Institut für Biometrie
Universität Ulm
Komm. Leiter: PD Dr. R. Muche

# Ersetzungsmethoden für fehlende Werte kategorialer Variablen in klinischen Datensätzen

Dissertation zur Erlangung des Doktorgrades der Humanbiologie
der Medizinischen Fakultät der Universität Ulm

Kathrin Hohl
geb. in Hamburg

Ulm 2007

| | |
|---|---|
| Amtierender Dekan: | Prof. Dr. K.-M. Debatin |
| 1. Berichterstatter: | PD Dr. R. Muche |
| 2. Berichterstatter: | Prof. Dr. U. Stadtmüller |
| Tag der Promotion: | 19.10.2007 |

# Inhaltsverzeichnis

**Abkürzungsverzeichnis** **iii**

**1 Einleitung** **1**

**2 Basiswissen fehlende Werte** **11**

2.1 Missing Data Diagnostic . . . . . 11

2.2 Klassifizierung von Ersetzungsmethoden . . . . . 17

2.3 Ersetzungsstrategien . . . . . 22

**3 Umgang mit fehlenden Werten** **32**

3.1 Ersetzungsmethoden für kategoriale Variablen . . . . . 32

3.2 Ersetzungs- vs. Auswertungsmodell . . . . . 45

3.3 Software zur Ersetzung . . . . . 49

3.4 Guidelines zu fehlenden Werten . . . . . 52

3.5 Sensitivitätsanalyse . . . . . 57

**4 Material und Methoden** **61**

4.1 Simulationsdatensatz . . . . . 64

4.2 Rahmenbedingungen der Simulation . . . . . 67

4.3 Parameter zum Vergleich der Ersetzungsmethoden . . . . . 72

4.4 Durchführung der Simulation . . . . . 90

**5 Ergebnisse** **96**

5.1 Vergleich der Übereinstimmung . . . . . 96

5.2 Vergleich der Validität . . . . . 108

**6 Diskussion** **121**

6.1 Interpretation der Simulationsergebnisse . . . . . 121

6.2 Limitierungen der Simulationsergebnisse . . . . . 131

6.3 Fazit für den Umgang mit fehlenden Werten in klinischen Datensätzen . . . . . 133

**7 Zusammenfassung** **137**

**Literaturverzeichnis** **139**

**Anhang** **151**

**Abbildungsverzeichnis** **153**

**Tabellenverzeichnis** **154**

# Abkürzungsverzeichnis

| | |
|---|---|
| fW | fehlende Werte |
| lin. Reg. | lineare Regression |
| log. Reg. | logistische Regression |
| ABB | Approximate Bayesian Bootstrap |
| BB | Bayesian Bootstrap |
| BMI | Body Mass Index |
| CCA | Complete Case Analysis |
| CPMP | Committee for Proprietary Medicinal Products |
| CRF | Case Report Form |
| DFM | Discriminant Function Method |
| ICH | International Conference on Harmonization |
| KI | Konfidenzintervall |
| MAR | Missing At Random |
| MCAR | Missing Completely At Random |
| MCMC | Markov Chain Monte Carlo |
| MDD | Missing Data Diagnostic |
| MDM | Missing Data Mechanism |
| MDP | Missing Data Pattern |
| MI | Multiple Imputation |
| MNAR | Missing Not At Random |
| MW | Mittlere Abweichung des Gesamtmittelwertes von dem wahren Mittelwert |
| RMSE | Root Mean Square Error |

| | |
|---|---|
| SA | Sensitivitätsanalyse |
| SE | Standardfehler |
| Sh | Steatosis hepatis |
| SI | Single Imputation |

# Kapitel 1

# Einleitung

Für eine erfolgreiche Forschung in der Medizin ist eine hohe Datenqualität der erhobenen Daten unabdingbar. Das heißt der auszuwertende Datensatz muss Informationen über alle für die jeweilige Fragestellung wichtigen Merkmale enthalten und die Daten müssen korrekt und vollständig sein. Je besser die Datenqualität ist, desto größer ist die Aussagekraft der Forschungsergebnisse. Aufgrund dieser Erkenntnis spielt die Datenqualität im Rahmen von klinischen Studien eine immer wichtigere Rolle. Es wird zunehmend größerer Aufwand betrieben, um die bestmögliche Datenqualität zu gewährleisten, nicht zuletzt, da dies von den internationalen Guidelines wie z.B. der Good Clinical Practice (GCP) [50] gefordert wird.

Zur Sicherung der Datenqualität wird meistens eine Doppeleingabe der Daten in elektronischer Form durchgeführt. Diese unterstützt eine korrekte Datenerfassung. Das gleiche Ziel hat die standardisierte Kodierung von Begleiterkrankungen und unerwünschten Ereignissen. Darüber hinaus werden nicht plausible Daten und fehlende Werte mit Hilfe eines Monitors der Studie und den Prüfzentren vor der teststatistischen Analyse geklärt.

Dennoch treten in den meisten klinischen Datensätzen fehlende Werte (fW) auf. Das Auftreten von fehlenden Werten kann zum Teil gar nicht verhindert

werden, etwa bei fehlerhaften Messinstrumenten, bei Drop-outs oder bei Ereignissen, die in der Vergangenheit liegen und nicht mehr rekonstruiert werden können.

Das Auftreten von fehlenden Werten in Datensätzen kann die Studienergebnisse auf verschiedene Arten beeinflussen und somit deren Aussagekraft stark einschränken. Was in dieser Arbeit unter dem Begriff fehlende Werte verstanden wird, wo diese auftreten, welche Probleme sie verursachen und wie diese z.T. gelöst werden können, wird in den nächsten Abschnitten kurz erläutert.

### Definition fehlender Werte

In dieser Arbeit wird gemäß [104] die folgende Definition für fW verwendet:

> Es liegt dann ein fehlender Wert vor, wenn der Wert theoretisch existiert, in dem betrachteten Datensatz aber wider Erwarten nicht vorliegt.

Diese Definition wird beispielhaft an der Variablen Rauchen verdeutlicht. Angenommen das Rauchverhalten der Patienten spielt eine wesentliche Rolle bei der Analyse einer medizinischen Fragestellung. Auf dem Erhebungsbogen, dem sogenannten Case Report Form (CRF), wird daher gefragt, ob der Patient raucht oder nicht. Für den Fall, dass der Patient raucht, wird zusätzlich die Menge der konsumierten Zigaretten erhoben. Wenn anschließend im Datensatz die Information von der Variablen Raucher/Nichtraucher fehlt, so liegt ein fehlender Wert gemäß obiger Definition vor. Fehlt allerdings die Angabe über den Zigarettenkonsum, so liegt nur bei den Rauchern ein fehlender Wert vor. Von den Nichtrauchern wird hierzu gar keine Angabe erwartet! Es ist folglich wichtig, den Grund für das Fehlen eines Wertes zu wissen, um entscheiden zu können, ob der Wert *wider* Erwarten fehlt oder nicht.

Es gibt viele verschiedene Bezeichnungen für „fehlende Werte“. In der Literatur werden fW auch als missing data, missing value, incomplete data, partial data, missing units, unit-nonresponse, item-nonresponse, intentional missing und

conditionally missing bezeichnet. Die Verwendung des jeweiligen Ausdrucks hängt stark von dem Forschungsumfeld und der Ursache des Fehlens ab. Die Bezeichnungen sollen hier als gleichwertig erachtet werden.

Der Umgang mit fehlenden Werten einer Variablen ist unter anderem abhängig vom Merkmalstyp der Variablen. In dieser Arbeit wird primär zwischen stetigen und kategorialen (diskreten) Variablen unterschieden. Eine stetige Variable kann jeden beliebigen Wert innerhalb eines Intervalls der reellen Zahlen annehmen. Beispiele für stetige Variablen sind Körpergröße, Blutdruck oder Insulinspiegel. Kategoriale Variablen hingegen nehmen nur eine begrenzte Anzahl an möglichen Werten an. Der kategoriale Merkmalstyp kann mit Hilfe des Skalenniveaus noch genauer spezifiziert werden in binär, nominal und ordinal. Wenn eine Variable nur zwei Werte annehmen kann (z.B. Geschlecht (männlich/weiblich)), so bezeichnet man sie als binär. Eine nominale Variable nimmt wenige Werte an und diese können nicht in eine Reihenfolge gebracht werden (z.B. Haarfarbe, Begleiterkrankung). Bei einer ordinalen Variablen haben die Merkmalsausprägungen eine eindeutige Reihenfolge oder Ordnung (z.B. Tumorstadium), wobei die Abstände zwischen den Merkmalsausprägungen unterschiedlich und meist auch gar nicht definiert sind.

In der Literatur sind Möglichkeiten für den Umgang mit fW stetiger Variablen in der statistischen Auswertung bereits eingehend diskutiert worden [63, 82, 83]. Deshalb konzentriert sich diese Arbeit auf den Umgang mit fehlenden Werten von kategorialen Variablen.

### Auftreten fehlender Werte kategorialer Variablen

Wichtige kategoriale Variablen, bei denen fehlende Werte auftreten, sind z.B. pathologische Laborwerte (ja/nein), Familienanamnese, aufgetretene Infektionen (ja/nein), Schulbildung (kein Abschluss, Sonderschulabschluss, Hauptschulabschluss/POS, Mittlere Reife, Abitur/Fachhochschulreife/FOS), Com-

pliance bei der angeordneten Medikamenteneinnahme oder in Klassen eingeteilte stetige Variablen jeglicher Art (Bluthochdruck, Diabetes, Cholesterin, Body Mass Index (BMI)).

Zusätzlich ist das Auftreten von fW kategorialer Variablen problematisch, wenn die Werte der Variablen Teil eines Summenscores sind. Beispielsweise wird der SF 12 (Instrument zur Erhebung gesundheitsbezogener Lebensqualität) als Summenscore basierend auf einer Reihe von kategorialen Variablen ausgewertet. Sobald eine der zugrunde gelegten kategorialen Variablen einen fehlenden Wert enthält, ist der Summenscore dieser Beobachtung nicht mit dem Summenscore anderer Beobachtungen, für die alle Werte beobachtet (erhoben) wurden, vergleichbar.

**Ursachen für fehlende Werte**

Die Gründe für das Auftreten von fehlenden Werten sind sehr vielfältig. Patienten, die gebeten werden einen Fragebogen auszufüllen, können leicht eine Frage übersehen oder absichtlich eine Frage unbeantwortet lassen. Eventuell ist Ihnen die Frage zu persönlich oder sie haben keine Motivation dazu, diese zu beantworten. Je länger der Fragebogen ist, desto geringer ist in der Regel die Motivation zur gewissenhaften Beantwortung aller Fragen. Eine fehlende Antwort kann aber auch daher rühren, dass der Patient aus intellektuellen oder sprachlichen Schwierigkeiten die Frage nicht verstanden und sie deshalb nicht beantwortet hat.

Fehlende Werte können auch vom klinischen Personal verursacht werden, wenn diese Personen den Erhebungsbogen (CRF) ausfüllen. Auch das Fachpersonal kann eine Frage auf dem CRF übersehen und diese deshalb unbeantwortet lassen. Ferner kommt es immer wieder vor, dass eine entnommene Probe nicht ausgewertet werden kann oder fehlerhafte Messinstrumente zu fehlenden Werten in der Diagnostik führen. In beiden Fällen können die Daten meistens nicht

nacherhoben werden.

### Probleme aufgrund von fehlenden Werten

Bei der Auswertung von Datensätzen mit teilweise fehlenden Werten beginnen die Probleme bereits beim Datenmanagement. Es muss darauf geachtet werden, dass die fehlenden Werte als solche für den jeweiligen Merkmalstyp der Variablen erkannt werden. Folglich sollte eine zusätzliche Kategorie eingeführt werden, die das Auftreten eines fW charakterisiert. In der Statistiksoftware SAS ist dies beispielsweise ein Punkt („."). Andernfalls würden die ursprünglich fehlenden Werte die Auswertungsergebnisse beeinflussen und zu nichtvaliden Ergebnissen führen.

Des Weiteren muss sichergestellt werden, dass, falls die Variable mit fW in einer Formel verwendet wird um eine abgeleitete Variable (z.B. BMI) zu generieren, die Formel von dem verwendeten Softwareprodukt korrekt implementiert wird. Das heißt, wenn in einer der in der Formel enthaltenen Variablen ein fW vorliegt, muss auch die daraus generierte Variable einen fW ausweisen. Vielfach verlassen sich die Anwender auf die implementierten Algorithmen, doch eine Untersuchung von Cordeiro [22] zeigt, dass in manchen Softwareprogammen, z.B. Minitab for Windows 14.20, die Algorithmen mit den fW nicht adäquat umgehen.

Die meisten Auswertungsmodelle, welche in den statistischen Softwareprodukten implementiert sind, basieren auf einer vollständigen Datenmatrix, d.h. zu jeder Variablen im Modell muss von jeder Beobachtung (Fall) ein Wert vorliegen. Fehlt bei einer Beobachtung der Wert von mindestens einer Variablen, so wird die gesamte Beobachtung in der Auswertung automatisch nicht berücksichtigt. Ein derartiges Vorgehen wird als *Complete Case Analysis* (CCA) bezeichnet. Die Durchführung einer CCA kann die Power von statistischen Analysen verringern, die Variabilität der Messwerte verändern und die Aussagekraft

der Studienergebnisse herabsetzten. Dennoch ist dies ein häufig angewandter Umgang mit fW zur Auswertung klinischer Datensätze [107].

Bei Vorgehen gemäß einer CCA verringert sich die Fallzahl z.T. drastisch. Zur Veranschaulichung wird ein Datensatz mit 25 Variablen betrachtet, wobei 3% der Werte je Variable zufällig fehlen. Angenommen, die fW treten gleichmäßig über den Datensatz verteilt auf. Dann werden bei einer CCA mehr als die Hälfte der Beobachtungen ($1 - 0.97^{25} = 0.53$) nicht berücksichtigt! Daher gilt: Je größer der Anteil an fW und die Anzahl an Variablen mit fW ist, desto größer ist die Wahrscheinlichkeit für einen Powerverlust bei statistischen Tests.

Patienten, die die Studie vorzeitig abbrechen, haben möglicherweise extremere Werte als die im Studienkollektiv verbliebenen. Therapiemisserfolg kann ebenso zu Drop-outs führen wie besonders gute Response, welche ebenfalls die Notwendigkeit der weiteren Teilnahme an der Studie hinfällig erscheinen lässt. Der Verlust dieser Drop-outs in der Auswertung kann daher zu einer Unterschätzung der Varianz führen und somit zu einem fälschlicherweise schmaleren Konfidenzintervall.

Die größte Problematik fehlender Werte ist die mögliche Verzerrung der Ergebnisse und somit Verringerung der Aussagekraft der Studie. Die Verzerrung kann sich auf die geschätzten Behandlungsunterschiede beziehen, die Vergleichbarkeit der Studienarme beeinflussen und die Repräsentativität des Auswertungskollektivs in Frage stellen (sogenannter Selektionsbias). Wenn beispielsweise alle Patienten mit einem geringen (keinem) Therapieerfolg in der Placebogruppe die Studie abbrechen und nur diejenigen in der Studie verbleiben, die sich zumindest teilweise verbessern, so kann der tatsächliche große Behandlungsunterschied bei einer CCA nicht festgestellt werden, da die Daten für den Behandlungsmisserfolg in der Auswertung nicht berücksichtigt werden.

Fehlende Werte führen vor allem dann zu nicht vergleichbaren Studienarmen

oder zu einem nichtrepräsentativen Auswertungskollektiv (im Vergleich zur Grundgesamtheit), wenn die fW systematisch auftreten. Die Aussagekraft der Ergebnisse ist in derartigen Situationen stark eingeschränkt.

Speziell in der Auswertung großer epidemiologischer Datensätze ist die Durchführung einer CCA aus zwei Gründen sehr problematisch. Zum einen enthalten epidemiologische Auswertungsmodelle in der Regel eine relativ große Anzahl von Einflussgrößen, um die Strukturgleichheit der primär interessierenden Risikogruppen zu sichern. Doch je mehr Variablen das Modell enthält, desto größer ist die Wahrscheinlichkeit, dass bei einer der Variablen ein fehlender Wert auftritt und somit die gesamte Beobachtung in der Auswertung nicht berücksichtigt wird. Mit zunehmender Anzahl an Einflussgrößen kann sich daher die Fallzahl reduzieren und folglich auch die Power.

Die zweite Problematik im Rahmen von epidemiologischen Datensätzen betrifft die Vergleichbarkeit der Ergebnisse von unterschiedlichen Analysen, die auf demselben ursprünglichen Datensatz basieren. Epidemiologische Datensätze werden oft von mehreren Forschern verwendet, um verschiedene Aspekte zu analysieren. Dazu werden unterschiedliche Variablen aus dem Datensatz in das jeweilige Auswertungsmodell mit aufgenommen. Wenn die Auswerter jeweils eine Complete Case Analyse durchführen, kann es daher passieren, dass die Ergebnisse auf unterschiedlichen Beobachtungen basieren und somit nicht mehr vergleichbar sind.

### Lösungen der Fehlenden-Werte-Problematik

Um die oben genannten Probleme von fehlenden Werten in der Datenmatrix zu umgehen, besteht die Möglichkeit, die fW zunächst mittels sogenannter Ersetzungsmethoden geeignet zu schätzen. Die Methoden versuchen die fehlenden Werte auf unterschiedliche Art und Weise zu schätzen, sodass dem Anwender anschließend eine vervollständigte Datenmatrix zur Auswertung zur Verfügung

steht. Diese Datenmatrix kann dann mit den üblichen statistischen Verfahren analysiert werden. Durch dieses Vorgehen werden alle Beobachtungen und die gesamten tatsächlich erhobenen Informationen berücksichtigt. Allerdings kann man nie sicher sein, ob der geschätzte Wert dem unbeobachteten tatsächlichen Wert entspricht. Es ist daher wichtig, diese Unsicherheit bei der Ersetzung bzw. in der Analyse zu berücksichtigen, da andernfalls die Ergebnisse wiederum verzerrt sein können. Diese Unsicherheit wird jedoch nur von wenigen Ersetzungsmethoden adäquat berücksichtigt.

In der Literatur exisiteren viele verschiedene Ersetzungsmethoden, um fW zu schätzen. Welche Ersetzungsmethode in einem speziellen Fall die Beste ist, hängt vom Grund für das Auftreten von fW, vom Ort der fW im Datensatz und vom Merkmalstyp (stetig/kategorial) der Variablen mit fW ab.

Wie bereits erwähnt, gibt es für den Umgang mit fW von stetigen Variablen (z.B. Körpergröße, Cholesterin) eine Anzahl von (anerkannten) Ersetzungsmethoden [63, 82, 83]. Diese Ersetzungsmethoden setzen vielfach voraus, dass die Variable mit fW normalverteilt ist. Bei Variablen, die diese Voraussetzung nicht erfüllen, können die Ersetzungsmethoden zu nicht validen Ergebnissen führen. Zudem erzeugen die Ersetzungmethoden für fW stetige Variablen zu ersetzende Werte aus einem breiten Intervall (z.B. den reellen Zahlen) und sind nicht begrenzt auf wenige plausible Ausprägungen, wie es bei kategorialen Variablen wünschenswert ist. Daher würde die Anwendung dieser Ersetzungsmethoden bei kategorialen Variablen (z.B. Tumorstadium) weder zu plausiblen noch zu validen Ergebnissen führen.

Spezielle Ersetzungsmethoden für kategoriale Variablen sind erst innerhalb des letzten Jahrzehnts entwickelt worden und daher erst in den letzten Jahren in manchen statistischen Softwarepaketen implementiert worden. Es gibt deshalb bis jetzt kaum Literatur über die Eigenschaften dieser Methoden. Ferner ist unklar, wie groß der Fehler ist, den man begeht, wenn man bei Vorliegen von

fW einer kategorialen Variablen leicht modifizierte Ersetzungsmethoden für stetige Variablen verwendet.

Das Ersetzen von fehlenden Werten ist nicht die einzige Lösung für die Fehlende-Werte-Problematik. Für die Auswertung von Longitudinalstudien gibt es statistische Methoden (z.B. Mixed Models oder die Überlebenszeitanalyse), die *keine* vollständige Datenmatrix voraussetzen. Um den Problemen einer CCA entgegenzuwirken, ist eine Ersetzung daher in diesen speziellen Situationen nicht nötig. Deshalb beschränkt sich diese Arbeit auf die Ersetzung von fW kategorialer Variablen bei Querschnittstudien, d.h. Studien, in denen die Untersuchungen nur einmal je Beobachtung/Fall durchgeführt werden.

**Ziele der Arbeit**

In der vorliegenden Arbeit werden folgende Fragestellungen untersucht:

1. Welche Ersetzungsmethoden für kategoriale Variablen gibt es?
2. Wie gut können ausgewählte Ersetzungsmethoden die fehlenden Werte kategorialer Variablen schätzen?
3. Welchen Einfluss hat die jeweilige Ersetzungsmethode auf die Ergebnisse der teststatistischen Analyse?
4. Welche Empfehlungen lassen sich aus diesen Kenntnissen für den Umgang mit fehlenden Werten kategorialer Variablen in klinischen Datensätzen ableiten?

Die Fragestellungen werden in den anschließenden Kapiteln wie folgt analysiert: In Kapitel 2 (Basiswissen fehlende Werte) wird zunächst die zugrunde liegende Theorie und Notation von fW erläutert. Dazu gehören auch allgemeine Hinweise für die Deskription und teststatische Auswertung bei Vorliegen von fW. Im 3. Kapitel werden Empfehlungen für den Umgang mit fehlenden Werten gegeben. Dies beinhaltet die Beschreibung verschiedener Ersetzungsmethoden für kategoriale Variablen, Angaben zur Wahl eines geeigneten

Ersetzungsmodells in Abhängigkeit vom Auswertungsmodell, welche Softwarepakete zur Schätzung fW verwendet werden können, eine Zusammenfassung von internationalen Guidelines bezüglich des Umgangs mit fW und eine kurze Erläuterung der Durchführung einer geeigneten Sensitivitätsanalyse im Kontext von fW.

Zur Auswertung der 2. und 3. Fragestellung wurden Simulationen mit der Statistiksoftware SAS durchgeführt. Der Aufbau der Simulationen wird in Kapitel 4 (Material und Methoden) beschrieben, während die Simulationsergebnisse in Kapitel 5 (Ergebnisse) dargestellt werden. In Kapitel 6 (Diskussion) werden die Ergebnisse interpretiert, die Verallgemeinerbarkeit der Ergebnisse diskutiert und daraus abgeleitete Empfehlungen für die Praxis zum Umgang mit fehlenden Werten kategorialer Variablen in klinischen Datensätzen gegeben. Eine Zusammenfassung der vorliegenden Arbeit wird in Kapitel 7 gegeben.

# Kapitel 2

# Basiswissen fehlende Werte

In diesem Kapitel werden grundlegende Aspekte im Kontext von fehlenden Werten vorgestellt. Dieses Basiswissen wird in den weiteren Kapiteln vorausgesetzt und ist daher für das Verständnis der Arbeit wichtig. Der erste Abschnitt befasst sich mit der Beschreibung der fW im Datensatz, welche Informationen notwendig sind, um die Aussagekraft der Studie zu beurteilen und welche Informationen bei der Wahl einer geeigneten Ersetzungsmethode benötigt werden. Im zweiten Abschnitt werden verschiedene Klassifikationen von Ersetzungsmethoden vorgestellt. Die Ersetzungsmethoden einer Klasse haben jeweils die gleichen Vor- und Nachteile bei der Schätzung der fW, sodass im folgenden Kapitel 3, in dem verschiedene Ersetzungsmethoden vorgestellt werden, nicht für jede einzelne die Vor- und Nachteile aufgelistet werden müssen. Die Schätzung der fW kann (abhängig von der gewählten Ersetzungsmethode) entweder zu einem oder mehreren vervollständigten Datensätzen führen. Der letzte Abschnitt dieses Kapitels beschreibt die Motivation und das Vorgehen zur Erzeugung mehrerer vervollständigter Datensätze.

## 2.1 Missing Data Diagnostic

Die Aussagekraft von Studienergebnissen basierend auf einem Datensatz mit (ursprünglich) fehlenden Werten hängt stark von den Ergebnissen der Missing

Data Diagnostic (MDD) ab. Ein Teil der MDD besteht aus der Beschreibung, bei welcher Variablen bzw. Beobachtung wie viele fW auftreten. Anhand dieser Ergebnisse können mögliche Fehler bei der Dateneingabe oder beim Datenmanagement erkannt werden, die sich eventuell korrigieren lassen.

Zusätzlich werden Unterschiede in der Zielgröße und den charakteristischen Eigenschaften zwischen Beobachtungen mit und ohne fW analysiert. Das bedeutet, es wird untersucht, ob fW vermehrt bei beispielsweise Alten, Männern oder Rauchern ect. auftreten.

Der andere Teil der Missing Data Diagnostic beschreibt die Anordnung der fW im Datensatz, dem so genannten Missing Data Pattern, und den (möglichen) Gründen für das Auftreten der fW, dem so genannten Missing Data Mechanism. Letzteres ist wichtig für die Wahl einer geeigneten Ersetzungsmethode.

### 2.1.1 Missing Data Pattern

Das Missing Data Pattern (MDP), auch bekannt als Missing Pattern oder Missing Value Pattern, beschreibt die globale Struktur der fW im Datensatz. Fasst man den Datensatz als Matrixform auf, wobei die Zeilen den einzelnen Beobachtungen/Fällen und die Spalten den $p$ betrachteten Variablen $X_1, \ldots, X_p$ entsprechen, so gibt es zwei Arten von Missing Data Pattern.

Bei einem *monotonen* MDP lassen sich die Variablen im Datensatz so anordnen, dass, wenn für eine Variable $X_j$ der Wert für diese Beobachtung nicht erhoben worden ist, auch die Werte der folgenden Variablen $X_{j+1}, \ldots, X_p$ fehlende Werte aufweisen. Dies ist grafisch in Abb. 1 (a) dargestellt, wobei mit „x“ die beobachteten Werte und mit „.“ die fehlenden Werte gekennzeichnet sind. Die einzelnen Zeilen stellen jeweils ein Pattern, ein Muster dar und repräsentieren die Gruppe von Beobachtungen, die in den gleichen Variablen beobachtete Werte aufweisen. Die detaillierte Beschreibung des MDP beinhaltet zusätzlich

zwei Spalten, in denen die absolute und relative Häufigkeit für das Auftreten der einzelnen Pattern angegeben wird. Ein Spezialfall des monotonen MDP ist das *univariate* MDP, bei dem nur in einer Variablen des Datensatzes fW auftreten [91].

| $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|
| x | x | x | x |
| x | x | x | . |
| x | x | . | . |
| x | . | . | . |

(a) monotones MDP

| $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|
| x | x | x | x |
| x | . | x | . |
| x | x | x | . |
| x | . | . | x |

(b) beliebiges MDP

Abb. 1: Missing Data Pattern

Ein monotones Missing Data Pattern tritt vor allem bei Longitudinalstudien auf, in denen die Werte zu manchen Beobachtungen zu einem späteren Zeitpunkt nicht mehr erhoben werden konnten. Eine andere denkbare Situation ist der Auswertungsdatensatz von (langen) Fragebögen. Möglicherweise hatten manche ausfüllende Personen zum Ende hin nicht mehr genügend Zeit oder ihre Motivation ließ nach, sodass die letzten Fragen unbeantwortet blieben.

Wenn die fehlenden Werte im Datensatz zufällig auftreten und die Variablen nicht in eine Reihenfolge gemäß eines monotonen MDP gestellt werden können, so bezeichnet man diese Anordnung der fW als *beliebiges* Missing Data Pattern, siehe Abb. 1 (b). Dieses MDP tritt in klinischen Datensätzen am häufigsten auf.

In den meisten großen Softwarepaketen, die Methoden zur Schätzung fW bereitstellen, siehe Abschnitt 3.3, kann das Missing Data Pattern automatisch erstellt werden. Allerdings kann die Ausgabe teilweise fehlerhaft sein. Pierchala [77] stellte fest, dass die Standardausgabe von SAS Version 9.1 nur numerische

Variablen zur Erstellung des MDP heranzieht und hat zur Behebung dieses Problems ein SAS Makro geschrieben.

### 2.1.2 Missing Data Mechanism

Annahmen über die Gründe für das Auftreten von fehlenden Werten werden üblicherweise als eine der 3 Arten des von Rubin [81] vorgestellten Missing Data Mechanism (MDM) formuliert. Da in den „modernen“ Ersetzungsmethoden das Auftreten der fW als Zufallsexperiment aufgefasst wird, wird der MDM in manchen Literaturstellen auch als *distribution of missingness* bezeichnet [91, 92]. Zur mathematischen Beschreibung der Mechanismen werden zunächst einige Notationen eingeführt.

Sei $\mathbf{Y}$ ein Datensatz, bestehend aus $N$ Zeilen, die die Beobachtungen repräsentieren und $p$ Spalten für die Variablen $X_1, \ldots, X_p$, so lässt sich der Datensatz partitionieren in

$$\mathbf{Y} = \{\mathbf{Y_b}, \mathbf{Y_f}\},$$

wobei $\mathbf{Y_b}$ den beobachteten Teil und $\mathbf{Y_f}$ den Teil mit fehlenden Werten bezeichnet. Zu beachten ist, dass in $\mathbf{Y}$ sowohl Zielgrößen als auch Einflussgrößen enthalten sein können. Zu jedem Wert von $\mathbf{Y}$ gibt es einen Indikator, der eine Zufallsvariable ist und dessen Realisierungen angeben, ob der Wert erhoben wurde (=1) oder nicht (=0). $\mathbf{R}$ sei die $N\mathtt{x}p$ Matrix bestehend aus den Indikatorvariablen für die einzelnen Werte in $\mathbf{Y}$. Die gemeinsame Verteilungsfunktion von $\mathbf{R}$ wird in der Literatur als *missingness mechanism* [2] oder *distribution of missingness* [91] bezeichnet.

#### MCAR

*Missing Completely At Random* (MCAR) ist der stringenteste MDM. Dieser liegt vor, wenn angenommen werden kann, dass die Wahrscheinlichkeit für das

Auftreten eines fW weder von einem der beobachteten noch von den unbeobachteten Werten abhängt, d.h. wenn gilt

$$P(\mathbf{R} \mid \mathbf{Y_b}, \mathbf{Y_f}) = P(\mathbf{R}).$$

Unter dieser Annahme bilden die beobachteten Werte einer Variablen eine repräsentative Stichprobe des Originaldatensatzes ohne fW für diese Variable. Für die Praxis bedeutet dies, dass unter MCAR die Auswertungergebnisse, die ausschließlich auf Beobachtungen mit vollständig erhobenen Werten (der so genannten Complete Case Analyse) basieren, zu validen Ergebnissen führen, wobei manche Information verloren geht.

MCAR liegt beispielsweise vor, wenn eine Materialprobe im Labor herunterfällt und somit der Wert für diese Beobachtung fehlt. In der Realität tritt MCAR äußerst selten auf, da nicht ausgeschlossen werden kann, dass die fW nicht doch von einer bestimmten Gegebenheit abhängen, wie das folgende Beispiel zeigt. Ein Patient einer klinischen Studie erscheint nicht mehr zur Nachuntersuchung, da er von einem Zug erfasst wurde. Wenn dies eine psychiatrische Studie ist, dann kann der Unfall möglicherweise nicht zufällig sondern eine Indikation für ein schlechtes Ansprechen der Behandlung sein und somit wäre die Vorraussetzung für MCAR verletzt.

Einige Autoren haben einen Test auf MCAR vorgestellt; Little [60] für den Datensatz bestehend aus nur stetigen Variablen und Park *et al.* [74] für den Datensatz, der nur kategoriale Variablen enthält. Aktuellere Literaturstellen über Tests auf MCAR sind in [13] enthalten.

### MAR

Wenn die Wahrscheinlichkeit für das Auftreten von fW nur von den tatsächlich beobachteten Werten nicht aber von den unbeobachteten abhängt, dann bezeichnet man dies als *Missing At Random* (MAR) [81]. Mathematisch lässt

sich dies ausdrücken durch

$$P(\mathbf{R} \mid \mathbf{Y_b}, \mathbf{Y_f}) = P(\mathbf{R} \mid \mathbf{Y_b}).$$

Anschaulich bedeutet es, dass Beobachtungen welche die gleichen Variablenwerte bei den beobachteten Werten aufweisen, auch bei den Variablen mit fW die gleichen statistischen Eigenschaften haben - aber nicht unbedingt die gleichen Werte!

Von Schafer *et al.* [91, 92] wird MAR auch als *ignorable nonresponse* bezeichnet, weil unter MAR der zugrundeliegende Missing Data Mechanism nicht explizit modelliert werden muss, d.h. ignoriert werden kann, um valide statistische Auswertungsergebnisse zu erzielen. Unter MAR sind per Definition alle Informationen über das Auftreten der fehlenden Werte in den beobachteten Werten enthalten, sodass die Berücksichtigung dieser Information in der statistischen Analyse ausreicht, um valide Ergebnisse zu erzielen. Ignorable nonresponse bedeutet aber nicht, dass die fW an sich ignoriert werden dürfen! Eine CCA unter MAR führt im allgemeinen zu nicht validen Ergebnissen. Ein Spezialfall von MAR ist, wenn die fW auch unabhängig von den beobachteten Werten auftreten, d.h. wenn MCAR vorliegt.

MAR liegt in klinischen Studien beispielsweise vor, wenn ein Patient aus der Studie herausgenommen wird, da sein Befinden nicht ausreichend gut kontrolliert wurde und dieses Kriterium vorab definiert war. Ein anderes Beispiel für MAR ist, wenn zwei Messwerte zur gleichen Zeit erhoben werden und in den Fällen, in denen sich diese Werte um mehr als einen bestimmten Wert unterscheiden, ein dritter Wert erhoben wird. Der dritte Messwert fehlt dann bei den Personen, bei denen die ersten beiden Werte nicht stark abwichen.

Viele Ersetzungsmethoden setzen das Vorliegen von MAR voraus. Bis jetzt gibt es aber noch keinen Test, der das Vorliegen von MAR bestätigt. Potthoff *et al.* [78] haben einen Test auf MAR+ vorgestellt. MAR+ impliziert MAR, sodass

wenn MAR+ nicht abgelehnt wird, man auch die Annahme, dass MAR gilt, gut vertreten kann. Wenn der Test allerdings darauf hindeutet, dass MAR+ nicht vorliegt, so ist auch das Vorliegen eines MAR fraglich. In derartigen Situationen müssen die Auswertungsergebnisse, die auf vervollständigten Datensätzen basieren, welche mittels einer Ersetzungsmethode generiert wurden, die MAR voraussetzt, noch vorsichtiger interpretiert werden.

#### MNAR

Ein Missing Data Mechanism wird als *Missing Not At Random* (MNAR) bezeichnet, wenn die Wahrscheinlichkeit für das Auftreten von fW abhängig von den unbeobachteten Werten ist. Dabei kann es sich bei den unbeobachteten Werten um die nichterhobenen Werte der betrachteten Variablen handeln oder auch um Werte von Variablen, die nicht im Datensatz enthalten sind [2]. MNAR wird auch als *nonignorable nonresponse* bezeichnet, da valide Auswertungsergebnisse nur unter einem gemeinsamen Modell von $\mathbf{Y}$ und $\mathbf{R}$, d.h. den beobachteten Daten und dem zugrundeliegenden Missing Data Mechanism, erzielt werden können.

Es sei auf die folgenden 2 Punkte hingewiesen. MAR und MNAR schließen sich gegenseitig aus. Ferner kann anhand der beobachteten Werte *allein* nicht beurteilt werden, ob MAR oder MNAR vorliegt.

## 2.2 Klassifizierung von Ersetzungsmethoden

Ersetzungsmethoden lassen sich abhängig von der Art, wie die fehlenden Werte geschätzt werden, klassifizieren. Innerhalb einer Klasse ist die Vorgehensweise zur Schätzung der fW ähnlich und die Ersetzungsmethoden haben die gleichen Vor- und Nachteile in Bezug auf die Güte der ersetzten Werte. Wenn die Klasse, der die Ersetzungsmethode zugeordnet wird, bekannt ist, kann daraus die Aussagekraft der darauf basierenden Auswertungsergebnisse abgeleitet werden. Zu

beachten ist, dass manche Ersetzungsmethoden mehreren Klassifizierungen zugeordnet werden können, da sich die unten aufgeführten Klassifizierungen nicht unbedingt gegenseitig ausschließen. In einem derartigen Fall müssen deshalb die Vor- und Nachteile der jeweiligen Klasse genau abgewogen werden.

### 2.2.1 Deterministische Ersetzungsmethoden

In deterministischen Ersetzungsmethoden erfolgt die Ersetzung von fW durch eine einfache, eindeutige Zuordnung eines festen Wertes. Die Zuordnung kann verschiedene Formen annehmen. Die ersetzten Werte können auf externen Informationen, Werten von anderen Beobachtungen aus der erhobenen Stichprobe zur gleichen Variablen oder auf den Werten von weiteren so genannten Hilfsvariablen im Datensatz basieren. Beispiele für deterministische Ersetzungsmethoden sind Cold deck oder Hot deck Ersetzung, siehe Abschnitt 3.1.2. Die Zuordnung ist eindeutig in dem Sinne, dass kein Zufall in der Bestimmung des ersetzten Wertes berücksichtigt wird, daher auch die Bezeichnung „deterministisch“.

Aufgrund der einfachen Durchführung verwenden viele Anwender deterministische Ersetzungsmethoden. Dies kann nicht empfohlen werden, da derartige Ersetzungsmethoden im Allgemeinen zu einer Unterschätzung der Varianz und einer verzerrten Verteilung der Variablen mit fW führen. Die Varianz kann aus zwei Gründen unterschätzt werden. Ein Grund ist, dass die ersetzten Werte je Variable zu wenig bzw. gar nicht streuen, die Fallzahl aber durch die Ersetzung zum Teil erheblich erhöht wird. Insgesamt betrachtet verringert sich somit die Varianz für die einzelne Variable mit ursprünglich fehlenden Werten. Diese Art von Varianzunterschätzung tritt vor allem bei stetigen Variablen auf. Bei kategorialen Variablen beeinflusst die deterministische Ersetzungsmethode die Häufigkeitsverteilung der einzelnen Ausprägungen, welche ebenfalls ein verzerrtes Bild des Kollektivs darstellen kann. In Abb. 2. ist beispielhaft für die Variable Stärke des Rauchens mit den drei Ausprägungen 0, 1 und 2 eine

mögliche deterministische Ersetzung der ursprünglich fW durchgeführt worden. Es wurde angenommen, dass die Information bezüglich dieses Merkmals fehlt, wenn die Person ein starker Raucher ist, d.h. Ausprägung 2 hat. Ersetzt man alle fW gemäß dieser Zuordnung, so ist der Anteil an starken Rauchern im vervollständigten Datensatz größer als die Anzahl von Personen mit den anderen beiden Ausprägungen. Dies ergibt ein anderes Bild, als das, welches das Kollektiv mit ursprünglich erhobenen Daten liefert.

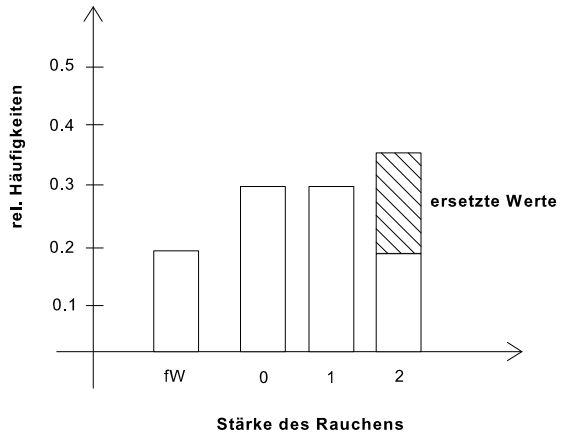


Abb. 2: Mögliche Verzerrung der Häufigkeitsverteilung durch deterministische Ersetzung.

Der zweite Grund für die Unterschätzung der Varianz besteht darin, dass die Unsicherheit über die Schätzung des ersetzten Wertes in der Auswertung nicht berücksichtigt wird [2]. Vielmehr wird der ersetzte Wert in der Auswertung als tatsächlich beobachtet aufgefasst.

Da die ersetzten Werte von deterministischen Ersetzungsmethoden geringfügig bzw. gar nicht streuen, beeinflussen sie mit zunehmendem Anteil an fehlenden Werten die Verteilung der Variablen. Die ersetzten Werte führen meist zu einer Verzerrung und können die Korrelation zwischen der Variablen mit fW und anderen Variablen in dem Datensatz verändern. In Abb. 3. ist dies beispielhaft dargestellt. Die Variable Stärke des Rauchens wird geschlechterspezifisch betrachtet. Ohne Berücksichtigung der ursprünglich fW ist der Anteil an Rau-

chern und starken Rauchern bei Frauen kleiner als bei den Männern. Unter der Annahme, dass die Personen, die keine Angabe bezüglich des Rauchens machen, starke Raucher sind, ist der Anteil an starken Rauchern unter Frauen nach der Ersetzung deutlich größer als unter Männern.

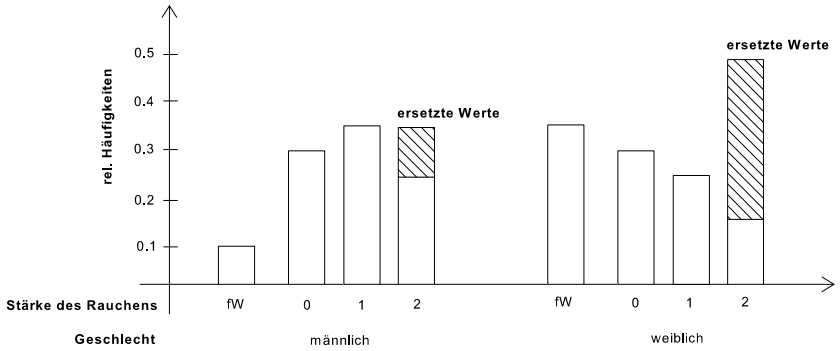


Abb. 3: Mögliche Verzerrung der Korrelation durch deterministische Ersetzung.

