Fehlende Werte in der (Regressions-) Analyse von Datensätzen: zwei SAS-Makros

Kathrin Hohl*,
Christina Ring*,
*Abt. Biometrie und Med. Dok.,
Universität Ulm
Schwabstraße 13
89075 Ulm
kathrin.hohl@medizin.uni-ulm.de

Rainer Muche*,
Christoph Ziegler‡
‡Boehringer Ingelheim Pharma
GmbH & Co. KG
Birkendorfer Straße 65
88397 Biberach/ Riß
rainer.muche@medizin.uni-ulm.de

Zusammenfassung

Ein generelles Problem bei der Auswertung von Datensätzen ist das Vorkommen von fehlenden Werten. Besonders bei multivariaten Regressionsmodellen kann es bei einer hohen Anzahl an fehlenden Werten zu einer drastischen Fallzahlreduktion kommen, da solche Modelle auf der Analyse einer vollständigen Datenmatrix beruhen. Etliche Ersetzungsmethoden sind in den letzten Jahren in der Literatur vorgeschlagen worden. Um eine geeignete Methode zu bestimmen, ist es notwendig, den Datensatz zuerst rein deskriptiv auf fehlende Werte zu untersuchen und Kenntnisse über die Struktur fehlender Werte zu gewinnen. Mögliche Ersetzungsstrategien sind die Single oder Multiple Imputation. Für die komplexe Auswertung von Datensätzen mit fehlenden Werten wurden von uns zwei SAS-Makros für eine detaillierte Deskription bzw. für die Ersetzung fehlender Werte entwickelt. Die Makros vereinfachen den Umgang mit fehlenden Werten erheblich und werden im Folgenden kurz vorgestellt.

Schlüsselworte: Listwise Deletion, Single Imputation, Multiple Imputation, SAS-Makro, Ersetzung fehlender Werte.

1 Fragestellung

Ein großes Problem bei der Analyse von Datensätzen ergibt sich, wenn viele Messwerte im Datensatz fehlen. Dabei können fehlende Werte aus ganz unterschiedlichen Gründen auftreten. Bei Fragebögen kann beispielsweise eine Antwort fehlen, da die Frage übersehen wurde oder die Person diese Frage nicht beantworten wollte. Bei Longitudinalstudien können bei späteren Follow-ups immer häufiger fehlende Werte auftreten, wenn Personen die Studie verlassen haben.
Die meisten statistischen Analysen in SAS, auch die Regressionsanalysen, basieren auf vollständigen Datensätzen. Das hat beim Auftreten von fehlenden Werten im

Datensatz eine Fallzahlreduktion und eine Verringerung der Power zur Folge. Ferner findet ein Informationsverlust statt und die Repräsentativität der ausgewerteten Stichprobe ist fraglich. Um diese Probleme zu umgehen, wurden in den letzten Jahren von Little und Rubin [4], Rubin [6] und Schafer [7] diverse Ersetzungsmöglichkeiten für fehlende Werte vorgeschlagen. Harrell gibt in seinem Lehrbuch zur Modellierung [3] den Rat, fehlende Werte zu schätzen, statt die gesamte Beobachtung nicht zu berücksichtigen, auch wenn Ersetzungen mit Vorsicht anzuwenden sind. Einige von den Ersetzungsmethoden sind in SAS implementiert und bilden die Grundlage für die von uns entwickelten Makros. Wichtig zur Bestimmung einer optimalen Ersetzungsmethode für einen vorliegenden Datensatz mit fehlenden Werten ist die Kenntnis über die Struktur der fehlenden Werte, d.h. die Gründe für das Auftreten von fehlenden Werten und die Anordnung dieser im Datensatz. Für diese Kenntnisgewinnung ist eine gute Deskription notwendig.

