Simulationsergebnisse zum Vergleich von Ersetzungsmethoden fehlender Werte von kategorialen Variablen in SAS mit PROC MI

Kathrin Hohl
Institut für Biometrie, Universität Ulm
Schwabstr. 13
89070 Ulm
kathrin.hohl@uni-ulm.de

Zusammenfassung

Fehlende Werte in Datensätzen können zu Fehlinterpretationen der Daten führen. Deshalb muss der Umgang mit fehlenden Werten wohlüberlegt gewählt werden. Eine Möglichkeit ist die Ersetzung dieser Werte durch plausible geschätzte Werte. Es stehen hierfür dem Anwender in SAS in der Prozedur PROC MI mehrere Methoden zur Verfügung. Die Eignung der Methoden hängt von mehreren Faktoren ab, unter anderem vom Merkmalstyp der Variablen mit fehlenden Werten. Seit SAS Version 9.1 sind die beiden Ersetzungsmethoden logistische Regression und Discriminant Function Method speziell für kategoriale Variablen implementiert. In einer Simulationsstudie wurde untersucht, welche Konsequenzen es hat, wenn fehlende Werte kategorialer Variablen durch die beiden neuen Ersetzungsmethoden oder durch Ersetzungsmethoden primär für stetige Variablen ersetzt werden. Die Simulationsergebnisse zeigen, dass sich die Ersetzungsmethoden nur geringfügig im Hinblick auf die Übereinstimmung der ersetzten mit den Originalwerten und der Validität der Ergebnisse einer statistischen Auswertungsmethode unterscheiden.

Schlüsselwörter: Ersetzungsmethoden, PROC MI, Simulation.

1 Einleitung

Die Datenqualität, d.h. die Vollständigkeit und Korrektheit der Informationen im Datensatz, beeinflusst in hohem Maße die Aussagekraft der statistischen Analyseergebnisse. Daher wird bei der Datenerhebung im Rahmen von klinischen Studien u.a. mit Hilfe von Monitoring und Doppelerhebung ein großer Aufwand betrieben, um eine bestmögliche Datenqualität zu gewährleisten. Dennoch treten in den meisten Datensätzen fehlende Werte auf, da man diese teilweise nicht verhindern kann (z.B. durch fehlerhafte Messinstrumenten, Dropouts, (un-)absichtliche Nichtbeant-wortung einer Frage, etc.).

Als fehlende Werte (fW) werden in diesem Artikel diejenigen bezeichnet, welche theoretisch existieren, in dem betrachteten Datensatz aber wider Erwarten nicht vorliegen [11]. Fehlende Werte können bei stetigen Variablen (z.B. Körpergröße oder Alter) und bei kategorialen Variablen (z.B. Geschlecht oder Tumorstadium) auftreten.
Das Auftreten von fehlenden Werten im Datensatz kann zu verzerrten Parameterschätzern, einem erhöhten Fehler 1. Art, inkorrekten Konfidenzintervallen und einem Verlust

von statistischer Power führen. Daher erhöht sich mit fW im Datensatz das Risiko, falsche Schlussfolgerungen aus den Daten zu ziehen. Um dieses Risiko zu minimieren, muss sorgfältig überlegt werden, wie mit den fW in der Auswertung umgegangen werden soll. Eine Vorgehensweise ist die Ersetzung fW.

Zur Ersetzung von fehlenden Werten stetiger Variablen existieren bereits einige Ersetzungsmethoden (z.B. lineare Regression, MCMC/Data Augmentation), die in SAS in der Prozedur PROC MI (siehe [8]) implementiert und in der Literatur bereits eingehend diskutiert worden sind, vergl. [7, 10]. Diese setzen voraus, dass die Variable mit fW normalverteilt ist. Letzteres trifft offensichtlich bei kategorialen Variablen nicht zu, so dass diese Ersetzungsmethoden, angewandt auf derartige Variablen, zu verzerrten Ergebnissen führen können. Seit SAS® Version 9 stehen dem Anwender in PROC MI speziell für die Ersetzung fehlender Werte von nominalen und ordinalen Variablen die beiden Ersetzungsmethoden Discriminant Function Method (DFM) und logistische Regression zur Verfügung.

Die lineare Regressionsmethode und die MCMC-Methode werden von Schafer [9] als robust gegenüber ihrer Vorraussetzung der multivariaten Normalverteilung bezeichnet. Ziel der Simulationen war daher zu untersuchen, wie groß der Vorteil der Verwendung von Ersetzungsmethoden speziell für kategoriale Variablen im Vergleich zu stetigen Variablen ist.

