988:205-214

[4] Rubin DB, Inference and missing data (with discussion). Biometrica 63, 1976:581-592

[5] Robins JM, Gill RD, Non-response models for the analysis of non-monotone ignorable missing data, Statistics in Medicine, Vol. 16, 39-56, 1997

[6] Poldermans D, Fioretti PM, Dobutamine-atropine stress echocardiography in patients with suspected or proven coronary artery disease: experience in 650 consecutive examinations. unpublished PhD. dissertation, Thorax center and Department of Vascular Surgery, University Hospital Rotterdam-Dijkzigt and Erasmus University Rotterdam, The Netherlands, 1994

[7] Poldermans D, Fioretti PM, Boersma E, Forster T, van Urk H, Cornel JH, Arnese MR, Roelandt JRTC, Safety of dobutamine-atropine stress echocardiography in patients with suspected or proven coronary artery disease: experience in 650 consecutive examinations. American Journal of Cardiology, Vol. 73, No. 7, 456-459

[8] Fioretti PM, Poldermans D, Salustri A, Bellotti P, Forster T, Boersma E, McNeill AJ, el-Said ES, Roelandt JRTC, Atropine increases the accuracy of dobutamine stress echocardiography in patients taking beta blockers. Eur Heart J, Vol. 15, No. 3,, 155-160

[9] Poldermans D, Boersma E, Fioretti PM, van Urk H, Boomsma F, Man in 't Veld AJ, Cardiac chronotropic responsiveness to β-receptor stimulation is not reduced in the elderly. J Am Coll Cardiol,Vol. 25, No. 5, 995-999

[10] Poldermans D, Fioretti PM, Forster T, Thomson IR, Boersma E, el-said ES, bu Bios NAJJ, Roelandt JRTC, van Urk H, Dobutamine stress echocardiography for assessment of perioperative cardiac risk in patients undergoing major vascular surgery. Circulation Vol. 87, 1993: 1506-1512

[11] Poldermans D, Fioretti PM, Forster T, Boersma E, Thomson IR, Arnese MR, van Urk H, Roelandt JRTC, Dobutamine-atropine stress echocardiography for assessment of perioperative and late cardiac risk in major vascular surgery. Eur J Vasc Surg, Vol. 8, No. 3, 1994: 286-293

[12] Poldermans D, Fioretti PM, Boersma E, Thomson IR, Arnese MR, van Urk H, Roelandt JRTC, Hemodynamics, safety and prognostics value of dobutamine-atropine stress echocardiography in 177 elderly patients unable to perform an exercise test. Unpublished PhD. dissertation, Thorax center and Department of Vascular Surgery, University Hospital Rotterdam-Dijkzigt and Erasmus University Rotterdam, The Netherlands, 1994

[13] Poldermans D, Fioretti PM, Boersma E, Cornel JH, Borst F, Vermeulen DGJ, Arnese MR, El-Hendy A, Roelandt JRTC, Dobutamine-atropine stress in echocardiography and clinical data for predicting late cardiac events in partients with suspected coronary artery disease. Am J Med, Vol. 97, No. 2, 1199-1225

[14] Greenland S, Finkle WD, A critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. American Journal of Epidemiology, Vol. 124, No. 12, 1995:1255-1264

Chapter 3

Existing approaches for statistical analysis of incomplete data

3.1 Introduction

Missing data are a pervasive problem in statistical analysis. Problems that are associated with missing data are:

1. The subsample of complete records may not be representative when the underlying missing data mechanism is not MCAR, possibly resulting in biased estimates, (see chapter 2 for definitions of MCAR, MAR and MNAR);

2. There is loss in power of statistical testing, so that conclusions are weaker than in case of complete data;

3. Statistical analysis is more complicated, e.g., when some covariates are incompletely observed, standard regression methods cannot be applied to the entire data set.

At least eight basic approaches to the problem of incomplete data exist [1-3]. These can be subdivided into simple and advanced approaches. Simple approaches are: listwise deletion, available-case analysis, single imputation, the indicator method and weighting; advanced approaches are: likelihood based approaches, posterior based approaches, and multiple imputation. Beside these general methods, many customized solutions exist for particular techniques.

See for example [4] for the analysis of variance applied to incomplete data. In this chapter, such customized methods will not be considered.

3.2 Existing approaches

Simple approaches:

- **Listwise deletion.** Listwise deletion, also known as complete-case analysis and complete-subject analysis, is the standard treatment in much statistical software. With listwise deletion, the incomplete cases are discarded and the remaining complete cases are analyzed. An advantage is that it is easy to implement and that standard statistical software can be used for the analysis. A serious disadvantage, however, is a potential waste of data. With listwise deletion, even a modest fraction of missing data entries may result in discarding a large fraction of the data set. Another disadvantage is the possibility of bias. In case of univariate statistical analysis, an underlying missing data mechanism which does not satisfy MCAR may result in bias, depending on the parameter of interest. For regression analysis with incompletely observed covariates, bias may occur when the probability of a covariate to be missing depends on the outcome variable. On the other hand, the problem of bias is considerably less serious when this probability is independent of the outcome variable. In [5] it is demonstrated that in the case of logistic regression with two categorical covariates and missing data in one covariate, estimates of regression coefficients are unbiased when the probability of missing data is independent of the outcome variable.

- **Available-case analysis.** With available-case analysis, also known as pairwise-deletion, the largest sets of available cases are used for the estimation of separate parameters. An example of available-case analysis is linear regression of a dependent variable y on independent variables $x_1, \ldots, x_p$ using a covariance matrix of which each element is estimated from the largest set of available cases.

  A disadvantage of linear regression based on available-case analysis is that the estimated covariance matrix may not be positive definite. This is especially the case for highly correlated data [2], since in this case some eigenvalues of the corresponding

covariance matrix are very close to zero; small changes in such covariance matrices often result in matrices which are not positive definite.

- **Single imputation.** With single imputation, each missing data entry is imputed (is filled in) once by an estimate and the resulting completed data set is analyzed. A simple method is unconditional mean imputation, where each missing value is imputed by the sample mean of the observed data of the corresponding variable. This method, however, may yield biased estimates of the covariance matrix [2] and is therefore not recommended. A better approach is conditional mean imputation, where for each missing data entry the expected value conditional on the observed values in the corresponding case is imputed. When the data are assumed to be multivariate normally distributed, a well-known method is the method proposed by Buck [6]. In this method, first estimates $\widehat{\mu}$ and $\widehat{\Sigma}$ of the mean vector μ and the covariance matrix Σ are derived from the complete cases. The regression coefficients for estimating the missing data entries are easily obtained from $\widehat{\mu}$ and $\widehat{\Sigma}$ by means of the sweep operator [1]. Although this method yields reasonable estimates of means, variances and covariances are underestimated, since the random noise with which observed data entries are distributed around their expected values is neglected. Better estimates of covariance matrices may be obtained by adding a random error-term to the imputed values.

  A disadvantage of any single imputation method is that standard errors are underestimated, confidence intervals are too narrow and p-values are too low [7], suggesting a higher precision and more evidence than in fact can be concluded from the observed data. This is due to the fact that the extra uncertainty due to missing data is not reflected since the imputed values are treated as if they are fixed known values. A simulation study in [8] illustrates that the underestimation of standard errors and p-values may be considerable.

- **Indicator method.** Another simple approach is the indicator method which is used in regression analysis. With this method, for each incomplete independent variable x_j, the regression term $\beta_j x_j$ is replaced by $\beta_{0j}(1-R_j)+\beta_j R_j x_j$, with R_j the response indicator of x_j. The term $\beta_j R_j x_j$ is zero when x_j is missing. The intercept is adjusted

by the additional term $\beta_{0j}(1-R_j)$. When only a single categorical covariate contains missing data entries, the indicator method is equivalent to creating an additional "missing" category for the covariate. Although by epidemiologists widely perceived as a formally correct method for handling missing data, the indicator method yields biased estimates under most conditions [3,9-11]. A reduction in bias with regard to the ordinary indicator method described above, may be obtained by a modified indicator method, where each regression term x_j is replaced by $\beta_{0j}(1-R_j)+R_j\beta_j x_j+\sum_{k\in mis;k\neq j} R_j(1-R_k)\beta_{jk}x_j$, where mis denotes the set of indices of the incomplete independent variables. The term $\sum_{k\in mis;k\neq j} R_j(1-R_k)\beta_{jk}x_j$ adjusts the regression coefficient β_j to the particular missing data patterns. Because of its larger number of extra parameters, it is not clear whether this method is more efficient than listwise deletion [3]. Simulations suggest that for logistic regression, there is no gain in efficiency with regard to listwise deletion [3];

- **Weighting.** Weighting is most commonly applied for the estimation of population means in case of unit nonresponse in sample surveys [1]. With unit nonresponse a case is either entirely observed or entirely missing. The principle of weighting is the same as the principle of probability sampling, where the population mean is estimated from a sample $y_1,\ldots,y_n$ with π_i the prior probability that y_i is contained in the sample. The weighed mean is then:

$$\overline{y}_w = \sum_{i=1}^{n} w_i y_i \quad , \tag{3.1}$$

where the weights w_i are given by

$$w_i = \frac{\pi_i^{-1}}{\sum_{j=1}^{n}\pi_j^{-1}}. \tag{3.2}$$

The weights w_i in Eq .3.2 are proportional to π_i^{-1} which is regarded as the number of units in the population represented by y_i [1]. Weighting is an extended form of probability sampling in which to each observed unit y_i, a weight is assigned inversely proportional to the prior probability of selection and response in combination. The

weights w_i in Eq .3.2 are replaced by

$$w_i = \frac{\pi_i^{-1}\phi_i^{-1}R_i}{\sum_{j=1}^{n}\pi_j^{-1}\phi_j^{-1}R_j} \quad , \tag{3.3}$$

where ϕ_i is the prior probability that y_i is observed given that y_i is selected and R_i is the response-indicator for this unit.

A standard approach to the estimation of the probabilities ϕ_j is to form adjustment cells on the basis of background variables Z, which are measured both for respondents and for non-respondents [1,12]. To remove nonresponse bias, the adjustment cells should be chosen in such a way that for each cell the observed units are a random subsample of the sampled units in this cell. In this case the probabilities ϕ_j are estimated by m_j/n_j, where n_j and m_j are the number of sampled and responding units in the j-th adjustment cell, respectively. When $y_1, \ldots, y_n$ is a random sample of Y, the probabilities π_i are equal to n^{-1}, so that according to Eq .3.1 and Eq .3.3 the population mean can be estimated by [1]

$$\overline{y}_{wc} = n^{-1}\sum_{j=1}^{J} n_j \overline{y}_{jR} \quad , \tag{3.4}$$

where J is the number of adjustment cells and $\overline{y}_{jR}$ is the average over the responding units in the j-th adjustment cell.

An advantage of weighting methods is that they are often easy to implement [12]. A limitation of weighting is, however, that generally methods are only well developed for unit nonresponse and simple problems, such as the estimation of the population meanby random sampling. For more complex problems with multiple missing data patterns, weights and standard errors may become extremely complex [3,13] and often ignore the component of variance due to estimating the weights from the data [12]. Another problem is that weighting methods are not always efficient in the sense that they can lead to estimates with unacceptably high standard errors [12].

Advanced approaches:

- **Likelihood based approaches (EM).** A common method for point estimation is maximum likelihood estimation (MLE). The idea behind MLE is that estimates of unknown model parameters of interest are found which maximize the likelihood of the observed data y under the assumed statistical model. More formally, let $f(y|\theta)$ be the probability, or for continuous data the probability density function, of the observed data y under the assumed statistical model given that θ is the true value of the unknown parameter of this model. The maximum likelihood estimate $\widehat{\theta}$ of θ is the value of θ which maximizes the likelihood $L(\theta|y)$, which is any function of y and θ proportional to $f(y|\theta)$. Usually, maximum likelihood estimates of θ are found by maximizing the log-likelihood $l(\theta|y)$, which is the natural logarithm of $L(\theta|y)$, rather than by maximizing $L(\theta|y)$. Maximizing $l(\theta|y)$ which is equivalent to maximizing $L(\theta|y)$ is generally easier. An asymptotic variance-covariance matrix for $\widehat{\theta}$ is given by $V(\widehat{\theta}) = I^{-1}(\widehat{\theta}|y)$, where $I(\theta|y)$ given by

$$I(\theta|y) = -\frac{\partial^2 l(\theta|y)}{\partial \theta^2}, \tag{3.5}$$

is the observed information about θ [1]. From the variance-covariance matrix $V(\widehat{\theta})$, component standard errors and a correlation matrix for $\widehat{\theta}$ is easily obtained.

For incomplete data y_{obs}, the corresponding log-likelihood $l(\theta|y_{obs})$ can be a complicated function with no obvious maximum and with a complicated form of the information matrix $I(\theta|y_{obs})$ [1]. In such cases, EM (Expectation - Maximization) is a popular iterative algorithm for MLE. Main advantages are that methods for complete data can often be used and convergence is stable [14]. The EM algorithm as proposed in [15], is a formalization of an old ad hoc idea: impute the missing data entries on the basis of an initial estimate $\theta^{(0)}$ of θ; re-estimate θ from the completed data by $\theta^{(1)}$; use $\theta^{(1)}$ to re-impute the missing data entries; use the re-completed data to re-estimate θ by $\theta^{(2)}$; and so on; iterate this until $\theta^{(t)}$ converges. Each iteration of EM consists of an E-step (Expectation) and an M-step (Maximization). In the E-step, the expected complete data log-likelihood $Q(\theta|\theta^{(t)}) = E[l(\theta|y)|y_{obs}, \theta = \theta^{(t)}]$ is estimated from the current estimate $\theta^{(t)}$. In the M-step, a new estimate $\theta^{(t+1)}$

is found, which maximizes the expected complete data likelihood $Q(\theta|\theta^{(t)})$ from the previous E-step for θ. In fact, in the E-step, not the missing data y_{mis} but functions of y_{mis} on which the complete data log-likelihood $l(\theta|y)$ depends are estimated. In [15] it is proved that in each iteration the incomplete data log-likelihood $l(\theta^{(t)}|y_{obs})$ increases. For special statistical models within the exponential family, such as the multivariate normal model, the M-step is similar to MLE for complete data.

Drawbacks of EM are:

- Convergence of EM can be very slow in cases of a large proportion of missing data [1];
- The convergence to a global maximum is not guaranteed;
- Standard errors and correlation matrices of point estimates are not directly available from EM and their calculation can be complicated. A general numerical procedure to obtain asymptotic sample variance-covariance matrices for point estimates is the supplemented EM (SEM) algorithm [16]. For a large number of parameters, however, SEM is computationally prohibitive [17];
- ML is essentially designed for large samples and has limitations for small samples [2].
- EM requires statistical expertise.

- **Posterior based approaches.** With posterior based approaches, the likelihood is extended by a prior component for θ and inference is based on the resulting posterior distribution of θ. In contrast to ML approaches posterior based approaches do not require a large sample size [2]. In the case of incomplete data, however, the posterior distribution of θ may be extremely complex, so that numerical integration or complex Monte Carlo techniques are required [2]. Examples of such Monte Carlo techniques are: data augmentation [18], Gibbs sampling [19], and importance sampling [20].
- **Multiple imputation.** Multiple imputation as proposed by Rubin [21], has the same main advantage as single imputation, i.e., the possibility of using statistical software for complete data. It does not have the main disadvantage of single imputation, i.e., not correctly reflecting the precision of point estimates. This is achieved

by replacing each missing data entry by m ($m \geq 2$) imputed values distributed around its expected value. The resulting m completed data sets are separately analyzed by the desired statistical method for complete data, and the m intermediate results are pooled into one final result by explicit procedures. For pooling completed data results, a complete-data analysis is represented by the tuple $(\hat{Q}, U)$, where $\hat{Q}$ is a point estimate of a parameter of interest Q and U is the variance-covariance matrix of $\hat{Q}$. From the m completed data results $(\hat{Q}_1^*, U_1^*), \ldots, (\hat{Q}_m^*, U_m^*)$, the unknown complete data results $(\hat{Q}, U)$ are estimated by $(\overline{Q}_m, \overline{U}_m)$, with $\overline{Q}_m$ given by

$$\overline{Q}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i^*, \tag{3.6}$$

and $\overline{U}_m$ given by

$$\overline{U}_m = \frac{1}{m} \sum_{i=1}^{m} U_i^*. \tag{3.7}$$

Extra inferential uncertainty about Q due to missing data is reflected by the between imputation variance-covariance matrix B_m given by

$$B_m = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i^* - \overline{Q}_m)(\hat{Q}_i^* - \overline{Q}_m)^T. \tag{3.8}$$

From the statistics $\overline{Q}_m, \overline{U}_m, B_m$, standard errors, confidence intervals, and p-values can be derived [8,21]. The underlying assumption is that the sample size is sufficiently large to approximate the distribution of $\overline{Q}_m$ by a multivariate normal distribution. The variance in $\overline{U}_m$ is neglected. In [21] it is proven that incomplete data analysis consisting of proper imputation (see chapter 4) followed by valid incomplete data analysis is also valid. In many practical cases, the required number of imputations m is modest; even with $m = 3$, multiple imputation works well [7,8].

Drawbacks of multiple imputation are:

- Similar to EM, multiple imputation is a large sample tool.
- Multiple imputation is laborious. For each missing data entry, m appropriate imputations must be generated. From these imputations and the observed data, m completed data sets must be generated and separately analyzed. Finally, the

Method	Statistical Validity		Efficiency	Complete data software	Small Samples
	Bias	Precision			
Listwise deletion	+/-	+	-	+	+
Available-case analysis	+/-	+	+/-	+	+
Single imputation	+	-		+	+
Ordinary indicator method	-			+	+
Modified indicator method	+	+	-	+	+
Weighting	+	+	+/-	-	+
EM	+	+	+	+/-	-
Posterior based methods	+	+	+	-	+
Multiple imputation	+	+	+	+	-

Table 3.1: Properties of the approaches to the analysis of incomplete data described in section 3.2 with regard to statistical validity, efficiency, use of software for complete data, and the applicability to small samples. The symbols '+', '+/-', '-' refer to good, reasonable and poor. When a cell is empty, the corresponding entry is regarded as meaningless.

m completed data results must be pooled into one result.

- Multiple imputation requires statistical expertise. This is especially the case for the generation of the imputations which requires an appropriate imputation model.

3.3 Discussion

In Table 3.1, an overview of the properties of the approaches to the analysis of incomplete data described in section 3.2 is given. In the table, statistical validity is subdivided into bias and precision. A point estimate is unbiased when it does not systematicaly deviate from the true value. Precision indicates that standard errors and confidence intervals are a correct reflection of the inferential uncertainty of the corresponding point estimates. In chapter 4, the concept behind statistical validity is described in more detail. Another criterion is efficiency. A point estimate is efficient when its corresponding standard error is minimal. The intuitive idea behind efficiency is that all information about the parameter of interest available in the data is used. When the statistical validity of a method is poor, its efficiency is considered meaningless and the corresponding cell in Table 3.1 is left empty. For the ordinary indicator method, the precision is also entered as meaningless, since in case of bias, the correct reflection of the precision of point estimates makes no sense.

From the table, it appears that multiple imputation is the only method with good properties regarding statistical validity, efficiency and use of statistical software for complete data, so that of the methods considered, multiple imputation can be regarded as the best approach to the analysis of incomplete data. It is clear that multiple imputation is the only method which combines the advantage of using statistical software for complete data with the advantage of a good statistical validity. With EM, existing methods for complete data can also be used, but EM is considerably less flexible in using complete data methods. This is due to the possibly large number of iterations to be carried out, and to the fact that, for some statistical models, the M-step is hard to formulate. Additional advantages of multiple imputation over other approaches are that it is very suitable for sensitivity analysis [21] (see chapter 6), and that diagnostic measures for the assessment of the extra inferential uncertainty due to missing data are easily available (see chapter 4).

A limitation of multiple imputation is, however, that it is essentially a large sample tool. Posterior based methods do not have this limitation, but the price which must be paid is that complex numerical integration and Monte Carlo techniques are required. In future simulation studies, it is important to investigate which sample size is required for multiple imputation to be acceptable in practice. Another topic to be investigated is, whether multiple imputation is robust against deviations from MAR, when used for regression analysis under the MAR assumption, with missing data in the covariates and when the probabilities of data entries to be missing are independent of the outcome variable.

Bibliography

[1] Little RJA, Rubin DB, Statistical analysis with missing data, 1987, New York, Wiley.

[2] Little RJA, Regression With Missing X's: A Review, Journal of the American Statistical Association, Vol. 87, No. 420, 1992: 1227-1237

[3] Greenland S, Finkle WD, A Critical Look At Methods for Handling Missing Covariates in Epidemiologic Regression Analysis. American Journal of Epidemiology, Vol. 142, No. 12, 1995:1255-1264

[4] Dodge Y, Analysis of experiments with missing data. J. Wiley & Sons, New York, 1985

[5] Vach W, Logistic regression with missing values in the covariates, New York, Springer Verlag, 1994

[6] Buck SF, A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. Journal of the Royal Statistical Society, Ser. B,22,1960:302-306

[7] Rubin DB, Schenker N, Multiple imputation in health-care databases: an overview and some applications. Statistics in Medicine, Vol. 10, 585-589,1991

[8] Li KH, Raghunathan TE, Rubin DB, Large-Sample Significance Levels From Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution. Journal of the American Statistical Association, Vol. 86, No. 1, 1991:65-92

[9] Mietinen OS, Theoretical epidemiology. New York, Wiley, 1985

[10] Vach W, Blettner M, Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiology, 1991;124:895,907

[11] Jones MP, Indicator and stratification methods for missing explanatory variables in multiple linear regression. Technical report no. 94-2 Ames, IA: Department of Statistics, University of Iowa, 1994

[12] Kessler RC, Little RJA, Groves RM, Advances in Strategies for Minimizing and Adjusting for Survey Nonresponse. Epidemiologic Reviews, Vol. 17, No.1, 1995:192-204

[13] Robins MR, Rotnitzky A, Zhao LP, Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. Journal of the American Statistical Association, Vol. 89, No.427, 1994:846-866

[14] Meng XL, Rubin DB, Maximum Likelihood Estimation via the ECM algorithm: A general framework, Biometrica, Vol.80, No.2, 1993:267-278

[15] Dempster AP, Laird NM, Rubin DB, Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. B39,1977:1-38

[16] Meng XL, Rubin DB, Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association, Vol. 86, No. 416, 1991:899-909

[17] Shafer JL, Analysis of Incomplete Multivariate Data. Chapman & Hall, London, 97

[18] Tanner M, Wong W, The Calculation of Posterior Distributions by Data Augmentation (with discussion). Journal of the American Statistical Association, Vol. 82, 1987:528-550

[19] Gelfland AE, Smith AFM, Sampling-Based Approaches to Calculating Marginal Densities, Journal of the American Statistical Association, Vol. 85,1990:398-409

[20] Rubin DB, Comment on "The Calculation of Posterior Distributions by Data Augmentation" by Tanner M, Wong W, Journal of the American Statistical Association, Vol. 82, 1987:528-550

[21] Rubin DB, Multiple imputation for nonresponse in surveys. Wiley New York, 1987

Chapter 4

Multiple imputation

4.1 Introduction

Single imputation as a strategy to cope with missing data results in confidence intervals that are too small and in p-values suggesting too much significance. Extra uncertainty due to missing data is neglected: imputed values are treated as if they were fixed known values. Uncertainty resulting from missing data can be reflected by replacing each missing data entry by an uncertain value. Uncertainty in a missing value is given by a probability distribution indicating how likely the possible values for a particular missing data entry are, given the observed data. The example below illustrates the representation of a missing data entry by an uncertain value:

Example 1 *In this example the unobserved value of ps (peak systolic blood pressure), represented by the third missing data entry in the case* $(*, *, *, 110)$*, belonging to the incomplete sample 'Dobutamine MAR' (see Appendix A) will be considered. The only observed value is the value 110 of pd (peak diastolic blood pressure). To represent the missing value of ps as an uncertain value, the observed value of pd 110 is used together with information about the relationship between ps and pd available from the observed data. Information about this relationship is found in the Tables 2.7 and 2.8 in chapter 2 and summarized in Figure 4.1. All information used is represented by bold numbers in this Figure.*

The observed value of 110 is a relatively high value for pd, as can be seen from the mean of pd 82.94 and the standard deviation of pd 15.85. From the strong correlation between ps

rs = systolic blood pressure in state of rest
rd = diastolic blood pressure in state of rest
ps = systolic blood pressure in state of maximal (peak) effort
pd = diastolic blood pressure in state of maximal (peak) effort

variable	nobs	mean	standard deviation
rs	294	140.59	22.74
rd	295	81.37	12.50
ps	221	**153.66**	**31.79**
pd	394	**82.94**	**15.85**

correlation matrix

	rs	rd	ps	pd
rs	1			
rd	0.67	1		
p	0.42	0.38	1	
pd	0.32	0.51	**0.71**	1

observation: **pd = 110**

Figure 4-1: Used information to represent *ps* as an uncertain value.

and pd (0.71), a relatively high value for the missing ps can be expected as well. Assuming that ps and pd have a bivariate normal distribution, the missing variable ps can be represented by a normal distribution with a mean of $153.66 + 0.71 * \frac{31.79}{15.85}(110 - 82.94) = 192.19$ *and a standard deviation of* $31.79 * \sqrt{(1 - (0.71)^2)} = 22.39$. *It can be seen that the observed value of 110 for pd provides information about the missing ps. Given the observed value of pd, the expected value of the missing ps is increased by 38.53 and its standard deviation is reduced by 30%* $(1 - \sqrt{(1 - (0.71)^2)} = 0.30)$*, as compared to the case when pd was not observed.*

We emphasize that by replacing the missing data entries by uncertain values, we do not add new information to the data. We merely reflect all the information about the missing data which is contained in the observed data. In example 1, this information is summarized by the means and standard deviations of the variables *ps* and *pd*, their correlation, and the observation for *pd* in the case where *ps* is missing. The assumption about normality cannot be regarded as adding extra information to the data, but is made in order to model the use of the available information in the data.

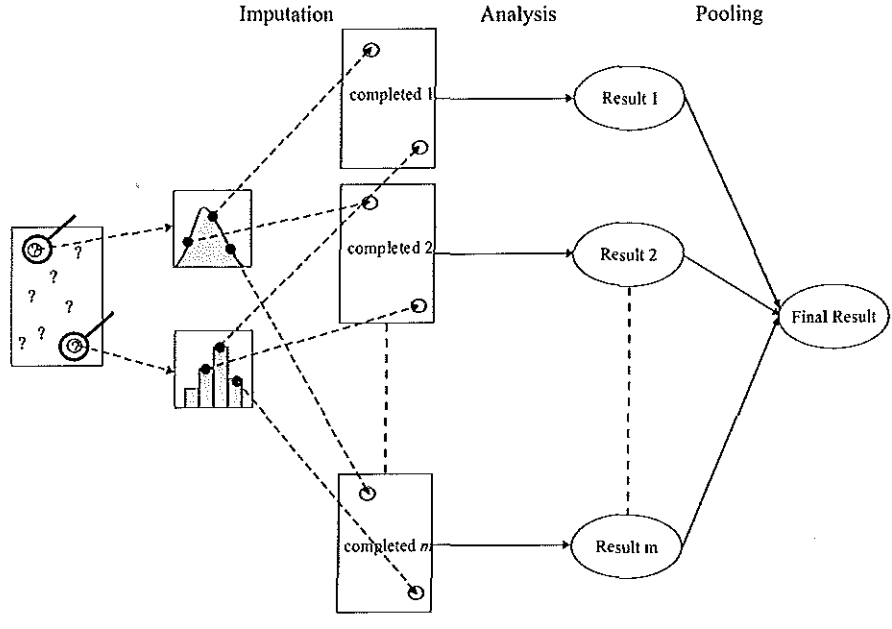


Figure 4-2: The concept of multiple imputation.

Analytical incorporation of the uncertainty due to missing data is generally very complicated. This is especially the case when there are incomplete cases which contain more than one missing data entry, so that simultaneous probability distributions for these missing data entries must be taken into account. Multiple imputation as proposed by Rubin [1] is a technique to perform the incorporation of the uncertainty about missing data by means of computer simulation. The concept behind multiple imputation is depicted in Figure 4.2.

In multiple imputation, for each missing data entry, m imputations ($m \geq 2$) are generated by simulated draws from the probability distribution representing the uncertainty of a missing value. The m resulting completed samples are separately analyzed by an appropriate statistical method for complete data, and the m intermediate analysis results are pooled into one final result by explicit procedures. Pooling of the results implies that the resulting point estimates are averaged over the m completed sample point estimates, and the resulting standard errors and p-values are adjusted according to the variance of the corresponding m completed sample

point estimates. This variance, also called the between imputation variance, provides a measure of the extra inferential uncertainty due to missing data. In the theoretical situation that m is infinite, the empirical probability distribution obtained from the imputations is exactly equal to the probability distribution representing the missing data entries as uncertain values, so that the multiple imputation results are identical to the results from the analytical incorporation of the extra uncertainty due to missing data. For modest m, the multiple imputation results are an approximation of analytical incorporation, so that the pooling procedures also reflect the extra uncertainty due to the simulation error. Even for small m, e.g. $m = 5$, the extra loss in power of statistical testing and in precision of point estimates (also called efficiency), due to this simulation error, is modest in most practical cases [2,3], so that $m = 5$ is generally a good choice for large samples.

Multiple imputation can be regarded as a bridge between the frequentistic and Bayesian schools. Inference from multiple imputation is based on Bayesian statistical inference and is validated according to a frequentistic criterion. Within the Bayesian framework, inference from multiple imputation is straightforward and much easier than within the frequentistic framework. The frequentistic framework is well known to statisticians and application researchers and is suitable for the validation of statistical procedures. A summary of the controversy between the Bayesian and the frequentistic school is given in [4].

For pooling completed sample results, it is sufficient to represent complete sample statistics by the tuple $\left(\widehat{Q}, U\right)$, where $\widehat{Q}$ is a point-estimator of a parameter of interest Q and U the complete sample variance-covariance matrix of $\widehat{Q}$. Usually, for univariate Q the standard error $\sqrt{U}$ is presented and for multivariate Q the correlation matrix together with the component standard errors of $\widehat{Q}$ are used. From the completed sample results of this tuple, the pooled results for point-estimators, standard errors, correlation matrices, test statistics and p-values, which are the basic statistics in statistical analysis, can be obtained. The underlying assumption is that the sample size is sufficiently large. Under this assumption, test statistics and p-values for a complete sample can also be determined from the tuple $\left(\widehat{Q}, U\right)$ and their relationship with this tuple is the same within the frequentistic- and the Bayesian framework.

The validation criterion for infinite multiple imputation is proper multiple imputation [1], which is similar to the criterion for valid complete data inference. Incomplete data inference

consisting of proper infinite multiple imputation and subsequent valid complete data inference is also valid [1]. Whether multiple imputation is proper depends on the underlying sampling mechanism generating the complete sample, the underlying missing data mechanism and the complete sample statistics $\left(\widehat{Q}, U\right)$, obtained from the m completed samples. Pooling procedures for modest m are separately validated assuming proper multiple imputation [1,3,5,6].

It is assumed, that if the imputations are independently generated from the Bayesian posterior predictive distribution of the missing data given the observed data $P(Y_{mis}|Y_{obs})$ under an appropriate statistical model for the complete sample and a given missing data mechanism, multiple imputation will be proper or at least approximately proper [1,7]. Rubin has formulated this as follows:

' *If imputations are drawn to approximate repetitions from a Bayesian posterior distribution of* Y_{mis} *under the posited response mechanism and an appropriate model for the data, then in large samples the imputation method is proper ... There is little doubt that if this conclusion were formalized in a particular way, exceptions to it could be found. Its usefulness is not as a general mathematical result, but rather as a guide to practice. Nevertheless, in order to understand why it may be expected to hold relatively generally, it is important to provide a general heuristic argument for it (Rubin 1987, pp. 125-126).* '

Generating imputations for multivariate data from the exact posterior predictive distribution $P(Y_{mis}|Y_{obs})$ is generally very difficult. An exception is a monotonous missing data pattern [1], where the generation of such imputations is straightforward. In the case of multivariate data with a non-monotonous missing data pattern, data augmentation (DA) [8] and sampling/importance resampling (SIR) [1] are well-known approaches for generating imputations from an approximate predictive distribution $P(Y_{mis}|Y_{obs})$ on the basis of a multivariate statistical model. Imputation methods by means of data augmentation have been developed for the multivariate normal model for entirely numerical data, for the saturated multinomial model and for the log-linear model for entirely categorical data, and for the general location model for mixed data [9].

A limitation of the approaches mentioned above is that they are problematic for large samples with many variables, since the algorithm may become numerically unstable for such samples, due to the large number of parameters of the imputation model. Our approach is

a variable-by-variable version of the Gibbs sampling algorithm, also applied in [10], in which the underlying statistical model is specified by separate regression models, each describing the statistical relationship between an imputation variable and a set of predictor variables. Gibbs sampling, also known as stochastic relaxation, is an iterative procedure which is used for Monte Carlo simulation of high dimensional stochastic systems. Applications of the Gibbs sampling algorithm can be found in [11,12].

In the variable-by-variable Gibbs sampling approach, it is possible to only include relevant predictor variables for the imputation variables, so that the number of parameters of the corresponding statistical model can be reduced. Another advantage of this approach is that it is easier to specify an appropriate statistical model. This is especially the case for samples consisting of a mixture of numerical and categorical variables (which is often the case in medical data sets), for which an adequate statistical model is often hard to find.

In section 4.2, the general Gibbs sampling algorithm and its application to multiple imputation is described. A strategy for specifying the parameters of the resulting imputation procedure is proposed. The multiple imputation procedures are developed under the MAR assumption and thus do not require modelling of the underlying missing data mechanism. The procedures for pooling the m intermediate completed sample results into one final result are described in section 4.3. In section 4.4, the validation of multiple imputation is described.

4.2 Generating proper imputations

The posterior predictive distribution $P(Y_{mis}|Y_{obs})$ of the missing data Y_{mis} given the observed data Y_{obs} is defined by

$$P(Y_{mis}|Y_{obs}) = \int_{\Theta} P(Y_{mis}|Y_{obs};\theta)\, P(\theta|Y_{obs})\, d\theta. \tag{4.1}$$

It is assumed, that if imputations $Y^{\star}_{mis}$ are independently drawn from the predictive distribution, multiple imputation will be proper or at least approximately proper. In Eq. 4.1, θ represents the parameters of a statistical model with parameter space Θ describing the hypothetical complete data Y, and the posterior $P(\theta|Y_{obs})$ reflects the uncertainty about θ given the observed data Y_{obs}. For multivariate data, drawing imputations from the exact predictive distribution in

Eq 4.1 is generally very difficult. Especially the incomplete data posterior $P(\theta|Y_{obs})$ is hard to evaluate in such cases. Imputations from an approximate predictive distribution $P(Y_{mis}|Y_{obs})$ can be obtained by the Gibbs sampling algorithm.

4.2.1 Gibbs sampling

Gibbs sampling, also known as stochastic relaxation, is a Monte Carlo technique to simulate a drawing from a multivariate probability density distribution by repeatedly drawing from conditional probability density distributions. In this subsection it is described how the general Gibbs sampling algorithm can be applied for generating imputations Y^*_{mis}.

Let Z be a p-dimensional random variable and let $\{Z(1), Z(2), \ldots, Z(k)\}$ $(k \leq p)$ be a partitioning of its components, i.e., $Z(1), \ldots, Z(k)$ are disjunct subsets of the components of Z and their union contains all components of Z. Let W be a possibly multi-dimensional variable and $P(Z|W)$ the probability distribution of interest. Starting with an initial value $Z^{(0)}$ of Z, the Gibbs sampling algorithm generates a sequence of values $Z^{(0)}, Z^{(1)}, Z^{(2)}, Z^{(3)}, \ldots$ where in iteration t, $t \geq 1$, $Z^{(t)}$ is generated from $Z^{(t-1)}$ by successively generating [13]:

$$
\begin{aligned}
Z(1)^{(t)} &\sim P\left(Z(1) \mid Z(2)^{(t-1)}, Z(3)^{(t-1)}, \ldots, Z(k)^{(t-1)}; W\right) \\
Z(2)^{(t)} &\sim P\left(Z(2) \mid Z(1)^{(t)}, Z(3)^{(t-1)}, \ldots, Z(k)^{(t-1)}; W\right) \\
&\vdots \\
Z(j)^{(t)} &\sim P\left(Z(j) \mid Z(1)^{(t)}, \ldots, Z(j-1)^{(t)}, Z(j+1)^{(t-1)}, \ldots, Z(k)^{(t-1)}; W\right) \\
&\vdots \\
Z(k)^{(t)} &\sim P\left(Z(k) \mid Z(1)^{(t)}, Z(2)^{(t)}, \ldots, Z(k-1)^{(t-1)}; W\right)
\end{aligned}
\tag{4.2}
$$

In Eq. 4.2, for each $Z(j)$, the value $Z(j)^{(t)}$ is generated conditionally on the values of the other variables, which are drawn most recently and a given value of W. According to Markov Chain theory [14], the distribution of $Z^{(t)}$ converges to the desired distribution $P(Z|W)$ under mild regularity conditions [15].

Let $y_1, \ldots, y_k$ be the variables containing missing data entries and $x_1, \ldots, x_q$ the completely observed variables in a sample. Figure 4.3 illustrates the application of the Gibbs sampling algorithm in Eq. 4.2 to generate imputations for this sample. The top of this Figure shows the

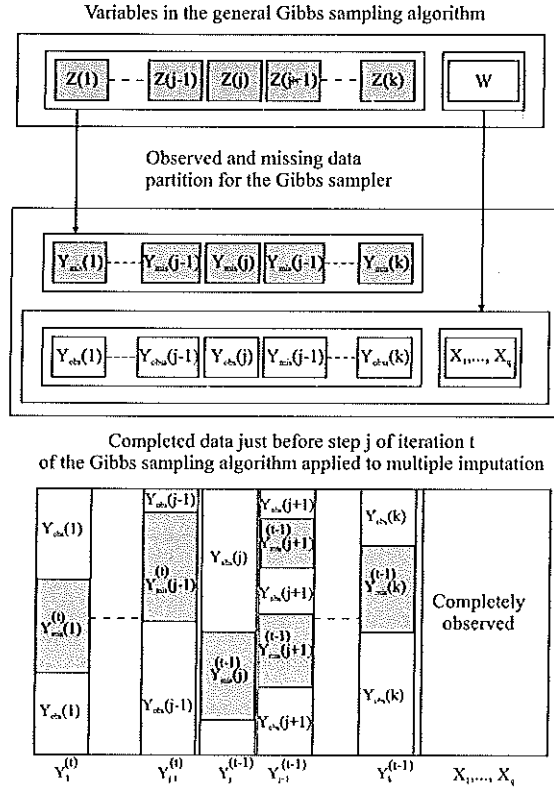


Figure 4-3: The Gibbs sampling algorithm for generating imputations.

mapping of the variables (Z, W) into the variables $(Y_{mis}; Y_{obs}, X)$. The probability distribution of interest here is $P(Y_{mis}|Y_{obs}; X_1, \ldots, X_q)$, where Y_{mis} represents the unknown missing data entries of $y_1, \ldots, y_k$, Y_{obs} the observed values for these variables, and X_j is the vector containing the observations for x_j. By mapping Z onto Y_{mis} and W onto the union of Y_{obs} and $X_1, \ldots, X_q$, the predictive distribution $P(Z|W)$ is translated into $P(Y_{mis}|Y_{obs}; X_1, \ldots, X_q)$. An obvious partitioning of Y_{mis} is $\{Y_{mis}(1), \ldots, Y_{mis}(k)\}$ with $Y_{mis}(j)$ the missing data entries for y_j, so that the following iterative algorithm for generating an imputation Y^*_{mis} is obtained.

Starting with an initial imputation $Y^{(0)}_{mis}$, a sequence of imputations $Y^{(1)}_{mis}, \ldots, Y^{(N)}_{mis}$ is generated by successively generating the imputations $Y^{(t)}_{mis}(i)$ conditional on the observed data and

on the most recently imputed data of $Y_{mis}(j), j \neq i$. In particular, $Y_{mis}^{(t)}(j)$ is generated from the predictive distribution
$P\left(Y_{mis}(j) \mid Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{obs}(j), Y_{j+1}^{(t-1)}, \ldots, Y_k^{(t-1)}, X_1, \ldots, X_q\right)$. In this predictive distribution $Y_1^{(t)}, \ldots, Y_{j-1}^{(t)}$ represent the completed data for $y_1, \ldots, y_{j-1}$ in the current iteration, $Y_{obs}(j)$ represents the observed data for y_j, $Y_{j+1}^{(t-1)}, \ldots, Y_k^{(t-1)}$ represent the completed data for $y_{j+1}, \ldots, y_k$ in the previous iteration, and $X_1, \ldots, X_q$ the completely observed data for $x_1, \ldots, x_q$. The completed data just before the generation of $Y_{mis}^{(t)}(j)$ is depicted at the bottom of Figure 4.3. In this Figure, the imputed data $Y_{mis}^{(t-1)}(j+1)$ in the $j+1$-th column is represented by two blocks to indicate that it is generally impossible to sort the cases of the sample such that in each column the missing data entries constitute a consecutive sequence.

When a regression model of y_j on $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_k, x_1, \ldots, x_q$ is specified and its parameters represented by ϕ_j are known, the predictive distribution of $Y_{mis}^{(t)}(j)$ can be defined. However, ϕ_j can only be estimated from the complete data. Using a point estimate $\widehat{\phi}_j$ of ϕ_j for generating $Y_{mis}^{(t)}(j)$ generally leads to improper imputation since the extra uncertainty about ϕ_j is not taken into account. To reflect the uncertainty about ϕ_j given the complete data, a value for ϕ_j is drawn from an appropriate posterior distribution for ϕ_j conditional on the most recently completed data, and this value is used to generate $Y_{mis}^{(t)}(j)$. Iteration t of the Gibbs sampling algorithm for generating $Y_{mis}^{(t)}$ from $Y_{mis}^{(t-1)}$ is then given by:

$$
\begin{aligned}
\phi_1^{(t)} &\sim P\left(\phi_1 \mid \left[Y_1^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q\right]_{obs(1)}\right)\\
Y_{mis}^{(t)}(1) &\sim P\left(Y_{mis}(1) \mid Y_2^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q;\phi_1^{(t)}\right)\\
\phi_2^{(t)} &\sim P\left(\phi_2 \mid \left[Y_1^{(t)},Y_2^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q\right]_{obs(2)}\right)\\
Y_{mis}^{(t)}(2) &\sim P\left(Y_{mis}(2) \mid Y_1^{(t)},Y_3^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q;\phi_2^{(t)}\right)\\
&\vdots\\
\phi_j^{(t)} &\sim P\left(\phi_j \mid \left[Y_1^{(t)},\ldots,Y_{j-1}^{(t)},Y_j^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q\right]_{obs(j)}\right)\\
Y_{mis}^{(t)}(j) &\sim P\left(Y_{mis}(j) \mid Y_1^{(t)},\ldots,Y_{j-1}^{(t)},Y_{j+1}^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q;\phi_j^{(t)}\right)\\
&\vdots\\
\phi_k^{(t)} &\sim P\left(\phi_k \mid \left[Y_1^{(t)},\ldots,Y_{k-1}^{(t)},Y_k^{(t-1)},X_1,\ldots,X_q\right]_{obs(k)}\right)\\
Y_{mis}^{(t)}(k) &\sim P\left(Y_{mis}(k) \mid Y_1^{(t)},\ldots,Y_{k-1}^{(t)},X_1,\ldots,X_q;\phi_k^{(t)}\right)
\end{aligned}
\tag{4.3}
$$

In Eq. 4.3, $\left[Y_1^{(t)},\ldots,Y_{j-1}^{(t)},Y_j^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q\right]_{obs(j)}$ are the rows of the completed data $Y_1^{(t)},\ldots,Y_{j-1}^{(t)},Y_j^{(t-1)},\ldots,Y_k^{(t-1)},X_1,\ldots,X_q$ corresponding to the observed values of y_j. In order to correct for possibly under- or overestimated associations between variables resulting from the initial imputation $Y_{mis}^{(0)}$, the parameter values $\phi_j^{(t)}$ are drawn from the posterior distribution of ϕ_j conditional on the completed data restricted to the cases for which y_j is observed rather than this posterior distribution conditional on the entire completed data.

Although convergence of the Gibbs sampling algorithm in Eq. 4.3 does not require the starting distribution of $Y_{mis}^{(0)}$ to be close to the target distribution $P(Y_{mis}|Y_{obs})$, such a close starting distribution is recommended, since the distribution of $Y_{mis}^{(t)}$ can remain heavily influenced by the starting distribution [16] for many iterations. A start imputation $Y_{mis}^{(0)}$ from a distribution which is close to the desired predictive distribution $P(Y_{mis}|Y_{obs})$ can be generated by ordering $y_1,\ldots,y_k$ according to ascending fractions of missing data. $Y_{mis}^{(0)}(1)$ can then be obtained conditional on the observed values $Y_{obs}(1)$ of y_1 and $X_1,\ldots,X_q$. Subsequently $Y_{mis}^{(0)}(2),\ldots,Y_{mis}^{(0)}(k)$ are generated conditional on the previously completed data. I.e., the imputation $Y_{mis}^{(0)}(j)$ is generated conditional on the completed data $Y_1^{(0)},\ldots,Y_{j-1}^{(0)}$ for $y_1,\ldots,y_{j-1}$, the data $X_1,\ldots,X_q$ and the observed values $Y_{obs(j)}$ for y_j. In example 2, an application of the Gibbs sampling

algorithm to the incomplete sample 'Dobutamine MAR' is described.

Example 2 *From table 2.8 and table 2.3 in chapter 2, it can be seen that pd is completely observed, and that the other variables rs, rd and ps contain 100, 99 and 173 missing data entries, respectively. For simplicity, it is assumed that all variables are linearly related according to the usual linear regression model, so that the resulting imputation model is given by:*

$$\begin{aligned} rs &= \beta_{10} + \beta_{11}rd + \beta_{12}ps + \beta_{13}pd + \epsilon_1 \;;\; \epsilon_1 \sim N\left(0,\sigma_1^2\right) \\ rd &= \beta_{20} + \beta_{21}rs + \beta_{22}ps + \beta_{23}pd + \epsilon_2 \;;\; \epsilon_2 \sim N\left(0,\sigma_2^2\right) \\ ps &= \beta_{30} + \beta_{31}rs + \beta_{32}rd + \beta_{33}pd + \epsilon_3 \;;\; \epsilon_3 \sim N\left(0,\sigma_3^2\right) \end{aligned} \tag{4.4}$$

In Eq. 4.4, ϵ_1, ϵ_2 and ϵ_3 are normally distributed error-terms, each of which is independent of the explanatory variables in the corresponding regression model.

