Möglichkeiten der Imputation fehlender Werte in SAS – eine Übersicht –

Benjamin Mayer, Rainer Muche
Institut für Biometrie
Schwabstraße 13
89075 Ulm
benjamin.mayer@uni-ulm.de

Zusammenfassung

SAS stellt für Studien aus dem klinischen und pharmazeutischen Bereich aufgrund der Validierung seiner Methoden eines der am meisten benutzten Auswertungsprogramme für statistische Analysen dar. Oftmals treten im Rahmen dieser Studien trotz umfangreicher Qualitätsanstrengungen bei der Datenerhebung fehlende Werte auf. Dieser Umstand stellt aus verschiedenen Gründen ein Problem dar, es können beispielsweise Informationsverlust, eine verminderte Aussagekraft der Studienergebnisse oder verzerrte Parameterschätzer resultieren. Zudem kann es zu einem Verlust an statistischer Power kommen, wenn für die Analyse von Daten beispielsweise Standardmethoden verwendet werden, da diese vollständiges Datenmaterial voraussetzen und deshalb die unvollständigen Fälle nicht berücksichtigen.

Aus diesem Grund besteht die Notwendigkeit, die fehlenden Werte eines Datensatzes angemessen zu ersetzen. Mittlerweile existieren zwar Richtlinien zum Umgang mit fehlenden Werten (CHMP „Guideline on Missing Data in Confirmatory Clinical Trials“, 2009), eine einheitliche Handhabung der Problematik kann darin jedoch nicht angegeben werden. Dennoch gibt es seit einigen Jahren Strategien (Single Imputation (SI) und Multiple Imputation (MI)) und Methoden zur Ersetzung fehlender Werte (Rubin, Schafer), um die angesprochenen Probleme anzugehen und teilweise zu lösen. In diesem Beitrag möchten wir eine Übersicht über die in SAS 9.2 verfügbaren Möglichkeiten (PROC MI und verschiedene Makros) zur Behandlung und Ersetzung fehlender Werte geben.

Schlüsselworte: fehlende Werte, Imputation fehlender Werte, PROC MI

1 Einleitung

In nahezu allen klinischen und pharmazeutischen Studien stellen fehlende Werte einen problematischen Aspekt dar. Das Ziel einer vollständigen Datenerhebung kann nur selten erreicht werden, da aufgrund verschiedenster Ursachen zumindest einzelne Fehlwerte nicht immer vermieden werden können. Gewöhnliche Auswertungsmodelle, wie sie in den meisten statistischen Standardsoftwareprodukten implementiert sind, basieren jedoch auf einem vollständigen Datensatz, der erfordert, dass für alle Variablen jeder einzelne Wert vorhanden ist. Im Falle eines fehlenden Wertes muss deshalb die betreffende Beobachtungseinheit, z.B. ein Patient, aus dem Auswertungskollektiv gestrichen werden, wenn die Daten mit Hilfe der Standardpakete analysiert werden sollen.

Dieser Ansatz bezeichnet die Vorgehensweise der so genannten Complete Case Analyse (CCA), bei der nur vollständig erhobene Beobachtungseinheiten für die Datenanalyse berücksichtigt werden (siehe Abbildung 1). Die CCA bringt jedoch eine ganze Reihe bedeutender Probleme mit sich, die ihre Verwendbarkeit mehr als in Frage stellen: Die Nichtberücksichtigung ganzer Beobachtungseinheiten kann dazu führen, dass die Fallzahl drastisch reduziert wird, die Variabilität der Merkmale sich verändert, die Aussagekraft der Studie vermindert wird und Parameterschätzer aufgrund der evtl. zerstörten Strukturgleichheit verzerrt sind. Darüber hinaus steht sie in Widerspruch zum Auswertungsprinzip der so genannten Intention-to-treat-Analyse (ITT), bei der alle Studienteilnehmer entsprechend der Randomisierung auszuwerten und für die Analyse zu berücksichtigen sind. Unter Beachtung der genannten Gründe ist es umso verwunderlicher, dass die CCA dennoch häufig angewandt wird. [8,15]