Die Wahl der geeigneten Ersetzungsmethode hängt nicht nur vom Merkmalstyp der Variablen mit fW, sondern auch von der Anordnung der fW im Datensatz (Missing Data Pattern, kurz: MDP) und den Ursachen für das Auftreten von fW (Missing Data Mechanism, kurz: MDM) ab. Wann welche Ersetzungsmethode aus theoretischer Sicht angezeigt ist, wurde bereits auf der 10. KSFE vorgestellt [5] und wird in Tabelle 1 für die in dieser Arbeit betrachteten Ersetzungsmethoden aufgelistet.

Tabelle 1: Anwendbarkeit ausgewählter Ersetzungsmethoden.

Ersetzungsmethoden	Merkmalstyp			Missing Data Pattern		Missing Data Mechanismus		
	nom.	ordinal	stetig	monoton	bel.	MCAR	MAR	MNAR
Regressionsersetzung			x	x		x	x	
MCMC			x	x	x	x	x	
Logistische Regression		x		x		x	x	
Discriminant Function M.	x			x		x	x	

In einer in SAS® Version 9.1 durchgeführten Simulationsstudie wurde untersucht, welche empirischen/praktischen Konsequenzen die Verwendung ausgewählter Ersetzungsmethoden zu Schätzung fW kategorialer Variablen unter verschiedenen Rahmenbedingungen mit sich bringen. Dabei wurde der Anteil an fW im Datensatz, der MDM und der Merkmalstyp der Variablen mit fW variiert.

Anhand der Simulationen wurde zum einen die Übereinstimmung der mittels verschiedener Ersetzungsmethoden geschätzten Werte mit den Originalwerten analysiert mit dem Ziel, die Plausibilität und Genauigkeit der geschätzten Werte zu vergleichen. Zum anderen wurde die Validität der Ergebnisse einer statistischen Auswertungsmethode basierend auf den vervollständigten Datensätzen untersucht.
Im folgenden Kapitel werden der Simulationsdatensatz, die betrachteten Rahmenbedingungen und die Parameter zum Vergleich der Ersetzungsmethoden vorgestellt. Das Kapitel 3 zeigt (einen Teil) der Simulationsergebnisse und im Kapitel 4 werden die Ergebnisse diskutiert und interpretiert.

2 Simulationsaufbau

Die Simulationen basierten auf einem realen Datensatz aus dem medizinischen Umfeld[1]. Damit ein Vergleich zwischen dem ersetzten und Originalwert möglich war, wurden nur die 464 der 541 Beobachtungen im Datensatz berücksichtigt, für die alle Werte der für die Simulationen interessierenden Variablen erhoben worden sind. Der Simulationsdatensatz bestand aus den 4 stetigen Variablen Insulin, Alter, Adiponektin und Taillenumfang, der binären Variablen Geschlecht und der ordinalen Variablen Steatosis hepatis (Fettleber), welche in 4 Grade untergliedert war. In den Simulationen wurde die Variable Steatosis hepatis (Sh) in einem Teil als ordinale Variable und in dem anderen als binäre Variable aufgefasst. Die entsprechenden Häufigkeitsverteilungen von Sh sind in der Tabelle 2 dargestellt.

Tabelle 2: Deskription der Variablen Steatosis hepatis (Sh).

Variable	Ausprägung	N	%
Sh_binär	Keine Fettleber	340	73.3
	Fettleber Grad 1-3	124	24.6
Sh_ordinal	Keine Fettleber	340	73.3
	Fettleber Grad 1	72	15.5
	Fettleber Grad 2	45	9.7
	Fettleber Grad 3	7	1.5

Da der Simulationsdatensatz ursprünglich keine fehlenden Werte enthielt, konnten verschiedene Anteile an fW im Datensatz und unterschiedliche Missing Data Mechanismen simuliert werden. Das Missing Data Pattern wurde nicht variiert, da die speziellen Ersetzungsmethoden für kategoriale Variablen nur bei einem monotonen MDP anwendbar sind. Deshalb wurden fW nur in der Variablen Sh erzeugt.
In Anlehnung an die oft in realen Datensätzen beobachteten Anteile an fW, wurden Datensätze mit 10%, 30% und 50% fW in der Variablen Sh generiert. 10% fehlende Werte ist ein eher geringer Anteil, bei dem gegebenenfalls eine Complete Case Analyse

[1] Der für die Simulationen verwendete Datensatz ist ein Teildatensatz von den Untersuchungen zur Ermittlung der Prävalanz von Typ-2-Diabetes von in Deutschland lebenden übergewichtigen Kindern und Jugendlichen. Für Details zu dieser Studie und deren Ergebnisse sei auf die Publikation [12] verwiesen.