2 Struktur fehlender Werte

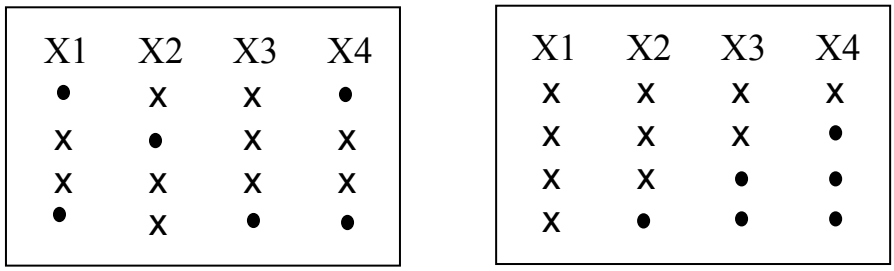
Unterschiedliche Gründe für das Auftreten von fehlenden Werten kann man formal durch die folgenden drei Strukturen beschreiben:

- **MCAR** (Missing completely at random):
  Die fehlenden Werte treten völlig unsystematisch und zufällig auf. Sie hängen weder vom unbeobachteten Wert noch von dem Wert einer anderen Variablen ab.
- **MAR** (Missing at random):
  Das Auftreten von fehlenden Werten einer Variablen ist zwar im stochastischen Sinne zufällig, dennoch aber systematisch und kann durch andere Variablen aus dem Datensatz erklärt werden. Nachdem für diese Variablen adjustiert wurde, treten die fehlenden Werte völlig unsystematisch auf. Diese Struktur wird bei vielen Ersetzungsmethoden vorausgesetzt.
- **MNAR** (Missing not at random):
  Das Auftreten eines fehlenden Wertes einer Variablen hängt nur von dem unbeobachteten Wert dieser Variablen ab.

Ein weiteres Merkmal von fehlenden Werten ist das so genannte Missing Pattern. Es beschreibt, in welcher Anordnung die fehlenden Werte im Datensatz auftreten. Bei einem beliebigen Missing Pattern (siehe Abbildung 1a) treten die Werte völlig unsystematisch im Datensatz auf. Ein solches Muster kann sowohl bei MCAR als auch bei MAR auftreten. Ein monotones Missing Pattern (siehe Abbildung 1b) tritt eher in

Longitudinalstudien auf, bei denen Patienten nach einer gewissen Zeit die Studie verlassen. Dieses Muster kann bei MAR und MNAR auftreten.

Abbildung 1: Missing Pattern.



a) beliebiges Missing Pattern b) monotones Missing Pattern.

3 Ersetzungsmethoden

Aufgrund der Kenntnisse über die Struktur fehlender Werte im Datensatz ist es möglich, eine geeignete Ersetzungsmethode für fehlende Werte zu wählen. Ersetzungsstrategien, die in SAS und in dem entwickelten Makro zur Verfügung stehen, umfassen die Listwise Deletion, Single Imputation und Multiple Imputation. Innerhalb der letzten beiden Strategien kann der Anwender zwischen verschiedenen Ersetzungsmethoden wählen.

3.1 Listwise Deletion (LD)

Die übliche Methode bei statistischen Analysen mit fehlenden Werten umzugehen, ist die Listwise Deletion. Sie ist auch bekannt als Case Deletion oder Complete Case Analysis. Hierbei werden nur vollständige Datensätze in der Auswertung berücksichtigt. Dies kann, vor allem bei einem beliebigen Missing Pattern von einem multivariaten Datensatz, zu einer drastischen Fallzahlreduktion, einem Informationsverlust und einer geringeren Power beim Testen führen.

Wenn die fehlenden Werte die MCAR Struktur haben, d.h. völlig unsystematisch im Datensatz auftreten, ist LD zwar zulässig, da unter diesem Umstand die vollständigen Datensätze eine repräsentative Zufallsstichprobe darstellen, aber die Schätzer können ineffizient sein. Gilt MCAR nicht, so können die Schätzer verzerrt sein. Dennoch ist LD die robusteste Methode für den Umgang mit fehlenden Werten, wenn eine Regressionsanalyse an einem Datensatz durchgeführt wird, bei dem fehlende Werte abhängig vom unbeobachteten Wert auftreten (Allison [1]).