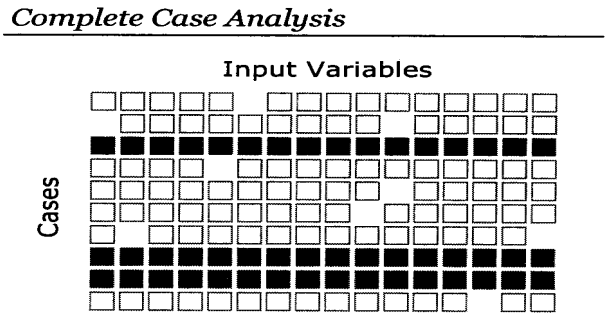

Abbildung 1: Complete Case Analyse

Beispielhaft für die z.T. enorme Reduktion der Fallzahl bei CCA betrachte man einen Datensatz mit 25 Variablen, wobei (nur) 3% der Werte je Variable zufällig fehlen mögen. Unter der Annahme, dass die fehlenden Werte über den Datensatz hinweg gleichverteilt sind, werden demnach $1 - 0.97^{25} = 0.53$ der Beobachtungen, also mehr als die Hälfte, nicht berücksichtigt [4]. Je größer der Anteil an fehlenden Werten und je größer die Anzahl an Variablen ist, desto größer ist die Wahrscheinlichkeit für einen Powerverlust bei statistischen Verfahren.

Die größte Problematik fehlender Werte ist die mögliche Verzerrung der Ergebnisse und die resultierende 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 (so genannter 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 fehlenden Werte systematisch auftreten. Die Aussagekraft der Ergebnisse ist in derartigen Situationen stark eingeschränkt.
Speziell in der Auswertung großer epidemiologischer Datensätze ist die Durchführung einer CCA sehr problematisch. Die epidemiologischen Auswertungsmodelle enthalten in der Regel eine relativ große Anzahl von Einflussgrößen, um die Strukturgleichheit der primär interessierenden Risikogruppen zu sichern. Je mehr Variablen das Modell jedoch 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 reduziert sich daher die Fallzahl entsprechend dem vorab genannten Beispiel, was sich unmittelbar auf die Power auswirkt. [4]
Dieser Beitrag soll eine Übersicht der für die Software SAS zur Verfügung stehenden Möglichkeiten für die Behandlung von fehlenden Werten sein. Die genannten Prozeduren und Makros geben den Stand von Februar 2011 und Erfahrungen der Autoren mit den verschiedenen SAS-Elementen wieder.
Es soll hier aber zunächst ein Überblick zur Diagnostik fehlender Werte gegeben werden, außerdem werden die wichtigsten Ersetzungsstrategien der Single Imputation (SI) und der Multiple Imputation (MI) vorgestellt. Anschließend werden dann die entsprechenden Lösungen der Missing Value Problematik für SAS beschrieben. Den Abschluss bildet eine kurze Zusammenfassung der vorgestellten theoretischen und praktischen Aspekte, zudem werden Empfehlungen zur Nutzung der Software gegeben.

2 Diagnostik und Ersetzungsstrategien

Die Aussagekraft von Studienergebnissen basierend auf einem Datensatz mit (ursprünglich) fehlenden Werten hängt stark von den Ergebnissen der so genannten **Missing Data Diagnostik (MDD)** ab. Ein Teil davon besteht aus der deskriptiven Beschreibung, bei welcher Variablen bzw. Beobachtung wie viele fehlende Werte 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 fehlende Werte analysiert. Das bedeutet, es wird untersucht, ob fehlende Werte vermehrt bei beispielsweise Alten, Männern oder Rauchern etc. auftreten.
Der andere Teil der Missing Data Diagnostik beschreibt die Anordnung der fehlenden Werte im Datensatz, dem so genannten **Missing Data Pattern (MDP)**, und den (möglichen) Gründen für das Auftreten der fehlenden Werte, dem so genannten **Missing Data Mechanism (MDM)**. Letzteres ist wichtig für die Wahl einer geeigneten Ersetzungsmethode.
Bei der Bestimmung des Patterns unterscheidet man im Wesentlichen zwischen zwei Mustern. Fehlen die Werte breit gestreut und mehr oder weniger vereinzelt über den ganzen Datensatz hinweg, so spricht man von einem beliebigen oder auch nicht-monotonen Muster. Im Gegensatz dazu steht ein monotones Muster, bei dem die Daten so angeordnet werden können, dass bis zum Beobachtungsende alle Werte eines Merkmals

ab einem bestimmten Zeitpunkt, zu dem ein Fehlwert das erste Mal aufgetreten ist, fehlen (siehe Abbildung 2).