### 2.2.2 Stochastische Ersetzungsmethoden

Stochastische Ersetzungsmethoden zeichnen sich dadurch aus, dass sie jedem geschätzten Wert zur Ersetzung eines fehlenden Wertes einen gewissen *zufälligen* Korrekturterm hinzufügen. Beispielsweise ist die Ersetzung mittels einer Regressionsersetzung, siehe Abschnitt 3.1.5, eine stochastische Methode, wenn zu jedem geschätzten Wert noch ein zufälliges Residuum (d.h. Fehlerterm) der Regressionsgleichung addiert wird.

Die mittels stochastischen Ersetzungsmethoden generierten ersetzten Werte weisen eine höhere Streuung auf als die von deterministischen Ersetzungsmethoden. In der Regel wird die Varianz daher weniger stark unterschätzt, wobei theoretisch auch eine Überschätzung der Varianz möglich ist. Sofern ein fehlender Wert aber nur durch einen Wert ersetzt wird, kann die Unsicherheit über den unbeobachteten Wert in der anschließenden Auswertung nicht berücksichtigt werden. Aus diesem Grund kann auch bei stochastischen Ersetzungsmethoden die (Ko-)Varianz unterschätzt werden.

### 2.2.3 Modellbasierte Ersetzungsmethoden

Modellbasierte Ersetzungsmethoden bestimmen die ersetzten Werte anhand von Modellannahmen. Die Annahmen betreffen die theoretische Verteilung der Variablen mit fehlenden Werten und den vorliegenden Missing Data Mechanismus. Wenn die Modellannahmen richtig sind - in der Praxis sind diese leider selten nachprüfbar - erzeugen modellbasierte Ersetzungsmethoden präzisere Schätzungen der fehlenden Werte als die nichtmodellbasierten, stochastischen Ersetzungsmethoden [48]. Modellbasierte Methoden können deterministisch (z.B. Regressionsersetzung ohne zufälligen Korrekturterm) oder stochastisch sein (z.B. Regressionsersetzung mit zusätzlichem zufälligen Korrekturterm).

Wie schon bei den deterministischen und stochastischen Ersetzungsmethoden bemerkt, führen auch modellbasierte Ersetzungsmethoden zu einer Unterschätzung der Varianz und können die Verteilung der Daten verzerren, wenn ein fehlender Wert nur durch *einen* Wert ersetzt wird.

### 2.2.4 Bayesianische Ersetzungsmethoden

Eine weitere Gruppe von Ersetzungsmethoden basiert auf der Theorie der bayesianischen Statistik (Bayesian-theory-based imputation). Diese Methoden fügen nicht nur den geschätzten Werten eine Variabilität zu, sondern auch den Parametern des zugrunde gelegten Modells, anhand dessen die zu ersetzenden Werte vorhergesagt werden [48]. Die zur Berechnung der ersetzten Werte verwendeten Parameter werden hierbei zufällig aus der posterioren Verteilung der Parameter gewählt. Die posteriore Verteilung ist diejenige Verteilung, die aus der Vorabannahme an die Verteilung eines unbekannten Parameters (der a-priori-Verteilung) und den Ergebnissen des Experiments resultiert. Bei einer modellbasierten Ersetzungsmethode wird beispielsweise die a-priori-Verteilung der unbekannten Modellparameter um die Information der anhand der erhobenen Daten berechneten Schätzer von den Parametern ergänzt und stellt dann eine posteriore Verteilung der Modellparameter dar. Beispiele für derartige Er-

setzungsmethoden sind Approximate Bayesian Bootstrap (ABB) [59], Bayesian Bootstrap (BB) [48] und Data Augmentation bzw. MCMC Algorithmus, siehe Abschnitt 3.1.7. Für die theoretischen Hintergründe von ABB und BB sei auf [82] verwiesen.

## 2.3 Ersetzungsstrategien

Bei der Schätzung der fW kann gemäß unterschiedlichen Strategien vorgegangen werden. Die Strategien unterscheiden sich im Hinblick auf die Anzahl der vervollständigten Datensätze, d.h. auf die Anzahl der plausiblen Werte, durch die die ursprünglich fehlenden Werte ersetzt werden.

### 2.3.1 Single Imputation

Bei der Single Imputation (SI) wird jeder fehlende Wert durch *einen* plausiblen Wert ersetzt und somit wird nur ein vervollständigter Datensatz erzeugt. Beispielsweise führen alle deterministischen Ersetzungsmethoden eine Single Imputation durch.

Die Vorteile der SI gegenüber einer Complete Case Analyse (CCA) sind zahlreich. Die Complete Case Analyse berücksichtigt nur die Informationen von den vollständig erhobenen Beobachtungen und ist daher im Allgemeinen nicht so effizient wie die Analyse eines vervollständigten Datensatzes. In der Auswertung, welche auf einem vervollständigten Datensatz basiert, werden alle tatsächlich erhobenen (beobachteten) Werte von jeder im Datensatz enthaltenen Beobachtung berücksichtigt, sodass bei einer SI keine erhobene Information verloren geht. Zusätzlich können Ersetzungsmethoden zur Schätzung der fehlenden Werte Informationen aus den tatsächlich erhobenen Werten zur Vorhersage der fW miteinbeziehen. Somit kann eine gute Validität der Auswertungsergebnisse bewahrt werden [91]. Bei der CCA ist die Validität der Ergebnisse nur bei Vorliegen eines Missing Data Mechanism gemäß Missing

Completely At Random gewährleistet, da andernfalls die Beobachtungen mit vollständig erhobenen Werten keine repräsentative Stichprobe des Kollektivs bilden. Ferner bleibt bei der Anwendung von einer Ersetzungsmethode im Vergleich zu der CCA die ursprüngliche Fallzahl erhalten.

Ein weiterer Vorteil der Single Imputation ist die Vergleichbarkeit der Ergebnisse, weil alle Personen, die den (epidemiologischen) Datensatz oder Teile des Datensatzes auswerten, die gleichen Beobachtungen in der Auswertung berücksichtigen.

Die Single Imputation hat allerdings auch Schwachstellen. Wie schon im Abschnitt 2.2 herausgestellt, führt die Ersetzung durch nur *einen* plausiblen Wert zu einer Unterschätzung der Varianz, da die Unsicherheit über den unbeobachteten tatsächlichen Wert nicht berücksichtigt wird. Aufgrund dieser Unterschätzung wird bei statistischen Tests der p-Wert unterschätzt und die SI führt somit zu einer Erhöhung des Fehlers 1. Art. Dies bedeutet, dass die Nullhypothese fälschlicherweise öfter abgelehnt wird als durch das ursprüngliche Signifikanzniveau festgelegt wurde. Da bei der SI die Unsicherheit über die ersetzten Werte nicht berücksichtigt wird, können die Korrelationen zwischen Variablen mit fehlenden Werten stark beeinflusst werden und zu Fehlinterpretationen führen.

Trotz dieser Nachteile der Single Imputation gibt es Situationen, in denen eine derartige Ersetzungsstrategie angemessen ist [91]. Wie schon im Einleitungskapitel (Kapitel 1) an einem Beispiel mit einem Datensatz bestehend aus 25 Variablen von denen jeweils 3% der Werte fehlen veranschaulicht wurde, kann schon ein kleiner Anteil an fW je Variable bei einer großen Anzahl an Variablen zu einer drastischen Fallzahlreduktion und Informationsverlust führen. Der Einfluss der ersetzten Werte auf die Varianzen und Korrelationen ist jedoch bei einem kleinen Anteil an fW im allgemeinen vernachlässigbar gegenüber den Vorteilen einer Ersetzungsmethode. Es wird allerdings empfohlen, dass auch

bei Vorliegen eines kleinen Anteils an fW die am besten geeignete und nicht die einfachste Ersetzungsmethode durchgeführt wird!

### 2.3.2 Multiple Imputation

Die Idee der Multiple Imputation (MI) wurde in den siebziger Jahren von Rubin [81] entwickelt. Die ursprüngliche Motivation war die Bewältigung der Probleme mit fehlenden Werten in Datensätzen, welche in groß angelegten Studien erhoben und anschließend von mehreren Personen unter verschiedenen Gesichtspunkten ausgewertet werden sollten [83]. Mittels MI sollte sichergestellt werden, dass alle Analysen auf den gleichen Beobachtungen und den gleichen Werten basierten. Nach der erstmaligen Publikation blieb das Vorgehen von MI unklar und wurde nur von Experten angewandt. Dies änderte sich mit der Entwicklung schnellerer Computer und Software, die dem Anwender eine einfache Durchführung der MI ermöglicht, siehe Abschnitt 3.3.

Bei der Multiple Imputation wird ein fehlender Wert durch mehrere ($m > 1$) plausible Werte ersetzt, sodass $m$ vervollständigte Datensätze aus der Ersetzung resultieren[1]. Diese Datensätze werden einzeln mit der gleichen Auswertungsmethode und einem üblichen Softwareprogramm basierend auf einem komplett erhobenen Datensatz ausgewertet. Anschließend werden die Ergebnisse dieser Analysen zu gemeinsamen Schätzern und Standardfehlern zusammengefasst. Das Vorgehen der MI ist in Abb. 4 grafisch dargestellt.

Die MI hat die gleichen Vorteile wie eine SI. Zusätzlich modelliert MI aufgrund der Variation zwischen den $m$ ersetzten Werten, die Unsicherheit über die unbeobachteten tatsächlichen Werte. Bei korrekter Durchführung wird daher die Varianz nicht unterschätzt. Weil mittlerweile in vielen großen Softwarepaketen die Durchführung einer MI implementiert ist, ist der erforderliche Aufwand für

[1] Offensichtlich können nur stochastische Ersetzungsmethoden eine MI durchführen, da deterministische Methoden nur identische Datensätze erzeugen würden.

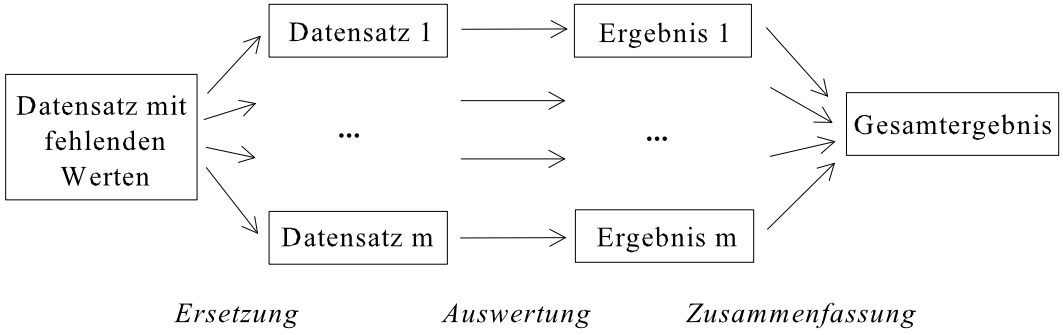


Abb. 4: Konzept einer Multiple Imputation

eine MI nur geringfügig größer als bei einer SI. Ferner ermöglicht eine MI, dass die Ersetzung der fW und die Auswertung der vervollständigten Datensätze von unabhängigen Personen bzw. Institutionen durchgeführt werden kann. Um valide Ergebnisse zu gewährleisten, sollte sich das verwendete Ersetzungsmodell, d.h. das Modell zur Schätzung der fW, am Auswertungsmodell orientieren und eher allgemeiner gehalten werden, d.h. mehr Variablen als das angestrebte Auswertungsmodell enthalten [20, 37, 90, 91]. Auf die Wahl eines geeigneten Ersetzungsmodells wird in Abschnitt 3.2 genauer eingegangen.

Ein weiterer Vorteil von MI ist, dass selbst bei einer kleinen Anzahl von 5 bis 10 vervollständigten Datensätzen die Schätzer asymptotisch effizient und präzise sind. Rubin [82] zeigte, dass sich die relative Effizienz des interessierenden Schätzers $Q$ (z.B. Mittelwert, Korrelation- oder Regressionskoeffizient) basierend auf $m$ vervollständigten Datensätzen im Vergleich zu unendlich vielen vervollständigten Datensätzen gemäß

$$\left(1 + \frac{\gamma}{m}\right)^{-\frac{1}{2}}$$

in % der Einheit der Standardabweichung verhält. Dabei bezeichnet $\gamma$ den Anteil an fehlender Information zur Schätzung von $Q$. Dieser Anteil drückt aus, um wie viel präziser der Schätzer wäre, wenn keine fehlenden Werte vorliegen würden. Angenommen $\gamma = 30\%$ der Information fehlt. Dann beträgt die Effizienz eines Schätzers, der auf $m = 3$ vervollständigten Datensätze basiert, nur

ca. 95% (in Einheiten der Standardabweichung) im Vergleich zu einem voll effizienten Schätzer, der auf unendlich vielen vervollständigten Datensätzen basiert. Die Tab. 1 gibt an, um wie viel sich die relative Effizienz des Schätzers $Q$ in Abhängigkeit von der Anzahl der vervollständigten Datensätze verändert, wenn der Anteil an fehlender Information variiert. Die in der Tabelle enthaltenen Werte sind in % der Einheit der Standardabweichung angegebenen.

Tab. 1: Relative Effizienz des Schätzers $Q$ basierend auf MI.

| | Anteil an fehlender Information $\gamma$ | | | | |
|---|---|---|---|---|---|
| m | 10% | 20% | 30% | 50% | 70% |
| 1 | 0.953 | 0.913 | 0.877 | 0.816 | 0.767 |
| 2 | 0.975 | 0.953 | 0.933 | 0.894 | 0.861 |
| 3 | 0.984 | 0.968 | 0.953 | 0.925 | 0.900 |
| 5 | 0.990 | 0.980 | 0.971 | 0.953 | 0.937 |
| 10 | 0.995 | 0.990 | 0.985 | 0.976 | 0.967 |
| 20 | 0.998 | 0.995 | 0.993 | 0.988 | 0.983 |

Nachdem die $m$ vervollständigten Datensätze generiert wurden, werden diese einzeln nach der gleichen Auswertungsmethode analysiert. Sei $Q$ der interessierende Parameter und $V$ dessen Varianz. Dann erzeugen die $m$ Auswertungen die Schätzer $\hat{Q}_1, \ldots, \hat{Q}_m$ mit den geschätzten Varianzen $\hat{V}_1, \ldots, \hat{V}_m$. Der im Endeffekt betrachtete gemeinsame Schätzer $\hat{Q}$ berechnet sich gemäß Rubin [82] aus dem Mittelwert der einzelnen Schätzer

$$\hat{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i.$$

Die zugehörige Varianz $\hat{V}$ setzt sich aus der Varianz zwischen den Ersetzungen (between-imputation variance)

$$B = \frac{1}{m-1} \sum_{i=1}^{m} \left( \hat{Q}_i - \hat{Q} \right)^2$$

und der Varianz innerhalb der Ersetzung (within-imputation variance)

$$W = \frac{1}{m} \sum_{i=1}^{m} \hat{V}_i$$

zusammen. Die insgesamt geschätzte Varianz $\hat{V}$ berechnet sich gemäß

$$\hat{V} = \left(1 + \frac{1}{m}\right) B + W.$$

Wenn die fehlenden Werte keine Information bzgl. des Parameters Q enthalten, so sind alle Schätzer $\hat{Q}_i$, $i = 1, \ldots, m$ identisch und die Gesamtvarianz $\hat{V}$ besteht lediglich aus dem Mittelwert der einzelnen Varianzen $\hat{V}_i$, $i = 1, \ldots, m$. Daher misst $r = \frac{(1+\frac{1}{m})B}{W}$ die relative Erhöhung der Varianz aufgrund von fehlender Information [90] und der Anteil an fehlender Information beträgt ungefähr $\gamma = \frac{r}{1+r}$.

Zur Ersetzung im Rahmen einer Multiple Imputation werden Informationen zum Ersetzungsmodell, dem Missing Data Mechanism und der a priori Verteilung der Modellparameter vorausgesetzt [95]. Die zugrunde liegende Theorie ist bayesianisch und wird im Folgenden erläutert. Eine detaillierte Beschreibung ist in [82], Kapitel 3 enthalten.

### Ersetzungsmodell

Die Validität der Ergebnisse von MI beruht auf der Durchführung der Ersetzungen und darauf, wie gut das Ersetzungsmodell mit dem Auswertungsmodell übereinstimmt. Sei $Y = (Y_b, Y_f)$ der komplette Datensatz bestehend aus den beobachteten Werten $Y_b$ und den fehlenden Werten $Y_f$. Dann erfolgt bei der *Single Imputation* die Ersetzung anhand der bedingten Verteilung

$$P(Y_f \mid Y_b, \hat{\theta}),$$

wobei $\hat{\theta}$ ein Schätzer der Modellparameter basierend auf den beobachteten Werten ist. Bei der MI hingegen wird der Parameter $\hat{\theta}$ nicht als Konstante,

sondern als Zufallsvariable aufgefasst. Daher werden bei der *Multiple Imputation* zunächst mehrere plausible Werte $\hat{\theta}_1, \ldots, \hat{\theta}_m$ simuliert. Die fW werden anschließend anhand der so genannten posterior Verteilung

$$P(Y_f \mid Y_b, \hat{\theta}_i),\ i = 1, \ldots, m$$

geschätzt. Das Ersetzungsmodell ist folglich wichtig, um eine Verteilung für die fehlenden Werte bedingt durch die beobachteten Werte zu finden.

Für gewöhnlich wird bei stetigen Variablen als Verteilung die multivariate Normalverteilung angenommen. Stetige Variablen, die nicht normalverteilt sind, sollten vor der MI zu normalverteilten Zufallsvariablen transformiert werden. Zur Auswertung müssen dann die beobachteten und ersetzten Werte zurück transformiert werden. Schafer [89] zeigte, dass selbst bei binären oder kategorialen Variablen die angenommene Normalverteilung zu akzeptablen Ergebnissen führen kann, wenn die ersetzten Werte anschließend auf die nächste beobachtete Ausprägung der jeweiligen Variablen gerundet werden. Die neueren Ersetzungsmethoden speziell für kategoriale Variablen basieren allerdings auf loglinearen Modellen.

Zu beachten ist, dass bei dem Ersetzungsmodell neben der Verteilungsannahme auch die enthaltenen Variablen unter Berücksichtigung des angenommenen Missing Data Mechanisms geeignet gewählt werden müssen. Empfehlungen zur Wahl des Ersetzungsmodells werden in Abschnitt 3.2 gegeben.

### Missing Data Mechanism

Um mittels MI valide Ergebnisse zu erzielen, muss ferner bei der Durchführung die Annahme über den vorliegenden MDM berücksichtigt werden. Die in der Software implementierten Methoden zur Durchführung einer MI setzen einen MDM gemäß MAR voraus. Dadurch vereinfacht sich der Algorithmus zur Bestimmung der $\hat{\theta}_1, \ldots, \hat{\theta}_m$. Die MI setzt aber nicht generell MAR voraus [90]! MAR gewährleistet, dass aufgrund der Abhängigkeitsstruktur im Datensatz

plausible Werte zur Ersetzung der fehlenden Werte anhand der beobachteten Daten vorhergesagt werden können, siehe Abschnitt 2.1.2. Falls MNAR vorliegt, kann MI zu falschen Ergebnissen führen [95].

Da die Modellierung von MNAR sehr komplizierte Berechnungen mit sich bringt, empfehlen Collins *et al.* [20] möglichst viele Merkmale zu erheben, die möglicherweise den Missing Data Mechanism gemäß MAR beschreiben, auch wenn nicht davon ausgegangen wird, dass diese Variablen einen direkten Einfluss auf die Zielgröße haben. Allerdings muss dieser vermutete Nutzen gegen den zusätzlichen Aufwand bei der Datenerhebung abgewogen werden. Aus statistischer Sicht muss zusätzlich die erhöhte Wahrscheinlichkeit für das Auftreten einer Multikollinearität durch die größere Anzahl an Einflussgrößen im Ersetzungsmodell beachtet werden. Als Multikollinearität wird die wechselseitige Abhängigkeit der Einflussgrößen im statistischen Modell bezeichnet. Diese kann zu hohen Standardfehlern der Modellschätzer führen und somit die Vorhersage der fehlenden Werte negativ beeinflussen.

### Prior Verteilung

Zusätzlich zu den Modellannahmen und dem MDM, verlangt die zugrundeliegende bayesianische Theorie der MI die Spezifizierung einer prior Verteilung der Modellparameter. Die unbekannten Parameter werden als Zufallsvariablen aufgefasst und mittels der prior Verteilung wird das Wissen des Anwenders über die Modellparameter vor der Datenerhebung quantifiziert [14]. Der Nachteil einer prior Verteilung ist die subjektive Wahl. Doch Sinharay *et al.* [95] weisen darauf hin, dass für viele Datenanalysen bei einer moderat großen Fallzahl jede sinnvolle prior Verteilung im Wesentlichen zu den gleichen Ergebnissen führt. Ferner scheinen in der Praxis die Ergebnisse stärker von dem gewählten Ersetzungsmodell als von der prior Verteilung abzuhängen [92].

Als prior Verteilung kann eine sogenannte informative oder nichtinformative

gewählt werden. Standardmäßig wird in der Software eine nichtinformative gewählt. Bei dieser Art von Verteilung ist über einen großen Wertebereich das Auftreten der einzelnen Werte gleich bzw. ähnlich wahrscheinlich, sodass die Verteilung kaum subjektiv, sondern primär von den erhobenen Werten beeinflusst wird [20]. Bei einer kleinen Fallzahl oder einem großen Anteil an fehlender Information kann es notwendig sein, eine informative prior Verteilung zu wählen. In einem solchen Fall ist es sinnvoll, verschiedene prior Verteilungen anzunehmen, um die Robustheit und die Validität der Ergebnisse beurteilen zu können.

Die prior Verteilung der Modellparameter wird unter Berücksichtigung der erhobenen Werte im Rahmen der MI korrigiert und führt somit zu der posterioren Verteilung. Aus dieser posterioren Verteilung werden dann Modellparameter simuliert, welche im Ersetzungsmodell zur Schätzung der fW enthalten sind. Die letzten beiden Schritte sind das Komplizierte an dem Vorgehen einer MI. Sie sind aber notwendig, um die Unsicherheit über die fW zu modellieren.

Eine Klasse von Methoden, welche die zufällige Wahl von Werten aus einer Verteilung via Markov Ketten simuliert, ist die Markov Chain Monte Carlo (MCMC). Eine Markov Kette ist eine Folge von Zufallsvariablen, deren Verteilung nur vom Wert der vorhergehenden Zufallsvariablen abhängt. Bei der Durchführung einer MCMC Methode wird eine Markov Kette generiert, bei der am Ende der Folge, d.h. am Ende der Iterationen, die Zufallsvariablen einer Iteration die gleiche Verteilung wie die Zufallsvariablen der vorhergehenden Iteration haben, d.h. eine stationäre Verteilung erreicht wurde. Diese Verteilung ist die interessierende posteriore Verteilung, aus der die Modellparameter zufällig gezogen werden.

Markov Ketten können auf vielerlei Arten konstruiert werden, sie sind aber alle Spezialfälle des Metropolis-Hastings Algorithmus [39, 67]. Eine umfassende Übersicht mit zugrunde liegender Theorie und Anwendungen von MCMC

Methoden ist im Buch von Gilks *et al.* [32] enthalten.

Zur Ersetzung der fW im Rahmen einer MI werden ausgehend von den kompletten Daten $Y = (Y_b, Y_f)$, dem Ersetzungsmodell $P(Y|\theta)$ und der prior Verteilung von $\theta$ die folgenden Schritte durchgeführt:

1. Gemäß einer SI Generierung der Startwerte von $Y_f^{(0)}$ und $\theta^{(0)}$.
2. Simulierung der fehlenden Werte $Y_f^{(t+1)} \sim P(Y_f|Y_b, \theta^{(t)})$.
3. Generierung der unbekannten Parameter $\theta^{(t+1)} \sim P(\theta|Y_b, Y_f^{(t+1)})$.
4. Wiederholung von 2. und 3. zur Generierung der Markov Kette

   $$(Y_f^{(1)}, \theta^{(1)}), (Y_f^{(2)}, \theta^{(2)}), \ldots, (Y_f^{(t)}, \theta^{(t)}).$$

   Anhand der stationären Verteilung wird der erste vervollständigte Datensatz erzeugt. Die Anzahl der Iterationen bis zu dieser Ersetzung werden auch als *burn-in* Iterationen bezeichnet. Bis zur Erzeugung des 2. vervollständigten Datensatzes werden bei einer *Single Chain Methode* eine gewisse Anzahl an weiteren Iterationen von Schritt 2 und 3 durchgeführt (überlicherweise zwischen 100 und 200 Iterationen). Bei einer *Parallel Chain Methode* werden zur Erzeugung des zweiten Datensatzes alle Schritte 1-4 erneut durchgeführt.

Trotz des anscheinend großen Aufwands bei der Durchführung einer Multiple Imputation wird dieses Vorgehen zur Schätzung fehlender Werte in der Literatur oft empfohlen [2, 6, 37, 45, 82, 91]. Da beispielsweise in SAS die einzelnen Schritte der Multiple Imputation bereits implementiert sind, ist der vom Anwender aufzubringende Aufwand nur geringfügig größer als bei der Durchführung einer Single Imputation.

Eine Multiple Imputation kann mit verschiedenen Ersetzungsmethoden durchgeführt werden. Die wesentlichen Methoden zur Ersetzung fehlender Werte kategorialer Variablen werden im folgenden Kapitel vorgestellt.

# Kapitel 3

# Umgang mit fehlenden Werten

In diesem Kapitel werden Grundlagen für den Umgang mit fehlenden Werten vermittelt. Im Abschnitt 3.1 werden mögliche Ersetzungsmethoden für kategoriale Variablen vorgestellt und im zweiten Abschnitt Empfehlungen für die Wahl eines geeigneten Ersetzungsmodells unter Berücksichtigung des angestrebten Auswertungsmodells gegeben. Abschnitt 3.3 stellt einige bekannte Softwarepakete und deren implementierte Ersetzungsmethoden vor. In Abschnitt 3.4 wird eine Zusammenfassung internationaler Guidelines für den Umgang mit fW präsentiert. Abschließend werden im Abschnitt 3.5 verschiedene Vorgehensweisen zur Durchführung einer Sensitivitätsanalyse vorgestellt, anhand derer die Validität der Auswertungsergebnisse basierend auf Datensätzen mit ursprünglich fW beurteilt werden kann.

## 3.1 Ersetzungsmethoden für kategoriale Variablen

In den folgenden Abschnitten werden oft angewandte Ersetzungsmethoden für kategoriale Variablen vorgestellt. Es werden jeweils die Voraussetzungen der Methode, die Vorgehensweise zur Ersetzung der fehlenden Werte und die wesentlichen Vor- und Nachteile aufgeführt.

### 3.1.1 Mittelwert- bzw. Modusersetzung

Die Mittelwert- bzw. Modusersetzung sind die einfachsten und daher die am häufigsten verwendeten Ersetzungsmethoden, die im Rahmen einer Single Imputation durchgeführt werden [75]. Die Methoden setzten einen Missing Data Mechanism gemäß MCAR voraus und können bei einem monotonen und beliebigen Missing Data Pattern durchgeführt werden.

Die Mittelwertersetzung ist zur Ersetzung stetiger Variablen entwickelt worden. Hierbei wird der Mittelwert aller für diese Variable beobachteten Werte als Schätzwert für die fW verwendet. Die Methode kann auch bei kategorialen Variablen angewandt werden, wobei der Mittelwert auf die nächstliegende Ausprägung der Variablen gerundet werden sollte. Bei kategorialen Variablen ist allerdings die Modusersetzung üblicher. Das heißt es wird von den beobachteten Werten der am häufigsten aufgetretene Wert ermittelt und die fW werden durch diesen ersetzt. Teilweise werden auch nur die beobachteten Werte von bestimmten Untergruppen (z.B. geschlechtsspezifisch) zur Berechnung des Mittelwertes bzw. Modus verwendet. Dies bezeichnet man als bedingte Mittelwert- bzw. Modusersetzung (conditional imputation) [33].

Aufgrund der Einfachheit sind die beiden Ersetzungsmethoden ein beliebtes Vorgehen zur Ersetzung fW von Teilfragen eines Fragebogens (z.B. des SF-12 oder APACHE II score), damit der zur Interpretation verwendete Summenscore auf der gleichen Anzahl an Fragen für alle Beobachtungen basiert [75, 76].

Die Motivation, fW durch den Mittelwert bzw. Modus zu ersetzten, ist, dass unter MCAR Beobachtungen mit fW eine repräsentative Stichprobe des betrachteten Kollektivs darstellen, sodass der Mittelwert bzw. Modus ein guter Schätzer für den Erwartungswert der fW ist. Da aber auch bei Vorliegen von MCAR durch diese Art von Ersetzung keine weitere Information dem Datensatz hinzugefügt, sondern nur die Fallzahl erhöht wird, können die Schätzer

für die Ko-Varianzen stark verzerrt sein [35, 63]. Diese deterministischen Ersetzungsmethoden, die nur eine Single Imputation durchführen können, sind daher selbst bei einem kleinen Anteil an fW nicht zu empfehlen.

### 3.1.2 Hot deck

Die Hot deck Ersetzung setzt voraus, dass die fehlenden Werte gemäß MAR aufgetreten sind. Sie kann bei einem monotonen und beliebigen Missing Data Pattern durchgeführt werden, wenngleich die Durchführung bei Vorliegen eines monotonen MDP unkomplizierter ist.

Bei dieser Methode wird der fW einer Variablen durch den aus einem Spenderdatensatz, welcher im Englischen auch als „deck“ bezeichnet wird, zufällig gezogenen Wert ersetzt. Der Spenderdatensatz besteht aus Beobachtungen, die ebenfalls im vorliegenden Datenansatz enthalten sind (deshalb der Begriff „hot“). Sie weisen ähnliche Charakteristiken wie die Beobachtung mit fW auf, haben aber für die interessierende Variable einen beobachteten Wert [82]. Die Wahl der Variablen, welche die Ähnlichkeit der Beobachtungen beurteilen und wie stark die Ausprägungen dieser Variablen von der Beobachtung mit fW abweichen können und dennoch als ähnlich bezeichnet werden, ist subjektiv wählbar und kann nicht standardisiert vorgegeben werden.

Die Hot deck Ersetzung wurde ursprünglich als eine Verbesserung der Mittelwert- bzw. Modusersetzung erachtet. Da die ersetzten Werte theoretisch alle Werte der tatsächlich erhobenen Werte annehmen können, wird eine größere Variabilität in den ersetzten Werten erzeugt und führt somit zu einem realistischeren Datensatz mit besseren Verteilungseigenschaften [75]. Ein weiterer großer Vorteil dieser nichtparametrischen Methode ist, dass sie keine Verteilungsannahmen voraussetzt. Doch in der Praxis überwiegen die Nachteile.

Die ersetzten Werte hängen stark von den nicht standardisiert wählbaren Va-

riablen ab, die den Spenderdatensatz definieren. Je mehr Variablen zur Erstellung des Spenderdatensatzes herangezogen werden, desto weniger fehlende Werte sollten im Datensatz auftreten, da sonst der Spenderdatensatz möglicherweise keine Beobachtungen enthält. Daher ist es einfacher, eine Hot deck Ersetzung bei einem monotonen anstatt einem beliebigen MDP durchzuführen. Sinharay *et al.* [95] weisen darauf hin, dass die Hot deck Methode zwar die Verteilung der einzelnen Variablen bewahrt, aber die Korrelation zwischen den Variablen im Datensatz zerstören kann. Ferner erfordert die Hot deck Ersetzung einen relativ großen programmiertechnischen Aufwand, da für jede Variable mit fW explizit ein Spenderdatensatz definiert werden muss und diese Methode bisher noch in keiner Software standardmäßig implementiert ist.

### 3.1.3 Cold deck

Die Cold deck Ersetzungsmethode ist universell anwendbar. Ähnlich wie die Hot deck Methode kann sie fW beliebigen Merkmalstyps ersetzen. Sie hat keine Voraussetzungen bezüglich des Missing Data Mechanism und des Missing Data Patterns.

Bei der Cold deck Methode wird der fW einer Variablen durch einen Wert aus einer externen (deshalb „cold“) Bezugsquelle derselben Population ersetzt (z.B. von einer vorhergehenden Untersuchung). Die fehlenden Werte einer Variablen können entweder alle durch den gleichen Wert ersetzt werden, oder ähnlich wie bei der Hot deck Methode, abhängig von den Werten anderer Hilfsvariablen durch möglicherweise verschiedene Werte ersetzt werden.

Wenn dem Analysten vergleichbare externe Daten vorliegen (was in der Praxis selten zutrifft), ist die Durchführung einer Cold deck Methode recht einfach. Little *et al.* [63] weisen darauf hin, dass auch diese deterministische Single Imputation Ersetzungsmethode die Unsicherheit über die ersetzten Werte nicht berücksichtigt und die Varianz dadurch unterschätzt wird.

### 3.1.4 Worst/ Best Case

Die Worst bzw. Best Case Ersetzungsmethode ist eine deterministische Ersetzungsmethode, die bei jedem Merkmalstyp angewandt werden kann und keinen bestimmten Missing Data Mechanism oder Missing Data Pattern voraussetzt.

Die beiden Ersetzungsmethoden wurden für Situationen entwickelt, in denen der Unterschied zwischen zwei Gruppen teststatistisch untersucht werden soll. Um konservativ zu testen, wird oft die Worst Case Methode durchgeführt. Hierbei werden die fehlenden Werte innerhalb der Gruppe mit den vermuteten besseren Werten durch den schlechtesten beobachteten Wert dieser Gruppe ersetzt. Die fehlenden Werte der anderen Gruppe hingegen werden durch den besten beobachteten Wert der Gruppe ersetzt. Dadurch wird der mögliche Unterschied zwischen den Gruppen absichtlich verkleinert und ein statistisch signifikanter Unterschied schwieriger nachweisbar.

Bei der Best Case Methode ist es genau umgekehrt. Fehlende Werte der Gruppe mit den vermutlich besseren Werten werden durch den besten beobachteten Wert ersetzt, während fehlende Werte der anderen Gruppe durch den schlechtesten beobachteten Wert innerhalb dieser Gruppe ersetzt werden.

Beide Methoden sind stark von den beobachteten Werten abhängig, zerstören die Korrelation zwischen den Variablen und berücksichtigen nicht die Unsicherheit in den ersetzten Werten. Diese Methoden werden in der Praxis oft im Rahmen von Sensitivitätsanalysen verwendet [101]. Doch selbst dafür ist deren Eignung fraglich, siehe Abschnitt 3.5.

### 3.1.5 Lineare Regression

Die lineare Regressionsmethode dient primär zur Ersetzung von fW stetiger Variablen. Sie setzt eine multivariate Normalverteilung, ein MDM gemäß MAR und ein monotones Missing Data Pattern voraus. Nichtnormalverteilte Varia-

blen müssen vor der Ersetzung geeignet transformiert werden. Schafer [89] weist darauf hin, dass die Methode in speziellen Situationen auch bei kategorialen Variablen angemessen ist, wenn anschließend die geschätzten Werte auf den dichtesten beobachteten Wert gerundet werden. Horton *et al.* [47] und Ake [1] haben mit Hilfe von Simulationsstudien allerdings gezeigt, dass bei binären Variablen und nominalen Variablen, die durch binäre Variablen im Ersetzungsmodell dargestellt werden, die lineare Regressionsmethode zu verzerrten Schätzern des Erwartungswertes führen kann.

Die Regressionsmethode schätzt die fW einer Variablen mittels eines linearen Regressionsmodells. Die abhängige Variable im Regressionsmodell ist die Variable mit teilweise fehlenden Werten und die unabhängigen Variablen sind vollständig beobachtete Variablen, von denen ein (deutlicher) Zusammenhang mit der Variablen mit fW angenommen wird. Basierend auf den beobachteten Werten wird eine Regressionsgleichung geschätzt, anhand derer die unbeobachteten Werte vorhergesagt werden. Befinden sich im Datensatz mit monotonem MDP mehrere Variablen mit fW, so wird die Regressionsmethode iterativ durchgeführt, wobei die vollständig beobachteten Variablen und Variablen mit bereits ersetzten Werten als Kovariablen im Modell enthalten sind.

Die Regressionsmethode besteht daher aus den folgenden Schritten. Der Datensatz mit monotonem MDP bestehe beispielsweise aus den drei Variablen $X_1, X_2$ und $X_3$, wobei $X_2$ und $X_3$ teilweise fW aufweisen.

1. Schätzung der fW von $X_2$ gemäß $\hat{X}_2 = \hat{\beta}_0 + \hat{\beta}_1 X_1$.
2. Schätzung der fW von $X_3$ durch $\hat{X}_3 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2^*$,
   wobei $X_2^*$ sowohl beobachtete als auch ersetzte Werte enthält.

Diese Art der Regressionsmethode ist deterministisch und kann daher nur eine Single Imputation durchführen. In den Softwareprogrammen ist allerdings überwiegend eine stochastische Regressionsmethode implementiert. Diese addiert zu jedem vorhergesagten Wert einen stochastischen (Korrektur-) Term

wie beispielsweise ein zufälliges Residuum der Regressionsgleichung. Damit vergrößert sich die Streuung der ersetzten Werte und es ist möglich, mehrere unterschiedliche vervollständigte Datensätze zu generieren, d.h. eine Multiple Imputation durchzuführen.

Auch wenn die Regressionsmethode in der Theorie eine sehr plausible und gute Ersetzungsmethode ist, hat sie einige Nachteile. Sie kann nur bei einem monotonen Missing Data Pattern durchgeführt werden, welches in der Praxis äußerst selten vorliegt. Ferner führt sie zu verzerrten Schätzern vor allem von den Ko-Varianzen [34]. Dabei hängt die Stärke der Verzerrung von der Eignung des Regressionsmodells zur Vorhersage der fW ab.

Um eine gemäß Rubin [82] korrekte Multiple Imputation mit Hilfe der stochastischen Regressionsmethode durchzuführen, müssen die $m$ vorhergesagten Werte je fehlendem Wert auf Regressionsmodellen mit verschiedenen Regressionskoeffizienten basieren. Dies kann dadurch erreicht werden, dass die Koeffizienten für jede Schätzung aus deren posterioren Verteilung generiert werden, vgl. Abschnitt 2.3.2.

### 3.1.6 EM Algorithmus

Der Expectation-Maximization (EM) Algorithmus ist ein allgemeines Verfahren, um Maximum-Likelihood-Schätzer von parameterischen Modellen basierend auf Datensätzen mit teilweise fW zu bestimmen [24]. Der in der Software implementierte EM Algorithmus im Kontext von fW setzt multivariat normalverteilte Daten und ein MDM gemäß MAR voraus. Im Gegenteil zur Regressionsmethode kann der EM Algorithmus sowohl bei einem monotonen als auch beliebigen Missing Data Pattern angewendet werden.

Der EM Algorithmus ist ein iterativer Prozess bestehend aus zwei Schritten. Im **Expectation Schritt** werden die fW gemäß einer bestimmten Funktion

geschätzt. Unter der Annahme einer multivariaten Normalverteilung ist dies im Wesentlichen eine Ersetzung gemäß der Regressionsmethode [2]. Im **Maximization Schritt** werden basierend auf dem vervollständigten Datensatz die Parameter wie Mittelwert, Kovarianzen und Regressionsmodellparameter neu berechnet. Mit Hilfe dieser neuen Parameterschätzer werden die ursprünglich fW erneut ersetzt, d.h. der Expectation Schritt wird erneut durchgeführt. Dieser iterative Prozess wird solange fortgeführt, bis die Parameterschätzer konvergieren, d.h. sie verändern sich nur minimal von Iteration zu Iteration. Die endgültigen Parameterschätzer sind Maximum-Likelihood-Schätzer. Eine detaillierte Beschreibung des EM Algorithmus ist in [63, 89] gegeben.

Der EM Algorithmus hat den gleichen Nachteil wie die Regressionsersetzung. Er ist eigentlich nur für stetige Variablen geeignet. Da er aber mit einer vollen Kovarianzmatrix startet und alle im Datensatz vorhandenen Variablen als abhängige Variablen im Regressionsmodell verwendet, kann er auch bei einem beliebigen MDP angewendet werden [2]. In manchen Statistikprogrammen, z.B. SPSS, wird der EM Algorithmus zur Ersetzung fW verwendet. In anderen Programmen, z.B. SAS oder S-Plus, dient er primär zur Generierung von Startwerten für andere Ersetzungsmethoden, wie z.B. die MCMC Methode.

### 3.1.7 MCMC

Markov Chain Monte Carlo (MCMC) Methoden dienen dazu, pseudo-zufällige Werte aus einer multidimensionalen Verteilungsfunktion mittels Markov-Ketten zu generieren. Eine Markov-Kette ist eine Folge von Zufallsvariablen, bei denen die Verteilung einer Zufallsvariablen abhängig vom Wert der vorhergehenden Zufallsvariablen ist. Eine spezielle MCMC Methode ist der Data Augmentation (DA) Algorithmus von Tanner und Wong [99]. Dieser Algorithmus wird vielfach zur Schätzung von fehlenden Werten eingesetzt. In der für diese Arbeit verwendeten Software SAS wird diese Ersetzungsmethode als MCMC Methode bezeichnet. Die Methode setzt ein MDM gemäß MAR vor-

aus und kann bei einem monotonem und beliebigem MDP angewendet werden.