(CCA) durchgeführt werden könnte, sofern diese nicht zu einem Selektionsbias führen würde. Bei einem Anteil von 30% fW ist die Fallzahlreduzierung bei einer möglichen CCA recht groß und kann daher einen deutlichen Einfluss auf die Power der angestrebten Auswertungsmethode haben. Der Anteil von 50% fW im Datensatz führt dazu, dass bei der Hälfte der Beobachtungen fW geschätzt werden müssen und die in der Auswertung berücksichtigte Information auf einen deutlichen Anteil an ersetzten Werten basiert. Eine Diskussion darüber, ob in diesem Fall überhaupt eine Auswertung durchgeführt werden sollte, ist hier sicher angebracht. Welche Konsequenzen aus der Ersetzung von fW in der jeweiligen Situation resultieren, sollten daher in den Simulationen untersucht werden.

Es wurden 5 verschiedene Missing Data Mechanism (MDM) simuliert: Missing Completely At Random (MCAR), zwei Arten von Missing At Random (MAR) und zwei Arten von Missing Not At Random (MNAR). Die Vorgehensweise in den Simulationen erfolgte in Anlehnung an Collins et al. [3] und wurde in SAS in Data Steps realisiert. Mit der Funktion `rantbl(seed,p1)` wurde eine binäre Variable miss generiert. Mit `seed` wurde explizit ein Seed für den Zufallszahlengenerator übergeben und das 2. Argument der Funktion (`p1`) gab die Wahrscheinlichkeit an, mit der die Variable die Ausprägung 1 annahm. Im nächsten Schritt wurden von den Beobachtungen, bei denen miss=1 auftrat, der Wert von Sh auf fehlend gesetzt.

Bei MCAR wurde 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 war abhängig von dem gewünschten Anteil an insgesamt fehlenden Werten im Datensatz.

Bei dem Missing Data Mechanism MAR fehlt der Wert einer Variablen, abhängig von dem Wert einer anderen Variablen im Datensatz, die ebenfalls im Ersetzungsmodell enthalten ist. Da davon ausgegangen wurde, dass es einen leichten, gleichsinnigen Zusammenhang zwischen Sh und dem Taillenumfang gibt, wurden im Simulationsdatensatz die fehlenden Werte von Sh abhängig von dem erhobenen Wert vom Taillenumfang je Beobachtung erzeugt. Die Intention bei diesem Vorgehen war, dass der Taillenumfang die Funktion der Variablen im Ersetzungsmodell übernahm, welche primär zur besseren Schätzung der fehlenden Werte im Modell enthalten ist.

Bei MAR können die fehlenden Werte der einen Variablen unter verschiedenen Abhängigkeitsfunktionen von der anderen Variablen auftreten. So ist es denkbar, dass mit wachsendem Taillenumfang die Wahrscheinlichkeit für das Fehlen der Werte von Sh abnimmt. In diesem Fall ist die Abhängigkeitsfunktion linear und wird im Folgenden als lineares MAR bezeichnet. Diese Abhängigkeit wurde modelliert, in dem für Beobachtungen mit Werten vom Taillenumfang, die kleiner als das 1. Quartil waren, der Wert von Sh mit der größten Wahrscheinlichkeit fehlte. Mit zunehmendem Quartil verringerte sich die Wahrscheinlichkeit für das Auftreten von fW von Sh. Die jeweiligen

Wahrscheinlichkeiten für das Fehlen der Werte innerhalb eines Quartils sind abhängig von dem insgesamt zu erzielenden Anteil an fW im Datensatz und sind in Tabelle 3 angegeben. Die angegebenen Wahrscheinlichkeiten für exemplarisch 10% fW im Datensatz gewährleisten, dass insgesamt 0.16 * 0.25 + 0.12 * 0.25 + 0.08 * 0.25 + 0.04 * 0.25 = 10 % der Werte fehlen.

Tabelle 3: Wahrscheinlichkeiten für das Fehlen der Werte von Sh abhängig vom Quartil des Taillenumfangs und dem Anteil an fW insgesamt zur Erzeugung eines linearen MAR.