By applying Eq. 4.3 to $y_1 = rs$, $y_2 = rd$, $y_3 = ps$, $k = 3$, $x_1 = pd$, $q = 1$, $\phi_i = (\beta_{i0},\beta_{i1},\beta_{i2},\beta_{i3},\sigma_i)$, $i = 1,\ldots,3$, and using the imputation model in Eq. 4.4, iteration t of

the Gibbs sampling algorithm is given by:

step 1: *(a)* *draw* $\beta_{10}^{(t)}$, $\beta_{11}^{(t)}$, $\beta_{12}^{(t)}$, $\beta_{13}^{(t)}$, $\sigma_1^{(t)}$ *from a posterior distribution of* β_{10}, β_{11},β_{12}, β_{13},σ_1, *conditional on the completed data for rs, rd, ps in the previous iteration and the complete data for pd, restricted to the cases for which rs is observed.*

(b) *impute each missing* rs_i *by* $rs_i^{(t)} = \beta_{10}^{(t)} + \beta_{11}^{(t)} rd_i^{(t-1)} + \beta_{12}^{(t)} ps_i^{(t-1)} + \beta_{13}^{(t)} pd_i + \epsilon_{1i}^{(t)}$, *with each error-term* $\epsilon_{1i}^{(t)}$ *independently drawn from* $N\left(0, \sigma_1^{(t)2}\right)$.

step 2: *(a)* *draw* $\beta_{20}^{(t)}$, $\beta_{21}^{(t)}$,$\beta_{22}^{(t)}$,$\beta_{23}^{(t)}$,$\sigma_2^{(t)}$ *from a posterior distribution of* β_{20}, β_{21}, β_{22}, β_{23}, σ_2, *conditional on the completed data for rs in step 1, the completed data for rd and ps in the previous iteration and the complete data for pd, restricted tot he cases for which rd is observed.*

(b) *impute each missing* rd_i *by* $rd_i^{(t)} = \beta_{20}^{(t)} + \beta_{21}^{(t)} rs_i^{(t)} + \beta_{22}^{(t)} ps_i^{(t-1)} + \beta_{23}^{(t)} pd_i + \epsilon_{2i}^{(t)}$, *with each error-term* $\epsilon_{2i}^{(t)}$ *independently drawn from* $N\left(0, \sigma_2^{(t)2}\right)$.

step 3: *(a)* *draw* $\beta_{30}^{(t)}$, $\beta_{31}^{(t)}$,$\beta_{32}^{(t)}$,$\beta_{33}^{(t)}$,$\sigma_3^{(t)}$ *from a posterior distribution of* β_{30}, β_{31}, β_{32}, β_{33}, σ_3, *conditional on the completed data of rs and rd in step 1 and step 2, the completed data for ps in the previous iteration and the complete data for pd, restricted to the cases for which ps is observed.*

(b) *impute each missing* ps_i *by* $ps_i^{(t)} = \beta_{30}^{(t)} + \beta_{31}^{(t)} rs_i^{(t)} + \beta_{32}^{(t)} rd_i^{(t)} + \beta_{33}^{(t)} pd_i + \epsilon_{3i}^{(t)}$, *with each error-term* $\epsilon_{3i}^{(t)}$ *independently drawn from* $N\left(0, \sigma_3^{(t)2}\right)$.

(4.5)

The substeps (a) in Eq. 4.5 correspond to the posterior draws of $\phi_j^{(t)}$ *in Eq. 4.3, and the substeps (b) correspond to the draws of* $Y_{mis}^{(t)}(j)$. *The posterior distributions from which the regression parameters* β *and* σ *in the substeps (a) are drawn, are derived from non-informative priors [17] about these parameters. The concept behind non-informative priors is outside the scope of this thesis, but it may be noted that these posterior distributions for* β *and* σ *resemble the sampling distributions of the least squares estimators* $\widehat{\beta}$ *and* $\widehat{\sigma}$ *from the completed data. Posterior draws* $\sigma^{(t)2}$ *are obtained from the distribution of* $\widehat{\sigma}^2 (n-4) / \chi_{n-4}^2$, *where* n *is the sample size of the hypothetical complete sample and* χ_{n-4}^2 *is a* χ^2 *random variable with* $n-4$ *degrees of freedom. Subsequently, posterior draws* $\beta^{(t)}$ *are obtained from* $N\left(\widehat{\beta}, \sigma^{(t)2} \left(X^T X\right)^{-1}\right)$, *with* X *a* $(n \times 4)$ *matrix with the first column consisting of ones and the other three columns consisting of completed data for the predictor variables. Similarity between sampling and posterior dis-*

tribution is most clear for β, where the posterior distribution of β is the sampling distribution of $\widehat{\beta}$ with β and σ^2 replaced by $\widehat{\beta}$ and $\sigma^{(t)2}$, respectively. For σ^2, however, the posterior and sampling distribution differ, but they both depend on the same term $(n-4)/\chi^2_{n-4}$.

Ordering rs, rd and ps according to their fractions of missing data results in the sequence rd, rs, ps. To generate a start-imputation for rd, rs and ps, the model in Eq. 4.4 is simplified to the model:

$$\begin{aligned} rd &= \beta_{20} + \beta_{23}pd + \epsilon_2 && ; \ \epsilon_2 \sim N\left(0,\sigma_2^2\right) \\ rs &= \beta_{10} + \beta_{11}rd + \beta_{13}pd + \epsilon_1 && ; \ \epsilon_1 \sim N\left(0,\sigma_1^2\right) \\ ps &= \beta_{30} + \beta_{31}rs + \beta_{32}rd + \beta_{33}pd + \epsilon_3 && ; \ \epsilon_3 \sim N\left(0,\sigma_3^2\right) \end{aligned} \tag{4.6}$$

Similar to Eq 4.5, a start imputation for rd, rs and ps is generated according to the following scheme:

step 1: *(a)* *draw $\beta_{20}^{(0)}$, $\beta_{23}^{(0)}$, $\sigma_2^{(0)}$ from a posterior distribution of β_{20}, β_{23}, σ_2, conditional on the observed data for rd and pd in cases where these two variables are simultaneously observed.*

(b) *impute each missing rd_i by $rd_i^{(0)} = \beta_{20}^{(0)} + \beta_{23}^{(0)}pd_i + \epsilon_{2i}^{(0)}$, with each error-term $\epsilon_{2i}^{(0)}$ independently drawn from $N\left(0,\sigma_2^{(0)2}\right)$.*

step 2: *(a)* *draw $\beta_{10}^{(0)}$, $\beta_{11}^{(0)}$, $\beta_{13}^{(0)}$, $\sigma_1^{(0)}$ from a posterior distribution of β_{10}, β_{11}, β_{13}, σ_1, conditional on the completed data for rd in step 1 and the observed data for rs and pd, restricted to the cases for which rs is observed.*

(b) *impute each missing rs_i by $rs_i^{(0)} = \beta_{10}^{(0)} + \beta_{11}^{(0)}rd_i^{(0)} + \beta_{13}^{(0)}pd_i + \epsilon_{1i}^{(0)}$, with each error-term $\epsilon_{1i}^{(0)}$ independently drawn from $N\left(0,\sigma_1^{(0)2}\right)$.*

step 3: *(a)* *draw $\beta_{30}^{(0)}$, $\beta_{31}^{(0)}$, $\beta_{32}^{(0)}$, $\beta_{33}^{(0)}$, $\sigma_3^{(0)}$ from a posterior distribution of β_{30}, β_{31}, β_{32}, β_{33}, σ_3, conditional on the completed data for rd and rs in step 1 and step 2, and the observed data of ps and of pd, restricted to the cases in which ps is observed.*

(b) *impute each missing ps_i by $ps_i^{(0)} = \beta_{30}^{(0)} + \beta_{31}^{(0)}rs_i^{(0)} + \beta_{32}^{(0)}rd_i^{(0)} + \beta_{33}^{(0)}pd_i + \epsilon_{3i}^{(0)}$, with each error-term $\epsilon_{3i}^{(0)}$ independently drawn from $N\left(0,\sigma_3^{(0)2}\right)$.*

(4.7)

4.2.2 Representation of imputation methods

When the number of variables in a sample is relatively large as compared to the number of cases in this sample, then applying Gibbs sampling according to Eq. 4.3, using for each imputation variable y every variable other than y as a predictor variable, may lead to numerical problems. In such situations it is better to use for each imputation variable y a subset $\{x\}$ of relevant predictor variables. According to Eq. 4.3, the imputations $Y_{mis}^{(t)}(j)$ for an imputation variable y_j during an iteration t can be generated according to different models, so that imputation methods can be formally represented by $\Pi = (\Pi_1, \ldots, \Pi_k)$, where $\Pi_j = (y_j, \{x_j\}, \pi_j)$ is an elementary imputation method generating imputations for the j-th imputation variable y_j by the method π_j conditionally on the completely observed set of predictor variables $\{x_j\}$. An imputation method Π generating imputations for more than one imputation variable is called a compound imputation method. An inventory of the different methods π for numerical and for categorical imputation variables y is made below. Technical details about these methods can be found in Appendix 4.B.

Categorical y:

The standard method is logistic or polytomous regression imputation, depending on whether y is a binary variable, or a polytomous variable (consisting of more than two categories). The logistic regression model with parameters $\phi = \beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is [18]:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p; \quad \text{with} \quad \pi = P(y = 1|x_1, \ldots, x_p) \tag{4.8}$$

Polytomous regression for s categories is a generalization of logistic regression and can be modelled as a series of separate logistic regression models of the categories $1, \ldots, s-1$ against a baseline category 0 according to [19]:

$$ln\left(\frac{P(y = j|x)}{P(y = 0|x)}\right) = \beta_{j0} + \beta_{j1} x_1 + \cdots + \beta_{jp} x_p \; ; \; j = 1, \ldots, s-1 \tag{4.9}$$

If y has an ordinal scale, regression imputation with the round off option (see numerical y) can be applied. For this situation, a polytomous regression model can also be defined [20], but

the estimation of its parameters is rather complicated and will not be considered here.

When most of the predictor variables are numerical an alternative method for logistic or polytomous regression imputation is Discriminant imputation. Categorical predictor variables are replaced by their corresponding dummy variables. The starting point of this method is the rule of Bayes

$$P(y=j|x) = \frac{P(x|y=j)P(y=j)}{\sum_{\nu=0}^{s-1} P(x|y=\nu)P(y=\nu)} \quad ; \quad j=0,\ldots,s-1. \tag{4.10}$$

Under the assumption that $x=(x_1,\ldots,x_p)$ given $y=j$ is normally distributed with a mean vector μ_j and covariance matrix Σ_j, the imputation model is

$$P(y=j|x) = \frac{f(x|\mu_j;\Sigma_j)\pi_j}{\sum_{\nu=0}^{s-1} f(x|\mu_\nu;\Sigma_\nu)\pi_\nu} \quad ; \quad j=0,\ldots,s-1. \tag{4.11}$$

In Eq. 4.11, π_j is the probability that $y=j$ and $f(.|\mu,\Sigma)$ is the probability density function of a multivariate normal distribution with a mean vector μ and a covariance matrix Σ. Let X be the currently completed data for $x_1,\ldots,x_p$, and X_{obs} and X_{mis} the rows of X corresponding to the observed values Y_{obs} and missing values Y_{mis} of y, respectively. In each of the three methods logistic regression imputation, polytomous regression imputation and Discriminant imputation, an imputation $Y_{mis}^{\star}$ is generated by first drawing $\phi^{\star}$ from an appropriate posterior distribution of the model parameters ϕ given Y_{obs} and X_{obs} and subsequently generating $Y_{mis}^{\star}$ from the predictive distribution $P\left(Y_{mis} \mid X_{mis};\phi^{\star}\right)$. Technical details about these three method are found in Appendix 4.B.

Numerical y:

The standard method is linear regression imputation, the model of which is similar to Eq. 4.4:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad ; \quad \varepsilon \sim N\left(0,\sigma^2\right) \tag{4.12}$$

In Eq. 4.12, ε is a normally distributed error-term, which is independent of $x_1,\ldots,x_p$. This is called the normal error-term variant. Linear regression imputation is applied to the imputation variables *rs*, *rd* and *ps* in example 2. Linear regression imputation can be easily adjusted for

non-linear relationships between y and $x_1, \ldots, x_p$, and for heteroscedastic, skew or heavy-tailed error-terms, which are the four basic deviations from the standard linear regression model in Eq. 4.12.

To adjust for non-linear relationships between y and $x_1, \ldots, x_p$, the linear regression model is generalized to a rich family of regression models by allowing higher order regression and by allowing transformations for one or more model variables. The first order regression model is the usual linear regression model. When this model is extended with all quadratic and cross-product terms, a second order regression model is obtained. Higher order regression models are rarely used. The dependent variable y and the numerical predictor variables x_j may be transformed using Box-Cox [21] and Power transformations [22], respectively. Both families of transformations are extended by a location parameter to avoid numerical problems. The only difference between these two families of transformations is the scaling factor and the constant term. This different treatment of the dependent variable and the predictor variables is common in literature [22].

Linear regression imputation is made robust against skew or heavy-tailed error terms by applying the hot-deck error-term variant [1]. In this variant, imputations y_i^* for the missing data entry y_i are generated according to $y_i^* = \widehat{y}_i^* + e_i^*$. The value $\widehat{y}_i^*$ is the predicted outcome for y_i from the regression coefficients β^* drawn from its posterior distribution. The error term e_i^* is drawn from the empirical distribution of a suitable subset of the residual error-terms $e_j = y_j - \widehat{y}_j$ for which y_j is observed with $\widehat{y}_j$ the predicted outcome of y_j from the coefficients $\widehat{\beta}$ estimated by ordinary least squares (OLS). When this subset is chosen as $\{e_{i_1}, \ldots, e_{i_q}\}$ such that the q corresponding predicted outcomes $\widehat{y}_{i_1}, \ldots, \widehat{y}_{i_q}$ are closest to the predicted outcome $\widehat{y}_i$ for y_i, linear regression imputation will be also robust against heteroscedastic error-terms. A reasonable choice for q may be $0.3 * n_{obs}$, with n_{obs} the number of observed values for y.

An extra option of regression imputation, which is useful to avoid the imputation of values outside the domain of y, is the round off option. With this option, the generated imputation y_i^* is replaced by $\widetilde{y}_i$, which is the observed value of y closest to y_i^*.

When relationships between y and $x_1, \ldots, x_p$ are too complex to specify a regression model that adequately fits the data, nearest neighbour imputation is an alternative method. With nearest neighbour imputation an imputation y_i^* for a missing data entry y_i is generated by

drawing y_i^* from an estimate $\widehat{P}\left(y_i \mid X_i^T\right)$ of the predictive distribution of y_i given the corresponding row X_i^T of X . In order to reflect the uncertainty about $P\left(y_i \mid X_i^T\right)$, the imputation y_i^* is generated from $\widehat{P}\left(y_i \mid X_i^T\right)$ according to the Bayesian bootstrap method. The estimate $\widehat{P}\left(y_i \mid X_i^T\right)$ is the empirical distribution of the observed values $y_{i_1}, \ldots, y_{i_q}$ of y which are chosen such that the corresponding rows $X_{i_1}^T, \ldots, X_{i_q}^T$ are the $q = [f_{dc} * n_{obs}]$ rows of X_{obs} closest to X_i^T, with f_{dc} the donor class fraction and [.] the entier function. In this context, 'close' is defined by a distance function $d = d\left(X_i^T, X_j^T\right)$, with X_i^T and X_j^T the i-th and j-th row of X. A reasonable value of f_{dc} may be 0.1.

For numerical X, a distance function d may be the Euclidean distance, but for reasons of efficiency the Hamming distance function is chosen:

$$d(X_i, X_j) = \sum_{t=1}^{p} \left| \widetilde{X}_{it} - \widetilde{X}_{jt} \right|, \tag{4.13}$$

where $\widetilde{X}_{it}$ and $\widetilde{X}_{jt}$ are the t-th components of the i-th and j-th rows of the matrix $\widetilde{X}$ obtained from X by standardizing its components to unit variance. Rather than X, the standardized matrix $\widetilde{X}$ is used to make d comparable across components. The Hamming distance in Eq. 4.13 is appealing by its simplicity, but has as a shortcoming that it does not take the correlations between predictor variables into account; if two predictor variables x_i and x_j are strongly positively correlated, they will be double counted by d in Eq. 4.13. A distance function which does take these correlations into account is the Mahalanobis distance function [23] ρ:

$$\rho(X_i, X_j) = (X_i - X_j)^T S^{-1} (X_i - X_j) \tag{4.14}$$

In Eq. 4.14, S is the sample covariance matrix of X. The associations between y and predictor variables x_t may be used as weight factors for the different components. The technical details of linear regression imputation and nearest neighbour imputation are given in Appendix 4.B. Distance functions for categorical predictor variables may also be developed but are outside the scope of this thesis.

4.2.3 Strategy for selecting an imputation method

In order to find an imputation method that efficiently uses existing relationships between variables and generates imputations from a statistical model that adequately fits to the data, a selection strategy in two steps is proposed. The initial set of imputation variables is the set of incomplete variables of interest, i.e., variables for which a target statistic $\left(\widehat{Q}, U\right)$ is requested, containing missing data entries. For each imputation variable y, relevant predictor variables are selected. Incomplete predictor variables other than imputation variables are added to the set of imputation variables and additional predictor variables are selected for them. Secondly, for each imputation variable, a method π is selected. In this chapter, step 2 is described for numerical imputation variables y only.

Step 1: Selection of predictor variables

Given a set of candidate predictor variables selected on grounds of plausibility and target statistics, for each imputation variable y, predictor variables x are selected according to the following four steps:

1. **Variables involved in a requested multivariate statistical analysis:** If x and y are involved in a requested multivariate target statistic $\left(\widehat{Q}, U\right)$ and x is not chosen as a predictor variable for y, then multiple imputation may be improper for $\left(\widehat{Q}, U\right)$ [7]. An example is the estimation of the correlation coefficient between two variables y and x, where y contains missing data and x is not used for the generation of imputations for y. If the correlation between these variables is high, then a large number of values, unlikely in combination with x, will be imputed for y. Consequently, the m completed samples yield estimates of the correlation coefficient biased toward zero. If y is involved in a multivariate target statistic $\left(\widehat{Q}, U\right)$, it is not necessary to select every variable x involved in $\left(\widehat{Q}, U\right)$. The following two cases are distinguished:

   (a) If $\widehat{Q}$ is an association measure, e.g., the correlation coefficient or multiple correlation coefficient, the variable(s) x involved in $\widehat{Q}$ should be selected.

   (b) If $\left(\widehat{Q}, U\right)$ is the result of a regression analysis of which y is an independent variable, the dependent variable of the regression should be selected. It is not necessary to

also select the independent variables x other than y since such variables are taken into account in step 3. When y is involved in an interaction term of the regression model, the other variables involved in this interaction term should be selected as well.

2. **Number and fraction of usable cases:** If a candidate predictor variable x has a large fraction of missing entries, then such a variable may be useless as a predictor variable. A sufficiently large number of cases $n_{obs}(y,x)$ with y and x simultaneously observed is necessary for fitting a model describing the relationship between the two variables. A sufficiently large fraction $f_p(y,x)$ of cases with x observed among all cases with y missing is necessary to be useful for the prediction of the missing data of y. Consequently, candidate predictor variables other than those selected in step 1 with $n_{obs}(y,x)$ and fraction $f_p(y,x)$ smaller than certain minimum values n_0 and f_0, are removed from the set of candidate predictor variables. Reasonable choices for n_0 and f_0 may be 50 and 0.3, respectively.

3. **Variables related to y:** Variables x, which are strongly related to y are useful for the prediction of missing values of y. Inclusion of such predictor variables may reduce the fraction of missing information due to missing data [1] and render a possibly MNAR missing data mechanism closer to MAR.

   An approach which is easy to implement is to select predictor variables x on the basis of bivariate association measures between y and x as estimated from the cases for which y and x are simultaneous observed. For numerical y, the absolute value of the Pearson product-moment correlation coefficient can be used as an association measure and for categorical y, the Cramer-C measure and Kruskal's lambda statistic L_B are useful [24]. Another useful association measure is the internal consistency which is related to homogeneity analysis [25]. To compare numerical predictor variables with categorical predictor variables, an adjusted correlation coefficient λ (see Appendix 4.A) can be used for numerical y, and for categorical y adjusted measures are obtained by discretizing the numerical predictor variables x and calculating association measures from the resulting contingency table. The association measure $\lambda(y,x)$ for numerical y and categorical x is comparable with the absolute value of the Pearson product-moment correlation coefficient between y and

$\widehat{y} = E[y \mid z]$ for another numerical predictor variable z. When y and z are linearly related, $\lambda(y,x)$ can be compared to the absolute value of the Pearson product-moment correlation coefficient between y and z.

Disadvantages of the approach mentioned above are, that a predictor variable x which is strongly related to y, may become redundant when other predictor variables are selected, and that a predictor variable x with a weak association with y, may be an important predictor variable conditionally on other predictor variables. A better approach which does not have these disadvantages is stepwise regression (stepwise linear regression [22] for numerical y and stepwise logistic regression [18] for binary y). When y is polytomous with s, $s \geq 3$, categories, predictor variables are selected by applying $s-1$ separate stepwise logistic regressions of the categories $1,\ldots,s-1$ against a baseline category 0 and joining the sets of predictor variables resulting from these stepwise regressions.

4. **Variables related to the nonresponse of y:** Such variables have a relevant association with the nonresponse indicator R_y of y, and are in this sense explanatory variables for the nonresponse of y. Not including such variables may yield invalid analysis under MAR. As a measure of association, the Pearson product-moment correlation coefficient between R_y and x is used. To compare numerical predictor variables with categorical predictor variables, the adjusted correlation coefficient λ (see Appendix 4.A) can be used. In this step, predictor variables x with a correlation coefficient between R_y and x larger than a certain value ρ_0, say $\rho_0 = 0.2$, and which are not included in the steps 1 and 3 and not excluded in step 2, are selected.

Step 2: Selection of elementary imputation methods

An obvious strategy is to first find an adequate regression model describing the relationship between y and its predictor variables. If this regression model fits the data adequately, regression imputation is chosen and a choice is made between the normal- and the hot-deck error term variant. When y is ordinal or the number of different values of y is small, it is obvious to apply the round off option as well. If no adequate regression model can be found, nearest neighbour imputation is the method of choice.

If y is not involved in a requested multivariate statistical analysis, then generally the standard linear regression model (first order and no transformations) will be sufficient. Otherwise, a regression model for y is chosen with more care. In the first place, the imputation model is chosen in such a way that it includes the model for the requested statistical analysis [7,26]. E.g., if second order regression is requested, then these second order terms are taken into account in the imputation model to avoid improper imputation for coefficients of second order terms significantly differing from zero. Secondly, a regression model which optimally describes the relationship between y and its predictors $x_1, \ldots, x_p$ is found. This is especially important when imputed data sets are to be created to which multiple statistical analyses will be applied. If imputations are generated according to this criterion, the existing structures in the incomplete sample and the uncertainty about these structures are preserved in the m completed samples. In case of a analysis model which does not correspond to the imputation model, the consequences are then less serious than when the structures are not preserved in the m completed data sets.

A popular optimality criterion for the regression model is the squared multiple correlation coefficient R^2. It is equal to the squared Pearson product-moment correlation coefficient between the observed and predicted outcome, and is regarded as a measure of the predictive power of a regression model. An optimal regression model can be found by finding transformations for y and $x_1, \ldots, x_p$, such that R^2 is maximized for second order regression. The value of R^2 is estimated from the cases where y and $x_1, \ldots, x_p$ are completely observed. If the number of such cases is too small, R^2 can be maximized for subsets of $x_1, \ldots, x_p$, for which the number of completely observed cases is sufficiently large. When the optimal transformations are found and the resulting value of R^2 for second order regression is significantly larger than for first order regression, then second order regression is chosen. Otherwise, it is sensible to opt for first order regression. The differences between the values of R^2 for first and second order regression can be examined by the F test.

Finding optimal transformations is a combinatorial problem; even a relatively small number of numerical predictor variables leads to a combinatorial explosion. A heuristic approach to this problem is, starting with an initial selection, to sequentially select for each variable the transformation which maximizes R^2, until R^2 no longer improves. A reasonable initial selec-

tion can be obtained by selecting for each numerical predictor variable x_j, the transformation which maximizes the correlation between y and the transformed x_j. If there is a theoretical or empirical basis for a set of transformations, of course this set should be taken as the initial set of transformations. Suppose there is empirical evidence to suggest that the logarithm of blood pressures is normally distributed as in [27]. It is then obvious to select the logarithmic transformation as the initial transformation for variables containing measurements of blood pressures.

To avoid incompatible regression models for different variables to be imputed in case of circularities, i.e., when there exists a pair of variables (z_1, z_2) such that z_2 is a predictor variable of z_1 and, in turn, z_1 is a predictor variable of z_2, the selection of transformations for such variables z_1 and z_2 should be made with care. If in the imputation model for z_1, transformations f_1 and f_2 are selected for z_1 and z_2, and in the imputation model for z_2 the selected transformations for z_2 and z_1 are given by g_2 and g_1, it can be verified that f_1, f_2, g_1 and g_2 must satisfy: $g_2 \equiv g_1 \circ f_1^{-1} \circ f_2$, where for two functions f and g and values x in the domain of g the function $f \circ g$ is evaluated by $(f \circ g)(x) = f(g(x))$.

A choice between regression- and nearest neighbour imputation is based in the first place on the R^2 statistic. If R^2 is relatively high, a good model fit is indicated and regression imputation is the method to be preferred. However, a relatively low R^2 does not automatically indicate an inadequate model, since it may also be due to a large residual variance. The adequacy of the regression model can be in general investigated by plotting the residuals $e_i = y_i - \widehat{y}_i$ against the predicted values $\widehat{y}_i$. With such plots, deviations from linearity and heteroscedasticity can be detected [22]. To choose between the normal and the hot-deck option, the distribution of the error-term is considered. Information about this distribution cannot be obtained from residual plots. Summary statistics, such as skewness and kurtosis of the residuals $\{e_i\}$ are useful for this purpose.

4.3 Inference from multiple imputation

4.3.1 Pooling of results

Pooled results of point-estimators, standard errors, confidence intervals, and p-values are calculated from the m completed sample results [1] $\left\{\left(\widehat{Q}_1^*, U_1^*\right), \ldots, \left(\widehat{Q}_m^*, U_m^*\right)\right\}$, where $\widehat{Q}_i^*$ is a point-estimate of a possibly multi-dimensional parameter of interest Q obtained from the i-th completed sample and U_i^* is the corresponding estimated variance of $\widehat{Q}_i^*$ (variance-covariance matrix for multivariate Q). Statistics $\left(\widehat{Q}, U\right)$ of the hypothetical complete sample, are estimated by $\left(\overline{Q}_m, \overline{U}_m\right)$, where $\overline{Q}_m$ given by

$$\overline{Q}_m = \frac{1}{m}\sum_{i=1}^{m}\widehat{Q}_i^* \tag{4.15}$$

is a pooled point-estimator, and $\overline{U}_m$ given by

$$\overline{U}_m = \frac{1}{m}\sum_{i=1}^{m}U_i^* \tag{4.16}$$

is the average completed data variance. Extra inferential uncertainty about Q due to missing data is reflected by the between imputation variance B_m given by

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{Q}_i^* - \overline{Q}_m\right)\left(\widehat{Q}_i^* - \overline{Q}_m\right)^T \tag{4.17}$$

On the basis of the lower order variability of U [1,5], the precision of $\overline{U}_m$ with regard to U needs not be reflected. Total inferential uncertainty about Q given the incomplete data is reflected by the total variance T_m given by

$$T_m = \overline{U}_m + B_m + m^{-1}B_m. \tag{4.18}$$

In the definition of T_m in Eq. 4.18, the following three sources of inferential uncertainty about Q are taken into account: (1) uncertainty with respect to the complete data represented by $\overline{U}_m$, (2) extra uncertainty due to missing data represented by B_m, and (3) extra uncertainty due to the simulation error represented by $m^{-1}B_m$. When m tends to infinity, this latter term vanishes.

Example 3 *For the same incomplete sample, parameter of interest Q and imputation method as treated in Example 2 (see subsection 4.2.1) for $m = 10$ and after 10 Gibbs sampling iterations, the corresponding multiple imputation estimates are given by $\overline{Q}_m = 149.04$, $\overline{U}_m = 1.42$, $B_m = 1.00$ and $T_m = 1.42 + 1.00 + 1.00/10 = 2.52$.*

For pooling of results, a distinction is made between a univariate- and a multivariate parameter of interest Q.

Univariate parameter of interest

For univariate Q, the three variances $\overline{U}_m$, B_m, and T_m are univariate. The pooled standard error is given by $\sqrt{T_m}$, the pooled $(1-\alpha)100\%$ confidence interval is given by

$$\overline{Q}_m \pm t_{\nu;1-\alpha/2}\sqrt{T_m}, \tag{4.19}$$

and the pooled p-value with regard to the hypothesis $H_0 : Q = Q_0$, with Q_0 a chosen value, is given by

$$\text{p-value} = P\left(F_{1,\nu} > \left(Q_0 - \overline{Q}_m\right)^2 / T_m\right). \tag{4.20}$$

In the equations 4.19 and 4.20, $t_{\nu;1-\alpha/2}$ is the $1-\alpha/2$ quantile of a Student t distribution with ν degrees of freedom and $F_{1,\nu}$ is an F distributed random variable with one and ν degrees of freedom, where ν is given by,

$$\nu = (m-1)\left(1 + r_m^{-1}\right)^2, \tag{4.21}$$

and r_m is given by

$$r_m = \left(1 + m^{-1}\right) B_m / U_m. \tag{4.22}$$

In Eq. 4.22, r_m is the relative increase in variance due to nonresponse. When the relative contribution of the missing data to the inferential uncertainty about Q increases, r_m also increases. The simulation error of r_m, with regard to r_∞, in case of an infinite number of imputations, is reflected by the factor $m^{-1}B_m/\overline{U}_m$. The t_ν and $F_{1,\nu}$ distributions in the equations 4.19 and 4.20 are approximations proposed in [1] to reflect the simulation error of B_m with regard to B_∞. When m tends to infinity, ν also tends to infinity, so that the corresponding distributions t_ν and $F_{1,\nu}$ will converge to the standard normal distribution and the χ^2_1 distribution, respectively

[28]. For an infinite number of imputations, the pooled $(1-\alpha)\,100\%$ confidence intervals and pooled p-values are then given by

$$\overline{Q}_\infty \pm z_{1-\alpha/2}\sqrt{T_\infty} \tag{4.23}$$

and

$$\text{p-value} = P\left(\chi_1^2 > \left(Q_0 - \overline{Q}_\infty\right)^2 / T_\infty\right), \tag{4.24}$$

where

$$T_\infty = \overline{U}_\infty + B_\infty. \tag{4.25}$$

In Eq. 4.23, $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Since the Student t distribution has thicker tails than the standard normal distribution, the value $t_{\nu;1-\alpha/2}$ is larger than $z_{1-\alpha/2}$ [28], so that the $(1-\alpha)\,100\%$ confidence interval for a finite number of imputations is indeed wider than for an infinite number of imputations. In Eq. 4.20, the random variable $F_{1,\nu}$ has the same probability distribution as the ratio $\frac{\chi_1^2}{\chi_\nu^2/\nu}$, where χ_1^2 and χ_ν^2 are two independent χ^2 random variables with 1 and ν degrees of freedom, respectively. When m tends to infinity, χ_ν^2/ν tends to 1, so that the loss in significance due to the simulation error is reflected by the variability of the denominator χ_ν^2/ν.

Multivariate parameter of interest

Methods for pooling of results for a multivariate parameter of interest Q described here concerns pooling of p-values with respect to Q. Three different cases are considered.

Pooling of p-values when corresponding completed data point-estimates and correlation matrices are given For a k-dimensional parameter of interest Q, a p-value for the hypothesis $H_0 : Q = Q_0$ [1,3], e.g., for a linear regression model, the hypothesis that the coefficients are simultaneously equal to zero, is given by

$$\text{p-value} = P\left(F_{k,w} > D_m\right). \tag{4.26}$$

In Eq. 4.26, $F_{k,w}$ is an F distributed random variable with k and w degrees of freedom and D_m is a test statistic given by

$$D_m = \left(\overline{Q}_m - Q_0\right)^T \overline{U}_m^{-1} \left(\overline{Q}_m - Q_0\right) / \left[k\left(1 + r_m\right)\right], \tag{4.27}$$

where r_m is given by

$$r_m = \left(1 + m^{-1}\right) \left(trB_m \overline{U}_m^{-1}\right) / k. \tag{4.28}$$

The definition of r_m in Eq. 4.28 is a multivariate extension of the definition of r_m in Eq. 4.22 for univariate Q. The best choice so far for w in Eq. 4.26 is given by [3]:

$$w = \begin{cases} 4 + (t-4)\left[1 + \left(1 - 2t^{-1}\right)\right] / r_m^2 & \text{if } t > 4 \\ (m-1)\left(\frac{k+1}{2}\right)\left(1 + r_m^{-1}\right)^2 & \text{otherwise} \end{cases} \tag{4.29}$$

with t given by

$$t = k\left(m - 1\right). \tag{4.30}$$

The factor $\left(\overline{Q}_m - Q_0\right)^T \overline{U}_m^{-1} \left(\overline{Q}_m - Q_0\right)$ in the expression for D_m in Eq. 4.27 is an estimate for the corresponding complete data test statistic $D = \left(\widehat{Q} - Q_0\right)^T U^{-1} \left(\widehat{Q} - Q_0\right)$ (see section 4.4). This factor is divided by the dimension k of Q since the reference distribution of D is an χ^2 distribution with k degrees of freedom and the F reference distribution of D_m is equal to the distribution of the random variable $\frac{\chi_k^2/k}{\chi_w^2/w}$, where χ_k^2 and χ_w^2 are two independent random variables with a χ^2 distribution with k and w degrees of freedom. The term $1 + r_m$ in the denominator of Eq. 4.27 reflects the extra uncertainty due to missing data. When r_m increases, D_m decreases so that the p-value in Eq. 4.26 increases and thus becomes less significant. The extra uncertainty due to the simulation error is reflected by the denominator degrees of freedom w of the F distribution of D_m. When the number of imputations m increases, w increases according to Eq. 4.29 and Eq. 4.30, which implies that the degree of concentration around 1 of the probability distribution of the term χ_w^2/w increases and thus the variability of the F reference distribution of D_m decreases.

The p-value in Eq. 4.26 is derived under the strong assumption that the theoretical between imputation variance-covariance matrix B (see section 4.4 for the definition of B) is proportional to U, i.e., $B = \overline{\lambda}U$ for a certain constant $\overline{\lambda}$. Nevertheless, a Monte Carlo study in [3] suggests

that even if this assumption is violated, performance of this procedure is acceptable in practice.

Pooling of p-values when only m completed data p-values are given The definition of D_m requires that the variance-covariance matrix U of $\widehat{Q}$ is available. For large statistical models, such as large log-linear models with possibly hundreds of parameters, however, presentation of the correlation matrices of $\widehat{Q}$ are cumbersome and may not be available in statistical software packages. For this purpose, a method using only the completed data p-values has been developed [4]. By this method, pooled p-values of the hypothesis $H_0 : Q = Q_0$ are derived from the corresponding completed data test statistics $\{d_1^*, \ldots, d_m^*\}$ of this hypothesis, where the i-th completed data statistic is equal to

$$d_i^* = \left(\widehat{Q}_i^* - Q_0\right)^T U_i^{*-1} \left(\widehat{Q}_i^* - Q_0\right). \tag{4.31}$$

If a statistical software package provides m completed data p-values but not m completed data test statistics, the latter can be derived from the relationship

$$i\text{-th completed data p-value} = P\left(\chi_k^2 > d_i^*\right). \tag{4.32}$$

The pooled p-value is given by

$$\text{p-value} = P\left(F_{k,w} > \widehat{\widehat{D}}_m\right). \tag{4.33}$$

In Eq. 4.33, $\widehat{\widehat{D}}_m$ is an estimate of D_m in Eq. 4.27 given by

$$\widehat{\widehat{D}}_m = \frac{\overline{d}_m/k - \left(\frac{m-1}{m+1}\right)\widehat{r}_m}{1+\widehat{r}_m}, \tag{4.34}$$

where $\widehat{r}_m$ is an estimate of r_m given by

$$\widehat{r}_m = \left(1+m^{-1}\right)\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(\sqrt{d_i^*} - \overline{\sqrt{d_m}}\right)^2\right]. \tag{4.35}$$

In Eq. 4.35, $\widehat{r}_m$ is equal to $\left(1+m^{-1}\right)$ times the sample variance of the square roots of the m complete data test statistics $\{d_1^*,\ldots,d_m^*\}$ and $\overline{\sqrt{d_m}}$ denotes the average over $\left\{\sqrt{d_1^*},\ldots,\sqrt{d_m^*}\right\}$. The denominator degrees of freedom w in Eq. 4.33 is given by

$$w = k^{-3/m}\,(m-1)\left(1+\widehat{r}_m^{-1}\right)^2 \tag{4.36}$$

In Eq. 4.34, the estimate of D_m is denoted by $\widehat{\widehat{D}}_m$ rather than by $\widehat{D}_m$ to indicate that D_m is approximated in two steps. First D_m is approximated by

$$\widehat{D}_m = \frac{\overline{d}_m - \left(\frac{m-1}{m+1}\right) r_m}{1+r_m}. \tag{4.37}$$

Subsequently, $\widehat{D}_m$ is approximated by $\widehat{\widehat{D}}_m$ by replacing r_m in Eq. 4.37 by the estimate $\widehat{r}_m$. The extra uncertainty due to missing data is reflected by the term $-\left(\frac{m-1}{m+1}\right)\widehat{r}_m$ in the numerator in Eq. 4.34 and the term $1+\widehat{r}_m$ in the denominator in Eq. 4.34. Similar to the method for pooling p-values on the basis of D_m, specified in the equations 4.26 through 4.30, the simulation error is reflected by the denominator degrees of freedom w of the reference F distribution given in Eq. 4.36 which increases when the number of imputations m increases.

However, performance of this method is inferior to the method using D_m, and it is advised to use this method only as a rough guide and to interpret its results as providing a range of p-values between one half and twice the calculated p-value [5,9]. That the performance of the method using $\widehat{\widehat{D}}_m$ is considerably less than the method using D_m can be understood from the fact that the completed data test statistics $\{d_1^*,\ldots,d_m^*\}$ only provide information about the magnitude of the differences between $\widehat{Q}_i^*$ and Q_0 but not of the direction of these differences [5].

Pooling of p-values from likelihood ratio tests For p-values from likelihood-ratio tests, a third method, which may be regarded as intermediate between the previous two [9], is proposed in [6]. Likelihood-ratio tests are used to decide if a given statistical model A describing a certain sample Y can be replaced by a more parsimonious statistical model B which is embedded in A. For the models A and B, the corresponding log-likelihood ratio test statistic d_L is given by $d_L\left(A,B \mid Y\right) = 2\left(ll\left(A;\widehat{\psi} \mid Y\right) - ll\left(B;\widehat{\psi}_0 \mid Y\right)\right)$, where $ll\left(M;\theta \mid X\right)$ is the log-likelihood

function for a statistical model M with parameter values θ and a sample X, and $\widehat{\psi}$ and $\widehat{\psi}_0$ are the maximum likelihood estimates of the parameters ψ and ψ_0 of A and B. The log-likelihood ll provides a measure for model fit. If $ll\left(A;\widehat{\psi} \mid Y\right) > ll\left(B;\widehat{\psi}_0 \mid Y\right)$, model A fits the data Y better than model B. Thus, $d_L(A,B \mid Y)$ measures the loss in modelfit to the data Y if model B rather than model A is chosen. The factor 2 in the definition of d_L is chosen, since for a sufficiently large size of sample Y, the corresponding p-value of d_L is given by

$$\text{p-value} = P\left(\chi_k^2 > d_L(A,B \mid Y)\right), \tag{4.38}$$

where k is the difference between the numbers of parameters for the models A and B, and χ_k^2 is a random variable with a χ^2 distribution with k degrees of freedom. If this p-value is significant, e.g., smaller than 0.05, it is obvious to choose model A, otherwise it is convenient to opt for model B.

An example of the use of p-values from likelihood ratio tests is to asses the individual contribution of an independent variable x to a logistic regression model (see chapter 7 for an application). In this example, model A is a logistic regression for which x is an independent variable, and model B is the reduced logistic regression model resulting from exclusion of x from A. The parameters $\widehat{\psi}$ and $\widehat{\psi}_0$ are the regression coefficients of the full and the reduced regression model obtained by maximum likelihood estimation. If x is numerical, $k = 1$, and if x is categorical k is equal to the number of categories of x minus 1. When the likelihood ratio p-value of x is significant, x is an important variable for the regression model.

When the sample Y is incompletely observed by Y_{obs}, the unknown complete data likelihood ratio test statistic $d_L = d_L(A,B \mid Y)$ is estimated by the average $\overline{d}_p$ over $d_p^{(1)}, \ldots, d_p^{(m)}$, with $d_p^{(i)} = 2\left(ll\left(A;\overline{\psi} \mid Y_i^*\right) - ll\left(B;\overline{\psi}_0 \mid Y_i^*\right)\right)$ the log-likelihood ratio test statistic for the i-th completed sample Y_i^* and the pooled point estimates $\overline{\psi}$ and $\overline{\psi}_0$ for ψ and ψ_0, respectively. The pooled log-likelihood ratio test statistic $\overline{d}_L$ is given by

$$\overline{d}_L = \frac{\overline{d}_p}{k(1+r_L)}. \tag{4.39}$$

In Eq. 4.39 r_L is an estimate of the relative increase in variance due to missing data and is

given by

$$r_L = \frac{m+1}{k\,(m-1)}\left(\overline{d}_c - \overline{d}_p\right), \tag{4.40}$$

where $\overline{d}_c$ is the average of $\overline{d}_c^{(1)}, \ldots, \overline{d}_c^{(m)}$, with $\overline{d}_c^{(i)} = 2\left(ll\left(A; \widehat{\psi}^{(i)} \mid Y_i^*\right) - ll\left(B; \widehat{\psi}_0^{(i)} \mid Y_i^*\right)\right)$ is the log-likelihood evaluated for the maximum likelihood estimates $\widehat{\psi}^{(i)}$ and $\widehat{\psi}_0^{(i)}$ for ψ and ψ_0 from the i-th completed sample Y_i^*. The difference $\overline{d}_c - \overline{d}_p$ indicates the imprecision of the estimate $\overline{d}_p$ of the complete data log-likelihood ratio d_L. The extra uncertainty due to missing data is reflected by dividing $\overline{d}_p$ by the term $1 + r_L$. The corresponding p-value is given by

$$\text{p-value} = P\left(F_{k,w} > \frac{\overline{d}_L}{}\right), \tag{4.41}$$

where $w = w(r_L)$ is obtained by replacing r_m by r_L in the expression for the denominator degrees of freedom w in Eq. 4.29 for the method using the corresponding pooled point estimates $\overline{Q}_m$ and average variance $\overline{U}_m$.

4.3.2 Missing information

Diagnostic measures for assessing the contribution of missing data to inferential uncertainty about point estimates are considered here only for a univariate parameter of interest Q. For multivariate Q, these measures can be separately evaluated for each component of Q. Generalizations of these measures for multivariate Q exist [1,3,5], but their interpretation is difficult.

The standard diagnostic measure is the fraction of information about Q missing due to nonresponse [1] given by

$$\gamma_m = \frac{\overline{U}_m^{-1} - \frac{\nu+1}{\nu+3} T_m^{-1}}{\overline{U}_m^{-1}}. \tag{4.42}$$

In Eq. 4.42, $\overline{U}_m^{-1}$ is an estimate of U^{-1}, which represents the information about Q contained in the complete sample. When U^{-1} increases, the precision of $\widehat{Q}$ with regard to Q also increases. The information about Q contained in the incomplete sample is given by $\frac{\nu+1}{\nu+3} T_m^{-1}$, so that the nominator in Eq. 4.42 represents the missing information about Q due to missing data. The factor $\frac{\nu+1}{\nu+3}$, with ν given in Eq. 4.21, is the correction factor for the finite number of imputations m. For an infinite number of imputations, the fraction of information missing due to missing

data is given by

$$\gamma_\infty = \frac{\overline{U}_\infty^{-1} - T_\infty^{-1}}{\overline{U}_\infty^{-1}}. \tag{4.43}$$

For a variety of univariate point estimators $\widehat{Q}$, such as the mean and the median, it can be verified from Eq. 4.43, that γ_∞ is equal to the fraction of missing data in the theoretical situation that no predictor variables are used for the generation of the imputations. Otherwise, γ_∞ will be smaller than or equal to this fraction.

Other diagnostic measures are the between imputation variance B_m and the relative increase in variance r_m given in Eq. 4.22. The following relationship between γ_m and r_m exists [1]

$$\gamma_m = \frac{r_m + 2/(\nu+3)}{r_m + 1}. \tag{4.44}$$

For γ_∞ and r_∞, this relationship is given by [1]

$$\gamma_\infty = \frac{r_\infty}{1 + r_\infty}. \tag{4.45}$$

The diagnostic measures mentioned above are defined in terms of loss in precision due to missing data. To assess the added value of multiple imputation with regard to complete-case analysis, it is useful to define similar measures for gain in precision of multiple imputation with regard to complete-case analysis. Let $\widetilde{U}$ be the variance of $\widehat{Q}$ when $\widehat{Q}$ is obtained by complete-case analysis. Similar to γ_m in Eq. 4.42, the fraction of information about Q, gained by multiple imputation with regard to complete-case analysis is defined by

$$\zeta_m = \frac{\frac{\nu+1}{\nu+3} T_m^{-1} - \widetilde{U}^{-1}}{\widetilde{U}^{-1}}. \tag{4.46}$$

Similar to the measures B_m and r_m, gain in precision is reflected by $\widetilde{U} - T_m$ and $\frac{\widetilde{U} - T_m}{\widetilde{U}}$, respectively.

Example 4 *From* $\overline{Q}_m = 149.04$, $\overline{U}_m = 1.42$, $B_m = 1.00$ *and* $m = 10$ *in example 3, the corresponding values of* r_m *and* ν *are 0.47 and 88.33, respectively. From Eq. 4.42 it follows that the fraction of information about* Q *missing due to missing data* γ_m *is equal to 0.33. Further,* $\widetilde{U}^{-1} = ((31.8)^2/221 * (1034-221)/1034)^{-1} = 0.279$ *(see Table 2.7 of chapter 2) and* $T_m^{-1} = 0.40$,

so that according to Eq. 4.44, the fraction of information about Q gained by multiple imputation ζ_m is given by 0.28. The factor $(1034-211)/1034$ *is the correction factor for finite populations. The fraction γ_m is lower than the corresponding fraction of missing data entries in ps given by* 0.47, *so that the selected predictor variables for ps provide information about the parameter of interest Q. This also appears from* $\zeta_m = 0.28$.

4.4 Validation

In the next chapter, the validation of some of the imputation methods developed in section 4.2 according to the conditions of proper imputation is described. In this section a non-technical explanation of proper imputation and its similarity with the validity conditions for complete and incomplete data inference is given. In subsection 4.4.1, a detailed description of the validity conditions for complete data inference is given. The similarity of the conditions for proper imputation with the conditions for valid complete data inference and for valid incomplete data inference are described in subsection 4.4.2, where the validation of pooling procedures for finite multiple imputation is also treated. In subsection 4.4.3 is described how the quality of imputations can be inspected for a real life data set.

4.4.1 Complete data inference

In this subsection, classical frequentistic validity criteria for complete data inference are formulated in terms of the tuple $(\widehat{Q}, U)$, where $\widehat{Q}$ is a point estimator of a parameter of interest and U the complete data variance of $\widehat{Q}$. The underlying assumption is that the sample size is sufficiently large. Under this assumption, confidence intervals, test statistics and p-values can be written as functions of the tuple $(\widehat{Q}, U)$.