Der DA Algorithmus ist ähnlich wie der EM Algorithmus ein iterativer Prozess bestehend aus zwei Schritten (siehe Abb. 5), wobei der DA Algorithmus ein stochastischer Prozess ist [89]. Bezeichne $Y = (Y_b, Y_f)$ den kompletten Datensatz partitioniert in den Teil $Y_b$ mit den beobachteten Werten und den Teil $Y_f$ mit fW. Mit $\theta$ sei der Parametervektor der Verteilungsfunktion $P(\theta|Y_b, Y_f)$ bezeichnet. Dann werden in dem **Imputation (I-) Schritt** basierend auf der derzeitigen Annahme von $\theta^{(t)}$ mittels der bedingten Verteilung $P(Y_f|Y_b, \theta^{(t)})$ ersetzte Werte für die fW $Y_f^{(t+1)}$ simuliert. In dem folgenden **Posterior (P-) Schritt** wird dann ein neuer Wert $\theta^{(t+1)}$ gemäß der posterioren Verteilung $P(\theta|Y_b, Y_f^{(t+1)})$ generiert. Diese beiden Schritte werden iterativ durchgeführt und führen zu einer Markov-Kette $\{\theta^{(t)}, Y_f^{(t)} : t = 1, 2, \ldots\}$, welche zu einer stationären Verteilung konvergiert, d.h. die *Verteilungen* der Parameter unterscheiden sich von Iteration zu Iteration nur noch minimal. Anhand der stationären Verteilung erfolgt die endgültige Ersetzung der ursprünglich fW.

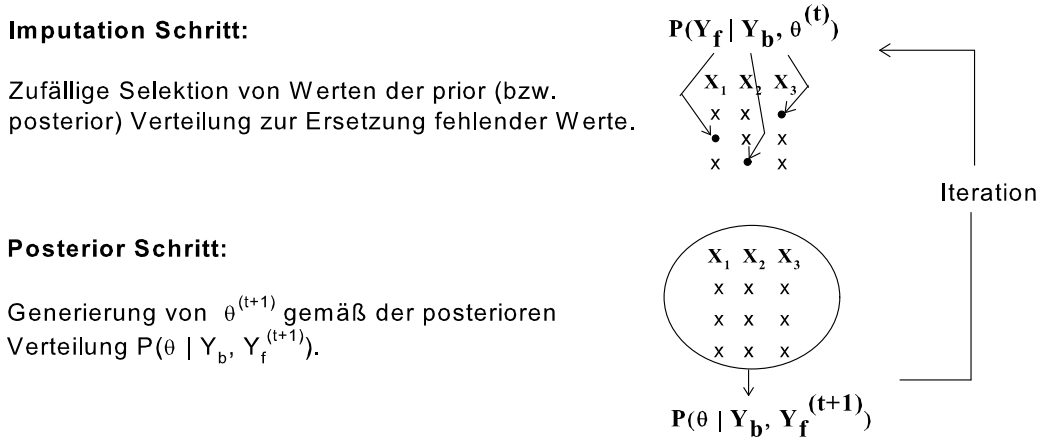


Abb. 5: MCMC Methode

Als Startwert der Parameter werden die Maximum-Likelihood-Schätzer vom EM Algorithmus empfohlen [2]. Je mehr Iterationen durchgeführt werden, de-

sto weniger hängen die simulierten Parameter von den ursprünglichen Parametern ab, so dass nach einer gewissen Anzahl $k$ an Iterationen die Parameter der $k$-ten Iteration und die ursprünglichen Parameter stochastisch unabhängig sind. Deshalb schlägt Schafer [89] als Konvergenzkriterium des Algorithmus die fehlende stochastische Abhängigkeit zwischen Parametern der Iteration $t$ und der Iteration $t + k$ vor.

Bei der Durchführung einer MI können die $m$ vervollständigten Datensätze entweder durch parallel laufende $m$ einzelne Markov-Ketten (*Parallel Chains*) oder anhand von einer langen Markov-Kette (*Sequential Chain*) generiert werden [92]. Angenommen, der DA Algorithmus konvergiert nach $k$ Iterationen. Dann wird bei der Parallel Chains Methode der DA $m$-mal durchgeführt und nach $k$ Iterationen werden die vervollständigten Datensätze für die weiteren Auswertungen abgespeichert. Hierbei ist es wichtig, dass die einzelnen Markov-Ketten nicht auf den gleichen Startwerten basieren [82]. Bei der Sequential Chain Methode hingegen werden $k*m$ Iterationen vom DA Algorithmus durchgeführt, bei denen nach jeweils $k$ Iterationen die vervollständigten Datensätze abgespeichert werden. Teilweise wird die Anzahl der Iterationen bis zum Abspeichern des ersten vervollständigten Datensatzes deutlich erhöht und als „burn-in period“ bezeichnet, um sicher zu gehen, dass der Prozess zur stationären Verteilung konvergiert ist [34].

### 3.1.8 Discriminant Function Method

Die Discriminant Function Method (DFM) ist speziell für die Ersetzung von nominalen Variablen geeignet. Sie setzt ein Missing Data Mechanism gemäß MAR und ein monotones MDP voraus. Ferner wird angenommen, dass die Kovariablen im Modell der Diskriminanzanalyse normalverteilt sind. Entwickelt wurde die Methode von Brand [7].

Sei $Y$ eine nominale Variable mit den Ausprägungen $1, \ldots, g$ und teilweise

fW. Die Variablen $X_1, \ldots, X_k$ seien stetige Variablen, welche innerhalb der einzelnen Ausprägungen von $Y$ multivariat normalverteilt sind und die Kovarianzen innerhalb der Ausprägungen seien ungefähr gleich. Die Informationen über die Mittelwerte, die gepoolte Kovarianzmatrix, die auf den einzelnen Kovarianzen innerhalb der Ausprägungen von $Y$ basiert, und den prioren Wahrscheinlichkeiten für das Auftreten der einzelnen Ausprägungen von $Y$ werden zur Erstellung linearer Diskriminanzfunktionen verwendet. Anhand dieser Funktionen werden die posterioren Wahrscheinlichkeiten für das Auftreten der jeweiligen Ausprägung von $Y$ berechnet. Für diese Wahrscheinlichkeiten gilt $p_1 + \cdots + p_g = 1$. Anschließend wird eine zwischen 0 und 1 gleichverteilte Zufallsvariable $u$ generiert. Falls $u < p_1$, so wird der fW der Beobachtung durch die erste Ausprägung ersetzt. Für $p_1 < u < p_1 + p_2$ entspricht der ersetzte Wert der 2. Auspägung von $Y$ usw.

Mit dieser stochastischen Ersetzungsmethode kann auch eine Multiple Imputation durchgeführt werden. Die wesentlichen Nachteile der Discriminant Function Method sind die Voraussetzung eines (unrealistischen) monotonen MDP und, dass nur stetige normalverteilte und keine kategorialen Variablen zur Schätzung der fW berücksichtigt werden können.

### 3.1.9 Logistische Regression

Rubin [82] hat die logistische Regressionsmethode zur Ersetzung kategorialer (binärer oder ordinaler) Variablen vorgeschlagen. Die Methode setzt Missing At Random und ein monotones MDP voraus.

Zur Schätzung der fW der kategorialen Variablen wird ein logistisches Regressionsmodell verwendet. Die Variable $Y$ mit teilweise fW ist die abhängige Variable und weitere vollständig beobachtete Variablen $X_1, \ldots, X_k$ im Datensatz dienen als Kovariablen im Modell. Für eine binäre 0,1 kodierte Variable beinhaltet die Ersetzungsmethode folgende Schritte:

1. Schätzen des logistischen Regressionsmodells

$$logit\, p = logit\, P(y = 0|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

2. Generierung von $\hat{\beta}_i$ aus der posterioren Verteilung von $\beta_i$, $i = 1, \ldots, k$.

3. Berechnung von $p = P(y = 0|x_1, \ldots, x_k)$.

4. Simulierung einer zwischen 0 und 1 gleichverteilten Zufallsvariablen $u$.

5. Falls gilt $p < u$, so wird der fW durch 0 andernfalls durch 1 ersetzt.

Analog wird bei der Schätzung fW ordinaler Variablen vorgegangen. Mit Hilfe dieser stochastischen Ersetzungsmethode kann eine korrekte Multiple Imputation gemäß Rubin [82] durchgeführt werden, da die Wahl der Regressionskoeffizienten eine hohe Variabilität der ersetzten Werte gewährleistet. Der große Nachteil dieser Methode ist wie bei der DFM die Voraussetzung eines monotonen MDP, welches in der Praxis selten auftritt.

Die wesentlichen Voraussetzungen der bisher vorgestellten Ersetzungsmethoden werden in der nachfolgenden Tab. 2 aufgelistet.

Tab. 2: Übersicht vorgestellter Ersetzungsmethoden.

| Ersetzungsmethode | Missing Data Pattern | Missing Data Mechanism | SI/MI | speziell für kat. Var. |
|---|---|---|---|---|
| Mittelwert/Modus | beliebig | MCAR | SI | nein |
| Hot/Cold deck | beliebig | MCAR | SI | ja |
| Worst/Best Case | beliebig | MCAR | SI | ja |
| lin. Regression | monoton | MAR | SI/MI | nein |
| EM-Algorithmus | beliebig | MAR | SI/MI | nein |
| MCMC Methode | beliebig | MAR | SI/MI | nein |
| DFM | monoton | MAR | SI/MI | ja |
| log. Regression | monoton | MAR | SI/MI | ja |

### 3.1.10 Ersetzungsmethoden bei anderen Studientypen

Die zuvor vorgestellten Ersetzungsmethoden sind nur zur Ersetzung fehlender Werte in Querschnittsstudien geeignet, da sie annehmen, dass die Beobachtungen im Datensatz unabhängig voneinander erhoben wurden. Sie eignen sich daher nicht zur Ersetzung fehlender Werte in Longitudinal- und Clusterstudien [91].

In Longitudinalstudien wird vielfach die deterministische Ersetzungsmethode Last Value Carried Forward (LVCF) auch bekannt unter Last Observation Carried Forward (LOCF) verwendet. Diese Methode hat ein hohes Potential zu verzerrten Auswertungsergebnissen zu führen, nicht zuletzt da sie die Variabilität der Daten unterschätzt. Sie wird daher in der Literatur nicht empfohlen [11, 53]. Weitere Ersetzungsmodelle speziell zur Schätzung fehlender Werte im Rahmen von Longitudinal- und Clusterstudien werden beispielsweise in [62], [64], [65] und [103] beschrieben.

Im Rahmen von Longitudinalstudien und bei potentiellem Vorliegen eines MDM gemäß Missing *Not* At Random wird zwischen zwei Arten von Modellen differenziert: dem *Selektionsmodell* von Little und Rubin [63] und dem *Pattern-Mixture Modell* entwickelt von Little [61, 62, 70]. In beiden Modellen werden nicht nachprüfbare Annahmen gemacht, wobei die Annahmen bei dem Pattern-Mixture Modell im Allgemeinen als ehrlicher im Sinne von vorbehaltloser erachtet werden [69].

Sei $Y_{ij}$ die Zielgröße für die Beobachtung $i, i = 1, \ldots, N$ zu dem Zeitpunkt $j, j = 1, \ldots, T$ und $\mathbf{Y}$ die Datenmatrix. Zusätzlich sei der Ausfallindikator $R_i$ gegeben, der je Beobachtung den Zeitpuntk für den Drop-out angibt. Dann kann die gemeinsame Verteilung von $\mathbf{Y}$ und der Matrix $\mathbf{R}$ mit den Ausfallindikatoren für ein statistisches Modell wie folgt dargestellt werden

$$P(\mathbf{Y}, \mathbf{R} \mid \theta, \psi) = P(\mathbf{Y} \mid \theta, \psi)\, P(\mathbf{R} \mid \mathbf{Y}, \psi),$$

wobei $\theta$ und $\psi$ Parameter des unter inhaltlichen Aspekten interessierenden Auswertungsmodells bzw. des Ausfallmodells bezeichnen.

In *Selektionsmodellen* wird angenommen, dass die Ausfälle von den (unbeobachteten) Werten von $\mathbf{Y}$ abhängen [37, 82], d.h. es gilt

$$P(\mathbf{Y}, \mathbf{R} \mid \theta, \psi) = P(\mathbf{Y} \mid \theta)\, P(\mathbf{R} \mid \mathbf{Y}, \psi).$$

Der zweite Faktor ist die Ausfallfunktion und beschreibt die bedingte Wahrscheinlichkeit von Ausfällen, die abhängig von $\mathbf{Y}$ und $\psi$ sind. Diese Annahme über die Ausfallfunktion kann nicht überprüft werden [26], da die wahren Werte nicht bekannt sind.

Bei dem *Pattern-Mixture Modell* wird die gemeinsame Verteilungsfunktion wie folgt faktorisiert

$$P(\mathbf{Y}, \mathbf{R} \mid \theta, \psi) = P(\mathbf{Y} \mid \mathbf{R}, \theta)\, P(\mathbf{R} \mid \psi).$$

Aus der bedingten Wahrscheinlichkeit für $\mathbf{Y}$ gegeben dem Ausfallindikator $\mathbf{R}$ folgt, dass das Pattern-Mixture Modell verschiedene Modelle für die einzelnen Missing Data Pattern formuliert [94], d.h. für die einzelnen Beobachtungsgruppen, welche bei den gleichen Variablen fW aufweisen. Bei diesem Modell kann die Annahme, dass die Zusammenhänge zwischen der abhängigen und den unabhängigen Variablen in der Gruppe mit Ausfall denen der Gruppe ohne Ausfall entsprechen, nicht überprüft werden [69].

## 3.2 Ersetzungs- vs. Auswertungsmodell

Die Wahl einer geeigneten Ersetzungsmethode ist abhängig vom Merkmalstyp der Variablen mit fW, dem Missing Data Pattern und dem vorliegenden Missing Data Mechanism. Für die anschließende Auswertung der Daten spielen diese Informationen keine Rolle, da die vervollständigten Datensätze mit gewöhnlichen Standardverfahren analysiert werden. Folglich können sich die

statistischen Modelle, die zur Generierung der vervollständigten Datensätze herangezogen werden, stark von den Annahmen und Modellen unter denen die anschließenden Auswertungen durchgeführt werden, unterscheiden. In der Literatur wird zwischen dem *Ersetzungsmodell* (Imputation Model) und dem *Auswertungsmodell* (Analysis Model) unterschieden. Fay [28], Meng [66], Rubin [83], Schafer [89] und Little [63] untersuchten die Auswirkung von Unterschieden zwischen Ersetzungsmodell und Auswertungsmodell auf die statistischen Auswertungsergebnisse. Um aussagekräftige Ergebnisse zu erhalten, sollte das Ersetzungsmodell so gut es geht mit dem Auswertungsmodell übereinstimmen.

Eine Ersetzungsmethode, die multivariat normalverteilte Daten annimmt, setzt voraus, dass jede Variable mit fW durch ein lineares Regressionsmodell mit allen anderen im Datensatz befindlichen Variablen als Kovariablen dargestellt werden kann. Dieses Ersetzungsmodell berücksichtigt nur Haupteffekte, keine Interaktionen und unterscheidet nicht zwischen abhängiger und unabhängiger Variable. Es dient lediglich dazu, die wesentlichen Merkmale der gemeinsamen Verteilung wie Mittelwertsvektor und Varianzen in den ersetzten Werten zu bewahren.

Schafer *et al.* [91] weisen darauf hin, dass eine Prozedur, die die gemeinsame Verteilung bewahrt, auch automatisch den linearen Zusammenhang von einer (abhängigen) Variablen mit den anderen (unabhängigen) Variablen im Datensatz bewahrt. Ein Ersetzungsmodell im multivariat normalverteilten Fall hat daher keinen Einfluss darauf, welche Variable in der nachfolgenden Auswertung als abhängige oder unabhängige Variable modelliert werden kann. Diese Entscheidung sollte daher erst im Auswertungsmodell getroffen werden.

Einen wesentlichen Einfluss auf die Gültigkeit der statistischen Auswertungsergebnisse haben die Variablen, welche im Ersetzungs- bzw. im Auswertungsmodell berücksichtigt werden. Schafer *et al.* [91] beschreiben die Wirkung von Unterschieden in dem Ersetzungs- und Auswertungsmodell anhand von den drei

Variablen $Y_1$, $Y_2$ und $Y_3$, wobei nur die Variable $Y_3$ fehlende Werte enthält. In der ersten Situation wird angenommen, dass die fehlenden Werte von $Y_3$ durch die Regressionsgleichung

$$Y_3 = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \epsilon \qquad (Ersetzungsmodell)$$

mit normalverteiltem Fehlerterm $\epsilon$ ersetzt werden. Die anschließende Auswertung berücksichtigt im Regressionsmodell von $Y_3$ jedoch nur die Variable $Y_1$:

$$Y_3 = \beta_0 + \beta_1 Y_1 + \epsilon \qquad (Auswertungsmodell)$$

Dieses Modell ist ein Spezialfall des Ersetzungsmodells, bei dem die zusätzliche Annahme $\beta_2 = 0$ gemacht wird.

Die Auswirkung dieses Unterschieds in dem Ersetzungs- und Auswertungsmodell hängt davon ab, ob die zusätzliche Annahme im Auswertungsmodell korrekt ist. Wenn in Wahrheit $Y_2$ keinen Einfluss auf $Y_3$ hat, so dass $\beta_2 = 0$ gilt, so ist das Ersetzungsmodell dennoch nicht ungültig. Die Auswertungsergebnisse können allerdings aufgrund des gewählten Ersetzungsmodells konservativer sein, da die zusätzliche Unsicherheit des Zusammenhangs zwischen $Y_3$ und $Y_2$ berücksichtigt wurde. In dem Fall, dass die zusätzliche Annahme $\beta_2 = 0$ nicht korrekt ist, werden die mit dem Auswertungsmodell berechneten Vorhersagen von $Y_3$ verzerrt sein. Die Verzerrung liegt allerdings an der schlechten Wahl des Auswertungsmodells, nicht an der Wahl des Ersetzungsmodells.

In der zweiten von Schafer *et al.* [91] untersuchten Situation entspricht das Ersetzungsmodell dem Auswertungsmodell aus der ersten Situation (s.o.) und das Auswertungsmodell dem vormaligen Ersetzungsmodell. Wenn die zusätzliche Annahme $\beta_2 = 0$ des Ersetzungsmodells korrekt ist, so haben auch die Auswertungsergebnisse eine gute Aussagekraft. Wenn allerdings tatsächlich ein Zusammenhang zwischen $Y_2$ und $Y_3$ existiert und dieser im Ersetzungsmodell nicht berücksichtigt wird, so wird der Schätzer für $\beta_2$ im anschließenden

Auswertungsmodell zur Null hin verzerrt sein und zu fehlerhaften Ergebnissen führen können. Daher sollten im Ersetzungsmodell so viele Variablen wie möglich berücksichtigt werden. Die nachteilige Verringerung der Präzision der Schätzer aufgrund der Berücksichtigung von z.T. unwichtigen Variablen im Modell ist gewöhnlich kleiner gegenüber dem Vorteil, dass der vervollständigte Datensatz für verschiedene Auswertungsmodelle verwendet werden kann [83].

Nun kann es natürlich sein, dass das Auswertungsmodell nichtlinear ist wie beispielsweise bei einer logistischen Regression. Schafer *et al.* [89] haben Simulationen diesbezüglich durchgeführt, die zeigen, dass die Verzerrung, die möglicherweise durch die unterschiedlichen funktionalen Modelle auftreten, für gewöhnlich nur gering sind und wenn überhaupt nur einen kleinen Einfluss auf die statistischen Auswertungsergebnisse hat.

Ein viel schwerwiegenderer Unterschied zwischen Ersetzungsmodell und Auswertungsmodell tritt auf, wenn das Auswertungsmodell Interaktionen von Variablen beinhaltet [92]. Das übliche Ersetzungsmodell im multivariat normalverteilten Fall berücksichtigt keine Interaktionen, sodass mögliche vorliegende Interaktionen abgeschwächt werden und in der anschließenden Auswertung schwerer entdeckt werden können. Wenn Interaktionen eine wesentliche Rolle in der Auswertung spielen, müssen diese Zusammenhänge daher im Ersetzungsmodell explizit berücksichtigt werden. Dies kann dadurch erreicht werden, dass das Produkt von Variablen, zwischen denen Interaktionen vermutet werden, zusätzlich zu den einzelnen Variablen ins Ersetzungsmodell aufgenommen wird.

Schafer *et al.* [91] haben eine weitere Möglichkeit, das Ersetzungsmodell in geeigneter Weise im Hinblick auf die anschließende Auswertung zu wählen, vorgestellt. Sollen in der anschließenden Auswertung primär die Mittelwerte oder Häufigkeitsverteilungen zweier Gruppen analysiert werden, so können Gruppenunterschiede bewahrt werden, indem eine zusätzliche Dummyvariable (Indikatorvariable) im Modell berücksichtigt wird, die den Wert 0 für die

eine Gruppe und den Wert 1 für die andere Gruppe annimmt. Um auch Interaktionen innerhalb einer Gruppe in der Ersetzung zu erhalten, könnte der Datensatz gemäß den beiden Gruppen aufgeteilt werden und dann separate Ersetzungsmodelle angepasst werden.

Stetige Variablen, die nicht normalverteilt sind, sollten bevor sie in das Ersetzungsmodell aufgenommen werden, geeignet transformiert werden. In Simulationsstudien haben Schafer [89] und Graham *et al.* [36] zwar gezeigt, dass die unter Normalverteilungsannahme durchgeführte Ersetzung in diesem speziellen Datensatz auch bei den binären oder ordinalen Variablen die statistischen Kenngrößen kaum beeinflusst haben. Doch es wird vermutet, dass die Ersetzungsmethoden speziell für kategoriale Variablen zu valideren Ergebnissen führen, wenn deren Annahmen erfüllt sind.

## 3.3 Software zur Ersetzung

Das Ziel dieser Arbeit ist es, mit Hilfe von Simulationen verschiedene Ersetzungsmethoden zur Schätzung fehlender Werte kategorialer Variablen zu vergleichen. Die Simulationen wurden in SAS Version 9.1 durchgeführt, da dieses Programm aktuelle, d.h. zum Teil erst 1999 entwickelte Ersetzungsmethoden auch speziell für kategoriale Variablen enthält und Marktführer unter den Programmen zur Auswertung von klinischen Datensätzen ist. Ferner ist es möglich, alle einzelnen Schritte der Simulationen mit diesem Programm durchzuführen: Simulierung von unterschiedlichen Missing Data Mechanism, Durchführung verschiedener Ersetzungsmethoden, anschließende teststatistische Auswertung und abschließend die Zusammenfassung dieser Ergebnisse.

Auch andere Software ermöglicht es fW zu schätzen, um für die teststatistische Auswertung einen vervollständigten Datensatz zur Verfügung zu haben. Ein kurzer Überblick wird in Tab. 3, S. 51 gegeben. Die Auflistung erhebt keinen Anspruch auf Vollständigkeit, führt aber die bekanntesten und allge-

mein anwendbaren Programme auf. Ausführlichere Vergleiche der aufgeführten Prozeduren können [45, 46] entnommen werden. Programme für spezielle Auswertungsverfahren (z.B. AMOS und Lisrel), werden in [68, 93] beschrieben.

Einige der in Tab. 3 angegebene Prozeduren und Bezeichnungen werden nur dort verwendet und sollen im Folgenden kurz erläutert werden. Die Prozedur ICE wird nach der von ihr durchgeführten Methode *Imputation using Chained Equations* bezeichnet. ICE ist eine in Stata implementiert Version von MICE, welches für *Multivariate Imputation by Chained Equations* steht.

Die Idee der Methode der Chained Equations ist, dass das Modell für die fehlenden Werte gegeben den beobachteten Daten nicht auf einer gemeinsamen Verteilung basiert (wie z.B. multivariate Normalverteilung bei MCMC), sondern auf einer Reihe von bedingten Verteilungen. Angenommen der Datensatz enhält eine stetige Variable X, eine poissonverteilte Variable Y und eine binäre Variable Z, wobei jede Variable teilweise fW aufweist. Dann wird zur Ersetzung der fW von X ein lineares Regressionsmodell mit Y und Z als unabhängige Variablen, zur Schätzung der Werte von Y ein Poisson Regressionsmodell und zur Ersetzung der fW von Z eine logistische Regression mit X und Y als unabhängige Variablen modelliert. Schritt für Schritt wird jedes Modell angepasst und die ersetzten mit den tatsächlich erhobenen Werte zur Schätzung der fW der folgenden Variablen verwendet, bis dieser Prozess konvergiert. Anschließend erfolgt die endgültige Ersetzung der fW.

Die Predictive Mean Matching (PMM) Methode ist eine Art Hot deck Methode. Zur Ersetzung eines fW wird aus einem Satz von tatsächlich beobachteten Werten, die dem mittels einer Regressionsersetzung vorhergesagtem Wert am ähnlichsten sind, zufällig einer gezogen. Die Methode kann bei jedem Merkmalstyp angewendet werden.

Tab. 3: Übersicht über Softwarepakete zur Schätzung fehlender Werte.

| Prozedur | Software-paket | Anbieter/ Entwickler | implementierte Verfahren | Bemerkungen | Nähere Informationen |
|---|---|---|---|---|---|
| Amelia II | R | Honaker, King und Blackwell | EM mit Bootstrap | MI für Quer- und Longitudinal-studien, nur zur Ersetzung | [43, 44, 56] |
| Hmisc | R und S-Plus | F. Harrell | Predicted Mean Matching (PMM),Regression | Geeignet für bin. und kat. Var. | [38] |
| ICE | Stata | P. Royston | Chained equations, logistische Regression | implementiert MICE, geeignet für bin. und kat. Var. | [80] |
| IVEware | SAS und unabhängig | University of Michigan | Logistische Regression | MI, unterstützt komplexere De-signs der Stichprobenerhebung | [79] |
| MICE | R und S-Plus | van Buuren *et al.* | Chained equations, PMM lin., log. und poly. Regr. | Geeignet für bin. und kat. Var. | [68] |
| Missing Data Library | S-Plus 7 | Insightful | EM und Data Augmenta-tion (DA) Algorithmus | Geeignet für bin. und kat. Var. | [24, 49, 99] |
| NORM | S-Plus und unabhängig | J. Schafer | EM und Data Augmenta-tion (DA) Algorithmus | aktuellere Ersetzungsmethoden sind in S-Plus | [34, 88] |
| PROC MI | SAS V9.1 | SAS Institute | MCMC, PMM, lin. und log. Regression, DFM | Geeignet für bin. und kat. Var. ermöglicht Zusammenfassung | [45, 84] |
| Missing Value Analysis | SPSS | SPSS Inc. | CCA, EM und lin. Regression | Führt zu verzerrten Schätzern [105] | [105, 96] |
| Solas | Statistical Solutions | D. Rubin | Predicitve Mean Matching, Discriminanz Analyse | Geeignet für bin. und kat. Var. | [23, 46, 97] |

## 3.4 Guidelines zu fehlenden Werten

Wichtige Guidelines für die klinische Forschung werden von der International Conference on Harmonization (ICH) und dem Committee for Proprietary Medicinal Products (CPMP) entwickelt; unter anderem zu dem Thema fehlende Werte. Die ICH Guideline E9 *Statistical Principles for Clinical Trials* [52] thematisiert in ihrem Abschnitt 5.3 kurz die Problematik fehlender Werte und welches Vorgehen als adäquat erachtet wird. Deutlich ausführlichere Erläuterungen werden in der Guideline der CPMP *Points to Consider on Missing Data* [21] gegeben. In Letzterer werden zunächst die potentiellen Einflüsse von fehlenden Werten auf die Studienergebnisse und mögliche Ersetzungsmethoden aufgeführt. Anschließend werden allgemeine Empfehlungen zur Vermeidung von fehlenden Werten und deren Berücksichtigung im Studienplan und dem Auswertungsbericht ausgesprochen. Beide Guidelines heben hervor, dass keine universal anwendbare Methode für den Umgang mit fehlenden Werten in der statistischen Analyse empfohlen werden kann, sondern die beste Methode von der Fragestellung und der Missing Data Diagnostic abhängt.

### 3.4.1 Empfehlungen zur Studienplanung

Zur Prävention von fehlenden Werten sollten in der Planung der Studie gemäß der ICH E9 und den Points to Consider on Missing Data folgende Aspekte berücksichtigt werden:

a) Festlegung einer Zielgröße, die genau und valide erhoben werden kann.

b) Möglichst Messwerte von nichtinvasiven Verfahren erheben.

c) Vorab spezifizierte Kriterien für den Abbruch der Studienteilnahme einzelner Patienten angeben.

d) Erhebung von Variablen zur Vorhersage möglicher fW anderer wichtiger Variablen.

e) Berücksichtigung detailierter Information zu (Gründen von) Drop-outs auf den Datenerhebungsbögen (CRFs).

Wenn die Zielgröße ohne viel Aufwand erhoben werden kann und kein unzureichend validiertes Messinstrument verwendet wird, so ist die Wahrscheinlichkeit für das Auftreten von fW und nicht plausiblen Werten, die gegebenenfalls als fehlend erachtet werden müssen, gering. Je weniger Aufwand betrieben werden muss, um die Zielgröße zu erheben (sei es die Länge des auszufüllenden Fragebogens für den Patienten oder die Komplexität des Messverfahrens), desto größer ist die Motivation der korrekten und vollständigen Erhebung dieser Variablen.

Der Aspekt c) kann nur bei bestimmten Fragestellungen im Studiendesign berücksichtigt werden [65]; beispielsweise in Studien, die Bluthochdruck oder Diabetes betreffen. Bei einem Placebo-kontrollierten Studiendesign zur Analyse derartiger Fragestellungen wird ein fehlender Behandlungsfortschritt erwartet. Diesen kann man explizit messen! Patienten ohne Behandlungsfortschritt neigen eher dazu die Studie abzubrechen. Bei diesen Studien sollte daher im Studienplan diese Art von Studienabbruch explizit berücksichtigt werden. Dies kann so gestaltet werden, dass Patienten ab einem gewissen Messwert, der kein Behandlungserfolg darstellt, gemäß dem Studienplan gezwungen werden die Studie abzubrechen. Die fW, welche durch dieses Design erzeugt werden, sind vorhersehbar, sodass der Missing Data Mechanism MAR ist. Durch Anwenden adäquater Ersetzungsmethoden für MAR können in diesen Fällen trotz Vorliegen von fW unverzerrte Ergebnisse über den Behandlungserfolg gewährleistet werden.

Der Aspekt d) betrifft Studien, in denen vorab schon eine gewisse Anzahl an fW in einer oder mehreren der untersuchten Variablen erwartet wird. In derartigen Situationen sollte vorab überlegt werden, welche anderen Variablen vollständig(er) erhoben werden und dann zur Schätzung der fehlenden Werte der eigentlich interessierenden Variable verwendet werden können. Beispielsweise könnten zur Schätzung möglicher fW zum Einkommen die Variablen Alter, Schulbildung und Wohnlage erhoben werden. Dieses Vorgehen ermöglicht

einen Missing Data Mechanism gemäß MAR vorliegen zu haben und die Aussagekraft der Studienergebnisse basierend auf einem geeignet vervollständigten Datensatz ist relativ gut.

Die Berücksichtigung detaillierter Information zu (Gründen von) Drop-outs (Aspekt e)) ermöglicht die Beurteilung welcher Missing Data Mechanism vorliegt und kann hilfreich sein für die Wahl einer geeigneten Ersetzungsmethode. Wenn beispielsweise bekannt ist, dass die Ursache für den Studienabbruch (Drop-out) unabhängig von der Studie entschieden wurde, da der Proband weggezogen ist oder eine andere schwere Krankheit hat, so kann ein Missing Data Mechanism gemäß MAR angenommen werden. In anderen Situationen, in denen systematisch in einer der Behandlungsgruppen fehlende Werte auftreten, können diese gegebenenfalls durch die erhobenen Informationen für den Grund zum Studienabbruch richtig eingeordnet werden.

Damit Studienergebnisse als valide angesehen werden, auch wenn fW in der ursprünglichen Datenmatrix vorhanden sind, ist es wichtig, dass die durchgeführte Ersetzungsmethode bereits vorab im Studienplan im Abschnitt über die Statistik festgelegt wurde [52]. In diesem Abschnitt muss die Methode detailliert beschrieben und deren Wahl begründet werden. Bei der Wahl der Methode ist darauf zu achten, dass sie die Daten konservativ ersetzt und nicht im Sinne der interessierenden Alternativhypothese. Bei Unterlegenheitsstudien sollten die ersetzten Werte daher den Unterschied eher vergrößern, während bei Überlegenheitsstudien die ersetzten Werte den Unterschied zwischen den Gruppen eher verkleinern sollten.

Zusätzlich zu dieser primär festgelegten Ersetzungsmethode wird empfohlen, die Auswertung anhand von weiteren Ersetzungsmethoden, die ebenfalls sinnvolle aber andere Annahmen an die fW stellen, durchzuführen. Diese sogenannte Sensitivitätsanalyse dient dazu die Robustheit der Studienergebnisse mit der vorab als primär festgelegte Ersetzungsmethode abschätzen und recht-

fertigen zu können. Selbstverständlich müssen auch die zur Sensitivitätsanalyse verwendeten Ersetzungsmethoden im Studienplan vorab erläutert und deren Auswahl begründet werden. Eine genauere Beschreibung von Sensitivitätsanalysen erfolgt in Abschnitt 3.5.

Aufgrund von unvorhersehbaren Problemen ist es möglich, während der Studie oder in der noch verblindeten Phase der Auswertung am Studienende, die Strategie für den Umgang mit fehlenden Werten im Studienplan zu ändern [21]. Diese Änderungen müssen dokumentiert und begründet werden.

### 3.4.2 Empfohlene Ersetzungsmethoden

Wie schon zuvor erwähnt, gibt es gemäß den Richtlinien keine Ersetzungsmethode, die als Goldstandard empfohlen werden kann. In den *Points to Consider on Missing Data* in CPMP [21] werden zu einigen Methoden für den Umgang mit fW Anmerkungen zu deren adäquater Einsetzbarkeit gemacht.

Die Complete Case Analyse wird im Allgemeinen nicht als primäre Auswertungsstrategie akzeptiert. Nur noch in Ausnahmefällen wie bei explorativen Studien oder zur Sensitivitätsanalyse von konfirmatorischen Studien kann die CCA geeignet sein.

In Longitudinalstudien, in denen die *Messwerte über die Zeit als konstant* erwartet werden, kann die oft verwendete Methode last observation carried forward (LOCF) verwendet werden. Bei dieser Methode wird der zuletzt beobachtete Wert einer Beobachtung zur Ersetzung der zeitlich nachfolgenden fehlenden Werte einer Variable verwendet. Zusammen mit der Ersetzung des besten bzw. schlechtesten Wertes im konservativen Sinne, ist die LOCF gemäß [21] eine geeignete Methode für die Sensitivitätsanalyse[1]. Auch andere einfache

[1] Dies steht im Widerspruch zu anderen Literaturstellen, in denen plausible Gründe gegen eine derartige Sensitivitätsanalyse dargelegt werden, siehe Abschnitt 3.5.

Ersetzungsmethoden (z.B. Mittelwertersetzung) können angewendet werden, wenn sie konservativ sind. Bei derartigen Ersetzungsmethoden ist allerdings zu berücksichtigen, dass viele die Variabilität unterschätzen, da sie die Unsicherheit über die ersetzten Werte nicht berücksichtigen. Das Risiko der Unterschätzung der Variabilität kann durch Maximum-Likelihood Methoden oder Multiple Imputation verringert werden.

Darüber hinaus werden in den *Points to Consider on Missing Data* von CPMP statistische Auswertungsmethoden erwähnt, die keine vollständige Datenmatrix voraussetzen. Eine Methode ist das Mixed effect Modell, welches u.a. zur Auswertung bei wiederholten Messungen der Zielgröße je Patient verwendet wird und den Zeitpunkt als zufälligen Effekt berücksichtigt. Eine andere Methode ist die Überlebenszeitanalyse. Drop-outs oder fW der Zielgröße am Studienende werden bei dieser Methode als sogenannte zensierte Beobachtungen in der Auswertung berücksichtigt. Die Überlebenszeitanalyse setzt allerdings voraus, dass kein Zusammenhang zwischen Behandlung und fehlender Zielgröße besteht, was im Allgemeinen nicht zutrifft.

### 3.4.3 Empfehlungen für den Auswertungsbericht

In den *Points to Consider on Missing Data* [21] wird gefordert, dass die Beschreibung der gewonnenen Daten neben der Deskription der erhobenen Daten auch eine detailierte Beschreibung der fehlenden Werte beinhaltet. Dies bedeutet, dass die Ergebnisse der ausführlichen Missing Data Diagnostic präsentiert werden müssen: Wie viele, wo und warum sind fW aufgetreten. Dazu gehört auch die Auflistung der ersetzten Werte.

Um mögliche Einflüsse der fehlenden Werte auf die Studienergebnisse diskutieren zu können, müssen weitere Untergruppen analysiert werden. Es gilt zu untersuchen, ob sich die Patienten mit fW von denen ohne fW unterscheiden. Zusätzlich muss analysiert werden, ob sich die Behandlungsgruppen hinsicht-

lich des Auftretens der fW unterscheiden.

Die Richtlinie ICH E3 (*Structure and Content of Clinical Study Reports*) [51] verlangt im Auswertungsbericht eine detaillierte Beschreibung der vorab definierten Ersetzungsmethoden, auch diejenigen, die für eine Sensitivitätsanalyse angewandt wurden. Eventuelle Amendments (Korrekturen) der verwendeten Ersetzungsmethoden müssen ebenfalls dokumentiert werden.

Die Robustheit der Studienergebnisse, die auf der primär festgelegten Ersetzungsmethode basieren, muss mit Hilfe der Angabe der Ergebnisse der Sensitivitätsanalyse diskutiert werden. Auf welche Weise Sensitivitätsanalysen durch geführt werden können, wird im folgenden Abschnitt erläutert.

## 3.5 Sensitivitätsanalyse

Bei der statistischen Auswertung von Datensätzen mit teilweise fW werden Annahmen über den Missing Data Mechanism gemacht. Diese können nicht mit Hilfe der vorhandenen Daten validiert werden. Daher ist die Durchführung einer Sensitivitätsanalyse (SA) zur Untersuchung der Auswirkung verschiedener Ersetzungsmethoden von fW auf die Ergebnisse der statistischen Analyse sinnvoll und wird von vielen Autoren gefordert [4, 52, 55, 71]. In diesem Abschnitt werden einige Ansätze zur Durchführung einer SA kurz vorgestellt. Für detailliertere Informationen werden Literaturstellen angegeben.

Eine Sensitivitätsanalyse kann auf unterschiedliche Weisen durchgeführt werden. Die zwei am meisten verwendeten Vorgehensweisen sind die folgenden.

1. **Explizite Variation der ersetzten Werte**. Die fehlenden Werte werden durch mehr als eine deterministische (meist *ad-hoc*) Ersetzungsmethode geschätzt. Ein Beispiel dafür ist die sogenannte *extreme value imputation*, d.h. es wird eine Worst und eine Best Case Ersetzung durch-

geführt. Ziel ist, einen Eindruck von der Sensitivität der Parameter im Auswertungsmodell in Bezug auf die geschätzten Werte zu erhalten.

2. **Modellierung verschiedener MDM**. Das gleiche Auswertungsverfahren wird mit unterschiedlichen Annahmen bezüglich des MDM durchgeführt. Es kann unter Annahme eines MCAR eine Complete Case Analyse durchgeführt werden oder weitere, für die Fragestellung nicht unbedingt relevante, Variablen ins Ersetzungsmodell aufgenommen werden, um einen MAR zu gewährleisten. Bei diesem Vorgehen wird primär der Einfluss der Annahmen der postulierten Ersetzungsmodelle auf die Auswertungsergebnisse untersucht, unabhängig davon wie die ersetzten Werte variieren.

Falls im Rahmen der durchgeführten Sensitivitätsanalyse alle Ersetzungsmethoden zu den gleichen Schlussfolgerungen/Ergebnissen führen, so deutet dies darauf hin, dass die nichterhobene Information (d.h. die fW) keinen Einfluss auf die Auswertungsergebnisse hat. Als konfirmatorische Auswertungsstrategie sollte eine Sensitivitätsanalyse dennoch nicht verwendet werden [100]. Denn im Falle von konsistenten Ergebnissen, wäre *eine* Ersetzungsmethode ausreichend. Wenn die Ergebnisse nicht konsistent sind, muss vorab definiert werden, welche Ersetzungsmethode als konfirmatorische Auswertung erachtet werden soll. Dann würde aber wiederum die Durchführung dieser einen Ersetzungsmethode ausreichen.

Die grundlegende Idee einer Sensitivitätsanalyse gemäß der 1. Vorgehensweise ist es, eine Ersetzungsmethode durchzuführen, die eher zu optimistischeren Ergebnissen führt und eine weitere durchzuführen, die zu eher pessimistischeren Ergebnissen führt. Daher ist eine übliche Vorgehensweise die Durchführung einer Worst und einer Best Case Ersetzung.

Unnebrink *et al.* [100] haben dieses Vorgehen in Simulationsstudien untersucht und sind zu folgenden Ergebnissen gekommen:

- Die Worst und Best Case Ersetzung sind zu extreme Ersetzungen, um eine sinnvolle Sensitivitätsanalyse durchzuführen. Nur bei einem sehr geringen Anteil an fehlender Information können nicht im Voraus die *gegensätzlichen* Ergebnisse der Analysen vorhergesagt werden. Es ist aber wenig aussagekräftig, wenn die Sensitivitätsanalyse zu dem Ergebnis führt, dass der wahre Unterschied zwischen den gegensätzlichen Ergebnissen (z.B. signifikant/ nicht signifikant) liegt.