Quartil Taillenumfang	Anteil an fehlenden Werten insgesamt		
	10%	30%	50%
bis 1.	0.16	0.48	0.8
1. – 2.	0.12	0.36	0.6
2. – 3.	0.08	0.24	0.4
ab 3.	0.04	0.12	0.2

Es könnte aber auch sein, dass fehlende Werte von Sh für Beobachtungen mit besonders großen oder kleinen Taillenumfängen (bezogen auf dieses Kollektiv) mit einer größeren Wahrscheinlichkeit als für Beobachtungen mit einem durchschnittlichen Taillenumfang auftreten. In einer derartigen Situation wird der MDM als konvexes MAR bezeichnet. Als Schranke für besonders kleine bzw. große Werte, wurde das 1. und 3. Quartil vom Taillenumfang herangezogen. Die verwendeten Wahrscheinlichkeiten zur Simulierung des konvexen MAR sind in Tabelle 4 enthalten.

Tabelle 4: Wahrscheinlichkeiten für das Fehlen der Werte von Sh abhängig vom Quartil des Taillenumfangs und dem Anteil an fW insgesamt zur Erzeugung eines konvexen MAR.

Quartil Taillenumfang	Anteil an fehlenden Werten insgesamt		
	10%	30%	50%
bis 1.	0.15	0.45	0.75
1. – 3.	0.05	0.15	0.25
ab 3.	0.15	0.45	0.75

Keine der hier betrachteten Ersetzungsmethoden ist theoretisch geeignet zur Ersetzung fW, wenn ein MDM gemäß MNAR vorliegt. Da aber nicht nachgewiesen werden kann, ob tatsächlich ein MAR vorliegt, wurde in den Simulationen untersucht, was für Konsequenzen die Ersetzung bei MNAR hat. Die Simulierung der 2 Arten von MNAR basierte auf den Datensätzen mit fehlenden Werten von Steatosis hepatis gemäß den 2 Arten des MAR, allerdings war die Variable Taillenumfang nicht im Ersetzungsmodell enthalten.

Die Ersetzungen der fW wurden mit der Prozedur PROC MI durchgeführt. Folgende Ersetzungsmethoden wurden verglichen: logistische Regression, Discriminant Function Method (DFM), lineare Regression und MCMC-Methode. Die Ergebnisse der statisti-

schen Auswertungsmethode wurden zusätzlich den Ergebnissen der Complete Case Analyse (CCA) gegenübergestellt, da diese eine häufige Umgangsweise mit fW ist. Theoretisch führt die CCA bei Vorliegen von MCAR zu keinen verzerrten Ergebnissen [9]. Da die DFM nur zur Ersetzung von fW nominaler Variablen entwickelt wurde [1], wurde sie nur zur Ersetzung der fW des binären Merkmalstyps von Sh angewandt.
Bei allen Ersetzungsmethoden wurde das gleiche *Ersetzungsmodell* gewählt:

$$Sh = Adiponektin + Alter + Insulin + Taillenumfang. \quad (1)$$

Bei der linearen Regressionsmethode und der MCMC-Methode wurden die ersetzten Werte anschließend auf plausible Werte der Variablen Sh gerundet. Bei den 2 Arten von MNAR war der Taillenumfang im Ersetzungsmodell nicht enthalten.

Als Maß für die zufallskorrigierte Übereinstimmung der geschätzten mit den Originalwerten wurde beim binären Merkmalstyp der einfache Kappa-Koeffizient [4] berechnet. Dieser berechnet sich aus der Differenz zwischen beobachteter und zufalls-bedingter erwarteter Übereinstimmung adjustiert für den theoretisch möglichen Anteil, der über den Zufall hinausgehenden übereinstimmenden Werten (Beurteilungen). Beim ordinalen Merkmalstyp wurde der gewichtete Kappa-Koeffizient mit Gewichten von Cicchetti und Allison [2] berechnet. Letzterer hat den Vorteil gegenüber dem ungewichteten Koeffizienten, dass er die größenmäßige Abweichung von der Übereinstimmung berücksichtigt.