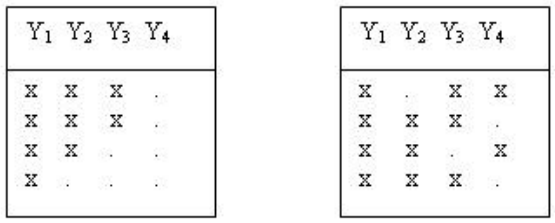

Abbildung 2: Formen des Missing Data Pattern

Die drei verschiedenen Ausprägungen des Missing Data Mechanismus seien hier nur kurz erwähnt, für eine genauere Beschreibung siehe [4] oder auch [11]. Man unterscheidet in drei Kategorien: Missing Completely At Random (MCAR), Missing At Random (MAR) und Missing Not At Random (MNAR). Bei MCAR ist die Drop-out-Wahrscheinlichkeit, also die Wahrscheinlichkeit aus der Studie auszuscheiden, in keinster Weise abhängig von den Werten der Zielgröße. MAR heißt, dass die Drop-out-Wahrscheinlichkeit nur von den beobachteten Werten abhängt, wobei MNAR bedeutet, dass die Wahrscheinlichkeit für einen Drop-out zusätzlich von den fehlenden Werten selbst abhängt. Allerdings ist es nahezu unmöglich, den vorliegenden Mechanismus explizit zu identifizieren und in den realen Daten nachzuweisen.

Missing Data Mechanismus	Beschreibung
Missing Completly At Random	Das Auftreten eines fehlenden Wertes in der Variable Y ist nicht abhängig a) von der Ausprägung der Variable Y selbst oder b) den restlichen Variablen X_1 bis X_n im Datensatz
Missing At Random	Das Auftreten eines fehlenden Wertes in einer Variable Y ist vollständig durch die Ausprägungen der restlichen Variablen X_1 bis X_n erklärbar
Missing Not At Random	Das Auftreten von fehlenden Werten in der Variable Y ist a) von der (unbekannten) Ausprägung der Variable Y abhängig b) nicht durch die Ausprägungen der übrigen Variabeln X_1 bis X_n erklärbar

Abbildung 3: Ausprägungen des Missing Data Mechanismus

Oftmals kann keine strikte Abgrenzung eines bestimmten Mechanismus vorgenommen werden, da es sich um eine Mischform handelt. Zusammen mit dem Pattern bildet dann der Mechanismus den so genannten Missing Data Prozess.
Um mit Standardverfahren der statistischen Datenanalyse arbeiten zu können, bedarf es also im Falle eines unvollständigen Datensatzes einer Ersetzung der fehlenden Werte, wenn man auf eine CCA verzichten möchte. Dafür bieten sich so genannte Single oder Multiple Imputation-Verfahren an. Bei der **Single Imputation (SI)** wird jeder fehlende Wert durch einen plausiblen Wert ersetzt und daher nur ein vervollständigter Datensatz erzeugt. Zum Beispiel führen alle deterministischen Ersetzungsmethoden eine Single Imputation durch. Das sind Methoden, bei denen die Ersetzung eines fehlenden Wertes durch eine einfache, eindeutige Zuordnung erfolgt. Denkbar sind in diesem Zusammenhang Ersetzungen auf Basis des Mittelwertes bzw. des Medians der beobachteten Daten.

Auch so genannte Hot Deck und Cold Deck-Techniken kommen ebenso zum Einsatz wie Regressionsverfahren oder stochastische Ersetzungsmethoden [4,8].
Bei der **Multiple Imputation (MI)** wird ein fehlender Wert durch mehrere ($m > 1$) plausible Werte ersetzt, so dass m vervollständigte Datensätze aus der Ersetzung resultieren. Diese Datensätze werden einzeln mit der gleichen Methode, basierend auf einem jeweils kompletten Datensatz, ausgewertet. Anschließend werden die Ergebnisse dieser Analysen zu gemeinsamen Schätzern und Standardfehlern zusammengefasst. Das Vorgehen der MI ist in der folgenden Abbildung graphisch dargestellt und in Little & Rubin [8] genauer erläutert.