3.2 Single Imputation

Bei der Single Imputation werden die fehlenden Werte einer stetigen Variablen im einfachsten Fall durch eine statistische Kenngröße, wie den Median oder Mittelwert der vorhandenen Beobachtungen, ersetzt. Ist bekannt, dass die fehlenden Werte einer Variablen von anderen Variablen abhängen, kann auch ein Regressionsansatz zur Ersetzung herangezogen werden. Für diskrete Variablen kann man eine zusätzliche Kategorie „Missing“ einführen. Durch den Regressionsansatz ist allerdings wegen der dann notwendigen Dummy-Codierung jeweils ein zusätzlicher Regressionskoeffizient je diskrete Variable im Modell zu schätzen.

Der Vorteil der Single Imputation ist die Einfachheit des Vorgehens. Dieses Vorgehen führt zu lediglich einem vervollständigten Datensatz, der dann für die Analyse zur Verfügung steht. Schafer [8] hat die Nachteile dieser Strategie aufgezeigt: Bei Vorliegen einer beliebigen Struktur fehlender Werte (MCAR, MAR oder MNAR) führt diese Ersetzungsstrategie zu verzerrten Schätzern von vielen Parametern. Die Verzerrung ist dadurch bedingt, dass die Unsicherheit der ersetzten Werte nicht berücksichtigt wird. Die Varianzen werden unterschätzt, die Teststatistiken folglich überschätzt und die Korrelationsstruktur zerstört.

3.3 Multiple Imputation

Die Multiple Imputation berücksichtigt die Unsicherheit fehlender Werte, indem die fehlenden Werte durch mehrere (m) plausible Werte ersetzt werden. Sie setzt allerdings voraus, dass die fehlenden Werte nicht abhängig vom unbeobachteten Wert aufgetreten sind, sondern entweder MAR oder MCAR vorliegt. Für die Multiple Imputation gibt es ebenfalls verschiedene Ansätze zur Ersetzung fehlender Werte. Wie in Abbildung 2 dargestellt, resultieren m parallele Datensätzen aus dieser Strategie. Für jeden dieser Datensätze wird dann die gewünschte Auswertung durchgeführt und anschließend die Ergebnisse zusammengefasst (siehe Abbildung 3). Unabhängig von der gewählten Auswertungsanalyse ergibt sich der endgültige Schätzer eines Parameters meist durch Mittelwertbildung. Die Schätzer für die Varianz beinhalten einen zusätzlichen Korrekturterm [2,5,6,9], um eine mögliche Unterschätzung der Varianz zu verhindern.

Nachdem die geeignete Ersetzungsmethode der Multiplen Imputation gewählt worden ist, muss die Anzahl der zu generierenden parallelen Datensätze festgelegt werden. Die abgebildete Tabelle 1 verdeutlicht, dass die relative Effizienz, d.h. die Effi-

zienz von *m* Ersetzungen im Vergleich zu unendlich vielen, abhängig vom Anteil fehlender Informationen bzgl. dem interessierenden Parameter Q bereits bei 5 bzw. 10 Ersetzungen mit bis zu 99% sehr hoch liegt. In der Literatur werden 5 bis 10 Ersetzungen empfohlen.

Abbildung 2: Prinzip der Multiplen Imputation (Rubin [6])