- Selbst bei Vorliegen eines MDM gemäß MCAR kann eine derartige Sensitivitätsanalyse zu inkonsistenten Ergebnissen führen, welches darauf hinweist, dass die Ersetzungsmethoden nicht geeignet gewählt wurden!

- Ferner ist durch dieses Vorgehen nicht garantiert, dass die Extreme der interessierenden Parameter des Auswertungsmodells durch diese Ersetzungsmethoden untersucht werden [10].

Eine Sensitivitätsanalyse basierend auf der Worst und Best Case Ersetzung wird daher in der Literatur nicht empfohlen [45, 54, 100].

Sensitivitätsanalysen gemäß der 2. Vorgehensweise sind moderner/aktueller und haben das Ziel, zu konsistenten Ergebnissen innerhalb einer gewissen Variation der Annahmen über den MDM zu führen [71]. Hierbei wird ein sogenannter Sensitivitätsparameters $\psi$ betrachtet, der den Einfluss des Missing Data Mechanism auf das Ersetzungsmodell beschreibt. Minini *et al.* [71] haben gezeigt, dass $\psi$ im Ersetzungsmodell explizit modelliert werden kann. Im Rahmen von Pattern-Mixture Modellen (siehe Abschnitt 3.1.10) modelliert $\psi$ die Unterschiede in den Verteilungen der Datenmatrix $\mathbf{Y}$ abhängig vom Missing Data Pattern. Bei Selektionsmodellen stellt $\psi$ den Einfluss von $\mathbf{Y}$ auf die Wahrscheinlichkeit für das Auftreten von fW dar.

Auch wenn der Sensitivitätsparameter im Ersetzungsmodell explizit berücksichtigt werden kann, ist dessen Schätzung fraglich, da dafür ebenfalls Annahmen getroffen werden müssen, die anhand der vorliegenden Daten nicht

verifiziert werden können. Daher empfiehlt Minini *et al.* [71] verschiedene Werte von $\psi$ explizit vorzugeben und mit diesen jeweils eine MI durchzuführen, um zu untersuchen, wie die Werte von $\psi$ die anderen Parameterschätzer des Ersetzungs- (und somit des Auswertungs-) Modells beeinflussen.

Vansteelandt *et al.* [102] und Kenward *et al.* [55] haben einen weiteren Ansatz zur Einschätzung des Einflusses der fW auf die Auswertungsergebnisse vorgestellt. Üblicherweise wird in der statistischen Auswertung die Ungenauigkeit (*imprecision*) aufgrund der erhobenen endlichen Stichprobe in Form von Konfidenzintervallen berücksichtigt. Diese Ungenauigkeit nimmt mit zunehmendem Stichprobenumfang ab. Bei Datensätzen mit fW muss aber zusätzlich die Unkenntnis (*ignorance*) über die nichterhobenen Werte berücksichtigt werden. Diese Unkenntnis kann ebenfalls durch ein Intervall reflektiert werden und nimmt bei gleichbleibendem Anteil an fW im Datensatz mit zunehmendem Stichprobenumfang *nicht* ab. In dem Konzept der Unsicherheit (*uncertainty*) werden sowohl die Ungenauigkeit als auch die Unkenntnis berücksichtigt. Zur Modellierung der Unsicherheit im Kontext von Kontingenztafeln haben Kenward *et al.* in [55] einen formalen Ansatz vorgestellt.

# Kapitel 4

# Material und Methoden

Die in Abschnitt 3.1 vorgestellten Ersetzungsmethoden schätzen die fehlenden Werte auf unterschiedliche Weise und können daher zu verschiedenen vervollständigten Datensätzen führen. Wie gut die Schätzung ist, kann auf zwei Arten beurteilt werden. Zum einen sollte der ersetzte Wert plausibel sein, d.h. einen Wert aus dem Wertebereich der Variablen mit fW annehmen und gut mit dem (unbeobachteten) tatsächlichen Wert übereinstimmen. Auf der anderen Seite werden die fehlenden Werte hauptsächlich ersetzt, um anschließend eine statistische Auswertungsmethode durchzuführen. Die Ergebnisse dieser Analyse sollten daher ähnlich sein wie die Ergebnisse von dem Datensatz, der die wahren (unbeobachteten) Werte enthält, d.h. die Auswertungsergebnisse sollten keinen systematischen Bias (Verzerrung) aufweisen und valide sein. Diese beiden Eigenschaften, Übereinstimmung der geschätzten mit den tatsächlichen Werten und die Validität der Analyseergebnisse, werden im Folgenden als die Güte einer Ersetzungsmethode bezeichnet.

Um die Güte von verschiedenen Ersetzungsmethoden zur Schätzung fehlender Werte von kategorialen Variablen zu vergleichen, wird eine Simulationsstudie durchgeführt. Basierend auf einem realen Datensatz werden potentielle Einflussfaktoren auf die Güte (Eignung) einer Ersetzungsmethode variiert, um verschiedene Situationen zu simulieren. Ziel der Vergleichsstudie ist, Emp-

fehlungen für die Wahl einer geeigneten Ersetzungsmethode zu formulieren. In diesem Kapitel wird der Aufbau und die Durchführung der Simulationen erläutert, während im nächsten Kapitel die Ergebnisse der Studie präsentiert werden.

Der Vergleich von Ersetzungsmethoden kann auf unterschiedlichen Datensätzen basieren. Es kann wie in [6, 33, 75] ein realer Datensatz mit bereits fehlenden Werten als Grundlage für den Vergleich herangezogen werden. Dieses Vorgehen ist sehr praxisnah und anwendungsorientiert, allerdings ist bei dieser Datenlage ein Vergleich mit den wahren Parameterschätzern oder dem Ergebnis basierend auf dem Originaldatensatz mit vollständig erhobenen Werten nicht möglich. Andere Vergleichsstudien basieren daher auf einem kompletten, realen Datensatz, bei dem gemäß bestimmter Missing Data Mechanismen fehlende Werte erzeugt werden [1, 5, 12, 14, 27, 40, 73, 98, 101]. Eine weitere Möglichkeit besteht darin, die Vergleiche an einem rein hypothetischen (simulierten) Datensatz durchzuführen [3, 19, 20, 48]. Das bedeutet, dass die Variablen im Datensatz gemäß gewisser Verteilungsannahmen simuliert werden und anschließend in diesem kompletten Datensatz fehlende Werte explizit erzeugt werden. Bei einem derartigen Datensatz können sehr viele potentielle Einflussfaktoren auf die Eignung von Ersetzungsmethoden kontrolliert werden, der Datensatz ist aber nicht anwendungsorientiert, da in der Realität keine Variable genau einer bestimmten Verteilung folgt.

In dieser Arbeit basiert der Vergleich von Ersetzungsmethoden auf einem realen, vollständig erhobenen Datensatz aus dem medizinischen Umfeld, um einen Bezug zur Praxis in der Medizinischen Fakultät herzustellen. Aufgrund der Vollständigkeit kann die Güte der verschiedenen Ersetzungsmethoden beurteilt werden und es können verschiedene Missing Data Mechanismen und Anteile an fW generiert werden, um unterschiedliche Situationen zu simulieren. Der verwendete Datensatz wird in Abschnitt 4.1 genauer beschrieben.

Simulationsstudien eignen sich gut zum Vergleich von Ersetzungsmethoden, da verschiedene Einflussfaktoren auf die Eignung (Güte) der Ersetzungsmethoden leicht kontrolliert werden können. Es wird vermutet, dass die Schätzung der fehlenden Werte abhängig ist von dem Merkmalstyp und der (Häufigkeits-) Verteilung der Ausprägungen der Variablen mit fehlenden Werten, dem Anteil an fW, der Anzahl der Beobachtungen im Datensatz, ob andere Variablen im Datensatz enthalten sind, die Informationen zur Schätzung der fW liefern, durch welchen Missing Data Mechanism die fW erzeugt wurden, welches Missing Data Pattern im Datensatz vorliegt und wie viele vervollständigte Datensätze bei Ersetzungsmethoden gemäß der Multiple Imputation erzeugt werden. In dieser Arbeit wird die Güte der Ersetzungsmethoden bei zwei verschiedenen Merkmalstypen, drei unterschiedlichen Anteilen von fW und fünf verschiedenen Missing Data Mechanism untersucht. Die genauen Ausprägungen der betrachteten Einflussfaktoren werden im Abschnitt 4.2 erläutert.

Studien zum Vergleich von Ersetzungsmethoden unterscheiden sich ferner anhand der Kriterien, mittels derer die Eignung (Güte) der Methoden beurteilt werden. Zum einen kann explizit die Übereinstimmung der ersetzten Werte mit den tatsächlichen Werten vom Originaldatensatz analysiert werden [73]. Zum anderen werden Ersetzungsmethoden auf Grundlage der in der Deskription verwendeten statistischen Kenngrößen, wie Schätzer für den Mittelwert, die Varianz oder die Breite des Konfidenzintervalles verglichen [1, 5, 12, 19, 27, 48, 98]. Als weiteres Vergleichskriterium werden die Ergebnisse (bzw. Parameterschätzer) bestimmter statistischer Auswertungsmethoden herangezogen [3, 20, 40, 73, 14, 75]. In dieser Arbeit werden die Methoden gemäß allen drei Möglichkeiten verglichen. Die genauen Kriterien werden im Abschnitt 4.3 vorgestellt.

In der Literatur wurden bereits Ersetzungsmethoden zur Schätzung fW von kategorialen Variablen verglichen [1, 3, 5, 12, 47]. Allerdings wurde noch nicht die Güte der beiden Ersetzungsmethoden speziell für kategoriale Variablen, die

logistische Regression und die Discriminant Function Method (DFM), systematisch und in einem adäquaten Kontext untersucht. Ake [1], Bernaards *et al.* [5] und Horton *et al.* [47] untersuchten den Einfluss von unterschiedlichen Arten des Rundens auf die Auswertungsergebnisse, wenn Ersetzungsmethoden für stetige Variablen auch zur Ersetzung fW von kategorialen Variablen verwendet werden. Allison [3] verglich mehrere Ersetzungsmethoden (unter anderem auch die logistische Regression und die DFM), aber nur bei einer binären Variablen, unter lediglich zwei verschiedenen MDM (MCAR und MAR) und verwendete im Gegenteil zu der vorliegenden Arbeit einen rein hypothetischen Datensatz in seinen Simulationen. Chen und Atebro [12] verglichen 6 Ersetzungsmethoden an einem realen Datensatz bestehend aus 2 kategorialen Variablen, wobei weder die logistische Regression noch die DFM darunter war.

Der genaue Aufbau der durchgeführten Simulationen mit der Anzahl der erzeugten Datensätze, dem Verfahren zur Erzeugung von fehlenden Werten gemäß bestimmter Missing Data Mechanismen, der Durchführung der einzelnen Ersetzungsmethoden und dem Zusammenfassen der Schätzer anhand derer die Ersetzungsmethoden verglichen werden, wird in Abschnitt 4.4 beschrieben.

## 4.1 Simulationsdatensatz

Die Simulationen zum Vergleich der Güte verschiedener Ersetzungsmethoden basieren auf einem realen Datensatz aus dem medizinischen Umfeld. Der Datensatz wurde im Rahmen einer Studie von der Forschungsgruppe Adipositas der Ulmer Universitätsklinik für Kinder- und Jugendmedizin in Zusammenarbeit mit der Klinik Hochried, Fachklinik für Kinder und Jugendliche, erhoben. Die Studie hat den Titel „Murnauer Komorbiditätsstudie“ und umfasst eine Serie von Querschnitts- und Verlaufsuntersuchungen zur Adipositas bei Kindern und Jugendlichen [25]. Der für die Simulationen verwendete Datensatz ist Grundlage für die Untersuchungen zur Ermittlung der Prävalenz von Typ-2-Diabetes von in Deutschland lebenden kaukasischen übergewichtigen Kindern

und Jugendlichen. Für Details zu dieser Studie und deren Ergebnisse sei auf die Publikation von Wabitsch *et al.* [106] verwiesen.

Der ursprüngliche Datensatz besteht aus 532 Kindern und Jugendlichen im Alter zwischen 8 und 20 Jahren, die aufgrund von Adipositas in die Klinik Hochried eingewiesen wurden. Von den Kindern wurden diverse Laborparameter, Maßzahlen und eine detaillierte Anamnese dokumentiert. Insgesamt umfasst der Datensatz 131 Variablen. Dabei weist ein Großteil der Variablen fehlende Werte auf.

Für den Vergleich der Ersetzungsmethoden wird ein Teildatensatz verwendet, der im Folgenden als Simulationsdatensatz bezeichnet wird. Die Wahl der im Simulationsdatensatz enthaltenen Variablen erfolgte nach zwei Gesichtspunkten. Zum einen sollte eine medizinisch sinnvolle Fragestellung teststatistisch untersucht werden können, um die Validität der Auswertungsergebnisse der jeweiligen Ersetzungsmethode beurteilen zu können. Zum anderen sollte mindestens eine kategoriale Variable enthalten sein, sowie eine stetige Variable, die zur Schätzung der fehlenden Werte von der kategorialen Variablen herangezogen werden kann.

Zusammen mit dem für die Murnauer Komorbiditätsstudie mitverantwortlichen Arzt von der Ulmer Universitätsklinik, Herrn Dr. C. Denzer, wurde aus den Variablen des ursprünglichen Datensatzes die Fragestellung extrahiert: Welchen Einfluss haben Alter, Geschlecht, Adiponektin, Steatosis hepatis und der Taillenumfang der Patienten auf den gemessenen Insulinspiegel. Diese Variablen bilden daher den Simulationsdatensatz. Die Wahl der Variablen wird im nächsten Absatz kurz erläutert. In Tab. 4 werden die Variablen mit Namen, Bezeichung, Ausprägungen und entsprechender Kodierung aufgelistet.

Es ist bekannt, dass das Insulin abhängig vom Alter und Geschlecht ist. Da Adiponektin ein Protein ist, das ausschließlich in den Fettzellen des Körpers

Tab. 4: Auflistung der Variablen im Simulationsdatensatz.

| Variable | Bezeichnung | Ausprägung | Kodierung |
|---|---|---|---|
| alter | Alter [Jahre] | stetig | |
| adiponectin | Adiponektin [$\mu$g/ml] | stetig | |
| insulin | Insulin [$\mu$U/ml] | stetig | |
| taille | Taillenumfang [cm] | stetig | |
| geschlecht | Geschlecht | männlich | 1 |
| | | weiblich | 2 |
| Sh_bin | Steatosis hepatis | keine Fettleber | 0 |
| | | Fettleber Grad 1-3 | 1 |
| Sh_kat | Steatosis hepatis | keine Fettleber | 0 |
| | | Fettleber Grad 1 | 1 |
| | | Fettleber Grad 2 | 2 |
| | | Fettleber Grad 3 | 3 |

produziert wird und empfindlich auf Insulin reagiert, soll dessen Einfluss auf das Insulin ebenfalls untersucht werden. Als kategoriale Variable wurde Steatosis hepatis (Sh) gewählt. Das Synonym für Sh ist Fettleber. Steatosis hepatis tritt oft bei Adipositas oder Diabetis mellitus auf. Die Variable Taillenumfang wurde primär ausgewählt, um Informationen zur Schätzung der fW von Sh zu liefern. Für detaillierte Erläuterungen der Variablen sei auf die entsprechende Fachliteratur verwiesen.

Die sonographische Klassifizierung von Steatosis hepatis erfolgte in dem Studiendatensatz nach den Empfehlungen von Saverymuttu *et al.* [87] in: keine Steatosis hepatis, Grad 1 (gering), Grad 2 (mittel) und Grad 3 (stark). Die Variable Sh ist daher ursprünglich im Datensatz ordinalskaliert. Da die Discriminant Function Method aber nur für die Ersetzung von fW binärer oder nominalskalierter Variablen geeignet ist, siehe Abschnitt 3.1.8, wird in einem weiteren Teil der Simulationen Sh als binäre Variable aufgefasst. Die binäre

Variable Sh (Sh_bin) differenziert, ob eine Fettleber vorliegt oder nicht.

Die Simulationen sollen neben der Validität der Auswertungsergebnisse, basierend auf vervollständigten Datensätzen, auch die Übereinstimmung der ersetzten Werte mit den tatsächlichen Werten untersuchen. Dies ist aber nur möglich, wenn für alle Beobachtungen die Werte bzgl. der oben aufgeführten Variablen vorliegen. Daher umfasst der Simulationsdatensatz nur die 464 der ursprünglich 532 Beobachtungen, für welche die Variablen des Simulationsdatensatzes vollständig erhoben wurden. In den Simulationen wird dann ein Teil der beobachteten Werte der Variablen Steatosis hepatis auf fehlend gesetzt, siehe Abschnitt 4.2.

Die Deskription der stetigen Variablen anhand von Mittelwert (MW), Standardabweichung (SD), Minimum, 1.- (Q1), 2.- (Median), 3.-Quartil (Q3) und Maximum sind in Tab. 5 dargestellt. Die Tab. 6 enthält die absoluten (N) und relativen (%) Häufigkeiten der kategorialen Variablen.

## 4.2 Rahmenbedingungen der Simulation

Der Vergleich der Ersetzungsmethoden wird unter verschiedenen Rahmenbedingungen durchgeführt. Zum einen wird die kategoriale Variable Steatosis hepatis einmal als ordinal skaliertes Merkmal und ein anderes Mal als nominal skaliertes (binäres) Merkmal betrachtet. Dies ermöglicht die adäquate Verwendung und Beurteilung der Ersetzungsmethoden logistische Regression und Discriminant Function Method. Zum anderen werden in den Simulationen der Einfluss vom Anteil an fehlenden Werten, vom Missing Data Mechanism und der verwendeten Ersetzungsmethode untersucht.

Tab. 5: Deskription der stetigen Variablen.

| Variable | N | MW | SD | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| alter | 464 | 14.0 | 2.0 | 8.8 | 12.7 | 14.0 | 15.1 | 20.3 |
| adiponectin | 464 | 6.3 | 3.2 | 1.2 | 4.1 | 5.5 | 7.7 | 24.5 |
| insulin | 464 | 16.5 | 10.0 | 3.2 | 9.6 | 14.1 | 20.0 | 64.2 |
| taille | 464 | 103.6 | 13.9 | 72.0 | 93.5 | 102.0 | 112.0 | 150.0 |

Tab. 6: Deskription der kategorialen Variablen.

| Variable | Ausprägung | N | % |
|---|---|---|---|
| geschlecht | männlich | 197 | 42.5 % |
| | weiblich | 267 | 57.5 % |
| Sh_bin | keine Fettleber | 340 | 73.3 % |
| | Fettleber Grad 1-3 | 124 | 26.7 % |
| Sh_kat | keine Fettleber | 340 | 73.3 % |
| | Fettleber Grad 1 | 72 | 15.5 % |
| | Fettleber Grad 2 | 45 | 9.7 % |
| | Fettleber Grad 3 | 7 | 1.5 % |

### 4.2.1 Anteil an fehlenden Werten

Der Anteil an fehlenden Werten innerhalb der Variablen Sh variiert in den Simulationen zwischen

- 10%,
- 30% und
- 50%.

Wenn in einem medizinischen Datensatz 10% der Werte fehlen, so ist dies eher ein geringer Anteil. Dieser Anteil wurde gewählt, um zu untersuchen, welchen Nutzen der Auswerter hat, wenn er trotz kleinem Anteil an fehlenden Werten eine Ersetzung durchführt (vorausgesetzt, dass es andernfalls zu keinem Selektionsbias führen würde).

Ein Anteil von 30% fW ist in medizinischen Datensätzen durchaus üblich und würde bei der Durchführung einer Complete Case Analyse zu einer deutlichen Fallzahlreduktion und somit zu einem Powerverlust führen. Bei einem Anteil von 50% fW innerhalb einer Variablen werden bereits die Hälfte der Daten zu dieser geschätzt („erfunden"). In den Simulationen soll untersucht werden, ob die Anwendung von Ersetzungsmethoden bei einem derart hohen Anteil an fW (dennoch) zu validen Ergebnissen führen kann.

### 4.2.2 Untersuchte Missing Data Mechanism

Die fünf untersuchten Missing Data Mechanismen (MDM) sind

- Missing Completey At Random (MCAR),
- lineares Missing At Random (lin. MAR),
- konvexes Missing At Random (konv. MAR),
- lineares Missing Not At Random (lin. MNAR) und
- konvexes Missing Not At Random (konv. MNAR).

Bei MCAR wird unabhängig von dem Wert (und dem Merkmalstyp) der Variablen Steatosis hepatis oder einer anderen Variablen im Simulationsdatensatz, der beobachtete Wert von Sh mit einer bestimmten Wahrscheinlichkeit auf fehlend gesetzt. Diese Wahrscheinlichkeit ist nur von dem gewünschten Anteil an insgesamt fehlenden Werten im Datensatz abhängig.

Bei dem Missing Data Mechanism MAR fehlt der Wert der Variablen Sh, abhängig von dem Wert einer anderen Variablen im Datensatz, die ebenfalls im Ersetzungsmodell enthalten ist. Da davon ausgegangen wird, dass es einen leichten, gleichsinnigen Zusammenhang zwischen Sh und dem Taillenumfang gibt, werden im Simulationsdatensatz die fW von Sh abhängig von dem erhobenen Wert vom Taillenumfang je Beobachtung erzeugt. Die Variable Taillenumfang hat im Ersetzungsmodell die primäre Aufgabe, gute Schätzungen für die fW von Sh zu liefern.

Bei MAR können die fehlenden Werte der einen Variablen unter verschiedenen Abhängigkeitsfunktionen von der anderen Variablen auftreten. Zur Veranschaulichung dieser Aussage werden die beiden Variablen body mass index (BMI) und sportliche Akitivität pro Woche mit den Ausprägungen gar nicht, wenig, viel und extrem viel betrachtet. Angenommen, die Wahrscheinlichkeit für das Auftreten von fW bei der sportlichen Aktivität nimmt mit dem BMI der Personen zu, so kann diese Abhängigkeit durch eine lineare Funktion beschrieben werden. Es liegt folglich ein lineares MAR vor. Wenn die Wahrscheinlichkeit für das Auftreten von fW bei der sportlichen Aktivität für Personen mit extrem kleinen bzw. großen BMI besonders groß ist, so bezeichnet man dies als ein konvexes MAR.

In dem Simulationsdatensatz dieser Arbeit werden ein linearer und konvexer MAR aus den zuvor beschriebenen inhaltlichen Gründen und aufgrund der unterschiedlichen Häufigkeitsverteilung der einzelnen Ausprägungen der Variablen Sh analysiert. Es wird vermutet, dass die Schätzung der fW auch davon abhängig ist, wie viele Beobachtungen mit der jeweiligen Ausprägung der kategorialen Variablen im Datensatz enthalten ist. Beim linearen MAR fehlen die Werte von Sh für Personen mit kleinen Taillenumfängen mit der größten Wahrscheinlichkeit und diese nimmt mit der Größe des Taillenumfangs ab. Beim konvexen MAR fehlen die Werte von Sh für Beobachtungen mit besonders großen bzw. kleinen Taillenumfängen (bezogen auf dieses Kollektiv) mit einer größeren Wahrscheinlichkeit als für Beobachtungen mit einem durchschnittlichen Taillenumfang. Das genaue Vorgehen und die verwendeten Wahrscheinlichkeiten werden im Abschnitt 4.4 beschrieben.

Bei einem MDM gemäß MNAR, werden die fehlenden Werte von Steatosis hepatis auf die gleiche Weise erzeugt wie bei MAR, allerdings ist die Variable Taillenumfang anschließend *nicht* im Ersetzungsmodell enthalten.

### 4.2.3 Verwendete Ersetzungsmethoden

In den Simulationsstudien werden die folgenden Ersetzungsmethoden verglichen:

- logistische Regression (log. Reg.),
- Discriminant Function Method (DFM),
- lineare Regression (lin. Reg.) und
- MCMC Methode (MCMC).

Zusätzlich wird die Güte der Complete Case Analyse (CCA) untersucht, da die CCA nach wie vor eine oft verwendete Umgangsweise mit fehlenden Werten im Datensatz ist. Theoretisch führt die Durchführung einer CCA unter MCAR zu keinen verzerrten Ergebnissen. Daher ist ein Vergleich der Ergebnisse, welche auf vervollständigten Datensätzen mittels unterschiedlicher Ersetzungsmethoden basieren, mit den Ergebnissen von der CCA unter MCAR interessant. Ferner sollte untersucht werden, wie sich die CCA im Vergleich zu Ersetzungsmethoden verhält, wenn kein MCAR vorliegt.

Primär soll mit den Simulationen die Güte der logistischen Regression und der DFM untersucht werden, da diese explizit für die Ersetzung von fW kategorialer Variablen empfohlen werden [7, 41, 86]. Deren Güte, d.h. die Validität der Auswertungsergebnisse und die Übereinstimmung der geschätzten mit den tatsächlichen (unbeobachteten) Werten, wurde in der Literatur aber noch nicht systematisch für kategoriale Variablen in Querschnittsstudien untersucht.

Das Vorgehen zur Schätzung von fW ist bei der linearen Regression und der MCMC Methode ähnlich. Diese Methoden wurden zur Schätzung fehlender Werte stetiger Variablen entwickelt, werden aber dennoch auch zur Ersetzung fW kategorialer Variablen verwendet [1, 47]. Deshalb ist es interessant zu prüfen, wie sich die Güte dieser beiden Ersetzungsmethoden von denen der Ersetzungsmethoden speziell für kategoriale Variablen unterscheiden.

Die Wahl des in den Simulationen verwendeten Ersetzungsmodells basiert auf den in Abschnitt 3.2 vorgestellten Argumenten. Für beide Merkmalstypen der Variablen Sh und alle Ersetzungsmethoden wird das identische Ersetzungsmodell zur Schätzung der fW angenommen[1]. Das *Ersetzungsmodell* ist gegeben durch

$$Sh = Adiponektin + Alter + Insulin + Taillenumfang. \tag{4.1}$$

Da in der verwendeten Statistiksoftware SAS die logistische Regression und die Discriminant Function Method im Ersetzungsmodell keine kategorialen Variablen berücksichtigen, ist die Variable Geschlecht nur im Auswertungsmodell und nicht im Ersetzungsmodell enthalten. In den Situationen, in denen Missing Not At Random simuliert wird, ist in dem Ersetzungsmodell (4.1) der Taillenumfang *nicht* enthalten.

## 4.3 Parameter zum Vergleich der Ersetzungsmethoden

Die Übereinstimmung zweier Messwerte von kategorialen Variablen kann am besten durch den (gewichteten) Kappa-Koeffizienten beurteilt werden [17]. Daher ist dies ein verwendetes Vergleichskriterium, um die Übereinstimmung der mittels verschiedenen Ersetzungsmethoden geschätzten Werte mit den Originalwerten zu beurteilen. Der Kappa-Koeffizient und dessen Interpretation wird in Abschnitt 4.3.1 vorgestellt.

In der Literatur wird auch bei kategorialen Variablen teilweise der Mittelwert und die quadratische Abweichung des Mittelwertes des vervollständigten Datensatzes von dem Originalmittelwert (root-mean-square error (RMSE)) zur Beurteilung der Übereinstimmung herangezogen [1, 3, 12]. Aus Gründen der

[1] Zur Erinnerung: Bei der Complete Case Analyse wird kein Ersetzungsmodell unterstellt, da die CCA keine Ersetzung der fW durchführt.

Vergleichbarkeit sind daher der Mittelwert und der RMSE auch in dieser Arbeit weitere Kriterien zum Vergleich der Ersetzungsmethoden, siehe Abschnitt 4.3.2.

Selbst wenn die Übereinstimmung der geschätzten mit den tatsächlichen (unbeobachteten) Werten recht gut erscheint, können sich die Ergebnisse der teststatistischen Analyse basierend auf dem vervollständigten Datensatz, deutlich von den Ergebnissen unterscheiden, die auf dem Originaldatensatz basieren. Dies kann beispielsweise bei einer (geringfügigen) systematischen Überschätzung der tatsächlichen Werte auftreten. Deshalb wird zusätzlich zur Übereinstimmung auch die Validität der Ergebnisse einer statistischen Auswertungsmethode von den verschiedenen Ersetzungsmethoden untersucht.

Eine häufig untersuchte Fragestellung in der medizinischen Forschung ist die Untersuchung des (simultanen) Einflusses von verschiedenen Variablen auf eine stetige Variable (Zielgröße). Fragestellungen von diesem Typ sind beispielsweise: Ist der neue Tumormarker [ng/ml] abhängig von der Größe des Primärtumors [mm], dem Geschlecht, dem Alter oder dem Stadium des Primärtumors? Ist die Anzahl an Asthmaanfällen abhängig von der Therapie, der Dauer der Behandlung oder dem Wissen über die Krankheit? Das adäquate statistische Auswertungsverfahren für derartige Fragestellungen ist ein multiples lineares Regressionsmodell.

Daher wird in dieser Arbeit mit Hilfe der multiplen linearen Regression der Einfluss von Alter, Geschlecht, Adiponektin, Steatosis hepatis und dem Taillenumfang auf den Insulinspiegel bei adipösen Kindern und Jugendlichen untersucht. In Abschnitt 4.3.3 wird eine kurze Einführung in die Theorie des multiplen linearen Regressionsmodells gegeben und die Kenngrößen (Parameter) des Regressionsmodells erläutert, anhand derer die Validität der Auswertungsergebnisse beurteilt werden.

### 4.3.1 Kappa-Koeffizient

Um in den Simulationen die Übereinstimmung der ersetzten mit den Originalwerten zu untersuchen, wird der Kappa-Koeffizient von Cohen [17] berechnet. Dieser misst die adjustierte Differenz zwischen beobachteter Übereinstimmung und zufallsbedingter erwarteter Übereinstimmung. Bei einem nominalen (z.B. binären) Merkmalstyp eignet sich der einfache (ungewichtete) Kappa-Koeffizient, während bei einem ordinalen Merkmalstyp der gewichtete Kappa-Koeffizient verwendet werden sollte [18].

Bei der Berechnung des Kappa-Koeffizienten basierend auf allen n=464 Beobachtungen im Datensatz stimmen automatisch die Werte der Beobachtungen mit tatsächlich erhobenen Werten überein. Auf diese Weise kann der Einfluss der geschätzten Werte auf die Übereinstimmung, wenn alle Beobachtungen insgesamt betrachtet werden, beurteilt werden. Allerdings kann bei einem derartigen Vorgehen keine unverzerrte Aussage über die tatsächliche Übereinstimmung der geschätzten Werte mit den tatsächlichen Werten getroffen werden. Deshalb wird zusätzlich der Kappa-Koeffizient ausschließlich basierend auf den geschätzten Werten berechnet. In diesem Fall variiert die zugrunde liegende Fallzahl abhängig vom Anteil an fehlenden Werten im Datensatz.

Im Folgenden werden zunächst der einfache und gewichtete Kappa-Koeffizient eingeführt und anschließend Paradoxien (Missverständnisse) bei der Interpretation des Kappa-Koeffizienten aufgezeigt. Letzteres ist wichtig, um die Aussagekraft und Interpretation dieses Parameters zum Vergleich der Ersetzungsmethoden richtig einschätzen zu können.

#### Einfacher Kappa-Koeffizient

Allgemein basiert der Kappa-Koeffizient auf den $k^2$ Zellhäufigkeiten einer Kontingenztafel. Die Kontingenztafel enthält die Beurteilung von $N$ Individuen durch zwei unabhängige Gutachter, wobei die erhobene kategoriale Variable

$k$ mögliche Ausprägungen hat. Im Kontext der durchgeführten Simulationen sind die beiden Gutachter zum einen der wahre Wert zu dieser Beobachtung und zum anderen der geschätzte Wert. Die interessierende kategoriale Variable ist Steatosis hepatis.

Für den binären Merkmalsytp von Steatosis hepatis (Sh_bin) mit den beiden Ausprägungen 0 und 1 basiert die Berechnung des einfachen Kappa-Koeffizienten auf der in Tab. 7 abgebildeten Kontingenztafel.

Tab. 7: Kontingenztafel zur Berechnung des Kappa-Koeffizienten.

| original Sh | Sh ersetzt 0 | Sh ersetzt 1 | $\sum$ |
|---|---|---|---|
| 0 | $p_{11}$ | $p_{12}$ | $p_{1\cdot}$ |
| 1 | $p_{21}$ | $p_{22}$ | $p_{2\cdot}$ |
| $\sum$ | $p_{\cdot 1}$ | $p_{\cdot 2}$ | 1 |

Bezeichne $p_{ij}$ den Anteil an Beobachtungen, die sich in der Zelle $i, j$ befinden und $p_{i\cdot} = \sum_{j=1}^{k} p_{ij}$ den Anteil an Beobachtungen, die der $i$-ten Zeile zugeordnet wurde. Analog dazu bezeichne $p_{\cdot j} = \sum_{i=1}^{k} p_{ij}$ den Anteil an Beobachtungen, die der $j$ten Spalte zugeordnet wurden. Die *beobachtete Übereinstimmung* $P_b$ ist dann gegeben durch

$$P_b = \sum_{i=1}^{k} p_{ii}$$

und die *zufällige Übereinstimmung* $P_e$ durch

$$P_e = \sum_{i=1}^{k} p_{i\cdot} p_{\cdot i}.$$

Aus diesen beiden Werten berechnet sich der *einfache Kappa-Koeffizient* gemäß

$$\kappa = \frac{P_b - P_e}{1 - P_e}.$$

Wenn die beobachtete Übereinstimmung genau der zufälligen Übereinstimmung entspricht, so ist $\kappa = 0$. Ist die beobachtete Übereinstimmung größer

als die zufällige, so nimmt Kappa positive Werte an, andernfalls negative. Bei einer perfekten Übereinstimmung gilt $\kappa = 1$. Der minimale Wert, den der Kappa-Koeffizient annehmen kann, hängt von den Randsummen $p_{i\cdot}$ bzw. $p_{\cdot j}$ der Kontingenztafel ab [17]; er beträgt aber mindestens -1. Eine Daumenregel [30] besagt, dass ein Kappa-Koeffizient von $\kappa \geq 0.75$ eine sehr gute Übereinstimmung ausdrückt, während $\kappa \leq 0.40$ eine schlechte Übereinstimmung kennzeichnet. Diese Werte sind allerdings aufgrund der in einem späteren Abschnitt aufgeführten Paradoxien nur als grobe Richtlinien zu erachten.

**Gewichteter Kappa-Koeffizient**

Für die Simulationen, in denen Steatosis hepatis als ordinale Variable mit den vier Ausprägungen 0, 1, 2 und 3 aufgefasst wird, wird der gewichtete Kappa-Koeffizient [18] berechnet. Dieser ist eine Verallgemeinerung des einfachen Kappa-Koeffizienten, wobei die jeweilige Abweichung von der genauen Übereinstimmung unterschiedlich gewichtet im Koeffizienten berücksichtigt wird. Die Gewichte sind unabhängig von den tatsächlich erhobenen Daten aus inhaltlichen Gründen für jede der $k^2$ Zellen der Kontingenztafel zu vergeben.

Folgende Anforderungen werden an die Gewichte $w_{ij}$ gestellt. Die Gewichte müssen im Intervall $[0, 1]$ liegen und für die Zellen auf der Hauptdiagonalen beträgt $w_{ii} = 1$. Das heißt der genauen Übereinstimmung wird das maximale Gewicht beigemessen. Für den Fall $i \neq j$ gilt $0 \leq w_{ij} < 1$, d.h. alle Abweichungen von der genauen Übereinstimmung erhalten weniger als das maximale Gewicht. Schließlich wird noch mit $w_{ij} = w_{ji}$ die Symmetrie der Gewichte vorausgesetzt.

Die *beobachtete Übereinstimmung* $P_{b(w)}$ berechnet sich nun gemäß

$$P_{b(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij},$$

wobei die $p_{ij}$ analog zum einfachen Kappa den Anteil an Individuen in Zelle $i, j$

bezeichnet. Die *zufallsbedingte erwartete Übereinstimmung* ist gegeben durch

$$P_{e(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i\cdot} p_{\cdot i}.$$

Der *gewichtete Kappa-Koeffizient* ist definiert als

$$\kappa = \frac{P_{b(w)} - P_{e(w)}}{1 - P_{e(w)}}.$$

Offensichtlich sind der gewichtete Kappa-Koeffizient und der einfache Koeffizient für Gewichte von $w_{ij} = 1 \quad \forall i, j = 1, \ldots, k$ identisch.

Es gibt in der Literatur mehrere Vorschläge für plausible Werte der Gewichte. Fleiss und Cohen [31] haben gezeigt, dass bis auf den Faktor $\frac{1}{N}$ der intraclass correlation coeffizient (ICC) mit dem gewichteten Kappa-Koeffizienten identisch ist, wenn die Gewichte gemäß

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$$

gewählt werden. Die Gewichte des Kappa-Koeffizienten gemäß Fleiss und Cohen für die ordinale Variable Sh_kat sind in Tab. 8 dargestellt.

Tab. 8: Gewichte $w_{ij}$ des Kappa-Koeffizienten gemäß Fleiss und Cohen.

| Original | Sh ersetzt | | | |
|---|---|---|---|---|
| Sh | 0 | 1 | 2 | 3 |
| 0 | 1 | $\frac{8}{9}$ | $\frac{5}{9}$ | 0 |
| 1 | $\frac{8}{9}$ | 1 | $\frac{8}{9}$ | $\frac{5}{9}$ |
| 2 | $\frac{5}{9}$ | $\frac{8}{9}$ | 1 | $\frac{8}{9}$ |
| 3 | 0 | $\frac{5}{9}$ | $\frac{8}{9}$ | 1 |

Cicchetti und Allison [15] haben unabhängig vom Kappa-Koeffizienten eine Statistik vorgestellt, die die Reliabilität zwischen Gutachtern misst. Diese Statistik ist formal identisch mit dem gewichteten Kappa-Koeffizienten, wobei die Gewichte gegeben sind durch

$$w_{ij} = 1 - \frac{|i-j|}{k-1}.$$

Der Unterschied zwischen den Gewichten von Fleiss und Cohen zu Cicchetti und Allison besteht darin, dass den Abweichungen von der genauen Übereinstimmung bei der Verwendung der Gewichte von Fleiss und Cohen ein größeres Gewicht beigemessen wird als bei Cicchetti und Allison. Für die ordinale Variable Sh_kat betragen die Gewichte des Kappa-Koeffizienten gemäß Cicchetti und Allison den in Tab. 9 abgebildeten.

Tab. 9: Gewichte $w_{ij}$ des Kappa-Koeffizienten gemäß Cicchetti und Allison.

| Original | Sh ersetzt | | | |
|---|---|---|---|---|
| Sh | 0 | 1 | 2 | 3 |
| 0 | 1 | $\frac{6}{9}$ | $\frac{3}{9}$ | 0 |
| 1 | $\frac{6}{9}$ | 1 | $\frac{6}{9}$ | $\frac{3}{9}$ |
| 2 | $\frac{3}{9}$ | $\frac{6}{9}$ | 1 | $\frac{6}{9}$ |
| 3 | 0 | $\frac{3}{9}$ | $\frac{6}{9}$ | 1 |

Da alle Ersetzungsmethoden mit dem gleichen gewichteten Kappa-Koeffizienten beurteilt werden, ist die Wahl der Gewichte nicht relevant. Die Gewichte von Cicchetti und Allison sind der Default in der für die Simulationen verwendeten Statistiksoftware SAS und werden vermutlich in der Praxis am meisten verwendet. Deshalb wird in den Simulationen der gewichtete Kappa-Koeffizient gemäß Cicchetti und Allison zur Beurteilung der Übereinstimmung berechnet.

**Paradoxien des Kappa-Koeffizienten**

Die Schwierigkeit bei der Verwendung des (gewichteten) Kappa-Koeffizienten liegt in der Interpretation. Es können 3 scheinbare Paradoxien auftreten. Zum einen ist es durchaus möglich, dass sich die Kappa-Koeffizienten stark unterscheiden, obwohl die beobachteten Übereinstimmungen $P_b$ zweier Kontingenztafeln gleich sind. Dieser Unterschied resultiert aus den verschiedenen Randsummen der beiden Kontingenztafeln, welche zu unterschiedlichen zufallsbedingten erwarteten Übereinstimmungen $P_e$ führen. Bei gleicher $P_b$ ist der Kappa-Koeffizient für große $P_e$ kleiner [29].

Ein weiteres Paradoxon besteht darin, dass die erwartete Übereinstimmung abhängig von der Größe des Unterschieds zwischen der $i$-ten Zeilensumme und der $i$-ten Spaltensumme ist. Immer unter der Voraussetzung, dass die beobachtete Übereinstimmung gleich ist, ist bei Kontingenztafeln mit sogenannten unbalancierten Randsummen, d.h. die jeweiligen Zeilen- und Spaltensummen unterscheiden sich (stark), die erwartete Übereinstimmung kleiner als bei Kontingenztafeln, die balancierte Randsummen aufweisen, d.h. deren Zeilen- und Spaltensummen jeweils betragsmäßig übereinstimmen. Folglich ist für Kontingenztafeln mit unbalancierten Randsummen die $P_e$ kleiner und der Kappa-Koeffizient somit größer als bei Kontingenztafeln mit balancierten Randsummen. Das Beispiel von [57] in Abb. 6 verdeutlicht dies nochmal.