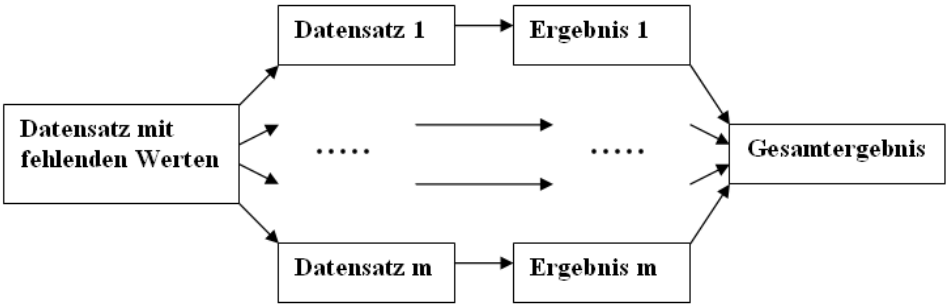

Abbildung 4: Schema der Multiple Imputation

Der entscheidende Vorteil der MI gegenüber den SI-Verfahren ist die korrekte Berücksichtigung des Standardfehlers. Allen SI-Verfahren ist gemein, dass sie von einer zu geringen Varianz ausgehen. Dem entgegen steht die MI, welche die eigentliche Ersetzung als zusätzliche Varianzquelle richtigerweise mit beachtet. Demzufolge werden auch Konfidenzintervalle und p-Werte korrekter berechnet, als das mit einer beliebigen SI-Methode möglich wäre.

3 Der Umgang mit fehlenden Werten in SAS

SAS (Version 9.2) (http://www.sas.com/offices/europe/germany/index.html) ist einer der Marktführer unter den Statistiksoftwarepaketen und wird häufig im Umfeld klinischer Forschung an Universitäten und der pharmazeutischen Industrie eingesetzt. Um die volle Leistungsfähigkeit auszuschöpfen (z.B. in Bezug auf individuelle Ideen zur Ersetzung fehlender Werte), muss die umfangreiche SAS-Syntaxsprache genutzt werden. In den mitgelieferten maus- und menügesteuerten Oberflächen sind die bestehenden Ersetzungsmethoden nicht abrufbar. Die SAS-Makro-Programmierung bietet die Möglichkeit, spezielle Auswertungsroutinen zusätzlich zur Verfügung zu stellen. Dies wird von SAS-Anwendern oft genutzt, so dass neben den offiziellen Prozeduren zur Bearbeitung fehlender Werte auch viele dieser Makros veröffentlicht und verfügbar sind. Die nach unserer Ansicht wichtigsten werden nach den Informationen zu SAS eigenen Lösungen hier vorgestellt.

3.1 Prozeduren für die Missing Data Diagnostik

Die Missing Data Diagnostik spielt insbesondere im Falle longitudinaler Daten eine wichtige Rolle hinsichtlich der Auswahl einer adäquaten Ersetzungsmethode. Neben den in den weiteren Abschnitten noch vorzustellenden speziellen Missing Data Prozeduren und Makros (siehe 3.3 bzw. 3.4), lassen sich erste MDD-Untersuchungen bereits mit SAS Standardprozeduren zur deskriptiven Beschreibung der Daten, wie beispielsweise PROC FREQ, PROC MEANS oder PROC UNIVARIATE, vornehmen. Man bekommt einen ersten Eindruck davon, welches Ausmaß der Anteil fehlender Werte im Datensatz insgesamt annimmt. Zudem sind Auswertungen möglich, die speziell darauf abzielen, den Anteil der Fehlwerte pro Beobachtungseinheit oder Variablen zu untersuchen.
Zudem bietet mittlerweile die Standardprozedur zur Ersetzung fehlender Werte, PROC MI (siehe Abschnitt 3.3), die Möglichkeit, grundlegende Untersuchungen des Missing Data Pattern vorzunehmen. Insbesondere wird dabei auf die Klassifizierung nach monotonem oder beliebigem Fehlwertmuster eingegangen. Detailliertere Analysen des MDP bietet das Makro %MISSDESCRIPTION (siehe Abschnitt 3.4).

3.2 Single Imputation Methoden

Einfache Ersetzungsmethoden wie z. B. die Mittelwertersetzung können in einigen weiteren SAS-Prozeduren schon lange durchgeführt werden. Die entsprechenden Ersetzungsmethoden finden sich u.a. in den Prozeduren PROC STANDARD / PROC STDIZE (Base SAS), PROC PRINQUAL (SAS/STAT) und PROC EXPAND (SAS/ETS). Für speziellere SI-Methoden wie Hot Deck oder Cold Deck-Ersetzung gibt es allerdings bis dato noch keine Möglichkeiten innerhalb des Prozedurumfangs in SAS.