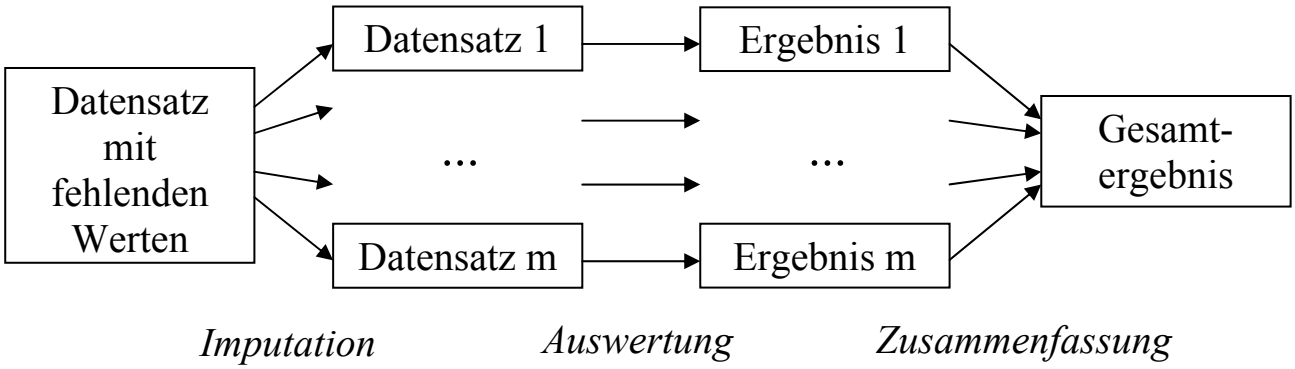


Abbildung 3: Zusammenführung der Ergebnisse (Yuan [9])

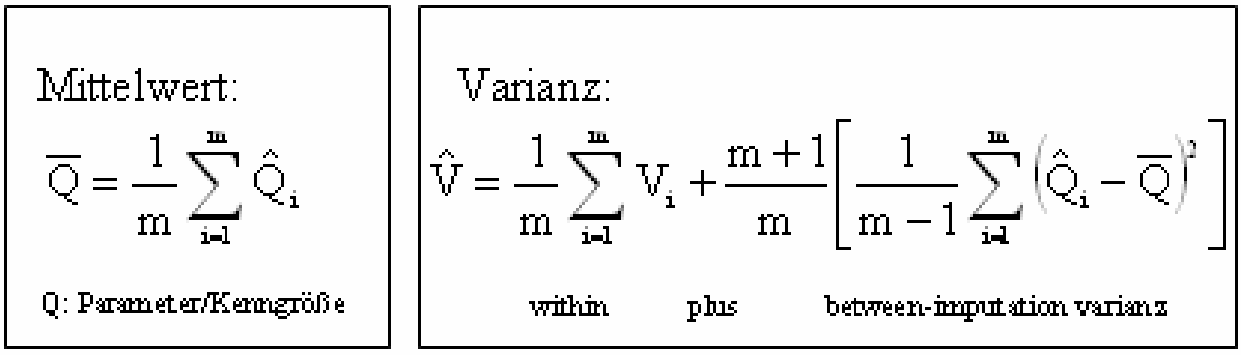


Tabelle 1: Anzahl notwendiger Ersetzungen (Rubin [6])

Relative Effizienz	**Anteil fehlender Informationen bzgl. Q**				
m	**10%**	**20%**	**30%**	**50%**	**70%**
5	0.980	0.962	0.943	0.909	0.877
10	0.990	0.980	0.971	0.952	0.935
20	0.995	0.990	0.985	0.976	0.966

Bei einer Multiplen Imputation kann in dem entwickelten Makro zur Ersetzung fehlender Werte, welches auf der SAS-Prozedur PROC MI basiert, zwischen der Regressionsmethode, der MCMC-Methode und dem EM-Algorithmus gewählt werden.

3.3.1 Regressionsmethode

Die Regressionsmethode setzt eine multivariate Normalverteilung voraus. Dies ist zwar eine recht starke Voraussetzung, aber zum einen betrifft sie nur Variablen mit fehlenden Werten und zum anderen hat Schafer [8] gezeigt, dass selbst bei Abweichungen dieser Annahme gute Ergebnisse mit dieser Methode erhalten werden. Dennoch sollten stetige Variablen, die nicht normalverteilt sind, vor dem Ersetzungsverfahren transformiert werden. Die Regressionsmethode ist besonders bei einem monotonen Missing Pattern sinnvoll, da ein Regressionsmodell für jede Variable mit fehlenden Werten angepasst wird, welches die vorhergehenden Variablen als Covariablen beinhaltet. Basierend auf den so geschätzten Regressionskoeffizienten wird ein neues Regressionsmodell mit Hilfe der posterioren Verteilung der Parameter simuliert und dieses wird benutzt, um die fehlenden Werte von anderen Variablen zu ersetzen. Dieser Prozess wird sequentiell für Variablen mit fehlenden Werten wiederholt, wobei für jede Ersetzung neue Regressionskoeffizienten und Varianzen von der posterioren Verteilung der Parameter simuliert werden.