Die Validität der Auswertungsergebnisse basierend auf den vervollständigten Datensätzen sollte anhand einer Auswertungsmethode analysiert werden, welche oft verwendet wird. Da eine häufig untersuchte Fragestellung in der medizinischen Forschung die Untersuchung des (simultanen) Einflusses von mehreren Variablen auf eine stetige Variable (Zielgröße) ist, wurde eine multiple lineare Regressionsanalyse durchgeführt. Das verwendete *Regressionsmodell* lautete:

$$Insulin = Alter + Geschlecht + Adiponektin + Taillenumfang + Steatosis\ hepatis. \quad (2)$$

Die Validität der Auswertungsergebnisse der einzelnen Ersetzungsmethoden in Abhängigkeit von den gegebenen Rahmenbedingungen wurde anhand von mehreren Parametern beurteilt. Aus Platzgründen werden jedoch im nächsten Kapitel nur die Simulationsergebnisse bezüglich der F-Statistik des zu Sh zugehörigen Regressionskoeffizienten präsentiert.

Es wurden je Kombination an Rahmenbedingungen (Anteil an fW, MDM und Merkmalstyp von Sh) 1000 Datensätze mit fehlenden Werten erzeugt. Diese wurden mit jeder Ersetzungsmethode gemäß einer Multiple Imputation vervollständigt, in dem für jeden Datensatz mit fW, 5 vervollständigte Datensätze generiert wurden. Eine Variation der Anzahl an vervollständigten Datensätze abhängig vom Anteil an fW wurde als nicht relevant erachtet, da selbst bei einem Anteil von 50% fW die relative Effizienz der Parameterschätzer mit 90% noch akzeptable erscheint, vgl. [6].
Für jeden der 5*1000 = 5000 vervollständigten Datensätze wurden der jeweilige Kappa-Koeffizient mit der Prozedur PROC FREQ und der Option AGREE berechnet und mit der Prozedur PROC REG die multiple lineare Regression durchgeführt. Die Zusammen-

fassung der berechneten Schätzer für die Kappa-Koeffizienten und deren Varianzen von den fünf vervollständigten Datensätzen je ursprünglichem Datensatz mit fW erfolgte gemäß den Empfehlungen von Rubin [6]. Zur Zusammenfassung der Parameterschätzer der linearen Regression wurde die Prozedur PROC MIANALYZE verwendet.

Da sich die Ergebnisse der linearen Regressionsmethode nur marginal von denen der MCMC-Methode unterscheiden, wird im Folgenden auf die Ergebnisdarstellung der linearen Regression aus Platzgründen verzichtet.

3 Simulationsergebnisse

Der mittlere Kappa-Koeffizient für den binären Merkmalstyp von Steatosis hepatis je Kombination an Rahmenbedingungen ist in Tabelle 5 abgebildet. Zusätzlich wird dessen Standardfehler (SE) angegeben, um die (Un-) Sicherheit der Schätzung beurteilen zu können und zu entscheiden, ob sich die Kappa-Koeffizienten signifikant zwischen den Ersetzungsmethoden unterscheiden.

Tabelle 5: Mittlerer Kappa-Koeffizient und dessen Standardfehler (SE) der *binären* Variablen Steatosis hepatis.

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
	MCMC	0.911	0.027	0.737	0.046	0.564	0.056
k. 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
	MCMC	0.913	0.027	0.742	0.046	0.573	0.058
k. 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
	MCMC	0.907	0.028	0.722	0.047	0.536	0.058
l. 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
	MCMC	0.919	0.026	0.758	0.044	0.599	0.055
l. 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
	MCMC	0.911	0.028	0.735	0.046	0.558	0.056

Die Simulationsergebnisse 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.

Betrachtet man die verschiedenen Missing Data Mechanismen, so lässt sich erkennen, dass unter dem linearen MAR alle Ersetzungsmethoden innerhalb jeden Anteils an fW die beste Übereinstimmung erzielen. Gleichzeitig ist hier der Standardfehler des Schätzers am kleinsten. 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 etwas besser, d.h. größer respektive kleiner, sind als von der MCMC-Methode, die primär zur Ersetzung von stetigen Variablen entwickelt wurde. Allerdings kann, unter Berücksichtigung der Größe der SE, kein signifikanter Unterschied zwischen den Kappa-Koeffizienten gezeigt werden. Daher weisen die beobachteten Unterschiede nur Tendenzen auf.

Die Ergebnisse zum Vergleich der Übereinstimmung bei dem ordinalen Merkmalstyp von Sh sind in der Tabelle 6 enthalten. Diese Simulationsergebnisse 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 der Standardfehler (SE) des Schätzers zunimmt.

Tabelle 6: Mittlerer gewichteter Kappa-Koeffizient und dessen Standardfehler (SE) der *ordinalen* Variablen Steatosis hepatis.