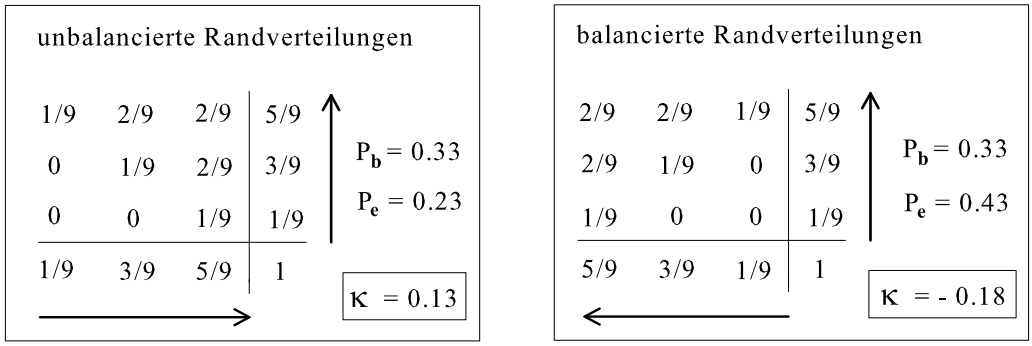


Abb. 6: Kappa Paradoxon Randverteilungen

Dass dritte Paradoxon tritt auf, wenn die beobachtete Übereinstimmung 0 beträgt, aber der Kappa-Koeffizient ungleich 0 ist, da die zufallsbedingte erwartete Übereinstimmung positiv ist. Insgesamt ist in solchen Situationen der Kappa-Koeffizient negativ. Dies mag als paradox erscheinen, da offensichtlich keine Übereinstimmung vorliegt und somit ein Kappa-Koefizient von $\kappa = 0$ erwartet wird.

Die Paradoxien lassen sich mit der Definition von [17] erklären: „Der Kappa-Koeffizient ist der Anteil an Übereinstimmung, nachdem der Anteil an zufälliger Übereinstimmung aus der Betrachtung entfernt wurde.“ Der suggerierte

Zusammenhang zwischen zufälliger und beobachteter Übereinstimmung ist, dass die zufällige Übereinstimmung Teil der beobachteten Übereinstimmung ist. Somit dürfte auch keine zufällige Übereinstimmung vorliegen, wenn schon keine beobachtete Übereinstimmung vorliegt. Dies ist allerdings ein Trugschluss, da in die Berechnung der zufälligen Übereinstimmung auch Zellhäufigkeiten miteingehen, die nichts zur beobachteten Übereinstimmung beitragen. Dies bedeutet, dass die zufällige Übereinstimmung nur teilweise in der beobachteten Übereinstimmung enthalten ist.

Der Kappa-Koeffizient korrigiert aber auch um den Teil der zufälligen Übereinstimmung, der nicht in der beobachteten Übereinstimmung berücksichtigt wird. Letztere basiert auf den Anteilen in den Zellen der Nebendiagonalen. Wenn man sich dies vor Augen führt, so sind die oben aufgeführten Paradoxien als Missverständnisse aufzufassen.

Um diesen Missverständnissen entgegenzuwirken, gibt es eine Reihe von Ansätzen. Es wird zum einen vorgeschlagen, die Berechnung der zufälligen Übereinstimmung zu modifizieren, so dass nur um den Teil korrigiert wird, der auch bei der beobachteten Übereinstimmung betrachtet wird. Andere Ansätze argumentieren, dass die Angabe des Kappa-Koeffizienten allein nicht aussagefähig ist, sondern noch weitere Parameter anzugeben sind, um den Wert des Kappa-Koeffizienten besser einschätzen zu können [9, 16, 58].

Trotz der aufgeführten Interpretationsschwierigkeiten wird in den Simulationen der Kappa-Koeffizient von Cohen berechnet, um einen Eindruck von der Übereinstimmung zwischen ersetztem und Originalwert zu erhalten. Im Vergleich zu den modifizierten Kappa-Koeffizienten ist der von Cohen ein vielfach verwendeter Parameter, auch zum Vergleich von Ersetzungsmethoden [73]. Somit sind die erzeugten Ergebnisse mit denen aus der Literatur vergleichbar. Ferner lassen sich anscheinend paradoxe Ergebnisse durch eine sachlich richtige Interpretation korrekt einordnen.

### 4.3.2 Mittelwert und RMSE

Die Parameter Mittelwert und root-mean-square error (RMSE) werden zur Vergleichbarkeit mit Ergebnissen aus der Literatur [5, 12, 27, 33] berechnet und um einen Eindruck zu gewinnen, ob die tatsächlichen Werte eher über- bzw. unterschätzt werden. Die Berechnung eines Mittelwertes bei einem kategorialen Merkmal ist statistisch gesehen nur selten sinnvoll. Der Mittelwert von einem binären 0/1 kodierten Merkmal hingegen drückt den Anteil an aufgetretenen Einsen aus und kann daher sinnvoll interpretiert werden. Allerdings gibt der Vergleich des Originalmittelwertes mit dem Mittelwert der vervollständigten Datensätze keinen Aufschluss darüber, ob die Einsen/Nullen auch bei den gleichen Beobachtungen wie im Originaldatensatz aufgetreten sind.

Der mittlere quadrierte Abstand (mean-square error, MSE) zwischen einem Punktschätzer $\hat{\theta}$ und seinem wahren Wert $\theta$ ist bei einem stetigen Merkmal ein nützlicher Messwert für die Genauigkeit der Schätzung. Dieser Messwert kombiniert die Idee von der Verzerrung (Bias $= E[\hat{\theta} - \theta]$) und der Effizienz, d.h. der Schätzer sollte eine möglichst kleine Varianz aufweisen. Es kann gezeigt werden, dass der MSE eines Schätzers gleich dessen quadriertem Bias plus der Varianz ist [20], d.h. es gilt

$$MSE = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - \theta] + Var(\hat{\theta}). \tag{4.2}$$

Für eine einfachere Interpretation wurde in den Simulationsergebnissen die Wurzel des MSE, also der RMSE abgebildet, da dieser denselben Maßstab wie die Variable (hier Steatosis hepatis) hat.

Die Größen vom Mittelwert und RMSE ermöglichen Aussagen bezüglich der Validität und der Genauigkeit der Ergebnisse. Wenn die Mittelwerte der vielen vervollständigten Datensätze je Situation in den Simulationen homogen um den Originalmittelwert streuen, weicht auch der Gesamtmittelwert, d.h. der Mittelwert der einzelnen Mittelwerte, kaum vom Originalmittelwert ab.

Man bezeichnet dies dann als *valides* Ergebnis, da kein systematischer Fehler vorliegt.

Im Vergleich dazu bezeichnet man ein Ergebnis als *genau*, wenn die Varianz des Punktschätzers klein ist. In Situationen, in denen der Gesamtmittelwert kaum vom Originalmittelwert abweicht, der RMSE aber relativ groß ist, deutet dies darauf hin, dass die Ersetzungsmethode valide, aber ungenaue Ergebnisse liefert. Auch wenn dies nicht optimal ist, ist diese Methode derjenigen vorzuziehen, die zwar einen kleineren RMSE hat, d.h. genaue Ergebnisse liefert, aber dessen Gesamtmittelwert stark vom Originalmittelwert abweicht, d.h. keine validen Ergebnisse erzielt.

### 4.3.3 Regressionsmodellparameter

Kenngrößen zur Messung der Übereinstimmung von ersetzten mit Originalwerten geben nicht unbedingt Auskunft darüber, welchen Einfluss die Ersetzung von fehlenden Werten auf die Ergebnisse einer statistischen Analyse hat. Deshalb wird zusätzlich eine multiple lineare Regression mit den vervollständigten Datensätzen durchgeführt und die Ergebnisse mit denen vom Originaldatensatz verglichen. Somit können die Auswirkungen von verschiedenen Ersetzungsmethoden auf die Schätzer der Regressionsparameter und deren Streuung untersucht werden.

Im Folgenden wird zunächst die zugrunde liegende Theorie der (multiplen) linearen Regression erläutert. Anschließend werden die Kenngrößen des Regressionsmodells vorgestellt, anhand derer die Ergebnisse der einzelnen Ersetzungsmethoden verglichen werden.

#### Lineare Regression

Eine lineare Regression wird durchgeführt, um den Zusammenhang (Einfluss) zwischen einer freien Variablen (Einflussgröße) und einer abhängigen Varia-

blen (Zielgröße) zu untersuchen. Dabei ist aus inhaltlichen Gründen bekannt, welche Variable die freie bzw. abhängige Variable ist. Es kann auch der Einfluss mehrerer Variablen auf die eine abhängige Variable untersucht werden. Im Gegenteil zur Korrelation wird bei der Regression nicht die Stärke des Zusammenhangs, sondern (primär) die Art des Zusammenhangs untersucht.

Das einfachste Modell der linearen Regression ist das einfache lineare Regressionsmodell. Einfach in dem Sinne, dass der lineare Einfluss von nur einer Variablen (Einflussgröße) X auf die abhängige, stetige Variable (Zielgröße) Y untersucht wird. Grafisch lässt sich dieser Einfluss am besten durch eine Punktewolke beurteilen, siehe Abb. 7.

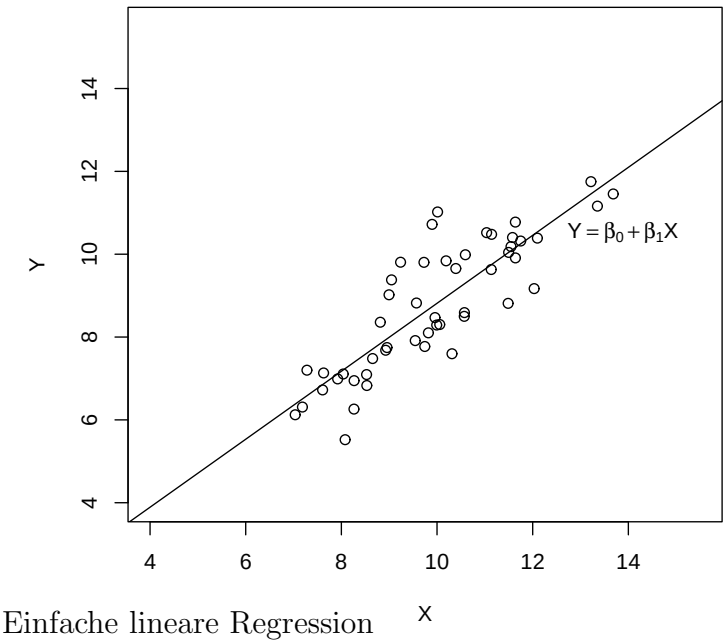


Abb. 7: Einfache lineare Regression

Die Art des Zusammenhangs wird durch die Regressionsgerade (das Regressionsmodell)

$$Y = \beta_0 + \beta_1 X$$

beschrieben. Die Parameter $\beta_0$ und $\beta_1$ werden als *Regressionskoeffizienten* bezeichnet und repräsentieren den Y-Achsenabschnitt respektive die Steigung der Regressionsgeraden.

Eine Verallgemeinerung der einfachen linearen Regression ist die multiple lineare Regression. Diese untersucht den linearen Einfluss von $p-1$ Einflussgrößen $x_1, \ldots, x_{p-1}$ auf die stetige Zielgröße $y$. Das *theoretische Modell* hat die Form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon.$$

Angenommen es liegen für jede Variable $n$ Beobachtungen vor. Dann ist das *Datenmodell* gegeben durch

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_{p-1} x_{1(p-1)} + \epsilon_1 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{p-1} x_{n(p-1)} + \epsilon_n, \end{aligned}$$

wobei

$y_i$ sind Zufallsvariablen,

$\beta_k$ , $k=0, \ldots, p-1$ sind die Regressionskoeffizienten,

$x_{ik}$ , $i=1, \ldots, n$, $k=0, \ldots, p-1$ sind bekannte Konstanten,

$\epsilon_k$ , $k = 0, \ldots, p-1$ sind normalverteilte Zufallsfehler (Residuen) mit Erwartungswert $E(\epsilon_i) = 0$ und Varianz $Var(\epsilon_i) = \sigma^2$; außerdem wird $Cov(\epsilon_i, \epsilon_j) = 0 \; \forall i \neq j$ vorausgesetzt.

In Matrixschreibweise lässt sich das Modell formulieren durch

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

Die Schätzer für die Regressionskoeffizienten ergeben sich durch Minimierung der Quadratsumme der Residuen mittels der Methode der kleinsten Quadrate. Als Lösung erhält man den Vektor $\mathbf{b}$, dessen Elemente Punktschätzer der Regressionskoeffizienten $\beta_k$, $k=0, \ldots, p-1$ sind. Dieser Vektor lässt sich darstellen gemäß

$$\mathbf{b} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}.$$

Die einzelnen $b_k$, $k = 1, \ldots, p-1$ geben an, um wie viel sich die Zielgröße verändert, wenn sich die Variable $\mathbf{x_k}$ um eine Einheit verändert.

Eine Variable $\mathbf{x_k}$ wird dann als einflussreich oder signifikant bezeichnet, wenn deren (geschätzter) Regressionskoeffizient $b_k$ signifikant unterschiedlich von 0 ist. Die hierfür zu testende Nullhypothese lautet $H_0 : \beta_k = 0$, $k = 1, \ldots, p-1$. Diese Hypothese wird mittels einer Teststatistik untersucht, die die Varianz des Schätzers beinhaltet.

Da der Vektor $\beta$ eine lineare Transformation von $\mathbf{y}$ und somit von $\epsilon$ ist, ist er normalverteilt gemäß $\mathcal{N}(\beta, \sigma^2(\mathbf{X^T X})^{-1})$. Setzt man für die unbekannte Varianz $\sigma^2$ von $\epsilon_k$ deren Schätzer ein, so ist die Kovarianzmatrix $\Sigma$ von $\beta$ gegeben durch $\Sigma = \hat{\sigma}^2(\mathbf{X^T X})^{-1}$ und die geschätzte Varianz $\widehat{Var(b_k)}$ eines Regressionskoeffizienten $b_k$ ist das k-te Diagonalelement der geschätzten Kovarianzmatrix.

Die Teststatistik zur Prüfung $H_0 : \beta_k = 0$ vs. $H_1 : \beta_k \neq 0$ ist gegeben durch

$$F = \frac{b_k^2}{\widehat{Var(b_k)}} \sim F_{1,n-p}.$$

Da die Teststatistik eine F-Verteilung hat, wird sie als *F-Statistik* bezeichnet. Falls gilt $F > F_{1,n-p;1-\alpha}$, so wird die Nullhypothese verworfen und geschlussfolgert, dass die Variable $\mathbf{x}_k$ einen signifikanten Einfluss auf die Zielgröße $\mathbf{y}$ hat, andernfalls kann ein deutlicher Einfluss von $\mathbf{x}_k$ auf $\mathbf{y}$ nicht gezeigt werden.

Am Regressionsmodell interessiert allerdings nicht nur der lineare Zusammenhang zwischen den einzelnen Einflussgrößen und der Zielgröße, sondern auch wie gut die mittels des Regressionsmodells geschätzten Werte mit den tatsächlich beobachteten Werten übereinstimmen. Ein Maß für die Güte dieser Anpassung ist das *multiple lineare Bestimmtheitsmaß* $R^2$ (multipler Determinationskoeffizient). Es ist definiert durch

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

wobei

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (mit\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i),$$

die quadrierte Abweichung der geschätzten Werte zum Gesamtmittelwert ist und mit *sums of squares due to regression* (SSR) bezeichnet wird,

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

die quadrierte Abweichung der beobachteten von den geschätzten Werten ist und mit *sums of squares due to error* (SSE) bezeichnet wird und

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

die gesamte quadrierte Abweichung der beoachteten Werte von dem Mittelwert von $y$ ist und daher mit *sums of squares total* (SST) bezeichnet wird.

Das $R^2$ nimmt Werte aus dem Intervall $[0, 1]$ an und drückt aus, um welchen Anteil die Variabilität der Zielgröße $y$ durch die Anpassung der $p-1$ Einflussgrößen $x_k$ verringert werden kann. Es gilt daher: je größer das $R^2$ desto besser ist das Modell. Bei einem $R^2 = 1$ liegen alle beobachteten Werte von $y$ in der geschätzten Regressionsebene.

Um zu überprüfen, ob das für die Stichprobe berechnete Bestimmtheitsmaß zufällig von 0 verschieden ist, kann ein sogenannter F-Test durchgeführt werden. Als Nullhypothese fungiert die Annahme, dass in der Grundgesamtheit kein linearer Zusammenhang zwischen der Zielgröße $y$ und den $p-1$ Einflussgrößen $x_k$ besteht, d.h. das Hypothesensystem besteht aus $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ vs. $H_1$: mindestens ein $\beta_k \neq 0$. Die Teststatistik ist gegeben durch

$$F \;=\; \frac{R^2}{p-1} \Big/ \frac{1-R^2}{n-p} \;=\; \frac{SSR}{p-1} \Big/ \frac{SSE}{n-p} \;=\; \frac{MSR}{MSE} \;\sim\; F_{p,n-p;\,1-\alpha}.$$

Zu beachten ist, dass ein großes $R^2$ nicht automatisch bedeutet, dass das angepasste Modell gut ist. Es gibt unterschiedliche Ursachen, die zu einem großen $R^2$ führen, wobei die Schätzer der Regressionskoeffizienten dennoch unbrauchbar sind. Zum einen kommt es darauf an, dass die erhobenen Werte den gesamten möglichen Wertebereich der Einflussgrößen abdecken. Andernfalls müsste

bei der Verwendung des Modells zur Vorhersage weiterer Beobachtungen oft extrapoliert werden und die Anpassung für Werte außerhalb des erhobenen Wertebereiches ist unbekannt.

Ferner lässt sich von einem großen $R^2$ nicht auf eine kleine Abweichung der geschätzten von den beobachteten Werten ($SSE$) schließen [72], welche aber notwendig ist, um möglichst präzise Vorhersagen mit dem Modell zu treffen. Dies kann vor allem bei Vorliegen von Ausreißern im Datensatz auftreten oder wenn die Streuung der beobachteten Werte von $y$ sehr groß ist.

Eine weitere Ursache für eine scheinbar gute Anpassung ist die Anzahl an Einflussgrößen im Modell. Je mehr Einflussgrößen im Modell enthalten sind, desto größer ist das $R^2$. Dies kann dadurch begründet werden, dass die $SSE$ durch Hinzufügen von weiteren Einflussgrößen nie kleiner werden kann und die insgesamte Streuung, $SST$, unabhängig von der Anzahl an Variablen im Modell, gleich bleibt. Deshalb wird bei der multiplen linearen Regression ein modifiziertes $R^2$ empfohlen, welches die Anzahl der Einflussgrößen im Modell berücksichtigt. Das *adjustierte lineare Bestimmtheitsmaß* berechnet sich gemäß

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}.$$

Dieses Maß kann sogar kleiner werden, wenn eine zusätzliche Variable ins Modell aufgenommen wird.

In den Simulationen wird der Einfluss von Alter, Geschlecht, Adiponektin, Steatosis hepatis (binär) und Taillenumfang auf die logarithmierten Werte von Insulin mittels des *Regressionsmodells*

$$\begin{aligned} log(insulin) &= \beta_0 + \beta_1\, alter + \beta_2\, geschlecht + \beta_3\, adiponectin \\ &\quad + \beta_3\, Sh_bin + \beta_4\, taille \end{aligned} \tag{4.3}$$

modelliert. Die logarithmische Transformation der Werte von Insulin ist notwendig, da ansonsten die Modellannahme der homogen streuenden Residuen

nicht erfüllt sein würde.

Da die Regressionskoeffizienten angeben, um wie viel sich der Wert der Zielgröße verändert, wenn sich die Einflussgröße (Kovariable) um eine Einheit verändert, ist es wichtig, dass die Einheiten der jeweiligen Einflussgröße interpretierbar sind. Dies trifft bei stetigen Variablen wie Alter, Adiponektin und Taillenumfang zu, selbst bei binären Variablen wie Geschlecht und Steatosis hepatis mit den beiden Ausprägungen {„keine Sh“, „eine Sh“}; nicht aber bei der Variablen Sh mit ordinalem Merkmalstyp.

Wenn die Variable Steatosis hepatis als ordinale Variable mit 4 Ausprägungen im Modell enthalten wäre, könnte der Wert des zugehörigen Regressionskoeffizienten nicht sinnvoll interpretiert werden. Daher wird zunächst eine Dummykodierung für diese Variable durchgeführt. Genauer, es wird eine Referenzkodierung durchgeführt, wobei die Ausprägung „keine Sh“ als Referenzkategorie gewählt wurde. Das Regressionsmodell lautet

$$\begin{aligned} log(insulin) &= \beta_0 + \beta_1\, alter + \beta_2\, geschlecht + \beta_3\, adiponectin \\ &\quad +\beta_3\, D1 + \beta_4\, D2 + \beta_5\, D3 + \beta_6\, taille, \end{aligned} \tag{4.4}$$

wobei die Dummyvariablen $D1, D2, D3$ den Unterschied der jeweiligen Ausprägung (Steatosis hepatis Grad 1, 2 bzw. 3) zur Referenzkategorie modellieren.

Um analog zu den anderen Einflussgrößen im Modell den Einfluss von der ordinalen Variablen Sh anhand von nur einer Teststatistik beurteilen zu können, wird zusätzlich die Hypothese $H_0 : D1 = D2 = D3 = 0$ vs. $H_1 : Di \neq 0$ für mindestens ein $i$, $i = 1, 2, 3$ getestet.

**Kenngrößen der Regressionsparameter zum Vergleich der Ersetzungsmethoden**

In den Simulationen werden die Ersetzungsmethoden anhand von mehreren Kenngrößen vom multiplen linearen Regressionsmodell verglichen. Da die Re-

gressionskoeffizienten die Art des Einflusses der jeweiligen Variablen auf die Zielgröße beschreiben, wird mit Hilfe der Simulationen untersucht, ob die Ersetzungsmethoden zu verzerrten Schätzern der Regressionskoeffizienten der Variablen Steatosis hepatis führen. Die Ersetzung der fW von Sh könnte zwar auch die Regressionskoeffizienten der anderen Einflussgrößen beeinflussen, doch aus Platzgründen wird darauf in dem Ergebnisteil nicht näher eingegangen.

Ebenso wichtig wie die genaue Beschreibung des Zusammenhangs zwischen Einfluss- und Zielgröße ist die Aussage, ob die jeweilige Einflussgröße einen teststatistisch signifikanten (deutlichen) Einfluss auf die Zielgröße hat. Daher werden in den Simulationen auch die Teststatistiken zu den jeweiligen Regressionskoeffizienten zum Vergleich der Ersetzungsmethoden herangezogen.

Als Maß für die Güte der Anpassung des Regressionsmodells stellt das lineare Bestimmtheitsmaß $R^2$ eine wesentliche Kenngröße von Regressionsmodellen dar. Aus diesem Grund wird es in den Simulationen für jede Rahmenbedingung erhoben und zum Vergleich der Ersetzungsmethoden verwendet. Ferner werden die Ersetzungsmethoden anhand des (globalen) F-Tests verglichen, der untersucht, ob das $R^2$ sich deutlich von 0 unterscheidet, d.h. ob mindestens eine der Einflussgrößen einen deutlichen Einfluss auf die Zielgröße hat.

Um die Validität der Ergebnisse des gewählten statistischen Auswertungsverfahrens zu beurteilen, werden nicht die Parameterschätzer (von den Regressionskoeffizienten, den Teststatistiken und dem $R^2$) an sich betrachtet, sondern die Abweichungen der Parameterschätzer basierend auf den vervollständigten Datensätzen zu den Parameterschätzern basierend auf dem Originaldatensatz. Zusätzlich werden zur Beurteilung der Effizienz (Genauigkeit) der Schätzungen die Standardabweichungen der Parameterschätzer berechnet. Die genaue Berechnung dieser Vergleichskenngrößen wird im folgenden Abschnitt beschrieben.

## 4.4 Durchführung der Simulation

Die Simulationen werden auf einem PC, auf dem das Betriebssystem Windows XP installiert ist, mit der Statistiksoftware SAS® Version 9.1 durchgeführt. Für eine genaue Beschreibung der verwendeten Funktionen und Prozeduren wird auf die Onlinehilfe von SAS [84] verwiesen.

Für jede Kombination der betrachteten Rahmenbedingungen (beide Merkmalstypen von Steatosis hepatis, den 3 Anteilen an fW (10%, 30% und 50%) und den Missing Data Mechanismen (MCAR, konvexes MAR und lineares MAR)) werden von dem ursprünglich vollständig beobachteten Simulationsdatensatz jeweils 1000 Datensätze mit fehlenden Werten erzeugt, welches eine übliche Anzahl an Wiederholungen bei Simulationen ist [8]. (Zur Erinnerung: Die Simulationen, in denen ein Missing Data Mechanism (MDM) gemäß MNAR (linear oder konvex) modelliert wird, basieren auf den 1000 Datensätzen, welche gemäß dem linearem bzw. konvexem MAR erzeugt wurden, nur ist die Variable Taillenumfang nicht im Ersetzungsmodell enthalten.)

Die Erzeugung von fehlenden Werten gemäß linearem und konvexem MAR lehnt sich an die Vorgehensweise von [20]. Die fW von Steatosis hepatis (Sh) werden bei diesen MDM abhängig von dem Wert der leicht mit Sh korrelierenden Variablen Taillenumfang erzeugt.

Beim linearen MAR nimmt die Wahrscheinlichkeit für das Auftreten von fehlenden Werten mit der Größe der Werte vom Taillenumfang (und somit auch von Sh) ab. Für Beobachtungen mit Werten vom Taillenumfang, die kleiner als das 1. Quartil sind, fehlt der Wert von Sh mit der größten Wahrscheinlichkeit. Mit jedem Quartil mehr verringert sich die Wahrscheinlichkeit für das Auftreten von fW von Sh um 1/4 der Wahrscheinlichkeit für Werte kleiner als das 1. Quartil. Die jeweiligen Wahrscheinlichkeiten für das Fehlen der Werte innerhalb eines Quartils sind abhängig von dem insgesamt zu erzielenden An-

teil an fW im Datensatz. Die genauen Wahrscheinlichkeiten sind in Tab. 10 angegeben.

Tab. 10: Wahrscheinlichkeiten für das Fehlen der Werte von Sh zur Erzeugung eines linearen MAR.

| Quartil | Anteil an fW insgesamt | | |
|---|---|---|---|
| Taillenumfang | 10% | 30% | 50% |
| bis 1. | 0.16 | 0.48 | 0.80 |
| 1. - 2. | 0.12 | 0.36 | 0.60 |
| 2. - 3. | 0.08 | 0.24 | 0.40 |
| 3. - 4. | 0.04 | 0.12 | 0.20 |

Sollen insgesamt 10% der Werte von Sh fehlen, so beträg die Wahrscheinlicheit für das Fehlen der Werte für Beobachtungen mit Werten vom Taillenumfang, die kleiner als das 1. Quartil sind, 0.16. Für Beobachtungen mit Taillenumfängen aus den folgenden Quartilen verringert sich diese Wahrscheinlichkeit jeweils um 0.04. D.h. für Beobachtungen mit Taillenumfängen zwischen dem 1. und 2. Quartil beträgt die Wahrscheinlichkeit für das Fehlen von Sh 0.12, für Taillenumfänge zwischen dem 2. und 3. Quartil 0.08 und für Beobachtungen mit Taillenumfängen, die größer als das 3. Quartil sind, 0.04. Dies gewährleistet, dass insgesamt 0.16 * 0.25 + 0.12 * 0.25 + 0.08 * 0.25 + 0.04 * 0.25 = 10% der Werte fehlen.

Beim konvexen MAR treten fehlende Werte von Steatosis hepatis mit einer größeren Wahrscheinlichkeit bei denjenigen Beobachtungen auf, die einen besonders kleinen oder großen Taillenumfang haben. Als Schranke für besonders kleine bzw. große Werte, wird das 1. und 3. Quartil vom Taillenumfang herangezogen. In Tab. 11 sind die Wahrscheinlichkeiten für das Fehlen der Werte in Abhängigkeit von dem insgesamten Anteil an fW und dem Quartil vom Taillenumfang zur Erzeugung des konvexen MAR angegeben.

Tab. 11: Wahrscheinlichkeiten für das Fehlen der Werte von Sh zur Erzeugung eines konvexen MAR.

| Quartil | Anteil an fW insgesamt | | |
|---|---|---|---|
| Taillenumfang | 10% | 30% | 50% |
| bis 1. | 0.15 | 0.45 | 0.75 |
| 1. - 2. | 0.05 | 0.15 | 0.25 |
| 2. - 3. | 0.05 | 0.15 | 0.25 |
| 3. - 4. | 0.15 | 0.45 | 0.75 |

Bei einem Anteil von insgesamt 10% fW von Sh, beträgt die Wahrscheinlichkeit für das Fehlen der Werte von Sh für Beobachtungen mit Werten vom Taillenumfang, die in den Extrembereich fallen, 0.15 und für Beobachtungen mit Taillenumfängen zwischen dem 1. und 3. Quartil 0.05. Daher fehlen insgesamt 0.15*0.5 + 0.05*0.5 = 10% der Werte.

Die 1000 Datensätze je Kombination der Rahmenbedingungen werden gemäß einer Multiple Imputation, siehe Abschnitt 2.3.2, vervollständigt. D.h. es werden für jeden Datensatz mit fW 5 vervollständigte Datensätze erzeugt. Eine Variation der Anzahl an vervollständigten Datensätzen abhängig vom Anteil an fehlenden Werten erscheint nicht relevant, da selbst bei einem Anteil von 50% fehlende Werte die relative Effizienz von 90% akzeptabel erscheint (vergl. Tab. 1, Seite 26).

Beim binären Merkmalstyp von Sh werden die unvollständigen Datensätze mit Hilfe der logistischen Regression, der Discriminant Function Method, der linearen Regression und der MCMC Methode vervollständigt. Da die DFM keine adäquate Ersetzungsmethode für ordinale Variablen ist, werden nur die anderen zuvor aufgezählten Ersetzungsmethoden bei dem ordinalen Merkmalstyp von Sh zur Schätzung der fehlenden Werte verwendet.

Für jeden der 5 * 1000 = 5000 vervollständigten Datensätze werden die in den Abschnitten zuvor aufgeführten Vergleichsparameter Kappa-Koeffizient, Mittelwert und RMSE zur Beurteilung der Übereinstimmung der geschätzten mit den tatsächlichen Werten berechnet. Darüber hinaus wird mit jedem vervollständigten Datensatz eine multiple lineare Regression durchgeführt, um die Validität der Auswertungsergebnisse zu untersuchten. Die Zusammenfassung der berechneten Parameterschätzer und deren Varianzen von den fünf vervollständigten Datensätzen je ursprünglichem Datensatz mit fW erfolgt gemäß Rubin [82], siehe auch Abschnitt 2.3.2.

Die folgenden Absätze beinhalten technische Hinweise zur Durchführung der Simulationen. Es wird dabei zum Teil auf die in den Simulationen verwendeten Programme verwiesen, die auf der beiliegenden CD im Anhang gespeichert sind. Die erzeugten Datensätze mit fW bzw. ersetzten Werten sind aufgrund der großen Datenmenge nicht auf der CD abgelegt.

In den Programmen Erzeugung datasets mit fW.sas und binär_Erzeugung datasets mit fW.sas werden die jeweils 1000 Datensätze mit den verschiedenen Missing Data Mechanism und Anteilen an fW für den jeweiligen Merkmalstyp von Sh erzeugt. Dazu wird eine binäre Variable *miss* mit Hilfe der Funktion `rantbl(seed, p1)` simuliert. Der explizit gewählte Seed, welcher der Anzahl an bereits simulierten Datensätzen mit fW entspricht, ermöglicht eine Wiederholung dieser Simulationsergebnisse. Als zweites Argument wird der Funktion die Wahrscheinlichkeit übergeben, mit der die Ausprägung 1 auftreten soll. Anschließend werden für Beobachtungen, bei denen die Variable *miss* die Ausprägung 1 hat, die Werte von Steatosis hepatis gelöscht.

Die in dieser Arbeit betrachteten Ersetzungsmethoden werden alle mit der Prozedur `PROC MI`, siehe [3, 77, 84, 86], durchgeführt. Diese Prozedur kann eine beliebige Anzahl an vervollständigten Datensätzen je ursprünglichem Datensatz mit fW erzeugen. Hier werden jeweils 5 vervollständigte Datensätze gene-

riert, welche separat unter verschiedenen Namen abgespeichert werden, damit sie später der Ersetzungsmethode und dem Anteil an ursprünglich fehlenden Werten zugeordnet werden können. Es werden alle Standardeinstellungen der SAS-Prozedur übernommen, d.h. bei der DFM wird Jeffrey's nichtinformative prior Verteilung verwendet, die MCMC-Methode benutzt die aus dem EM-Algorithmus resultierenden Schätzer als Startwerte und die fünf Ersetzungen je Datensatz mit fW erfolgen gemäß einer Single-Chain mit einer Anzahl von 200 burn-in Iterationen und 100 Iterationen zwischen den einzelnen Ersetzungen.

Die (gewichteten) Kappa-Koeffizienten werden mit der Prozedur `PROC FREQ` und der Option `AGREE` berechnet, siehe [85]. Mit dem Makro kappafuenf.mac.sas kann der einfache Kappa-Koeffizient basierend auf allen Beobachtungen bzw. basierend nur auf den geschätzten Werten berechnet werden. Das Makro wkappafuenf_07.mac.sas dient zur Berechnung des gewichteten Kappa-Koeffizienten basierend auf allen Beobachtungen. Wenn nur die Übereinstimmung der geschätzten Werte bei dem ordinalen Merkmal untersucht werden soll, so tritt teilweise das Problem auf, dass die zugrunde liegende Kontingenztafel nicht quadratisch ist und somit der Kappa-Koeffizient nicht standardmäßg berechnet werden kann. Deshalb wird in diesen Situationen der Prozedur `PROC FREQ` mit dem `WEIGHT-` Statement die einzelnen Zellhäufigkeiten einer künstlich erstellten quadratischen Kontingenztafel übergeben und zusätzlich die Optionen `Zeroes` und `Sparse` gewählt, um dennoch den Kappa-Koeffizienten berechnen zu können, siehe Makro wkappaers_07.mac.sas auf der beiliegenden CD.

Die Zusammenführung der Kappa-Koeffizienten und zugehörigen Standardfehler der jeweils fünf vervollständigten Datensätze erfolgt gemäß Rubin [82] und wird mit Hilfe der Prozedur `PROC UNIVARIATE` und einigen sogenannten Data Steps durchgeführt, innerhalb derer explizit die Varianz zwischen den Ersetzungen bzw. innerhalb einer Ersetzung berechnet werden.

Das Makro msemarsh01.mac.sas enthält die Berechnung der Mittelwerte der Parameterschätzer mit Hilfe der Prozedur `PROC MEANS`. Der RMSE wird anschließend in einem Data Step berechnet, in dem die Formel 4.2 zur Berechnung des RMSE implementiert ist. Die Zusammenführung der 5 Parameter je ursprünglichem Datensatz mit fW und dessen Mittelwert wird ebenfalls in einem Data Step durchgeführt.

Die multiple lineare Regression, siehe Abschnitt 4.3.3, wird mit der Prozedur `PROC REG` in dem Makro linregmodelli.mac.sas durchgeführt. Anschließend werden mit der Prozedur `PROC MIANALYZE` die Schätzer der Regressionskoeffizienten und deren Standardabweichungen der jeweils 5 vervollständigten Datensätze zusammengeführt.

Für jede Kombination an Rahmenbedingungen (Merkmalstyp von Sh, Ersetzungsmethode, MDM und Anteil an fW) sind die einzelnen Schritte der Simulationen in einem Programm auf der beiliegenden CD abgespeichert. Anhand des Namens kann auf die jeweilige Kombination zurückgeschlossen werden. Beispeilsweise werden in dem Programm simsh01logregconvmar30.sas die Simulationen für den binären (0,1) Merkmalstyp von Sh (Sh01), bei dem mittels der logistischen Regression die fW unter einem MDM gemäß eines convexen MAR und 30% fW ersetzt werden, durchgeführt.

# Kapitel 5

# Ergebnisse

In diesem Kapitel werden die Ergebnisse der Simulationen zum Vergleich der 4 betrachteten Ersetzungsmethoden und der CCA präsentiert. Im ersten Abschnitt werden die Ersetzungsmethoden anhand der Übereinstimmung der geschätzten mit den tatsächlichen Werten verglichen. Die Gegenüberstellung der Validität der Ergebnisse von einem statistischen Auswertungsverfahren basierend auf den vervollständigten Datensätzen mit der jeweiligen Ersetzungsmethode, wird im zweiten Abschnitt gezeigt.

Bei jedem betrachteten Vergleichskriterium werden zunächst die Ergebnisse für den binären Merkmalstyp von Steatosis hepatis (Sh) angegeben und anschließend die Ergebnisse vom ordinalen Merkmalstyp der Variablen Sh.

## 5.1 Vergleich der Übereinstimmung

Die Übereinstimmung der geschätzten mit den tatsächlichen Werten wird anhand des Kappa-Koeffizienten, siehe Abschnitt 5.1.1, und des Mittelwertes, siehe Abschnitt 5.1.2, beurteilt. Der Kappa-Koeffizient wird sowohl basierend auf der Gesamtfallzahl N = 464 unter Berücksichtigung aller beobachteten und geschätzten Werte, als auch ausschließlich basierend auf den geschätzten Werten berechnet.

### 5.1.1 Kappa-Koeffizient

In den Ergebnistabellen Tab. 12, Tab. 13, Tab. 14, Tab. 15 werden jeweils die Mittelwerte der Kappa-Koeffizienten von den 1000 vervollständigten Datensätzen je Kombination der Rahmenbedingungen (Missing Data Mechanism und Anteil an fW) und je Ersetzungsmethode[1] dargestellt. Zusätzlich wird der mittlere Standardfehler (SE) des jeweiligen Schätzers angegeben. Letzterer ermöglicht Aussagen bezüglich der (Un-)Sicherheit der Schätzung zu treffen und ob sich die Kappa-Koeffizienten deutlich (signifikant) unterscheiden.

Bei dem **binären Merkmalstyp** von Steatosis hepatis wird der einfache Kappa-Koeffizient zur Beurteilung der Übereinstimmung berechnet. In Tab. 12 werden die Simulationsergebnisse **basierend auf allen Beobachtungen** (N = 464) präsentiert. Die Ergebnisse zeigen, dass die Tendenzen im Hinblick auf den Einfluss des Anteils an fW im Datensatz und dem Vorliegenden MDM für alle Ersetzungsmethoden gleich sind. Den größten Einfluss auf die Höhe des Kappa-Koeffizienten hat der Anteil an fW im Datensatz. Je mehr fW im Datensatz auftreten, desto kleiner ist der Kappa-Koeffizient, d.h. desto schlechter ist die Übereinstimmung. Für die Anteile von 10% und 30% an fW betragen die Kappa-Koeffizienten im Mittel 0.92 bzw. 0.75, welche auf eine gute Übereinstimmung hindeuten. Bei einem Anteil von 50% fW beträgt der Kappa-Koeffizient nur noch durchschnittlich 0.59, welcher eher nur auf eine leichte Übereinstimmung hinweist.

Betrachtet man die verschiedenen Missing Data Mechanismen, so zeigt sich, dass unter dem linearen MAR alle Ersetzungsmethoden innerhalb jeden Anteils an fW den größten Kappa-Koeffizienten erzielen. Beispielsweise betragen bei einem Anteil von 10% fW die Kappa-Koeffizienten unter Vorliegen eines linearen MAR abhängig von der Ersetzungsmethode Werte zwischen 0.918 und

[1] Zur Erinnerung: Die Complete Case Analyse wird hier nicht in den Vergleich miteinbezogen, da sie keine Ersetzung durchführt.

Tab. 12: Kappa-Koeffizienten und deren Standardfehler (SE) für den binären Merkmalstyp basierend auf allen Werten.

| MDM | Ersetzungs-methode | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | | 10% | | 30% | | 50% | |
| | | Kappa | SE | Kappa | SE | Kappa | SE |
| MCAR | | | | | | | |
| | log. Reg. | 0.920 | 0.026 | 0.759 | 0.044 | 0.597 | 0.055 |
| | DFM | 0.921 | 0.026 | 0.762 | 0.044 | 0.602 | 0.055 |
| | lin. Reg. | 0.911 | 0.028 | 0.736 | 0.046 | 0.563 | 0.057 |
| | MCMC | 0.911 | 0.027 | 0.737 | 0.046 | 0.564 | 0.056 |
| lin. MAR | | | | | | | |
| | log. Reg. | 0.929 | 0.025 | 0.786 | 0.042 | 0.642 | 0.054 |
| | DFM | 0.930 | 0.024 | 0.790 | 0.041 | 0.644 | 0.054 |
| | lin. Reg. | 0.918 | 0.026 | 0.758 | 0.044 | 0.599 | 0.055 |
| | MCMC | 0.919 | 0.026 | 0.758 | 0.044 | 0.599 | 0.055 |
| konv. MAR | | | | | | | |
| | log. Reg. | 0.922 | 0.026 | 0.765 | 0.044 | 0.607 | 0.057 |
| | DFM | 0.923 | 0.025 | 0.768 | 0.044 | 0.611 | 0.056 |
| | lin. Reg. | 0.914 | 0.027 | 0.742 | 0.045 | 0.573 | 0.057 |
| | MCMC | 0.913 | 0.027 | 0.742 | 0.046 | 0.573 | 0.058 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.919 | 0.026 | 0.756 | 0.044 | 0.591 | 0.055 |
| | DFM | 0.922 | 0.026 | 0.762 | 0.044 | 0.595 | 0.056 |
| | lin. Reg. | 0.911 | 0.027 | 0.735 | 0.045 | 0.557 | 0.056 |
| | MCMC | 0.911 | 0.028 | 0.735 | 0.046 | 0.558 | 0.056 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.913 | 0.027 | 0.738 | 0.045 | 0.560 | 0.058 |
| | DFM | 0.915 | 0.027 | 0.742 | 0.045 | 0.565 | 0.057 |
| | lin. Reg. | 0.906 | 0.028 | 0.721 | 0.047 | 0.536 | 0.058 |
| | MCMC | 0.907 | 0.028 | 0.722 | 0.047 | 0.536 | 0.058 |

0.930, was der größte Kappa-Koeffizient der jeweiligen Methode ist. Gleichzeitig ist bei linearem MAR der Standardfehler des Schätzers am kleinsten; bei 10% fW beträgt er $\leq 0.026$. Etwas schlechter sind die Ergebnisse unter konvexem MAR, MCAR und linearem MNAR. Am schlechtesten ist die Übereinstimmung, wenn die fW gemäß konvexem MNAR vorliegen.