Validity conditions for complete data inference

Complete sample statistics $(\widehat{Q}, U)$ are validated under the assumption of a true underlying sampling mechanism. The validity conditions for complete data inference under this assumption are given by:

$$\widehat{Q} \sim N(Q, U_0) \tag{4.47}$$

$$U \approx U_0 \tag{4.48}$$

The condition in Eq. 4.47 means that $\widehat{Q}$ is an unbiased estimator of Q and that the sample size is large enough to approximate the sampling distribution of $\widehat{Q}$ by a normal distribution with a mean equal to Q and the theoretical variance equal to U_0. Equation Eq. 4.48 can be interpreted as a correct reflection of the precision of the point-estimator $\widehat{Q}$. The symbol $\approx$ here indicates equality in the sense of lower order variability [1,3]. The distinction between U and U_0, and the frequentistic concept behind the precision of $\widehat{Q}$ will be outlined in the next subsection.

On the basis of the validity conditions in Eq. 4.47 and Eq. 4.48, a $(1-\alpha)100\%$ confidence interval for univariate Q is given by [1]

$$\widehat{Q} \pm z_{1-\alpha/2}\sqrt{U}, \tag{4.49}$$

and, for multivariate Q, a p-value for the hypothesis $H_0 : Q = Q_0$, with Q_0 a chosen value, is given by

$$\text{p-value} = P(\chi^2_k > D), \tag{4.50}$$

where the corresponding test-statistic D is given by

$$D = (\widehat{Q} - Q_0)^T U^{-1} (\widehat{Q} - Q_0). \tag{4.51}$$

In Eq. 4.49, $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. In Eq. 4.50, χ^2_k is a χ^2 random variable with k degrees of freedom. The test statistic D in Eq. 4.51 can be interpreted as a distance function of $\widehat{Q}$ and Q_0 in which the component distances between $\widehat{Q}$ and Q_0 are weighed with the component standard errors of $\widehat{Q}$ and the correlations between these components.

Reflection of the precision of a point estimate

This subsection is illustrated by example 5.

Example 5 *This example is a simulated sample survey. The objective of a sample survey can be stated as statistical inference about a certain population quantity Q, which cannot be determined since the population is too large to be investigated entirely. Rather than from the*

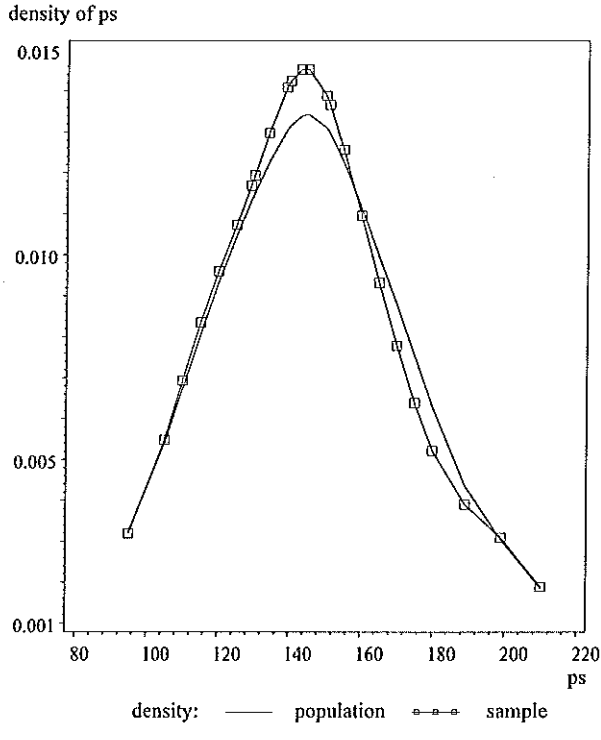


Figure 4-4: Kernel estimates of population and sample.

entire population, statistical inference about Q is obtained from a sample. This sample contains uncertain information about Q, in the sense that Q cannot be exactly determined from it.

In the simulated sample survey, the population Y consists of $n_p = 1034$ *observations of the variable ps (see Example 1) and the parameter of interest Q is the population mean* $Q = \overline{Y}$. *The sample y from Y, is the sample 'Dobutamine Complete', which is randomly drawn from Y by Monte Carlo techniques. In the sampling mechanism by which y is drawn, each entry in the population Y has a probability of* n_s/n_p *to be included in the sample with* n_s *equal to 400. The size of samples by this mechanism is therefore not necessarily exactly equal to 400, but has an expected value of 400. For this reason, the sample size of 'Dobutamine Complete' is not equal to 400 but equal to 394, as can be read from table 2.3. In sample surveys, the sample y is called*

a representative sample since the frequency distribution of y is approximately the same as the frequency distribution of the population Y. This is shown in Figure 4.4, where kernel estimates (see chapter 2) of the probability density distributions of y and Y are displayed. Throughout this section the parameter of interest Q will be the population mean $\overline{Y} = 147.7$. *A point estimator* $\widehat{Q}$ *of Q is the sample mean* $\overline{y}_{n_s} = 146.4$ *(see table 2.3 of chapter 2).*

In frequentistic statistical inference, the precision of $\widehat{Q}$ is indicated by the amount of fluctuation of $\widehat{Q}$ under repetition of sampling by the sampling mechanism from which the available data is assumed to be generated. A general impression of this fluctuation is given by the probability density distribution of $\widehat{Q}$ under this sampling mechanism. In Figure 4.5, these probability density distributions are displayed for the sample mean $\overline{y}_{n_s}$ as obtained from samples with different expected sample sizes given by $n_s = 10, 50, 100, 400$; The underlying sampling mechanism here is similar to the sampling mechanism from which the sample 'Dobutamine Complete' is drawn as described in example 5. For each expected sample size, the corresponding probability density distribution is approximated by generating $N = 1,000$ independent samples from the sampling mechanism by Monte Carlo techniques and calculating a kernel estimate from the corresponding point estimates $\widehat{Q}^{(1)}, \ldots, \widehat{Q}^{(N)}$.

In Figure 4.5, for low n_s, the sample mean $\overline{y}_{n_s}$ appears to be a very imprecise estimator of the population mean Q. According to the probability density distribution of y_{ns} with $n_s = 10$, point estimates with a value lower than 135 or a value larger than 165 are possible; such values of $\widehat{Q}$ deviate considerably from $Q = 147.7$. For increasing n_s the precision of $\overline{y}_{n_s}$ increases, as can be seen from the increasing degree of concentration around Q of the corresponding probability density functions. This is in accordance with the fact that with increasing sample size, the amount of information about Q, contained in the sample, also increases.

A generally accepted summary measure for the precision of $\widehat{Q}$ is the variance U_0 of $\widehat{Q}$ given by:

$$U_0 = Var(\widehat{Q}) = E\left[(\widehat{Q} - E[\widehat{Q}])^2\right] = \lim_{N\to\infty} \frac{1}{N-1} \sum_{i=1}^{N} (\widehat{Q}^{(i)} - \overline{Q}_N)^2 \tag{4.52}$$

$$\overline{Q}_N = \frac{1}{N} \sum_{i=1}^{N} \widehat{Q}^{(i)} \tag{4.53}$$

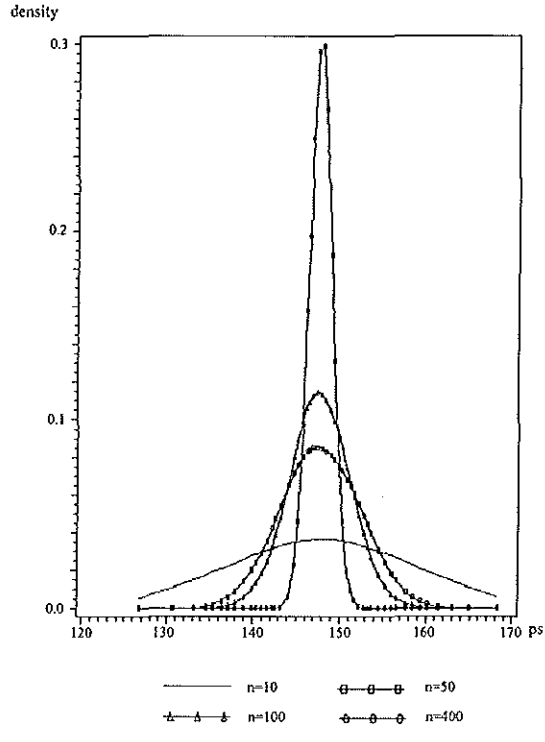


Figure 4-5: Distributions of sample means for different sample sizes.

In Eq. 4.52, $\widehat{Q}^{(1)}, \widehat{Q}^{(2)}, \ldots$ is a sequence of estimates of Q resulting from samples, which are independently generated by the assumed sampling mechanism. The quantity $\overline{Q}_N$ is the average of $\widehat{Q}^{(i)}$ over the N samples, which is an approximation of $E(\widehat{Q})$, the expectation of $\widehat{Q}$ which is usually interpreted as the long-term average of $\widehat{Q}$ under repeated sampling. The variance U_0 can be interpreted as the long-term average squared deviation of $\widehat{Q}$ from its long-term average $E(\widehat{Q})$. To facilitate the interpretation of U_0, the standard error of $\widehat{Q}$ given by $SE_0 = \sqrt{U_0}$ is usually presented. From the probability density function of $\widehat{Q}$ in Figure 4.5, the corresponding standard errors SE_0 are given by 10.25, 4.26, 2.92 and 1.20 for $n_s = 10, 50, 100, 400$, respectively.

The quantity U_0 is a theoretical variance which is not available from the sample. Therefore, it is estimated by the complete data variance U using assumptions about the underlying sampling mechanism. For a sample y from the population Y according to the sampling mechanism in example 5, this complete sample variance U is given by:

$$U = \frac{1}{n_s}\left(\frac{n_p - n_s}{n_p}\right) S_{n_s}^2 \tag{4.54}$$

In Eq. 4.54, S_{n_s} is the standard deviation of y and the factor $\left(\frac{n_p - n_s}{n_p}\right)$ is the so-called correction factor for finite populations. For the sample 'Dobutamine Complete', n_s, n_p and S_{n_s} are given by 394, 1034 and 30.16 (See table 2.3 of chapter 2), so that U is equal to 1.47 and SE is equal to 1.21. This standard error is close to its corresponding theoretical value for $n_s = 400$ given by $SE_0 = 1.20$.

Monte Carlo evaluation of complete data inference

Rather than the equations 4.47 and 4.48, the simplified conditions,

$$E[\widehat{Q}] = Q \tag{4.55}$$

and

$$E[U] = U_0, \tag{4.56}$$

are verified by a Monte Carlo study as depicted in Figure 4.6. In such a study, from a population Y, a large number N of independent samples $y^{(1)}, \ldots, y^{(N)}$ are generated by the assumed sampling mechanism, and the corresponding complete sample results $(\widehat{Q}^{(1)}, U^{(1)}), \ldots, (\widehat{Q}^{(N)}, U^{(N)})$ are calculated from these N samples. From these N results $E[Q]$ is approximated by $\overline{Q}$, the average of $\widehat{Q}^{(i)}$, $E[U]$ is approximated by $\overline{U}$, the average of $U^{(i)}$, and U_0 is approximated by $\widetilde{U}_0$, the sample variance over the $\widehat{Q}^{(i)}$. When N tends to infinity, $\overline{Q}$, $\overline{U}$ and $\widetilde{U}_0$ will tend to $E[Q]$, $E[U]$ and U_0, respectively.

For univariate Q, an indication of the normality condition in Eq. 4.47 can be obtained by verifying whether the actual coverage of the confidence interval is equal to its nominal coverage. The actual coverage is approximated by calculating the fraction of confidence intervals which

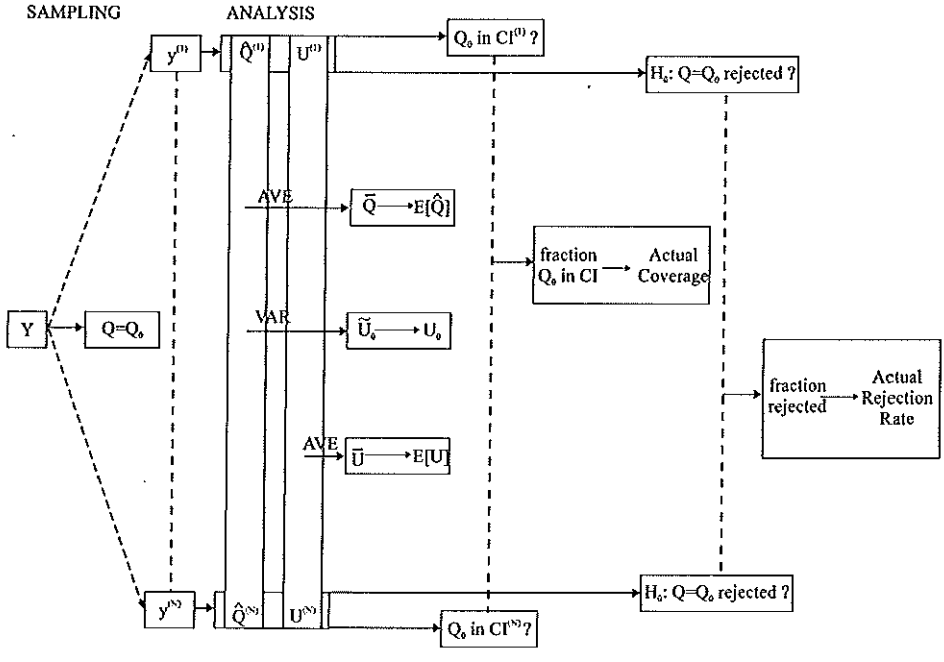


Figure 4-6: Setup for a Monte Carlo evaluation of complete data inference. The terms AVE and VAR represent the sample average and variance, respectively.

include the parameter of interest $Q = Q_0$ over the N confidence intervals $CI^{(i)}$ corresponding to the N complete sample results $(\widehat{Q}^{(i)}, U^{(i)})$. When N tends to infinity, this fraction will tend to the probability that under the sampling mechanism a confidence interval will include the parameter of interest Q_0. The nominal coverage of a confidence interval is this probability under the assumptions on the basis of which this interval is constructed. For the confidence interval given in Eq. 4.49, these assumptions are given by the validity conditions in Eq. 4.55 and Eq. 4.56. The actual and nominal rejection rates for statistical tests are defined in a similar way as depicted in Figure 4.6; the actual rejection rate is the probability that H_0 will be rejected under the assumed sampling mechanism given that H_0 is true. The nominal rejection rate is this probability under the assumptions on the basis of which this test is constructed. In the next example, a small evaluation study of complete data inference is illustrated:

Example 6 *In this example, the sample mean y_{n_s} and its corresponding standard error given*

sample size	*bias*	SE	SE_0	$SE_0 - SE$	$(SE_0 - SE)/SE_0$	*Coverage*
10	*-0.033*	*9.286*	*10.250*	*0.964*	*0.094*	*89.13*
20	*-0.031*	*6.659*	*6.924*	*0.265*	*0.0383*	*92.68*
30	*-0.140*	*5.442*	*5.622*	*0.181*	*0.032*	*93.28*
50	*0.011*	*4.199*	*4.264*	*0.065*	*0.015*	*93.90*
75	*-0.011*	*3.399*	*3.450*	*0.060*	*0.017*	*94.17*
100	*0.016*	*2.903*	*2.917*	*0.014*	*0.005*	*94.18*
150	*-0.005*	*2.304*	*2.288*	*-0.016*	*-0.007*	*94.69*
200	*0.013*	*1.940*	*1.953*	*0.0138*	*0.007*	*94.71*
250	*0.004*	*1.168*	*1.694*	*0.010*	*0.006*	*94.83*
300	*-0.000*	*1.487*	*1.490*	*0.003*	*0.002*	*94.75*
350	*-0.001*	*1.329*	*1.318*	*-0.011*	*-0.008*	*94.84*
400	*0.014*	*1.197*	*1.200*	*0.003*	*0.002*	*94.99*

Table 4.1: Results of a small scale Monte Carlo evaluation of complete data analysis.

in Eq. 4.54 are validated for different sample sizes, ranging from 10 to 400. The population Y, the population quantity of interest Q, and the underlying sampling mechanism are the same as in example 5. The number of generated samples is given by $N = 10.000$. The results are given in Table 4.1. In this table, the bias (condition in Eq. 4.55) is measured by the quantity $Q - \overline{Q}$, where $\overline{Q}$ is an approximation of $E[\widehat{Q}]$. The under/over estimation of the theoretical standard error $\sqrt{U_0}$ is measured by the relative difference $(SE_0 - SE)/SE_0$, where $SE_0 = \sqrt{\widetilde{U}_0}$ and $SE = \sqrt{\overline{U}}$ are approximations of $\sqrt{U_0}$ and $E[\sqrt{U}]$, respectively. In the rightmost column, the actual coverages of 95% confidence intervals according to Eq. 4.49 are given.

The results show that the simplified validity conditions in Eq. 4.55 and Eq. 4.56 hold for all sample sizes. The normality condition of Eq. 4.47, however, is not satisfied for smaller sample sizes ($n_s < 100$), since for these sample sizes the corresponding coverages are clearly lower than the nominal coverage of 95%. For increasing sample size, the actual coverages are increasing toward 95%, which is in agreement with the fact that with increasing sample size the distribution of $\widehat{Q}$ converges to the normal distribution in Eq. 4.47. The undercoverage for the smaller sample sizes follows from the fact that for smaller sample sizes U varies under repetition of sampling, while according to the validity condition $U \approx U_0$ in Eq. 4.48 it is assumed that it does not. For confidence intervals, this variability in U is reflected by replacing the $1 - \alpha/2$ quantile of the standard normal distribution $z_{1-\alpha/2}$ in Eq. 4.49 by the corresponding quantile $t_{n-1;1-\alpha/2}$ of the Student t distribution with $n - 1$ degrees of freedom, where n is the sample size; the quantile

$t_{n-1;1-\alpha/2}$ is larger than $z_{1-\alpha/2}$ and goes to $z_{1-\alpha/2}$ when n goes to infinity. If in this study the actual coverage of confidence intervals on the basis of $t_{n-1;1-\alpha/2}$ is approximated, it can be expected that the actual coverages for smaller sample sizes are close to the nominal coverage of 95%.

4.4.2 Incomplete data inference

Just as the validity conditions for complete data inference in the previous section, proper multiple imputation and valid incomplete data inference require unbiased point estimators, a correct reflection of their precision, and a sample sufficiently large to make valid assumptions about normality. The main distinction between these three validity criteria is that they take different levels of inferential uncertainty into account; valid complete data inference takes into account the inferential uncertainty about Q given the complete sample, proper imputation takes into account the inferential uncertainty about the statistics $(\widehat{Q}, U)$ of the hypothetical complete sample given the incomplete data, and valid incomplete data inference takes into account the inferential uncertainty about Q given the incomplete sample.

Similar to Figure 4.5, these three levels of inferential uncertainty are reflected by probability density distributions in Figure 4.7. In this Figure, the population Y, parameter of interest Q, point estimator $\widehat{Q}$ and the sampling mechanism, with $n_s = 400$, are the same as in example 5, and the missing data mechanism is the MAR mechanism described in chapter 2. The solid graph is a kernel estimate of the sampling distribution of $\widehat{Q}$ from $N = 1,000$ point estimates $\widehat{Q}^{(1)}, \ldots, \widehat{Q}^{(N)}$ corresponding to N independent samples generated by the sampling mechanism. Extra uncertainty due to missing data is represented by the probability density distribution of $\overline{Q}_\infty$ ($\overline{Q}_\infty$ is approximated here with $\overline{Q}_m$ where $m = 100$) under repeated generation of incomplete samples from a fixed complete sample. This probability distribution is approximated by a kernel estimate (graph with squares) from $M = 1000$ multiple imputation estimates $\overline{Q}_\infty^{(1)}, \ldots, \overline{Q}_\infty^{(M)}$ corresponding to M incomplete samples independently generated from a fixed complete sample by the same missing data mechanism. $\overline{Q}_\infty$ is distributed around the estimate $\widehat{Q}$ from this complete sample, and under repeated sampling $\widehat{Q}$ is distributed around the population quantity Q. This implies that the sampling distribution of $\widehat{Q}$ and the distribution of $\overline{Q}_\infty$ under repeated generation of incomplete samples differ in location. This difference appears in Figure

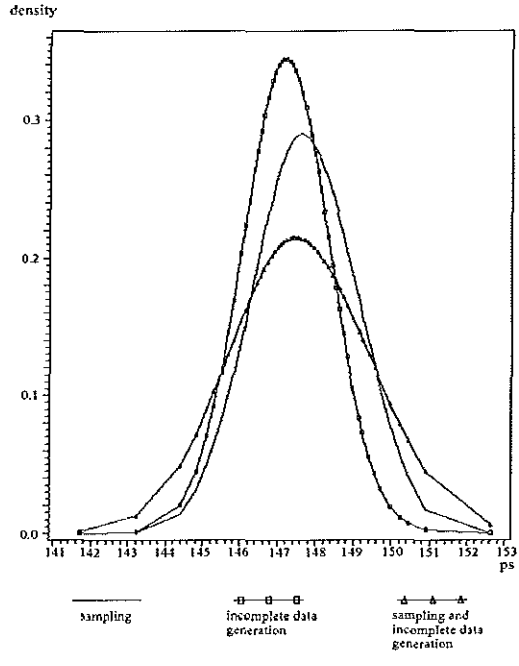


Figure 4-7: Kernel estimates of probability density distributions of $\widehat{Q}$ under repetition of sampling, $\overline{Q}_\infty$ under repetition of generating incomplete data sets from a fixed complete data set, and $\overline{Q}_\infty$ under repetition of sampling and generation of incomplete data sets, represented by solid graphs, graphs with squares, and with triangles, respectively.

4.7, where Q is equal 147.7 and $\widehat{Q}$ from the fixed complete sample is equal to 146.95. While the degree of concentration of $\widehat{Q}$ around Q indicates the precision of $\widehat{Q}$ with regard to Q, the increase in inferential uncertainty due to missing data is indicated by the degree of concentration of $\overline{Q}_\infty$ around $\widehat{Q}$.

The probability distribution of $\overline{Q}_\infty$ under repetition of sampling and generation of incomplete samples reflects the total inferential uncertainty about Q from the incomplete sample. This probability density distribution is approximated by a kernel estimate (graph with triangles) from $\overline{Q}_\infty^{(1)}, \ldots, \overline{Q}_\infty^{(M)}$, where each $\overline{Q}_\infty^{(i)}$ is independently obtained by generating from Y a complete sample $y^{(i)}$ by the sampling mechanism, generating one incomplete sample $y_{obs}^{(i)}$ from

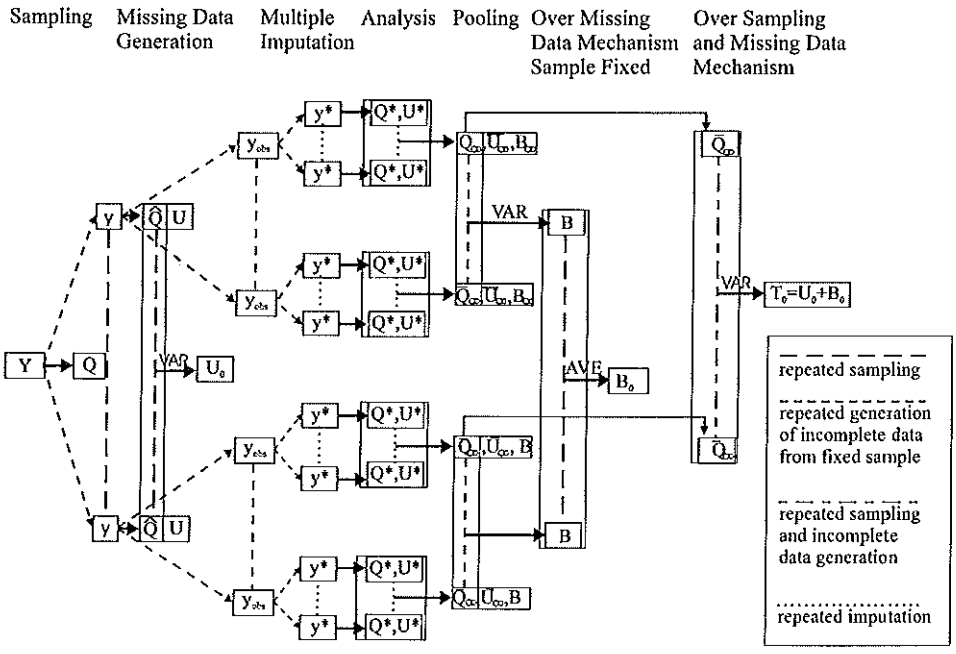


Figure 4-8: An overview of the definitions of the variance-covariance matrices U_0, B, B_0 and T_0.

$y^{(i)}$ by the missing data mechanism and obtaining $\overline{Q}_\infty^{(i)}$ from $y_{obs}^{(i)}$ by multiple imputation. The M different incomplete samples, here, are generated from different complete samples rather than from a fixed complete sample. The probability density distributions in Figure 4.7 indicates that an incomplete sample contains less information about Q than the corresponding complete sample; the probability density distribution of $\overline{Q}_\infty$ under repetition of sampling and generation of incomplete data (graph with squares) has a lower degree of concentration around Q than the probability density distribution of $\hat{Q}$ under repetition of sampling.

An overview of the theoretical variances corresponding to the three different levels of inferential uncertainty mentioned above, is given in Figure 4.8. This scheme is an extension of the scheme for Monte Carlo evaluation of complete data inference in Figure 4.6 by incorporating the missing data mechanism. In the left of Figure 4.8, the approximation of the theoretical

complete data variance U_0 is depicted. From a population Y with parameter of interest Q, a large number of samples y is independently generated by a certain sampling mechanism. By calculating the sample variance over the different point estimates $\hat{Q}$ corresponding to these samples, U_0 is approximated. When the number of generated samples goes to infinity, this sample variance goes to U_0. The theoretical between imputation variance B is approximated in the same way as depicted in the middle of Figure 4.8. From a fixed complete sample y, a large number of incomplete samples y_{obs} are independently generated by a certain missing data mechanism. To each incomplete sample y_{obs}, infinite multiple imputation is applied resulting in the triple $(\overline{Q}_\infty, \overline{U}_\infty, B_\infty)$. How this triple is obtained from an incomplete sample y_{obs} is also depicted in Figure 4.8. The completed sample and the completed sample results, here, are represented by y^* and Q^*, U^*, respectively. Similar to U_0, the variance B is approximated by calculating the sample variance over the multiple imputation estimates $\overline{Q}_\infty$ corresponding to the generated incomplete samples.

The between imputation variance B_∞ is an estimator of B. The variance B depends on the complete sample y so that the expectation B_0 of B under repetition of sampling is also defined. This variance B_0 is approximated by first approximating B for a large number of independently generated complete samples y and calculating the average over the resulting variances B. The theoretical total variance T_0 is the variance of $\overline{Q}_\infty$ under repetition of sampling and missing data generation. This variance is also approximated by calculating the variance over a large number of multiple imputation estimates $\overline{Q}_\infty$. The incomplete samples are generated by first independently generating a large number of complete samples and subsequently generating from each of these complete samples only one incomplete sample. The main difference in the approximation of B and the approximation of T_0, is that for B the incomplete samples are generated from a fixed complete sample and for T_0 the incomplete samples are generated from different complete samples. In [1] it is proved that T_0 equals $U_0 + B_0$. The theoretical variances U_0, B and T_0 corresponding to the three probability density distribution in Figure 4.8 are given by $U_0 = 1.46$, $B = 1.02$ and $T_0 = 2.65$. The variances U_0 , B and T_0 are close to their estimates $\overline{U}_m = 1.42$, $B_m = 1.00$ and $T_m = 2.52$ for $m = 10$ in example 3. The theoretical variance T_0 differs slightly from $T_m = 2.43$ in example 3.

An overview of the validity criteria for complete data inference, multiple imputation, and

incomplete data inference is given in Figure 4.9. The different levels of inferential uncertainty about Q are depicted in the top of this Figure. From the population Y, there is complete certainty about Q in the sense that Q can be fully determined from Y. When only a smaller complete sample y from Y is available, there is loss of information about Q, since Q cannot be fully determined from y. Therefore, Q is estimated by $\widehat{Q}$ and the precision of $\widehat{Q}$ with respect to Q is reflected by U. When only an incomplete sample y_{obs} of y is observed, a further loss of information about Q results, since the complete data statistics $(\widehat{Q}, U)$, to be derived from y, cannot be derived from y_{obs}. Estimates of $(\widehat{Q}, U)$ from the incomplete data y_{obs} by infinite multiple imputation are given by $(\overline{Q}_\infty, \overline{U}_\infty)$ and the precision of $\overline{Q}_\infty$ with respect to $\widehat{Q}$ is reflected by B_∞. By the lower order variability of U [1], the precision of $\overline{U}_\infty$ needs not be reflected. Similar to complete data statistics $(\widehat{Q}, U)$, incomplete data statistics are represented by $(\overline{Q}_\infty, T_\infty)$, where $T_\infty = U_\infty + B_\infty$ reflects the precision of $\overline{Q}_\infty$ with respect to Q. A third component of information loss about Q results from finite multiple imputation, since from finite multiple imputation $(\overline{Q}_\infty, \overline{U}_\infty, B_\infty)$ cannot be determined. Incomplete data statistics by finite multiple imputation are represented by $(\overline{Q}_m, T_m)$.

The similarity between the different validity criteria is illustrated in the bottom of Figure 4.9. The validity conditions for complete data inference are given by Eq. 4.47 and Eq. 4.48 and displayed in the left of the Figure. The conditions for proper imputation are given by

$$\overline{Q}_\infty \sim N\left(\widehat{Q}, B\right), \tag{4.57}$$

$$\overline{U}_\infty \approx U, \tag{4.58}$$

$$B_\infty \approx B, \tag{4.59}$$

$$B \approx B_0. \tag{4.60}$$

For valid incomplete data inference the conditions are given by

$$\overline{Q}_\infty \sim N\left(Q, T_0\right), \tag{4.61}$$

$$T_\infty \approx T_0. \tag{4.62}$$

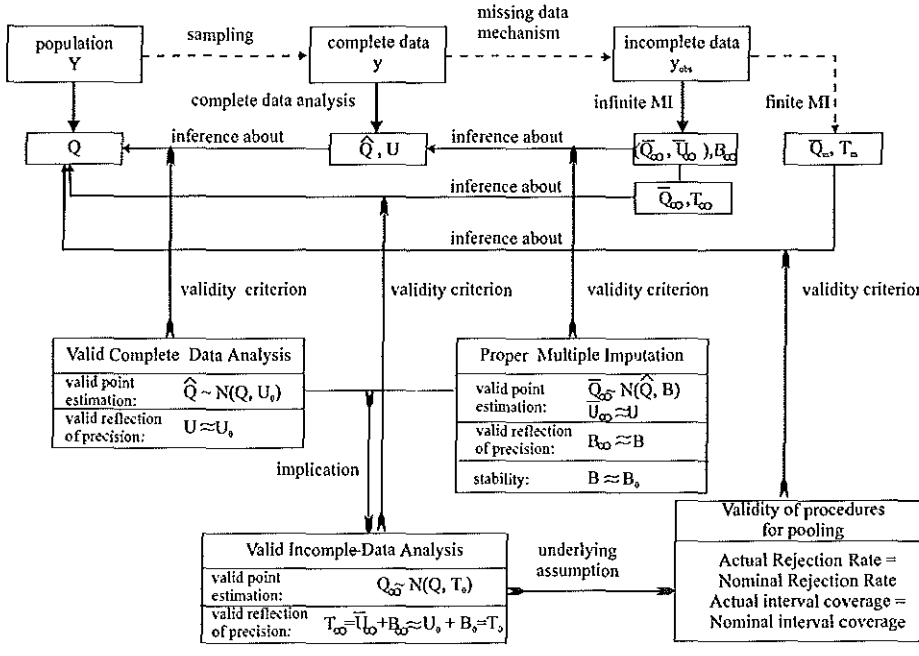


Figure 4-9: An overview of the validity criteria for complete data inference, incomplete data inference and multiple imputation.

In Eq. 4.62, $T_\infty = U_\infty + B_\infty$ and $T_0 = U_0 + B_0$.

The conditions for proper imputation are similar to those for valid complete data inference, in the sense that both validity criteria require a valid point estimation and a valid reflection of the precision of point estimators. The condition in Eq. 4.47 for valid complete data inference requires an unbiased estimation of Q by $\hat{Q}$ and the conditions in Eq. 4.57 and Eq. 4.58 can be interpreted as a valid estimation of the tuple $(\hat{Q}, U)$ by the tuple $(\overline{Q}_\infty, \overline{U}_\infty)$. Similar to Eq. 4.47, the condition in Eq. 4.57 requires that $\overline{Q}_\infty$ is an unbiased estimate of $\hat{Q}$ and that the sample size is large enough to approximate the distribution of $\overline{Q}_\infty$ under repeated missing data generation by a normal distribution. The condition in Eq. 4.58 is stronger than the other two and requires that $\overline{U}_\infty$ is approximately equal to U in the sense of lower order variability [3]. The conditions in Eq. 4.48 and Eq. 4.49 require that the precision of $\hat{Q}$ with regard to Q

and the precision of $\overline{Q}_\infty$ with regard to $\widehat{Q}$ is correctly reflected by U and B_∞, respectively. An additional condition in proper imputation is Eq. 4.60, which means that the variance B must be stable under repeated sampling; this may be regarded as a minor technical condition [7].

When multiple imputation is proper and the complete data statistics $(\widehat{Q}, U)$ obtained from completed samples are valid, the resulting incomplete data statistics $(\overline{Q}_\infty, T_\infty)$ are also valid according to the conditions in Eq. 4.61 and Eq. 4.62 [1]. Pooling procedures applied to confidence intervals and p-values are approximate procedures and thus require a separate validation under the assumption of valid incomplete data inference. The validity conditions for these procedures involve the validity of confidence intervals or the validity of p-values: for valid confidence intervals the actual coverage must be equal to the nominal coverage, for valid p-values the actual rejection rate must be equal to the nominal rejection rate. Studies in which the validity of these pooling procedures is investigated can be found in [1,3,5,6].

For the Monte Carlo evaluation of proper imputation, simplified validity conditions similar to the conditions in [7] are given by

$$E[\overline{Q}_m] = \widehat{Q} \tag{4.63}$$

$$E[\overline{U}_m] = U, \tag{4.64}$$

$$Var(\overline{Q}_m) = (1 + m^{-1})E[B_m]. \tag{4.65}$$

In the equations 4.63 through 4.65, $E[.]$ and $Var[.]$ are the expectation and variance under the assumed missing data mechanism. The equations 4.63 and 4.64 require that the multiple imputation estimates $(\overline{Q}_m, \overline{U}_m)$ are unbiased estimates of the statistics $(\widehat{Q}, U)$ of the hypothetical complete data set. The condition in Eq. 4.65 is similar to the property $Var[\overline{Q}_m] = (1 + m^{-1})B$ and thus requires that the extra inferential uncertainty about Q due to missing data is correctly reflected.

4.4.3 Inspection of generated imputations

Once imputations have been generated, their quality can be inspected by answering the following two questions:

1. Does the imputed data fit to the observed data?

2. Can the quality of the fit be explained by the underlying missing data mechanism?

A general answer of question 1 can be obtained by inspection of scatterplots of all variables involved in the imputation model in which the observed and imputed data points are marked by different colours. Question 1 can be answered further by comparing for each imputation variable y for several univariate statistics $\widehat{Q}$ the following:

1. The result $\widehat{Q}(y_{obs})$ obtained from the observed data y_{obs}, with the average of the results $\widehat{Q}\left(y_{mis}^{*(1)}\right), \ldots, \widehat{Q}\left(y_{mis}^{*(m)}\right)$ obtained from the m vectors $y_{mis}^{*(1)}, \ldots, y_{mis}^{*(m)}$ of imputed values for y;

2. The result $\widehat{Q}(y_{obs})$ obtained from the observed data y_{obs}, with the average of the results $\widehat{Q}\left(y^{*(1)}\right), \ldots, \widehat{Q}\left(y^{*(m)}\right)$ obtained from the m vectors $y^{*(1)}, \ldots, y^{*(m)}$ of completed data for y;

3. The fraction γ of information missing due to missing for $\widehat{Q}$ with the fraction of missing data entries in y. When γ is larger than this fraction it can be concluded that multiple imputation is improper;

To examine whether the relations between imputation variables and predictor variables are preserved, for each pair (y,x), with y an imputation variable and x a predictor variable, it can be useful to compare the result obtained after listwise deletion and the multiple imputation result for one or more measures for the association between y and x.

When the answer to question 2 is no, then there is strong evidence that multiple imputation is improper. Question 2 can be assessed for an imputation variable y, by verifying whether the estimated distribution $\widehat{P}_{\text{imp}(y)}$ of imputed values for y is approximately the same as the estimated distribution $P_{\text{mis}(y)}^{\text{MAR}}$ of the unobserved value of y under the MAR assumption. When y has two or more predictor variables, estimation of $P_{\text{mis}(y)}^{\text{MAR}}$ may be difficult. A less rigorous but easier way to examine question 2, is to compare for each predictor variable x of y, the distribution $\widehat{P}_{\text{imp}(y)}$ with the distribution of unobserved values of y under the restricted MAR(x) assumption that the nonresponse of y depends on the observed values of x only. For categorical x, this MAR(x) assumption is equivalent to the assumption that the missing data mechanism is stratified MCAR within each category of x when x is observed, and that the missing data

mechanism is MCAR when x is missing. Consequently, if missing is considered as a separate category, the distributions of the observed and of the unobserved values of y are approximately the same within each category of x. Let $P_{\text{mis}(y)}^{\text{MAR}(x)}$ be the distribution of the unobserved values of y under the MAR(x). In Appendix 4.C, it is shown that for a categorical predictor variable x, the distribution $P_{\text{mis}(y)}^{\text{MAR}(x)}$ can be estimated by:

$$\widehat{P}_{\text{mis}(y)}^{\text{MAR}(x)} = \widehat{w}_0 \widehat{P}_{\text{obs}(y)|x=.} + \sum_{j=1}^{c} \widehat{w}_j P_{\text{obs}(y)|x=j}, \tag{4.66}$$

where c is the number of categories of y, $\widehat{P}_{\text{obs}(y)x=.}$ is the estimated distribution of the observed values of y given that x is missing, $P_{\text{obs}(y)|x=j}$ is the conditional distribution of y given that $x = j$, as estimated from the cases with y and x simultaneously observed, $\widehat{w}_0$ is the fraction of cases with x missing among all cases with y missing, and $\widehat{w}_j$ is the fraction of cases with $x = j$ and observed among all cases with y missing. The right hand side of Eq. 4.66 can be interpreted as the weighed average of the distributions $\widehat{P}_{\text{obs}(y)R_x=0}$, $P_{\text{obs}(y)|x=1}, \ldots, P_{\text{obs}(y)|x=c}$ with weights $\widehat{w}_0, \widehat{w}_1, \ldots, \widehat{w}_k$ where $\widehat{w}_0$ and $\widehat{w}_j$ are the fractions of missing data entries of y having the assumed distributions $\widehat{P}_{\text{obs}(y)|x=.}$ and $P_{\text{obs}(y)|x=j}$, respectively.

4.5 Discussion

The variable-by-variable Gibbs sampling approach for generating imputations, as proposed in this chapter, is especially suitable for efficiently using existing relationships between variables, when the number of variables of a sample is large. Relationships between variables are efficiently used by selecting predictor variables for each imputation variable according to the strategy in subsection 4.2.2. This strategy is straightforward and can be automated to a large extent. The selection of a set of candidate predictor variables, from which predictor variables are selected in the steps 2,3 and 4 of this strategy require judgement from the analysis, and hence, are more difficult to automate.

Another advantage of the variable-by-variable Gibbs sampling approach is its flexibility in selecting a statistical model which adequately fits the data. An illustration of this flexibility is the selection strategy for an optimal elementary imputation method for a numerical imputation

variable. This strategy, however, is laborious so that its usefulness depends on future simulation studies regarding the robustness against deviations from the corresponding statistical model of imputation methods. How such studies can be carried out is discussed in more detail in chapter 5.

Contrary to the imputation methods on the basis of a multivariate statistical model, the Gibbs sampling approach is not without theoretical flaws. When a compound imputation method Π contains circularities, i.e., when there exists a pair of imputation variables (z_1, z_2) such that z_1 is a predictor variable for z_2 and in turn z_2 is a predictor variable of z_1, the statistical model corresponding to Π may be redundant in the sense that for this model parameter values $\widetilde{\phi}$ exist for which the corresponding probability density function is not defined. E.g., it can be easily verified that in the circular bivariate model

$$\begin{array}{rcl} y_1 = \beta_{10} + \beta_{11} y_2 + \epsilon_1 & ; & \epsilon_1 \sim N\left(0, \sigma_1^2\right) \\ y_2 = \beta_{20} + \beta_{21} y_1 + \epsilon_2 & ; & \epsilon_2 \sim N\left(0, \sigma_2^2\right) \end{array}, \tag{4.67}$$

where y_1 and y_2 are two numerical imputation variables, the following relationship holds:

$$\beta_{11}\beta_{21} = \rho^2, \tag{4.68}$$

where ρ is the Pearson product-moment correlation coefficient between y_1 and y_2. A consequence of this redundancy is that during the Gibbs sampling iterations, values $\phi^{(t)}$ of the statistical model corresponding to Π may be generated, for which the predictive distribution $P\left(Y_{mis} \mid Y_{obs}; \phi^{(t)}\right)$ cannot be defined, so that convergence to the desired posterior predictive distribution $P\left(Y_{mis} \mid Y_{obs}\right)$ is not guaranteed.

This problem of convergence merely concerns the posterior draws of the regression parameters $\phi_j^{(t)}$ to reflect the uncertainty about these parameters given the currently completed data, and is not regarded as a serious problem. In fact, when the true parameters ϕ of the statistical model corresponding to Π are known and the posterior draws of the regression parameters $\phi_j^{(t)}$ are replaced by their true values ϕ_j, the predictive distribution of $Y_{mis}^{(t)}$ will converge to the predictive distribution $P\left(Y_{mis} \mid Y_{obs}; \phi\right)$. In chapter 5, the effect of circularities on the properness of imputation methods will be investigated for a few representative cases.

Appendix 4.A An adjusted correlation coefficient for a continuous and a categorical variable

A well known property of two bivariate normally distributed variables Y and X is the relationship

$$Var[Y|X = x] = Var[Y](1 - \rho^2), \tag{A.4.1}$$

where ρ is the Pearson product-moment correlation coefficient. Eq.A.4.1 can be rewritten as

$$\rho^2 = 1 - \frac{E[Var[Y \mid X]]}{Var[Y]} \tag{A.4.2}$$

Let Y be a continuous variable, X a categorical variable and $1, \ldots, s$ the categories of X, then

$$E[Var[Y|X]] = \sum_{i=1}^{s} Var[Y|X = i]P(X = i). \tag{A.4.3}$$

By substituting Eq. A.4.3 into Eq. A.4.2 an adjusted correlation coefficient $\lambda(.,.)$ can be constructed as follows:

$$\lambda(Y, X) = \sqrt{1 - \frac{\sum_{i=1}^{s} Var[Y|X = i]P(X = i)}{Var[Y]}}. \tag{A.4.4}$$

Appendix 4.B Imputation methods

The imputation methods π presented in subsection 4.2.2 are described below in more detail for categorical and numerical imputation variables, respectively.

Categorical imputation variable y

Logistic regression imputation

Let y be a binary imputation variable and $x_1, \ldots, x_p$ the set of numerical predictor variables resulting from replacing any categorical predictor variable x of y by its corresponding dummy variables. The underlying statistical model of logistic regression imputation is

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p; \text{ with } \pi = P\left(y = 1 \mid x_1, \ldots, x_p\right), \tag{B.4.1}$$

where π is the conditional probability that $y = 1$ given the observed values of the predictor variables $x_1, \ldots, x_p$ of y. With logistic regression imputation, an imputation Y^*_{mis} is drawn according to the following scheme:

$$
\begin{aligned}
&\text{(i)} \quad \text{draw } \beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*) \text{ from } N\left(\widehat{\beta}, V\left(\widehat{\beta}\right)\right) \\
&\text{(ii)} \quad \text{let } \pi_i^* = 1 \big/ \left(1 + \exp\left(-\left(\beta_0^* + \beta_1^* X_{i1} + \cdots + \beta_p^* X_{ip}\right)\right)\right), \text{ for } i = 1, \ldots, n_{mis} \\
&\text{(iii)} \quad \text{let } y_i^* = 1 \text{ and } y_i^* = 0 \text{ with probabilities} \\
&\qquad \pi_i^* \text{ and } 1 - \pi_i^*, \text{ respectively, for } i = 1, \ldots, n_{mis}
\end{aligned}
\tag{B.4.2}
$$

In step (i) of Eq. B.4.2, the vector of regression coefficients β^* is drawn from the approximate posterior distribution $N\left(\widehat{\beta}, V\left(\widehat{\beta}\right)\right)$ [1], where $\widehat{\beta}$ is an estimate of β obtained from Y_{obs} and X_{obs}, and $V(\widehat{\beta})$ is the estimated covariance matrix of $\widehat{\beta}$. The estimates $\widehat{\beta}$ and $V(\widehat{\beta})$ can be obtained by the iterative weighted least squares algorithm described in [29]. In step (ii), for each missing data entry y_i, the corresponding probability π_i^* resulting from the statistical model in Eq. B.4.1 is calculated from the coefficients β^* and the i-th row $(X_{i1}, \ldots, X_{ip})$ of X_{mis}. In step (iii), for each missing data entry y_i, an imputation y_i^* is generated, such that $y_i^* = 1$ and $y_i^* = 0$ with probabilities π_i^* and $1 - \pi_i^*$, respectively.

Polytomous regression imputation

Let y be a polytomous imputation variable with categories $0, \ldots, s-1$ and $x_1, \ldots, x_p$ be the set of predictor variables of y resulting from replacing any categorical predictor variable of y by its corresponding dummy variables. Polytomous regression can be modelled as a series of separate logistic regression models of the categories $1, \ldots s-1$ against a baseline category 0 according to

$$
\ln\left(\frac{P(y = j \mid x)}{P(y = 0 \mid x)}\right) = \beta_{j0} + \beta_{j1} x_1 + \cdots + \beta_{jp} x_p, \text{ for } j = 1, \ldots s-1 \tag{B.4.3}
$$

With polytomous regression imputation, an imputation Y^*_{mis} is generated according the following imputation scheme:

(i) draw β^* from $N\left(\widehat{\beta}, V\left(\widehat{\beta}\right)\right)$

(ii) let $\pi_{ij}^* = \frac{\exp\left(-\left(\beta_{j0}^* + \beta_{j1}^* X_{i1} + \cdots + \beta_{jp}^* X_{ip}\right)\right)}{1 + \sum_{\nu=1}^{s-1} \exp -\left(\beta_{\nu 0}^* + \beta_{\nu 1}^* X_{i1} + \cdots + \beta_{\nu p}^* X_{ip}\right)}$ (B.4.4)

for $i = 1, \ldots, n_{mis}; j = 0, \ldots s-1$, with $\beta_0^T = \left(\beta_{00}, \ldots, \beta_{0p}\right) = 0$

(iii) let $y_i^* = j$, with probabilities π_{ij}^* for $i = 1, \ldots, n_{mis}; j = 0, \ldots s-1$

In step (i) of Eq. B.4.4, the regression parameters are represented by the vector $\phi = \beta = [\beta_1 | \ldots, | \beta_{s-1}]$ with $\beta_j^T = \left(\beta_{j0}, \ldots, \beta_{jp}\right)$. An estimator $\widehat{\beta}$ of β can be obtained by estimating each β_j by logistic regression restricted to the cases with $y = 0$ or $y = j$. The covariance matrix $V(\widehat{\beta})$ of this estimator $\widehat{\beta}$ is given in [19]. In step (ii), $\pi_{i,j}^*$ is the probability that the i-th missing data entry is equal to the j-th category of y corresponding to the drawn regression coefficients β_j^*. In step (iii), for each missing data entry y_i an imputation y_i^* is generated such that $y_i^* = j$ with probability π_{ij}^*.