3.3 PROC MI und PROC MIANALYZE

Seit der Version 8.2 hat SAS eine Prozedur zur Durchführung einer Multiple Imputation experimentell eingeführt. Mit dieser Prozedur PROC MI (aktuell in 9.2) lässt sich mittlerweile das Missing Data Pattern ausgeben und die fehlenden Werte mit den Methoden EM-Algorithmus, MCMC-Algorithmus (Data Augmentation), Regressionsersetzung, Logistic Regression Method, Predictive Mean Matching, Propensity Score Method und Discriminant Function Method ersetzen. Dabei wird auf die Vorarbeiten von Allison [1,2,3] sowie auf Rubin [12] und Schafer [13] zurückgegriffen. Die Methoden Logistic Regression Method und Discriminant Function Method eignen sich speziell zur Ersetzung von fehlenden Werten binärer bzw. kategorialer Variablen [4]. Die Beschreibung der Prozedur mit ihren Optionen kann in der Online-Dokumentation von SAS (Version 9.2) unter der Internetdresse
http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#mi_toc.htm nachgelesen werden. Ein Auszug aus der Prozedurbeschreibung erklärt:
The MI procedure performs multiple imputation of missing data ... Multiple imputation does not attempt to estimate each missing value through simulated values. Instead, it draws a random sample of the missing values from its distribution. This process results

in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, confidence intervals with the correct probability coverage.
Die Veröffentlichung von Yuan [16] beschreibt die Möglichkeiten von PROC MI (http://www.sas.com/rnd/app/papers/multipleimputation.pdf), allerdings noch in der Version 9.0.
Zusätzlich zu PROC MI bietet SAS die Prozedur PROC MIANALYZE an, welche die Ergebnisse einer mit PROC MI durchgeführten Multiple Imputation geeignet zusammenführt. Eine ausführliche Beschreibung der Prozedur findet sich ebenfalls in der SAS 9.2 Online-Dokumentation unter der Internetadresse
http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#mianalyze_toc.htm, wo die wesentlichen Möglichkeiten und Optionen zusammengefasst sind. Die Prozedur MIANALYZE kombiniert die mittels Standardprozeduren (z.B. PROC REG) erzeugten Parameterschätzer und zugehörigen Standardfehlern bzw. Kovarianzmatrizen, die auf Basis der vervollständigten Datensätze berechnet wurden, zu einem MI-Schätzer.

3.4 SAS Makros zur Bearbeitung fehlender Werte

Im Folgenden werden die bekanntesten und am häufigsten zitierten SAS-Makros zur Bearbeitung fehlender Werte sowie die zum Teil eigenen Entwicklungen der Autoren aufgelistet und kurz beschrieben.

SAS Makros von Allison

Einer der wichtigsten Autoren zur Methodik und Anwendung zur Bearbeitung fehlender Werte, Allison [1,2,3] stellt seit langem SAS-Makros für Multiple Imputation zur Verfügung. Diese Makros stammen aus der Zeit vor PROC MI und waren u.a. Grundlage bei der Entwicklung der Prozedur. Die folgenden Makros sind von seiner Internetseite http://www.ssc.upenn.edu/~allison/ #Macros abrufbar:

MISS (version 1.05) uses the EM algorithm to estimate the parameters of a multivariate normal distribution when data are missing, and optionally generates multiply imputed data sets using the methods of Schafer.
COMBINE (version 1.03) takes estimates based on multiply imputed data sets and combines them into a single set of estimates and associated statistics.
COMBCHI (version 1.0) takes chi-square statistics from multiply imputed data sets and produces a single p-value.

SAS Makros von Hohl und Muche: %MISSDESCRIPTION und %MISSING

Das Makro %MISSDESCRIPTION dient zur Beschreibung eines vorliegenden Datensatzes speziell in Bezug auf fehlende Werte. Zunächst wird der Anteil an fehlenden Werten je Variable und im gesamten Datensatz, optional die Beobachtungen mit den meisten fehlenden Werten und anschließend das Missing Data Pattern (aus PROC MI) angegeben. Darüber hinaus erfolgt eine (gewöhnliche) Deskription aller Variablen [5].