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
	MCMC	0.906	0.025	0.724	0.041	0.546	0.049
k. MAR	Log. Reg.	0.917	0.026	0.750	0.043	0.584	0.054
	MCMC	0.907	0.025	0.725	0.041	0.549	0.050
k. MNAR	Log. Reg.	0.909	0.027	0.726	0.045	0.538	0.056
	MCMC	0.899	0.026	0.702	0.042	0.509	0.051
l. MAR	Log. Reg.	0.934	0.023	0.804	0.039	0.675	0.051
	MCMC	0.918	0.023	0.757	0.039	0.602	0.049
l. MNAR	Log. Reg.	0.923	0.026	0.770	0.043	0.615	0.054
	MCMC	0.908	0.025	0.728	0.041	0.553	0.049

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 schlechter ist die Übereinstimmung bei linearem MNAR, MCAR und konvexem MAR. Analog zu den Ergebnissen des binären

Merkmalstyps, ist bei Vorliegen eines konvexen MNAR die Übereinstimmung am schlechtesten.

Generell ist die Übereinstimmung zwischen ersetzten und Originalwerten bei der logistischen Regression etwas besser als bei der MCMC-Methode, aber ein signifikanter Unterschied in den Kappa-Koeffizienten ist aufgrund der Größe der SE nicht erkennbar.

Zur Beurteilung der Validität der Auswertungsergebnisse basierend auf den vervollständigten Datensätzen wurden diese mit den Auswertungsergebnissen des Originaldatensatzes verglichen. Zur schnelleren Beurteilung der unter- bzw. Überschätzung der Vergleichsparameter wurde die Differenz zwischen geschätztem und wahrem Parameterschätzer vom Originaldatensatz berechnet. Eine negative Abweichung deutet daher auf eine Unterschätzung des Parameters hin.

In Tabelle 7 wird die mittlere Abweichung der zum Regressionskoeffizienten von Sh zugehörigen F-Statistik von der F-Statistik des Originaldatensatzes präsentiert. Die „wahre“ F-Statistik für den binären Merkmalstyp von Sh beträgt 16.574. Da der kritische Wert 5.057 ist, hat Sh einen deutlichen Einfluss auf Insulin im Originalmodell. Bei einer Unterschätzung von bis zu 11.517 der F-Statistik, würde das Ergebnis basierend auf dem vervollständigten Datensatz noch zur gleichen Interpretation bzgl. des Einflusses von Sh auf Insulin führen.

Wie schon bei den Kappa-Koeffizienten gesehen wurde, hat der Anteil an fehlenden Werten im Datensatz 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 zu und der zugehörige Standardfehler der F-Statistik vergrößert sich. Bei einem Anteil von 50% fW im Datensatz könnte sogar der Regressionskoeffizient der Variablen Sh (unter Berücksichtigung des Standardfehlers) eine F-Statistik haben, die kleiner als der kritische Wert ist, sodass geschlussfolgert werden würde, dass Sh keinen Einfluss auf Insulin zu haben scheint!

Bei diesem Vergleichskriterium ist die mittlere Abweichung der F-Statistik innerhalb der jeweiligen Ersetzungsmethode unter einem MDM gemäß MNAR am kleinsten. Am stärksten wurde die F-Statistik bei konvexem MAR unterschätzt. Die Standardfehler hingegen waren bei MNAR deutlich größer als bei MAR oder MCAR.

Bei jeder Kombination der Rahmenbedingungen war die F-Statistik basierend auf den ersetzten Werten mittels DFM am wenigsten unterschätzt, aber deren SE am größten unter den betrachteten Ersetzungsmethoden und der Complete Case Analyse (CCA).

Tabelle 7: Mittlere Abweichung der F-Statistik und deren Standardfehler (SE) des binären Merkmaltyps von Sh.

MDM	Ersetzungs-methode	Anteil an fehlenden Werten					
		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
	MCMC	-2.558	2.758	-6.078	4.054	-8.776	4.541
k. 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
	MCMC	-2.871	2.655	-6.589	3.883	-9.400	3.955
k. 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
	MCMC	-2.049	2.887	-4.837	4.340	-7.465	4.836
l. 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
	MCMC	-2.371	2.653	-5.380	3.992	-7.890	4.280
l. 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
	MCMC	-1.572	3.023	-3.657	4.442	-5.278	5.269

Abgesehen von den MDM gemäß MNAR führte CCA zu den nächst kleinsten Unterschätzungen der F-Statistik. Beachtenswert ist die relativ große Unterschätzung von CCA bei Vorliegen von MCAR. Die mittels der logistischen Regression vervollständigten Datensätze unterschätzten die F-Statistik etwas mehr als die von DFM und deutlich geringer als die von der MCMC-Methode. Doch aufgrund der großen SE sind die beschriebenen Unterschiede nicht signifikant.