3.3.2 MCMC-Methode

Die MCMC-Methode setzt ebenfalls eine multivariate Normalverteilung voraus. Diese Methode eignet sich zur Ersetzung von fehlenden Werten mit beliebigem Missing Pattern. Sie besteht aus einem iterativen Zwei-Schritt-Verfahren. Beim Start werden die Informationen der beobachteten Daten oder der EM-Algorithmus zur Startwertsuche von Mittelwertsvektor und Kovarianzmatrix genutzt, welche die prior Verteilung beschreiben. Im ersten Schritt (Imputation-Step) werden zufällig selektierte Werte dieser prior-Verteilung benutzt, um die fehlenden Werte zu ersetzen. Im nächsten Schritt (Posterior-Step) werden, basierend auf dem vervollständigten Datensatz, die Parameter der Verteilung neu geschätzt. Daraus resultiert die posterior-Verteilung. Diese Verteilung wird nun wiederum im Imputation-Step zur Ersetzung der fehlenden Werte im Originaldatensatz zu Grunde gelegt. Das Verfahren wird so lange fortgeführt, bis die Verteilung stationär ist. Mit dieser stationären Verteilung erfolgt dann die endgültige Ersetzung der fehlenden Werte.

3.3.3 EM-Algorithmus

Der EM-Algorithmus ist ebenfalls ein iterativer Prozess, bestehend aus 2 Schritten. Der erste Schritt (Expectation-Step) entspricht unter multivariater Normalverteilung der Regressionsmethode. Im nächsten Schritt (Maximization-Step) erfolgt die Berechnung des Mittelwertvektors und der Kovarianzmatrix, basierend auf dem vervollständigten Datensatz. Der iterative Prozess endet, wenn sich das Modell nicht mehr substantiell verändert bzw. zu Maximum-Likelihood Schätzern konvergiert.

4 Zwei SAS-Makros

Das allgemeine Vorgehen bei der Auswertung von Datensätzen mit fehlenden Werten sollte zunächst mit einer Analyse der Gründe für die fehlenden Werte beginnen. Daraus kann z. T. die Struktur fehlender Werte abgeleitet werden. Als nächstes sollte eine Deskription der fehlenden Werte folgen, um zu erkennen, welche Variablen fehlende Werte haben und welche Beobachtungen wie viele fehlende Werte aufweisen. Da viele in SAS implementierte Ersetzungsmethoden nicht in der Lage sind, fehlende Werte von stark abhängigen Variablen zu ersetzten, sollten ferner die Variablen im Datensatz auf Multikollinearität überprüft werden. Aufgrund der gewonnenen Erkenntnisse mit Hilfe der Deskription kann dann eine „optimale“ Ersetzungsmethode für die fehlenden Werte bestimmt und durchgeführt werden. Nach der Ersetzung erfolgt die Auswertung des neuen Datensatzes. Sollte die vielfach in der Literatur empfohlene Ersetzungsstrategie Multiple Imputation gewählt worden sein, so müssen die *m* Ergebnisse in einem letzten Schritt zu einem Gesamtergebnis zusammengeführt werden. Die zwei SAS-Makros, die wir zur Verfügung stellen, unterstützen die Deskription und die Ersetzung fehlender Werte. Im Folgenden werden die wesentlichen Parameter der Makros angegeben. Eine genaue Definition der Makroparameter ist jeweils im Kopf der Makros enthalten.

4.1 %MISSDESCRIPTION

Das Makro MISSDESCRIPTION führt eine Deskription der fehlenden Werte durch. Es werden die Anzahlen an fehlenden Werten pro Variable und Beobachtung ausgegeben. Da die Berechnung von fehlenden Werten je Beobachtung sehr rechenintensiv ist, werden allerdings nach Benutzerangabe nur die `MOSTEXTREME=x` Beobachtungen mit den häufigsten fehlenden Werten ausgegeben.