Mit dem Makro %MISSING können Single – und Multiple Imputation durchgeführt werden. Eine Single Imputation bei stetigen Variablen wird unter Nutzung der SAS-Prozedur PROC STDIZE durchgeführt. Fehlende Werte können hierbei durch den Median oder Mittelwert der vorhandenen Beobachtungen ersetzt werden. Bei kategorialen Variablen ist zudem die Erzeugung einer eigenen Missing-Kategorie möglich [5].
Der Leistungsumfang des Makros %MISSING in Bezug auf die im jeweiligen Fall sinnvollen Ersetzungsmethoden ist in der folgenden Abbildung aufgelistet.

	Merkmalstyp			Missing Pattern		Missing Data Mechanism		durchführbar als	
Methoden	nominal	ordinal	stetig	monoton	beliebig	MCAR	MAR	SI	MI
Mittelwert/Medianersetzung			X	X	X	X		X	
Predictive Mean Matching			X	X			X	X	
Regressionsersetzung			X	X		X	X	X	X
EM Algorithmus			X	X	X	X	X	X	X
MCMC (Data Augmentation)			X	X	X	X	X	X	X
Logistische Regression		X		X		X	X	X	X
Discriminant Function Method	X			X		X	X	X	X

Abbildung 5: Ersetzungsmöglichkeiten in %MISSING (Version 9) [6]

SAS Makropaket von Müller: Analyse und Ersetzung von Missing Data

Verschiedene SAS Makros zur Diagnostik und Ersetzung fehlender Werte werden von Müller auf seiner Internetseite zur Verfügung gestellt, erreichbar unter http://www.joergmmueller.de/AuswahlEntwickelterAnwendungssoftware.htm.

%INDIKAT (2000) Erstellung einer Missing-Data Indikatormatrix
%MISSING (2000) Analyse der Missing-Data nach Personen und Variablen
%KATPAT (2000) Analyse der bivariaten Verteilung von Missing Data
%IMPUTAT (1999) Multivariaten Datenersetzung
%CHECKIMP(1999) Kontrolle der ersetzten Werte

SAS Makro von Little und Yau: Multiple Imputation in Zeitverläufen

Little und Yau haben 1996 eine Methode zur Ersetzung fehlender Werte in der speziellen Auswertungssituation longitudinaler Daten mit Drop-outs (ITT-Analyse in klinischen Studien) vorgeschlagen [9] und dokumentieren die entsprechenden SAS-Programme auf der Internetseite http://www.sph.umich.edu/~rlittle/jobs2.htm

SAS Makro von Houck et al.: Missing Value Ersetzung via LOCF

Unter http://www.nesug.org/proceedings/nesug98/post/p137.pdf wird ein SAS-Programm zur Verfügung gestellt, mit dem eine Ersetzung gemäß der so genannten Last Observation Carried Forward Methode (LOCF) möglich ist. Patricia Houck et al. entwickelten dieses Makro, um Drop-outs im Falle longitudinaler Daten zu ersetzen mit dem letzten beobachteten Wert pro Beobachtungseinheit.

SAS Makro von van Buuren: MISTRESS

MISTRESS ist eine spezielle Methode zur Ersetzung fehlender kategorialer Daten [14]. Das SAS-IML-Makro MISTRESS V. 1.17 steht zur Verfügung unter der URL http://www.stefvanbuuren.nl/mistress/index.html.

SAS Makro von Gregorich: EM_COVAR

Steve Gregorich stellt unter http://lib.stat.cmu.edu/general/em_covar.sas ein SAS-Programm EM_COVAR zur Verfügung, mit dem durch die Anwendung des EM-Algorithmus eine ML-Kovarianzmatrix und ein zugehöriger Mittelwertvektor geschätzt werden kann.

SAS Makro von Raghunathan et al.: IVEWARE

IVEWARE (Imputation and Variance Estimation) ist ein SAS basiertes Softwarepaket (URL http://www.isr.umich.edu/src/smp/ive/). Mit IVEWARE kann eine Multiple Imputation durchgeführt werden:

1. *Perform single or multiple imputations of missing values using the Sequential Regression Imputation Method*
2. *Perform a variety of descriptive and model based analyses accounting for such complex design features as clustering, stratification and weighting.*
3. *Perform multiple imputation analyses for both descriptive and model-based survey statistics.*

4 Zusammenfassung

Es wurde eine Übersicht gegeben über die vorhandenen Möglichkeiten zur Behandlung fehlender Werte in klinischen Datensätzen mit der SAS-Software.