Discriminant imputation

Let y be a polytomous imputation variable with categories $0, \ldots, s-1$ and $x_1, \ldots, x_p$ the set of predictor variables resulting from replacing any categorical predictor variable of y by its corresponding dummy variables. Let n_j be the number of values of Y_{obs} in category j, $f(. \mid \mu, \Sigma)$ the probability density function of the multivariate normal distribution with mean vector μ and variance Σ, respectively. Under the assumption that the conditional probability distribution of $x = (x_1, \ldots, x_p)$ given $y = j$ is a multivariate normal distribution with mean vector μ_j and covariance matrix Σ_j the underlying statistical model of discriminant imputation is given by

$$P(y = j \mid x) = \frac{f\left(x \mid \mu_j; \Sigma_j\right) \pi_j}{\sum_{v=0}^{s-1} f\left(x \mid \mu_v; \Sigma_v\right) \pi_v} \tag{B.4.5}$$

The model in Eq. B.4.5 follows directly from substitution of $P(x \mid y = v) = f\left(x \mid \mu_v; \Sigma_v\right)$ and $P(y = v) = \pi_v$ into the formula of Bayes. With discriminant imputation, an impu-

tation Y^*_{mis} is generated according to the following scheme:

(i) let $\alpha_j = 1/2 + n_j$, for $j = 0, \ldots, s-1$

(ii) draw $\theta^*_0, \ldots, \theta^*_{s-1}$ from the standard gamma distribution with parameters given by $\alpha_0, \ldots, \alpha_{s-1}$

(iii) let $\pi^*_j = \theta^*_j \Big/ \left(\sum_{\nu=0}^{s-1} \theta^*_\nu \right)$ for $j = 0, \ldots, s-1$

(iv) draw Σ^*_j from an Inverted Wishart distribution $\left(n_j - 1, (n_j S_j)^{-1}\right)$ for $j = 0, \ldots, s-1$

(v) draw μ^*_j from $N\left(\widehat{\mu}_j, \Sigma^*_j\right)$ for $j = 0, \ldots, s-1$

(vi) let $p^*_{ij} = \frac{f\left(X_{i1}, \ldots, X_{ip} | \mu^*_j, \Sigma^*_j\right)\pi^*_j}{\sum_{\nu=0}^{s-1} f(X_{i1}, \ldots, X_{ip} | \mu^*_\nu, \Sigma^*_\nu)\pi^*_\nu}$

for $i = 1, \ldots, n_{mis}$ and $j = 0, \ldots, s-1$

(vii) let $y^*_i = j$, with probabilities p^*_{ij} for $i = 1, \ldots, n_{mis}$ for $j = 0, \ldots, s-1$

(B.4.6)

The parameters θ^*_j (steps (i) through (iii)) and the parameters μ^*_j and Σ^*_j (steps (v) and (iv)), are drawn from posterior distributions with non-informative priors for these parameters, as described in the chapters 5 and 7 of [9], respectively. In the steps (vi) and (vii), the posterior draws in the steps (i) through (v) are used to generate the imputations according to the statistical model.

Numerical imputation variable y

Linear regression imputation: The underlying statistical model of linear regression imputation is given by

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p + \varepsilon \; ; \; \varepsilon \sim N\left(0, \sigma^2\right) \tag{B.4.7}$$

Let $\widehat{\beta}$ and $\widehat{\sigma}^2$ be the least squares estimators of β and σ^2 from Y_{obs} and X_{obs}, V given by $([1|X_{obs}][1|X_{obs}]^T)^{-1}$ and $V^{1/2}$ the Choleski decomposition [30] of V. The notation $[1|X_{obs}]$ indicates that X_{obs} is concatenated with a column of ones to account for the intercept. The imputation scheme of regression imputation is similar to the method in

Example 5.1 of [1] and is given by:

(i) draw a $\chi^2_{n_{obs}-p}$ random variable g

(ii) let $\sigma^{*2} = \widehat{\sigma}^2(n_{obs} - p)/g$

(iii) draw p independent random variables $Z_1, \ldots, Z_p$ from $N(0,1)$ and let $Z = (Z_1, \ldots, Z_p)$ (B.4.8)

(iv) let $\beta^* = \widehat{\beta} + \sigma^*[V]^{1/2}Z$

(v) draw n_{mis} independent random variables $z_1, \ldots, z_{n_{mis}}$ from $N(0,1)$

(vi) let $y_i^* = \beta_0^* + \beta_1^* X_{i1} + \cdots + \beta_p^* X_{ip} + z_i\sigma^*$ for $i = 1, \ldots, n_{mis}$

In Eq. 4.5 of example 2, the steps (a) correspond to the steps (i) through (iv) in Eq. B.4.8, and the steps (b) correspond to the steps (v) and (vi). The factor $(n_{obs} - p)/g$ in step (ii) represents the sampling variability of $\widehat{\sigma}$ and the term σ^* in step (iv) represents the sampling variability of $\widehat{\beta}$. For second order regression or regression with transformed variables, the regression coefficients β and the residual variance σ are estimated by the Ordinarily Least Squares (OLS) method, similar to first order regression with untransformed variables. If the dependent variable is untransformed, the imputation scheme in Eq. B.4.8 can be applied by adjusting the matrix X to the second order terms or to the transformed variables. If the dependent variable y is also transformed by a certain transformation f, then imputations z^* for the transformed variable z are generated first. The generated imputations z^* are then transformed back into imputations $y^* = f^{-1}(z^*)$ for y if $f(.)$ is invertible; for the Box-Cox and Power transformations this is always the case. With the hot-deck error-term variant, the n_{mis} draws $\{z_i\}$ in step (v) of Eq. B.4.8 are not drawn from standard normal distributions, but randomly drawn without replacement from the observed residuals $e_1, \ldots, e_{n_{obs}}$, standardized to unit variance and given by

$$e_i = (\widehat{y}_i - y_i)(1 - p/n_{obs})^{-1/2}/\widehat{\sigma}_{obs} \quad ; \quad \text{for } i = 1, \ldots, n_{obs} \tag{B.4.9}$$

In Eq. B.4.5, $\widehat{\sigma}_{obs}$ is an estimator of σ from Y_{obs} and X_{obs}. Similar to regression imputation with a normally distributed error-term, the hot-deck error-term variant can also be applied to second order regression, or regression with transformed variables. When the round off

option is applied, the generated imputation y_i^* is replaced by $\tilde{y}_i$ which is the observed value of y closest to y_i^*.

Nearest Neighbour imputation: With nearest neighbour imputation, an imputation y_i^* for a missing data entry y_i is generated as follows:

(i) select from X_{obs} the q rows $X_{i_1}^T, \ldots, X_{i_q}^T$ which are closest to X_i^T as measured by a certain distance function d.

(ii) draw $q-1$ uniform(0,1) random numbers and let $a_1, \ldots, a_{q-1}$ be their ordered values. Also let $a_0 = 0$ and $a_q = 1$

(iii) let $p_j = a_j - a_{j-1}$ for $j = 1, \ldots, q$

(iv) draw y_i^* from $y_{i_1}, \ldots, y_{i_q}$ with probabilities $p_1, \ldots, p_q$

(B.4.10)

The number q in Eq. B.4.10 is chosen in such a way, that only a small donor class fraction $f_o = q/n_{obs}$ of Y_{obs} is selected; a reasonable value for this fraction may be $f_o = 0.1$. When the predictive distribution $P\left(y_i|X_i^T\right)$ is a smooth continuous function of X_i^T, then for a small donor class fraction f_o, the empirical distribution of $\left\{y_{i_1}, \ldots, y_{i_q}\right\}$ provides information about this predictive distribution. In the steps (ii) through (iv), the imputation y_i^* is drawn from this empirical distribution according to the Bayesian bootstrap method[1]. The probabilities p_j in step (iii) are drawn to reflect uncertainty about the predictive distribution $P\left(y_i|X_i^T\right)$, given the observed values $\left\{y_{i_1}, \ldots, y_{i_q}\right\}$.

Appendix 4.C Estimation of the probability distribution of the unobserved values of an imputation variable y under the MAR(x) assumption with respect to a categorical variable x

The probability distribution of interest is $P_{\text{mis}(y)}^{\text{MAR}(x)} = P\left(y \mid R_y = 0\right)$, where R_y is the response indicator of y. Let R_x be the response indicator of x. The MAR(x) assumption is formulated as

1. $P\left(y \mid R_x = 0; R_y = 0\right) = P\left(y \mid R_x = 0; R_y = 1\right)$
2. $P\left(y \mid x = j; R_x = 1; R_y = 0\right) = P\left(y \mid x = j; R_x = 1; R_y = 1\right)$

Factorization of $P(y \mid R_y = 0)$ with respect to R_x yields

$$P_{\text{mis}(y)}^{\text{MAR}(x)} = P(y \mid R_y = 0) = \begin{array}{l} P(R_x = 0 \mid R_y = 0)\, P(y \mid R_x = 0; R_y = 0) + \\ P(R_x = 1 \mid R_y = 0)\, P(y \mid R_x = 1; R_y = 0) \end{array} \tag{C.4.1}$$

Let c be the number of categories of x. Factorization of $P(y \mid R_x = 1; R_y = 0)$ with respect to the different categories of x results in

$$\begin{array}{l} P(R_x = 0 \mid R_y = 0)\, P(y \mid R_x = 1; R_y = 0) = \\ \sum_{j=1}^{c} P(x = j \mid R_x = 1; R_y = 0)\, P(R_x = 1 \mid R_y = 0)\, P(y \mid x = j; R_x = 1; R_y = 0) \end{array} \tag{C.4.2}$$

The term $P(x = j \mid R_x = 1; R_y = 0)\, P(R_x = 1 \mid R_y = 0)$ can be written as

$$P(x = j \mid R_x = 1; R_y = 0)\, P(R_x = 1 \mid R_y = 0) = P(x = j; R_x = 1 \mid R_y = 0) \tag{C.4.3}$$

Substituting condition 2 of the MAR(x) assumption and Eq. C.4.3 into Eq. C.4.2 yields

$$\begin{array}{l} P(y \mid R_x = 1; R_y = 0)\, P(R_x = 1 \mid R_y = 0) = \\ \sum_{j=1}^{c} P(y \mid x = j; R_x = 1; R_y = 1)\, P(x = j; R_x = 1 \mid R_y = 0) \end{array} \tag{C.4.4}$$

Substitution of condition 1 of the MAR(x) assumption into Eq. C.4.1 results in

$$P_{\text{mis}(y)}^{\text{MAR}(x)} = P(y \mid R_y = 0) = \begin{array}{l} P(R_x = 0 \mid R_y = 0)\, P(y \mid R_x = 0; R_y = 1) + \\ \sum_{j=1}^{c} P(y \mid x = j; R_x = 1; R_y = 1) \\ P(x = j; R_x = 1 \mid R_y = 1) \end{array} \tag{C.4.5}$$

Let $P_{\text{obs}(y)|x=.} = P(y \mid R_y = 1; R_x = 0)$, $P_{\text{obs}(y)|x=j} = P(y \mid x = j; R_x = 1; R_y = 1)$, $w_0 = P(R_x = 0 \mid R_x = 0; R_y = 0)$, $w_j = P(x = j; R_x = 1 \mid R_y = 0)$, for $j = 1, \ldots, c$, then Eq. C.4.6 can be rewritten as

$$P_{\text{mis}(y)}^{\text{MAR}(x)} = w_0 P_{\text{obs}(y)|x=.} + \sum_{j=1}^{c} w_j P_{\text{obs}(y)|x=j} \tag{C.4.6}$$

The probability distribution $P_{\text{mis}(y)}^{\text{MAR}(x)}$ can be estimated by

$$\widehat{P}_{\text{mis}(y)}^{\text{MAR}(x)} = \widehat{w}_0 \widehat{P}_{\text{obs}(y)x=.} + \sum_{j=1}^{c} \widehat{w}_j \widehat{P}_{\text{obs}(y)|x=j} \tag{C.4.7}$$

where c is the number of categories of y, $\widehat{P}_{\text{obs}(y)|x=.}$ is the estimated conditional distribution of the observed values of y given that x is missing, $P_{\text{obs}(y)|x=j}$ is the conditional distribution of y given that $x = j$, as estimated from the cases with y and x simultaneously observed, $\widehat{w}_0$ is the fraction of cases with x missing among all cases with y missing, and $\widehat{w}_j$ is the fraction of cases with $x = j$ and observed among all cases with y missing.

Bibliography

[1] Rubin DB, Multiple imputation for nonresponse in surveys. Wiley New York. 1987

[2] Rubin DB, Schenker N, Multiple imputation in health-care databases: an overview and some applications. Statistics in Medicine, Vol. 10, 585-589, 1991

[3] Li KH, Raghunathan TE, Rubin DB, Large-Sample Significance Levels From Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution. Journal of the American Statistical Association. Vol.86, No.416, 1991:1065-1073

[4] Carlin BP, Louis TA, Bayes and Empirical Bayes: methods for data analysis. Chapman & Hall, 1996

[5] Li KH, Raghunathan TE, Rubin DB, Significance Levels from Repeated p-values with Multiply Imputed Data. Statistica Sinica, Vol.1. No.1, 1991:65-92

[6] Meng XL, Rubin DB, Performing likelihood ratio tests with multiply-imputed data sets, Biometrika, Vol 79, No.1, 1992:103-111

[7] Rubin DB, Multiple Imputation After 18+ Years. Journal of the American Statistical Association, Vol.91, No.434, 1996:473-489

[8] Tanner MA, Wong WH, The calculation of the Posterior Distribution by Data Augmentation. Journal of the American Statistical Association, Vol.82, No.398, 1987:528-550

[9] Schafer JL, Analysis of incomplete multivariate data. Chapman & Hall, London, 1997

[10] Kennickell AB, Imputation of the 1989 Consumer Finances: Stochastic Relaxation and Multiple Imputation. American Statistical Association. Proceedings of the Section on Survey Research Methods. 1991:1-10

[11] Geman S, Geman D, Stochastic relaxation, Gibbs distributions and Bayesian restoriation of images, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.6, 1984:721-741

[12] Vardi Y, Shepp LA, Kaufman L, A statistical model for position emission tomography., Journal of the American Statistical Association, Vol.80, 1985:8-37

[13] Rubin DB, Schafer JL, Efficiently creating multiple imputations for incomplete multivariate normal data. American Statistical Association, Proceedings of the Statistical Computing Section, 1990:83-88

[14] Li KH, Imputation Using Markov Chains. Journal of Statistical Computation and Simulation, Vol.30, 1988:57-79

[15] Rubin DB, EM and BEYOND. Psychometrica, Vol.56, No.2, 1991:241-254

[16] Gelman A, Rubin DB, Inference from Iterative Simulation Using Multiple Sequences. Statistical Science, Vol.7, No.4, 1992:457-511

[17] Box GEP, Tiao GC, Bayesian Inference in Statistical Analysis. Reading MA Adisson Wesley, 1973

[18] Hosmer DW, Lemeshow S, Applied Logistic Regression, Wiley and Sons, New York,1989

[19] Begg CB. Gray R, Calculation of polytomous regression parameters using individualized regression. Biometrica, Vol.71, No.1, 1984:11-19

[20] Engel J, Polytomous logistic regression, Statistica Neerlandica 42, nr. 4, 1988

[21] Box GEP, Cox DR, An analysis of transformations, Journal of Royal Statistical Association, B 26, 1964: 221-252

[22] Draper NR, Smith H, Applied regression analysis, Second edition, Wiley & Sons, 1981

[23] Chatfield C, Collins AJ, Introduction to Multivariate Analysis. Chapman and Hall, London, 1989

[24] Siegel S, Castellan NJ, Nonparametric statistics for the behavorial sciences. McCraw-Hill Book Company, second edition. New York, 1988

[25] Gifi A, Nonlinear multivariate analysis. Wiley Chichester, 1990

[26] Meng XL, Multiple-Imputation Inferences with Uncongenial Sources of Input, Statistical Sciences, Vol.9, No.4, 1994:538-573

[27] Makuch RW, Freeman DH, Johnson MF, Justification for the lognormal distribution as a model for blood pressure, Journal Chronical Disease, Vol.32, 1979:245-250

[28] Stuart A, Ord JK, Kendall's advanced theory of statistics. Vol.1 of fifth edition: Distribution theory, Charles Griffin & Co, London 1987

[29] McCullagh P, Nelder JA, Generalized Linear Models. Chapman & Hall, 1990

[30] Press WH, Teukolsky SA, Vetterling WT, Flannery BP, Numerical Recipes in C: The art of Scientific Computing. Second Edition, Cambridge University Press, 1992

Chapter 5

Validation of methods for the generation of imputations

5.1 Introduction

This chapter describes the validation of some of the imputation methods developed in chapter 4. An imputation method is validated by establishing whether the empirical multivariate distribution of the variables involved as obtained from a complete data set is adequately recovered from this data set when artificially made incomplete and subsequently completed by imputation. For multivariate data this is achieved by verifying for each imputation variable y the following:

1. Whether the empirical distribution of y as obtained from the complete data set is adequately recovered from a number of data sets completed by imputation;
2. Whether the relationships between y and its predictor variables x as obtained from the complete data set are adequately recovered from a number of data sets completed by imputation;
3. Whether the extra uncertainties due to missing data in the quantities mentioned above are correctly reflected.

The statements mentioned above can be indirectly verified by establishing whether the imputation method is *proper* according to Rubin's definition [1,2] (see also chapter 4) for several

descriptive statistics of y and for several association measures between y and its predictor variables x.

The validation of imputation methods is restricted to the situation that the underlying assumptions of the methods are true. These assumptions are:

1. The imputation method is based on the statistical model that generated the complete data set.

2. The underlying missing data mechanism is MAR.

A further restriction is that only imputation methods using linear regression imputation with a normally distributed error-term, logistic regression imputation, polytomous regression imputation and discriminant imputation are validated. Validating all imputation methods developed in chapter 4, or investigating the robustness of these methods against deviations from their underlying assumptions is outside the scope of this thesis.

The validation of imputation methods consists of two steps. First, elementary methods that generate imputations for one imputation variable conditionally on completely observed predictor variables are studied. Next, compound methods generating imputations for more than one imputation variable, i.e., for multivariate missing data, are validated.

5.2 Design

The properness of an imputation method [1,2] mainly depends on the hypothetical complete data set, the underlying missing data mechanism and the target statistic derived from the m completed data sets. To avoid effects of model misspecification, the complete data sets in this study are generated according to the statistical model for the imputation method to be validated. That the properness of an imputation method also depends on the target statistic is clear from the following example:

Let x_1 and x_2 be two strongly correlated numerical variables, where x_1 contains missing data and x_2 is completely observed, and let Π be an imputation method generating imputations for x_1 without using x_2 as a predictor variable. The method Π may be proper for the mean of x_1, but is clearly improper for the correlation coefficient between x_1 and x_2. This is so because

a large number of values, unlikely in combination with x_2, will be imputed for x_1, resulting in an estimate of the correlation coefficient which is biased toward zero.

Rather than verifying proper imputation for a few target statistics, a better insight in the quality of an imputation method is obtained by establishing whether the empirical distribution of the variables involved as obtained from the complete data set is adequately recovered from a number of data sets completed by imputation, and whether the extra uncertainties about this distribution due to missing data are correctly reflected. To establish this directly is difficult, but a general impression can be obtained by verifying the properness of an imputation method for an appropriate set of target statistics describing the distribution.

A formal definition of the imputation methods to be considered in this study is given in subsection 5.2.1. The generation of a complete data set according to the statistical model of an imputation method is described in subsection 5.2.2. Subsection 5.2.3 describes a class of MAR missing data mechanisms to be used to generate incomplete data sets and in subsection 5.2.4 the target statistics to be considered are listed. Verification of the properness of an imputation method for a given complete data set, missing data mechanism, and target statistic is described in subsection 5.2.5. A plan of attack is described in subsection 5.2.6.

5.2.1 Imputation methods

The imputation methods developed in chapter 4 are formally represented by $\Pi = (\Pi_1, \ldots, \Pi_k)$, where $\Pi_j = (y_j, \{x_j\}, \pi_j)$ is an elementary imputation method, generating imputations for the j-th imputation variable y_j by the method π_j, conditionally on the completely observed set of predictor variables $\{x_j\}$. An imputation method Π generating imputations for more than one imputation variable is called a compound imputation method. When for some imputation variables y_i, predictor variables $\{x_i\}$ are incompletely observed, imputations are generated by means of the Gibbs sampling algorithm (see chapter 4). In each iteration of this algorithm, imputations are sequentially generated for the imputation variables $y_1, \ldots, y_k$ by the methods $\pi_1, \ldots, \pi_k$, conditionally on the observed data and the imputations generated most recently for $\{x_1\}, \ldots, \{x_k\}$. In the special case that for each variable y_i the predictor variables $\{x_i\}$ are completely observed, imputations can be generated in a single iteration.

In this study, the elementary methods π_i are restricted to Linear Regression imputation with

the Normal error-term variant (LRN), LOgistic Regression imputation (LOR), POlytomous Regression imputation (POR), and DIScriminant imputation (DIS), which are considered as the basic methods. The methods POR and DIS are applied to a nominal imputation variable y with three or more categories.

5.2.2 Complete data sets

For each imputation method Π to be validated, a complete data set is generated according to the corresponding statistical model of Π. Basic material for the generation of such model based data sets, are a raw data set consisting of daily average wind speeds in knots (1 knot = 0.5148 m/s) at 12 synoptic meteorological stations in the Republic of Ireland during the period 1961-1978, and a raw data set from a study which was undertaken to assess factors associated with women's knowledge, attitude, and behaviour toward mammography [3]. The Irish wind speeds data consists of 6574 cases and is analyzed in detail in [4]. The data can be retrieved from the website: http://lib.stat.cmu.edu/datasets/ under wind.data. Descriptive statistics and a correlation matrix of this data set are given in the tables 5.1 and 5.2. The existence of strong correlations makes this data set suitable for the construction of systematic MAR missing data mechanisms, since for MAR mechanisms the probability of a data entry to be missing depends on values of other observed variables. In this validation study, different days at the same measurement station are treated as independent observations. A code sheet for the variables used in the Mammographic Experience data set is given in table 5.3. The number of records in this data set is 412. The variable PB (Perceived Benefit) is treated as a numerical variable.

Complete data generation for an elementary imputation method

Complete data (Y, X) for the validation of an elementary imputation method $\Pi = (y, \{x\}, \pi)$ is generated in the following two steps:

1. Complete data X for the predictor variables $\{x\}$ is selected from one of the two "raw" data sets Z described above. When Z is the Mammographic Experience data set, all columns for all 412 records of Z corresponding to $\{x\}$ are selected for X. In the case of the Irish windspeeds data set, X is a random sample from these columns. To use the raw

variable	full name	mean	standard deviation	minimum	maximum
RPT	Roche's Pt.	15.73	8.80	1.00	31.00
VAL	Valentia	12.36	5.62	0.67	35.80
ROS	Rosslare	10.64	5.27	0.21	33.37
KIL	Kilkenny	11.66	5.01	1.50	33.84
SHA	Shannon	10.46	4.94	0.13	37.54
BIR	Birr	7.09	3.97	0.00	26.16
DUB	Dublin	9.78	4.98	0.00	30.37
CLA	Claremorris	8.49	4.50	0.00	31.08
MUL	Mullingar	8.50	4.17	0.00	25.88
CLO	Clones	8.71	4.50	0.04	28.21
BEL	Belmullet	13.12	5.84	0.13	42.38
MAL	Malin Head	15.60	6.70	0.67	42.54

Table 5.1: Descriptive statistics for the daily average windspeeds.

	RPT	VAL	ROS	KIL	SHA	BIR	DUB	CLA	MUL	CLO	BEL	MAL
RPT	1.00											
VAL	0.84	1.00										
ROS	0.73	0.60	1.00									
KIL	0.87	0.77	0.74	1.00								
SHA	0.83	0.86	0.59	0.86	1.00							
BIR	0.81	0.81	0.65	0.87	0.90	1.00						
DUB	0.74	0.67	0.66	0.81	0.79	0.83	1.00					
CLA	0.76	0.80	0.61	0.82	0.87	0.89	0.79	1.00				
MUL	0.78	0.74	0.62	0.85	0.85	0.90	0.88	0.87	1.00			
CLO	0.75	0.72	0.61	0.84	0.81	0.87	0.84	0.88	0.88	1.00		
BEL	0.64	0.75	0.47	0.70	0.77	0.78	0.70	0.86	0.77	0.81	1.00	
MAL	0.62	0.61	0.48	0.66	0.67	0.71	0.77	0.74	0.77	0.81	0.76	1.00

Table 5.2: Correlation matrix for the daily average windspeeds.

variable	full name	codes
ME	Mammographic Experience	0=never 1=during the past year 2=over one year ago
SYMPD	Don't need a Mammogram unless you develop symptoms	0=strongly agree or agree 1=strongly disagree or disagree
PB	Perceived Benefit of Mammography	5-20
HIST	Mother or Sister with a History of Breast Cancer	0=no 1=yes
BSE	Has anyone taught you how to examine your own breast; that is BSE?	0=no 1=yes
DETC	How likely is it that a mammogram could find a new case of breast cancer?	0=not likely 1=somewhat likely 2=very likely

Table 5.3: Code Sheet for the variables in the Mammographic Experience data set.

Irish windspeeds data set for categorical variables in $\{x\}$, the data for these variables is discretized such that each category has the same number of observations and the ordering of the categories corresponds to the original values of the discretized variable.

2. Complete data Y for y is generated from $P_\pi\left(Y \mid X; \widehat{\phi}\right)$, where, for the methods LRN, LOR or POR, $P_\pi(\cdot \mid \cdot)$ is the conditional probability function according to the standard linear, logistic or polytomous regression model, and $\widehat{\phi}$ is the vector of parameters of the regression of y on $\{x\}$, as estimated from the raw data set.

Complete data generation for a compound imputation method

When a compound imputation method Π does not contain circularities, i.e., when there exists no pair of variables (z_1, z_2) such that z_1 is a predictor variable for z_2 and in turn z_2 is a predictor variable for z_1, the corresponding statistical model of Π can be written as

$$P_\Pi(y_1, \ldots, y_k \mid x_1, \ldots, x_v; \phi) = \prod_{i=1}^{k} P_{\pi_i}(y_i \mid \{x_i\}; \phi_i). \tag{5.1}$$

Eq. 5.1 is an application of a well known property in the theory of probabilities. In this equation, $y_1, \ldots, y_k$ are the imputation variables of Π sorted such that for each y_i only y_j's, with $j < i$, are possible predictor variables; $x_1, \ldots, x_v$ are the completely observed predictor variables

of Π. The parameters of the statistical model corresponding to Π are represented by ϕ and ϕ_i represents the parameters of the statistical model corresponding to the elementary method $\Pi_i = (y_i, \{x_i\}, \pi_i)$. According to Eq. 5.1, complete data generation for Π is straightforward. Starting with complete data X for the complete predictor variables involved in Π, complete data for Π is generated by successively generating complete data $Y_1, \ldots, Y_k$ for $y_1, \ldots, y_k$ according to the statistical models corresponding to the elementary imputation methods $\Pi_1, \ldots, \Pi_k$. In particular, Y_i is generated from $P_{\pi_i}\left(Y_i \mid X_i; \widehat{\phi}_i\right)$, where X_i is the complete data already generated for $\{x_i\}$, and $\widehat{\phi}_i$ are the parameters of the regression model of y_i on $\{x_i\}$ corresponding to π_i, as estimated from the raw data set.

If Π does contain circularities, Eq. 5.1. cannot be applied, so that generating complete data according to the statistical model of Π may be difficult or impossible. One approach is to generate data from a closely related statistical model of $\widetilde{\Pi}$, with $\widetilde{\Pi}$ obtained from Π by removing all circularities. When for Π, each elementary method π_i is equal to LRN and each variable involved is numerical (as is the case in the Irish windspeeds data), complete data for Π is generated from the multivariate normal distribution with mean vector and covariance matrix estimated from the raw data.

5.2.3 Missing data mechanisms

In this validation study, only MAR and MCAR (a special case of MAR) missing data mechanisms are considered. Below, a class of MAR and MCAR missing data mechanisms is introduced, which is considered general enough for the validation study. This class also contains missing data mechanisms where for some variables the probability of occurrence of a missing data entry depends on the observed part of incompletely observed other variables. The basic parameters are the fraction α of incomplete cases, p predefined missing data patterns $t_1^T, \ldots, t_p^T$, with t_i^T indicating the i-th row of a $(p \times q)$ matrix t, with q the number of variables and t_{ij} given by

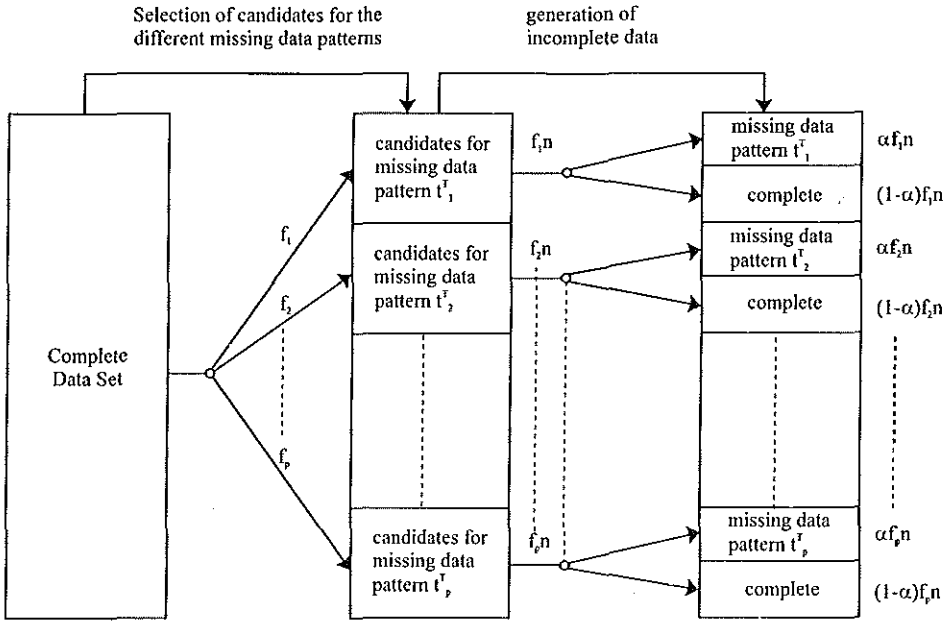


Figure 5-1: Schematic overview of the generation of incomplete data sets.

$$t_{ij} = \begin{cases} 1 & \\ \quad \text{if in the } i\text{-th missing data pattern } t_i^T \text{ the } j\text{-th variable is} & \begin{matrix}\text{observed} \\ \\ \text{missing}\end{matrix} \\ 0 & \end{cases}, \qquad (5.2)$$

and the p-vector f with f_i the fraction of incomplete cases with missing data pattern t_i^T. For example, for the three missing data mechanisms in example 2 of chapter 2, $\alpha = 0.455$, $t = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}$, and $f = \begin{pmatrix} 0.35 \\ 0.25 \\ 0.25 \\ 0.15 \end{pmatrix}$.

In Figure 5.1, a schematic overview of the generation of incomplete data sets is given. A complete data set is made incomplete in two steps. First, each case is nominated for a missing data pattern t_i^T with probability f_i. As shown in the middle of the Figure, the cases are

subdivided into p groups of candidates for the p different missing data patterns, where the expected number of candidates for t_i^T is nf_i, with n the total number of cases in the data set. In the next step, for each missing data pattern t_i^T, a subset of cases with an expected fraction of α from the set of cases nominated for t_i^T is made incomplete according to t_i^T. Consequently, the expected number of incomplete cases with missing data pattern t_i^T generated by the missing data mechanism is $\alpha f_i n$, as shown in the right hand side of Figure 5.1.

A MCAR missing data mechanism (a special case of MAR) is determined by the parameters α, t and f. In the MCAR mechanisms, for each missing data pattern t_i^T each case nominated for t_i^T is made incomplete according to t_i^T with probability α.

For a MAR missing data mechanism which is not MCAR, the probability that a nominated case z is made incomplete according to t_i^T, depends on a linear combination of the observed values of z according to t_i^T in a step-wise manner. A general MAR missing data mechanism may be formally defined as follows:

Let $z = (z_1, \ldots, z_q)$ be the values of a case nominated for t_i^T, $A = \{a_{ij}\}$ a $(p \times q)$ matrix of arbitrarily chosen weights, and $s_i = \sum_j a_{ij} t_{ij} z_j$ a linear combination of the observed values of z according to t_i^T. Let further $c_i(\theta_{ij})$ be the θ_{ij}-th quantile of s_i for $0 = \theta_{i0} < \theta_{i1} < \ldots < \theta_{i,r} < \theta_{i,r+1} = 1$. Finally, let $G = \{g_{ij}\}$ be a $(p \times r)$ matrix of arbitrarily chosen weights. In our missing data mechanisms, the probability that a nominated case z is made incomplete depends on s_i in a step-wise manner via A, $\Theta = \{\theta_{ij}; i = 1, \ldots, p; j = 1; \ldots, q\}$, and G, as follows:

$$\frac{P\left(z \text{ incomplete} \mid c_i(\theta_{ij}) \leq s_i < c_i(\theta_{i,j+1}) \text{ and } z \text{ nominated for } t_i^T\right)}{P\left(z \text{ incomplete} \mid s_i < c_i(\theta_{i,1}) \text{ and } z \text{ nominated for } t_i^T\right)} = g_{ij}. \qquad (5.3)$$

When each entry of G is equal to 1, the missing data mechanism is MCAR, and the more the entries of G differ from 1, the more the missing data mechanism deviates from MCAR. The MAR missing data mechanism in example 2 of chapter 2 has parameters A, Θ, and G given by

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \Theta = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, G = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 3 \end{pmatrix}.$$

According to this mechanism, the expected fraction of incomplete cases among cases with the fourth variable (pd) larger than its median is three times as large as this expected fraction

among cases with *pd* smaller than or equal to its median.

In Appendix 5.A, it is shown that Eq. 5.3 can be rewritten as

$$P\left(z \text{ incomplete} \mid c_i(\theta_{ij}) \leq s_i < c_i(\theta_{i,j+1}) \text{ and } z \text{ nominated for } t_i^T\right) = \lambda_i g_{ij}, \qquad (5.4)$$

where λ_i is given by

$$\lambda_i = P\left(z \text{ incomplete} \mid s_i < c_i(\theta_{i,1}) \text{ and } z \text{ nominated for } t_i^T\right) = \frac{\alpha}{\sum_{j=0}^{r} (\theta_{i,j+1} - \theta_{i,j})\, g_{i,j}}, \qquad (5.5)$$

and $g_{i,0} = 1$. According to the equations 5.4 and 5.5, each case nominated for t_i^T is made incomplete with probability λ_i when $s_i < c_i(\theta_{i,1})$ and with probability $\lambda_i g_{ij}$ when $c_i(\theta_{ij}) \leq s_i < c_i(\theta_{i,j+1})$.

5.2.4 Target statistics

Target statistics are chosen to answer the following two central questions:

1. Are the empirical distributions of the imputation variables y as obtained from the complete data set adequately recovered from a number of data sets completed by imputation?

2. For each imputation variable y, are the relationships between y and its predictor variables x as obtained from the complete data set adequately recovered from a number of data sets completed by imputation?

With respect to the first question, the distribution of a numerical imputation variable y is characterized by the mean and several quantiles (the 25% quantile, median and 75% quantile are used here) and for categorical y by the proportions of its different categories.

Table 5.4 gives the statistics which are of interest for the second question. For numerical y and numerical x it is sufficient to consider the Pearson product-moment correlation coefficient since the method LRN assumes linearity between y and x. When y is numerical and x is categorical, relationships between y and x are determined by the conditional probability density functions $f_{y|x=s}$, with $s = 0, \ldots, S_x - 1$ the categories of x. To limit the number of different

	numerical x	categorical x
numerical y	Pearson product-moment correlation coefficient.	conditional mean $\overline{y}_{x=s}$
categorical y	conditional mean $\overline{x}_{y=s}$	- log-OR - Cramer-C

Table 5.4: The target statistics to be considered for y and x.

statistics, proper multiple imputation is verified for the conditional means $\overline{y}_{x=s}$. For categorical y and numerical x, proper multiple imputation is verified for the conditional mean $\overline{x}_{y=s}$ of x given that $y = s$ with $s = 0, \ldots, S_y - 1$ the categories of y. When y and x are both categorical, with at least one of these variables, say y, binary, the target statistic is the logarithm of the odds-ratio, or log-OR. The log-OR is chosen as a target statistic, since for large samples its sampling distribution is approximately normal. The sample log-OR is the slope of the logistic regression of y on x when x is binary. When the variable x has three or more categories, the log-OR between y and x is calculated for the categories s and 0, with $s = 1, \ldots, S_x - 1$ and 0 the reference category. When both y and x have three or more categories, the target statistic is the Cramer-C measure [5].

5.2.5 Verification of proper multiple imputation

The simplified version of properness given by the equations 4.63 through 4.65 in chapter 4, is used as a starting point. Figure 5.2 depicts how proper multiple imputation is verified for a given complete data set, missing data mechanism and target statistic. From the initial complete data set, $(\widehat{Q}, U)$ is computed by the complete data analysis of interest (see in the left of the Figure 5.2). Here $\widehat{Q}$ plays the role of the target statistic and U is the variance of $\widehat{Q}$. Subsequently, $N = 500$ incomplete data sets are independently generated by the missing data mechanism. Multiple imputation is applied to each incomplete data set, resulting in the sequence $\left(\overline{Q}_m^{(1)}, \overline{U}_m^{(1)}, B_m^{(1)}\right), \ldots, \left(\overline{Q}_m^{(N)}, \overline{U}_m^{(N)}, B_m^{(N)}\right)$, where m is the number of imputations for each incomplete data set and $\left(\overline{Q}_m^{(i)}, \overline{U}_m^{(i)}, B_m^{(i)}\right)$ are the pooled analysis results for the i-th series of m completed data sets, as illustrated in the middle of Figure 5.2. Estimates $\widehat{E}[\overline{Q}_m]$, $\widehat{E}[\overline{U}_m]$, and $\widehat{E}[B_m]$ of $E[\overline{Q}_m]$, $E[\overline{U}_m]$ and $E[B_m]$ are obtained by averaging of $\overline{Q}_m^{(i)}$, $\overline{U}_m^{(i)}$, and $B_m^{(i)}$ over all i. The variance $Var[\overline{Q}_m]$ is estimated by the variance $\widehat{Var}[\overline{Q}_m]$ of $\overline{Q}_m^{(i)}$. Finally, properness is established by verifying whether $\widehat{E}[\overline{Q}_m] \approx \widehat{Q}$, $\widehat{E}[\overline{U}_m] \approx U$, and $\widehat{Var}[\overline{Q}_m] \approx \left(1 + m^{-1}\right) \widehat{E}[B_m]$,

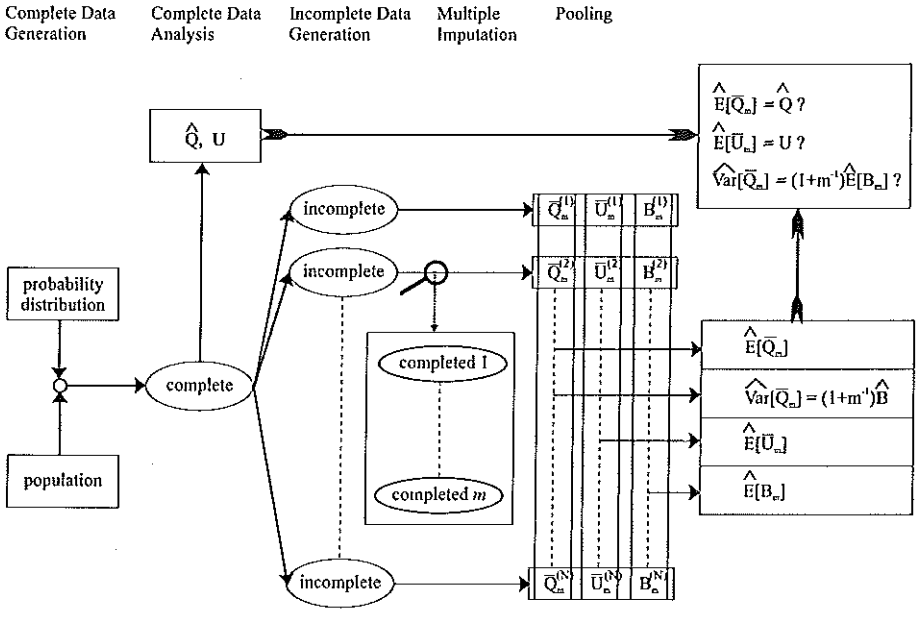


Figure 5-2: A schematic overview of the verification of proper multiple imputation.

where '$\approx$' means approximately equal. Whether two quantities are approximately equal is judged according to an intuitive criterion.

A first indication of properness is obtained by the actual coverage of 95% confidence intervals of $\widehat{Q}$ from the completed data sets given by $\overline{Q}_m \pm \left(\sqrt{(1+m^{-1})\,B_m}\right) t_{m-1;0.975}$, where $t_{m-1;0.975}$ is the 0.975 quantile of the Student t distribution with $m-1$ degrees of freedom. This interval is based on the fact that for proper multiple imputation, $\overline{Q}_m$ is normally distributed with a mean given by $\widehat{Q}$ and a variance given by $(1+m^{-1})\,B$, where the estimate B_m of B has the same distribution as $\left(\chi^2_{m-1}/(m-1)\right) B$ with χ^2_{m-1} a χ^2 random variable with $m-1$ degrees of freedom [1]. The actual coverage of the interval is estimated by calculating the percentage of confidence intervals which include $\widehat{Q}$ over the N confidence intervals obtained from each series of m completed data sets. This estimate of the actual coverage is considerably less precise than estimates such as $\widehat{E}\left[\overline{Q}_m\right]$ of $E\left[\overline{Q}_m\right]$ and should be accompanied by a confidence interval. When the estimated coverage lies in the interval $95\% \pm \left(\frac{1.96}{\sqrt{500}}\sqrt{0.95 * (1-0.95)}\right) 100\% = 95\% \pm 1.9\%$, it can be concluded that the actual coverage does not significantly differ from 95%. This is

because in case of a true actual coverage of 95% the probability that the estimated actual coverage is included in the interval $95\% \pm 1.9\%$ is approximately equal to 95%.

The actual coverage of the confidence intervals of $\widehat{Q}$ should be interpreted with care. An actual coverage of 95% does not automatically imply properness, since a bias of $\overline{Q}_m$ with respect to $\widehat{Q}$ in combination with an overestimation of the between imputation variance B by B_m may also result in an actual coverage of 95% or more. Thus the actual coverage of the confidence interval of $\widehat{Q}$ can only be properly interpreted when $\overline{Q}_m$ is approximately unbiased with respect to $\widehat{Q}$.

The complete data variance U is the squared standard error of the mean when the target statistic is the mean. For the $\theta - th$ quantile t_θ, this variance is given by $\theta(1-\theta)n^{-1}\widehat{f^{-1}(t_\theta)}$ [6] where $\widehat{f^{-1}(t_\theta)}$ is an estimate of the inverse density $f^{-1}(t_\theta)$ according to the method in [7]. For category proportions of y, the complete data variance U is a function of $\widehat{Q}$ and of the sample size n, and for the Pearson product-moment correlation coefficient this variance depends on n only , so that for these statistics the quantities U and $\overline{U}_m$ are not presented. The sampling distribution of the correlation coefficient r is very skew, especially if the corresponding population correlation coefficient ρ is close to -1 or 1. Multiple imputation estimates $\overline{Q}_m(r)$ and confidence intervals for r are calculated via the Fisher transformation [8] z of r given by

$$z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right). \tag{5.6}$$

For large samples, the sampling distribution of z is approximately normal with variance $1/(n-3)$. Multiple imputation estimates $\overline{Q}_m(r)$ and confidence intervals for r are then obtained by back transforming the corresponding estimates $\overline{Q}_m(z)$ and confidence intervals for z via the inverse Fisher transformation

$$r(z) = \frac{e^{2z}-1}{e^{2z}+1}. \tag{5.7}$$

Since the sampling distribution of the Cramer-C measure is complex and also very skew for population values of this measure close -1 or 1, only the quantities $\widehat{Q}$ and $\overline{Q}_m$ are presented for the Cramer-C measure.

method	imputation variable	predictor variables	data set	sample size
LRN	ROS	RPT, SHA, DUB, CLO	Irish windspeeds	400
LOR	VAL	RPT, ROS, SHA, DUB	Irish windspeeds	400
POR and DIS	ME	SYMPD, PB, HIST, BSE, DETC	Mammographic Experience	412

Table 5.5: Imputation variables, predictor variables and sample sizes of the elementary imputation methods LRN, LOR and POR.

5.2.6 Methods

Each compound imputation method consists of two or more elementary imputation methods π_i. Further, each elementary imputation method consists of a series of numerical procedures such as random number generators and matrix inversion, so that validation of the imputation methods is carried out in the following three steps:

1. Validation of numerical procedures in the Gibbs sampling algorithm;
2. Validation of elementary imputation methods for one imputation variable conditional on completely observed covariates;
3. Validation of compound imputation methods with more than one imputation variable and incomplete predictor variables.

The first step is straightforward, the steps 2 and 3 are described below.

Validation of elementary imputation methods

The elementary imputation methods and the data sets to which they are applied are presented in Table 5.5. For the methods LRN and LOR, one complete data set generated from the Irish Windspeeds data set according to the corresponding regression model that is considered. The methods POR and DIS are compared using the Mammographic Experience data set. The raw data set of Irish windspeeds is not suitable for the method POR since the associations between the variables are too strong to generate complete data according to the polytomous regression model.