Darüber hinaus erfolgt eine (gewöhnliche) Deskription aller Variablen (im Wesentlichen PROC FREQ und PROC UNIVARIATE). Dies kann die Ausgabe zugehöriger Grafiken beinhalten. Ein weiterer Parameter (XVARCAT) bestimmt, ab wie vielen Ausprägungen eine stetige Variable als kategorial betrachtet werden soll. Der Makroaufruf ist wie folgt:

```
%MISSDESCRIPTION(   DATA          =&dset,
                    XVAR          =&stetigvarlist,
                    XVARCAT       =10,
                    CVAR          =&katvarlist,
                    MISS          =0,
                    EXTREME       =0,
                    MOSTEXTREME   = 5,
                    OBS_ID        =&idvar,
                    GRAPHICS      =1,
                    EXT           =tiffp,
                    PATH          ='c:\temp',
                    MACRO_PATH    =&makropfad    )
```

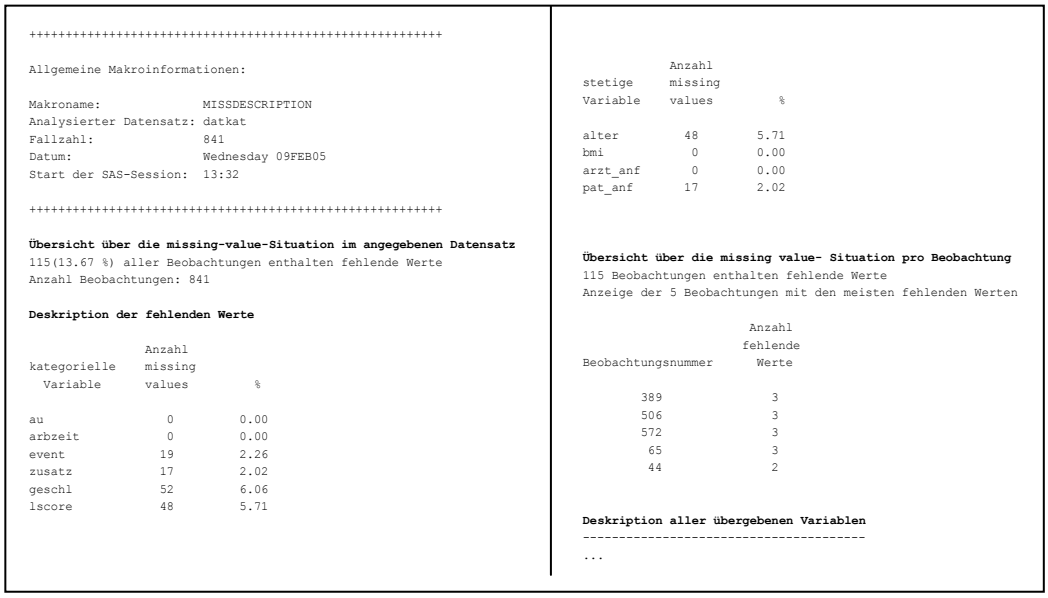
Alle fettgedruckten Parameter sind beim Makroaufruf zwingend erforderlich, für die anderen Variablen gibt es Voreinstellungen. Der Anwender kann im Aufruf entscheiden, ob Beobachtungen mit fehlenden Werten in der üblichen Deskription der Variablen berücksichtigt werden sollen oder nicht (listwise deletion entspricht MISS=0). Ein weiterer wichtiger Parameter ist OBS_ID, welchem eine numerische Variable übergeben werden muss, die die Beobachtungen eindeutig identifiziert. (Eine solche Variable kann beispielsweise in einem Datastep über ID=_N_; erzeugt werden. Hierbei dürfen die Beobachtungen aber nur ein Mal im Datensatz auftreten!)
Ein beispielhafter Output ist in Abbildung 4 dargestellt.

4.2 %MISSING

Das Makro MISSING kann zur Erstellung des Missing Patterns des übergebenen Datensatzes verwendet werden (IMPUTATION_ART=0). Darüber hinaus dient es zur Ersetzung fehlender Werte mittels diverser Ersetzungsstrategien und –methoden.
Eine Single Imputation bei stetigen Variablen wird unter Nutzung der SAS-Prozedur PROC STDIZE durchgeführt. Fehlende Werte können hierbei durch den Median oder Mittelwert (REPLACE_METHOD) der vorhandenen Beobachtungen ersetzt werden. Bei kategoriellen Variablen ist die Erzeugung einer eigenen Missing-Kategorie (MISS_CAT=99) möglich.