Die Abweichungen von der F-Statistik des Originaldatensatzes bei dem ordinalen Merkmalstyp von Sh sind in Tabelle 8 aufgelistet. Die wahre F-Statistik anhand der untersucht wird, ob die Variable insgesamt betrachtet einen Einfluss auf Insulin hat, beträgt 6.55. Der zugehörige kritische Wert ist 3.15. Somit würde eine Unterschätzung von mehr als 3.4 zu einer anderen Schlussfolgerung als im Originaldatensatz führen.

Die Simulationsergebnisse aus Tabelle 8 zeigen, dass die mittleren Abweichungen der F-Statistiken betragsmäßig bei der ordinalen Variablen nicht so groß sind wie bei der binären Variablen. Das gleiche gilt für die SE. Die Abweichungen nehmen daher mit zunehmendem Anteil an fW im Datensatz (absolut betrachtet) nicht so deutlich zu, wie bei dem binären Merkmalstyp von Sh. Ferner unterscheiden sich die Ergebnisse in

Tabelle 8: Mittlere Abweichung der F-Statistik und deren Standardfehler (SE) des ordinalen Merkmaltyps von Sh.

MDM	Ersetzungs-methode	Anteil an fehlenden Werten					
		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
	MCMC	-0.192	1.062	-0.515	1.836	-0.721	2.574
k. MAR	CCA	-0.624	0.902	-1.789	1.274	-3.020	1.514
	Log. Reg.	0.007	1.054	-0.010	1.899	-0.208	2.643
	MCMC	-0.268	1.063	-0.653	1.801	-1.106	2.254
k. MNAR	Log. Reg.	0.616	1.235	1.650	2.239	2.015	3.198
	MCMC	0.158	1.176	0.417	2.010	0.418	2.662
l. 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
	MCMC	-0.076	0.975	-0.118	1.656	0.132	2.286
l. MNAR	Log. Reg.	0.872	1.158	2.837	2.280	5.395	3.546
	MCMC	0.378	1.115	1.325	1.999	2.749	2.794

Hinblick auf den Merkmalstyp in dem bei der ordinalen Variablen sowohl Unter- als auch Überschätzungen abhängig von der betrachteten Ersetzungsmethode auftreten.

Beim ordinalen Merkmalstyp von Sh sind betragsmäßig die mittleren Abweichungen der F-Statistiken innerhalb der einzelnen Ersetzungsmethode unter MCAR und den beiden Arten von MAR kleiner als bei dem MDM gemäß MNAR, auch die zugehörigen SE sind geringfügig kleiner.

Generell führt die logistische Regression eher zu einer Überschätzung der F-Statistik und die MCMC und CCA zu einer Unterschätzung. Dabei ist die Abweichung von CCA größer als bei MCMC. Außer bei dem konvexem MAR ist die mittlere Abweichung der F-Statistik von der logistischen Regression betragsmäßig größer als bei der MCMC-Methode und hat auch einen marginal größeren SE.

4 Diskussion und Schlussfolgerungen

Die Simulationen sind im Rahmen einer Dissertation durchgeführt worden, in der eine Reihe von Vergleichsparameter untersucht worden sind. Aus Platzgründen wurde hier nur ein Teil der Vergleichsparameter und deren Ergebnisse vorgestellt. Es ist daher nicht möglich anhand dieser Ergebnisse allgemeingültige Aussagen abzuleiten, aber Tendenzen können aufgezeigt werden.

Die deutliche Abhängigkeit der Kappa-Koeffizienten von dem Anteil an fehlenden Werten im Datensatz ist plausibel vor dem Hintergrund, dass bei dem kleinsten Anteil

an fW der Anteil an eindeutiger Übereinstimmung am größten ist, da alle nichtersetzten Werte automatisch genau übereinstimmen.