Missing data mechanisms For each of the combinations of complete data sets and imputation methods described above, four different missing data mechanisms, i.e., one MCAR and three MAR mechanisms are considered. Each of these missing data mechanisms generates missing data in the imputation variable y with an expected fraction of incomplete cases $\alpha = 0.5$. This fraction is chosen relatively large, since small fractions α are not interesting. The three MAR missing data mechanisms are MARRIGHT, MARTAIL and MARMID, with for MARRIGHT, $\Theta = 0.5$ and $G = 4$, and for MARTAIL and MARMID, $\Theta = (0.33, 0.67)$ and G given by $(0.25, 1)$ and $(4, 1)$, respectively. The missing data mechanism MARRIGHT generates a relatively large fraction of incomplete cases among cases with relatively large values of the imputation variable y. For MARTAIL and MARMID, the fraction of incomplete cases is relatively large among cases with values of y in the tails and in the middle of its probability distribution, respectively. Values of G corresponding to a strongly systematic MAR missing data mechanism are chosen here, since it is plausible that if multiple imputation is proper under strongly systematic MAR missing data mechanisms, it will also be proper in less systematic cases. To optimize the effect of the three MAR missing data mechanisms, the $(1 \times q)$ row vector of weights A (q is the number of predictor variables) is chosen equal to the regression coefficients of the imputation variable y on its predictor variables. When y and/or some of its predictor variables are categorical (ordinal), the values of these variables in this regression are replaced by their category numbers. Since the ordering of the categories of the discretized variables x correspond to their original values, the nonresponse indicator R_y of y will also strongly depend on x in case of a strong association between y and x.

Validation of compound imputation methods

The three compound imputation methods to be considered are a method for exclusively numerical imputation variables Π_{num}, a method for imputation variables of mixed type Π_{mix}, and a method for exclusively categorical imputation variables Π_{cat}. For each method, the imputation variables $y_1, \ldots, y_4$ and complete predictor variables x_1, x_2, x_3 with their corresponding data types (num=numerical, bin=binary, tri=trichotomous) are given in Table 5.6. To assess the effect of circularities on the properness of multiple imputation, for each method and each imputation variable y_j, the complete predictor variables and the imputation variables other than

variables		imputation methods					
		$\Pi_{\text{num}}, n = 400$		$\Pi_{\text{mix}}, n = 600$		$\Pi_{\text{cat}}, n = 412$	
		name	type	name	type	name	type
imputation	y_1	RPT	num	RPT	num	ME	tri
	y_2	ROS	num	VAL	bin	SYMPD	bin
	y_3	SHA	num	BIR	tri	BSE	bin
	y_4	DUB	num	DUB	num	DETC	tri
complete predictor	x_1	CLO	num	MUL	num	PB	num
	x_2	MAL	num	CLO	bin	HIST	bin
	x_3			BEL	tri		

Table 5.6: Imputation variables and complete predictor variables for the three compound imputation methods. The types num, bin, and tri refer to numeric, binary and trichotomous variables.

y_j are used as predictor variables. For instance, for the method Π_{mix}, the predictor variables of the imputation variable y_2=VAL are the variables y_1=RPT, y_3=BIR, y_4=DUB, x_1=MUL, x_2=CLO, and x_3=BEL. The elementary imputation methods LRN and LOR are used for numeric and binary imputation variables. Since the simulation results for the elementary methods indicate a superiority of POR over DIS (see subsection 5.3.1), only POR will be considered for the trichotomous imputation variables in the methods Π_{mix} and Π_{cat}.

The associations between the variables in the raw data set of Irish windspeeds are too strong to use this data set for the generation of a complete data set for the method Π_{mix} as described in subsection 5.2.2. This is because after the discretization of y_3, the joint distributions of the other variables for $y_3 = 0$ and for $y_3 = 2$ are discretely different. Thus, a polytomous regression model of y_3 on the other variables is inappropriate, since such a model assumes that the log-probability ratio $\log\left(P\left(y_3 = 0 \mid y_1, y_2, y_4, x_1, x_2, x_3\right) / P\left(y_3 = 2 \mid y_1, y_2, y_4, x_1, x_2, x_3\right)\right)$ is a smooth and continuous function of y_1, y_2, y_4, x_1, x_2, and x_3. To use the raw data set of Irish windspeeds for the generation of a complete data set for Π_{mix}, this data set is adjusted as follows:

First, the imputation variable y_3 is uniformly discretized into three categories. Further, the observations X of the other variables y_1, y_2, y_4, x_1, x_2, and x_3 are replaced by $\widetilde{X}(\eta)$, with the i-th row $\widetilde{x}_i^T(\eta)$ of $\widetilde{X}(\eta)$ given by

$$\widetilde{x}_i^T(\eta) = \begin{cases} x_i^T + \left(\mu_1^T - \mu_0^T - \eta\sigma_0^T\right) & \text{if } y_{3i} = 0 \\ x_i^T & \text{if } y_{3i} = 1 \\ x_i^T - \left(\mu_2^T - \mu_1^T - \eta\sigma_2^T\right) & \text{if } y_{3i} = 2 \end{cases} \tag{5.8}$$

In Eq. 5.8, x_i^T is the i-th row of X, μ_j^T and σ_j^T are two row vectors consisting of the conditional means and standard deviations of the variables $y_1, y_2, y_4, x_1, x_2, x_3$ given that $y_3 = j$, and $\eta > 0$ is a free parameter. According to Eq. 5.8, the two centroids of the rows of $\widetilde{X}(\eta)$ corresponding to $y_3 = 0$ and the rows of $\widetilde{X}(\eta)$ corresponding to $y_3 = 2$, are moving toward the centroid of the rows of $\widetilde{X}(\eta)$ corresponding to $y_3 = 1$, when η decreases toward zero. In the simulation study $\eta = 0.9$ is chosen. Finally, the observations of $\widetilde{X}(\eta)$ for y_2, x_2 and x_3 are uniformly discretized into two, two and three categories, respectively.

For each of the three methods Π_{num}, Π_{mix}, and Π_{cat}, two different complete data sets are considered. For Π_{num}, the first complete data set is generated from the multivariate normal distribution with mean vector and covariance matrix estimated from the raw data set of Irish windspeeds, and for Π_{mix} and Π_{cat} this data set is generated according to the statistical model of the corresponding non-circular methods $\widetilde{\Pi}_{\text{mix}}$ and $\widetilde{\Pi}_{\text{cat}}$ (see subsection 5.2.2) given by

$$\begin{aligned} \widetilde{\Pi}_{\text{mix}} = \; & ((y_1, \{x_1, x_2, x_3\}, \text{LRN}), (y_2, \{y_1, x_1, x_2, x_3\}, \text{LOR}), \\ & (y_3, \{y_1, y_2, x_1, x_2, x_3\}, \text{POR}), (y_4, \{y_1, y_2, y_3, x_1, x_2, x_3\}, \text{LRN})) \\ \widetilde{\Pi}_{\text{cat}} = \; & ((y_1, \{x_1, x_2\}, \text{POR}), (y_2, \{y_1, x_1, x_2\}, \text{LOR}) \\ & (y_3, \{y_1, y_2, x_1, x_2, \}, \text{LOR}), (y_4, \{y_1, y_2, y_3, x_1, x_2\}, \text{POR})) \end{aligned} \tag{5.9}$$

The second complete data set is generated as a sample from the raw data set of Irish Windspeeds for Π_{num}, as a sample from the discretized data set of Irish Windspeeds for Π_{mix} and is the raw Mammographic Experience data set for Π_{cat}. By comparing simulation results for the first and the second complete data set, a first impression is obtained about the robustness of the methods against deviations from assumptions about the underlying statistical model of the complete data set.

For each of the $6 = 3 \times 2$ combinations of methods and complete data sets, one MAR missing data mechanisms similar to MARRIGHT is considered. The parameters of this mechanism are

$\alpha = 0.625$, $t = \begin{pmatrix} 0 & 1 & 0 & 1 & 1\cdots1 \\ 0 & 0 & 1 & 1 & 1\cdots1 \\ 1 & 1 & 0 & 0 & 1\cdots1 \\ 1 & 0 & 1 & 0 & 1\cdots1 \end{pmatrix}$, $f = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$, $\Theta = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}$, $G = \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \end{pmatrix}$, and $A = \{a_{ij}\}$ a matrix with the same dimension as t and $a_{ij} = \begin{cases} 0 & \text{if } t_{ij} = 0 \\ \beta_{ij} & \text{if } t_{ij} = 1 \end{cases}$, where β_{ij} is the regression coefficient for the j-th variable of the regression of y_i on the observed variables according to the i-th missing data pattern t_i^T of t. To assess the convergence of the Gibbs sampling algorithm, the missing data patterns according to the matrix t of this mechanism are chosen to be non-monotonic. Similar to the three MAR missing data mechanisms for the elementary imputation methods, this MAR missing data mechanism also strongly deviates from MCAR and generates a large fraction of incomplete cases (62.5%).

5.3 Results

The complete data sets for the elementary imputation methods LRN and LOR generated from the raw Irish wind speeds data set are called LID and LOD, respectively. The missing data mechanisms MCAR, MARRIGHT, MARTAIL, MARMID generate incomplete data with an expected percentage of missing data entries of 50% in the first column of these data sets. For the compound imputation methods the corresponding complete data sets are called NUMRAWD, NUMMODD, MIXRAWD, MIXMODD, CATRAWD, CATMODD, where 'NUM', 'MIX' and 'CAT" refer to numerical, mixed and categorical data, and 'RAW' and 'MOD' refer to the raw data set and the data set generated according to the underlying statistical model of the imputation method. The MAR missing data mechanism generates incomplete data with an expected fraction of missing data entries of 31.25% in four columns in these data sets. For each of the considered combinations of complete data sets, missing data mechanisms and imputation methods, the number of generated incomplete data sets is 500 and the number of imputations m is equal to 10. For the compound imputation methods, the number of Gibbs sampling iterations is equal to 5. The simulation program was written in SAS/IML. The results are presented in the tables 5.7 through 5.14 in Appendix 5.B.

5.3.1 Elementary imputation methods

The results for the methods LRN, LOR, POR and DIS are presented in the Tables 5.7, 5.8, 5.9 and 5.10, respectively. In each row of these tables, the results for a certain target statistic under a certain missing data mechanism are given. The missing data mechanisms and target statistics are represented by MDM and statistic in the first and second column. The terms Q_1 and Q_3 represent the first (25%) and the third (75%) quartile. In the third, and fifth through ninth column, the results for $\widehat{Q}$, $\widehat{E}[\overline{Q}_m]$, U, $\widehat{E}[\overline{U}_m]$, $\widehat{B} = \widehat{Var}\,[\overline{Q}_m] \,/\, (1 + m^{-1})$, and $\widehat{E}[B_m]$ are given. The average estimate $\widehat{E}[\widehat{Q}_{inc}]$ of $\widehat{Q}$ obtained by complete-case analysis is presented in the fourth column to assess the added value of multiple imputation with respect to complete-case analysis. In the last column, the actual coverages of the 95% confidence intervals of $\widehat{Q}$ are given. For each target statistic under each missing data mechanism, properness is established by verifying whether $\widehat{E}[\overline{Q}_m] \approx \widehat{Q}$, $\widehat{E}[\overline{U}_m] \approx U$, $\widehat{E}[B_m] \approx \widehat{B}$, and whether the actual coverage of the confidence interval for $\widehat{Q}$ lies in the interval $95\% \pm 1.9\%$. The bias of $\overline{Q}_m$ with respect to $\widehat{Q}$ is measured by $\left|\widehat{Q} - \widehat{E}[\overline{Q}_m]\right|$. The added value of multiple imputation with regard to complete-case analysis is assessed by comparing the difference between $\widehat{E}[\widehat{Q}_{inc}]$ and $\widehat{Q}$ with the difference between $\widehat{E}[\overline{Q}_m]$ and $\widehat{Q}$.

Results for LRN

From Table 5.7, it appears that the pooled point estimates $\overline{Q}_m$ are approximately unbiased in most of the cases. For 27 of the 32 target statistics, the difference between $\widehat{Q}$ and $\widehat{E}[\overline{Q}_m]$ is smaller than or equal to 0.05, while for 18 target statistics this difference is smaller than or equal to 0.01. However, the pooled median under MARTAIL and MARMID is biased with differences between $\widehat{Q}$ and $\widehat{E}[\overline{Q}_m]$ of 0.2 and 0.1. It is surprising that the median is considerably more biased than the first and the third quartile under these missing data mechanisms, since estimates of the median are more stable than estimates of the first and third quartile. It is also surprising that under MARRIGHT the pooled median is approximately unbiased while the differences between $\widehat{Q}$ and $\widehat{E}[\widehat{Q}_{inc}]$ under this mechanism are considerably larger than under MARTAIL and MARMID.

For each of the three MAR missing data mechanisms, the differences between $\widehat{Q}$ and $\widehat{E}[\widehat{Q}_{inc}]$ are larger than the corresponding differences between $\widehat{Q}$ and $\widehat{E}[\overline{Q}_m]$. Thus, under MAR, multi-

ple imputation has added value over complete-case analysis. This is most clear from the results for the mechanism MARRIGHT, where the pooled estimates $\overline{Q}_m$ are approximately unbiased and the estimates $\widehat{Q}_{inc}$ are strongly biased with respect to $\widehat{Q}$.

In most of the cases, the pooled estimate $\overline{U}_m$ is approximately unbiased with respect to the complete data variance U. For the first quartile under MARMID and the third quartile under MARRIGHT and MARTAIL, $\overline{U}_m$ clearly overestimates U. This overestimation can be explained by the fact that under MARRIGHT and under MARTAIL the fraction of missing data entries is relatively large for values of y larger than its median and under MARMID this fraction is relatively large for values of y smaller than its median. For MARRIGHT this is clear from its definition. For MARTAIL and MARMID this can be concluded from the fact that the median $\widehat{Q} > \widehat{E}[\widehat{Q}_{inc}]$ under MARTAIL and $\widehat{Q} < \widehat{E}[\widehat{Q}_{inc}]$ under MARMID.

The actual coverage of the confidence intervals for $\widehat{Q}$ lies in the interval 95% ± 1.9% or is significantly larger than 95% in most of the cases. This coverage is underestimated for the correlation coefficients between y and x_1 and between y and x_2 under MARRIGHT with estimated values of 93.0 and 84.2, respectively. Each of the coverages under the missing data mechanism MARMID is strongly overestimated.

Results for LOR

In Table 5.8, prop $(y = j)$ stands for the proportion of $y = j$ and mean$(x_j) \mid y = r$ stands for the conditional mean of x_j given $y = r$. The large differences between $\widehat{Q}$ and $E[\overline{Q}_{inc}]$ for the proportion under MARRIGHT are striking. Under MCAR the conditional means are approximately unbiased and under the three MAR mechanisms they are slightly biased with bias ranging from 0.02 (conditional mean of x_1 given $y = 0$ under MARTAIL) to 0.23 (conditional mean of x_1 given $y = 0$ under MARRIGHT). Under each missing data mechanism, $\overline{U}_m$ is approximately unbiased for each statistic. In 23 of the 40 cases, the actual interval coverage lies in the interval 95% ± 1.9%. Under MARRIGHT, the actual interval coverages are quite low (estimated actual interval coverage <90%) for the proportions and some of the conditional means. For the proportions this is not a serious problem since $\overline{Q}_m$ is approximately unbiased and $\widehat{Q}_{inc}$ is strongly biased.

Results for POR and DIS

The results for POR and DIS are presented in the Tables 5.9 and 5.10. The first thing which attracts the attention is that the coverages of POR are close to 95% and the coverages of DIS are far below 95%. This undercoverage of DIS is due to an underestimation of B by B_m. In many cases, this underestimation is serious. For instance, for the conditional mean of *pb* given $y = 1$ under MARRIGHT, $\widehat{B}$ and $\widehat{E}[B_m]$ are 0.19 and 0.02, respectively.

For both methods POR and DIS, $\overline{Q}_m$ is approximately unbiased under MCAR and MARTAIL for the proportions of y and for the conditional means of *pb* given $y = 0$, and given $y = 1$. For the conditional mean of *pb* given $y = 2$, however, $\overline{Q}_m$ is biased for POR under MCAR and strongly biased for POR and DIS under MARTAIL. Under MARRIGHT, the bias of $\overline{Q}_m$ is considerably larger for DIS than for POR. This is especially the case for the conditional means. Under MARMID, $\overline{Q}_m$ is approximately unbiased for both methods DIS and POR in most of the cases. An exception is the conditional mean of *pb* given $y = 1$ for DIS, where $\overline{Q}_m$ has a bias of 0.52.

5.3.2 Compound imputation methods

The results for the compound imputation methods are presented in the Tables 5.11, through 5.14. To reduce the number of tables, results are only given for $\widehat{Q}$, $\widehat{E}[\widehat{Q}_{inc}]$, $\widehat{E}[\overline{Q}_m]$ and the actual interval coverage. In each table, results are presented for the complete data set generated according to the statistical model corresponding to the method and for the raw data set. The number of Gibbs sampling iterations is 5.

Numerical data

Results for Π_{num} are found in Table 5.11. In this Table, correl(x_i, x_j) stands for the correlation coefficient between x_i and x_j.

Results for the multivariate normal data set For the multivariate normal data set NUMMODD, Π_{num} is approximately proper for most statistics. The pooled point estimate $\overline{Q}_m$ is approximately unbiased or slightly biased with a bias ranging from 0.00 (mean of x_1) to 0.13 (third quartile of x_3). For 18 of the 30 target statistics, the estimated actual interval coverage

lies in the interval 95% ±1.9%. The actual interval coverages for the first quartile of x_4, and for the correlation coefficient between x_2 and x_3 are quite low with values of 88.8% and 87.6%. The intervals for the first quartile of x_3, the mean of x_4, and the correlation between x_3 and x_4 are slightly undercovered with values of 91.4, 92.4 and 90.6. The intervals for the median of x_1, the first quartile of x_2, the median of x_3, the third quartile of x_4, and the correlation coefficients between x_1 and x_3, between x_1 and x_4, and between x_4 and x_5, are overcovered (estimated actual interval coverage > 96.9%) with estimated actual coverages of 97.4, 98.4, 98.0, 97.8, 97.4, 98.0, and 97.0, respectively.

For the data set NUMMODD and any target statistic, the difference between $\widehat{Q}$ and $\widehat{E}[\widehat{Q}_{inc}]$ is considerably larger than the difference between $\widehat{Q}$ and $\widehat{E}[\overline{Q}_m]$. In this respect, the results for the median of x_1 ($\widehat{Q} = 13.06, \widehat{E}[\overline{Q}_{\text{inc}}] = 11.86, \widehat{E}[\overline{Q}_m] = 13.05$), the median of x_3 ($\widehat{Q} = 11.18, \widehat{E}[\overline{Q}_{\text{inc}}] = 9.97, \widehat{E}[\overline{Q}_m] = 11.17$), the correlation coefficient between x_1 and x_4 ($\widehat{Q} = 0.71, \widehat{E}[\overline{Q}_{\text{inc}}] = 0.44, \widehat{E}[\overline{Q}_m] = 0.70$), and the correlation coefficient between x_2 and x_3 ($\widehat{Q} = 058, \widehat{E}[\overline{Q}_{\text{inc}}] = 0.26, \widehat{E}[\overline{Q}_m] = 0.55$) are striking. In view of this performance, the fact that for the correlation coefficient between x_2 and x_3 the actual interval coverage is lower than 90% and $\overline{Q}_m$ is slightly biased should not be considered as a serious problem.

Robustness against non-normal data According to Table 5.11, Π_{num} seems to be robust against non-normal data. Similar to the results for the data set NUMMODD, for the data set NUMRAWD, $\overline{Q}_m$ is approximately unbiased or slightly biased, most of the actual interval coverages are close to 95%, and the differences between $\widehat{Q}$ and $\widehat{E}[\widehat{Q}_{inc}]$ are considerably larger than the differences between $\widehat{Q}$ and $\widehat{E}[\overline{Q}_m]$. However, the performance of π_{num} for NUMRAWD is slightly worse than for NUMMODD. This is most clearly seen from the estimated actual interval coverages of 73.8 and 54.6 for the correlation coefficients between x_1 and x_3 and between x_1 and x_4, which are much lower than any actual interval coverage for the data set NUMMOD. Moreover, for the data set NUMRAWD, the number of estimated actual interval coverages lying in the interval 95% ± 1.9% is 13, as compared to 18 for the data set NUMMODD. Furthermore, $\overline{Q}_m$ is slightly more biased for the raw data set than for the multivariate normal data set. For NUMRAWD and all univariate target statistics, the bias of $\overline{Q}_m$ ranges from 0.00 (first quartile of x_3) to 0.19 (median of x_2), while for the NUMMODD and the same target statistics this

bias ranges from 0.00 (the mean of x_1) to 0.13 (third quartile of x_3). This is also the case for the correlation coefficients, where for data set NUMRAWD the bias of $\overline{Q}_m$ ranges from 0.00 (correlation coefficient between x_3 and x_5) to 0.05 (correlation coefficient between x_1 and x_4) and for the data set NUMMODD, this bias ranges from 0.00 (correlation coefficient between x_3 and x_5) to 0.03 (correlation coefficient between x_2 and x_3).

Mixed data

The results for the method Π_{mix} are presented in the Tables 5.12 and 5.13. In these Tables, cramc(x_i,x_j) stands for the Cramer-C measure. When x_i and x_j are two binary variables, l-OR(x_i,x_j) stands for the log-OR between x_i and x_j. For a binary variable x_i and a trichotomous variable x_j, l-OR(x_i,$x_j - s$), with $s = 1, 2$, stands for the log-OR between x_i and x_j for the categories 0 and s of x_j, where 0 is the reference category.

Results for the data set according to the imputation model The performance of Π_{mix} for the data set according to the imputation model MIXMODD is good. In most of the cases, $\overline{Q}_m$ is approximately unbiased. Biased estimates $\overline{Q}_m$ are obtained for the third quartile of x_1 (bias = 0.23), the conditional means of x_1 given $x_3 = 1$ (bias = 0.16), x_4 given $x_3 = 1$ (bias = 0.12), x_4 given $x_3 = 2$ (bias = 0.16), x_5 given $x_3 = 1$ (bias = 0.15), and x_5 given $x_3 = 2$ (bias = 0.13), and the log-OR between x_2 and x_3 for the categories 0 and 1 of x_3 (bias = 0.15). For every target statistic inspected, the bias of $\widehat{Q}_{inc}$ is larger than the bias of $\overline{Q}_m$.

For 36 of the 47 target statistics for which the actual interval coverage has been estimated, this coverage lies in the interval 95%±1.9%. The estimated actual coverage is low with a value of 74.4 for the third quartile of x_1. This is partially due to the bias of $\overline{Q}_m$ for this target statistic. Slight underestimates (90% < estimated actual interval coverage < 93.1%) are obtained for confidence intervals for the mean of x_1, the first quartile of x_4, and the conditional means of x_1 given that $x_7 = 2$ and of x_5 given that $x_2 = 0$. Overestimates (estimated actual interval coverage > 96.9%) are obtained for the confidence intervals for the first quartile of x_1, for the median and the third quartile of x_4, for the proportions of the categories 1 and 2 of x_3, and for the correlation coefficient between x_1 and x_4.

Robustness According to the Tables 5.12 and 5.13, Π_{mix} seems to be robust against data generated from a probability distribution deviating from the statistical model corresponding to Π_{mix}, although for the data set MIXMODD generated according to the statistical model corresponding to $\widetilde{\Pi}_{\text{mix}}$, the performance of Π_{mix} is somewhat better than for the raw data set MIXRAWD. This is most clearly seen from the actual interval coverages. For MIXRAWD, the number of estimated interval coverages lying in the interval $95\% \pm 1.9\%$ is 19, while for MIXMODD this number is 36. Furthermore, for MIXRAWD, $\overline{Q}_m$ is biased for the correlation coefficients whereas for MIXMODD it is not. For the other target statistics the degree of bias of $\overline{Q}_m$ is comparable for both data sets.

Categorical data

The results for Π_{cat} are found in Table 5.14.

Results for data set according to imputation model For the data CATMODD, generated according to the statistical model corresponding to $\widetilde{\Pi}_{\text{cat}}$, $\overline{Q}_m$ is approximately unbiased for the proportions and the conditional means, except the conditional mean of x_3 given $x_5 = 0$, where $\overline{Q}_m$ is slightly biased with a bias of 0.11. This bias is probably due to the low proportion (0.11) of the category 0 of x_5. Due to the weak associations between the variables, the differences between $\widehat{Q}$ and $\widehat{E}[\widehat{Q}_{inc}]$ are very small for the proportions.

For the Cramer-C measure between x_1 and x_6, the bias of $\overline{Q}_m$ is larger than the bias of $\widehat{Q}_{inc}$. The performance of Π_{cat} is particularly bad for the log-OR between x_4 and x_5 and the two log-OR between x_5 and x_6, where the bias of $\overline{Q}_m$ is considerably larger than the bias of $\widehat{Q}_{inc}$. An explanation of this bad performance may be the skewness of x_5 and x_6, probably causing low bivariate cell frequencies. The high actual interval coverages, each larger than 99%, for these three log-OR may be regarded as a compensation mechanism, where B_m is boosted to compensate for the loss of information in the data set.

Robustness Contrary to the results for Π_{num} and for Π_{mix}, the performance of Π_{cat} is better for the raw data set CATRAWD than for the data set CATMODD. This is especially the case for the log-OR, where for CATRAWD the performance of Π_{cat} is bad only for the two log-OR between x_4 and x_5 and between x_5 and x_6 for the categories 0 and 1. For the other three

log-OR (between x_2 and x_4, between x_2 and x_5 and between x_4 and x_5), the bias of $\overline{Q}_m$ for the data set CATMODD is considerably larger than for the data set CATRAWD. Further, for the Cramer-C measure between x_1 and x_6, $\overline{Q}_m$ is approximately unbiased for CATRAWD and biased for CATMODD. Finally, the number of estimated actual coverages lying in the interval $95\% \pm 1.9\%$ is 21 for CATRAWD set and 19 for CATMODD.

5.4 Conclusion

From the results described in the previous section, the following conclusions can be drawn:

1. The performance of the elementary imputation methods LRN, LOR and POR is generally good.

2. The elementary method POR appears to be superior over DIS. For the method DIS, the between imputation variance is strongly underestimated while this is not the case for POR. The bias of pooled point estimates $\overline{Q}_m$ of $\widehat{Q}$ for DIS is considerably larger than for POR under some MAR missing data mechanisms.

3. When the association between variables is strong and the distribution of the categorical variables is not too skew, the performance of imputation methods for entirely numerical or mixed data sets is generally good. For such types of data sets, an imputation method Π seems to be robust against a deviation of the probability distribution from the statistical model corresponding to Π, although the performance of Π appears to be somewhat better for complete data sets generated according to the statistical model corresponding to Π.

4. When the associations between the variables in a data set are strong, the presence of circularities in the imputation model seems to have no or only little effect on the properness of an imputation method.

5. When for a compound imputation method the number of imputation variables is modest, five Gibbs sampling iterations appear to be sufficient.

6. When the probability distributions of some categorical variables in a data set are skew, the performance of multiple imputation for the log-OR corresponding to such variables

may be bad.

5.5 Discussion and future research

The results of this study are encouraging and illustrate that multiple imputation is suitable for bias reduction if incomplete cases differ systematically from complete cases. These results also clearly stress the added value of multiple imputation with regard to complete-case analysis.

The robustness against deviations of the probability distribution from the corresponding statistical model for the compound imputation methods Π_{num}, Π_{mix} and Π_{cat} is in accordance with the results of other simulation studies concerning multiple imputation [2,9,10]. The simulation study of [9], also reported in chapter 4 of [2], shows that, even for univariate populations which are skewed or heavy-tailed, the intervals for the population mean resulting from imputation on the basis of a univariate normal model have an actual coverage which is very close to the nominal coverage. In an American study [10], designed to imitate the process of data collection and the underlying missing data mechanism in the Third National Health and Nutrition Examination Survey (NHANES-III), it has been illustrated that the performance of model based multiple imputation appears to be good in the situation of sample surveys (sample surveys are also described in chapter 4).

Strongly systematic MAR missing data mechanisms have been applied, since it is plausible that if multiple imputation is proper under strongly systematic MAR mechanisms, it will also be proper for less systematic and thus more realistic MAR missing data mechanisms. In this way, a good impression of the validity of imputation methods can be obtained with a relatively small number of combinations of complete data sets and missing data mechanisms. Furthermore, with strongly systematic MAR mechanisms, the added value of multiple imputation with respect to complete-case analysis can also be stressed. Only complete data sets with strong associations between the different variables are considered, since for complete data sets with weak associations, it is not possible to construct a MAR missing data mechanism generating incomplete data sets with incomplete cases differing systematically from complete cases. Topics for future research are discussed below.

5.5.1 Future research

Robustness

For the automation of a selection strategy of imputation methods described in chapter 4, it is important to investigate whether the performance of compound imputation methods consisting of the elementary methods LRN, LOR and POR is sufficient under the MAR assumption in most practical cases. As in the simulation study of multiple imputation for the Third National Health and Nutrition Examination Survey (NHANES-III) in [10,11], this can be achieved by evaluating these methods under repetition of sampling and generation of incomplete data for several sampling and missing data mechanisms mimicked from real sample survey situations. The sampling mechanism can be imitated by drawing random samples from an artificial population created by drawing a large random sample without replacement, say $n = 2000$, from all complete cases in the survey. Under the MAR assumption, the parameters of the missing data mechanisms can be estimated from the entire survey. In such a case, statistical inference from multiple imputation with respect to the population quantity Q, rather than with respect to $\widehat{Q}$, is evaluated as in the validation study described in the previous sections. This is so because when evaluating the robustness of imputation methods, effects of model misspecification are relevant, and for an end-user the validity of statistical inference with respect to Q is more interesting than validity with respect to $\widehat{Q}$. To investigate the robustness against deviations from MAR, it is useful to use MNAR mechanisms generating approximately the same fraction of incomplete cases with approximately the same distribution of the different missing data patterns over the incomplete cases, as the MAR missing data mechanism applied here.

By mimicking sample survey situations, generally the robustness of model-based multiple imputation against moderate deviations of the probability distribution from the statistical model corresponding to the imputation method is examined [11]. It is worthwhile to investigate the robustness of imputation methods against more extreme departures. Departures which can be considered for numerical imputation variables y and numerical predictor variables x are non-linear or non-monotonic relationships between y and x, and heteroscedastic, skew or heavy-tailed error-terms in the regression model of y on x. Data sets with such departures can be artificially generated in the same way as complete data sets corresponding to the statistical

model of compound imputation methods as described in subsection 5.2.2.

When in some situations the compound imputation methods consisting of LRN, LOR and POR are not robust against deviations of the probability distribution from the statistical model corresponding to the imputation method, it is worthwhile to investigate whether in these situations better compound imputation methods can be constructed from a more extended set of elementary imputation methods. For numerical imputation variables the selection strategy of elementary imputation methods proposed in step 2 of subsection 4.2.3 of chapter 4 can be tried. For categorical imputation variables it can be investigated whether the performance of multiple imputation can be improved by extending the corresponding elementary imputation methods LOR or POR by interaction terms and by transforming numerical predictor variables.

Other topics

- **Convergence.** In the simulation study described here, only data sets with a small number of imputation variables are considered. Since convergence of the Gibbs sampling algorithm is reported to be slow for a large number of variables [12], it should be investigated how the required number of Gibbs sampling iterations is related to the number of imputation variables;

- **Added value of the variable-by-variable Gibbs sampling approach with respect to existing imputation methods using a multivariate statistical model.** The following issues are relevant to consider:

  - The performance of multiple imputation for large data sets with many variables. It is assumed that if the number of imputation variables is large, multiple imputation on the basis of a multivariate statistical model may become numerically unstable (see chapter 4);
  - Whether for the variable-by-variable Gibbs sampling approach, multiple imputation is more robust against deviations of the probability distribution from the statistical model corresponding to the imputation method, than imputation methods using a multivariate statistical model.

- **Constraints with respect to the validity of multiple imputation.** An important question here is: is it possible to construct diagnostic measures as a function of the observed data set and the target statistic $\left(\widehat{Q}, U\right)$, on the basis of which it can be concluded whether valid statistical inference with multiple imputation under the MAR assumption is possible? It is plausible that given the target statistic $\left(\widehat{Q}, U\right)$, the validity of multiple imputation will depend on the following factors:

  - The number of complete cases in the data set. In order to estimate the parameters of the imputation model this number should be sufficiently large;
  - The fraction of missing data entries among all entries in the data set;
  - The number of observed data entries per variable. This measure is important for univariate statistical analyses. If a certain variable has only few observations, multiple imputation will be improper for the mean of this variable;
  - For each pair of variables x and y, the number of cases for which x and y are simultaneously observed. If this number is small, then the performance of multiple imputation is bad for the correlation coefficient between x and y;
  - The deviation from MCAR of the underlying missing data mechanism;
  - The deviation of the probability distribution from the statistical model corresponding to the imputation method.

Appendix 5.A Proof of Equations 5.4 and 5.5

Let n_i be the number of cases nominated for missing data pattern t_i^T and w_i the number of incomplete cases with missing data pattern t_i^T. Eq. 5.4 directly follows from Eq. 5.3. Eq. 5.5 is proved by writing $E[w_i]$ in two ways. First $E[w_i] = \alpha E[n_i] = \alpha f_i n$, since the cases nominated for missing data pattern t_i^T are randomly selected with probability f_i from the n cases in the data set, and an expected fraction α of these nominated cases is made incomplete according to missing data pattern t_i^T. Let n_{ij} be the number of cases nominated for missing data pattern t_i^T with $c_i(\theta_{ij}) \leq s_i < c_i(\theta_{i,j+1})$ and w_{ij} the number of incomplete cases with missing data pattern t_i^T and $c_i(\theta_{ij}) \leq s_i < c_i(\theta_{i,j+1})$. From the definition of w_{ij} it follows that $w_i = \sum_j w_{ij}$. The cases nominated for missing data pattern t_i^T are randomly selected from the cases of the data

set, so that $E[n_{ij}] = (\theta_{i,j+1} - \theta_{i,j})\, E[n_i]$. From the cases nominated for missing data pattern t_i^T with $c_i(\theta_{ij}) \leq s_i < c_i(\theta_{i,j+1})$, a fraction of $\lambda_i g_{ij}$ is made incomplete, which implies that $E[w_{ij}] = \lambda_i g_{ij} E[n_{ij}]$. Consequently,

$$\begin{aligned} E[w_i] &= E\left[\textstyle\sum_j w_{ij}\right] \\ &= \textstyle\sum_j E[w_{ij}] \\ &= \textstyle\sum_j \lambda_i g_{ij} E[n_{ij}] \\ &= \textstyle\sum_j \lambda_i g_{ij} (\theta_{i,j+1} - \theta_{i,j})\, E[n_i] \\ &= \textstyle\sum_j \lambda_i g_{ij} (\theta_{i,j+1} - \theta_{i,j})\, n f_i. \end{aligned}$$

Both expressions of $E[w_i]$ imply

$$\alpha f_i n = \sum_j \lambda_i g_{ij} (\theta_{i,j+1} - \theta_{i,j})\, n f_i,$$

so that

$$\lambda_i = \frac{\alpha}{\sum_j (\theta_{i,j+1} - \theta_{i,j})\, g_{ij}}.$$

Appendix 5.B Simulation results

MDM	statistic	$\widehat{Q}$	$\widehat{E}[\widehat{Q}_{inc}]$	$\widehat{E}[\overline{Q}_m]$	U	$\widehat{E}[\overline{U}_m]$	$\widehat{B}$	$\widehat{E}[B_m]$	cover
MCAR	mean	11.44	11.46	11.45	0.06	0.06	0.03	0.03	97.2
	Q1 (25%)	8.04	8.12	8.04	0.09	0.10	0.05	0.05	97.2
	median	11.41	11.40	11.34	0.09	0.10	0.04	0.05	96.8
	Q3 (75%)	14.64	14.64	14.68	0.10	0.11	0.05	0.06	95.2
	corr(y, x_1)	0.70	0.70	0.69			0.03	0.00	96.8
	corr(y, x_2)	0.57	0.57	0.57			0.01	0.00	96.0
	corr(y, x_3)	0.64	0.64	0.63			0.01	0.00	98.6
	corr(y, x_4)	0.58	0.58	0.57			0.01	0.00	98.0
MARRIGHT	mean	11.44	9.69	11.43	0.06	0.06	0.05	0.05	96.2
	Q1 (25%)	8.04	6.86	7.96	0.09	0.10	0.02	0.03	96.8
	median	11.41	9.55	11.37	0.09	0.10	0.06	0.08	97.8
	Q3 (75%)	14.64	12.79	14.66	0.10	0.12	0.10	0.13	95.8
	corr(y, x_1)	0.70	0.60	0.67			0.02	0.00	93.0
	corr(y, x_2)	0.57	0.42	0.53			0.01	0.00	84.2
	corr(y, x_3)	0.64	0.56	0.63			0.01	0.00	97.2
	corr(y, x_4)	0.58	0.46	0.56			0.01	0.00	95.0
MARTAIL	mean	11.44	11.34	11.42	0.06	0.06	0.03	0.03	97.2
	Q1 (25%)	8.04	8.37	8.04	0.09	0.09	0.04	0.05	97.8
	median	11.41	11.15	11.22	0.09	0.10	0.03	0.04	93.4
	Q3 (75%)	14.64	14.22	14.67	0.10	0.12	0.05	0.06	95.2
	corr(y, x_1)	0.70	0.61	0.69			0.03	0.00	96.4
	corr(y, x_2)	0.57	0.47	0.56			0.01	0.00	96.8
	corr(y, x_3)	0.64	0.55	0.63			0.01	0.00	95.6
	corr(y, x_4)	0.58	0.48	0.57			0.01	0.00	96.6
MARMID	mean	11.44	11.61	11.49	0.06	0.06	0.02	0.03	97.6
	Q1 (25%)	8.04	7.52	8.03	0.09	0.12	0.03	0.06	99.8
	median	11.41	11.82	11.51	0.09	0.09	0.02	0.05	98.6
	Q3 (75%)	14.64	15.27	14.70	0.10	0.10	0.03	0.07	99.0
	corr(y, x_1)	0.70	0.77	0.70			0.03	0.00	98.2
	corr(y, x_2)	0.57	0.66	0.57			0.01	0.00	97.8
	corr(y, x_3)	0.64	0.72	0.64			0.01	0.00	98.8
	corr(y, x_4)	0.58	0.67	0.57			0.01	0.00	99.0

Table 5.7: Simulation results for the elementary imputation method LRN, for the complete data set LID, and for the missing data mechanisms MCAR, MARRIGHT, MARTAIL and MARMID.

MDM	statistic	$\widehat{Q}$	$\widehat{E}[\widehat{Q}_{inc}]$	$\widehat{E}[\overline{Q}_m]$	U	$\widehat{E}[\overline{U}_m]$	$\widehat{B}$	$\widehat{E}[B_m]$	cover
MCAR	prop (y=0)	0.47	0.47	0.47			0.00	0.00	93,0
	prop (y=1)	0.53	0.53	0.53			0.00	0.00	93.0
	mean(x_1) | y=0	8.63	8.62	8.65	0.08	0.08	0.04	0.04	95.0
	mean(x_1) | y=1	15.29	15.31	15.29	0.12	0.13	0.04	0.04	94.4
	mean(x_2) | y=0	9.10	9.09	9.11	0.10	0.10	0.04	0.04	94.6
	mean(x_2) | y=1	13.18	13.18	13.18	0.10	0.11	0.04	0.03	92.2
	mean(x_3) | y=0	7.12	7.12	7.13	0.06	0.06	0.03	0.03	95.4
	mean(x_3) | y=1	13.32	13.34	13.32	0.10	0.10	0.03	0.03	94.4
	mean(x_4) | y=0	7.08	7.06	7.10	0.07	0.07	0.03	0.03	94.6
	mean(x_4) | y=1	11.66	11.66	11.66	0.13	0.13	0.03	0.03	94.4
MARRIGHT	prop (y=0)	0.47	0.63	0.45			0.00	0.00	81.6
	prop (y=1)	0.53	0.37	0.55			0.00	0.00	81.6
	mean(x_1) | y=0	8.63	7.72	8.40	0.08	0.08	0.08	0.07	83.6
	mean(x_1) | y=1	15.29	12.97	15.23	0.12	0.12	0.01	0.0	93.0
	mean(x_2) | y=0	9.10	8.55	9.01	0.10	0.10	0.05	0.05	92.4
	mean(x_2) | y=1	13.18	11.74	13.10	0.10	0.10	0.02	0.02	93.0
	mean(x_3) | y=0	7.12	6.34	6.94	0.06	0.06	0.05	0.05	86.2
	mean($x3$) | y=1	13.32	11.31	13.24	0.10	0.10	0.01	0.01	88.8
	mean(x_4) | y=0	7.08	6.46	6.94	0.07	0.07	0.05	0.05	88.6
	mean(x_4) | y=1	11.66	9.96	11.60	0.13	0.12	0.01	0.02	97.6
MARTAIL	prop (y=0)	0.47	0.45	0.46			0.00	0.00	95.0
	prop (y=1)	0.53	0.55	0.54			0.00	0.00	95.0
	mean(x_1) | y=0	8.63	9.56	8.65	0.08	0.08	0.04	0.04	95.2
	mean(x_1) | y=1	15.29	14.07	15.16	0.12	0.13	0.04	0.05	96.0
	mean(x_2) | y=0	9.10	9.77	9.07	0.10	0.09	0.02	0.03	95.0
	mean(x_2) | y=1	13.18	12.53	13.13	0.10	0.10	0.02	0.03	96.8
	mean(x_3) | y=0	7.12	8.01	7.18	0.06	0.06	0.02	0.03	96.8
	mean(x_3) | y=1	13.32	12.21	13.16	0.10	0.10	0.03	0.03	92.8
	mean(x_4) | y=0	7.08	7.62	7.10	0.07	0.07	0.02	0.03	96.6
	mean(x_4) | y=1	11.66	10.75	11.57	0.13	0.13	0.02	0.03	97.0
MARMID	prop (y=0)	0.47	0.45	0.46			0.00	0.00	95.0
	prop (y=1)	0.53	0.55	0.54			0.00	0.00	95.0
	mean(x_1) | y=0	8.63	9.56	8.65	0.08	0.08	0.04	0.04	95.2
	mean(x_1) | y=1	15.29	14.07	15.16	0.12	0.13	0.04	0.05	96.0
	mean(x_2) | y=0	9.10	9.77	9.07	0.10	0.09	0.02	0.03	95.0
	mean(x_2) | y=1	13.18	12.53	13.13	0.10	0.10	0.02	0.03	95.0
	mean(x_3) | y=0	7.12	8.01	7.18	0.06	0.06	0.02	0.03	96.8
	mean(x_3) | y=1	13.32	12.21	13.16	0.10	0.10	0.03	0.03	92.8
	mean(x_4) | y=0	7.08	7.62	7.10	0.07	0.07	0.02	0.03	96.6
	mean(x_4) | y=1	11.66	10.75	11.57	0.13	0.13	0.02	0.03	97.0

Table 5.8: Simulation results for the elementary imputation method LOR, for the complete data set LOD, and for the missing data mechanisms MCAR, MARRIGHT, MARTAIL, and MARMID.

MDM	statistic	$\widehat{Q}$	$\widehat{E}[\widehat{Q}_{inc}]$	$\widehat{E}[\overline{Q}_m]$	U	$\widehat{E}[\overline{U}_m]$	B	$\widehat{E}[B_m]$	cover
MCAR	prop (y=0)	0.56	0.56	0.54			0.00	0.00	94.4
	prop (y=1)	0.27	0.27	0.27			0.00	0.00	95.4
	prop (y=2)	0.18	0.18	0.19			0.00	0.00	96.2
	mean(pb) (y=0)	8.08	8.07	8.04	0.02	0.02	0.01	0.01	94.0
	mean(pb) (y=1)	6.95	6.95	7.00	0.03	0.03	0.02	0.03	96.0
	mean(pb) (y=2)	6.84	6.85	6.97	0.05	0.05	0.05	0.06	95.0
MARRIGHT	prop (y=0)	0.56	0.64	0.53			0.00	0.00	89.8
	prop (y=1)	0.27	0.19	0.28			0.00	0.00	96.2
	prop (y=2)	0.18	0.16	0.19			0.00	0.00	97.8
	mean(pb) (y=0)	8.08	8.64	8.11	0.02	0.02	0.01	0.01	95.8
	mean(pb) (y=1)	6.95	7.81	6.95	0.03	0.03	0.02	0.03	95.2
	mean(pb) (y=2)	6.84	7.51	6.99	0.05	0.05	0.03	0.04	96.4
MARTAIL	prop (y=0)	0.56	0.54	0.53			0.00	0.00	90.8
	prop (y=1)	0.27	0.29	0.29			0.00	0.00	91.4
	prop (y=2)	0.18	0.16	0.18			0.00	0.00	98.0
	mean(pb) (y=0)	8.08	8.14	8.03	0.02	0.02	0.01	0.01	94.8
	mean(pb) (y=1)	6.95	7.39	6.99	0.03	0.03	0.02	0.02	98.6
	mean(pb) (y=2)	6.84	7.25	7.09	0.05	0.06	0.06	0.07	92.0
MARMID	prop (y=0)	0.56	0.58	0.56			0.00	0.00	98.6
	prop (y=1)	0.27	0.23	0.25			0.00	0.00	96.0
	prop (y=2)	0.18	0.19	0.19			0.00	0.00	98.0
	mean(pb) (y=0)	8.08	8.00	8.09	0.02	0.02	0.00	0.01	97.8
	mean(pb) (y=1)	6.95	6.15	6.86	0.03	0.03	0.03	0.07	97.6
	mean(pb) (y=2)	6.84	6.33	6.83	0.05	0.04	0.03	0.05	98.8

Table 5.9: Simulation results for the elementary method POR, for the Mammographic Experience data set, and for the missing data mechanisms MCAR, MARRIGHT, MARTAIL and MARMID.

