Multiple Imputation in SAS

Stephanie Roll
Institut für Sozialmedizin
Epidemiologie und Gesundheitsökonomie
Charité - Universitätsmedizin Berlin
10098 Berlin
stephanie.roll@charite.de

Zusammenfassung

Mit Hilfe der Multiplen Imputation kann dem Problem unerwünscht fehlender Werte in einem Datensatz begegnet werden. Dazu stehen in SAS mit Version 9 zwei neue Prozeduren zur Verfügung: PROC MI und PROC MIANALYZE (in SAS V8 deren 'experimental versions'). In einem ersten Schritt werden mehrere vollständige Datensätze erzeugt. PROC MI bietet dafür verschiedene Verfahren an. Die vollständigen Datensätze werden dann mit gängigen statistischen Methoden analysiert (z. Bsp. PROC REG). Die daraus resultierenden Ergebnisse werden im Anschluß durch PROC MIANALYZE zu einem Ergebnis zusammengefaßt. Eine Beschreibung der Anwendung, der Syntax und deren Ergebnisse soll hier vorgestellt werden.

Keywords: missing data, multipe imputation.

1. Problem der fehlenden Werte

In vielen Fällen der Datenanalyse liegt ein unvollständiger Datensatz vor, d.h. er enthält fehlende Werte (missing data). Verschiedene Ursachen können dafür verantwortlich sein. Fehlende Werte können beabsichtigt sein (entsprechendes Studiendesign) oder unbeabsichtigt (wenn Angaben nicht vollständig erfaßt wurden) und damit oft unerwünscht. Fehlende Werte führen hauptsächlich zu folgenden Problemen:

- Informationsverlust
- Geringere Fallzahlen für die statistische Analyse
- Bias

Wie schwerwiegend das Problem der fehlenden Werte jeweils ist, bestimmen Anzahl und das Muster, in dem die Werte fehlen (missing pattern). Man unterscheidet drei Arten dieses Musters ([1]):

- MCAR (missing completely at random)
- MAR (missing at random)
- non-ignorable

Es stehen verschiedene Möglichkeiten zu Verfügung, dem Problem der fehlenden Werte zu begegnen. Dazu gibt es je nach Anzahl der fehlenden Werte, missing pattern, Skalenniveau der Variablen mit fehlenden Werten, zu untersuchender Fragestellung, Art der statistischen Analyse, usw. unterschiedliche Verfahren.

Falls nur ein sehr geringer Anteil der Werte fehlt, und man sicher davon ausgehen kann, daß MCAR vorliegt, kann man Beobachtungen mit unvollständigen Werten von der Analyse ausschließen. Es kommt dann weder zu einem Fallzahlproblem noch werden die Ergebnisse verzerrt. Dies ist, auch wenn die Anforderungen dafür nicht erfüllt sind, die standardmäßige und wohl am häufigsten angewandte Methode.

Möchte man fehlende Werte ersetzen, gibt es Verfahren, die sich unter der Bezeichnung 'single imputation' zusammenfassen lassen. Dazu gehören z. B. Mittelwertersetzung, Last Value Carried Forward (LVCF), Hot Deck-Verfahren. Diese Verfahren haben den Vorteil, daß nach Ersetzen der Werte ein vollständiger Datensatz vorliegt. Ein Nachteil jedoch ist, daß die Unsicherheit über die ersetzten Werte nicht berücksichtigt wird. Beobachtete Werte und ersetzte Werte werden gleich behandelt. Will man diese Unsicherheit miteinbeziehen, bietet sich die Methode der Multiplen Imputation (MI) an: hierbei wird jeder fehlende Wert mehrmals ersetzt ([2]).

2. Multiple Imputation

Das Multiple Imputations-Verfahren erfolgt in drei Schritten:

1. *m* Ersetzungen (Imputationen) für jeden fehlenden Wert.
   Dies resultiert in *m* vollständigen Datensätzen.
2. Identische, separate statistische Analyse dieser *m* Datensätze.
   Dies resultiert in m Ergebnissen.
3. Kombination der *m* Ergebnisse zu einem Endergebnis.