Darüber hinaus zeigen die Simulationsergebnisse, dass bei jeder Kombination der Rahmenbedingungen die Kappa-Koeffizienten und deren SE von den Ersetzungsmethoden speziell für kategoriale Variablen (log. Reg. und DFM) etwas

besser, d.h. größer respektive kleiner, sind als von den Ersetzungsmethoden, die primär zur Ersetzung von stetigen Variablen entwickelt wurden (lin. Reg. und MCMC).

Ob signifikante (deutliche) Unterschiede zwischen den Ersetzungsmethoden vorliegen, kann mit Hilfe eines 95% Konfidenzintervalles (KI) um den Kappa-Koeffizienten beurteilt werden. Davon ausgehend, dass der Schätzer des Kappa-Koeffizienten approximativ normalverteilt ist, berechnet sich das 95% KI gemäß Schätzer $\pm$ 1.96 * SE. Unter Berücksichtigung der Größe der SE von mindestens 0.024, kann kein deutlicher Unterschied zwischen den Kappa-Koeffizienten gezeigt werden, da sich die 95% KI von den einzelnen Ersetzungsmethoden je Rahmenbedingung überlappen. Die beobachteten Unterschiede in den Ersetzungsmethoden weisen daher nur Tendenzen auf.

Um den Effekt der automatisch übereinstimmenden Werte zu eliminieren wird zusätzlich der Kappa-Koeffizient ausschließlich **basierend auf den geschätzten Werten** berechnet. Die Ergebnisse sind in Tab. 13 dargestellt. Sie zeigen, dass zur Beurteilung der Übereinstimmung ausschließlich der geschätzten mit den tatsächlichen Werten, der Anteil an fW hauptsächlich die Größe der Standardfehler (SE) beeinflusst, nicht aber den Kappa-Koeffizienten selber. Mit zunehmendem Anteil an fW vergrößert sich die Fallzahl, auf der die Schätzung basiert und der SE nimmt ab, während der Kappa-Koeffizient tendenziell geringfügig größer wird. Die Kappa-Koeffizienten nehmen Werte zwischen 0.08 und 0.24 an und sind unter Berücksichtigung der SE nicht deutlich von 0 verschieden, da das 95% KI von jedem Kappa-Koeffizienten den Wert 0 beinhaltet. Dies deutet auf eine schlechte bzw. rein zufällige Übereinstimmung hin!

Jede Ersetzungsmethode weist den größten Kappa-Koeffizient bei Vorliegen eines konvexen MAR auf, gefolgt von MCAR, linearen MAR, konvexen MNAR und linearen MNAR.

Tab. 13: Kappa-Koeffizienten und deren Standardfehler (SE) für den binären Merkmalstyp basierend auf den geschätzten Werten.

| MDM | Ersetzungs-methode | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | | 10% | | 30% | | 50% | |
| | | Kappa | SE | Kappa | SE | Kappa | SE |
| MCAR | | | | | | | |
| | log. Reg. | 0.200 | 0.210 | 0.198 | 0.125 | 0.196 | 0.098 |
| | DFM | 0.201 | 0.211 | 0.203 | 0.125 | 0.202 | 0.098 |
| | lin. Reg. | 0.145 | 0.212 | 0.146 | 0.124 | 0.147 | 0.097 |
| | MCMC | 0.146 | 0.210 | 0.150 | 0.123 | 0.149 | 0.097 |
| lin. MAR | | | | | | | |
| | log. Reg. | 0.159 | 0.218 | 0.160 | 0.129 | 0.160 | 0.101 |
| | DFM | 0.158 | 0.214 | 0.163 | 0.128 | 0.157 | 0.101 |
| | lin. Reg. | 0.112 | 0.209 | 0.118 | 0.124 | 0.116 | 0.097 |
| | MCMC | 0.113 | 0.211 | 0.116 | 0.125 | 0.117 | 0.097 |
| konv. MAR | | | | | | | |
| | log. Reg. | 0.234 | 0.210 | 0.234 | 0.126 | 0.232 | 0.102 |
| | DFM | 0.240 | 0.210 | 0.241 | 0.125 | 0.236 | 0.101 |
| | lin. Reg. | 0.176 | 0.210 | 0.178 | 0.124 | 0.176 | 0.100 |
| | MCMC | 0.177 | 0.211 | 0.178 | 0.126 | 0.176 | 0.100 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.104 | 0.209 | 0.109 | 0.124 | 0.108 | 0.095 |
| | DFM | 0.115 | 0.212 | 0.113 | 0.124 | 0.109 | 0.096 |
| | lin. Reg. | 0.080 | 0.204 | 0.082 | 0.119 | 0.078 | 0.092 |
| | MCMC | 0.077 | 0.205 | 0.081 | 0.120 | 0.079 | 0.093 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.139 | 0.213 | 0.139 | 0.125 | 0.131 | 0.100 |
| | DFM | 0.150 | 0.213 | 0.146 | 0.125 | 0.138 | 0.099 |
| | lin. Reg. | 0.106 | 0.213 | 0.107 | 0.125 | 0.099 | 0.097 |
| | MCMC | 0.107 | 0.212 | 0.108 | 0.125 | 0.101 | 0.098 |

Vergleicht man die Ersetzungsmethoden innerhalb des jeweiligen MDM, so führt die Schätzung der fW mittels der DFM zu dem größten Kappa-Koeffizienten. Geringfügig kleiner ist der Kappa-Koeffizient bei der Verwendung einer logistischen Regression. Die Kappa-Koeffizienten basierend auf vervollständigten Datensätzen mittels Ersetzungsmethoden speziell für kategoriale Variablen sind um ungefähr 0.05 größer als diejenigen, bei denen die lineare Regression bzw. die MCMC Methode angewandt wurde. Unter Berücksichtigung der relativ großen SE gibt es jedoch keine deutlichen Unterschiede in den Kappa-Koeffizienten innerhalb eines MDM und Anteils an fW zwischen den verschie-

denen Ersetzungsmethoden.

Bei dem **ordinalen Merkmalstyp** von Steatosis hepatis wird der gewichtete Kappa-Koeffizient von Cicchetti und Allison zur Beurteilung der Übereinstimmung herangezogen. Aufgrund der Gewichte wird in der Berechnung des Kappa-Koeffizienten berücksichtigt, um wie viele Einheiten sich die geschätzten Werte von den tatsächlichen Werten unterscheiden. Je Größer diese Abweichung ist, desto kleiner ist der Kappa-Koeffizient. In Tab. 14 werden die gewichteten Kappa-Koeffizienten und die zugehörigen Standardfehler (SE) je Ersetzungsmethode[2] und Kombination der Rahmenbedingungen **basierend auf allen Beobachtungen** (N = 464) präsentiert.

Die Ergebnisse der Simulationen für den ordinalen Merkmalstyp von Sh deuten auf die gleichen Tendenzen hin, wie die Simulationsergebnisse für den binären Merkmalstyp. Die Übereinstimmung wird primär durch den Anteil an fW im Datensatz beeinflusst. Mit zunehmendem Anteil an fW verringert sich der gewichtete Kappa-Koeffizient, während dessen Standardfehler zunimmt. Für die Anteile von 10% und 30% an fW betragen die Kappa-Koeffizienten durchschnittlich 0.9 bzw. 0.75, welche auf eine gute Übereinstimmung hinweisen. Bei einem Anteil von 50% fW beträgt der Kappa-Koeffizient im Mittel nur noch 0.55, welcher eher nur auf eine leichte Übereinstimmung hinweist.

Für jede Ersetzungsmethode und innerhalb des gleichen Anteils an fW treten die größten gewichteten Kappa-Koeffizienten mit den gleichzeitig kleinsten SE bei einem MDM gemäß linearem MAR auf. Etwas kleiner sind die Kappa-Koeffizienten bei linearem MNAR, MCAR und konvexem MAR. Am kleinsten sind die Kappa-Koeffizienten mit den gleichzeitig größten SE bei Vorliegen eines konvexen MNAR.

[2]Zur Erinnerung: die Discriminant Function Method wird nur zur Ersetzung fW nominaler Variablen empfohlen. Deshalb wird sie bei dem ordinalen Merkmalstyp von Sh nicht angewendet.

Tab. 14: Kappa-Koeffizienten und deren Standardfehler (SE) für den ordinalen Merkmalstyp basierend auf allen Werten.

| MDM | Ersetzungs-methode | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | | 10% | | 30% | | 50% | |
| | | Kappa | SE | Kappa | SE | Kappa | SE |
| MCAR | | | | | | | |
| | log. Reg. | 0.918 | 0.025 | 0.756 | 0.043 | 0.594 | 0.052 |
| | lin. Reg. | 0.906 | 0.025 | 0.723 | 0.041 | 0.546 | 0.049 |
| | MCMC | 0.906 | 0.025 | 0.724 | 0.041 | 0.546 | 0.049 |
| lin. MAR | | | | | | | |
| | log. Reg. | 0.934 | 0.023 | 0.804 | 0.039 | 0.675 | 0.051 |
| | lin. Reg. | 0.918 | 0.023 | 0.757 | 0.039 | 0.601 | 0.049 |
| | MCMC | 0.918 | 0.023 | 0.757 | 0.039 | 0.602 | 0.049 |
| konv. MAR | | | | | | | |
| | log. Reg. | 0.917 | 0.026 | 0.750 | 0.043 | 0.584 | 0.054 |
| | lin. Reg. | 0.907 | 0.025 | 0.724 | 0.040 | 0.549 | 0.050 |
| | MCMC | 0.907 | 0.025 | 0.725 | 0.041 | 0.549 | 0.050 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.923 | 0.026 | 0.770 | 0.043 | 0.615 | 0.054 |
| | lin. Reg. | 0.908 | 0.025 | 0.728 | 0.041 | 0.600 | 0.049 |
| | MCMC | 0.908 | 0.025 | 0.728 | 0.041 | 0.553 | 0.049 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.909 | 0.027 | 0.726 | 0.045 | 0.538 | 0.056 |
| | lin. Reg. | 0.899 | 0.026 | 0.701 | 0.043 | 0.509 | 0.051 |
| | MCMC | 0.899 | 0.026 | 0.702 | 0.042 | 0.509 | 0.051 |

Generell ist die Übereinstimmung zwischen ersetzten und Originalwerten bei der logistischen Regression etwas besser als bei den Ersetzungsmethoden für stetige Variablen, aber da sich die 95% KI aufgrund der Größe der SE überlappen, gibt es keinen signifikanten Unterschied in den Kappa-Koeffizienten.

Um ausschließlich die Übereinstimmung der geschätzten mit den tatsächlichen Werten zu beurteilen, werden zusätzlich die gewichteten Kappa-Koeffizienten **basierend auf den geschätzten Werten** berechnet. Die Ergebnisse sind in Tab. 15 dargestellt. Sie zeigen, dass auch bei dem ordinalen Merkmalstyp von Sh, wenn nur die geschätzten Werte betrachtete werden, der Anteil an fW primär die Größe des SE beeinflusst und nicht den Kappa-Koeffizienten. Der Standardfehler nimmt mit zunehmendem Anteil an fW -und folglich mit zu-

nehmender Fallzahl- ab, während der Kappa-Koeffizient nur marginal größer wird. Die Kappa-Koeffizienten nehmen ähnlich zum binären Merkmalstyp Werte zwischen 0.07 und 0.22 an, welche eher auf eine schlechte (bzw. rein zufällige) Übereinstimmung hinweisen.

Tab. 15: Kappa-Koeffizienten und deren Standardfehler (SE) für den ordinalen Merkmalstyp basierend auf den geschätzten Werten.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | Kappa | SE | Kappa | SE | Kappa | SE |
| MCAR | | | | | | | |
| | log. Reg. | 0.190 | 0.182 | 0.197 | 0.109 | 0.196 | 0.086 |
| | lin. Reg. | 0.125 | 0.161 | 0.128 | 0.094 | 0.129 | 0.074 |
| | MCMC | 0.125 | 0.162 | 0.129 | 0.094 | 0.129 | 0.074 |
| lin. MAR | | | | | | | |
| | log. Reg. | 0.153 | 0.191 | 0.160 | 0.115 | 0.165 | 0.092 |
| | lin. Reg. | 0.101 | 0.165 | 0.105 | 0.097 | 0.107 | 0.076 |
| | MCMC | 0.101 | 0.166 | 0.105 | 0.097 | 0.106 | 0.076 |
| konv. MAR | | | | | | | |
| | log. Reg. | 0.220 | 0.181 | 0.225 | 0.109 | 0.220 | 0.088 |
| | lin. Reg. | 0.158 | 0.160 | 0.161 | 0.095 | 0.160 | 0.077 |
| | MCMC | 0.158 | 0.159 | 0.163 | 0.095 | 0.161 | 0.077 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.101 | 0.182 | 0.108 | 0.107 | 0.107 | 0.083 |
| | lin. Reg. | 0.062 | 0.153 | 0.066 | 0.090 | 0.106 | 0.076 |
| | MCMC | 0.059 | 0.155 | 0.064 | 0.090 | 0.066 | 0.069 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.126 | 0.180 | 0.123 | 0.108 | 0.110 | 0.085 |
| | lin. Reg. | 0.085 | 0.159 | 0.083 | 0.095 | 0.079 | 0.074 |
| | MCMC | 0.086 | 0.160 | 0.084 | 0.093 | 0.079 | 0.075 |

Analog zum binären Merkmalstyp tritt bei jeder Ersetzungsmethode die größte Übereinstimmung bei Vorliegen eines konvexen MAR, gefolgt von MCAR, linearen MAR, konvexen MNAR und linearen MNAR auf.

Die logistische Regression führt unter jeder Rahmenbedingung zu etwas größeren Kappa-Koeffizienten und SE als die lineare Regression und die MCMC-Methode. Doch unter Berücksichtigung der SE gibt es keine deutlichen Unterschiede in den Kappa-Koeffizienten der einzelnen Ersetzungsmethoden.

### 5.1.2 Mittelwert/ RMSE

Da in vielen Literaturstellen [5, 12, 27, 33] als Maß der Übereinstimmung der Mittelwert betrachtet wird (selbst bei ordinalen Variablen), werden aus Gründen der Vergleichbarkeit und um zu beurteilen ob die geschätzten Werte die tatsächlichen Werte eher über- bzw. unterschätzten, die Gesamtmittelwerte und deren RMSE für die beiden Merkmalstypen von Sh in den Tabellen Tab. 16 und Tab. 17 angegeben. In den Tabellen sind jeweils die mittleren Abweichungen des Gesamtmittelwertes von dem wahren Mittelwert und der über die 1000 Datensätze mit ursprünglich fW gemittelte RMSE aufgelistet. Negative Werte bei der Abweichung des Gesamtmittelwertes entsprechen einer Unterschätzung des wahren Mittelwertes und positive Werte einer Überschätzung.

Der Mittelwert des **binären Merkmalstyps von Sh** beträgt im Originaldatensatz 0.2672, d.h. bei 26.72% = 124/464 der Beobachtungen nimmt Sh den Wert 1 an. Die Ergebnisse in Tab. 16 zeigen, dass der Mittelwert generell eher etwas überschätzt wird. Nur die DFM führt in den meisten Situationen zu einer Unterschätzung des Mittelwertes[3].

Wie zu erwarten war, hat der Anteil an fehlenden Werten einen deutlichen Einfluss auf die Abweichung der Mittelwerte. Mit zunehmendem Anteil an fW im Datensatz nimmt die Abweichung von im Mittel 0.008 (d.h. $0.008 * 464 \approx 4$ Beobachtungen mehr mit Sh=1) bei 10% fW, über die Abweichung von durchschnittlich 0.02 (d.h. 9 Beobachtungen mehr mit Sh=1) bei 30% fW auf eine Abweichung von im Mittel 0.04 (d.h. 18 Beobachtungen mehr mit Sh=1) bei 50% fW zu. Bei 50% fW haben nun nicht mehr 27% der Beobachtungen eine Fettleber (Sh=1), sondern -abhängig von der Ersetzungsmethode und dem MDM- bis zu 33% der Beobachtungen. Auch der RMSE vergrößert sich deutlich mit dem Anteil an fW.

[3]Dies bedeutet inhaltlich, dass durch die Ersetzung mittels DFM bei Personen mit fW von Sh eher das Vorliegen keiner Fettleber vermutet wird.

Tab. 16: Mittlere Abweichung des Gesamtmittelwertes vom wahren Mittelwert (MW) und mittlerer RMSE für den binären Merkmalstyp von Sh. Der wahre Mittelwert beträgt 0.2672.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | MW | RMSE | MW | RMSE | MW | RMSE |
| MCAR | | | | | | | |
| | CCA | 0.0002 | 0.0067 | 0.0002 | 0.0133 | 0.0003 | 0.0205 |
| | log. Reg. | 0.0008 | 0.0067 | 0.0013 | 0.0134 | 0.0021 | 0.0201 |
| | DFM | -0.0003 | 0.0067 | -0.0018 | 0.0133 | -0.0033 | 0.0208 |
| | lin. Reg. | 0.0042 | 0.0081 | 0.0117 | 0.0179 | 0.0202 | 0.0289 |
| | MCMC | 0.0041 | 0.0079 | 0.0116 | 0.0178 | 0.0195 | 0.0282 |
| lin. MAR | | | | | | | |
| | CCA | 0.0074 | 0.0097 | 0.0276 | 0.0300 | 0.0642 | 0.0669 |
| | log. Reg. | 0.0010 | 0.0064 | 0.0018 | 0.0126 | 0.0042 | 0.0208 |
| | DFM | -0.0001 | 0.0066 | -0.0009 | 0.0125 | 0.0004 | 0.0200 |
| | lin. Reg. | 0.0061 | 0.0089 | 0.0186 | 0.0224 | 0.0333 | 0.0387 |
| | MCMC | 0.0058 | 0.0088 | 0.0180 | 0.0218 | 0.0331 | 0.0383 |
| konv. MAR | | | | | | | |
| | CCA | -0.0015 | 0.0069 | -0.0068 | 0.0141 | -0.0156 | 0.0234 |
| | log. Reg. | 0.0003 | 0.0066 | -0.0004 | 0.0123 | -0.0006 | 0.0187 |
| | DFM | -0.0005 | 0.0067 | -0.0023 | 0.0126 | -0.0053 | 0.0192 |
| | lin. Reg. | 0.0033 | 0.0073 | 0.0085 | 0.0149 | 0.0124 | 0.0218 |
| | MCMC | 0.0032 | 0.0073 | 0.0083 | 0.0149 | 0.0119 | 0.0218 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.0050 | 0.0083 | 0.0173 | 0.0215 | 0.0385 | 0.0433 |
| | DFM | 0.0037 | 0.0077 | 0.0135 | 0.0185 | 0.0333 | 0.0392 |
| | lin. Reg. | 0.0092 | 0.0114 | 0.0308 | 0.0332 | 0.0623 | 0.0653 |
| | MCMC | 0.0092 | 0.0115 | 0.0306 | 0.0332 | 0.0620 | 0.0650 |
| konv. MNAR | | | | | | | |
| | log. Reg. | -0.0010 | 0.0068 | -0.0047 | 0.0137 | -0.0115 | 0.0217 |
| | DFM | -0.0018 | 0.0070 | -0.0073 | 0.0143 | -0.0159 | 0.0246 |
| | lin. Reg. | 0.0027 | 0.0074 | 0.0064 | 0.0142 | 0.0066 | 0.0202 |
| | MCMC | 0.0026 | 0.0072 | 0.0060 | 0.0143 | 0.0063 | 0.0199 |

Abhängig von der jeweiligen Ersetzungsmethode führt mal der eine und mal der andere Missing Data Mechanism zu der geringsten Abweichung des Mittelwertes. Die CCA hat die geringste Abweichung unter MCAR und etwas größere bei Vorliegen eines konvexen bzw. linearen MAR. Die Ersetzungsmethoden speziell für kategoriale Variablen (logistische Regression und DFM)

schätzen den Mittelwert am besten unter MAR, schlechter unter MCAR und am schlechtesten, wenn MNAR vorliegt. Die Mittelwertschätzer der linearen Regression und MCMC Methode haben die geringsten Abweichungen bei Vorliegen eines konvexen M(N)AR, gefolgt von MCAR und sind am schlechtesten bei Vorliegen eines linearen M(N)AR.

Vergleicht man die Ersetzungsmethoden innerhalb einer Rahmenbedingung, so zeigt sich, dass diejenigen für kategoriale Variablen die geringsten Abweichungen mit den tendenziell kleinsten RMSE haben, gefolgt von der CCA. Die Ersetzungsmethoden für stetige Variablen weisen unter jeder Rahmenbedingung die größten Abweichungen mit den ebenfalls größten RMSE auf. Die Unterschiede zwischen den Ersetzungsmethoden sind jedoch nicht signifikant, da sich die 95% KI um die Mittelwertschätzer aufgrund der Größe der Standardfehler überlappen.

Der wahre Mittelwert des **ordinalen Merkmalstyps von Sh** mit den 4 Ausprägungen 0, 1, 2 und 3 beträgt 0.3944. Die Simulationsergebnisse mit den Abweichungen vom wahren Mittelwert sind in Tab. 17 enthalten. Analog zum binären Merkmalstyp wird der Mittelwert eher über- als unterschätzt und der Anteil an fW im Datensatz hat den deutlichsten Einfluss auf die mittleren Abweichungen und den RMSE.

Der Einfluss des Missing Data Mechanism auf die Abweichungen ist abhängig von der Ersetzungsmethode. Die CCA führt unter MCAR im Mittel zu einer vernachlässigbar geringen Abweichung. Bei Vorliegen eines konvexen MAR unterschätzt sie den Mittelwert leicht und bei linearen MAR weist sie ihre größte Abweichung in Form einer mittleren Überschätzung auf. Die logistische Regression und die beiden Ersetzungsmethoden für stetige Variablen führen zu den geringsten Abweichungen unter konvexem M(N)AR und zu etwas größeren unter MCAR und linearem MAR. Am größten ist die Abweichung unter linearem MNAR.

Tab. 17: Mittlere Abweichung des Gesamtmittelwertes vom wahren Mittelwert (MW) und mittlerer RMSE für den ordinalen Merkmalstyp von Sh. Der wahre Mittelwert beträgt 0.3944.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | MW | RMSE | MW | RMSE | MW | RMSE |
| MCAR | | | | | | | |
| | CCA | 0.0002 | 0.0111 | 0.0004 | 0.0219 | 0.0003 | 0.0337 |
| | log. Reg. | 0.0019 | 0.0108 | 0.0061 | 0.0216 | 0.0097 | 0.0339 |
| | lin. Reg. | 0.0158 | 0.0190 | 0.0462 | 0.0511 | 0.0778 | 0.0846 |
| | MCMC | 0.0156 | 0.0189 | 0.0463 | 0.0511 | 0.0767 | 0.0836 |
| lin. MAR | | | | | | | |
| | CCA | 0.0134 | 0.0164 | 0.0513 | 0.0546 | 0.1198 | 0.1237 |
| | log. Reg. | 0.0024 | 0.0094 | 0.0067 | 0.0184 | 0.0119 | 0.0302 |
| | lin. Reg. | 0.0196 | 0.0218 | 0.0618 | 0.0643 | 0.1109 | 0.1145 |
| | MCMC | 0.0196 | 0.0218 | 0.0622 | 0.0647 | 0.1099 | 0.1136 |
| konv. MAR | | | | | | | |
| | CCA | -0.0039 | 0.0123 | -0.0162 | 0.0267 | -0.0376 | 0.0473 |
| | log. Reg. | 0.0014 | 0.0113 | 0.0035 | 0.0218 | 0.0028 | 0.0354 |
| | lin. Reg. | 0.0129 | 0.0172 | 0.0356 | 0.0412 | 0.0491 | 0.0592 |
| | MCMC | 0.0130 | 0.0173 | 0.0348 | 0.0407 | 0.0489 | 0.0587 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.0094 | 0.0135 | 0.0432 | 0.0383 | 0.0715 | 0.0777 |
| | lin. Reg. | 0.0263 | 0.0281 | 0.0867 | 0.0889 | 0.1125 | 0.1162 |
| | MCMC | 0.0264 | 0.0283 | 0.0868 | 0.0889 | 0.1649 | 0.1676 |
| konv. MNAR | | | | | | | |
| | log. Reg. | -0.0025 | 0.0116 | -0.0111 | 0.0238 | -0.0284 | 0.0420 |
| | lin. Reg. | 0.0121 | 0.0170 | 0.0320 | 0.0390 | 0.0396 | 0.0511 |
| | MCMC | 0.0121 | 0.0169 | 0.0317 | 0.0386 | 0.0387 | 0.0509 |

Die logistische Regression führt zu den absolut betrachtet kleinsten Abweichungen und RMSE. Die CCA führt außer bei MCAR zu etwas größeren Abweichungen und am schlechtesten und ungenauesten sind die Mittelwertschätzungen basierend auf mittels linearer Regression oder der MCMC Methode vervollständigten Datensätzen. Aufgrund der RMSE lassen sich aber keine deutlichen Unterschiede in der Schätzung zwischen den Ersetzungsmethoden unter der jeweiligen Rahmenbedingung erkennen.

## 5.2 Vergleich der Validität

Dieser Abschnitt stellt die Ergebnisse bezüglich der Validität der Auswertungsergebnisse basierend auf den mittels verschiedener Ersetzungsmethoden vervollständigten Datensätze bzw. der CCA vor. Als ein typisches Auswertungsverfahren in der medizinischen Forschung wird eine multiple lineare Regression durchgeführt. Die Beurteilung der Validität erfolgt anhand der mittleren Abweichung des jeweiligen Vergleichsparameters zu dem wahren Parameterwert. Negative Abweichungen bedeuten eine Unterschätzung des wahren Wertes, positive Abweichungen eine Überschätzung. Die Validität wird anhand dem zu der Variablen Sh gehörigen Regressionskoeffizienten, der zugehörigen F-Statistik und dem $R^2$, welches die Modellanpassung bewertet, beurteilt. Bei jedem Vergleichparameter werden zunächst die Ergebnisse für den binären Merkmalstyp und anschließend die für den ordinalen Merkmalstyp von Sh vorgestellt.

### 5.2.1 Regressionskoeffizient

In Tab. 18, Tab. 19 und Tab. 20 sind die mittleren Abweichungen der 1000 Regressionskoeffizienten von dem wahren Regressionskoeffizienten und der mittlere Standardfehler (SE) der Regressionskoeffizienten für den binären und ordinalen Merkmalstyp von Sh aufgelistet. Da die ordinale Variable Sh durch 3 Dummyvariablen im Modell repräsentiert wird, werden bei diesem Merkmalstyp auch drei Regressionskoeffizienten je Rahmenbedingung auf deren Validität überprüft.

Die Ergebnisse für den **binären Merkmalstyp** von Sh sind in der Tab. 18 abgebildet. Der wahre Regressionskoeffizient beträgt 0.225, d.h. Personen mit Fettleber haben im Mittel einen 0.225 Einheiten größeren Insulinspiegel als Personen ohne Fettleber. Die mittleren Abweichungen sind betragsmäßig nur gering, doch mit zunehmendem Anteil an fW nehmen diese und vorallem der SE des Regressionskoeffizienten von durchschnittlich 0.005 auf 0.03 bzw. von

durchschnittlich 0.018 auf 0.045 zu. Die Ersetzungsmethoden für stetige Variablen führen generell zu einer Unterschätzung des Regressionskoeffizienten, während die Ersetzungsmethoden für kategoriale Variablen und die CCA diesen Parameter mal über und mal unterschätzen.

Tab. 18: Mittlere Abweichung der Regressionskoeffizienten von dem wahren Koeffizienten und der mittlere Standardfehler (SE) des binären Merkmalstyps. Der wahre Regressionskoeffizient beträgt 0.225.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | Reg.koeff. | SE | Reg.koeff. | SE | Reg.koeff. | SE |
| MCAR | | | | | | | |
| | CCA | < -0.001 | 0.018 | < 0.001 | 0.035 | < -0.001 | 0.056 |
| | log. Reg. | -0.002 | 0.018 | -0.005 | 0.035 | -0.008 | 0.054 |
| | DFM | 0.002 | 0.019 | 0.009 | 0.036 | 0.015 | 0.056 |
| | lin. Reg. | -0.008 | 0.019 | -0.022 | 0.032 | -0.035 | 0.047 |
| | MCMC | -0.008 | 0.019 | -0.021 | 0.033 | -0.036 | 0.047 |
| lin. MAR | | | | | | | |
| | CCA | < 0.001 | 0.017 | 0.002 | 0.032 | 0.003 | 0.045 |
| | log. Reg. | < -0.001 | 0.018 | -0.001 | 0.033 | < -0.001 | 0.045 |
| | DFM | 0.003 | 0.018 | 0.009 | 0.033 | 0.012 | 0.045 |
| | lin. Reg. | -0.007 | 0.018 | -0.020 | 0.031 | -0.032 | 0.041 |
| | MCMC | -0.007 | 0.018 | -0.019 | 0.031 | -0.031 | 0.039 |
| konv. MAR | | | | | | | |
| | CCA | -0.003 | 0.016 | -0.011 | 0.033 | -0.026 | 0.048 |
| | log. Reg. | -0.005 | 0.017 | -0.012 | 0.034 | -0.020 | 0.046 |
| | DFM | < -0.001 | 0.018 | 0.001 | 0.035 | < 0.001 | 0.050 |
| | lin. Reg. | -0.011 | 0.018 | -0.027 | 0.031 | -0.045 | 0.043 |
| | MCMC | -0.010 | 0.018 | -0.027 | 0.033 | -0.043 | 0.044 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.004 | 0.019 | 0.015 | 0.033 | 0.024 | 0.043 |
| | DFM | 0.009 | 0.019 | 0.026 | 0.034 | 0.031 | 0.044 |
| | lin. Reg. | -0.003 | 0.019 | -0.008 | 0.031 | -0.015 | 0.037 |
| | MCMC | -0.003 | 0.019 | -0.008 | 0.030 | -0.015 | 0.037 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.002 | 0.019 | 0.006 | 0.033 | 0.005 | 0.047 |
| | DFM | 0.006 | 0.019 | 0.019 | 0.035 | 0.027 | 0.050 |
| | lin. Reg. | -0.006 | 0.018 | -0.016 | 0.032 | -0.027 | 0.042 |
| | MCMC | -0.006 | 0.019 | -0.015 | 0.032 | -0.026 | 0.043 |

Abhängig von der Ersetzungsmethode führen unterschiedliche MDM zu den

geringsten mittleren Abweichungen. Die CCA führt unter MCAR nur zu minimalen mittleren Abweichungen, selbst bei 30% und 50% fW von Sh. Auch bei Vorliegen von MAR sind die Abweichungen unter CCA nur sehr klein. Die Ersetzungsmethoden für kategoriale Variablen zeigen geringfügige Abweichungen bei Vorliegen von MCAR, und den beiden Arten von MAR. Etwas größer sind die Abweichungen unter konvexem und linearem MNAR. Die Ersetzungsmethoden für stetige Variablen führen erstaunlicherweise unter MNAR zu den geringsten Abweichungen, gefolgt von den Abweichungen unter MCAR, linearem MAR und konvexem MAR.

Betragsmäßig sind die Abweichungen von der linearen Regression und der MCMC Methode etwas größer als bei den anderen Methoden. Doch aufgrund der Größe der SE, die sich bis auf bei dem Anteil von 50% fW nur marginal zwischen den Ersetzungsmethoden und der CCA unterscheidet, überlappen sich die 95% KI um die Abweichungen, sodass kein deutlicher Unterschied in der Validität des Regressionskoeffizienten abhängig von der Ersetzungsmethode bzw. CCA bei dem binären Merkmalstyp von Sh zu erkennen ist. (Die beobachteten geringfügigen Abweichungen für den Schätzer des Regressionskoeffizienten führen bei dem vorliegenden Datensatz zu keinen relevanten Fehlinterpretationen.)

Bei dem **ordinalen Merkmalstyp** von Sh werden die Ersetzungsmethoden anhand der Regressionskoeffizienten zu den 3 Dummyvariablen der Referenzkodierung (mit keine Fettleber als Referenz) verglichen. Die Ergebnisse sind in Tab. 19 und Tab. 20 enthalten. Für den Originaldatensatz betrug der Regressionskoeffizient zu der Variablen D1 0.170, zu D2 0.311 und zu D3 0.396. Inhaltlich bedeutet dies, dass je größer der Grad der Fettleibigkeit ist, desto größer ist der Insulinspiegel. Allerdings sind diese Unterschiede nur minimal. Alle betrachteten Methoden führen meistens zu einer Unterschätzung des wahren Regressionskoeffizienten.

Tab. 19: Mittlere Abweichung der Regressionskoeffizienten von den wahren Koeffizienten der drei Dummyvariablen und die mittleren Standardfehler (SE) des ordinalen Merkmalstyps. Der wahre Regressionskoeffizient für D1 beträgt 0.170, für D2 0.311 und für D3 0.396. (**Teil I**)

| MDM | Ersetzungs- | Sh_ | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|---|
| | methode | kat | 10% | | 30% | | 50% | |
| | | | Reg.koeff. | SE | Reg.koeff. | SE | Reg.koeff. | SE |
| MCAR | | | | | | | | |
| | CCA | D1 | < -0.001 | 0.021 | < -0.001 | 0.042 | < -0.001 | 0.063 |
| | CCA | D2 | < 0.001 | 0.029 | < 0.001 | 0.058 | < -0.001 | 0.092 |
| | CCA | D3 | -0.002 | 0.091 | -0.003 | 0.178 | 0.002 | 0.274 |
| | log. Reg. | D1 | -0.003 | 0.021 | -0.009 | 0.035 | -0.013 | 0.044 |
| | log. Reg. | D2 | -0.004 | 0.028 | -0.013 | 0.052 | -0.020 | 0.075 |
| | log. Reg. | D3 | < 0.001 | 0.088 | 0.002 | 0.144 | 0.022 | 0.189 |
| | lin. Reg. | D1 | -0.013 | 0.023 | -0.030 | 0.033 | -0.034 | 0.039 |
| | lin. Reg. | D2 | -0.006 | 0.029 | -0.016 | 0.053 | -0.029 | 0.079 |
| | lin. Reg. | D3 | -0.007 | 0.094 | -0.018 | 0.176 | -0.021 | 0.256 |
| | MCMC | D1 | -0.012 | 0.023 | -0.029 | 0.032 | -0.040 | 0.039 |
| | MCMC | D2 | -0.006 | 0.029 | -0.017 | 0.055 | -0.029 | 0.079 |
| | MCMC | D3 | -0.006 | 0.094 | -0.017 | 0.177 | -0.022 | 0.253 |
| lin. MAR | | | | | | | | |
| | CCA | D1 | < -0.001 | 0.021 | -0.001 | 0.040 | -0.005 | 0.057 |
| | CCA | D2 | -0.001 | 0.025 | -0.006 | 0.043 | -0.013 | 0.057 |
| | CCA | D3 | 0.003 | 0.059 | 0.008 | 0.109 | 0.014 | 0.149 |
| | log. Reg. | D1 | 0.002 | 0.022 | 0.008 | 0.038 | 0.015 | 0.049 |
| | log. Reg. | D2 | -0.005 | 0.025 | -0.012 | 0.043 | -0.018 | 0.055 |
| | log. Reg. | D3 | -0.003 | 0.061 | -0.010 | 0.100 | -0.007 | 0.123 |
| | lin. Reg. | D1 | -0.014 | 0.023 | -0.029 | 0.032 | -0.033 | 0.036 |
| | lin. Reg. | D2 | -0.006 | 0.026 | -0.018 | 0.042 | -0.027 | 0.054 |
| | lin. Reg. | D3 | -0.004 | 0.061 | -0.008 | 0.109 | -0.005 | 0.142 |
| | MCMC | D1 | -0.012 | 0.023 | -0.027 | 0.033 | -0.032 | 0.036 |
| | MCMC | D2 | -0.006 | 0.026 | -0.017 | 0.042 | -0.027 | 0.055 |
| | MCMC | D3 | -0.003 | 0.061 | -0.007 | 0.107 | -0.008 | 0.144 |
| konv. MAR | | | | | | | | |
| | CCA | D1 | -0.005 | 0.018 | -0.015 | 0.037 | -0.036 | 0.053 |
| | CCA | D2 | 0.001 | 0.029 | 0.006 | 0.059 | 0.016 | 0.091 |
| | CCA | D3 | -0.015 | 0.111 | -0.051 | 0.217 | -0.161 | 0.324 |
| | log. Reg. | D1 | -0.005 | 0.019 | -0.014 | 0.035 | -0.027 | 0.045 |
| | log. Reg. | D2 | -0.004 | 0.029 | -0.012 | 0.051 | -0.025 | 0.074 |
| | log. Reg. | D3 | -0.016 | 0.104 | -0.048 | 0.163 | -0.094 | 0.190 |
| | lin. Reg. | D1 | -0.016 | 0.023 | -0.036 | 0.030 | -0.053 | 0.036 |
| | lin. Reg. | D2 | -0.005 | 0.029 | -0.011 | 0.055 | -0.015 | 0.079 |
| | lin. Reg. | D3 | -0.018 | 0.113 | -0.052 | 0.210 | -0.130 | 0.287 |
| | MCMC | D1 | -0.016 | 0.023 | -0.036 | 0.032 | -0.051 | 0.035 |
| | MCMC | D2 | -0.004 | 0.029 | -0.010 | 0.055 | -0.015 | 0.079 |
| | MCMC | D3 | -0.018 | 0.113 | -0.053 | 0.213 | -0.128 | 0.289 |

Tab. 20: Mittlere Abweichung der Regressionskoeffizienten von den wahren Koeffizienten der drei Dummyvariablen und die mittleren Standardfehler (SE) des ordinalen Merkmalstyps. Der wahre Regressionskoeffizient für D1 beträgt 0.170, für D2 0.311 und für D3 0.396. (**Teil II**)

| MDM | Ersetzungs- | Sh_ | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|---|
| | methode | kat | 10% | | 30% | | 50% | |
| | | | Reg.koeff. | SE | Reg.koeff. | SE | Reg.koeff. | SE |
| lin. MNAR | | | | | | | | |
| | log. Reg. | D1 | 0.010 | 0.022 | 0.029 | 0.037 | 0.048 | 0.046 |
| | log. Reg. | D2 | 0.004 | 0.026 | 0.013 | 0.041 | 0.023 | 0.053 |
| | log. Reg. | D3 | 0.005 | 0.061 | 0.014 | 0.098 | 0.029 | 0.121 |
| | lin. Reg. | D1 | -0.007 | 0.024 | -0.013 | 0.032 | -0.034 | 0.035 |
| | lin. Reg. | D2 | -0.001 | 0.026 | $< 0.001$ | 0.040 | -0.028 | 0.056 |
| | lin. Reg. | D3 | -0.002 | 0.062 | 0.003 | 0.108 | -0.007 | 0.144 |
| | MCMC | D1 | -0.007 | 0.024 | -0.014 | 0.032 | -0.012 | 0.034 |
| | MCMC | D2 | $< -0.001$ | 0.026 | 0.002 | 0.042 | 0.006 | 0.051 |
| | MCMC | D3 | -0.001 | 0.062 | 0.003 | 0.108 | 0.021 | 0.138 |
| konv. MNAR | | | | | | | | |
| | log. Reg. | D1 | 0.003 | 0.020 | 0.010 | 0.035 | 0.009 | 0.044 |
| | log. Reg. | D2 | 0.005 | 0.029 | 0.013 | 0.050 | 0.014 | 0.071 |
| | log. Reg. | D3 | -0.014 | 0.109 | -0.040 | 0.177 | -0.095 | 0.220 |
| | lin. Reg. | D1 | -0.010 | 0.022 | -0.20 | 0.030 | -0.028 | 0.035 |
| | lin. Reg. | D2 | $< 0.001$ | 0.030 | 0.007 | 0.055 | 0.016 | 0.077 |
| | lin. Reg. | D3 | -0.018 | 0.115 | -0.056 | 0.219 | -0.138 | 0.310 |
| | MCMC | D1 | -0.009 | 0.022 | -0.020 | 0.030 | -0.028 | 0.036 |
| | MCMC | D2 | $< 0.001$ | 0.030 | 0.006 | 0.056 | 0.016 | 0.079 |
| | MCMC | D3 | -0.019 | 0.115 | -0.056 | 0.220 | -0.140 | 0.310 |

Den größten Einfluss auf die mittleren Abweichungen hat der Anteil an fW. Mit zunehmendem Anteil vergrößert sich die mittlere Abweichung der geschätzten Regressionskoeffizienten von den tatsächlichen Regressionskoeffizienten und die Standardfehler der Schätzer. Zusätzlich ist die Größe der Abweichungen deutlich abhängig von dem jeweiligen Regressionskoeffizienten der 3 Dummyvariablen. Tendenziell ist die Schätzung des Regressionskoeffizienten von der Variablen D2 valider und genauer als von den beiden anderen Dummyvariablen, da die Abweichungen geringer und die SE kleiner sind.