MDM	statistic	$\widehat{Q}$	$\widehat{E}$	$\widehat{Q}_{inc}$	$\widehat{E}[\overline{Q}_m]$	U	$E[\overline{U}_m]$	B	$\widehat{E}[B_m]$	cover
MCAR	prop (y=0)	0.56	0.56		0.58			0.01	0.00	52.8
	prop (y=1)	0.27	0.27		0.27			0.01	0.00	28.8
	prop (y=2)	0.18	0.18		0.16			0.00	0.00	37.6
	mean(pb) (y=0)	8.08	8.08		8.04	0.02	0.02	0.02	0.00	74.0
	mean(pb) (y=1)	6.95	6.95		6.91	0.03	0.03	0.02	0.01	69.4
	mean(pb) (y=2)	6.84	6.84		6.96	0.05	0.07	0.08	0.02	59.6
MARRIGHT	prop (y=0)	0.56	0.64		0.64			0.01	0.00	50.4
	prop (y=1)	0.27	0.19		0.21			0.01	0.00	42.2
	prop (y=2)	0.18	0.16		0.15			0.00	0.00	41.4
	mean(pb) (y=0)	8.08	8.64		7.85	0.02	0.02	0.05	0.01	52.0
	mean(pb) (y=1)	6.95	7.80		7.29	0.03	0.05	0.19	0.02	51.8
	mean(pb) (y=2)	6.84	7.50		7.23	0.05	0.09	0.10	0.02	49.8
MARTAIL	prop (y=0)	0.56	0.54		0.55			0.00	0.00	71.0
	prop (y=1)	0.27	0.29		0.31			0.00	0.00	28.8
	prop (y=2)	0.18	0.16		0.13			0.00	0.00	42.2
	mean(pb) (y=0)	8.08	8.14		8.06	0.02	0.02	0.02	0.01	85.0
	mean(pb) (y=1)	6.95	7.38		6.88	0.03	0.03	0.04	0.01	66.4
	mean(pb) (y=2)	6.84	7.25		7.21	0.05	0.09	0.08	0.03	44.8
MARMID	prop (y=0)	0.56	0.57		0.62			0.00	0.00	64.0
	prop (y=1)	0.27	0.23		0.18			0.00	0.00	38.2
	prop (y=2)	0.18	0.19		0.21			0.00	0.00	50.0
	mean(pb) (y=0)	8.08	8.00		8.10	0.02	0.02	0.01	0.00	75.8
	mean(pb) (y=1)	6.95	6.15		6.43	0.03	0.03	0.01	0.00	41.4
	mean(pb) (y=2)	6.84	6.32		6.76	0.05	0.04	0.09	0.03	78.4

Table 5.10: Simulation results for elementary method DIS, for the Mammographic Experience data set, and for the missing data mechanisms MCAR, MARRIGHT, MARTAIL and MARMID.

data	NUMMODD					NUMRAWD				
statistic	$\widehat{Q}$	$\widehat{E}$	$\widehat{Q}_{inc}$	$\widehat{E}[\overline{Q}_m]$	cover	$\widehat{Q}$	$\widehat{E}$	$\widehat{Q}_{inc}$	$\widehat{E}[\overline{Q}_m]$	cover
mean (x_1)	12.93	11.98	12.93	95.0		12.05	11.11	12.08	94.2	
Q1 (x_1) (25%)	9.02	8.10	9.08	94.2		7.67	6.81	7.83	86.0	
med (x_1) (50%)	13.06	11.86	13.05	97.4		11.63	10.33	11.68	97.4	
Q3 (x_1) (75%)	16.87	15.77	16.76	96.2		15.71	14.67	15.77	96.8	
mean (x_2)	12.16	11.40	12.14	95.6		11.45	10.74	11.45	93.2	
Q1 (x_2) (25%)	8.48	7.96	8.53	98.4		7.83	7.35	7.74	91.2	
med (x_2) (50%)	12.31	11.50	12.27	94.4		10.54	9.78	10.73	83.4	
Q3 (x_2) (75%)	15.59	14.80	15.62	94.0		14.54	13.51	14.63	92.0	
mean (x_3)	11.08	10.18	11.10	95.8		10.23	9.39	10.23	94.0	
Q1 (x_3) (25%)	7.48	6.83	7.54	91.4		6.92	6.22	6.92	97.0	
med (x_3) (50%)	11.18	9.97	11.17	98.0		10.04	9.14	10.01	93.2	
Q3 (x_3) (75%)	14.38	13.41	14.51	94.0		13.25	11.92	13.22	97.6	
mean (x_4)	10.64	9.75	10.60	92.4		9.70	8.85	9.72	93.2	
Q1 (x_4) (25%)	6.96	6.46	7.07	88.8		6.17	5.54	6.19	92.2	
med (x_4) (50%)	10.47	9.26	10.53	94.4		9.21	8.27	9.25	96.6	
Q3 (x_4) (75%)	13.96	12.89	13.93	97.8		12.96	11.34	12.84	95.8	
correl (x_1,x_2)	0.69	0.65	0.69	96.8		0.69	0.65	0.70	93.0	
correl (x_1,x_3)	0.83	0.81	0.83	97.4		0.82	0.80	0.84	73.8	
correl (x_1,x_4)	0.71	0.44	0.70	98.0		0.71	0.41	0.76	54.6	
correl (x_1,x_5)	0.73	0.72	0.72	94.2		0.73	0.71	0.72	94.2	
correl (x_1,x_6)	0.61	0.61	0.61	95.2		0.56	0.54	0.56	96.0	
correl (x_2,x_3)	0.58	0.26	0.55	87.6		0.57	0.27	0.61	89.0	
correl (x_2,x_4)	0.68	0.64	0.68	96.8		0.62	0.58	0.64	87.0	
correl (x_2,x_5)	0.61	0.60	0.61	94.6		0.56	0.54	0.55	95.8	
correl (x_2,x_6)	0.49	0.49	0.49	96.4		0.37	0.36	0.38	91.8	
correl (x_3,x_4)	0.77	0.72	0.75	90.6		0.79	0.73	0.77	90.6	
correl (x_3,x_5)	0.81	0.80	0.81	94.8		0.81	0.80	0.81	96.0	
correl (x_3,x_6)	0.67	0.67	0.67	96.8		0.65	0.63	0.64	89.4	
correl (x_4,x_5)	0.81	0.81	0.81	97.0		0.83	0.82	0.84	90.2	
correl (x_4,x_6)	0.74	0.74	0.74	95.0		0.76	0.75	0.76	93.4	

Table 5.11: Simulation results for the compound imputation method Π_{num}, for the data sets NUMMODD and NUMRAWD and for a MAR missing data mechanism. The data set NUMMODD is generated according to the multivariate normal distribution and the data set NUMRAWD is generated by sampling from the raw data set of Irish Windspeeds.

data	MIXMODD				MIXRAWD			
statistic	$\widehat{Q}$	$\widehat{E}\left[\widehat{Q}_{inc}\right]$	$\widehat{E}\left[\overline{Q}_m\right]$	cover	$\widehat{Q}$	$\widehat{E}\left[\widehat{Q}_{inc}\right]$	$\widehat{E}\left[\overline{Q}_m\right]$	cover
mean (x_1)	12.08	11.33	12.09	92.6	12.29	11.50	12.29	89.4
Q1 (x_1) (25%)	8.78	8.14	8.78	98.4	8.96	8.15	8.92	95.2
med (x_1) (50%)	11.90	11.07	11.90	96.0	11.67	10.81	11.77	89.8
Q3 (x_1) (75%)	14.93	14.20	15.16	74.4	15.11	14.30	15.31	85.8
prop (x_2=0)	0.48	0.55	0.47	96.0	0.49	0.56	0.49	95.2
prop (x_2=1)	0.52	0.45	0.53	96.0	0.51	0.44	0.51	95.2
prop (x_3=0)	0.34	0.41	0.34	95.2	0.32	0.39	0.32	95.2
prop (x_3=1)	0.35	0.33	0.35	97.8	0.34	0.34	0.34	97.6
prop (x_3=2)	0.32	0.26	0.31	98.0	0.34	0.27	0.34	97.8
mean (x_4)	9.67	9.01	9.67	96.2	9.84	9.18	9.86	93.0
Q1 (x_4) (25%)	6.85	6.23	6.81	92.6	6.84	6.32	6.91	93.0
med (x_4) (50%)	9.37	8.66	9.35	97.2	9.49	8.79	9.52	91.4
Q3 (x_4) (75%)	12.20	11.36	12.21	97.0	12.12	11.27	12.21	92.4
mean1 (x_2=0)	9.06	8.77	9.07	93.6	9.16	8.90	9.18	96.4
mean1 (x_2=1)	14.81	13.99	14.80	94.6	15.31	14.41	15.30	89.2
mean1 (x_3=0)	8.70	8.46	8.67	95.4	8.73	8.56	8.79	97.2
mean1 (x_3=1)	12.18	11.39	12.34	96.2	11.89	11.10	12.05	96.6
mean1 (x_3=2)	15.53	14.76	15.48	94.2	16.03	15.05	15.81	88.6
correl (x_1,x_4)	0.56	0.21	0.56	97.8	0.61	0.33	0.65	74.0
correl (x_1,x_5)	0.67	0.66	0.67	95.4	0.66	0.66	0.68	79.4
mean1 (x_6=0)	10.17	9.77	10.20	93.6	10.35	9.92	10.37	95.0
mean1 (x_6=1)	14.22	13.67	14.21	95.0	14.27	13.61	14.25	89.4
mean1 (x_7=0)	9.54	9.35	9.52	95.4	9.77	9.49	9.77	96.2
mean1 (x_7=1)	11.90	11.34	11.97	94.6	11.89	11.29	11.89	95.0
mean1 (x_7=2)	14.80	14.43	14.79	93.0	15.01	15.46	15.01	87.4

Table 5.12: Simulation results for the compound imputation method Π_{mix}, for the data sets MIXRAWD and MIXMODD, and for a MAR missing data mechanism. The data set MIXMODD is generated according to the underlying statistical model of the method $\tilde{\Pi}_{mix}$ obtained from Π_{mix} by removing the circularities, and MIXRAWD is the data set obtained by sampling from the discretized data set of the Irish Windspeeds.

data	MIXMODD					MIXRAWD				
statistic	$\widehat{Q}$	$\widehat{E}$	$\widehat{Q}_{inc}$	$\widehat{E}[\overline{Q}_m]$	cover	$\widehat{Q}$	$\widehat{E}$	$\widehat{Q}_{inc}$	$\widehat{E}[\overline{Q}_m]$	cover
l-OR (x_2,x_3-1)	1.87	1.26	2.02	94.0		1.44	0.83	1.54	94.2	
l-OR (x_2,x_3-2)	1.68	2.66	1.67	94.8		1.62	2.60	1.54	91.6	
mean4 (x_2=0)	7.89	7.38	7.85	95.4		8.06	7.61	8.03	97.2	
mean4 (x2=1)	11.29	10.15	11.30	96.6		11.56	10.46	11.62	94.8	
mean5 (x2=0)	6.91	6.63	6.87	91.8		7.06	6.82	7.07	97.4	
mean5 (x2=1)	9.92	9.49	9.94	95.8		10.04	9.61	10.04	98.0	
l-OR (x_2,x_6)	1.64	1.63	1.67	96.0		1.37	1.33	1.39	95.2	
l-OR (x_2,x_7-1)	1.65	1.55	1.64	95.0		1.30	1.18	1.27	95.6	
l-OR (x_2,x_7-2)	1.79	2.16	1.82	96.0		1.46	1.81	1.46	93.4	
mean4 (x_3=0)	7.11	6.97	7.15	95.2		7.14	6.97	7.12	92.8	
mean4 (x_3=1)	9.72	8.70	9.84	96.6		9.27	8.65	9.43	97.8	
mean4 (x_3=2)	12.34	11.48	12.18	95.4		12.94	12.36	12.86	95.2	
mean5 (x_3=0)	6.34	6.21	6.34	95.8		6.39	6.27	6.38	90.4	
mean5 (x3=1)	8.49	8.16	8.64	95.6		8.36	8.10	8.49	98.2	
mean5 (x3=2)	10.75	10.43	10.62	94.8		10.85	10.56	10.73	94.0	
l-OR (x_6,x_3-1)	1.73	1.64	1.76	96.2		1.30	1.27	1.37	94.2	
cramc (x_3,x_7)	0.33	0.32	0.32			0.35	0.35	0.35		
correl (x_4,x_5)	0.84	0.83	0.84	96.2		0.81	0.79	0.80	94.2	
mean4 (x_6=0)	7.43	7.23	7.44	95.6		7.58	7.38	7.60	97.0	
mean4 (x_6=0)	12.20	11.87	12.17	96.4		12.17	11.81	12.18	93.4	
mean4 (x_7=0)	7.15	7.06	7.16	93.8		7.46	7.30	7.49	93.8	
mean4 (x_7=0)	9.22	8.80	9.21	95.2		9.41	9.01	9.45	97.0	
mean4 (x_7=0)	12.65	12.35	12.63	95.6		12.47	12.10	12.44	91.4	

Table 5.13: The simulation results for the compound method Π_{mix} and for a MAR missing data mechanism.

data	CATMODD				CATRAWD			
statistic	$\widehat{Q}$	$\widehat{E}\,[\widehat{Q}_{inc}]$	$\widehat{E}\,[\overline{Q}_m]$	cover	$\widehat{Q}$	$\widehat{E}\,[\widehat{Q}_{inc}]$	$\widehat{E}\,[\overline{Q}_m]$	cover
prop (x_1=0)	0.58	0.60	0.57	94.0	0.57	0.61	0.57	96.4
prop (x_1=1)	0.23	0.21	0.23	96.8	0.25	0.22	0.25	93.8
prop (x_1=2)	0.20	0.19	0.20	95.8	0.18	0.17	0.18	96.6
prop (x_2=0)	0.24	0.25	0.24	95.2	0.27	0.30	0.28	96.0
prop (x_2=1)	0.76	0.75	0.76	95.2	0.73	0.70	0.70	96.0
prop (x_5=0)	0.11	0.13	0.13	97.2	0.13	0.14	0.13	96.4
prop (x_5=1)	0.89	0.87	0.87	97.2	0.87	0.86	0.87	96.4
prop (x_6=0)	0.06	0.06	0.06	96.4	0.04	0.05	0.05	99.0
prop (x_6=1)	0.20	0.21	0.20	95.2	0.25	0.28	0.26	98.4
prop (x_6=2)	0.75	0.72	0.74	95.4	0.70	0.67	0.69	97.8
l-OR (x_2,x_1-1)	0.14	0.04	0.17	96.6	2.23	2.06	2.15	95.6
l-OR (x_2,x_1-2)	1.19	1.38	1.18	95.0	2.06	2.31	2.14	96.6
mean3 (x_1=0)	8.00	8.37	8.01	95.4	8.06	8.35	8.05	95.8
mean3 (x_1=1)	6.72	7.04	6.79	94.8	6.69	7.04	6.74	97.6
mean3 (x_1=2)	7.23	7.51	7.17	94.2	7.19	7.52	7.17	96.4
l-OR (x_4,x_1-1)	1.41	1.44	1.36	95.0	1.26	1.31	1.35	93.0
l-OR (x_4,x_1-2)	1.19	1.26	1.18	94.0	1.25	1.52	1.27	96.4
l-OR (x_5,x_1-1)	1.16	1.01	0.98	97.4	1.52	1.52	1.62	95.4
l-OR (x_5,x_1-2)	1.79	2.04	1.66	97.8	2.17	2.40	2.08	95.4
cramc (x_1,x_6)	0.06	0.08	0.10		0.14	0.11	0.15	
mean3 (x_2=0)	8.19	8.59	8.19	95.0	8.22	8.52	8.21	96.6
mean3 (x_2=1)	7.36	7.73	7.36	95.4	7.31	7.70	7.31	94.6
l-OR (x_2,x_4)	0.55	0.84	0.82	90.6	0.28	0.21	0.24	98.0
l-OR (x_2,x_5)	0.11	-0.32	-0.02	93.4	0.79	0.35	0.77	95.8
l-OR (x_2,x_6-1)	-0.04	-0.03	-0.02	97.4	0.39	0.43	0.37	97.4
l-OR (x_2,x_6-2)	-2.46	-2.16	-2.40	98.2	-1.70	-1.41	-1.67	97.0
mean3 (x_5=0)	8.43	8.76	8.32	95.8	8.46	8.75	8.43	95.8
mean3 (x_5=1)	7.45	7.82	7.45	96.8	7.42	7.79	7.43	95.4
l-OR (x_5,x_4)	1.82	1.68	1.07	99.4	0.45	0.65	0.64	93.8
l-OR (x_5,x_6-1)	-1.12	-1.16	-0.94	99.6	1.57	1.48	1.39	96.4
l-OR (x_5,x_6-2)	-3.58	-3.34	-3.19	99.0	-1.35	-1.23	-1.36	95.6

Table 5.14: Simulation results for the compound imputation method Π_{cat}, for the data sets CATMODD and CATRAWD, and for a MAR missing data mechanism. The data set CATMODD is generated according to the underlying statistical model of the imputation method $\tilde{\Pi}_{cat}$ obtained from Π_{cat} by removing the circularities. The data set CATRAWD is the Mammographic Experience data set. The variables are represented by: $x1$ = me, $x2$ = sympd, $x3$ = pb, $x4$ = hist, $x5$ = bse, $x6$ = detcn

Bibliography

[1] Rubin DB, Multiple Imputation after 18+ Years, Journal of the American Statistical Association, Vol. 91, No. 434, 1996:473-489

[2] Rubin DB, Multiple imputation for nonresponse in surveys. Wiley New York, 1987

[3] Hosmer DW, Lemeshow S, Applied Logistic Regression, Wiley and Sons, New York, 1989

[4] Haslett J, Raftery AE, Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource. Appl. Statist., Vol. 38, No.1, 1989:1-50

[5] Siegel S, Castellan NJ, Nonparametric statistics for the behavorial sciences, McCraw-Hill Book Company, second edition. New York, 1988

[6] Stuart A, Keith Ord J, Kendall's Advanced Theory of Statistics. Volume 1 Charles Griffing & CO., London, 1987

[7] Cox DR, Hinkley DV, Theoretical Statistics, Chapman & Hall, New York, 1974

[8] Fisher RA, On the probable error of a correlation coefficient (1921) ...Reprinted in Collected Papers of R.A.. Fisher, Vol. 1, (ed. J.H. Bennet) University of Adelaide Press, Adelaide, South Australia, 1971

[9] Rubin DB, Schenker N, Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, Vol. 81, 366-374, 1986

[10] Rice E, Johnson W, Khare M, Little RJA, Rubin DB, Schafer JL, A simulation study to evaluate the performance of model-based multiple imputation in NCHS health examination

surveys. Proceedings of the Annual Research Conference, 257-266, Bureau of the Census, Washington DC, 1995

[11] Schafer JL, Analysis of Incomplete Multivariate Data. Chapman & Hall, London, 1997

[12] Kennickell AB, Imputation of the 1989 Consumer Finances: Stochastic Relaxation and Multiple Imputation. American Statistical Association. Proceedings of the Section on Survey Research Methods. 1991:1-10

Chapter 6

The implementation of multiple imputation as a missing data engine in HERMES

6.1 Introduction

One advantage of multiple imputation is that it may be regarded as a preprocessing step prior to the application of existing statistical software for complete data. Commercial statistical packages, such as BMDP and SPSS, have been extensively tested for reliability. Despite its advantages, multiple imputation has been applied on a small scale only, due to the following reasons:

1. **Multiple imputation is laborious:** Multiple imputation requires the generation of m imputations for each missing value. Each of the m completed data sets is separately analyzed by the desired statistical method for complete data. Finally, the m results are combined into one result. This requires more work than a simple ad hoc method.

2. **Unfamiliarity with numerical techniques:** To make efficient use of the information available in the incomplete data, the generation of multiple imputations should be based on an adequate statistical model, which requires adequate statistical expertise. Generation of imputations is based on several numerical techniques such as random number

generation, matrix inversion and Choleski decomposition, which are not familiar to the average BMDP or SPSS user.

3. **No standard multiple imputation software available:** Statistical analysis is often performed by an applied researcher who uses standard software. However, no standard multiple imputation software is available in the main statistical packages.

To make multiple imputation available to a larger group of users, it should be implemented in a transparent way, so that users can apply it for the statistical analysis of incomplete data sets on their own without too much trouble with the technical problems. The software package to be developed for this purpose, is called a *Missing Data Engine.* The main goal of this chapter is to give a blueprint for a missing data engine, and to describe what has been achieved so far, and what remains to be done. The design of the missing data engine distinguishes two different types of users: statistically *experienced* users (expert users) and statistically *inexperienced* users. For an expert user, it is relevant to provide a missing data engine with interfaces to select all parameters for an imputation method, to obtain diagnostic information relevant for making optimal selections and to inspect the quality of the imputations. For a statistically inexperienced user, automatic selection of an imputation method is to be preferred.. Another useful feature is the possibility to use data sets containing imputations previously created by an expert user. Related work about missing data engines can be found in [2-4].

A prototype missing data engine has been implemented in the HERMES Medical Workstation environment [5-8]. This is a client-server based environment, developed at the Department of Medical Informatics of the Erasmus University Rotterdam, The Netherlands. The main objectives of HERMES are:

- Network integration of existing databases and applications in an end-user graphical workstation without the need to adapt them;
- User friendly and transparent access to existing databases and applications without the need to know the details about the underlying different data formats and command languages;
- Reusability of newly developed modules;

The main advantages of using HERMES for the implementation of multiple imputation are:

- The possibility to encapsulate existing statistical software for complete data analysis as autonomous entities. Encapsulation of a statistical software package consists of input generation, execution of the corresponding statistical module, and filtering of the output;
- Easy access to data. Data sets from different multiple database systems can be selected.

In section 6.2, requirements for a missing data engine are listed. In these requirements a distinction is made between the two types of users. Section 6.3 describes the translation of the requirements into a conceptual model in which the different execution steps of the multiple imputation cycle are presented in a chronological order. The interactive steps are distinguished from those performed automatically. It is also indicated which steps have been realized and which steps remains to be realized in future. The main principles of the HERMES Medical Workstation are outlined in section 6.4. The client-server architecture, the data and language format for the communication between different modules and the available statistical functionality are described here. Section 6.5 outlines the architecture of a missing data engine in HERMES and in section 6.6, the validation of this missing data engine is described. Finally, section 6.7 discusses the status of the currently realized missing data engine.

6.2 Requirements

This section contains an inventory of requirements for a missing data engine which we consider as important issues.

1. **Interactive specification of statistical analysis:** Specification of statistical analysis comprehends:
   - Opening of a data set;
   - Selection of variables, model and options;
   - Specification of the missing data symbol and idle symbol per variable. The idle symbol indicates a missing data entry for which imputation is undesirable;

- Additional options such as transformation of variables and creation of new variables as functions of existing variables.

The interactive selection can be realized by means of a graphical interface or via a script interface. It is convenient to use the same format in the script as in statistical software packages such as BMDP, SPSS or SAS;

2. **Encapsulation of existing statistical software for the analysis of complete data:** One of the main advantages of multiple imputation is the possibility to perform statistically valid analysis, using existing statistical software for complete data, under less severe assumptions than those required for simple methods such as complete case analysis. In order to take full advantage of commercial statistical packages, such as BMDP or SPSS, such packages should be encapsulated in the missing data engine. To this end, the specified statistical analysis must be translated into the input format of the statistical package and its execution and output filtering should be automated;

3. **Analysis of the missing data mechanism:** The main goal of the analysis of the missing data mechanism is to get a general impression of the seriousness of the missing data problem. Relevant issues are:

   - Investigation of the relationship between the occurrence of missing entries in variables and the observed values of other variables;
   - Statistical tests for the MCAR assumption [9-11]. these are useful when the MAR assumption is plausible;
   - Statistical tests for the MAR assumption in case of external information. External information can be, e.g.:
     - An additional sample among the non-respondents;
     - Assumptions about the sampling distribution such as normality, or symmetry. Performance of statistical tests of the MAR assumption on the basis of distributional assumptions in a statistically sound way, requires a solid basis for these assumptions.

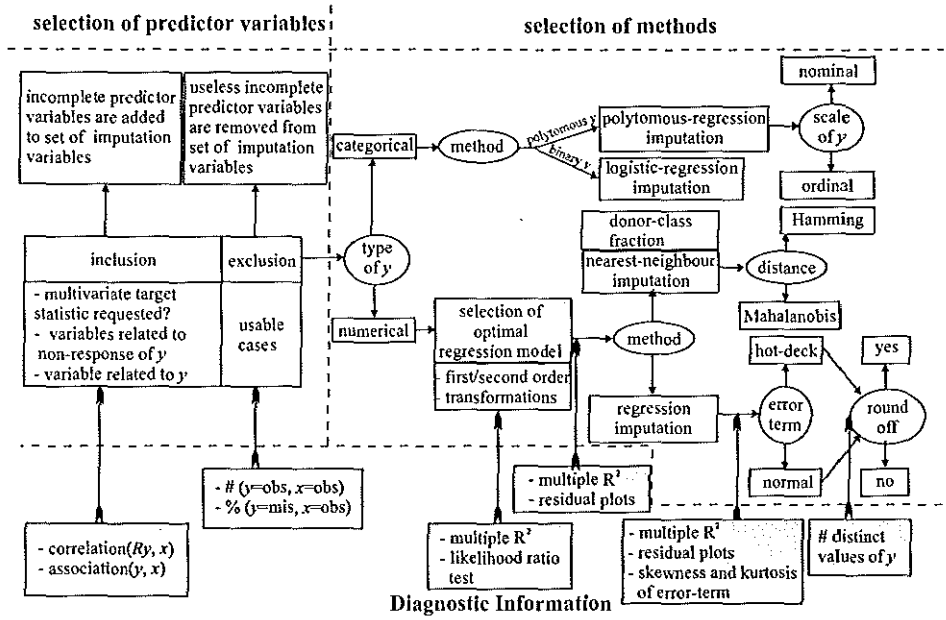


Figure 6-1: Schematic overview of the interactive selection of a multiple imputation procedure.

- Analysis of the missing data pattern. A possibility is to check whether there is a special (e.g., monotonous [1]) missing data pattern;
- Comparison of complete cases with incomplete cases. If, for instance, in survival analysis a certain covariate x is incomplete, it is useful to examine the difference in survival between cases with x observed and with x not observed. A significant difference could indicate bias when complete case analysis is carried out. Comparison of complete- and incomplete cases can be performed outside the missing data engine;
- Calculation of several diagnostic measures to indicate the influence of missing data on the results of the analysis.

4. **Selection of parameters for an imputation method:** Parameters for the imputation methods developed in chapter 4 mainly consist of the set of imputation variables $y_1, \ldots, y_k$, and for each imputation variable y_j a set of predictor variables $\{x_j\}$ and a method π_j. For

linear regression imputation, additional parameters to be specified are those of possible transformations for the imputation variable and for numerical predictor variables, a choice between the normal- or hot-deck error-term variant, and the round off option. Additional parameters for nearest neighbour imputation are the donor class fraction and the distance function.

A distinction is made between automatic and interactive selection. For interactive selection, a graphical user interface or a script interface may be chosen. The automatic selection, and the graphical and script interface are described below:

Automatic selection: Automatic selection is useful for making multiple imputation available to statistically inexperienced users. Starting point for automatic selection is the selection strategy for predictor variables (subsection 4.2.3 of chapter 4), using stepwise regression for the selection of predictor variables x related to the imputation variable y (step 3). The question whether it is sufficient to use linear regression imputation for numerical y and logistic or polytomous regression imputation for binary or polytomous y (consisting of three or more categories) is hard to answer in general. Future research regarding the robustness against deviations from the corresponding statistical model of imputation methods as discussed in chapter 5 may shed light into the matter. To inspect and modify an automatically selected imputation method, it would be convenient to visualize the model by a graphical or script representation;

Graphical user interface: A graphical user interface for the selection of predictor variables and methods is useful for data sets with a relatively small number of imputation variables. It should be implemented in such a way that imputation methods can be selected according to the selection strategy as proposed in subsection 4.2.3 of chapter 4, and that useful diagnostic information on which this selection can be based is available. A schematic overview of an interactive selection process is depicted in Figure 6.1. Selection of predictor variables (left of the Figure) consists of some inclusion and exclusion steps. In particular, if an incomplete predictor variable other than an imputation variable is selected, this variable should be added to the list of

imputation variables. Conversely, if the user deselect a predictor variable, it should be checked if other predictor variables become obsolete and can be removed from the imputation model. A predictor variable x becomes obsolete if it is not related with an imputation variable of interest. To aid the selection of predictor variables, the following diagnostic information is useful for each pair (y, x), were y is an imputation variable and x is a candidate predictor variable of y: (i) the correlation between x and the nonresponse indicator R_y of y, (ii) measures of the association between y and x, (iii) the fraction of cases for which y and x are simultaneously observed, (iv) the fraction of cases with x observed among all cases where y is missing.

The process of interactive selection of methods is depicted in the right of Figure 6.1. Since Discriminant imputation appeared to be inferior to polytomous or logistic regression imputation (see chapter 5), it is not implemented as an option for imputing a categorical variable y. With polytomous regression imputation the user can choose between a nominal or an ordinal scale of y. For categorical imputation variables there is no additional information to be made available. In order to find optimal regression models for numerical imputation variables, the multiple R^2 statistic and the likelihood ratio test of the first order versus the second order regression model could be useful as a diagnostic measure. A choice between nearest neighbour imputation and linear regression imputation can be based on the multiple R^2 statistic and residual plots [12]. When regression imputation is chosen, a further choice between the normal- and the hot-deck error-term can be based on the multiple R^2 statistic, residual plots $e_i = \widehat{y}_i - y_i$ with $\widehat{y}_i$ and y_i the predicted and observed outcome, and the skewness and kurtosis of these residuals. An additional option for regression imputation is to round off imputations (see chapter 4).

Script interface: For data sets with many imputation variables, it is more convenient to use a script interface. A script interface should be equipped with a set of tests to guarantee that a complete and valid imputation procedure is specified.

5. **Specification of logical conditions for imputations:** Logical singular and plural conditions can be specified to prevent imputation of (combinations of) impossible values.

A singular condition is, e.g., the statement that the diastolic blood pressure should be higher than a given minimum value. A plural condition is, e.g., the statement that the systolic blood pressure should be higher than the corresponding diastolic blood pressure.

6. **Inspection of generated imputations:** Once imputations have been generated, a user should have the possibility to inspect their quality. Therefore diagnostic information should be presented such that the user can obtain an answer to the following two questions:

   (a) Does the imputed data fit to the observed data?

   (b) Can the quality of the fit be explained by the underlying missing data mechanism?

   How the questions (a) and (b) can be assessed is explained in section 4.4.3 in chapter 4;

7. **Saving imputations:** If an expert user is given the possibility to generate and save multiple imputations, other users can use these for their own statistical analyses.

8. **Display of measures to assess the contribution of the missing data to the inferential uncertainty of point estimates:** Such measures are usually defined in terms of loss of precision due to missing data. To assess the added value of multiple imputation, it is useful to also define measures for the gain in precision with respect to complete case analysis. For the reflection of these influences on precision, a distinction is made between point estimates and standard errors. Mathematical definitions for measures for point estimates and standard errors are given in chapter 4.

   - **Point estimates:** Loss in precision is reflected by the fraction of information missing due to missing data. The gain in precision with regard to complete case analysis can be defined in a similar way and is developed in chapter 4;
   - **Standard errors:** Loss in precision is reflected by the relative increase in variance r_m and the between imputation variance B_m [1]. It is sensible to display $\sqrt{r_m}$ and $\sqrt{B_m}$ rather than r_m and B_m, since standard errors of point estimates are more commonly used than the corresponding variances. Similar measures for the gain in precision are developed in chapter 4.

9. **Comment generation:** Generation of comment by the missing data engine may be useful if for a certain statistical analysis, the pooled results have a different interpretation than the complete data results. An example is the Analysis of Variance table with standard linear regression, where the pooled test-statistic and corresponding p-value are no longer directly related to the other pooled results of this table.

10. **Sensitivity analysis under MNAR missing data mechanisms:** If the underlying missing data mechanism is MAR, no further specifications for the multiple imputation procedures are required. This is different for an MNAR missing data mechanism, which requires the specification of a probability model of the nonresponse indicator given the hypothetical complete data set. Although it is possible to adjust the generation of imputations to a specified missing data mechanism, it is very hard to specify an adequate model for this mechanism. Unless external knowledge is available, such as an additional sample among the non-respondents, it is impossible to verify such a model empirically or to estimate its parameters. The approach to be followed in this case is sensitivity analysis, the main goal of which is to investigate the robustness of multiple imputation against deviations from the MAR assumption. Sensitivity analysis is performed by repeated application of multiple imputation for several more or less realistic MNAR missing data mechanisms. To apply sensitivity analysis, a missing data engine should be provided with interfaces for the specification of several MNAR missing data mechanisms.

11. **On-line HELP:** Depending on statistical expertise and experience with the missing data engine, different levels of on-line help should be provided to the user.

6.3 A Conceptual model for a missing data engine

A conceptual model for a missing data engine is given in Figure 6.2. In this model, the different steps that are executed during a multiple imputation cycle are presented in a chronological order. A distinction is made between interactive and automatic actions. Steps that have been realized so far are distinguished from steps that remain to be realized in future by their shading. The multiple imputation cycle starts with the collection of data sets represented by the dark-shaded hexagon.

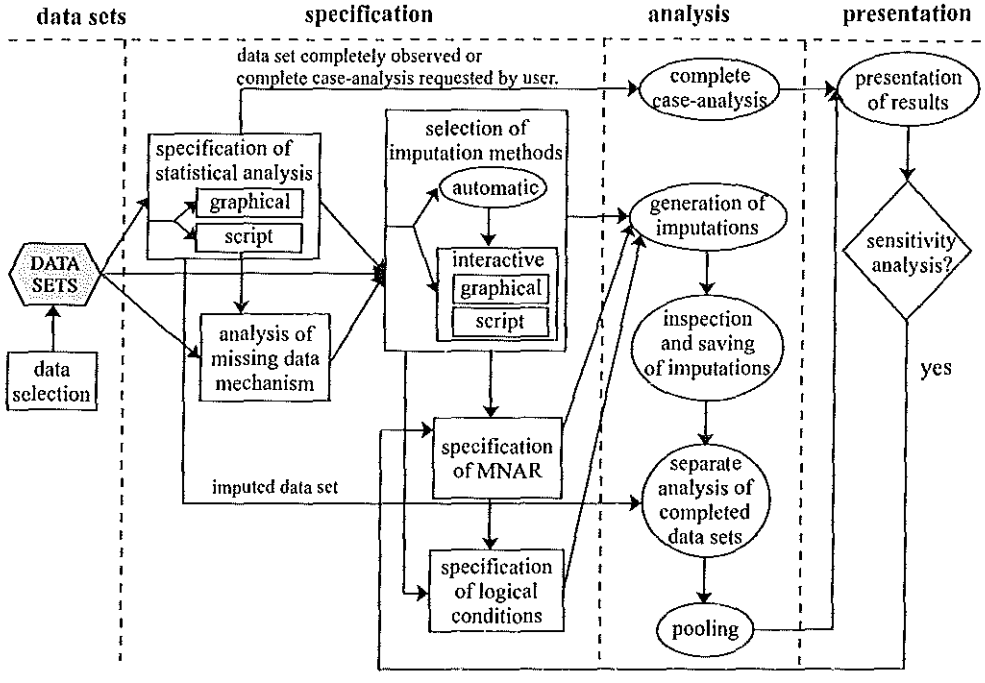


Figure 6-2: A conceptual model for a missing data engine. Rectangular objects represent steps that are executed interactively, and oval objects those that are executed without interaction. Steps that have been realized so far and those that remain to be implemented in future are represented by shaded and unshaded objects, respectively. The collection of available data sets is represented by the dark-shaded hexagon.

The statistical analysis can be specified by means of a graphical or a script interface. The graphical specification has been realized for several statistical modules, the specification via a script file has not been realized thus far.

When the request for statistical analysis has been submitted, the missing data engine checks for missing data entries. If the data set is complete, then the requested analysis is performed. For an incomplete data set, both multiple imputation and listwise deletion can be applied as well. The latter option is usually acceptable if the fraction of incomplete cases is small, e.g., smaller than 10%. An extra option is to create a data set containing the imputations, also called a multiply imputed data set, to be used for future statistical analyses. When statistical

analysis is requested for such a multiply imputed data set, for each of the m completed data sets, a separate statistical analysis is carried out, and the m intermediate results are pooled into one result and presented to the user. An option not yet realized is to analyze the missing data mechanism before the imputation model is specified.

The parameters of an imputation method can be selected either interactively or automatically. The interactive selection by means of a graphical interface has been realized. The automatic selection and interactive selection by means of a script interface have not yet been implemented. Another option to be realized in future is to present an automatically selected imputation method in the graphical or script form for inspection and possible modification.

Options not yet realized are the specification of an MNAR missing data mechanism and logical conditions for the imputations prior to an imputation request. If the missing data mechanism is not specified, imputations are generated on the basis of the MAR assumption. Generated imputations can be inspected and saved. Saved imputed data sets identify the imputation method and its parameters.

To investigate the robustness of the final results against violations of the assumed missing data mechanism (usually MAR), an option to be realized in future is the application of sensitivity analysis. Steps that are executed during sensitivity analysis are shown in the cycle at the right hand side of Figure 6.2. The results as obtained from the various specified MNAR missing data mechanisms can be compared. If under each missing data mechanism the same conclusions are drawn, it may be concluded that the results are robust against violations of the assumed missing data mechanism.

6.4 The HERMES Medical Workstation

6.4.1 Objective

The developments of computer technology of the last decades have resulted in a large variety of powerful computer applications for health care. However, optimal use of these applications requires computer expertise: applications and database systems are generally located on different computers in a network and have different file formats, command languages, and interfaces. Consequently, a user has to exchange data between different computer systems and different file

formats, learn the different command languages, and understand the different user interfaces.

The main goal of HERMES (HEalth care and Research MEdiating System) is to offer clinical users a transparent access to existing database systems and applications.

6.4.2 Integration Issues

In the design of HERMES, the following integration issues have been considered [6]:

- **Shareability:** Applications and databases residing on a server in the network can be used by different users from different computers;
- **Connectivity:** To achieve shareability, the different applications and databases on different computers are connected in a network and the dataflow between them is automated;
- **Modularity:** Modularity is an important issue in the development of large and complex systems. When developing large systems, one divides the system into functionally independent modules. One solution is to implement each module as a library that can be developed and tested independently. However, each modification in a module requires relinking of the whole system. A second solution is an extended notion of modularization in terms of dynamic linking libraries (DLL), which are identified and loaded at application load time. After modification of a module which is available as a DLL, the system needs not to be relinked. In the design of HERMES, the concept of modularization is further extended and modules are implemented as autonomous entities, called servers or facilities;
- **Encapsulation:** One of the most important features in the design of HERMES is the encapsulation of the functionality of existing applications without having to modify them. Encapsulation can be achieved by attaching a wrapping layer around an application, which translates between the input and output format of the application and the HERMES workstation environment;
- **Extensibility:** The HERMES workstation can be dynamically extended with new or changed applications to take full advantage of the latest software developments;
- **User-friendliness:** Integrated applications and databases are presented in a manner which is easy to learn and to handle.

6.4.3 Indirect Client-Server architecture

In the HERMES architecture, a system can be decomposed into independent modules, each of which can act both as a client and as a server. A client sends a request for a task to be executed to a server. The server handles the request and sends its results back to the client. Clients and servers may be located on different computers and may run under different operating systems.

It is clear that the issues of shareability, connectivity and modularity can be easily realized by a client-server architecture. Encapsulation of an application can be realized by building an application server translating between the input and output format of the application and the HERMES environment. User friendliness can be realized by building user interfaces for a client. To realize the issue of extensibility, the communication between clients and servers is indirect. When a client sends a request, a special broker server consults a database to find the most appropriate server for the request and binds the request to this server. The broker database can be edited by a system manager and contains for each identified request the name of the corresponding server with additional information such as the name of the host on which the server is running. In this database, servers and requests can be added or removed and for each request the corresponding server can be changed. As an example, a system manager can extend the HERMES broker base by the request "linear regression" and decide whether linear regression will be carried out by the BMDP server or by the SPSS server.

6.4.4 Message language

To standardize the communication between clients and servers, a special message language, the ISF (Internal Storage Format) language, has been developed. The syntax of this language is given in Figure 6.3 [8]. An ISF message consists of one or more statements. Each statement consists of the name of the application which inserted it, a keyword for its identification and a value. To achieve standardization, several keywords are reserved. For instance, the keyword 'request' is reserved for the broker server to find the most appropriate server. Keywords which are commonly used can be dynamically specified in a special include file.

The four basic types of values are 'simple', 'list', 'struct' and 'raw'. The type 'simple', represents values such as integer, char, string and file, which are atomic values. Lists are composite values consisting of a series of values of the same type. Structures are composite values

```
request     :=   message+

message     :=   "begin" statement+ "end"

statement   :=   creator "." keyword[type] "=" value [";"]

creator     :=   -- a name of a proces --

keyword     :=   -- a string --

type        :=   "[" "%" [size] [construct] typemark "]"

typemark    :=   "d"   /* integer*/
                 "f"   /* float*/
                 "c"   /* char */
                 "s"   /* string */
                 "q"   /* quoted string */
                 "b"   /* boolean */
                 "w"   /* file */
                 "r"   /* raw */

construct   :=   "n"   /* list *
                 "m"   /* structure */

value       :=   list | struct | simple | raw

list        :=   "(" value ("," value)* ")"

struct      :=   "{" statement+ "}"

simple      :=   -- a string --    /* for integer until boolean */
                 -- raw bytes --  /* for file and raw types */
```

Example

```
begin
  client.request[%s] =          "select data"
  client.user[%9s] =            mulligen
  client.host[%s] =             mwiv
  client.variables[%ns] =       ( age, sex, birthdate, diagnosis)
  broker.service[%s] =          ingres
  broker.servicehost[%s] =      mwi
  rsmgr.language[%s] =          english
end
```

Figure 6-3: Definition of the message language syntax and an example message. A '+' after a word in the language syntax means that the element may be repeated. Words between square brackets indicate an option. Words between double quotes are literally included in messages.

consisting of one or more statements. The values in a list may be structures and a structure may contain one or more lists. Thus the ISF language is suitable for storing information hierarchically. To include binary data and data which is not structured according to the ISF syntax, such as images and signals, values can be represented by the type 'raw'.

In the HERMES environment, messages can be internally represented and stored into a file. Standard procedures have been developed for reading, saving, composing and sending messages. These procedures are available as libraries.

6.4.5 Data access

Data sets are also represented in ISF format and will be called isf-data sets. Additional information, such as storage format, declaration of the variables, and the missing data symbol which is the same for each variable is contained in isf-data sets. The HERMES data dictionary service contains information about types and codes of the variables as contained in a data model that is created and edited with an interactive graphical tool belonging to the data model server. In a data model, information about attributes and the organization of these attributes in a database is stored. Each data set for which a data model has been defined, contains a key referring to this model. All information about a set of variables, contained in the data dictionary, can be obtained within an application, by sending their attribute addresses and model key to the data model server.

Other services for data access are a file manager and a data selection server. The file manager can be used for interactively opening and saving data sets and can be easily incorporated within applications. With the data selection server, data sets from one or more database systems can be interactively selected [7]. After the specification of the selection has been completed, the data selection server will forward a request containing an SQL$^+$ query (SQL$^+$ is an extension of SQL for multiple database systems) to the multi-database server. This server decomposes the SQL$^+$ query into the individual SQL queries for the various database management systems involved. Subsequently, these queries are passed to their corresponding database servers and the resulting data sets are joined into one data set. Although joining of data sets seems straightforward, in practice many problems, such as incompleteness, may be encountered.

6.4.6 Complete data analysis

An overview of the functionality for complete data analysis available within HERMES is given in Figure 6.4. The following statistical modules have been realized within HERMES:

- Simple descriptive statistics, including mean, standard deviation, standard error of the mean, coefficient of variation, minimum, maximum and range;

- Standard linear and logistic regression;

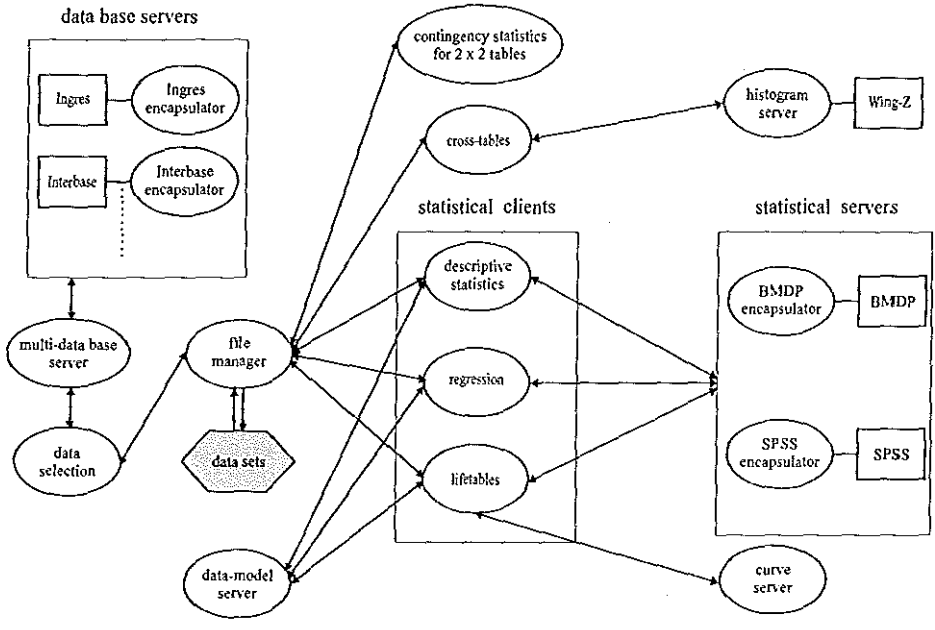


Figure 6-4: An overview of the functionally for complete data analysis within HERMES. The dark shaded hexagon represents all available data sets. An oval object represents a module which has been implemented within HERMES and a rectangular object represents a commercial software package. Client-server interaction between two modules is represented by double arrows, the arrows pointing to and from the hexagon represent saving and opening a data set. The modules which have been realized and the modules which may be realized in future are represented by the shaded and unshaded objects, respectively.

- Survival curves according to the Kaplan-Meier method, or actuarial life table method. Curves can be estimated for several classes within a data set. The curves are visualized by a specific curve server.

- $(k \times r)$ cross-tables. A cross-table can be visualized as a histogram by the histogram server. As an additional option, the histogram server can export into the spreadsheet WingZ format.

- Estimation of rate ratios, rate differences, odds-ratios and their corresponding p-values (one- and two sided) and confidence intervals from (2×2) contingency tables are implemented as stand-alone modules. For $(k \times 2)$ tables, with $k > 2$, the statistics mentioned above

can be presented for several (2×2) tables including a given reference row. The input for this module is generated by the cross-tables module. In Figure 6.4, this module is not connected to the data model server since no data dictionary has been defined for cross-tables.

The calculations for the descriptive statistics, life tables, linear and logistic regression are carried out by the BMDP [13] modules 1D, 1L, 1R and LR, respectively. The modules for cross-tables and (2×2) contingency tables are implemented standalone and act as servers. Graphical interfaces for the interactive specification of the statistical analysis for the BMDP modules are implemented as statistical clients and the encapsulation of the corresponding BMDP modules has been realized by statistical servers. Encapsulation of BMDP by a statistical server is outlined below and schematically represented in Figure 6.5. It has been realized in the following three steps:

1. Conversion of the statistical analysis request to BMDP input. To this end, the data set is converted into an appropriate format and the script is generated;
2. Execution of BMDP with converted data and generated script as input;
3. Conversion of BMDP output to ISF format.

The BMDP script generation is driven by a special table which is parsed together with the request. Tables have been written for the BMDP modules 1D (simple descriptive statistics), 1L (lifetables and survival curves), 1R (linear regression), and LR (logistic regression). Generation of script for other modules can be easily realized by defining the appropriate tables. The table to be used for the generation of script is identified in the request.

The organization of the knowledge for the generation of script as stored in the table is represented in Figure 6.5. A module in BMDP consists of several paragraphs, each consisting of a series of commands. The knowledge is stored for each command separately and these knowledge units are grouped into knowledge chunks for each paragraph. Generally, knowledge about a command is represented by its name, its location and data structure in the request, and the syntax in the BMDP-script. Other statistical packages, such as SPSS may be implemented in future.