MDD Missing Data Diagnostik | SI Single Imputation | MI Multiple Imputation

Modul	www	MDD	SI	MI
Prozedur PROC MI	http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#mi_toc.htm	▪	▪	▪
Makros von Allison	http://www.ssc.upenn.edu/~allison/ #Macros			▪
MISSING und MISSDESCRIPTION	http://www.uni-ulm.de/med/med-biometrie/forschung/sas-makros-fuer-missing-values.html	▪	▪	▪
Makro von Little und Yau	http://www.sph.umich.edu/~rlittle/jobs2.htm			▪
Makro MISTRESS	http://www.stefvanbuuren.nl/mistress/index.html		▪	
Makro EM_COVAR	http://lib.stat.cmu.edu/general/em_covar.sas		▪	
Makro IVEWARE	http://www.isr.umich.edu/src/smp/ive/		▪	▪

Abbildung 6: Übersicht der Möglichkeiten in SAS

Mit der Angabe verschiedener SAS Makros können die Möglichkeiten der SAS Prozedur PROC MI ergänzt bzw. einfacher aufgerufen werden. Deren jeweilige Handhabung und Anwendbarkeit ist neben der Erfahrung des Nutzers zusätzlich stets von der betreffenden Analysesituation abhängig. Die vorgestellten Makros eignen sich besonders in speziellen Situationen, wie z.B. der Analyse von kategorialen oder longitudinalen Daten, für die es derzeit innerhalb von PROC MI noch keine detailspezifischen Lösungsansätze gibt. Insgesamt jedoch deckt PROC MI als Standardprozedur zur Ersetzung fehlender Werte mittlerweile ein sehr breites Spektrum ab, so dass eine universelle Grundbehandlung fehlender Werte mit PROC MI durchaus empfohlen werden kann. Die Prozedur wird dabei standardmäßig mit dem MCMC-Ansatz (Markov Chain Monte Carlo) ausgeführt, welcher eine Ersetzung unabhängig des vorliegenden Missing Data Patterns ermöglicht.

Literatur

[1] Allison P. (2000) Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods Research. 28: 301-309

[2] Allison P. (2001) Missing Data. Thousand Oaks, CA: Sage

[3] Allison P. (2005) Fixed Effects Regression Methods for Longitudinal Data Using SAS. SAS Publishing

[4] Hohl K. (2008) Umgang mit fehlenden Werten – Ersetzungsmethoden für fehlende Werte kategorialer Variablen in klinischen Datensätzen. Vdm Verlag Dr. Müller, Saarbrücken, Seite 105-116

[5] Hohl K., Muche R., Ring C., Ziegler C. (2005) Fehlende Werte in der (Regressions-)Analyse von Datensätzen: zwei SAS-Makros. 9. KSFE, Shaker Verlag, Aachen, Seiten 105-116

[6] Hohl K., Muche R., Brodrecht K., Ziegler C. (2006) Ersetzung fehlender Werte in SAS: zwei weiterentwickelte SAS-Makros. 10. KSFE, Shaker Verlag, Aachen,

[7] Horton N.J., Lipsitz S.R. (2001) Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Values. The American Statistician. 55(3):244-254.

[8] Little R.J.A., Rubin D.B. (1987) Statistical Analysis with Missing Data. J. Wiley & Sons, New York

[9] Little R.J.A., Yau L. (1996) Intention-to-treat-Analysis for Longitudinal Studies with Drop-outs. Biometrics. 52,4: 1324-1333

[10] Mayer B., Muche R., Hohl K. (2009) Software zur Behandlung und Ersetzung fehlender Werte. GMS Med Inform Biom Epidemiol. 5(2):Doc15

[11] Molenberghs G., Kenward M.G. (2007) Missing Data in Clinical Studies. J. Wiley & Sons, Chichester

[12] Rubin D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York

[13] Schafer J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London

[14] van Buuren S. (1992) Mistress 1.17 documentation. Statistiekreeks 92/07, Leiden: NIPG-TNO

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

[16] Yuan Y.C. (2000) Multiple Imputation for Missing Data: concepts and new development. SAS Institute Inc., Rockville MD.
URL: http://support.sas.com/rnd/app/papers /multipleimputation.pdf
(aufgerufen am 18.02.2011*)*