Eine Multiple Imputation wird mit Hilfe der Prozedur PROC MI (ab SAS 8.2, [9]) durchgeführt. Der Anwender kann mit dem Parameter MI_METHOD zwischen der Regressionsmethode, der MCMC-Methode und dem EM-Algorithmus wählen.

Abbildung 4: Beispielhafter Output von %MISSDESCRIPTION



++

Allgemeine Makroinformationen:

Makroname: MISSDESCRIPTION
Analysierter Datensatz: datkat
Fallzahl: 841
Datum: Wednesday 09FEB05
Start der SAS-Session: 13:32

++

Übersicht über die missing-value-Situation im angegebenen Datensatz
115(13.67 %) aller Beobachtungen enthalten fehlende Werte
Anzahl Beobachtungen: 841

Deskription der fehlenden Werte

kategorielle Variable	Anzahl missing values	%
au	0	0.00
arbzeit	0	0.00
event	19	2.26
zusatz	17	2.02
geschl	52	6.06
lscore	48	5.71

stetige Variable	Anzahl missing values	%
alter	48	5.71
bmi	0	0.00
arzt_anf	0	0.00
pat_anf	17	2.02

Übersicht über die missing value- Situation pro Beobachtung
115 Beobachtungen enthalten fehlende Werte
Anzeige der 5 Beobachtungen mit den meisten fehlenden Werten

Beobachtungsnummer	Anzahl fehlende Werte
389	3
506	3
572	3
65	3
44	2

Deskription aller übergebenen Variablen

...

Standardeinstellung ist die MCMC-Methode (MI_METHOD=MCMC). Zur Steuerung des Verfahrens kann die Anzahl der zu erzeugenden vollständigen Datensätze (NIMPUTE) und die Rundung (XVAR_ROUND) der stetigen Variablen im Makro angegeben werden.

Der Makroaufruf von %MISSING umfasst eine überschaubare Anzahl an Parametern, die benutzerfreundlich ist, aber dem Anwender dennoch die Möglichkeit gibt, zwischen unterschiedlichen Einstellungen zu wählen.

```
%MISSING(       DATA                =&dset,
                XVAR                =&stetigvarlist,
                CVAR                =&katvarlist,
                OBS_ID              =&idvar,
                IMPUTATION_ART      =0,
                MISS_CAT            =99,
                REPLACE_METHOD      =1,
                IMPUTATION_FILE     =work.simputed,
                MI_METHOD           =MCMC,
                XVAR_ROUND          =&rundunglist,
                NIMPUTE             =5,
                RANDOM              =0,
```

```
                MIMPUTATION_FILE    =work.mimputed,
                MACRO_PATH          =&pfad    )
```

Auch bei diesem Markoaufruf sind alle fettgedruckten Parameter zwingend erforderlich, während es für die anderen Variablen Voreinstellungen gibt.
Unabhängig von der gewählten Ersetzungsstrategie werden, im Gegenteil zur Prozedur PROC MI, die vervollständigten Datensätze zur weiteren Bearbeitung separat gespeichert, was die anschließenden einzelnen Auswertungen erleichtert.
Abbildung 5 zeigt einen beispielhaften Output für ein Missing Pattern. In Abbildung 6 werden die Originalwerte und die 5 mittels der MCMC-Methode ersetzten Werte einer bestimmten Variablen für einige Beobachtungen gegenübergestellt.