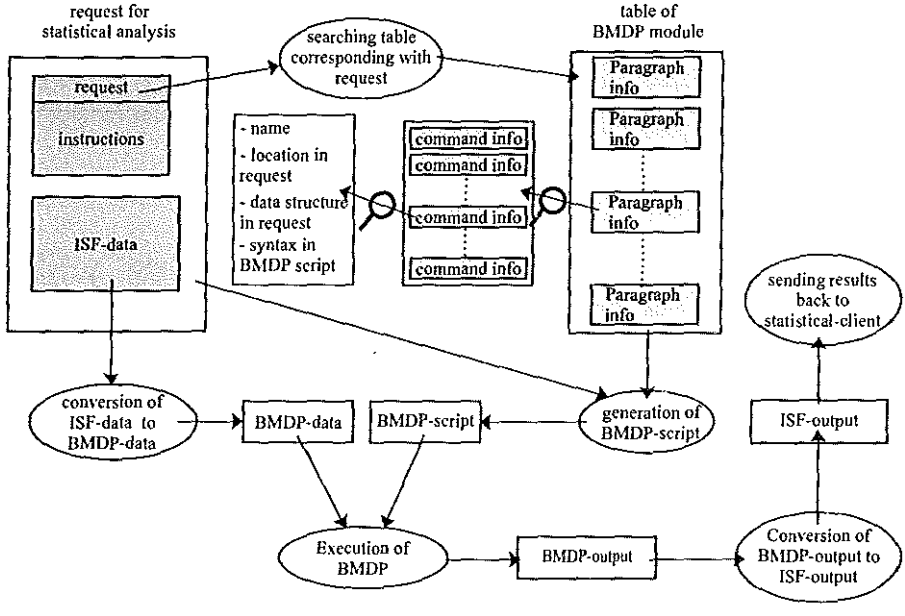


Figure 6-5: A schematic overview of the encapsulation of BMDP. The rectangular objects represent files and the oval objects represent processes. An arrow pointing from a rectangular object to an oval object means that the corresponding file is input for the corresponding process and an arrow pointing from an oval to a rectangular object means that the corresponding file is output of the corresponding process. An arrow with a magnifying glass means 'consists of'.

6.5 The missing data engine in HERMES

The realization of the missing data engine according to its conceptual model, is illustrated in Figure 6.6; the dashed double arrow denotes client-server communication, that has not been realized so far. Specification of missing data symbols and idle symbols which may be different for different variables has been implemented as a separate module. After these symbols have been specified for an isf-data set, the corresponding data entries are replaced by the symbols specified in this data set.

The core of the missing data engine has been implemented as three functionally independent modules: the missing data server, the imputation server and the pooling server. The missing data server coordinates the multiple imputation cycle and mediates between the statistical

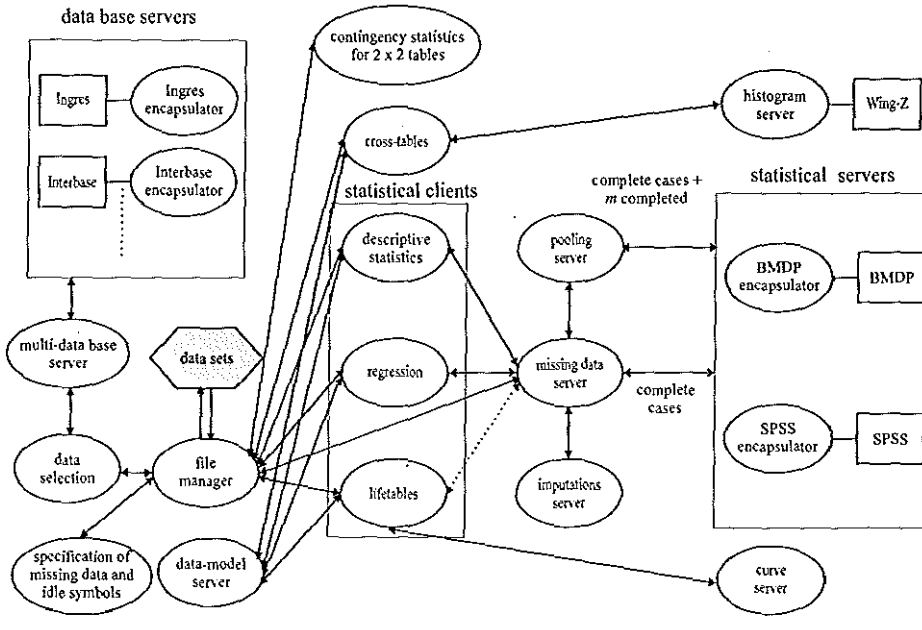


Figure 6-6: Architecture of the missing data engine in HERMES.

clients and the statistical servers. It supports interactive selection of an imputation method, specification of logical conditions and the inspection of imputations. Automatic selection of an imputation method, analysis of the missing data mechanism, and performance of sensitivity analysis may also be implemented in this server, but these options have not yet been realized. The missing data server is also directly connected to the statistical servers to perform complete data analysis or listwise deletion.

The imputation server generates m imputations for each missing data entry which is not idle. The request from the missing data server contains the parameters of the method for the imputations to be generated. Generated imputations are sent back as m separate files to the missing data server.

The pooling server generates m completed data sets and subsequently forwards each of these to the statistical server as requested by the statistical client. For initialization, the incomplete data set is forwarded as well. The m intermediate results are then combined into one final

result which is returned together with several diagnostic quantities to assess the contribution of the missing data to the inferential uncertainty.

For pooling of results, seven classes of statistics are considered. Apart from the four basic classes: - point-estimates, standard errors, confidence intervals, and p-values -, additional classes included are univariate and multivariate test-statistics and auxiliary statistics. Univariate and multivariate test-statistics are associated with p-values belonging to a hypothesis involving a univariate and a multivariate parameter of interest, respectively. The class of auxiliary statistics consists of the statistics, the function of which is merely to give additional information, rather than representing statistical inference about an estimand. Examples of auxiliary statistics are the number of distinct covariate patterns in logistic regression and the number of non-survivors, and the number of subjects remaining at risk as presented in survival analysis by means of the Kaplan-Meier method. Auxiliary statistics are not pooled from the m completed data results, but directly calculated from the incomplete data set.

An additional complexity is that for some of the pooling classes, the m completed data results of other statistics must be involved as well. E.g., when pooling standard errors, the m completed data results of the corresponding point-estimates are to be taken into account. To deal with this complexity, pooling of the results has been indirectly implemented via a table. A syntax for the output of a statistical server has been developed and the knowledge in the pooling table is structured accordingly. Pooling of results has been realized for descriptive statistics, linear and logistic regression. In future, other statistical modules may be incorporated as well by extending the table.

6.5.1 Graphical interface for the interactive selection of an imputation method

The graphical interface for the interactive selection is given in Figure 6.7. The names 'PEAKDBP', 'PEAKSBP', 'RESTDBP' and 'RESTSBP' refer to the variables names 'pd', 'ps', 'rd' and 'rs' in chapter 2. The substring 'HR' refers to the heart rate and the substring 'WMSC' refers to the wall motion score of the heart. A wall motion score of 1 is considered normal, a wall motion score much larger than 1 is considered seriously abnormal. Similar to SBP and DBP, for HR and WMSC, a distinction between rest and maximal effort is made. The initial imputation variables (here, PEAKSBP and WMSCREST) are presented in the

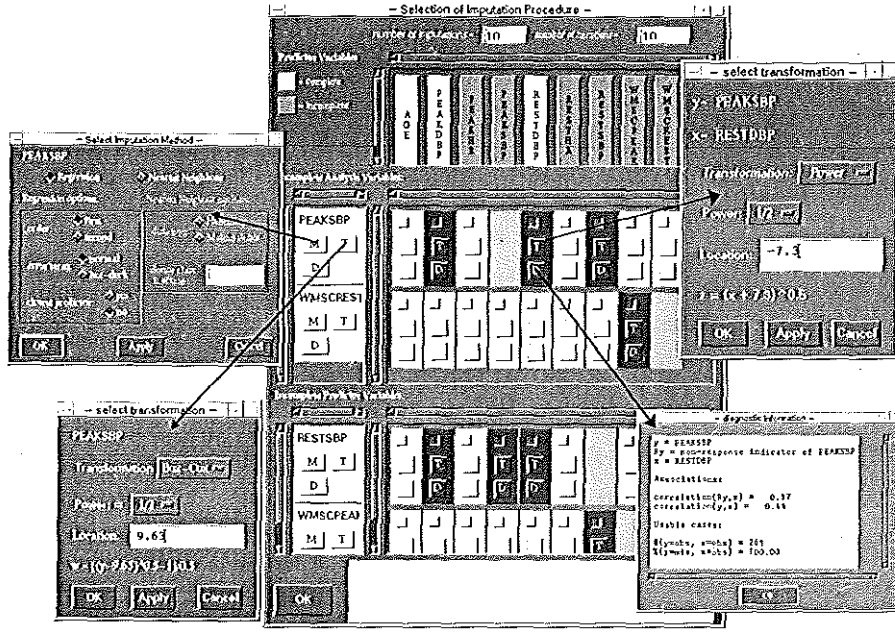


Figure 6-7: The graphical user interface for the interactive specification of an imputation method.

middle left scrollable window of the main interface. If a statistical analysis is requested, the initial imputation variables are the incomplete variables involved in this analysis. Otherwise, the initial imputation variables are chosen by the user. Predictor variables for these incomplete analysis-variables can be selected and deselected in the middle right scrollable window. The dark coloured blocks indicate which predictor variables are selected for each of the incomplete analysis variables. 'PEAKDBP', 'RESTDBP' and 'RESTSBP' are selected as predictor variables for 'PEAKSBP' and 'WMSCPEAK' is selected as a predictor variable for 'WMSCREST'.The variables presented in the upper scrollable window are all candidate predictor variables.

In the user interface, the complete and incomplete candidate predictor variables are presented in different colours. In the example of Figure 6.7, only 'AGE', 'PEAKDBP' and 'RESTDBP' are complete as indicated by the light colour. Additional information about a candidate predictor variable, such as the type, codes and number of missing data entries, can be obtained

by pressing the corresponding button.

All selected incomplete predictor variables other than initial imputation variables are presented in the lower left scrollable window. For these incomplete variables, predictor variables can be selected in the lower right scrollable window. If for an imputation variable an incomplete predictor variable other than an imputation variable is selected, this incomplete predictor variable is added to the two lower scrollable windows. If an incomplete predictor variable is deselected, it is checked whether incomplete predictor variables become obsolete. A predictor variable x is obsolete when it is not related to an initial imputation variable, i.e., there exists no sequence of variables $z_1, \ldots, z_s$ with z_s an initial imputation variable, such that x is a predictor for z_1, z_i is a predictor variable for z_{i+1}, for $i = 1, \ldots, s-1$. The incomplete variables which become obsolete are then removed from the list of incomplete predictor variables in the two lower scrollable windows. If for instance, in Figure 6.7, the variable RESTSBP is deselected as predictor variable for the variable PEAKSBP, this predictor variable becomes obsolete and will be removed from the list of incomplete predictor variables.

To reduce the complexity of the user interface, the selection of imputation methods and transformations, and the display of available information is performed in separate windows. An imputation method for an imputation variable is selected by pressing the 'M' button of the corresponding variable. In Figure 6.7, the interface for a numerical imputation variable is depicted (left, upper window). As an example, first order regression imputation with a normally distributed error term and without closest predictor has been selected. A different interface has been developed for categorical imputation variables (not shown). Transformations can be selected by pressing the appropriate 'T' button. Different interfaces have been developed for the transformations of imputation variables and of predictor variables, respectively. A transformation for an imputation variable is selected from the family of Box-Cox transformations (see lower left window) and a transformation for a predictor variable is selected from the family of power transformations (see top right window). In both interfaces, the currently selected transformations are visualized. Additional diagnostic information can be obtained by pressing the appropriate 'D' button. In Figure 6.7, information about the predictor variable RESTDBP for the imputation variable PEAKSBP is presented (see bottom right window). Information for different selections can be compared by subsequently pressing the 'Apply' button. Each

time the 'Apply' button is pressed, the information is adjusted to the current selection. On-line information is displayed in the field at the bottom of the main interface. The current selections are displayed in this field when the mouse pointer is moved to the appropriate 'M' or 'T' button.

6.6 Validation of the missing data engine in HERMES

The missing data engine consists of the three functionally independent modules: the missing data server, the imputation server, and the pooling server. Therefore, validation of the missing data engine consists of validating each of these modules separately and verifying whether the exchange of messages between the different modules is correct. The validation of each module is outlined below:

6.6.1 Missing data server

For the missing data server the following is checked::

- Interactive selection of an imputation method. It should be checked whether the request for the multiple imputation server contains all parameters for the imputation method to be selected. Parameters are the imputation variables and for each of these, the predictor variable(s) and imputation method for each;
- Correctness of displayed available information for selecting an imputation method.

6.6.2 Imputation server

The validation of the imputation server consists of two steps. First, each component of the imputation server is validated separately. Second, the multiple imputation results are validated. These two steps are described below:

1. **Validation of separate components:** In this step a distinction is made between reading of the request and generation of the imputations.

   Reading of request: The following is considered:

- Reading and storing of the incomplete data set and the parameters of the imputation method;
- Correspondence of the created data structure for storage of the generated imputations with the missing data patterns in the incomplete data set.

Generation of the imputations: The m imputations are sequentially generated by m Gibbs sampling runs. Each run consists of a certain number of iterations specified by the user. During an iteration of the i-th run, the i-th imputations are sequentially generated for the imputation variables and the completed data set is sequentially updated with these values for the imputation variable y. In this iteration, imputations for an imputation variable y are generated by an elementary imputation method with as input parameters the vector Y_{obs} of observed values of y and the matrices X_{mis} and X_{obs}. The matrices X_{mis} and X_{obs} are submatrices of the currently completed data X for the predictor variables of y and consist of the rows of X corresponding to the missing and observed values of y, respectively. Consequently the following procedures are checked:

- All numerical procedures and random number generators used in the elementary imputation methods;
- Generation of the input parameters Y_{obs}, X_{obs} and X_{mis} for elementary imputation methods during a Gibbs sampling run;
- Updating the completed data set with the imputations for an imputation variable y generated by an elementary imputation method.

2. **Validation of results:** The results of the imputation server are compared with the results of the simulation program in SAS/IML described in chapter 5 for the same incomplete data set, target statistics and imputation methods. A total of 10 imputations is generated. The imputation methods to be investigated are the compound imputation methods Π_{num} for exclusively numerical imputation variables, Π_{mix} for imputation variables of mixed type, and Π_{cat} for exclusively categorical imputation variables, as described in subsection 5.2.6 of chapter 5. The data set used for each method is generated in the same way as the raw data set used in chapter 5 for that method. The results for Π_{num} are displayed in

Table 6.1 and the results for Π_{mix} and Π_{cat} are given in Table 6.2. Each row of these tables corresponds with the results of a certain imputation method for a certain target statistic. The terms Q_1 and Q_3 represent the first (25%) and the third (75%) quartile. In the second and third column of the two Tables, the point estimates $\hat{Q}$ obtained from the complete data set and the point estimates $\hat{Q}_{\text{inc}}$ obtained by listwise deletion from the incomplete data set are given. The pooled point estimates $\overline{Q}_m^{(\text{S})}$ obtained by the simulation program in SAS/IML and these point estimates $\overline{Q}_m^{(\text{H})}$ obtained by the imputation server of the missing data engine in HERMES are given in the fourth and fifth column, respectively.

Table 6.1 shows that for Π_{num}, the differences between $\overline{Q}_m^{(\text{S})}$ and $\overline{Q}_m^{(\text{H})}$ are quite small. For the univariate target statistics these differences range from 0.00 (mean of x_3) to 0.17 (third quartile of x_1), and for the correlations these differences are at most 0.01. The same conclusions can be drawn for Π_{mix} and Π_{cat} from Table 6.2. For Π_{mix} and Π_{cat}, the differences between $\overline{Q}_m^{(\text{S})}$ and $\overline{Q}_m^{(\text{H})}$ for the proportions are at most 0.02. For Π_{mix}, these differences range from 0.01 (first quartile of x_4) to 0.12 (third quartile of x_3) for the univariate target statistics obtained for the numerical variables x_1 and x_4, and these differences are at most 0.02 for the three correlation coefficients.

The results described above show that two implementations of the same imputation algorithm on two different hardware configurations and different operating systems give approximately the same results. Although the validation of the imputation server requires a more extended study, it can be concluded that the imputation server is reliable since the SAS/IML implementations of Π_{num}, Π_{mix} and Π_{cat} are validated in chapter 5.

statistic	$\widehat{Q}$	$\widehat{Q}_{inc}$	SAS/IML $\overline{Q}_m^{(S)}$	HERMES $\overline{Q}_m^{(H)}$
mean ($x1$)	12.37	11.34	12.48	12.55
Q1 ($x1$)	8.00	6.92	7.93	8.01
med ($x1$)	12.00	10.92	12.26	12.32
Q3 ($x1$)	16.29	14.67	16.12	16.29
mean ($x2$)	11.86	11.05	11.88	11.90
Q1 ($x2$)	8.00	7.25	7.77	7.81
med ($x2$)	10.92	10.13	11.16	11.17
Q3 ($x2$)	15.16	14.12	15.31	15.45
mean ($x3$)	10.57	9.56	10.39	10.39
Q1 ($x3$)	7.04	6.25	7.01	7.02
med ($x3$)	10.08	9.08	10.03	10.08
Q3 ($x3$)	13.59	12.54	13.39	13.47
mean ($x4$)	9.81	9.02	9.83	9.82
Q1 ($x4$)	6.00	5.41	6.08	6.08
med ($x4$)	9.13	8.00	9.33	9.23
Q3 ($x4$)	13.21	11.75	12.98	13.01
correl ($x1$,$x2$)	0.74	0.69	0.78	0.79
correl ($x1$,$x3$)	0.84	0.82	0.84	0.84
correl ($x1$,$x4$)	0.76	0.46	0.80	0.79
correl ($x1$,$x5$)	0.78	0.75	0.78	0.78
correl ($x1$,$x6$)	0.69	0.66	0.71	0.71
correl ($x2$,$x3$)	0.59	0.27	0.64	0.64
correl ($x2$,$x4$)	0.71	0.69	0.71	0.72
correl ($x2$,$x5$)	0.64	0.64	0.65	0.65
correl ($x2$,$x6$)	0.58	0.58	0.60	0.60
correl ($x3$,$x4$)	0.80	0.78	0.78	0.78
correl ($x3$,$x5$)	0.83	0.82	0.82	0.82
correl ($x3$,$x6$)	0.72	0.72	0.72	0.72
correl ($x4$,$x5$)	0.85	0.88	0.87	0.87
correl ($x4$,$x6$)	0.81	0.83	0.82	0.82

Table 6.1: The results for a compound method for exclusively numerical variables and a MAR missing data mechanism.

data set	statistic	$\hat{Q}$	$\hat{Q}_{inc}$	SAS/IML $\bar{Q}_m^{(S)}$	HERMES $\bar{Q}_m^{(H)}$
raw mixed	mean ($x1$)	12.27	11.65	12.28	12.32
data	Q1 ($x1$)	9.00	8.45	9.09	9.16
	med ($x1$)	12.06	11.29	12.11	12.18
	Q3 ($x1$)	14.82	14.17	14.97	15.09
	prop ($x2$=0)	0.49	0.57	0.50	0.50
	prop ($x2$=1)	0.51	0.43	0.50	0.50
	prop ($x3$=0)	0.32	0.37	0.32	0.32
	prop ($x3$=1)	0.34	0.28	0.33	0.35
	prop ($x3$=2)	0.34	0.28	0.33	0.33
	mean ($x4$)	9.76	9.06	9.80	9.77
	Q1 ($x4$)	6.92	6.42	6.97	6.96
	med ($x4$)	9.20	8.58	9.32	9.32
	Q3 ($x4$)	12.03	11.08	12.37	12.31
	correl ($x1$,$x4$)	0.59	0.34	0.66	0.64
	correl ($x1$,$x5$)	0.63	0.65	0.66	0.66
	correl ($x4$,$x5$)	0.79	0.78	0.79	0.79
raw categorical	prop ($x1$=0)	0.57	0.59	0.56	0.55
data	prop ($x1$=1)	0.25	0.24	0.28	0.27
	prop ($x1$=2)	0.18	0.17	0.16	0.18
	prop ($x2$=0)	0.27	0.31	0.29	0.29
	prop ($x2$=1)	0.73	0.69	0.71	0.71
	prop ($x5$=0)	0.13	0.16	0.15	0.16
	prop ($x5$=1)	0.87	0.84	0.85	0.84
	prop ($x6$=0)	0.04	0.05	0.05	0.05
	prop ($x6$=1)	0.25	0.27	0.25	0.26
	prop ($x6$=2)	0.70	0.68	0.69	0.69

Table 6.2: The results for a compound imputation method for variables of mixed type and a compound method for exclusively categorical variables. The two incomplete data sets for these two methods are artificially generated by a MAR missing data mechanism

6.6.3 Pooling server

For the validation of the pooling server, the following issues have been considered:

- Parsing of the complete data analysis results from the statistical server, using the pooling table;
- Reading, storing and pooling of m completed data results;
- Comparison of the pooled p-values with those obtained from the corresponding m completed data sets. It is plausible that pooled p-values are larger than or equal to the average of the m completed data p-values, since in the calculation of the latter, the imputed values are assumed to fixed values, while in the pooled p-values the extra inferential uncertainty due to missing data is incorporated;
- For likelihood-ratio p-values, the validation of the procedures used for the calculation of the log-likelihoods of the pooled point-estimates for the parameters of the two models to be compared. These log-likelihoods are not available from standard statistical software for complete data;
- Validation of the numerical procedures for determining the cumulative probability distribution function and its inverse of the standard normal, chi-square, student t and F distribution, to be used for pooling the m completed data results. These procedures are validated by comparing their results with existing tables for the quantiles of these distributions [14].

6.7 Discussion

This chapter describes the design of a missing data engine and its realization in the HERMES Medical Workstation. Thus far the following has been realized: a prototype of a missing data engine, including the interactive selection of imputation methods, and the application of multiple imputation in combination with descriptive statistics, linear regression and logistic regression. Important issues that still await implementation in future are sensitivity analysis, i.e.,

examining the robustness against deviations of the MAR assumption, and automatic selection of an imputation method.

An advantage of HERMES is, that statistical software packages can be encapsulated as autonomous entities. In this way, an underlying software package can be substituted by another package without having to modify the software of the missing data engine. This advantage is, however, relative, since changes in the output format of a new version of a statistical software package requires that the software for filtering this output must be adapted. Another advantage of HERMES is its client-server architecture, which makes it possible to subdivide the missing data engine into functionally independent modules. Each module can be modified and recompiled without having to recompile the entire missing data engine.

Despite these advantages of HERMES, encapsulation of existing statistical software is laborious. This is due to the large number of potential different statistical modules to be encapsulated. Script generation can be relatively easily implemented by a driver table. Conversion of output of a statistical software package to ISF-format is, however, much harder.

Consequently, to make multiple imputation available to a larger group of users, it is necessary that in future existing statistical software packages will be equipped with an option for multiple imputation. The missing data engine presented here may serve as a prototype for this.

Bibliography

[1] Rubin DB, Multiple imputation for nonresponse in surveys. Wiley New York, 1987

[2] Brand J, Buuren S, Van Mullingen EM, Timmers T, Gelsema E, Multiple Imputation as a Missing Data Machine . Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care, 194:303-306

[3] Van Buuren S, Van Mulligen EM, Brand J, Routine multiple imputation in statistical databases. Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management, Charlottesville, Virginia, Los Alamitos, California: IEEE Computer Society Press, 1994: 74-78

[4] Van Buuren S, Van Mulligen EM, Brand J, Omgaan met ontbrekende gegevens in statistische databases: Multiple imputatie in HERMES, Kwantitatieve methoden, 1995 nummer 50: 5-13

[5] Van Mulligen EM, Timmers T, Van Bemmel JH, Accessors As a Framework For Integration of Data and Software. chapter 5 of PhD thesis, Department of Medical Informatics, Erasmus University Rotterdam, The Netherlands, 1993

[6] Van Mulligen EM, Timmers T, Van Bemmel JH, New Perspectives on an Integrated Medical Workstatition. chapter 6 of PhD thesis, Department of Medical Informatics, Erasmus University Rotterdam, The Netherlands, 1993

[7] Van Mulligen EM, Timmers T, Van Bemmel JH, A New Architecture for Integration of Heterogeneous Software Components. Methods of Information in Medicine. 1993: 32-292-301

[8] Van Mulligen EM, Timmers T, Brand JPL, Cornet R, Van den Heuvel F, Kalshoven K, Van Bemmel JH, HERMES a health care workstation integration archictecture. International Journal of Bio-Medical Computing, 1994:24: 267-275

[9] Diggle PJ, Testing for Random Dropouts in Repeated Measurement Data. Biometrics, vol. 45, 1989:1198-1202

[10] Little RJA, A Test of Missing Completely At Random. Technometrics, vol. 30, no.2, 1988:205-214

[11] Simonoff JS, Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression. Technometrics, vol.30, no.2, 1998:205-214

[12] Draper NR, Smith H, Applied regression analysis, Second edition, Wiley & Sons, 1981

[13] Dixon WJ, Engelman L, Hill MA, Jennrich RI, BMDP Statistical Software Manual, Vol.1 and 2, University of California Press, London, 1988

[14] Stuart A, Ord JK, Kendals advanced theory of statistics, Vol. 1 of fifth edition: Distribution theory, Charles Griffin & Co, London 1987

Chapter 7

Application of multiple imputation in the Leiden Safety Observed Study

7.1 Introduction

The occurrence of accidents among elderly people (55+) is a serious problem in The Netherlands. The following facts [1] give an impression of the extent of this problem:

- The annual mortality rates associated with accidents among elderly people (55+) incurred at home and during leisure activities are about 50, and in traffic about 15 per 100,000 elderly people;
- The annual hospitalization rates of elderly people due to accidents incurred at home and during leisure activitiesare about 907, and in traffic about 139 per 100,000 elderly people.

It is expected that in the next years, without extra preventive effort, the impact of this problem will increase due to ageing of the Dutch population. The ultimate goal of the Leiden Safety Observed Study [2] is to formulate targeted prevention measures. There is epidemiologic evidence that several health aspects are important risk factors for accidents among the elderly [3]. One of the objectives of the Leiden Safety Observed Study (Veiligheid in de Peiling) is to investigate health aspects as possible risk factors for different types of accidents among the elderly.

Due to missing data and the large number of candidate risk factors, up to now, health aspects have been investigated only by means of univariate analyses. In the present study, application of listwise deletion prior to multivariate analysis would result in a 55% reduction of cases from 907 to 405. This implies that, because of the low frequencies of some categories, many cannot be included. Moreover, the nonresponse appears to be dependent on age, gender, education and several disabilities which are candidate risk factors, so that the reduction of cases through listwise deletion raises questions about the validity and reliability of the results.

For an investigation of the relationships between the different types of accidents and the various candidate risk factors, multivariate analyses are necessary since only in that way can mutual relationships between the candidate risk factors be taken into account. When each health aspect is analyzed separately, some of such aspects may wrongly emerge or submerge as important risk factors.

This chapter describes the application of multiple imputation prior to a multivariate statistical analysis in the Leiden Safety Observed Study. One particular problem treated in this analysis is the application of multiple imputation to stepwise regression. The purpose of this chapter is twofold:

1. Description and illustration of the methodology;

2. Examination of the added value of multiple imputation with respect to listwise deletion.

Conclusions and recommendations for future preventive measures emerging from this analysis are not considered in this chapter but will be described in full detail elsewhere. The added value of multiple imputation with respect to listwise deletion is examined by comparing the statistical models found by multiple imputation with those found by listwise deletion.

Section 7.2 describes the data collection, the variables involved in the multivariate analysis and the methodology. In section 7.3, the methodology is illustrated by a few examples, the added value of multiple imputation is examined, and for some imputation variables, the quality of the imputations is inspected according to the criteria as proposed in requirement 6 of section 6.2 of chapter 6.

7.2 Methods

7.2.1 Data description

A random sample of 3500 elderly people (65-85 years) was drawn, stratified by age, living independently in Leiden, The Netherlands. The highest age category was overrepresented in this sample, because in this category a relatively high nonresponse rate was expected. In November 1993, the 3500 participants were contacted to fill out a questionnaire (pre-measurement) concerning the following variables:

- the demographic variables gender, age, social-economic status, and living situation;
- subjective health;
- health variables concerning disabilities in hearing/vision/locomotion from the OESO disability questionaire [4];
- health variables concerning pain/trouble per body part;
- other health variables concerning dizziness, loss of physical strength, and tiredness;
- variables concerning attitude and behaviour with respect to prevention.

From the sample of 3500 persons, 1055 persons started participating in the follow-up in which, from March 1994 through May 1995, accident data were monthly collected telephonically. The 907 people who completed the entire follow-up were contacted in June 1995 for a second questionnaire, the post-measurement. A total of 775 persons responded to this second questionnaire in which the pre-measurements were repeated and which contained additional questions about the presence of 32 chronic diseases [5] and the use of medication for such diseases. It appears that among the nonrespondents to the post-measurement, there was a relatively large number of persons with primary education only and a relative large number of persons with one or more serious disabilities [2]. The variables involved in the multivariate analysis are described below.

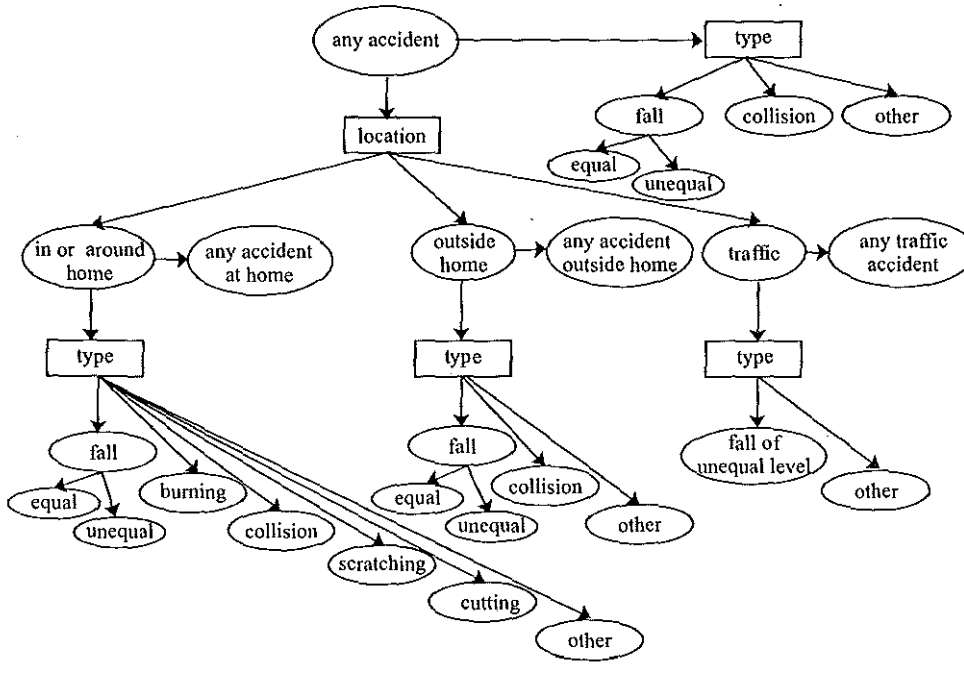


Figure 7-1: Combinations of type and location of accidents considered in the study.

Definition of variables

A distinction is made between accident variables and variables measured in the pre- and post-measurement.

accident variables An accident variable y is defined such that $y = 1$ if the corresponding accident occurred during the follow-up and $y = 0$ if it did not. Accidents are qualified as follows:

- **type:** falling, falling from equal level, falling from unequal level, burning, collision, cutting, scratching, other;
- **location:** in or around home, outside home, traffic;
- **treatment:** medically treated, not medically treated, unknown;
- **cause:** one or more external factors, no external factors.

variables considered		# variables
unqualified	(i.e., any accident)	1
qualified by type	(any fall, fall from equal height, fall from unequal height, collision, other)	5
qualified by location	(in or around home, outside home, traffic)	3
qualified by location and type	(in or around home (8 variables), outside home (5 variables), traffic, (2 types))	15

Table 7.1: 24 accident variables qualified according to type, and/or location.

According to type and/or location, 24 accident variables were defined as listed in Table 7.1 and illustrated in Figure 7.1. In some cases, the accidents burning, scratching and cutting are taken into the category "other", since the frequencies of these accidents in the follow-up were too low. The same 24 variables, as depicted in Figure 7.1, were considered, but qualified according to treatment (3 classes), in principle giving rise to $3 \times 24 = 72$ accident variables. Also a distinction by cause (2 classes) was made, in principle giving rise to $2 \times 24 = 48$ accident variables.

The definitions above would yield $6 \times 24 = 144$ accident variables. However, qualified accidents occurring fewer than 20 times in the sample of 907 persons were disregarded, leaving 100 accident variables to be investigated.

For validation purposes, 21 additional accident variables z, qualified as above, were defined, z taking the value of 1 for a person incurring the corresponding accident at least twice. Since such multiple occurrences are much rarer than single occurrences (as required for y to take the value 1), the number of z-variables is much smaller than the number of y-variables (21 versus 100) because of the restriction that z is considered only if for at least 20 persons in the sample of 907 a z-value of 1 is recorded. If for a z-variable a similar model is found as for its corresponding y-variable, it is unlikely that the relationships between y and its risk factors in the model are due to random variation.

Variables in the pre and post-measurement The demographic variables and variables concerning health aspects measured in the pre-measurement are listed in Table 7.2. The second column gives the number (nobs) and percentage (%) of observed values. The initial letter 'p' (pre) indicates that these variables are measured in the pre-measurement. In the post-measurement, the same health variables as listed in Table 7.2 were measured. Names for these

Demographic variables and general health			
name	nobs (%)	description	categories
age	907 (100)	age in years	
gender	907 (100)	gender	male, female
education	899 (99.1)	education	primary, lower general or vocational, intermediate general, intermediate vocational, higher vocational, higher general, university
household	897 (98.9)	number of persons in household	1,2,3,4, $\geq$ 5 persons
income	861 (94.9)	net income in NLG/month	$\leq$1700, 1700-2150, 2150-2750, 2750-3450, 3450-5000, $\geq$5000
live_time	895 (98.8)	number of years of living in residence	$\leq$1/2 year, 1/2-1 year, 1-5 years, $\geq$5 years
p_gen_h	896 (98.8)	general health	very good, good, reasonable, variable, bad

Health variables concerning disability divided into the four categories: (without difficulty, with some difficulty, with much difficulty, impossible)					
name	nobs (%)	description	name	nobs (%)	description
p_hear1	896 (98.8)	following groups conversation	*p_locom1*	898 (99.0)	carrying 5 kg 10 m
			p_locom2	898 (99.0)	stooping
p_hear2	887 (97.8)	conversation with one person	*p_locom3*	898 (99.0)	walking 400 m
			p_locom4	899 (99.1)	dressing/undressing
p_vision1	897 (98.9)	reading small letters	*p_locom5*	900 (99.2)	getting in/out bed
			p_locom6	897 (98.9)	moving to another room
p_vision2	899 (99.1)	face recognition			

Health variables concerning pain/trouble in body parts divided into the four categories: (no, yes sometimes, yes regular, yes for a long time)					
p_neck	862 (95.0)	in neck	*p_wrist*	876 (96.6)	in wrists or hands
p_u_back	829 (91.4)	in upper back	*p_hip*	866 (95.5)	in hips or thighs
p_l_back	882 (97.2)	in lower back	*p_knee*	879 (96.9)	in knees
p_shoulder	871 (96.0)	in shoulders	*p_foot*	873 (96.3)	in feet or ankles
p_elbow	852 (93.9)	in elbows			

Other health variables divided into the three categories (never or rarely, sometimes, often)					
p_dizzy	889 (98.0)	dizzyness	*p_strength*	886 (97.7)	loss of strength in legs
p_tired	895 (98.7)	tired at daytime			

Attitude variables divided into the five categories (absolutely never, almost never, seldom, occasionally, often)					
p_fear1	890 (98.1)	afraid to fall at home	*p_fear2*	888 (97.9)	afraid to fall outside home

Table 7.2: Names and categories of variables concerning health aspects and attitude in pre-measurement. Also listed are the number (nobs) and percentage (%) of observed values for these variables.

name	nobs (%)	name	nobs (%)	name	nobs (%)
a_gen_h	764 (84.2)	*a_locom5*	765 (84.3)	*a_knee*	757 (83.5)
a_hear1	764 (84.2)	*a_locom6*	765 (84.3)	*a_foot*	750 (82.7)
a_hear2	761 (83.9)	*a_neck*	750 (82.3)	*a_dizzy*	768 (84.7)
a_vision1	767 (84.6)	*a_u_back*	742 (81.8)	*a_strength*	765 (84.3)
a_vision2	768 (84.7)	*a_l_back*	756 (83.4)	*a_tired*	772 (85.1)
a_locom1	770 (84.9)	*a_shoulder*	757 (83.5)	*a_fear1*	758 (83.6)
a_locom2	765 (84.3)	*a_elbow*	741 (81.7)	*a_fear2*	753 (83.0)
a_locom3	762 (84.0)	*a_wrist*	753 (83.0)		
a_locom4	764 (84.2)	*a_hip*	757 (83.5)		

Table 7.3: Number (nobs) and percentage (%) of health variables in the post-measurement.

new variables are defined by replacing the initial letter '*p*' in Table 7.2 by an '*a*' (after). The number (nobs) and percentage (%) of observed values for these variables are listed in Table 7.3. In addition, variables concerning 32 chronic diseases were measured in the post-measurement. Due to low frequencies of categories of some variables, only 29 chronic diseases, as listed in Table 7.4, are considered in the multivariate statistical analysis presented in this chapter. Most of these variables concerning chronic diseases are trichotomous (see top of Table 7.4). Due to low frequencies of some of the categories, some variables are coded binary (see Table 7.4).

7.2.2 Analysis strategy

The ultimate goal of the analysis is to obtain, for each of the 121 accident variables, a logistic regression model, containing the most important risk factors. Candidate risk factors are:

- The demographic variables *age*, *gender*, *education*, *household*, *income* and *live_time* (Table 7.2);
- The health aspects and attitudes toward prevention as measured in the pre-measurement (Table 7.2);
- The 29 chronic diseases as measured in the post-measurement (Table 7.4).

In case of completely observed data, the selection of risk factors would be fairly straightforward: the models are derived by means of stepwise forward logistic regression and the model fit is assessed by the Hosmer-Lemeshow test [6].

coded by (no, yes without medication, yes with medication)		
name	description	nobs (%)
asthma	asthma/ COPD	748 (82.5)
sinusiti	sinusitis	742 (81.8)
myocard	myocardial infarction or cardiac disease	746 (82.2)
hyperten	hypertension	745 (82.1)
intestine	intestine disorder	747 (82.4)
diabetes	diabetes	745 (82.1)
thyroid	thyroid diseases	745 (82.1)
hernia	hernia or other back problems	743 (81.9)
arthrosi	arthrosis	749 (82.6)
arthriti	arthritis	745 (82.1)
rheuma	chronic other rheuma	746 (82.2)
paralysi	paralysis or loss of strength	749 (82.6)
dizzy	dizziness with falling	749 (82.6)
migraine	migraine	746 (82.2)
skin	skin disease	748 (82.5)
cancer	malignant cancer	748 (82.5)
incont	incontinence	746 (82.2)
sleep	sleeplessness	754 (83.1)
stress	severe stress	741 (81.7)
coded by (no, yes)		
stroke	stroke	736 (81.1)
stroke_c	stroke consequences	736 (81.1)
gall_sto	gall stone or inflammation of gall	747 (82.4)
liver_di	liver disease	746 (82.2)
prolaps	prolaps	809 (89.2)
parkinso	Parkinsons disease	749 (82.6)
forget	forgetfulness	740 (81.6)
coded by (no or yes without medication, yes with medication)		
stom_ulc	stomach or dusdenal ulcer	743 (81.9)
cystitis	chronic cystitis	747 (82.4)
coded by (no other disease, one other disease, more than one other disease)		
other_di	other chronic diseases	580 (63.9)

Table 7.4: Variable names, categories and number (nobs) and percentage (%) of observed values for 29 chronic diseases.

However, 10% of the data is missing. Due to the large number of candidate risk factors involved, only 55% of all cases is completely observed. Therefore, listwise deletion is an unattractive option. For this reason, multiple imputation has been applied in order to try (as best as possible) to recover the information hidden in the data set. Multiple imputation results in m ($m \geq 2$) completed data sets, and the selection of risk factors from a set of candidate risk factors based on such a collection of data sets is not straightforward. In the next subsections, the following steps in the analysis chain will be described:

1. Selection of predictor variables for each imputation variable;
2. Specification of an imputation model for each imputation variable and generating the imputations;
3. Selection of risk factors for each accident variable from the candidate risk factors on the basis of the m completed data sets;
4. Fitting of these logistic regression models to each of the m completed data sets and pooling of the results.

Throughout this chapter the following definitions are used:

Accident variables: Any accident variable is the outcome variable of one of the 121 logistic regression models and refers to a qualified accident considered in this study;

Risk factors: The independent variables in a logistic regression model with an accident variable as outcome variable. Different accident variables will generally have different risk factors;

Candidate risk factors: A set of variables from which, for each accident variable, the risk factors in the corresponding regression model are selected. The candidate risk factors are: the demographic variables *age*, *gender*, *education*, *household*, *income* and *live_time* (Table 7.2), health aspects and attitude variables measured in the pre-measurement (Table 7.20, and 29 chronic diseases measured in the post-measurement (Table 7.4);

Imputation variables: The 82 incomplete variables for which imputations are to be generated. Besides the 57 incomplete candidate predictor variables, the imputation variables

A	*p_gen_h, p_hear1, p_hear2, p_vision1, p_vision2, p_locom1, ..., p_locom6, p_neck, p_u_back, p_l_back, p_shoulder, p_elbow, p_wrist, p_hip, p_knee, p_foot, p_dizzy, p_strength, p_tired, p_fear1, p_fear2*
B	*a_gen_h, a_hear1, a_hear2, a_vision1, a_vision2, a_locom1, ..., a_locom6, a_neck, a_u_back, a_l_back, a_shoulder, a_elbow, a_wrist, a_hip, a_knee, a_foot, a_dizzy, a_strength, a_tired, a_fear1, a_fear2*
C	*asthma, sinusiti, myocard, stroke, hyperten, stroke_c, stom_ulc, intestine, gall_sto, liver_di, cystite, prolaps, diabetes, thyroid, hernia, arthrosi, arthriti, rheuma, paralysi, dizzy, migraine, skin, cancer, parkinso, incont, sleep, stress, forget, other_di*
D	*education, household,* live_time, *income*

Table 7.5: The blocks A, B, C and D of imputation variables.

also include the health aspects and attitude variables measured in the post-measurement (Table 7.3). Imputations for this latter group of variables may be useful for future research;

Predictor variables: The independent variables in an imputation model for a certain imputation variable;

Candidate predictor variables: The set of variables from which the predictor variables for a certain imputation variable are selected. Different imputation variables will generally have different candidate predictor variables.

Selection of predictor variables

The variables *age* and *gender*, and the accident variables are completely observed. The imputation variables (variables for which imputations are requested) are divided into the four blocks A, B, C and D as listed in Table 7.5. Blocks A and B consist of the health and attitude variables as measured in the pre- and post-measurement, respectively. Block C contains 29 chronic disease variables; block D consists of the incomplete demographic variables. The variables in block B are not candidate risk factors. However, imputations for these variables may be useful in future analyses.

For the imputation variables in each of the four blocks, the set of candidate predictor variables is listed in Table 7.6. Because of the high correlations, variables from block B are used as predictors for those in block A and vice versa. For block C, variables from block A,

block of imputation variables	candidate predictor variables
A	A, B, *age*, *gender*, 100 accident variables
B	A, B, *age*, *gender*, 100 accident variables
C	A, *age*, *gender*, 100 accident variables
D	A, *age*, *gender*, 100 accident variables

Table 7.6: Candidate predictor variables for the imputation variables.

rather than from block B are used as predictors, since the fraction of missing data entries in block A is smaller than in block B.

Table 7.6 shows that the first 100 accident variables are used as predictors for all four blocks of imputation variables. This ensures consistency of the logistic regression analyses for the various accident variables. In order to reduce the number of predictor variables, these 100 primary accident variables are reduced to the first 9 HOMALS object scores [7]. HOMALS, which is an extention of correspondence analysis to a multivariate crosstable, is a technique for dimension reduction of categorical data. For each of the imputation variables, the following strategy (see also chapter 4) is used to select predictor variables from the candidate predictor variables:

1. Include the first 9 components of the 100 primary accident variables as predictor variables;

2. Exclude from the candidate predictor variables other than accident variables those variables x, with a percentage of usable cases smaller than or equal to 30%. For a pair (y,x) of an imputation variable y and a candidate predictor variable x, the percentage of usable cases $f_p(y,x)$ is:

$$f_p(y,x) = \left(\frac{\text{number of cases for which } y \text{ is missing and } x \text{ is observed}}{\text{number of cases for which } y \text{ is missing}}\right) 100\%; \quad (7.1)$$

3. The set of remaining candidate predictors is now used as a pool to select predictors from. This selection is done by forward stepwise regression, using the data set obtained from listwise deletion. For numerical or ordinal y, stepwise linear regression is applied, and for binary y stepwise logistic regression is applied. When y is polytomous with s categories, $s-1$ separate stepwise logistic regressions of the categories $1,\ldots,s-1$ against a baseline category 0 are applied and the union of the resulting sets of predictor variables

is taken. These stepwise regressions were performed using the statistical package SAS with its default p-values for inclusion and exclusion (0.05 and 0.15, respectively). When stepwise linear regression is applied, only the predictor variables with a partial $R^2 > 1\%$ are included. In the stepwise regressions, the candidate predictor variables are treated as numerical variables.

Specification of an imputation model and generation of the imputations

The ordinal imputation variables in the blocks A, B and D are imputed by linear regression imputation with the normal error-term and round off variant. For the binary variables in block C, logistic regression imputation is applied. Since use or no use of medication for a chronic disease cannot be interpreted in terms of an ordinal ranking of severity, the trichotomous variables in block C (no, yes without medication, yes with medication) are imputed using a polytomous regression model. The $m = 5$ imputations are generated according to the variable-by-variable Gibbs sampling approach described in chapter 4. First, imputations for the variables in the blocks A and B are generated in 10 Gibbs sampling iterations. Subsequently, the imputations for the variables in the blocks C and D are generated in one iteration using the previously completed data for the variables in block A.

Selection of risk factors on the basis of the m=5 completed data sets

Having obtained the $m = 5$ completed data sets, forward stepwise logistic regression is applied to each of them. For each of the accident variables this gives rise to five logistic regression models possibly with different independent variables. There is no standard theory available to combine the information of such different models. This problem is solved in an ad-hoc way as follows:

1. A new super model is constructed incorporating those candidate risk factors which appear in at least three of the five regression models. The rationale for the three out of five criterion is to exclude candidate risk factors which may be selected accidentally and can be considered as a coarse correction for multiple testing;

2. Starting with the super model found in step 1, a stepwise procedure of backward elimination is initiated. In each step, for each candidate risk factor x of the current model the pooled likelihood ratio p-value of this model versus the model excluding x according to the method proposed in [8] (see also chapter 4) is calculated from the 5 completed data sets. If for one or more x, this p-value is larger than 0.05, the candidate risk factor with the largest p-value is discarded and a new step is made, otherwise the procedure stops.