Die CCA führt unter MCAR bei jeder Dummyvariablen und jedem Anteil an fW zu nur marginalen Abweichungen. Etwas größer sind die Abweichungen bei Vorliegen von linearem MAR und noch größer unter konvexem MAR. Die logistische Regression weist ebenfalls bei Vorliegen von MCAR die geringsten Abweichungen für die jeweilige Dummyvariable auf. Etwas größer sind die Abweichungen unter den anderen MDM. Die Ersetzungsmethoden speziell für stetige Variablen führen in den meisten Situationen zu einer (geringfügigen) Unterschätzung der Regressionskoeffizienten. Am kleinsten sind die Abweichungen unter MNAR, etwas größer unter MCAR und den beiden Arten von MAR. Tendenziell sind die Abweichungen für die Dummyvariable D3 am größten, welche die äußerst selten auftretende Ausprägung Fettleber Grad 3 repräsentiert. Aufgrund der realtiv großen SE der Schätzer überlappen sich die 95% KI um die Schätzer und es lassen sich keine deutlichen (signifikanten) Unterschiede zwischen den Ersetzungsmethoden erkennen.

### 5.2.2 F-Statistik

Die zum Regressionskoeffizienten zugehörige F-Statistik prüft den Einfluss der Variablen auf die Zielgröße. Bei dem ordinalen Merkmalstyp von Sh testet die F-Statistik den Einfluss der 3 Dummyvariablen gleichzeitig mit Hilfe des Hypothesenpaares $H_0: \beta_{D1} = \beta_{D2} = \beta_{D3} = 0$ vs. $H_1:$ mindestens ein $\beta_{Di} \neq 0$, $i = 1, 2, 3$. In diesem Fall wird die F-Statistik mit dem kritischen Wert $F_{3,n-p,1-\alpha}$ verglichen. Die mittleren Abweichungen der 1000 F-Statistiken von der wahren F-Statistik mit dem Standardfehler der F-Statistiken werden in Tab. 21 für den binären Merkmalstyp und in Tab. 22 für den ordinalen Merkmalstyp von Steatosis hepatis gezeigt.

Die F-Statistik für den **binären Merkmalstyp** von Steatosis hepatis basierend auf dem Originaldatensatz beträgt 16.574. Der zugehörige kritische Wert beträgt $F_{1,n-p,1-\alpha} = F_{1,464-5,1-0.05} = 5.057$ und folglich gilt $F > F_{1,n-p,1-\alpha}$, sodass der binäre Merkmalstyp von Sh einen deutlichen Einfluss auf die Zielgröße

hat, vergl. Abschnitt 4.3.3. Bei einer Unterschätzung von bis zu 11.517 der F-Statistik würde das Ergebnis basierend auf dem vervollständigten Datensatz noch zu der gleichen Interpretation bzgl. des Einflusses von Sh führen.

Da die CCA auf einem verkleinerten Datensatz basiert, sind die ensprechenden kritischen Werte abhängig vom Anteil an fW. Bei einem Anteil von 10% fW beträgt der kritische Wert $F_{1,418-5,1-0.05} = 5.060$, bei 30% fW 5.071 und bei 50% fW 5.090. Diese minimalen Abweichungen lassen vermuten, dass auch die F-Statistik des Originaldatensatzes, wenn nur ein Teilkollektiv betrachtet worden wäre, sich nicht stark von der F-Statistik basierend auf dem kompletten Originaldatensatz unterscheidet. Daher werden in den Ergebnistabellen auch bei der CCA die Abweichungen zur F-Statistik des gesamten (N=464) Originaldatensatzes berechnet.

Der Anteil an fehlenden Werten im Datensatz hat wiederum den größten Einfluss auf die Güte (hier die Validität) der Ersetzungsmethoden. Die F-Statistik wird fast ausschließlich unterschätzt und mit zunehmendem Anteil an fW nimmt diese Unterschätzung von durchschnittlich -2 auf -7 zu, während sich der zugehörige Standardfehler der F-Statistik von durchschnittlich 2.7 auf 4.5 vergrößert. Bei einem Anteil von 50% fW im Datensatz könnte sogar (unter Berücksichtigung der Standardfehler) die F-Statistik kleiner als der kritische Wert sein, sodass geschlussfolgert werden würde, dass Sh keinen Einfluss auf die Zielgröße hat!

Für jede Ersetzungsmethode zeigt sich die kleinste mittlere Abweichung von der wahren F-Statistik bei Vorliegen eines MDM gemäß linearen MNAR, gefolgt von konvexen MNAR, linearen MAR und MCAR. Am stärksten wurde die F-Statistik bei konvexem MAR unterschätzt. Die Standardfehler hingegen waren bei MNAR um ca. 1 Einheit größer als bei MAR oder MCAR.

Bei jeder Kombination der Rahmenbedingungen wird die F-Statistik basierend

Tab. 21: Mittlere Abweichung der F-Statistiken von der wahren F-Statistik (F-Stat.) und der mittlere Standardfehler (SE) beim binären Merkmalstyp. Die wahre F-Statistik beträgt 16.574 und der kritische Wert $F_{1,464-5,0.95} = 5.057$.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | F-Stat. | SE | F-Stat. | SE | F-Stat. | SE |
| MCAR | | | | | | | |
| | CCA | -1.598 | 2.376 | -4.682 | 3.763 | -7.872 | 4.182 |
| | log. Reg. | -1.756 | 2.083 | -4.741 | 4.667 | -7.255 | 5.665 |
| | DFM | -1.160 | 2.928 | -3.034 | 5.101 | -5.316 | 6.685 |
| | lin. Reg. | -2.636 | 2.786 | -6.126 | 4.053 | -8.541 | 4.875 |
| | MCMC | -2.558 | 2.758 | -6.078 | 4.054 | -8.776 | 4.541 |
| lin. MAR | | | | | | | |
| | CCA | -1.239 | 2.381 | -3.831 | 3.719 | -6.474 | 3.990 |
| | log. Reg. | -1.475 | 2.678 | -4.084 | 4.372 | -6.298 | 5.293 |
| | DFM | -0.958 | 2.781 | -2.687 | 4.645 | -4.955 | 5.604 |
| | lin. Reg. | -2.302 | 2.728 | -5.685 | 4.059 | -7.887 | 4.450 |
| | MCMC | -2.371 | 2.653 | -5.380 | 3.992 | -7.890 | 4.280 |
| konv. MAR | | | | | | | |
| | CCA | -1.979 | 2.175 | -5.669 | 3.341 | -9.484 | 3.367 |
| | log. Reg. | -2.203 | 2.648 | -5.653 | 4.297 | -8.183 | 4.775 |
| | DFM | -1.451 | 2.753 | -4.020 | 4.792 | -6.379 | 5.820 |
| | lin. Reg. | -2.945 | 2.624 | -6.714 | 3.584 | -9.374 | 4.018 |
| | MCMC | -2.871 | 2.655 | -6.589 | 3.883 | -9.400 | 3.955 |
| lin. MNAR | | | | | | | |
| | log. Reg. | -0.502 | 3.062 | -1.275 | 5.306 | -2.147 | 6.892 |
| | DFM | 0.221 | 3.262 | 0.514 | 5.830 | -1.287 | 7.184 |
| | lin. Reg. | -1.438 | 3.016 | -3.686 | 4.485 | -5.078 | 5.384 |
| | MCMC | -1.572 | 3.023 | -3.657 | 4.442 | -5.278 | 5.269 |
| konv. MNAR | | | | | | | |
| | log. Reg. | -0.963 | 3.002 | -2.936 | 4.876 | -5.558 | 5.865 |
| | DFM | -0.243 | 3.134 | -0.876 | 5.595 | -3.258 | 6.971 |
| | lin. Reg. | -2.103 | 2.851 | -5.075 | 4.436 | -7.520 | 4.826 |
| | MCMC | -2.049 | 2.887 | -4.837 | 4.340 | -7.465 | 4.836 |

auf den mittels der DFM ersetzten Werte am wenigsten unterschätzt, aber sie führt zu größeren SE als die anderen betrachteten Ersetzungsmethoden und der Complete Case Analyse. Die logistische Regression und die CCA führen zu den nächst größeren Abweichungen mit den kleinsten SE. Die größten Abweichungen basieren auf den mittels der linearen Regression bzw. der MCMC

Methode vervollständigten Datensätzen. Auch bei diesem Vergleichskriterium sind die Unterschiede in den Abweichungen aufgrund der Größe der SE nicht signifikant, da sich die 95% KI überlappen.

Die Abweichungen der F-Statistiken von der jeweiligen F-Statistik des Originaldatensatzes für den **ordinalen Merkmalstyp** von Sh sind in der Tab. 22 enthalten. Die F-Statistik des Originaldatensatzes beträgt 6.55 und der kritische Wert $F_{3,n-p,1-\alpha} = F_{3,464-7,1-0.05} = 3.145$. Im Originaldatensatz hat Sh folglich auch als ordinaler Merkmalstyp im Regressionsmodell einen deutlichen Einfluss auf die Zielgröße und bei einer Unterschätzung um bis zu 3.045 Einheiten, ist die Interpretation basierend auf dem vervollständigten Datensatz die gleiche. Wenn eine CCA durchgeführt wird, ist der kritische Wert bei einem Anteil von 10% fW 3.147, von 30% fW 3.157 und von 50% fW 3.174. Die F-Statistik des verkleinerten Originaldatensatz wird sich vermutlich auch nur geringfügig von der F-Statistik des kompletten Originaldatensatzes unterscheiden. Daher werden in der Ergebnistabelle für alle Methoden die Abweichungen zur original F-Statistik angegeben, ohne Korrektur für die kleinere Fallzahl bei der CCA. Bei einem MDM gemäß MNAR wird die F-Statistik überschätzt, während unter den anderen MDM die F-Statistik eher unterschätzt wird.

Der Anteil an fW hat auch hier einen deutlichen Einfluss auf die Güte der Ersetzungsmethoden. Betragsmäßig tritt bei einem Anteil von 10% fW durchschnittlich eine Abweichung von 0.4 mit durchschnittlichem SE von 1.1, bei 30% fW eine durchschnittliche Abweichung von 1 mit SE von 1.8 und bei 50% fW eine durchschnittliche Abweichung von 1.5 mit mittlerem SE von 2.5 auf. Eine Abweichung deutlich über diesem Durchschnitt tritt bei der CCA auf. Zu einer deutlich geringeren Abweichung und somit valideren Schätzung führt die Ersetzung gemäß der logistischen Regression.

Abhängig von der jeweiligen Ersetzungsmethode führt mal der eine und mal der andere MDM zu der geringsten Abweichung von der wahren F-Statistik.

Tab. 22: Mittlere Abweichung der F-Statistiken von der wahren F-Statistik des gemeinsamen Einflusses der 3 Dummyvariablen und der mittlerer Standardfehler (SE) beim ordinalen Merkmalstyp. Die wahre F-Statistik beträgt 6.55 und der kritische Wert $F_{3,464-7,0.95} = 3.145$.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | F-Stat. | SE | F-Stat. | SE | F-Stat. | SE |
| MCAR | | | | | | | |
| | CCA | -0.538 | 0.915 | -1.567 | 1.451 | -2.665 | 1.638 |
| | log. Reg. | 0.108 | 1.041 | 0.371 | 2.010 | 0.877 | 3.155 |
| | lin. Reg. | -0.199 | 1.064 | -0.503 | 1.841 | -0.691 | 2.572 |
| | MCMC | -0.192 | 1.062 | -0.515 | 1.836 | -0.721 | 2.574 |
| lin. MAR | | | | | | | |
| | CCA | -0.428 | 0.852 | -1.334 | 1.286 | -2.251 | 1.430 |
| | log. Reg. | 0.143 | 0.993 | 0.527 | 1.835 | 1.119 | 2.672 |
| | lin. Reg. | -0.100 | 0.992 | -0.171 | 1.653 | 0.162 | 2.298 |
| | MCMC | -0.076 | 0.975 | -0.118 | 1.656 | 0.132 | 2.286 |
| konv. MAR | | | | | | | |
| | CCA | -0.624 | 0.902 | -1.789 | 1.373 | -3.020 | 1.514 |
| | log. Reg. | 0.007 | 1.054 | -0.010 | 1.899 | -0.208 | 2.643 |
| | lin. Reg. | -0.277 | 1.059 | -0.672 | 1.785 | -1.156 | 2.259 |
| | MCMC | -0.268 | 1.063 | -0.653 | 1.801 | -1.106 | 2.254 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.872 | 1.158 | 2.837 | 2.280 | 5.395 | 3.546 |
| | lin. Reg. | 0.367 | 1.133 | 1.307 | 1.919 | 0.099 | 2.301 |
| | MCMC | 0.378 | 1.115 | 1.325 | 1.999 | 2.749 | 2.794 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.616 | 1.235 | 1.650 | 2.239 | 2.015 | 3.198 |
| | lin. Reg. | 0.134 | 1.189 | 0.434 | 2.015 | 0.413 | 2.645 |
| | MCMC | 0.158 | 1.176 | 0.417 | 2.010 | 0.418 | 2.662 |

Alle Ersetzungsmethoden haben gemein, dass sie bei MNAR die F-Statistik überschätzen, während sie bei MAR oder MCAR die F-Statistik meistens unterschätzen.

Bei der CCA treten unter MCAR und MAR ähnlich große Abweichungen und zugehörige Standardfehler auf. Dabei sind die Abweichungen deutlich größer als die von den Ersetzungsmethoden. Die logistische Regression führt bei Vor-

liegen von MCAR und MAR zu den kleinsten Abweichungen, auch im Vergleich zu den anderen Methoden. Zu deutlich größeren Abweichungen führt die logistische Regression, wenn ein MDM gemäß MNAR vorliegt. Die Ersetzungsmethoden speziell für stetige Variablen weisen die geringsten Abweichungen bei linearem MAR, gefolgt von konvexem MNAR, MCAR, konvexem MAR und linearem MNAR auf. Aufgrund der SE zeigen die Ergebnisse nur Tendenzen auf und es gibt weder deutliche Unterschiede zwischen den Ersetzungsmethoden innerhalb einer Rahmenbedingung noch innerhalb einer Ersetzungsmethode zwischen den betrachteten Rahmenbedingungen.

### 5.2.3 $R^2$

Die Validität der Auswertungsergebnisse wurde ferner anhand des $R^2$ beurteilt, welches ein Maß für die Güte der Anpassung der anhand der Regressionsgleichung des Auswertungsmodells geschätzten mit den beobachteten Werten der Zielgröße darstellt. Das $R^2$ ermöglicht somit eine globale Beurteilung des Regressionsmodells und ist nicht nur auf eine Einflussgröße des Modells fixiert. In Tab. 23 und Tab. 24 werden die mittleren Abweichungen der 1000 $R^2$ von dem $R^2$ des Originaldatensatzes und die zugehörigen Standardfehler für die jeweilige Rahmenbedinung angegeben.

Das $R^2$ basierend auf dem Originaldatensatz und dem Regressionsmodell, in dem Sh als **binäres Merkmal** aufgefasst wird, beträgt 0.234, welches eine schlechte Anpassung ausdrückt. Die mittleren Abweichungen davon sind in Tab. 23 enthalten. Die Abweichungen vom wahren $R^2$ sind generell nur minimal, auch die SE sind sehr klein, selbst bei einem Anteil von 50% fW. Daher scheint auf das $R^2$ weder der Anteil an fW noch der MDM oder die gewählte Ersetzungsmethode einen (deutlichen) Einfluss zu haben.

Ein ähnliches Bild zeigen die Simulationsergebnisse für den **ordinalen Merkmalstyp** von Sh, siehe Tab. 24. Das $R^2$ basierend auf dem Originaldatensatz

beträgt 0.239 und die mittleren Abweichungen sind für alle Anteile an fW, MDM und Ersetzungsmethoden minimal. Einzig die CCA führt bei einem Anteil von 50% fW von Sh zu einer etwas größeren Abweichung und tendenziell größeren SE.

Tab. 23: Mittlere Abweichung der $R^2$ vom wahren $R^2$ des Regressionsmodells und die mittleren Standardfehler (SE) beim binären Merkmalstyp. Das wahre $R^2$ beträgt 0.234.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | $R^2$ | SE | $R^2$ | SE | $R^2$ | SE |
| MCAR | | | | | | | |
| | CCA | < 0.001 | 0.012 | 0.003 | 0.024 | 0.008 | 0.036 |
| | log. Reg. | < -0.001 | 0.004 | < 0.001 | 0.009 | < 0.001 | 0.013 |
| | DFM | < 0.001 | 0.005 | 0.003 | 0.009 | 0.006 | 0.014 |
| | lin. Reg. | -0.001 | 0.005 | -0.003 | 0.008 | -0.004 | 0.011 |
| | MCMC | -0.001 | 0.005 | -0.003 | 0.008 | -0.004 | 0.011 |
| lin. MAR | | | | | | | |
| | CCA | 0.004 | 0.012 | 0.013 | 0.023 | 0.022 | 0.035 |
| | log. Reg. | < 0.001 | 0.004 | < 0.001 | 0.008 | 0.002 | 0.011 |
| | DFM | 0.001 | 0.005 | 0.003 | 0.009 | 0.005 | 0.012 |
| | lin. Reg. | -0.007 | 0.018 | -0.020 | 0.031 | -0.032 | 0.041 |
| | MCMC | < -0.001 | 0.004 | -0.002 | 0.008 | -0.003 | 0.009 |
| konv. MAR | | | | | | | |
| | CCA | -0.004 | 0.013 | -0.017 | 0.025 | -0.038 | 0.036 |
| | log. Reg. | < -0.001 | 0.004 | -0.002 | 0.008 | -0.002 | 0.010 |
| | DFM | < 0.001 | 0.004 | 0.001 | 0.008 | 0.002 | 0.012 |
| | lin. Reg. | -0.002 | 0.004 | -0.004 | 0.007 | -0.006 | 0.009 |
| | MCMC | -0.002 | 0.004 | -0.004 | 0.007 | -0.006 | 0.009 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.002 | 0.005 | 0.008 | 0.009 | 0.014 | 0.014 |
| | DFM | 0.003 | 0.005 | 0.011 | 0.010 | 0.016 | 0.014 |
| | lin. Reg. | 0.001 | 0.005 | 0.003 | 0.008 | 0.004 | 0.011 |
| | MCMC | < 0.001 | 0.005 | 0.003 | 0.008 | 0.004 | 0.010 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.001 | 0.005 | 0.004 | 0.009 | 0.006 | 0.013 |
| | DFM | 0.003 | 0.005 | 0.008 | 0.010 | 0.011 | 0.014 |
| | lin. Reg. | < -0.001 | 0.005 | < -0.001 | 0.008 | -0.001 | 0.010 |
| | MCMC | < -0.001 | 0.005 | < 0.001 | 0.008 | < -0.001 | 0.011 |

Tab. 24: Mittlere Abweichung der $R^2$ von dem wahren $R^2$ des Regressionsmodells und die mittleren Standardfehler (SE) beim ordinalen Merkmalstyp. Das wahre $R^2$ beträgt 0.239.

| MDM | Ersetzungs- | Anteil an fehlenden Werten | | | | | |
|---|---|---|---|---|---|---|---|
| | methode | 10% | | 30% | | 50% | |
| | | $R^2$ | SE | $R^2$ | SE | $R^2$ | SE |
| MCAR | | | | | | | |
| | CCA | 0.001 | 0.012 | 0.005 | 0.025 | 0.013 | 0.037 |
| | log. Reg. | $<$ 0.001 | 0.005 | 0.002 | 0.010 | 0.004 | 0.015 |
| | lin. Reg. | $<$ -0.001 | 0.005 | -0.003 | 0.009 | -0.004 | 0.012 |
| | MCMC | $<$ -0.001 | 0.005 | -0.003 | 0.009 | -0.004 | 0.012 |
| lin. MAR | | | | | | | |
| | CCA | 0.004 | 0.012 | 0.014 | 0.023 | 0.026 | 0.035 |
| | log. Reg. | $<$ 0.001 | 0.005 | 0.002 | 0.009 | 0.005 | 0.012 |
| | lin. Reg. | $<$ -0.001 | 0.005 | $<$ -0.001 | 0.008 | $<$ 0.001 | 0.011 |
| | MCMC | $<$ -0.001 | 0.005 | $<$ -0.001 | 0.008 | $<$ 0.001 | 0.011 |
| konv. MAR | | | | | | | |
| | CCA | -0.003 | 0.013 | -0.014 | 0.025 | -0.031 | 0.037 |
| | log. Reg. | $<$ 0.001 | 0.005 | $<$ -0.001 | 0.009 | -0.001 | 0.013 |
| | lin. Reg. | -0.001 | 0.005 | -0.003 | 0.009 | -0.006 | 0.011 |
| | MCMC | -0.001 | 0.005 | -0.003 | 0.009 | -0.006 | 0.011 |
| lin. MNAR | | | | | | | |
| | log. Reg. | 0.004 | 0.005 | 0.013 | 0.010 | 0.024 | 0.016 |
| | lin. Reg. | 0.002 | 0.005 | 0.006 | 0.009 | $<$ 0.001 | 0.011 |
| | MCMC | 0.002 | 0.005 | 0.006 | 0.009 | 0.013 | 0.013 |
| konv. MNAR | | | | | | | |
| | log. Reg. | 0.003 | 0.006 | 0.008 | 0.010 | 0.009 | 0.015 |
| | lin. Reg. | $<$ 0.001 | 0.006 | 0.002 | 0.010 | 0.002 | 0.012 |
| | MCMC | 0.001 | 0.006 | 0.002 | 0.010 | 0.002 | 0.013 |

# Kapitel 6

# Diskussion

Dieses Kapitel ist in 3 Teile aufgeteilt. Im ersten Teil werden die wesentlichen Ergebnisse der Simulationsstudie vorgestellt und interpretiert. Anschließend wird die Verallgemeinerbarkeit der Ergebnisse erläutert. Im dritten Teil werden Empfehlungen für den Umgang mit fehlenden Werten, speziell auch bei kategorialen Variablen, in klinischen Studien gegeben.

## 6.1 Interpretation der Simulationsergebnisse

Die Simulationsergebnisse für den binären und ordinalen Merkmalstyp von Sh zur **Beurteilung der Übereinstimmung** der geschätzten mit den tatsächlichen Werten gemessen durch den Kappa-Koeffizienten ähneln sich stark. Betrachtet man die tatsächlich beobachteten und geschätzten Werte gemeinsam, so deutet die Größe des Kappa-Koeffizienten auf eine gute Übereinstimmung hin, wobei sich diese deutlich mit zunehmendem Anteil an fW verringert. Dies ist plausibel, da bei einem größeren Anteil an fW der Anteil an automatisch übereinstimmenden Beobachtungen kleiner ist und sich somit der Kappa-Koeffizient deutlicher von 1 unterscheiden kann. Aus dem gleichen Grund nimmt der Standardfehler (SE), d.h. die Unsicherheit der Schätzung des Kappa-Koeffizienten, mit dem Anteil an fW zu.

Berücksichtigt man hingegen nur die geschätzten Werte zur Berechnung des Kappa-Koeffizienten, so betragen die durchschnittlichen Kappa-Koeffizienten sowohl bei dem binären als auch bei dem ordinalen Merkmalstyp abhängig von der jeweiligen Rahmenbedingung nur noch zwischen 0.1 und 0.24. Da die 0 in jedem 95% KI um den Kappa-Koeffizienten enthalten ist, ist die Übereinstimmung nicht signifikant besser als rein zufällig! Bei der Auswertung eines klinischen Datensatzes muss immer das Ganze betrachtet werden: Wie groß ist der Einfluss der (in der Praxis nicht nachprüfbaren!) ungenauen Schätzung der fW auf die Kollektivbeschreibung und die teststatistischen Analyseergebnisse? Hierzu sei auf die weiteren Vergleichsparameter verwiesen.

Der Anteil an fW scheint in der Situation, in der nur die geschätzten Werte betrachtet werden, keinen Einfluss auf die Größe des Kappa-Koeffizienten zu haben. Folglich wird vermutlich die Wahrscheinlichkeit, verzerrte Parameterschätzer zu erhalten, mit zunehmendem Anteil an fW größer, da die (fehlerhaft) ersetzten Werte einen immer größeren Teil der berücksichtigten Werte in der Schätzung darstellen.

Zur Interpretation der unterschiedlich großen Kappa-Koeffizienten je nach vorliegendem Missing Data Mechanism (MDM), sei zunächst auf Folgendes hingewiesen: Unterschiedliche Randsummen in den Kontingenztafeln führen zu verschiedenen zufallsbedingten erwarteten Übereinstimmungen $P_e$, sodass bei gleicher beobachteter Übereinstimmung $P_b$ unterschiedliche Kappa-Koeffizienten berechnet werden können, vergl. Abschnitt 4.3.1. Wenn nur die geschätzten Werte berücksichtigt werden, können sich die Randsummen zwischen den Missing Data Mechanism stark unterscheiden.

Bei einem linearen MAR treten beispielsweise vermehrt für Beobachtungen mit Sh=0 fehlende Werte auf, sodass die Ausprägung Sh=1 seltener in der Kontingenztafel zur Berechnung des Kappa-Koeffizienten auftritt als bei den Kontingenztafeln basierend auf den Datensätzen, bei denen ein MDM gemäß

MCAR oder konvexem MAR erzeugt wurde. Der kleinere Kappa-Koeffizient bei Vorliegen eines linearen MAR bedeutet, dass der Anteil an Übereinstimmung, nachdem der Anteil an zufälliger Übereinstimmung aus der Betrachtung entfernt wurde, kleiner ist als bei einem MDM gemäß MCAR oder konvexem MAR. Es bedeutet aber nicht unbedingt, dass unter linearem MAR weniger Beobachtungen genau übereinstimmen (d.h. auf der Hauptdiagonalen der Kontingenztafel liegen). Ein Vergleich der Ersetzungsmethoden anhand der absoluten Größe des Kappa-Koeffizienten ist daher in dieser Situation nur innerhalb eines MDM und eines Anteils an fW möglich.

Die Abhängigkeit des Kappa-Koeffizienten von den Randsummen kann der Grund sein, dass unter Berücksichtigung aller Werte der eine Missing Data Mechanism zu dem größten Kappa-Koeffizienten führt, während bei der Berechnung basierend nur auf den geschätzten Werten ein anderer MDM zu dem größten Kappa-Koeffizienten führt.

Ein wesentliches Ergebnis der Simulationen ist, dass die mittleren Kappa-Koeffizienten unter Verwendung von Ersetzungsmethoden speziell für kategoriale Variablen (logistische Regression und Discriminant Function Method (DFM)) größer sind, als wenn die fehlenden Werte mittels Ersetzungsmethoden für stetige Variablen (lineare Regression und MCMC Methode) geschätzt worden wären. Aufgrund der beobachteten Streuung der Kappa-Koeffizientenschätzer ist allerdings kein statistisch signifikanter Unterschied erkennbar, da sich die 95% Konfidenzintervalle um die Kappa-Koeffizienten der einzelnen Ersetzungsmethoden überlappen. Ob der beobachtete Unterschied im Einzelfall relevant ist, sollte sich durch die Simulationsergebnisse bezüglich der weiteren Vergleichsparameter herausstellen.

Newgard [73] hat ebenfalls den Kappa-Koeffizienten zur Beurteilung der Übereinstimmung ausschließlich der geschätzten mit den tatsächlichen Werten kategorialer Variablen für den kleinen Anteil von 10% = 191 Beobachtungen mit

fW berechnet. Der Kappa-Koeffizient basierend auf einem mittels der linearen Regression vervollständigten Datensatz ist mit 0.66 (95% KI: [0.40; 0.92]) erstaunlich hoch im Vergleich zu den Kappa-Koeffizienten aus der vorliegenden Arbeit.

Mit Hilfe der Abweichung des Mittelwertes basierend auf den vervollständigten Datensätzen im Vergleich zu dem Originalmittelwert wird in dieser Arbeit untersucht, ob die ersetzten Werte die tatsächlichen Werte eher über- oder unterschätzten. Bei beiden Merkmalstypen von Sh führen die Ersetzungsmethoden und die CCA im vorliegenden Datensatz meistens zu einer Überschätzung des Mittelwertes. Nur die DFM (im binären Fall) führt überwiegend zu einer Unterschätzung. Inhaltlich bedeutet die Überschätzung, dass fälschlicherweise mehr Personen mit Steatosis hepatis (Sh) bzw. einem größeren Grad an Sh im Kollektiv des vervollständigten Datensatzes enthalten sind.

Die Abweichung vom Originalmittelwert hängt primär von dem Anteil an fehlenden Werten ab. Auch die Streuung der Schätzer nimmt mit dem Anteil an fW zu. Dies ist plausibel, denn je mehr fW ersetzt werden müssen, desto größer ist der Einfluss der (fehlerhaft) ersetzten Werte auf den Mittelwertschätzer und der Genauigkeit der Schätzung. Außer bei einem Anteil von 50% fW scheinen in dem vorliegenden Datensatz die Abweichungen nicht relevant zu sein.

Im Unterschied zum Kappa-Koeffizienten zeigen die Abweichungen vom Originalmittelwert der einzelnen Ersetzungsmethoden keine einheitliche Tendenz in Bezug auf die MDM. Zumindest bei der logistischen Regression ist die Rangfolge der Abweichungen unter dem jeweiligen MDM abhängig von dem Merkmalstyp der Variablen Sh. Dies zeigt, dass die Abweichungen vom Originalmittelwert nicht nur vom MDM und dem Anteil an fehlenden Werten abhängig sind, sondern vermutlich auch vom Merkmalstyp der Variablen und möglicherweise deren Häufigkeitsverteilung.

Generell führen die Ersetzungsmethoden für kategoriale Variablen im vorliegenden Datensatz bei dem binären und kategorialen Merkmalstyp von Steatosis hepatis zu den betragsmäßig geringsten Abweichungen von dem Originalmittelwert gefolgt von der CCA. Deutlich größere Abweichungen und Streuungen in der Schätzung (RMSE) weisen die Ersetzungsmethoden für stetige Variablen auf. Diese Unterschiede sind allerdings unter Berücksichtigung der Streuung nicht statistisch signifikant.

In den Literaturstellen [1, 5, 19], in denen ebenfalls die Abweichung statistischer Kenngrößen abhängig von der gewählten Ersetzungsmethode verglichen wurden, zeigte sich analog zu den vorliegenden Ergebnissen, dass der Anteil an fW einen deutlichen Einfluss auf die Abweichung hat. Gemäß Rubin [82] führen die Ersetzungsmethoden zumindest bei Vorliegen ihrer jeweiligen Voraussetzungen zu keinen verzerrten Parameterschätzern. Die vorliegenden Simulationsergebnisse zeigen, dass selbst bei Nichterfüllung aller Voraussetzungen die Abweichungen vom Originalparameter (hier Mittelwert) einer Variablen irrelevant sein können.

**Zusammengefasst** zeigen die **Simulationsergebnisse** bezüglich der **Übereinstimmung** der geschätzten mit den fehlenden Werten, dass die Übereinstimmung für alle Ersetzungsmethoden nur schwach zu sein scheint, wobei sie unter Verwendung der logistischen Regression und DFM etwas besser ist als bei der linearen Regression oder der MCMC Methode. Die beobachteten geringen aber systematischen Abweichungen vom Originalmittelwert zeigen, dass alle hier betrachteten Ersetzungsmethoden zu verzerrten Parameterschätzern führen können. Anhand der Simulationsergebnisse kann allerdings nicht beurteilt werden, ob die Ersetzungsmethoden auch bei anderen Datensätzen die tatsächlichen Werte überwiegend überschätzen würden. Es ist denkbar, dass die Häufigkeitsverteilung des Merkmals und die Korrelationen zwischen den anderen Variablen im Ersetzungsmodell mit der Variablen mit fW die Verzerrung nach oben bzw. unten beeinflussen.

Im Folgenden werden die wesentlichen Ergebnisse der Simulationen in Bezug auf die **Validität der Auswertungsergebnisse** basierend auf den vervollständigten Datensätzen für den binären und ordinalen Merkmalstyp von Steatosis hepatis erläutert.

Die mittleren Abweichungen der zu Sh gehörenden Regressionskoeffizienten von den Originalwerten unterscheiden sich nur geringfügig von Null (0). Mit zunehmendem Anteil an fW werden die Abweichungen und Standardfehler (SE) zwar betragsmäßig größer und je nach Ersetzungsmethode führen unterschiedliche MDM zu den betragsmäßig größten Abweichungen, doch für den vorliegenden Datensatz sind die beobachteten Unterschiede so minimal, dass sie nicht relevant sind. Eine Systematik in der Über- bzw. Unterschätzung der Regressionskoeffizienten ist unter Berücksichtigung der relativ großen SE im Vergleich zu den Abweichungen nicht erkennbar.

Die Validität, d.h. mögliche systematische Fehler in der Schätzung, des Regressionskoeffizienten der ordinalen Variablen wird anhand dreier Dummyvariablen beurteilt. Da sich innerhalb einer Rahmenbedingung die Abweichungen zum Teil stärker innerhalb einer Ersetzungsmethode zwischen den 3 Dummyvariabeln unterscheiden als zwischen den Ersetzungsmethoden für die jeweilige Dummyvariable, untermauert dies die Vermutung, dass die Häufigkeitsverteilung der Variablen eine wesentliche Rolle bei der Schätzung der fW spielt. Die Simulationsergebnisse lassen keine Schlussfolgerung zu, wie die Häufigkeitsverteilung die Schätzung der fW beeinflusst, da sich die relativen Häufigkeiten für das Auftreten der 3 Dummyvariablen nur geringfügig unterscheiden. Die im Mittel geringsten Abweichungen treten bei der Variablen D2 auf, welche ca. 10% der Beobachtungen im Originaldatensatz charakterisiert. Die Regressionskoeffizienten der Variablen D1 und D3 werden deutlich schlechter geschätzt. Diese Variablen treten im Originaldatensatz mit einer Häufigkeit von ca. 16% bzw. 2% auf.

Basierend auf diesen Simulationsergebnissen kann in Bezug auf die Validität des Regressionskoeffizienten keine Methode für den Umgang mit fW favorisiert bzw. als ungeeignet angesehen werden. Selbst die CCA führt bei diesem Vergleichsparameter nicht zu deutlich größeren Abweichungen vom Originalwert!

Während der Regressionskoeffizient die Art des Einflusses der Variablen Sh auf die Zielgröße beschreibt, drückt die zugehörige F-Statistik aus, ob die Variable Sh unter den betrachteten Einflussgrößen im linearen Regressionsmodell einen signifikanten (wesentlichen) Einfluss auf die Variabilität der Zielgröße hat. Je größer die F-Statistik im Vergleich zum kritischen Wert ist, desto wichtiger ist die Variable aus statistischer Sicht in diesem Modell. Wird die F-Statistik unterschätzt, so spricht man von konservativem Testen, da ein tatsächlich existierender Einfluss schwieriger zu erkennen ist. Bei einer Überschätzung hingegen wird möglicherweise ein Einfluss suggeriert, der basierend auf den Originalwerten nicht vorliegt.

Die Simulationsergebnisse bezüglich der Validität der F-Statistik stimmen nur teilweise für die zwei betrachteten Merkmalstypen der Variablen Sh überein. Bei beiden treten betragsmäßig relativ große Abweichungen mit großen SE auf. Auch der Einfluss des Anteils an fW auf die Validität ist bei beiden Merkmalstypen deutlich. Doch das Auftreten von Über- bzw. Unterschätzungen der F-Statistik ist abhängig von dem Merkmalstyp und dem MDM. Bei dem binären Merkmalstyp wird die F-Statistik fast ausschließlich unterschätzt und bei einem Anteil von 50% fW kann sich die Interpretation von einem signifikanten Einfluss zu einem nicht signifikanten Einfluss verändern. Bei dem ordinalen Merkmalstyp von Sh hingegen treten bei einem MDM gemäß MNAR auch deutliche Überschätzungen der F-Statistik auf. Die durchgeführten Simulationen ermöglichen allerdings keine allgemeingültige Vorhersage unter welchem Missing Data Mechanism bzw. Merkmalstyp der Variablen mit fW die F-Statistik über- oder unterschätzt wird, vergl. Abschnitt 6.2.

Bei dem binären Merkmalstyp treten für alle Ersetzungsmethoden die absolut betrachtet kleinsten Abweichungen bei Vorliegen eines MDM gemäß MNAR auf, wobei die SE deutlich größer sind als bei den anderen MDM. Doch bei dem ordinalen Merkmalstyp führt ein MDM gemäß MAR gefolgt von MCAR und MNAR zu den betragsmäßg kleinsten Abweichungen. Auch bei diesem Vergleichsparameter scheint daher ein MDM gemäß MNAR nicht zwangsläufig zu verzerrteren Auswertungsergebnissen zu führen als bei MCAR oder MAR.

Die Ersetzungsmethoden speziell für kategoriale Variablen führen bei beiden Merkmalstypen zu betragsmäßig deutlich kleineren mittleren Abweichungen als die stetigen Ersetzungsmethoden. Bei dem binären Merkmalstyp sind die Abweichungen basierend auf den mittels der DFM vervollständigten Datensätzen wesentlich kleiner als von den anderen Methoden und beim ordinalen Merkmalstyp weist die CCA erheblich größere Abweichungen als die Ersetzungsmethoden auf. Unter Berücksichtigung der Standardfehler sind diese Unterschiede aber nicht statistisch signifikant.

Auf die Schätzung des Vergleichsparameters $R^2$ als Maß für die Güte der Anpassung der vorhergesagten mit den beobachteten Werten der Zielgröße des Auswertungsmodells scheinen die gewählten Rahmenbedingungen keinen Einfluss zu nehmen. Bei beiden Merkmalstypen von Steatosis hepatis sind die mittleren Abweichungen vom Originalwert und die zugehörigen Standardfehler nur geringfügig unterschiedlich von Null, selbst bei einem großen Anteil an fW. Es lassen sich auch keine Tendenzen in der Größe der Abweichungen abhängig vom MDM oder der verwendeten Methode für den Umgang mit fW erkennen. Möglicherweise würden sich differenziertere Ergebnisse zeigen, wenn das Modell für den Originaldatensatz eine gute Anpassung hätte, d.h. ein $R^2$ von mindestens 0.6 und nicht wie im vorliegenden Fall von 0.23.

**Zusammengefasst** zeigen die **Simulationsergebnisse** bezüglich der **Validität** der Auswertungsergebnisse, dass zwar deutliche Tendenzen in den mittle-

ren Abweichungen zwischen den Ersetzungsmethoden und der CCA unter den einzelnen Rahmenbedingungen erkennbar sind, diese aber aufgrund der Streuung der Vergleichsparameterschätzer (Regressionskoeffizient(en), F-Statistik und $R^2$) nicht statistisch signifikant sind. Ähnliche Ergebnisse werden auch in der Literatur [3, 20, 40, 75] beschrieben. Tendenziell führen die Ersetzungsmethoden für kategoriale Variablen zu betragsmäßig kleineren Abweichungen als die Ersetzungsmethoden primär für stetige Variablen und die CCA. Bei dem vorliegenden Datensatz beeinflussen die Methoden für den Umgang mit fW und die betrachteten Rahmenbedingungen die Schätzer für die Regressionskoeffizienten und das $R^2$ kaum. Doch die F-Statistik wurde zum Teil deutlich über- bzw. unterschätzt, sodass möglicherweise eine Fehlinterpretation des Einflusses von Sh auf die Zielgröße aus dem einzelnen vervollständigten Datensatz erstellt worden wäre. Ferner zeigen die Simulationen, dass auch MNAR unter gewissen Umständen zu den betragsmäßig kleinsten Abweichungen vom Vergleichsparameter (hier der F-Statistik) führen kann.

Wie bereits erwähnt traten in den Simulationsergebnissen keine statistisch signifikanten Unterschiede in der Validität der betrachteten Parameter zwischen den Methoden für den Umgang mit fW auf. Die fehlende statistische Signifikanz unter Betrachtung der 1000 Datensätze je Rahmenbedingung muss allerdings nicht zwangsläufig als fehlende Relevanz der beobachteten Unterschiede aufgefasst werden. Die Relevanz ist abhängig von der zu untersuchenden Fragestellung und der angestrebten Auswertung zu beurteilen. Die Simulationen zeigen, dass die etwas größeren Kappa-Koeffizienten bei der Verwendung von Ersetzungsmethoden für kategoriale Variablen dazu führen, dass diese Methoden auch bei den anderen Vergleichsparametern zu tendenziell valideren Schätzungen führen.

Auch wenn keine statistisch signifikanten Unterschiede zwischen den Ergebnissen der CCA und denen der Ersetzungsmethoden festgestellt werden konnten, empfehle ich vor einer teststatistischen Auswertung von klinischen Da-

tensätzen eine geeignete Ersetzungsmethode gemäß einer Multiple Imputation durchzuführen. So kann sichergestellt werden, dass alle tatsächlich erhobenen Informationen (Werte) berücksichtigt werden und kein Selektionsbias auftritt. Wie die Simulationen zeigen sind bei einem derartigen Vorgehen in den meisten Fällen keine systematischen Fehler in der Parameterschätzung zu befürchten.

Aus medizinischer Sicht können die Ergebnisse wie folgt interpretiert werden. In dem Regressionsmodell basierend auf den N=464 vollständig erhobenen Beobachtungen haben das Geschlecht, der Taillenumfang, Steatosis hepatis und das Adiponektin einen deutlichen (signifikanten) Einfluss auf die Zielgröße. Um homogen streuende Residuen des Regressionsmodells zu erhalten, ist die Zielgröße der Logarithmus vom Insulin (loginsulin). Das Alter scheint keinen signifikanten Einfluss auf loginsulin zu haben, da der p-Wert des Regressionskoeffizienten mit 0.064 größer als das Signifikanzniveau von 0.05 ist. Aus inhaltlichen Gründen sollte das Alter dennoch im Modell enthalten sein.

Die Modellanpassung ist mit $R^2 = 0.234$ nicht gut, da die im Modell enthaltenen Variablen nur ca. 23% der Variabilität des loginsulins erklären. Mindestens eine wesentliche Einflussgröße fehlt demnach im Modell, sodass das in den Simulationen verwendete Modell nicht zur Vorhersage der Werte des (logarithmierten) Insulins geeignet wäre. Es ist fraglich, ob die momentan signifikanten Einflussgrößen den deutlichen Einfluss auf die Zielgröße behalten werden, wenn die anderen wesentlichen Einflussgrößen im Modell enthalten sind.