Abbidung 5: Beispielhaftes Missing Pattern

Übersicht über die missing value-Struktur im angegebenen Datensatz
115 (13.67 %) aller Beobachtungen enthielten fehlende Werte
Anzahl Beobachtungen: 841
'X' = Wert vorhanden / '.' = missing
Multiple Imputation- Methode: MCMC

missing pattern

Group	AU	ARBZEIT	EVENT	ZUSATZ	GESCHL	LSCORE	ALTER	BMI	ARZT_ ANF	PAT_ ANF	Freq	Percent
1	X	X	X	X	X	X	X	X	X	X	711	84.54
2	X	X	X	X	.	X	X	X	X	X	47	5.59
3	X	X	X	X	X	.	X	X	X	X	44	5.23
4	X	X	X	X	.	.	X	X	X	X	4	0.48
5	X	X	.	X	X	X	.	X	X	X	2	0.24
6	X	X	X	X	.	X	.	X	X	X	1	0.12
7	X	X	.	.	X	X	X	X	X	.	17	2.02

Abbildung 6: Ersetzte Werte mittels MCMC-Methode

		Leistungsscore: 1 - 12				
obs	***original***	***1***	***2***	***3***	***4***	***5***
361	3	3	3	3	3	3
367	•	**11**	**10**	**10**	**11**	**11**
369	•	**4**	**3**	**3**	**4**	**3**
377	•	**7**	**8**	**8**	**7**	**8**
389	•	**3**	**2**	**2**	**2**	**3**
402	•	**5**	**5**	**4**	**4**	**5**
409	•	**1**	**1**	**1**	**2**	**2**
418	2	2	2	2	2	2
421	•	**4**	**4**	**4**	**4**	**3**

5 Diskussion

Die hier vorgestellten SAS-Makros erleichtern die Auswertung von Datensätzen mit fehlenden Werten. Mit diesen Makros ist es ohne Aufwand möglich, die Anzahl an fehlenden Werten je Beobachtung und je Variable zu berechnen, sowie das Missing Pattern des vorliegenden Datensatzes zu erhalten. Diese Informationen sind wichtig für die Wahl der geeigneten Ersetzungsmethode. Das Makro MISSING kann dann zur Durchführung der gewählten Ersetzungsmethode benutzt werden, was die anschließenden Auswertungen erleichtert, indem es die vervollständigten Datensätze einzeln abspeichert.

Die Analysen, die auf diesen so erzeugten Datensätzen beruhen, können im Sinne einer Sensitivitätsanalyse mit der sonst üblicherweise genutzten Listwise Deletion verglichen werden. Bei großen Abweichungen der Ergebnisse sollten sowohl die Ergebnisse der Listwise Deletion als auch die aus den vervollständigten Datensätzen resultierenden Ergebnisse nur vorsichtig interpretiert werden.

Die Makros sind in der SAS Version 8.2 programmiert. Da sich die Syntax von PROC MI, das innerhalb des Makros MISSING aufgerufen wird, von Version 8.2 und Version 9 unterscheidet, ist dieses Makro auf der neuen Version noch nicht einsetzbar. Eine Aktualisierung des Makros wird angestrebt.

Die Makros sind im Rahmen eines Forschungsprojektes im Reha-Forschungsverbund Ulm und der Diplomarbeit von Christoph Ziegler [10] für die Nutzung in logistischen Regressionsmodellen entstanden und können von den Autoren angefordert werden.

Literatur

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

[2] Chantala, K., Suchindran, C. (2003): Multiple imputation for missing data. http://www.cpc.unc.edu/services/computer/presentations/mi_presentation2.pdf (aufgerufen am 16.02.2005)

[3] Harrell, F.E.Jr. (2001): Regression modeling strategies. Springer, New York

[4] Little, R.J.A., Rubin, D.B. (1987): Statistical analysis with missing data. John Wiley and Sons Inc., New York

[5] Muche, R., Ring, Ch. Ziegler, Ch. (2005): Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Shaker-Verlag, Aachen

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

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

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

[9] Yuan, Y.C. (2000): Multiple imputation for missing data: concepts and new development. SAS Institute Inc., Cary NC http://support.sas.com/rnd/app/papers/multipleimputation.pdf (aufgerufen am 16.02.2005)

[10] Ziegler, C. (2003): Ein SAS-Makro-Paket zur Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Diplomarbeit, Med. Dokumentation und Informatik, FH Ulm. Download unter: http://www.uni-ulm.de/uni/fak/medizin/biodok/v2004/prognosemakros.htm