Pooling of the results

The goodness-of-fit for each of the regression models for the 121 accident variables is assessed by the Hosmer-Lemeshow test [6], in which observed and predicted frequencies of the accident variable for up to 10 different groups of cases in the data set are compared. These groups are constructed on the basis of the corresponding probability distributions of the accident variable calculated from the fitted logistic regression model (see [7,10] for more details). The p-values of the Hosmer-Lemeshow test are pooled using the procedure proposed in [10] (see also chapter 4). This procedure is theoretically justifiable, since the test statistic of the Hosmer-Lemeshow test has an approximate χ^2 distribution. The number of degrees of freedom is two less than the number of groups of cases taking part in the comparison of observed and predicted frequencies of the accident variable. The individual contributions of the risk factors x to the model fit are assessed by the pooled likelihood ratio p-value [8] of the full model versus the model excluding x. Odds-ratios are pooled by averaging the five corresponding completed data regression coefficients (the logarithm of the odds-ratio) and taking the exponential transformation of this average. Their confidence intervals are pooled by first calculating the confidence interval for the corresponding regression coefficient [11], followed by an exponential transformation of this interval.

7.3 Results

In this section, for some imputation variables, the steps in the analysis chain are illustrated, the added value of multiple imputation with respect to listwise deletion is examined, and the quality of the imputations is inspected. Throughout this section, results are described for the

imputation variable	predictor variables	R^2
education (ord)	*household, income*	0.38
household (ord)	*gender, age, education, income, live_room*	0.30
p_hear1 (ord)	*p_elbow, a_hear1*	0.60
p_neck (ord)	*p_locom6, p_l_back, p_shoulder, p_tired, a_neck*	0.53
p_shoulder (ord)	*p_locom4, p_locom5, p_neck, p_u_back, a_shoulder*	0.51
p_wrist (ord)	*p_locom4, p_foot, a_wrist*	0.45
stroke_c (bin)	gender, p_locom5	
parkinso (bin)	*p_locom4*	
hernia (tri)	*p_u_back, p_l_back*	
dizzy (tri)	*age, p_gen_h, p_vision1, p_locom2, p_dizzy*	
skin (tri)	*p_locom3, p_hip*	
cancer (tri)	*age, p_vision2, p_u_back, p_shoulder, p_wrist*	
incont (tri)	*p_u_back, p_wrist, p_knee, p_foot, p_dizzy*	
sleep (tri)	*gender, p_gen_h, p_strength, p_fear1*	

Table 7.7: Predictor variables for imputation variables.

three accident variables *y_*acc, *y_*home, and *y_*fall, concerning any accident, accidents in or around home, and fall accidents, respectively. For these variables, the number of cases for which the outcome is 1 is equal to 406, 284 and 258, respectively.

7.3.1 Illustration of the analysis chain

Predictor variables and imputation models

Imputation variables considered here are the incomplete candidate risk factors involved in the super models for *y_*acc, *y_*home, and *y_*fall obtained in step 1 of the ad-hoc strategy for construction of a logistic regression model. These imputation variables and the predictor variables are listed in Table 7.7. Ordinal, binary and trichotomous variables are indicated by "ord", "bin", and "tri", respectively. For the ordinal imputation variables, the multiple R^2 is listed in the rightmost column of this Table. For each of the imputation variables *p_hear1*, *p_neck*, *p_shoulder* and *p_wrist*, the corresponding value obtained in the post-measurement is one of the predictor variables, which results in the high R^2 values.

The ordinal imputation variables are imputed by linear regression imputation with the normal error-term and round off variant. The binary and trichotomous imputation variables are imputed by logistic regression imputation and polytomous regression imputation, respectively.

accident variables	candidate riskfactors of separate models
y_acc	**age(5), gender(5), p_hear1(5), p_shoulder(5), dizzy(5), incont(5), education(4), parkinso(4), stroke_c(3), sleep(3),** *p_locom4*(1), *intestine*(1), *rheuma*(1), *other_di*(1)
y_home	**hernia(5), age(4), dizzy(4), cancer(4), incont(4), sleep(4), p_wrist(3), parkinso(3),** *income*(2), *p_tired*(2), *p_gen_h*(1), *p_shoulder*(1), *gender*(1), *education*(1), *stroke_c*(1), *intestine*(1), *gall_sto*(1), *liver_di*(1), *diabetes*(1), *rheuma*(1), *paralysi*(1)
y_fall	**gender(5), parkinso(5), incont(5), dizzy(5), p_neck(4), skin(4), household(3),** *p_locom*(2), *p_u_back*(2), *stroke_c*(2), *stroke_c*(1), *arthriti*(1), *cancer*(1)

Table 7.8: For the three accident variables, the independent variables appearing in at least one completed data regression model together with the number of such models in which they appear.

Selection of risk factors

For each of the accident variables y_acc, y_home and y_fall, Table 7.8 presents the candidate risk factors appearing in at least one of the five models obtained by applying forward stepwise regression separately to the five completed data sets. The candidate risk factors are ordered according to the number of completed data regression models in which they appear. Candidate risk factors appearing in three or more of such models (presented in bold) are included in the super model. A relatively large number of candidate risk factors (13) appears in only one or two completed data regression models of the accident variable y_home. For the accident variables y_acc, and y_fall, these numbers are four and six, respectively.

The iterative process of backward elimination, starting with the super models for y_acc, y_home and y_fall, is presented in Table 7.9. In each iteration, the independent variables of the current regression models, together with the corresponding pooled likelihood ratio p-values [7] calculated from the 5 completed data sets are listed in this table. When for at least one candidate risk factor, this p-value is larger than 0.05, the one with the largest p-value is excluded, as listed in the right most column of the table, and a new iteration is started.

The selection of risk factors for y_fall requires two iterations, where in the first iteration the candidate risk factor skin with a likelihood ratio p-value of 0.173 is removed. In the second iteration, all p-values are smaller than 0.05. For the accident variables y_acc and y_home, four

step number	**accident variable:** y_acc candidate risk factors with likelihood ratio p-values	exclude variable
step 1	*age*(0.001), *gender*(<0.001), *education*(0.003), *p_hear1*(0.015), *p_shoulder*(0.007), *stroke_c*(0.15), *dizzy*(0.040), *parkinso*(0.102), *incont*(0.002), *sleep*(0.104)	*stroke_c*
step 2	*age*(0.002), *gender*(<0.001), *education*(0.003), *p_hear1*(0.023), *p_shoulder*(0.008), *dizzy*(0.035), *parkinso*(0.078), *incont*(0.004), sleep(0.102)	*sleep*
step 3	*age*(0.002), *gender*(<0.001), education(0.003), p_hear1(0.021), p_shoulder(0.005), dizzy(0.019), *parkinso*(0.069), *incont*(0.004),	*parkinso*
step 4	*age*(0.002), *gender*(<0.001), *education*(0.003), *p_hear1*(0.021), *p_shoulder*(0.005), *dizzy*(0.021), *incont*(0.003),	
	accident variable: y_home candidate risk factors with likelihood ratio p-values	
step 1	*age*(0.020), *p_wrist*(0.062), *hernia*(0.003), *dizzy*(0.060), *cancer*(0.112), parkinso(0.103), incont(0.088), sleep(0.016)	*cancer*
step 2	age(0.029), *p_wrist*(0.035), *hernia*(0.006), *dizzy*(0.063), *parkinso*(0.105), *incont*(0.072), *sleep*(0.017)	*parkinso*
step 3	*age*(0.031), *p_wrist*(0.036), *hernia*(0.007), *dizzy*(0.061), *incont*(0.064), *sleep*(0.014)	*incont*
step 4	*age*(0.041), *p_wrist*(0.011), *hernia*(0.006), *dizzy*(0.042), *sleep*(0.010)	
	accident variable: y_fall candidate risk factors with likelihood ratio p-values	
step 1	*gender*(0.004), *household*(0.009), *p_neck*(0.038), *dizzy*(0.012), *skin*(0.173), *parkinso*(0.015), *incont*(0.021)	*skin*
step 2	*gender*(0.002), *household*(0.007), *p_neck*(0.037), *dizzy*(0.017), *parkinso*(0.014), *incont*(0.024)	

Table 7.9: Backward elimination for the three accident variables.

accident variable: y_acc						
Hosmer-Lemeshow						
test-statistic				df2	p-value	
0.154				25.6	0.995	
odds-ratios						
risk factor	category	reference category or unit	adjusted		unadjusted	
			OR	CI	OR	CI
age		1 year	0.95	(0.93,0.98)	0.98	(0.95,1.00)
gender	female	male	1.86	(1.38,2.49)	1.74	(1.33,2.26)
education	lower general or vocational	primary	1.30	(0.86,1.96)	1.05	(0.72,1.54)
	interm. general		1.40	(0.93,2.11)	1.14	(0.78,1.68)
	interm. vocational		1.72	(1.04,2.86)	1.37	(0.86,2.20)
	higher vocational		1.63	(0.88,3.01)	1.47	(0.83,2.62)
	higher general		2.20	(1.20,4.05)	1.81	(1.02,3.20)
	university		3.64	(1.83,7.24)	2.56	(1.33,4.91)
p_heart	with some difficulty	without difficulty	1.50	(1.06,2.11)	1.07	(0.63,1.82)
	with much difficulty		2.39	(1.16,4.89)	1.24	(0.17,8.85)
	impossible		1.77	(0.68,4.57)	1.24	(0.08.19.91)
p_shoulder	yes sometimes	no	1.82	(1.29,2.56)	1.95	(1.41,2.70)
	yes regular		1.52	(0.92,2.48)	1.83	(1.17,2.83)
	yes for a long time		1.60	(0.74,3.42)	1.97	(0.96,4.01)
dizzy	yes without medication	no	2.90	(1.06,7.59)	3.00	(1.22,7.22)
	yes with medication		3.12	(1.01,9.24)	3.39	(1.17,9.48)
incont	yes without medication	no	1.65	(1.02,2.67)	2.03	(1.29,3.20)
	yes with medication		4.25	(1.54,11.61)	4.50	(1.72,11.69)

Table 7.10: The pooled results for the accident variable y_acc.

iterations were required. For y_acc, the candidate risk factors *stroke_c* (0.150), *sleep* (0.102) and *parkinso* (0.069) were successively excluded, and for y_home, the sequence of excluded candidate risk factors is *cancer* (0.112), *parkinso* (0.105) and *incont* (0.064).

Pooling of the results

The pooled results of the Hosmer-Lemeshow tests, adjusted and unadjusted odds-ratios and their corresponding 95% confidence intervals are presented for the accident variable y_acc in Table 7.10 and for the two accident variables y_home and y_fall in Table 7.11. The pooled test-statistic of the Hosmer-Lemeshow test has an F reference distribution with a numerator number of degrees of freedom of 8 (10 groups of cases minus 2) and a denominator number of

accident variable: $y_$home						
Hosmer-Lemeshow						
test-statistic				df2	p-value	
0.355				13.6	0.927	
odds-ratios						
risk factor	category	reference category	adjusted		unadjusted	
		or unit	OR	CI	OR	CI
age		1 year	0.97	(0.94,1.00)	0.98	(0.95,1.01)
p_wrist	yes sometimes	no	1.68	(1.15,2.47)	1.85	(1.27,2.70)
	yes regular		1.76	(1.14,2.73)	2.17	(1.44,3.29)
	yes for a long time		1.55	(0.68,3.53)	1.78	(0.82,3.86)
hernia	yes without medication	no	0.88	(0.44,1.76)	1.00	(0.51,1.91)
	yes with medication		3.18	(1.61,6.27)	4.12	(2.12,7.93)
dizzy	yes without medication	no	3.67	(1.32,9.67)	4.29	(1.78,9.99)
	yes with medication		1.78	(0.55,5.41)	2.18	(0.76,5.94)
sleep	yes without medication	no	1.40	(0.81,2.40)	1.52	(0.90,2.54)
	yes with medication		1.94	(1.27,2.95)	2.30	(1.55,3.39)
accident variable: $y_$fall						
Hosmer-Lemeshow						
test-statistic				df2	p-value	
0.281				260.9	0.972	
odds-ratios						
risk factor	category	reference category	adjusted		unadjusted	
		reference	OR	CI	OR	CI
gender	female	male	1.69	(1.20,2.38)	2.08	(1.54,2.80)
household	2 persons	1 person	0.72	(0.52,1.01)	0.58	(0.43,0.79)
	3 persons		0.50	(0.18,1.38}	0.35	(0.13,0.93)
	4 persons		14.60	(1.57,135)	7.30	(0.82,66.83)
	more than 5 persons		2.17	(0.13,36.4)	1.85	(0.11,29.79)
p_neck	yes sometimes	no	1.46	(1.03,2.07)	1.59	(1.14,2.23)
	yes regular		1.84	(1.15,2.96)	2.13	(1.36,3.33)
	yes for a long time		1.49	(0.58,3.82)	2.09	(0.84,5.21)
dizzy	yes without medication	no	2.24	(0.96,5.16)	3.01	(1.38,6.49)
	yes with medication		3.03	(1.14,7.84)	3.76	(1.41,9.67)
parkinso	yes	no	5.63	(1.40,20.7)	5.17	(1.26,19.14)
incont	yes without medication	no	1.56	(0.95,2.52)	2.05	(1.31,3.20)
	yes with medication		2.87	(1.08,7.37)	3.58	(1.43,8.71)

Table 7.11: The pooled results for the accident variables $y_$home and $y_$fall.

degrees of freedom listed in the Tables 7.10 and 7.11 under "df2". The p-values close to 1 for each of the three accident variables indicate an adequate model fit. An unadjusted odds-ratio for a risk factor x and an accident variable y is estimated from the data of y and x only. An adjusted odds-ratio for x is the odds-ratio corresponding to the entire logistic regression model and can be interpreted as the individual contribution of x to the accident risk. The unit for the odds-ratio for *age* is 1 year, i.e., the odds-ratio for *age* indicates the relative increase in risk per year. For the other risk factors, the odds-ratios of the categories relative to the first category as reference category are presented. When a confidence interval does not include the value 1, the corresponding odds-ratio differs significantly from 1.

For most of the risk factors, the adjusted odds-ratio is closer to 1 than the corresponding unadjusted odds-ratio which signifies that in these cases the association between a risk factor and an accident variable can be partly explained by the other risk factors. For the risk factors *age, gender, education, p_heart* for the accident variable y_acc, and the risk factor *age* for the accident variable y_home, this is not the case. The confidence intervals for the odds-ratios for the categories "4 persons" and "more than 5 persons" of the risk factor *household* for the accident variable y_fall are very large. This is due to random error since the observed frequencies for these categories are very skew (5 and 2).

The two odds-ratios for *age* for the accident variables y_acc and y_home are smaller than 1, indicating a decreasing risk with increasing *age*, while *age* is generally considered as an important risk factor. An explanation for this may be the healthy worker effect. As expected, the chronic diseases hernia, dizziness with falling (*dizzy*), sleeplessness (*sleep*), Parkinsons disease (*parkinso*) and incontinence (*incont*) indicate a higher accident risk. In most of the cases, use of medication for a chronic disease indicates a higher risk than no use of medication. An exception is the risk factor dizzy for the accident variable y_home.

7.3.2 Added value of multiple imputation with respect to listwise deletion

The regression models found by multiple imputation are compared with those found by listwise deletion for the accident variables y_acc, y_home and y_fall. With the listwise deletion approach, forward stepwise regression is applied to the 405 cases for which each candidate risk factor is observed. Starting with the resulting model, a similar process of backward elimination,

independent variable	y_acc		y_home		y_fall	
	MI	LWD	MI	LWD	MI	LWD
age	0.95		0.97			
gender	1.86				1.69	
education	1.30					
	1.40					
	1.72					
	1.63					
	2.20					
	3.64					
household					0.72	0.62
					0.50	0.49
					14.60	8.25
					2.17	2.08
income		0.73				
		1.07				
		1.02				
		0.76				
		2.05				
p_hear1	1.50					
	2.39					
	1.77					
p_neck					1.46	
					1.84	
					1.49	
p_shoulder	1.82					
	1.52					
	1.60					
p_wrist			1.68	1.93		
			1.76	2.10		
			1.55	1.32		
hernia			0.88	1.07		
			3.18	3.88		
dizzy	2.90		3.67		2.24	
	3.12		1.78		3.03	
parkinso		5.84		6.12	5.63	7.57
incont	1.65				1.56	
	4.25				2.87	
sleep			1.40			
			1.94			

Table 7.12: Comparison of the models found by multiple imputation (MI) and listwise deletion. (LWD)

as is used for the construction of a regression model from the five completed data sets, is applied. In this process, the corresponding likelihood ratio p-values are calculated from the cases for which each candidate risk factor of the current model is observed.

In Table 7.12, the odds-ratios for the risk factors of the regression models for the three accident variables resulting from multiple imputation (MI) and from listwise deletion (LWD) are listed. A blank entry indicates that the corresponding risk factor is not included in the corresponding model. The models resulting from listwise deletion and from multiple imputation *are very different* for the three accident variables *y_*acc, *y_*home and *y_*fall. For *y_*acc, multiple imputation and listwise deletion result in two models for which the two sets of independent variables are even disjunct. For the two accident variables *y_*home and *y_*fall, the two models have two risk factors in common. For each of the three accident variables *y_*acc, *y_*home and *y_*fall, *the models resulting from multiple imputation contain considerably more risk factors than the models resulting from listwise deletion.* This is in concordance with the fact that multiple imputation uses more information available in the data set than listwise deletion, so that with multiple imputation risk factors are identified earlier and better.

7.3.3 Quality inspection of generated imputations

For some imputation variables, the quality of the generated imputations has been inspected by examining whether the differences or similarities between the distributions of the observed values and of the imputed values in the various categories, can be explained by the underlying missing data mechanism under the MAR assumption (see also sub section 4.4.3 of chapter 4)..

In Table 7.13, frequency distributions in the various categories (cg) of the imputation variables listed in Table 7.7, except *education, household,* and *p_hear1*, which have only 10, 7 and 11 missing data entries, are presented. Table 7.13 lists the frequency distribution of the following data sets (from left to right):

- the incomplete data set (obs);
- averaged over the generated imputations (imp);
- averaged over the five completed data sets (cmpav);

mputation variable	cg	obs	(%)	imp	(%)	cmpav	(%)
p_neck	1	478	(55.5)	13.6	(30.2)	491.6	(54.2)
	2	255	(29.6)	23.2	(51.6)	278.2	(30.7)
	3	109	(12.7)	7.0	(15.6)	116.0	(12.8)
	4	20	(2.3)	1.2	(2.7)	21.2	(2.3)
p_shoulder	1	548	(62.9)	12.0	(33.3)	560.0	(61.7)
	2	194	(22.3)	15.4	(42.8)	209.4	(23.1)
	3	97	(11.1)	6.6	(18.3)	103.6	(11.4)
	4	32	(3.7)	2.0	(5.6)	34.0	(3.8)
p_wrist	1	579	(66.1)	15.0	(48.4)	594.0	(65.5)
	2	154	(17.6)	12.0	(38.7)	166.0	(18.3)
	3	114	(13.0)	3.6	(11.6)	117.6	(13.0)
	4	29	(3.3)	0.4	(1.3)	29.4	(3.2)
stroke_c	1	723	(98.2)	165.6	(96.8)	888.6	(98.0)
	2	13	(1.8)	5.4	(3.2)	18.4	(2.0)
hernia	1	665	(89.5)	143.6	(87.6)	808.6	(89.2)
	2	44	(5.9)	10.2	(6.2)	54.2	(6.0)
	3	34	(4.6)	10.2	(6.2)	44.2	(4.9)
dizzy	1	710	(94.8)	143.4	(90.8)	853.4	(94.1)
	2	22	(3.0)	8.6	(5.4)	30.6	(3.4)
	3	17	(2.3)	6.0	(3.8)	23.0	(2.5)
skin	1	729	(96.3)	149.4	(94.0)	878.4	(96.9)
	2	6	(0.8)	5.0	(3.1)	11.0	(1.2)
	3	13	(1.7)	4.6	(2.9)	17.6	(1.9)
cancer	1	720	(96.3)	146.8	(92.3)	866.8	(95.6)
	2	17	(2.3)	7.8	(4.9)	24.8	(2.7)
	3	11	(1.5)	4.4	(2.8)	15.4	(1.7)
parkinso	1	738	(98.5)	152.6	(96.6)	890.6	(98.2)
	2	11	(1.5)	5.4	(3.4)	16.4	(1.8)
incont	1	652	(87.4)	134.4	(83.5)	786.4	(86.7)
	2	75	(10.1)	20.2	(12.6)	95.2	(10.5)
	3	19	(2.6)	6.4	(4.0)	25.4	(2.8)
sleep	1	568	(75.3)	109.4	(71.5)	677.4	(74.7)
	2	72	(9.6)	15.2	(9.9)	87.2	(9.6)
	3	114	(15.1)	28.4	(18.6)	142.4	(15.7)

Table 7.13: Frequency distribution of observed data, imputations, and completed data sets for some imputation variables.

imputation variable: *p_ neck*								
cg	$\widehat{P}_{\text{obs}(y)}$	$\widehat{P}_{\text{imp}(y)}$	$\widehat{P}^{\text{MAR}(x)}_{\text{mis}(y)}$ for the predictor variables: *p_ locom6*, *p_ l_ back*, *p_ shoulder*, *p_ tired* and *a_ neck*					
1	55.5	30.2	54.5	44.2	48.8	49.7	56.2	
2	29.6	51.6	31.1	35.6	32.0	31.6	28.7	
3	12.7	13.6	12.2	17.2	15.9	15.0	13.1	
4	2.3	2.7	2.3	3.0	3.4	3.8	2.1	
imputation variable: *p_ shoulder*								
cg	$\widehat{P}_{\text{obs}(y)}$	$\widehat{P}_{\text{imp}(y)}$	$\widehat{P}^{\text{MAR}(x)}_{\text{mis}(y)}$ for the predictor variables: *p_ locom4*, *p_ locom5*, *p_ neck*, *p_ u_ back*, *p_ l_ back*, *a_ shoulder*					
1	62.9	33.3	61.0	61.8	58.1	57.9	57.0	55.0
2	22.3	42.8	22.2	22.8	21.9	25.3	23.1	24.7
3	11.1	18.3	12.1	12.0	14.6	12.4	15.7	15.1
4	3.7	5.6	4.8	3.5	5.4	4.5	4.2	5.3
imputation variable: *p_ wrist*								
cg	$\widehat{P}_{\text{obs}(y)}$	$\widehat{P}_{\text{imp}(y)}$	$\widehat{P}^{\text{MAR}(x)}_{\text{mis}(y)}$ for the predictor variables: p_ locom4, p_ foot, a_ wrist					
1	66.1	48.4	68.6	62.0	68.2			
2	17.6	38.7	17.7	19.1	17.0			
3	13.0	11.6	11.3	14.8	11.8			
4	3.3	1.3	2.4	4.2	3.1			

Table 7.14: Estimated distribution of the unobserved values of some imputation variables.

In general, the distributions of the observed and of the imputed data fit quite well, though it appears that differences are fairly large for the imputation variables *p_ neck*, *p_ shoulder* and *p_ wrist*. For the first category of these imputation variables the proportion of imputed values is much *smaller* than the proportion of observed values. This is mainly compensated with the second category. From the column of the average completed data frequencies (cmpav), it appears differences have little effect on the distributions of the completed data sets.

The distributions $\widehat{P}_{\text{obs}(y)}$, $\widehat{P}_{\text{imp}(y)}$, $\widehat{P}^{\text{MAR}(x)}_{\text{mis}(y)}$ for the imputation variables *p_ neck*, *p_ shoulder* and *p_ wrist* and their predictor variables x are listed in Table 7.14. The distributions $\widehat{P}_{\text{obs}(y)}$, $\widehat{P}_{\text{imp}(y)}$ are the distributions of the observed and of the imputed values. The distribution $\widehat{P}^{\text{MAR}(x)}_{\text{mis}(y)}$ is the estimated distribution of the unobserved values of y under the MAR(x) assumption that the nonresponse in y depends on the observed values of x only (see subsection 4.4.3 of chapter 4). When for some predictor variables x the difference between $\widehat{P}_{\text{imp}(y)}$ and

$\widehat{P}^{\mathrm{MAR}(x)}_{\mathrm{mis}(y)}$ are small, the difference between the imputed and of the observed values for y can be explained by nonresponse.

The nonresponse appears to be systematic for the imputation variable *p_neck* but not for the other two. For *p_neck*, the difference between the distributions $\widehat{P}_{\mathrm{obs}(y)}$ and $\widehat{P}^{\mathrm{MAR}(x)}_{\mathrm{mis}(y)}$ is large for the predictor variable *p_l_back*. Although for *p_l_back*, the distribution $\widehat{P}^{\mathrm{MAR}(x)}_{\mathrm{mis}(y)}$ deviates from $\widehat{P}_{\mathrm{obs}(y)}$ in the same direction as the distribution $\widehat{P}_{\mathrm{imp}(y)}$ deviates from $\widehat{P}_{\mathrm{obs}(y)}$, the difference between the distributions of the imputed values and of the observed values seem to be too large to be explained by systematic nonresponse. This is especially the case for *p_shoulder* and *p_wrist*, where for each predictor variable x, the distribution $\widehat{P}^{\mathrm{MAR}(x)}_{\mathrm{mis}(y)}$ is approximately equal to the distribution $\widehat{P}_{\mathrm{obs}(y)}$ or slightly different from this distribution, while the distributions $\widehat{P}_{\mathrm{obs}(y)}$ and $\widehat{P}_{\mathrm{imp}(y)}$ differ considerably. Thus there is some evidence that the method used for *p_shoulder*, *p_wrist* and *p_neck* might be improper, but the exact cause for these differences is still an open question.

7.4 Discussion

This chapter describes the application of multiple imputation, using the missing data engine in HERMES, prior to a multivariate analysis in the Leiden Safety Observed Study. The main purpose of this chapter is to illustrate the methodology. Conclusions and future preventive measures emerging from the analysis will be published elsewhere.

The selection of risk factors for an accident variable from a set of candidate risk factors on the basis of five completed data sets requires some thought. In case of a completely observed data set, such risk factors are usually selected by means of stepwise regression. Application of stepwise regression to each of the five completed data sets, however, will generally result in five regression models with possibly different independent variables, and there is no standard theory available for pooling such different models. In this chapter, an ad hoc approach to this problem was taken. An alternative, theoretically justifiable approach would be a stepwise regression algorithm in which in each iteration the inclusion or exclusion of independent variables x is based on the pooled likelihood ratio p-value [8] (see also subsection 4.3.1 of chapter 4) of the current regression model versus this model excluding x, as calculated from the five completed

data sets. However, this would require that pooling of multiple imputation results is build into the stepwise algorithm which would be cumbersome to implement.

Although this is an example in which the application of multiple imputation is not particularly easy, applying the EM algorithm [12] in this case is even more difficult, since EM does not directly provide the likelihood ratio p-values that are necessary for the selection of the risk factors from the set of candidate risk factors. Also, using the variable-by-variable Gibbs sampling approach, it is conceptually simpler to find an appropriate imputation model than it would be using the data augmentation approach [13], which starts from a multivariate statistical model.

In general, regression models found by multiple imputation appear to be different from those found by listwise deletion. For the accident variables $y_$acc, $y_$home, $y_$fall, the models resulting from multiple imputation contain considerably more independent variables than the models resulting from listwise deletion. This suggests that with multiple imputation the information available in the data set is used more efficiently than with listwise deletion.

Bibliography

[1] Hertog P Den, Toet H, Ongevallen bij ouderen: een analyse van ongevalsgegevens met betrekking tot personen van 55 jaar en ouder. Amsterdam: Stichting Consument en Veiligheid, 1995

[2] Wijlhuizen GJ, Staats PGM, Radder JJ, Veiligheid in de peiling; Een epidemiologisch onderzoek naar determinanten van ongevallen die in- en om huis plaatsvinden bij ouderen (65-84). TNO-PG, Leiden, 1996

[3] Berg RL, Cassels JS, eds. Falls in older persons: risk factors and prevention. In: The second fifty years; promoting health and preventing disability. Washington DC: Institute of Medicine, 1990

[4] Sonsbeek, van JLA, Methodische en inhoudelijke aspecten van OESO-indicator betreffende langdurige beperkingen in het lichamelijk functioneren. Mndber gezondheid (CBS), 1998;6-4-17

[5] CBS, Centraal Bureau voor de Statistiek. Gezondheidsenquete. 's-Gravenhage, SDU, 1993

[6] Hosmer DW, Lemeshow S, Goodness-of-fit tests for the multiple logistic regression model. Commun. Statist.-Part A Theor. Methd. A9(10), 1980:1043-1069

[7] Gifi A, Nonlinear multivariate analysis. Wiley Chichester, 1990

[8] Meng XL, Rubin DB, Performing likelihood ratio tests with multiply-imputed data sets. Biometrika, Vol 79, No.1, 1992:103-111

[9] Hosmer DW, Lemeshow S, Applied Logistic Regression, Wiley and Sons, New York, 1989

[10] Li KH, Raghunathan TE, Rubin DB, Significance Levels from Repeated p-values with Multiply Impued Data. Statistica Sinica, Vol.1. No.1, 1991:65-92

[11] Rubin DB, Multiple imputation for nonresponse in surveys. Wiley New York, 1987

[12] Dempster AP, Laird NM, Rubin DB, Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), J. Roy. Stat. Soc., B39, 1997: 1-38

[13] Tanner MA, Wong WH, The calculation of the posterior distribution by data augmentation. Journal of the American Statistical Association, Vol. 82, No. 398, 1987: 528-550

Chapter 8

Summary and Conclusions

This thesis is concerned with scientific aspects of the development of an interactive system for the statistical analysis of incomplete data sets by embedded multiple imputation. This system is called the missing data engine and it is implemented in the indirect client-server based HERMES (HEalth care and Research Mediating System) Medical Workstation environment developed at the Department of Medical Informatics of the Erasmus University Rotterdam, The Netherlands. More in particular these aspects concern the statistical method for the analysis of incomplete data in particular multiple imputation, construction of the interface and other implementation aspects.

Chapter 1 addresses problems associated with missing data and motivates the desirability of the missing data engine.

Assumptions about the occurrence of missing data are formulated as a missing data mechanism and incorporated in the analysis. Typically, it is assumed that a hypothetical complete data set exists which is only partly observed due to an unknown process. Missing data mechanisms are classified into the three basic classes: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). The definitions of the three classes of missing data mechanisms, especially the distinction between MCAR and MAR, are often misunderstood. The main purpose of **chapter 2** is to illustrate the concept behind these classes by means of simple numerical examples.

Chapter 3 argues that multiple imputation is the best approach known at this time for the statistical analysis of incomplete data. With multiple imputation, for each missing data

entry, m statistically likely values are filled in. In this chapter eight basic approaches, subdivided into simple and advanced approaches, are discussed. Simple approaches considered are: listwise deletion, available-case analysis, single imputation, the indicator method, and weighting. Advanced approaches are: EM (Expectation Maximization), posterior based approaches, and multiple imputation. These approaches are evaluated with respect to statistical validity, efficiency, possibility of the use of standard software for complete data, and suitability for statistical inference from small samples. Statistical validity is subdivided into bias and precision. This chapter concludes that multiple imputation is the method of choice since it has good properties regarding statistical validity, efficiency, and the use of standard statistical software for complete data. A limitation of multiple imputation is, however, that it is, similar to EM, a large sample tool, and that its properties for small samples are not yet well understood.

Chapter 4, proposes the variable-by-variable Gibbs sampling algorithm for the generation of imputations and a strategy for the selection of its parameters. In this approach, a separate imputation model is specified for each imputation variable, including the specification of predictor variables and the form of the model (linear regression, logistic regression, polytomous regression imputation, etc.). If an incomplete predictor variable, other than an imputation variable of interest, is selected, this variable is added to the list of imputation variables and additional predictor variables and model form are selected for it. A distinction is made between elementary imputation methods generating imputations for one imputation variable and compound imputation methods generating imputations for two or more imputation variables. The main advantages of the variable-by-variable Gibbs sampling algorithm are:

- The variable-by-variable Gibbs sampling algorithm is suitable for imputation of large data sets with many variables;
- Specification of an imputation model variable-by-variable is conceptually easier than the specification of an encompassing multivariate model.

On the other hand, no definite proof of convergence can be given, but the results given in chapter 5 suggest that the algorithm works quite well in practice. All imputation methods concerning the variable-by-variable Gibbs sampling algorithm are developed under the MAR assumption and thus do not require explicit modelling of the underlying missing data mecha-

nism. It is also possible to develop imputation methods incorporating an MNAR missing data mechanism, but this is outside the scope of this thesis. Furthermore, chapter 4 gives an overview of existing methods for pooling of the m completed data results, and provides a non-technical explanation of Rubin's validation criterion of proper multiple imputation. Simplified conditions for Monte Carlo evaluation of proper multiple imputation are given.

Chapter 5 describes a simulation study in which some of the imputation methods using the variable-by-variable Gibbs sampling algorithm are validated according to the simplified conditions for proper imputation given in chapter 4. It is concluded that elementary and compound imputation methods consisting of linear regression , logistic regression, or polytomous regression imputation are generally approximately proper when their underlying assumptions are true. It also appears that these methods are robust against deviations of the underlying statistical model which is in line with other findings in literature.

Chapter 6 describes the design, implementation and validation of the missing data engine in HERMES. Eleven requirements are formulated and translated into a conceptual model. The core of the missing data engine is implemented in HERMES as three functionally independent modules: the missing data server, the imputation server and the pooling server. The missing data server coordinates the entire multiple imputation cycle and uses the imputation server for the generation of the imputations. The pooling server coordinates the repeated analysis of the m completed data sets and the pooling of the m intermediate results into one final result. The analysis of a completed data set is carried out by a statistical server that encapsulates the statistical package BMDP. The missing data engine is tested by separately validating the missing data server, the imputation server and the pooling server, and by verifying whether the exchange of messages between these modules is correct. The imputation server is tested by comparing its results with the results of the simulation program in SAS/IML described in chapter 5 for the same incomplete data set, target statistics and imputation methods. The results of the missing data engine and the simulation program appear to be approximately the same. Although the complete validation of the imputation server requires a more extensive study, the results indicate that the imputation server is reliable.

Chapter 7 describes the application of the missing data engine to a multivariate statistical analysis of the Leiden Safety Observed study conducted at the TNO Prevention and Health in

Leiden, The Netherlands. The main purpose of this study is to find risk factors for several types of accidents incurred by elderly people. For each type of accident considered, the risk factors are selected from a set of 60 candidate risk factors consisting of demographic variables, health aspects and chronic diseases. One particular problem treated here concerns the application of multiple imputation to stepwise methods. This study mainly considers the methodology. Conclusions and recommendations emerging from this analysis are to be published elsewhere.

Chapter 9

Samenvatting en conclusies

Dit proefschrift houdt zich bezig met de wetenschappelijke aspecten van de ontwikkeling van een interaktief systeem voor de statistische analyse van incomplete gegevensbestanden met behulp van multiple imputation. Dit systeem wordt de missing data machine genoemd en is geimplementeerd in de op een indirect client-server model gebaseerde HERMES (Health care and Research Mediating System) medische werkstation omgeving ontwikkeld op de vakgroep Medische Informatica aan de Erasmus Universiteit te Rotterdam, in Nederland. Meer in het bijzonder, deze aspecten betreffen de statistische methoden voor de analyse van incomplete data sets, de constructie van het interface en andere implementatie aspecten.

Hoofdstuk 1 noemt het met missing data samenhangende probleem en motiveert de wenselijkheid van de missing data machine.

Aannamen over het optreden van missing data zijn geformuleerd als een missing data mechanisme en verwerkt in de statistische analyse. In het algemeen is verondersteld dat een hypothetische complete data set bestaat die alleen gedeeltelijk is geobserveerd tengevolge van een onbekend proces. Missing data mechanismen zijn geklassificeerd in de drie basis klassen: Missing Completely At Random (MCAR), Missing At Random (MAR) en Missing Not At Random (MNAR). Over de definities van de drie klassen van missing data mechanismen, in het bijzonder het onderscheid tussen MCAR en MAR, bestaan veel misverstanden. Het doel van **hoofdstuk 2** is het concept achter deze klassen te illustreren met behulp van eenvoudige numerieke voorbeelden.

Hoofdstuk 3 beargumenteert dat multiple imputatie tot dusver de beste strategie is voor

de analyse van incomplete data. Met multiple imputatie worden voor elke missing data entry, m statistisch waarschijnlijke waarden ingevuld. In dit hoofdstuk worden acht basis methoden, onderverdeeld in eenvoudige en geavanceerde methoden, bediscussieerd. Eenvoudige methoden die worden beschouwd zijn: list-wise deletion, available-case analysis, single imputation, de indicator methode en weging. Geavanceerde methoden zijn: EM (Expectation Maximisation), op een posterior verdeling gebaseerde methoden en multiple imputation. Deze methoden zijn geëvalueerd tenopzichte van statistische validiteit, de mogelijkheid om standaard statistische software te gebruiken, en de geschiktheid voor de statistische analyse van kleine steekproeven. Statistische validiteit is onderverdeeld in bias en precisie. De conclusie van dit hoofdstuk is dat multiple imputatie de meest geschikte methode is omdat het goede eigenschappen heeft ten opzichte van statistische validiteit, efficiency en het gebruik van standaard statistische software voor complete data. Een beperking van multiple imputatie is echter, dat het evenals EM, alleen goed geschikt is voor grote steekproeven, en dat de eigenschappen van multiple imputatie voor kleine steekproeven niet goed bekend zijn.

Hoofdstuk 4, introduceert het variable-by-variable Gibbs sampling algoritme voor het genereren van de imputaties en een strategie voor de selectie van de parameters van dit algoritme. Bij Gibbs sampling is voor elke imputatie variabele een afzonderlijk imputatie model gespecificeerd, wat bestaat uit de selectie van predictor variabelen en de vorm van het statistische model (lineaire regressie, logistische regressie, polytome regressie imputatie, etc.). Indien een incomplete variabele anders dan een predictor variabele is geselecteerd, wordt deze variabele toegevoegd aan de lijst van imputatie variabelen. Voor deze incomplete predictor variabele worden dan additionele predictor variabelen geselecteerd. Onderscheid is gemaakt tussen elementaire imputatie methoden die imputaties genereren voor één imputatie variabele en samengestelde imputatie methoden die imputaties genereren voor twee of meer imputatie variabelen. De belangrijkste voordelen van het variable-by-variable Gibbs sampling algoritme zijn:

- Het variable-by-variable Gibbs sampling algoritme is geschikt voor het imputeren van grote data sets met veel variabelen;
- De specificatie van een imputatie model per variabele is conceptueel eenvoudiger dan de

specificatie van een omvattend multivariate model.

Aan de andere kant kan geen waterdicht bewijs van convergentie worden gegeven, maar de resultaten in hoofdstuk 5 suggereren dat dit algoritme goed werkt in de praktijk. Alle imputatie methoden betreffende het variable-by-variable Gibbs sampling algoritme zijn ontwikkeld onder de MAR aanname en vereisen geen expliciete modellering van het onderliggende missing data mechanisme. Het is ook mogelijk om imputatie methoden te ontwikkelen waarin een MNAR missing data mechanisme is verwerkt, echter dit valt buiten het bereik van dit proefschrift. Verder geeft hoofdstuk 4 een overzicht van bestaande methoden voor het poolen van de m gecompleteerde data resultaten en is dit hoofdstuk voorzien van een niet-technische beschrijving van Rubin's validatie criterium proper multiple imputation. Vereenvoudigde voorwaarden voor Monte Carlo evaluatie van multiple imputatie zijn gegeven.

Hoofdstuk 5 beschrijft een simulatie onderzoek waarin sommige imputatie methoden gebruik makende van het variabele-by-variable Gibbs sampling algoritme zijn gevalideerd volgens de vereenvoudigde voorwaarden voor proper multiple imputatie beschreven in hoofdstuk 4. De conclusie is dat elementaire en samengestelde imputatie methoden bestaande uit lineaire regressie, logistische regressie of polytome regressie imputatie in het algemeen bij benadering proper zijn wanneer de onderliggende aannamen waar zijn. Ook blijkt het dat deze methoden robuust zijn tegen afwijkingen van het onderliggende statistische model, wat overeenkomt met andere bevindingen in de literatuur.

Hoofdstuk 6 beschrijft het ontwerp, de implementatie en validatie van de missing data machine in HERMES. Elf vereisten zijn geformuleerd en vertaald in een conceptueel model. De kern van de missing data machine is geimplementeerd in HERMES als drie functioneel onafhankelijke modulen: de missing data server, de imputatie server en de pooling server. De missing data server coördineert de gehele multiple imputation cyclus and gebruikt de imputatie server voor het genereren van de imputaties. De pooling server coördineert de herhaalde analyse van de m gecompleteerde data sets en het poolen van de m tussenresultaten tot één eindresultaat. De missing data machine is gevalideerd door het afzonderlijk valideren van de missing data server, imputatie server en pooling server, en door te verifieren dat het berichten verkeer tussen de modules correct is. De imputatie server is gevalideerd door de resultaten van deze server te vergelijken met de resultaten van het simulatie programma in SAS/IML, beschreven

in hoofdstuk 5 voor dezelfde incomplete data set, statistieken en imputatie methoden. De resultaten van de missing data machine en het simulatie programma bleken bij benadering dezelfde te zijn. Ofschoon de volledige validatie van de imputatie server een uitgebreidere studie vereist, wijzen de resultaten er op dat de imputatie server betrouwbaar is.

Hoofdstuk 7 beschrijft de toepassing van de missing data machine op de multivariate statistische analyse in de studie 'Veiligheid In de Peiling'. Deze study is uitgevoerd op het instituut TNO Preventie en Gezondheid in Leiden in Nederland. Het hoofddoel van deze studie is het vinden van risicofactoren voor verschillende typen van ongevallen die voorkomen bij oudere mensen. Voor elk type van de beschouwde ongevallen zijn risicofactoren geselecteerd van een verzameling van 60 kandidaat risicofactoren bestaande uit demografische variabelen, gezondheidsaspecten en chronische ziekten. Een specifiek probleem betreft de toepassing van multiple imputatie op stapsgewijze methoden. Deze studie beschrijft hoofdzakelijk de methodologie. Conclusies en aanbevelingen aan de hand van deze analyse worden elders gepubliceerd.

Acknowledgement

My Ph.D. research can be compared very well with a long and arduous voyage of discovery whose goal is to map out a large and unknown area. During this long voyage many assistants and guides were involved. Without their help and support I would certainly have not been able to complete this voyage and accomplish this thesis.

This voyage of discovery was a unique cooperation between TNO Prevention and Health (TNO-PG) in Leiden, The Netherlands, and the Department of Medical Informatics at the Erasmus University in Rotterdam, The Netherlands. I thank TNO-PG for funding this research, and especially, ir. C. Zeelenberg and Dr. J.L.A. van Rijckevorsel for their efforts to make it possible to realize this research, which was a bit unusual for the Department of Medical Informatics. Although not easy, this period was also a unique and extremely valuable investment in myself.

My main guides during this voyage were my supervisor Prof. dr. E.S. Gelsema and my co-supervisor Dr. S. van Buuren. Dear Edzard, your endless patience in correcting myriad's of versions of my chapters, the very pleasant way in which you gave guidance to my research, your sense of humour, and the great freedom you gave in carrying out this research in my own way, made a very deep impression on me. Thanks to your help, I was able to develop in a relatively short period all skills I needed for describing clearly in my thesis all my discoveries and insights I acquired during my voyage of discovery. Dear Stef, especially by you I learned to judge my own research critically and to argue the choices of the used methods, which was not always easy. Your contribution to the quality of my thesis was very considerable.

Beside these two important guides, there were many coaches, who also contributed to this thesis. Therefore, I thank Dr. E.M. van Mulligen for his efforts in assisting me to learn programming in C, HERMES and X-windows, skills needed for the development of the missing-data machine (chapter 6), and for helping me with any kind of practical problems. I thank Ronald Cornet and Martin Kalshoven for their willingness to help me with programming during the beginning of my research. When after many segmentation violations, bus errors, HtNerror messages, and other errors, I was desperate for a short moment, I could always ask for help, and nearly always we found a solution. I thank Dr. P.G.H. Mulder for his very useful contribution to my thesis. Paul, it was very valuable that my thesis was also reviewed from a more mathematical-statistical point of view so that I was able to cross the t's and dot the i's. I thank Gert-Jan Wijlhuizen and Jan Radder for the very pleasant cooperation we had during carrying out the SECVIP project (chapter 7). I thank Ton Rövekamp for his efforts regarding the successful transfer of the missing data machine to TNO-PG. I thank Dr. C.G.M. Oudshoorn and Prof. dr. Th. Stijnen for their comments on the final phase of my thesis.

Although I carried out this research for a very large part according to my own vision, I was glad to have the opportunity to learn from Prof. dr. D.B. Rubin during a month at Harvard University in Boston, USA. Donald, I am very grateful for this unique opportunity. Thanks to the discussions I had with you, I gained a much deeper

understanding in multiple imputation, which was very essential for carrying out my voyage of discovery.

Beside al these guides and coaches I thank al my colleagues for the pleasant time at TNO and at the Department of Medical Informatics. I especially thank my colleague Dr. A.M. van Ginneken with whom I found a very warm friendship during my research. Astrid, a friendship with you is almost as with Orca's that are inspiring you so much, based on sincerity, honesty and mutual respect and the transfer of the deepest possible thoughts and feelings. As you know, this research was not an easy period in my life and required a lot of struggling. This period was a voyage to my own awareness of self, and as such, taught me to know and appreciate myself better. Astrid, your contribution to this phase in my life was extremely important. For my deepest thoughts and feelings you always had an open ear.

I thank paranymphs Joris van der Koogh and Astrid van Ginneken for coordinating the celebration after the defense of my thesis. You are really the best friends I have. I thank Désirée de Jong, Rosa Scholte and Prof. dr. J.H. van Bemmel for their great help during the last preparations for the defense of my thesis and the following celebration.

I thank my new employer Statistical Solutions Ltd. in Cork, Ireland for the unique opportunity they offered me to bring in practice the expertise I developed during my research. The very long time I hope to live and work in Ireland are a splendid opportunity to explore this beautiful country with its interesting and exciting history. Although it is uncertain whether I ever return to The Netherlands, I will always love my fatherland, being proud to be a Dutchman. At the cover of this thesis, this new phase in my life is represented in a magnificent way by the designer Richard Smithers of Allied Print Ltd. Richard, I am very grateful for the efforts you put in this design without any charge. I thank my new colleague Helen Murphy for the coordination of the realization of this design.

Finally, last, but not least, I thank my father and mother for their unconditional support, love, ardor, care and sympathy. You are really the best parents I ever could wish. My gratitude to you is impossible to describe and stretches much and much beyond this Ph.D. research. With thanks to your love, I was able to reach this success in my life and to extend this in the future.

Curriculum Vitae

Jacob Pieter Laurens Brand was born on May 26, 1965 in Zaandam. He attended the Montessori Mavo in Amsterdam from 1979 till 1983, HAVO from 1983 till 1985 and VWO at the Montessori Lyceum in Amsterdam. In 1987, he started his study in mathematics at the Free University in Amsterdam with specialization in statistics and he received his degree in 1992. His final-term project concerns single imputation of entirely categorical data sets and the construction of the confidence interval for sumscores of dichotomous items in a questionnaire. In 1993, he started his PhD. research as described in this thesis at the Department of Medical Informatics at the Erasmus University in Rotterdam in coorperation with TNO-Prevention and Health in Leiden. Since September 1, 1998 he is working for the statistical software company Statistical Solutions Ltd. in Cork in Ireland.