Die beobachteten geringfügigen Unterschiede in den Kappa-Koeffizienten zwischen den Ersetzungsmethoden deuten darauf hin, dass die Ersetzungsmethoden speziell für kategoriale Variablen nicht zu einer deutlich besseren Übereinstimmung der geschätzten mit den tatsächlichen Werten führen, im Vergleich zu der allgemein einsetzbareren Ersetzungsmethode, der MCMC-Methode. Dadurch, dass man bei der MCMC-Methode die Möglichkeit hat die ersetzten Werte zu runden, sind die ersetzten Werte ebenso plausibel wie die mittels der logistischen Regression oder DFM geschätzten Werte. Unabhängig von dem Merkmalstyp der kategorialen Variablen konnte daher die Übereinstimmung betreffend, kein Vorteil der beiden neuen Methoden entdeckt werden.

Bei der Beurteilung der Validität der Auswertungsergebnisse wurde hier nur die Abweichung der F-Statistik von der wahren F-Statistik vorgestellt. Dieser Parameter allein reicht nicht aus, um sichere Aussagen über die Validität zu treffen. Es sollten ferner die Abweichung der Regressionskoeffizienten und des R^2 in der Beurteilung berücksichtigt werden (wie dies in der Dissertation der Fall sein wird).

Die beobachteten Abweichungen der F-Statistiken zeigen allerdings, dass die CCA auch bei Vorliegen eines MCAR zu einer deutlichen Unterschätzung eines Parameters führen kann. Ferner scheinen die Ersetzungsmethoden speziell für kategoriale Variablen den Einfluss der kategorialen Variablen nicht ganz so stark zu unterschätzen, wie Ersetzungsmethoden für stetige Variablen angewandt auf kategoriale Variablen. Dies bedeutet, dass zumindest in dem verwendeten Datensatz die Ersetzungsmethoden für stetige Variablen noch konservativer waren als die für kategoriale Variablen.

Darüber hinaus zeigen die Simulationsergebnisse, dass die Abweichung der F-Statistiken zwar abhängig von dem Vorliegenden MDM sind, aber das Vorliegen von MNAR (zumindest in diesem Beispieldatensatz) nicht notwendigerweise zu einer schlechteren Schätzung führt.

Während bei dem binären Merkmalstyp von Sh die Abweichungen der F-Statistiken von den Ersetzungsmethoden für kategoriale und stetige Variablen in die gleiche Richtung zeigten, traten bei dem ordinalen Merkmalstyp von Sh durchaus gegenläufige Abweichungen auf. Dies zeigt, dass die Ersetzungsmethoden zumindest bei Vorliegen von MCAR oder MAR geeignet sind, um Sensitivitätsanalysen durchzuführen. Bei dem Vorliegen von MNAR haben alle Ersetzungsmethoden die F-Statistik überschätzt, d.h. der Fehler 1. Art könnte tendenziell größer sein, als ursprünglich festgelegt.

Vorläufiges Fazit der Simulationen: Es konnten in Bezug auf die Übereinstimmung der geschätzten mit den Originalwerten und der Validität von Auswertungsergebnissen keine eindeutigen Unterschiede zwischen den betrachteten Ersetzungsmethoden logistische Regression, Discriminant Function Method, lineare Regression und MCMC-Me-

thode festgestellt werden. Der große Vorteil von der MCMC-Methode gegenüber den anderen Methoden ist allerdings deren allgemeine Anwendbarkeit, da sie auch bei einem beliebigen Missing Data Pattern durchgeführt werden kann.

Literatur

[1] Brand, J.P.L. (1999). Development, Implementation and Evaluation of Multiple Imputation: Strategies for the Statistical Analysis of Incomplete Data Sets. Ph.D. Thesis, TNO Prevention and Health/Erasmus University Rotterdam. ISBN 90-74479-08-1.

[2] Cicchetti, D.V.; Allison, T. (1971): A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, 11, 101-109.

[3] Collins LM, Schafer JL, Kam CM (2001): A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. 6(4), 330-51.

[4] Fleiss, JL (1981): Statistical methods for rates and proportions, John Wiley & Sons, Inc., 2. Auflage.

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

[6] Rubin, D.B. (1987): *Multiple imputation for nonresponse in surveys*. John Wiley and Sons Inc., New York.

[7] Rubin, D.B. (1996): Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

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

[9] Schafer, J.L. (1997): Analysis of incomplete multivariate data. Chapman & Hall, London.

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

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

[12] Wabitsch, M.; Hauner, H.; Hertrampf, M.; Muche, R.; Hay, B.; Mayer, H.; Kratzer, W.; Debatin, K.-M.; Heinze, E. (2004): Type II diabetes mellitus and impaired glucose regulation in Caucasian children and adolescents with obisity living in Germany. *Int J Obes Relat Metab Disord.,* 28, 307-313.