In dem ursprünglichen Datensatz von N=532 Beobachtungen treten fehlende Werte beim Insulin (N=2), Taillenumfang (N=17) und Adiponektin (N=53) auf. Da diese Variablen alle stetig sind und die fehlenden Werte gemäß einem beliebigen Missing Data Pattern auftreten, empfehle ich zur Schätzung der fehlenden Werte die MCMC Methode anzuwenden und 5 vervollständigte Datensätze zu erzeugen. Diese müssen einzeln ausgewertet werden und anschließend zu einem Ergebnis zusammengefassten werden. Dieses Vorgehen habe

ich in SAS mittels den Prozeduren `PROC MI`, `PROC REG` und `PROC MIANALYZE` umgesetzt. Es zeigt sich, dass unter Berücksichtigung der N=532 Beobachtungen auch das Alter einen knapp signifikanten Einfluss auf die Zielgröße hat (p=0.048), während sich die p-Werte der anderen Einflussgrößen nur minimal verändern.

## 6.2 Limitierungen der Simulationsergebnisse

Die Simulationen basierten auf einem realen, vollständig erhobenen Datensatz mit einigen hundert Beobachtungen. Sie untersuchten den Einfluss von 5 verschiedenen Missing Data Mechanism und 3 verschiedenen Anteilen an fehlenden Werten (fW) auf die Übereinstimmung der geschätzten mit den tatsächlichen Werten und die Validität der Ergebnisse einer multiplen linearen Regression, wobei 5 verschiedene Methoden für den Umgang mit den fW verwendet wurden. Aufgrund des gewählten Simulationsaufbaus ist eine Übertragung der Simulationsergebnisse nur auf Datensätze möglich, die ähnlich groß sind und ebenfalls die häufig angewandte lineare Regression als statistisches Auswertungsverfahren durchführen.

Die Ergebnisse bzgl. der Regressionskoeffizienten der 3 Dummyvariablen deuten darauf hin, dass die Häufigkeitsverteilung der Variablen mit fehlenden Werten eine wesentliche Rolle bei der Validität der Parameterschätzer basierend auf vervollständigten Datensätzen spielt. Da in den Simulationen nur eine Häufigkeitsverteilung des binären bzw. ordinalen Merkmalstyps der Variablen mit fehlenden Werten betrachtet wurde, können Simulationen, in denen die Variable mit fW eine andere Häufigkeitsverteilung aufweist, zu anderen als den vorliegenden Ergebnissen führen.

Ferner wurde in den durchgeführten Simulationen nur eine Korrelationsmatrix der im Ersetzungsmodell enthaltenen Einflussgrößen betrachtet. Eine Variation der Korrelationen zwischen den Einflussgrößen, speziell der Korrelationen

mit der Variablen mit fW, war nicht möglich, da ein realer Datensatz verwendet wurde. Vermutlich würde eine Korrelationsmatrix, bei der die Variablen im Ersetzungsmodell einen stärkeren Zusammenhang mit der Variablen mit fW aufweisen, zu einer besseren Schätzung der fW und so eventuell zu kleineren Abweichungen in der F-Statistik führen.

Die fehlenden Werte von Steatosis hepatis (Sh) wurden bei dem MDM gemäß MAR in Abhängigkeit vom Taillenumfang der jeweiligen Beobachtung erzeugt. Die Korrelation zwischen den beiden Variablen war mit 0.4 nur mäßig, sodass der Taillenumfang nur wenig Information zur Schätzung der fW von Sh lieferte. Ein MDM gemäß MNAR lag dann vor, wenn die Variable Taillenumfang nicht im Ersetzungsmodell enthalten war. Der geringe Unterschied zwischen den Simulationsergebnissen basierend auf einem MDM gemäß MAR bzw. MNAR ist möglicherweise auf die geringe Korrelation zwischen dem Taillenumfang und der Variablen Sh zurückzuführen. Vermutlich wären deutlichere Unterschiede erkennbar, wenn der Taillenumfang mehr Informationen zur Schätzung der fW von Sh liefern könnte.

Folglich kann nicht ausgeschlossen werden, dass bei einer anderen Häufigkeitsverteilung der binären oder ordinalen Variablen mit fehlenden Werten, einer anderen Korrelationsmatrix der Einflussgrößen im Ersetzungsmodell und dem Nichterheben einer anderen wichtigen Variablen zur Schätzung der fW, andere Ergebnisse bzgl. der Übereinstimmung und der Validität der Auswertungsergebnisse beobachtet werden. Sowohl die Größe der Abweichungen von den Originalwerten könnte sich ändern als auch die Tendenz zur Über- bzw. Unterschätzung der Originalwerte. Beispielsweise könnte die F-Statistik dann primär über- und nicht unterschätzt werden. Um den Einfluss der hier aufgeführten konstant gehaltenen Störgrößen beurteilen zu können, wäre die Durchführung weiterer Simulationen, in denen diese Störgrößen geeignet variiert werden, sinnvoll.

## 6.3 Fazit für den Umgang mit fehlenden Werten in klinischen Datensätzen

Da die durchgeführten Simulationen „leider“ zeigen, dass unter den hier betrachteten Situationen keine der 5 Methoden (Complete Case Analyse, logistische Regression, Discriminant Function Method, lineare Regression und MCMC Methode) für den Umgang mit fehlenden Werten kategorialer Variablen als allgemein am besten geeignet zu sein scheint, wird auch in Zukunft eine hohe Datenqualität eine wesentliche Rolle für die Aussagekraft klinischer Datensätze spielen. Die sicherste Methode eine gute Aussagekraft zu erreichen ist die Durchführung einer gewissenhaften vollständigen Datenerhebung[1].

Nach wie vor muss daher das primäre Ziel des Datenmanagements eine korrekte und vollständige Erhebung aller Werte der betrachteten Variablen sein. Hierfür gilt es die folgenden Aspekte zu beachten: Bei Variablen mit medizinisch inhaltlich gleichwertigem/ähnlichem Informationsgehalt bezüglich der Fragestellung sollte diejenige Variable bevorzugt erhoben werden, von der bei der Erhebung ein kleinerer Anteil an fehlenden Werten erwartet wird.

In dem Abschnitt 3.2 über das Ersetzungs- vs. Analysemodell wurde empfohlen, möglichst viele Variablen ins Ersetzungsmodell mitaufzunehmen, um möglichst einen Missing Data Mechanism (MDM) gemäß Missing At Random (MAR) zu erhalten. Die im Abschnitt 3.4 vorgestellten Guidelines schlagen vor, weitere Variablen zu erheben, die primär zur Modellierung eines MAR geeignet wären, auch wenn diese nicht für die Auswertung der Fragestellung relevant sind. Die Simulationsergebnisse lassen den Vorteil von der Erhebung zusätzlicher Variablen aus diesem Grund nur bedingt erkennen, da die Ergebnisse bei Vorliegen eines MNAR nur teilweise etwas schlechter als bei Vorliegen

[1] Natürlich wird die Aussagekraft klinischer Studien auch durch eine geeignete Wahl des Studiendesigns, geeignet gewählten Ein- und Ausschlusskriterien und der Wahl der geeigneten Variablen zur Beantwortung der Fragestellung beeinflusst.

eines MAR waren. Daher empfehle ich, nur bei einem relativ kleinen angestrebten Datensatz zusätzliche Variablen zur möglichen Modellierung eines MDM gemäß MAR zu erheben. Bei einem größeren Datensatz muss dieser Nutzen dem Risiko einer abgeschwächten Motivation zur gewissenhaften, lückenlosen Erhebung aller Werte gegenübergestellt werden.

Wird eine Studie mit einer konfirmatorischen, d.h. Hypothesen beweisenden, statistischen Auswertung geplant, so ist es wichtig, vorab im Studienplan den Umgang mit möglichen fW festzulegen. Eine zusätzliche Sensitivitätsanalyse, in welcher mindestens zwei weitere Methoden für den Umgang mit fW durchgeführt werden, ist sinnvoll, um den Einfluss der gewählten Methode auf die Auswertungsergebnisse zu beurteilen. Doch kann mit einer Sensitivitätsanalyse keine konfirmatorische Auswertung durchgeführt werden. Von einer Sensitivitätsanalyse mit Worst- bzw. Best Case Ersetzung ist abzuraten, siehe Abschnitt 3.5. Stattdessen sollten verschiedene Ersetzungsmethoden angewendet werden, die eine Multiple Imputation durchführen können.

Bei der Datenerfassung ist es wichtig, dass die fehlenden Werte als solche im Datensatz zu erkennen sind. Der Erhebungsbogen muss daher so gestaltet werden, dass zwischen einer unbeantworteten Frage und einer anderen Ausprägung der zu erfassenden Variablen differenziert werden kann. Beispielsweise sollten zur Beantwortung einer ja/nein Frage beide Antwortmöglichkeiten gegeben sein. Wenn beide nicht ausgefüllt wurden, kann so auf einen fehlenden Wert geschlossen werden. Damit die Auswertungsergebnisse eindeutig interpretierbar sind, dürfen ferner die fW nicht zusammen mit anderen Ausprägungen einer Variablen in einer Kategorie enthalten sein.

Darüber hinaus muss darauf geachtet werden, dass bei neudefinierten Variablen, welche aus anderen Variablen gemäß einer bestimmten Vorschrift hervorgehen (z.B. BMI), diese Vorschrift von dem verwendeten Softwareprogramm gerade bei Vorliegen von fW adäquat implementiert ist [22]. Weist eine der

originären Variablen einen fW für eine Beobachtung auf, so muss auch die neudefinierte Variable für diese Beobachtung einen fW haben.

Für die Wahl der für den vorliegenden Datensatz geeigneten Ersetzungsmethode von fehlenden Werten und zur Beurteilung der Aussagekraft der Studienergebnisse ist eine ausführliche Missing Data Diagnostic (MDD), siehe Abschnitt 2.1, im Rahmen der Deskription des Datensatzes wichtig. Die MDD beinhaltet eine genaue Beschreibung, bei welcher Variablen wie viele fW auftreten, welches Missing Data Pattern (MDP) vorliegt und welcher Missing Data Mechanism (MDM) vermutet wird. In der Software SAS® kann dies auf einfache Weise mittels des von Hohl *et al.* [42] vorgestellten Makros oder direkt mit der Prozedur PROC MI durchgeführt werden.

Auch wenn in den Simulationsergebnissen keine statistisch signifikanten Unterschiede zwischen den betrachteten (Ersetzungs-) Methoden in Bezug auf die Übereinstimmung der geschätzten mit den tatsächlichen Werten und der Validität der Auswertungsergebnisse erkennbar waren, deuten die Tendenzen darauf hin, dass bei fehlenden Werten nominaler Variablen die Discriminant Function Method (DFM) bzw. bei fehlenden Werten ordinaler Variablen die logistische Regression bevorzugt angewandt werden sollten. Ihre mittleren Abweichungen vom Originalwert waren meistens deutlich geringer als die von den anderen betrachteten Methoden.

Die logistische Regression und DFM haben gegenüber einer Complete Case Analyse (CCA) den Vorteil, dass sie zu keinem Selektionsbias führen. Nur bei Vorliegen eines MDM gemäß Missing Completely At Random (MCAR) führt die CCA, die nur auf Beobachtungen mit vollständig erhobenen Werten basiert, zu validen Ergebnissen. Das betrachtete Teilkollektiv ist nur unter MCAR eine repräsentative Stichprobe der Grundgesamtheit. Andernfalls kann die Verallgemeinerung der Studienergebnisse gehemmt sein.

Neben der Schätzung der tatsächlichen Werte haben die Ersetzungsmethoden primär für stetige Variablen (z.B. lineare Regression oder MCMC Methode) das Problem, dass deren geschätzten Werte nicht unbedingt den Ausprägungen der kategorialen Variablen mit fW entsprechen. Daher müssen die geschätzten Werte gegebenenfalls vor der Ersetzung gerundet werden. Dies birgt eine weitere Gefahr, verzerrte Ergebnisse zu erhalten. Allerdings haben die lineare Regression und MCMC Methode den Vorteil, dass sie auch bei einem beliebigen MDP angewendet werden können, während die logistische Regression und DFM ein monotones Missing Data Pattern voraussetzen.

Mittels einer gemäß Rubin [82] korrekt durchgeführten Multiple Imputation (MI), siehe Abschnitt 2.3.2, wird die Unsicherheit über die geschätzten fehlenden Werte zum Teil berücksichtigt. (Die in dieser Arbeit betrachteten Ersetzungsmethoden haben alle eine MI durchgeführt.) Deterministische Ersetzungsmethoden, wie beispielsweise eine Modusersetzung, die nur eine Single Imputation durchführen können, berücksichtigen diese Unsicherheit nicht und sind daher im Allgemeinen nicht zu empfehlen.

Das Ersetzungsmodell zur Schätzung der fW sollte in Anlehnung an das angestrebte Auswertungsmodell gewählt werden, siehe Abschnitt 3.2. Generell sollte das Ersetzungsmodell eher mehr Variablen als das Auswertungsmodell enthalten, um möglichst unverzerrte Ergebnisse zu erhalten [63]. Ferner ist es wichtig, dass eventuelle Interaktionsterme, die im Auswertungsmodell enthalten sein werden, auch im Ersetzungsmodell berücksichtigt werden.

Insgesamt kann festgehalten werden, dass der zusätzliche Aufwand, der betrieben werden muss, um trotz fW im Datensatz eine relativ gute Aussagekraft der Studienergebnisse zu gewährleisten ziehmlich groß ist! Die „neuen“ Ersetzungsmethoden für kategoriale Variablen führen tendenziell zu valideren Ergebnissen als die anderen betrachteten Methoden, aber sie können nur bei dem eher selten auftretenden monotonen Missing Data Pattern angewandt werden.

# Kapitel 7

# Zusammenfassung

Das Vorliegen von fehlenden Werten (fW) in klinischen Datensätzen führt zu einer Reihe von Problemen. Die fehlenden Werte können die Power einer Studie reduzieren, zu verzerrten Studienergebnissen führen und die Verallgemeinerung der Ergebnisse hemmen. Die beste Lösung für diese Probleme ist die Vermeidung von fW. Spezielle Studiendesigns, eine gewissenhafte Datenerhebung und ein gutes Monitoring können das Auftreten von fW auf ein Minimum reduzieren. Dennoch beinhalten fast alle klinischen Datensätze fehlende Werte.

Zum Umgang mit fehlenden Werten in der statistischen Auswertung gibt es verschiedene Methoden. Die Complete Case Analyse (CCA) berücksichtigt nur Beobachtungen (Fälle) mit vollständig erhobenen Werten in der Auswertung. Dies ist die Standardmethode der meisten Auswertungsverfahren in statistischen Softwareprogrammen. Da die CCA in vielen Situationen zu den oben aufgeführten Problemen führt, wurden in den letzten 2 - 3 Jahrzehnten Ersetzungsmethoden entwickelt, mit denen die fW geschätzt werden und anschließend, basierend auf dem vervollständigten Datensatz, die Standardverfahren zur statistischen Auswertung verwendet werden können.

Es wurden primär Ersetzungsmethoden zur Schätzung fW stetiger normalverteilter Variablen entwickelt. Die Anwendung dieser Methoden bei kategorialen

Variablen führt meist zu ersetzten Werten, die nicht eine der beobachteten Ausprägungen der Variablen annehmen. Das erforderliche Runden auf plausible Werte kann zu verzerrten Ergebnissen führen. Daher wurden im letzten Jahrzehnt auch Ersetzungsmethoden speziell für kategoriale Variablen entwickelt, welche dieses Problem nicht haben. Sie wurden kürzlich in größeren Softwarepaketen (z.B. SAS) implementiert. Der Nachteil dieser Methoden gegenüber den anderen ist jedoch, dass sie nur bei Vorliegen einer eher selten auftretenden Anordnung der fehlenden Werte im Datensatz angewandt werden können.

In der vorliegenden Arbeit wurde im Rahmen einer Simulationsstudie untersucht, wie stark sich die Übereinstimmung der geschätzten mit den tatsächlichen Werten und die Validität der Ergebnisse eines Auswertungsverfahrens zwischen verschiedenen Ersetzungsmethoden und der CCA unter verschiedenen Rahmenbedingungen unterscheiden. Als Ersetzungsmethoden für kategoriale Variablen wurden die logistische Regression und Discriminant Function Method betrachtet und für stetige Variablen die lineare Regression und MCMC Methode. Die Simulationen wurden mit der oft verwendeten Statistiksoftware SAS® Version 9 durchgeführt, die dem Anwender bereits einige Ersetzungsmethoden zur Verfügung stellt.

Die Simulationsergebnisse zeigen keine statistisch signifikanten Unterschiede. Tendenziell führen die Ersetzungsmethoden für kategoriale Variablen zu valideren Ergebnissen. Doch wenn deren Voraussetzung nicht erfüllt ist, können vermutlich die betrachteten Ersetzungsmethoden für stetige Variablen verwendet werden, ohne zu stark verzerrten Ergebnissen zu führen. Eine geeignete Ersetzung ist auf jeden Fall einer CCA vorzuziehen, um die angestrebte Power der Studie zu gewährleisten und einem Selektionsbias entgegenzuwirken, welcher die Aussagekraft der Studienergebnisse stark beeinträchtigen kann.

# Literaturverzeichnis


[1] Ake, Ch.F. Rounding After Multiple Imputation With Non-binary Categorical Covariates, SUGI 30, Paper 112-30. 2005, http://www2.sas.com/proceedings/sugi30/112-30.pdf. (Stand: 11.06.2007).

[2] Allison, P.D. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136., Thousand Oaks, CA: Sage, 2001.

[3] Allison, P.D. Imputation of Categorical Variables with PROC MI, SUGI 30, Paper 113-30. 2005, http://www2.sas.com/proceedings/sugi30/113-30.pdf. (Stand: 11.06.2007).

[4] Baker, S.G., Ko, C.-W., and Graubard, B.I. A sensitivity analysis for nonnormal missing categorical data arising from a national health disability survey. *Biostatistics*, 4:41–56, 2003.

[5] Bernaards, C.A., Belin, T.R., and Schafer, J.L. Robustness of multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26:1368–1382, 2007.

[6] Bernaards, C.A., Farmer, M.M., Qi, K., Dulai, G.S., Ganz, P.A., and Kahn, K.L. Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey. *Journal of Data Science*, 1:1–20, 2003.

[7] Brand, J.P.L. *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data*

*Sets*. Med. Dissertation, Erasmus University Rotterdam, ISBN 90-74479-08-1, 1999.

[8] Burton, A., Altman, D.G., Royston, P., and Holder, R.L. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25:4279–4292, 2006.

[9] Byrt, T., Bishop, J., and Carlin, J. B. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 56(5):423–9, 1993.

[10] Carpenter, J. and Kenward, M. Guidelines for handling missing data in Social Research. Technical report, http://www.lshtm.ac.uk/msu/missingdata/guidelines.pdf. (Stand: 11.06.2007).

[11] Carpenter, J., Kenward, M., Evans, S., and White, I. Last Observation Carry-Forward and Last Obervation Analysis. *Statistics in Medicine*, 23:3241–3244, 2004.

[12] Chen, G. and Astebro, T. How to deal with Missing Categorical Data: Test of a Simple Bayesian Method. *Organizational Research Methods*, 6(10):1–19, 2003.

[13] Chen, H.Y. and Little, R. A test of missing completely at random for generalized estimating equations with missing data. *Biometrika*, 86:1–13, 1999.

[14] Choi, Y.J., Nam, C.M., and Kwak, M.J. Multiple Imputation Technique Applied to Appropriateness Ratings in Cataract Survey. *Yonsei Medical Journal*, 45(5):829–837, 2004.

[15] Cicchetti, D.V. and Allison, T. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Jorunal of EEG Technology*, 11:101–109, 1971.

[16] Cicchetti, D.V. and Feinstein, A.R. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558, 1990.

[17] Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[18] Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psycholigal Bulletin*, 70:213–220, 1968.

[19] Cole, S.R., Chu, H., and Greenland, S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35:1074–1081, 2006.

[20] Collins, L.M., Schafer, J.L., and Kam, C.-M. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, 6(4):330–351, 2001.

[21] Committee for Proprietary Medicinal Products. Points to Consider on Missing Data. November 2001, http://www.emea.eu.int/pdfs/human/ewp/177699EN.pdf. (Stand: 11.06.2007).

[22] Cordeiro, J.A. Logical operators of some statistical computing packages and missing values. *Computational Statistics and Data Analysis*, 51:2783–2787, 2007.

[23] Deal, K. Missing Something: Multiple imputation software might help find missing value data. *Journal of Marketing Research*, 41:44–46, 2004.

[24] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39:1–22, 1977.

[25] Denzer, C., Muche, R., Mayer, H., Heinze, E., Debatin, K.-M., and Wabitsch, M. Serum uric acid levels in obese children and adolescents:

linkage to testosterone levels and pre-metabolic syndrome. *Journal of pediatric endocrinology & metabolism*, 16(9):1225–32, 2003.

[26] Dmitrienko, A., Molenbergs, G., Chuang-Stein,C., and Offen, W. *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Press, 2005.

[27] Engels, J.M. and Diehr, P. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56:968–976, 2003.

[28] Fay, R.E. When are Inference From Multiple Imputation Valid? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 227–232, 1992.

[29] Feinstein, A.R. and Cicchetti, D.V. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

[30] Fleiss, J.L. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York, second edition, 1981.

[31] Fleiss, J.L. and Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as a measure of reliability. *Educational and Psychological Measurement*, 33:613–619, 1973.

[32] Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[33] Gmel, G. Imputation of missing values in the case of multiple item instrument measuring alcohol consumption. *Statistics in Medicine*, 20:2369–2381, 2001.

[34] Graham, J.W., Cumsille, P.E., and Elek-Fisk, E. Methods for handling missing data. In Schinka, J.A. and Velicer, W.F., editors, *Research Methods in Psychology*, pages 87–114. John Wiley and Sons, New York, 2002.

[35] Graham, J.W., Hofer, S.M., and Piccini, A.M. Analysis with missing data in drug prevention research. In Collins,L.M. and Seitz, L., editors, *Advances in data analysis for prevention intervention research*, pages 13–63. National Institute on Drug Abuse Research Monograph Series, Washington DC, 1994.

[36] Graham, J.W. and Schafer, J. On the performance of multiple imputation for multivariate data with small sample size. In Hoyle, R., editor, *Statistical Strategies for Small Sample Research*, pages 1–29. Thousand Oaks, California, 2002.

[37] Harel, O. and Zhou, X.-H. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26:3057–3077, 2007.

[38] Harrell, F.E. Hmisc package. 2006, http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/Hmisc. (Stand: 11.06.2007).

[39] Hastings, W.K. Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97–106, 1970.

[40] Hawthorne, G. and Elliot, P. Imputing cross-sectional missing data: comparison of common techniques. *Australian and New Zealand Journal of Psychiatry*, 39(7):583–590, 2005.

[41] Hohl, K., Muche, R., Brodrecht, K., and Ziegler, C. Ersetzung fehlender Werte in SAS: zwei weiterentwickelte SAS-Makros. In Kaiser, K. and Bödeker, R.-H. , editors, *Statistik und Datenanalyse mit SAS. Proceedings der 10. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE)*, pages 99–108. Shaker-Verlag, Aachen, 2006.

[42] Hohl, K., Ring, Ch., Muche, R., and Ziegler, C. Fehlende Werte in der (Regressions-) Analyse von Datensätzen: zwei SAS-Makros. In Rödel, E. and Bödeker, R.-H. , editors, *SAS: Verbindung von Theorie und Praxis. Proceedings der 9. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE)*, pages 105–116. Shaker-Verlag, Aachen, 2005.

[43] Honaker, J. and King, G. What to do about Missing Values in Time Series Cross-Section Data. 2006, http://gking.harvard.edu/files/pr.pdf. (Stand: 11.06.2007).

[44] Honaker, J., King, G., and Blackwell, M. Amelia Software Web Site. 2006, http://gking.harvard.edu/amelia. (Stand: 11.06.2007).

[45] Horton, N.J. and Kleinman, K.P. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician*, 61(1):79–90, 2007.

[46] Horton, N.J. and Lipsitz, S.R. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. *The American Statistician*, 55(4):244–254, 2001.

[47] Horton, N.J., Lipsitz, S.R., and Parzen, M. A Potential for Bias When Rounding in Multiple Imputation. *The American Statistician*, 57(4):229–232, 2003.

[48] Hu, M., Salvucci, S., and Cohen, M.P. Evaluation of Some Popular Imputation Algorithms. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 308–313, 1998.

[49] Insightful. Missing Data Library for S-Plus. 2007, http://www.insightful.com. (Stand: 11.06.2007).

[50] International Conference on Harmonisation. ICH Topic E6 (R1): Guideline for Good Clinical Practice. Juli 2002, http://www.emea.eu.int/pdfs/human/ich/013595en.pdf. (Stand: 11.06.2007).

[51] International Conference on Harmonisation. ICH Topic E3: Structure and Content of Clinical Study Reports. Juli 1996, http://www.emea.eu.int/pdfs/human/ich/013795en.pdf. (Stand: 11.06.2007).

[52] International Conference on Harmonisation. ICH Topic E9: Statistical Principles for Clinical Trials. September 1998, http://www.emea.eu.int/pdfs/human/ich/036396en.pdf. (Stand: 11.06.2007).

[53] Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. Analyzing Incomplete Discrete Longitudinal Clinical Trial Data. *Statistical Science*, 21:52–69, 2006.

[54] Jones, M.P. Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

[55] Kenward, M.G., Goetghebeur, E.J.T., and Molenberghs, G. Sensitivity analysis for incomplete categorical data. *Statistical Modelling*, 1:31–48, 2001.

[56] King, G., Honaker, J., Joseph, A., and Scheve, K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, 95:49–69, 2001.

[57] Kutschmann, M. Der Kappa-Koeffizient: Diskussion eines Missverständnisses und ein Modifizierungsvorschlag. 2002, http://epi.klinikum.uni-muenster.de/StatMethMed/2002/Berlin/Kutschmann.pdf. (Stand: 11.06.2007).

[58] Lantz, C.A. and Nebenzahl, E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology*, 49(4):431–4, 1996.

[59] Lavori, P.W., Dawson, R., and Shera, D. A multiple imputation strategy for clinical trails with truncation of patient data. *Statistics in Medicine*, 14:1913–1925, 1995.

[60] Little, R.J.A. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83:1198–1202, 1988.

[61] Little, R.J.A. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.

[62] Little, R.J.A. Modeling the Drop-Out Mechanism in Longitudinal Studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.

[63] Little, R.J.A. and Rubin, D.B. *Statistical analysis with missing data*. John Wiley & Sons, New York, 2002.

[64] Liu, M., Taylor, J.M.G., and Belin, T.R. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, 56:1157–1163, 2000.

[65] Liu, M., Wei, L., and Zhang, J. Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study. *Pharmaceutical Statistics*, 5:7–18, 2006.

[66] Meng, X.L. Multiple Imputation Inferences with Uncongenial Sources of Input (with discussion). *Statistical Science*, 9:538–574, 1994.

[67] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. Equation of State Calculations by Fast Computing Machines. *Statistics in Medicine*, 21:1087–1092, 1953.

[68] MICE. MICE library. 2007, http://www.multiple-imputation.com. (Stand: 11.06.2007).

[69] Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Journal fo Chemical Physics*, 21:1023–1041, 2002.

[70] Michiels, B., Molenberghs, G., and Lipsitz, S.R. Selection Models and Pattern-Mixture Models for Incomplete Data with Covariates. *Biometrics*, 55:978–983, 1999.

[71] Minini, P. and Chavance, M. Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics*, 5:531–544, 2004.

[72] Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wassermann, W. *Applied Linear Regression Models*. Irwin book team, Chicago, 1990.

[73] Newgard, C.D. The Validity of Using Multiple Imputation for Missing Out-of-hospital Data in a State Trauma Registry. *Academic Emergency Medicine*, 13(3):314–324, 2006.

[74] Park, T. and Davis, C.S. A Test of the Missing Data Mechanism for Repeated Categorical Data. *Biometrics*, 49:631–638, 1993.

[75] Perez, A., Dennis, R.J., Gil, J.F., Rondon, M.A., and Lopez, A. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. *Statistics in Medicine*, 21:3885–3896, 2002.

[76] Perneger, T.V. and Burnand, B. A simple imputation algorithm reduced missing data in SF-12 health surveys. *Journal of Clinical Empidemiology*, 58(2):142–149, 2005.

[77] Pierchala, C.E. (Northeast SAS Users Group Inc.). PROC MI as the Basis for a Macro for the Study of Patterns of Missing Data. 2003, http://www.nesug.org/proceedings/nesug03/st/st009.pdf. (Stand: 11.06.2007).

[78] Potthoff, R.F., Tudor, G.E., Pieper, K.S., and Hasselblad, V. Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, 15:213–234, 2006.

[79] Raghunathan et al. IVEware: Imputation and Variance Estimation Software. 2007, http://www.isr.umich.edu/src/smp/ive. (Stand: 11.06.2007).

[80] Royston, P. Multiple Imputation of Missing Values. *Stata Technical Journal*, 5:527–536, 2005.

[81] Rubin, D.B. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[82] Rubin, D.B. *Multiple Imputation for nonresponse in surveys*. Wiley, New York, 1987.

[83] Rubin, D.B. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91:473–489, 1996.

[84] SAS Institute Inc. SAS OnlineDoc 9.1. 2003, http://support.sas.com/91doc/docMainpage.jsp. (Stand: 11.06.2007).

[85] SAS Institute Inc. SAS OnlineDoc 9.1, FREQ Procedure. 2003, http://support.sas.com/91doc/getDoc/statug.hlp/freq_index.htm. (Stand: 11.06.2007).

[86] SAS Institute Inc. SAS OnlineDoc 9.1, MI Procedure. 2003, http://support.sas.com/91doc/getDoc/statug.hlp/mi_index.htm. (Stand: 11.06.2007).

[87] Saverymuttu, S.H., Joseph, A.E., and Maxwell, J.D. Ultrasound scanning in the detection of hepatic fibrosis and steatosis. *Lancet*, 292(6512):13–5, 1986.

[88] Schafer, J. NORM, CAT, MIX und PAN. 2007, http://www.stat.psu.edu/~jls/misoftwa.html. (Stand: 11.06.2007).

[89] Schafer, J.L. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, New York, 1997.

[90] Schafer, J.L. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15, 1999.

[91] Schafer, J.L. and Graham J.W. Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):147–177, 2002.

[92] Schafer, J.L. and Olsen, M.K. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, 33:545–571, 1998.

[93] Schnell, R. Programme zur Imputation und multiplen Imputation fehlender Daten. 2001, http://www.uni-konstanz.de/FuF/Verwiss/Schnell/mdprogs.html. (Stand: 08.05.2007).

[94] Shen, C. and Weissfeld, L. Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. *Biostatistics*, 6(2):333–347, 2005.

[95] Sinharay, S., Stern, H.S., and Russell, D. The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Methods*, 6(4):317–329, 2001.

[96] SPSS Inc. SPSS Missing Value Analysis. 2007, http://www.spss.com/missing_value/. (Stand: 11.06.2007).

[97] Statistical Solutions. SOLAS. 2007, http://www.statsol.ie/html/solas/solas_home.html. (Stand: 11.06.2007).

[98] Tang, L., Song, J., Belin, T.R., and Unützer, J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24:2111–2128, 2005.

[99] Tanner, M.A. and Wong, W.H. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

[100] Unnebrink, K. and Windeler, J. Sensitivity analysis by worst and best case assessment: Is it really sensitive? *Drug Information Journal*, 33:835–839, 1999.

[101] Unnebrink, K. and Windeler, J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Statistics in Medicine*, 20:3931–3946, 2001.

[102] Vansteelandt, S., Goetghebeur, E., Kenward, M.G., and Molenberghs, G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16:953–979, 2006.

[103] Verbeke, G. and Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000.

[104] Völkner, T. *Der Einfluss des Umgangs mit fehlenden Werten auf die Evaluation von Behandlungseffekten in Messwiederholungsdesigns*. Diplomarbeit, Universität Freiburg, 2005.

[105] von Hippel, P.T. Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician*, 58:160–164, 2004.

[106] Wabitsch, M., Hauner, H., Hertrampf, M., Muche, R., Hay, B., Mayer, H., Kratzer, W., Debatin, K.-M., and Heinze, E. Type II diabetes mellitus and impaired glucose regulation in Caucasian children and adolescents with obesity living in Germany. *International Journal of Obesity & Related Metabolic Disorders*, 28:307–313, 2004.

[107] Wood, A.M., White, I.R., and Thompson, S.G. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368–376, 2004.

# Anhang

Die beiliegende CD enthält jeweils für den binären und ordinalen Merkmalstyp der Variablen Steatosis hepatis einen Ordner mit den drei Unterordnern *Programme*, *Makros* und *Ergebnisse*.

Für jede Kombination der betrachteten Rahmenbedingungen (Ersetzungsmethode, Missing Data Mechanism und Anteil an fehlenden Werten) ist ein Programm geschrieben worden. Anhand des Namens kann auf die jeweilige Kombination zurückgeschlossen werden. Beispielsweise werden in dem Programm simsh01logregconvmar30.sas die Simulationen für den binären (0,1) Merkmalstyp von Sh (Sh01), bei dem mittels der logistischen Regression die fehlenden Werte unter einem Missing Data Mechanism gemäß eines convexen MAR und 30% fW ersetzt werden, durchgeführt.

Die in den Simulationsprogrammen verwendeten Makros sind in dem Ordner *Makros* abgespeichert. Die Ergebnisse der Simulationen sind für den jeweiligen Missing Data Mechanism in tabellarischer Form in Word-Dokumenten im Ordner *Ergebnisse* enthalten. Diesen Tabellen können auch die geringfügigen Abweichungen der Regressionskoeffizienten von den anderen Einflussgrößen entnommen werden.

Die Datensätze mit den simulierten fehlenden Werten sind aufgrund der großen Datenmenge auf der beiliegenden CD nicht gespeichert. Diese Datensätze kön-

nen aber mittels den Programmen Erzeugung datasets mit fW.sas bzw. binär_Erzeugung datasets mit fW.sas generiert werden.

Zu beachten ist ferner, dass vor der Ausführung der Simulationsprogramme die Pfade, auf welche zugegriffen wird, um die Makros oder die Datensätze mit fehlenden Werten einzulesen, angepasst werden müssen.

# Abbildungsverzeichnis

Abb. 1 Missing Data Pattern . . . . . 13
Abb. 2 Mögliche Verzerrung der Häufigkeitsverteilung durch deterministische Ersetzung. . . . . 19
Abb. 3 Mögliche Verzerrung der Korrelation durch deterministische Ersetzung. . . . . 20
Abb. 4 Konzept einer Multiple Imputation . . . . . 25
Abb. 5 MCMC Methode . . . . . 40
Abb. 6 Kappa Paradoxon Randverteilungen . . . . . 79
Abb. 7 Einfache lineare Regression . . . . . 83

# Tabellenverzeichnis

Tab. 1 Relative Effizienz des Schätzers $Q$ basierend auf MI. . . . . 26
Tab. 2 Übersicht vorgestellter Ersetzungsmethoden. . . . . 43
Tab. 3 Übersicht über Softwarepakete zur Schätzung fehlender Werte. 51
Tab. 4 Auflistung der Variablen im Simulationsdatensatz. . . . . 66
Tab. 5 Deskription der stetigen Variablen. . . . . 68
Tab. 6 Deskription der kategorialen Variablen. . . . . 68
Tab. 7 Kontingenztafel zur Berechnung des Kappa-Koeffizienten. . . 75
Tab. 8 Gewichte $w_{ij}$ des Kappa-Koeffizienten gemäß Fleiss und Cohen. 77
Tab. 9 Gewichte $w_{ij}$ des Kappa-Koeffizienten gemäß Cicchetti und Allison. . . . . 78
Tab. 10 Wahrscheinlichkeiten für das Fehlen der Werte von Sh zur Erzeugung eines linearen MAR. . . . . 91
Tab. 11 Wahrscheinlichkeiten für das Fehlen der Werte von Sh zur Erzeugung eines konvexen MAR. . . . . 92
Tab. 12 Kappa-Koeffizienten und deren Standardfehler (SE) für den binären Merkmalstyp basierend auf allen Werten. . . . . 98
Tab. 13 Kappa-Koeffizienten und deren Standardfehler (SE) für den binären Merkmalstyp basierend auf den geschätzten Werten. . . 100
Tab. 14 Kappa-Koeffizienten und deren Standardfehler (SE) für den ordinalen Merkmalstyp basierend auf allen Werten. . . . . 102
Tab. 15 Kappa-Koeffizienten und deren Standardfehler (SE) für den ordinalen Merkmalstyp basierend auf den geschätzten Werten. . 103
Tab. 16 Mittlere Abweichung des Gesamtmittelwertes vom wahren Mittelwert (MW) und mittlerer RMSE für den binären Merkmalstyp von Sh. Der wahre Mittelwert beträgt 0.2672. . . . . 105

Tab. 17 Mittlere Abweichung des Gesamtmittelwertes vom wahren Mittelwert (MW) und mittlerer RMSE für den ordinalen Merkmalstyp von Sh. Der wahre Mittelwert beträgt 0.3944. . . . . 107
Tab. 18 Mittlere Abweichung der Regressionskoeffizienten von dem wahren Koeffizienten und der mittlere Standardfehler (SE) des binären Merkmalstyps. Der wahre Regressionskoeffizient beträgt 0.225. . . . . 109
Tab. 19 Mittlere Abweichung der Regressionskoeffizienten von den wahren Koeffizienten der drei Dummyvariablen und die mittleren Standardfehler (SE) des ordinalen Merkmalstyps. Der wahre Regressionskoeffizient für D1 beträgt 0.170, für D2 0.311 und für D3 0.396. (**Teil I**) . . . . . 111
Tab. 20 Mittlere Abweichung der Regressionskoeffizienten von den wahren Koeffizienten der drei Dummyvariablen und die mittleren Standardfehler (SE) des ordinalen Merkmalstyps. Der wahre Regressionskoeffizient für D1 beträgt 0.170, für D2 0.311 und für D3 0.396. (**Teil II**) . . . . . 112
Tab. 21 Mittlere Abweichung der F-Statistiken von der wahren F-Statistik (F-Stat.) und der mittlere Standardfehler (SE) beim binären Merkmalstyp. Die wahre F-Statistik beträgt 16.574 und der kritische Wert $F_{1,464-5,0.95} = 5.057$. . . . . 115
Tab. 22 Mittlere Abweichung der F-Statistiken von der wahren F-Statistik des gemeinsamen Einflusses der 3 Dummyvariablen und der mittlerer Standardfehler (SE) beim ordinalen Merkmalstyp. Die wahre F-Statistik beträgt 6.55 und der kritische Wert $F_{3,464-7,0.95} = 3.145$. . . . . 117
Tab. 23 Mittlere Abweichung der $R^2$ vom wahren $R^2$ des Regressionsmodells und die mittleren Standardfehler (SE) beim binären Merkmalstyp. Das wahre $R^2$ beträgt 0.234. . . . . 119
Tab. 24 Mittlere Abweichung der $R^2$ von dem wahren $R^2$ des Regressionsmodells und die mittleren Standardfehler (SE) beim ordinalen Merkmalstyp. Das wahre $R^2$ beträgt 0.239. . . . . 120

## Danksagung

Ich danke allen Personen, die mich bei der Erstellung dieser Arbeit tatkräftig und moralisch unterstützt haben.

Meinem Doktorvater, PD Dr. Rainer Muche, danke ich für die anregenden Diskussionen und Tipps zur Erstellung der Arbeit sowie seine zeitnahe Beantwortung aufgetretener Fragen.

Herrn Professor Ulrich Stadtmüller danke ich, dass er trotz des zeitaufwendigen „Nebenjobs“ als Vizepräsident für die Lehre sich bereit erklärt hat, mein Zweitgutachter zu sein und somit eine Verbindung zu meinem Studium der Wirtschaftsmathematik an der Universität Ulm darstellt.

Meinem Mann, Ralf Hohl, und allen meinen KollegInnen im Institut für Biometrie danke ich für deren Verständnis und Unterstützung bei der Anfertigung meiner Arbeit.

# Lebenslauf

**Kathrin Hohl**

(geb. Kötting)

geboren am 21.08.1978

in Hamburg

**Schulbildung**

| | |
|---|---|
| 1984 - 1988 | Grundschule Hinsbleek, Hamburg |
| 1988 - 1994 | Carl-von-Ossietzky-Gymnasium, Hamburg |
| 1994 - 1995 | Douglas County High School, Georgia, USA |
| 1994 - 1997 | Carl-von-Ossietzky-Gymnasium, Hamburg |

**Hochschulstudium**

| | |
|---|---|
| Okt. 1997 - Mai 2003 | Studium der Wirtschaftsmathematik, Universität Ulm |
| Aug. 2001 - Juli 2002 | Studium der Mathematik, University of Missouri-Rolla, USA |
| | Abschluss: Master of Science in Applied Mathematics |
| Mai 2003 | Abschluss: Diplom-Wirtschaftsmathematikerin |

**Bisherige Tätigkeiten**

| | |
|---|---|
| Okt. 1997 - Juli 2001 | Abendpersonal, Theater Ulm |
| Okt. 1999 - März 2003 | Wissenschaftliche Hilfskraft, Universität Ulm |
| Feb. 2000 - April 2000 | Praktikantin, Allianz, Stuttgart |
| Aug. 2001 - Juli 2002 | Teaching Assistant, University of Missouri-Rolla, USA |
| Seit Juli 2003 | Wissenschaftliche Mitarbeiterin, Institut für Biometrie, Universität Ulm |