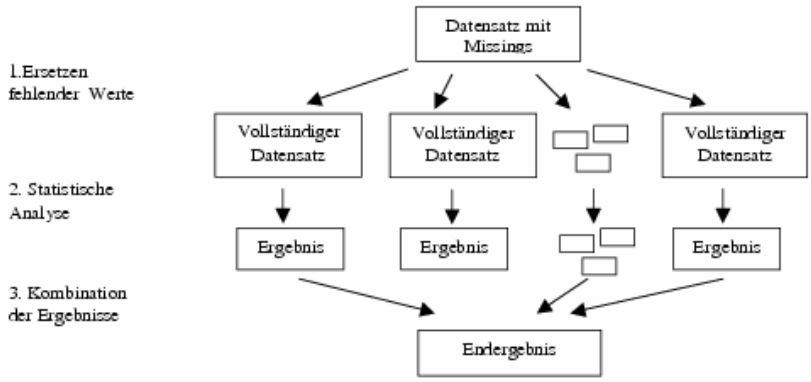
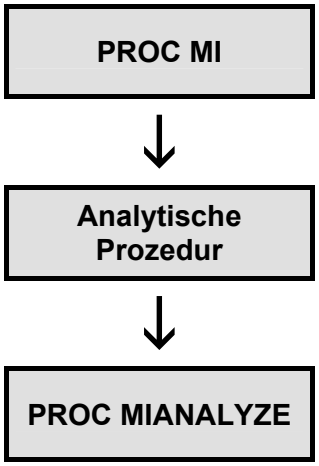


Abbildung 1: Grafische Darstellung des MI-Verfahrens

Dabei genügt i. d. R. im ersten Schritt eine relativ geringe Anzahl an Imputationen, z.B. $m = 5$ ([3]). In diesem Fall würde jeder fehlende Wert durch 5 plausible Werte ersetzt. Unabhängig welches Ersetzungsverfahren dafür gewählt wird, erfolgt die Kombination der Ergebnisse in Schritt 3 üblicherweise auf die gleiche Art (z.B. arithmetisches Mittel).

3. Multiple Imputation in SAS

Für die Durchführung der Multiplen Imputation in SAS stehen zwei Prozeduren zu Verfügung: PROC MI und PROC MIANALYZE. Sie sind als 'experimental versions' ab SAS V8 und offiziell ab SAS V9 implementiert (es gibt jedoch Unterschiede in der Syntax zwischen den Versionen in V8 und V9). Die Anwendung von PROC MI und PROC MIANALYZE erfolgt analog der oben beschriebenen Struktur des MI-Verfahrens:



Erzeugung mehrerer Datensätze mit unterschiedlich ersetzten Werten

Analyse der vollständigen Datensätze mittels üblicher statistischer Verfahren

Kombination der Ergebnisse der statistischen Analyse zu einem Endergebnis

Im Folgenden wird die Syntax von PROC MI und PROC MIANALYZE beschrieben (für weitere Details siehe [4] und [5], bzw. [6]). Für die Anwendung dieser Prozeduren ist die Struktur der fehlenden Werte von Bedeutung. Man spricht von einem 'monotonen missing pattern', wenn folgendes gilt: für einen Datensatz mit den Variablen $X_1, X_2, ..., X_n$ (in dieser Reihenfolge) impliziert das Fehlen des Wertes für Variable X_j einer Beobachtung, daß die Werte aller nachfolgenden Variablen X_k, $k>j$, für diese Beobachtung ebenfalls fehlen. Anders ausgedrückt: wenn der Wert der Variablen X_j beobachtet ist, so müssen auch die Werte aller vorherigen Variablen X_i, $i < j$, beobachtet sein. Man beachte, daß für diese Definition die Reihenfolge der Variablen im Datensatz von Bedeutung ist.

Beispiel: Datensatz mit monotonem missing pattern

Obs	X_1	X_2	X_3	X_4
1	x	x	x	x
2	x	.	.	.
3	x	x	x	.
4	.	.	.	.
5	x	x	.	.

'x': Wert beobachtet,' . ': fehlender Wert

Würden in diesem Datensatz die Variablen X_3 und X_4 vertauscht, läge kein monotones missing pattern mehr vor; die Voraussetzung dafür wäre für die dritte Beobachtung verletzt.

Mit der Prozedur PROC MI werden fehlende Werte in einem Datensatz ersetzt (Schritt 1 der Multiplen Imputation).

```
proc mi data = data_missings
        out = mi_out
        seed = 3000 nimpute = 5
        minimum = 0 maximum = 100;
            var x1-x5;
            monotone method = regression;
run;
```

Optionen des proc mi-Statements (Auswahl)	
`data =`	benennt den Input-Datensatz, der fehlende Werte enthält
`out =`	benennt Output-Datensatz; die fehlenden Werte sind ersetzt
`nimpute =`	Anzahl der Imputationen
`seed =`	Seed für den Start des Zufallszahlengenerators
`minimum =`	Minimal möglicher Wert für ersetzte Werte
`maximum =`	Maximal möglicher Wert für ersetzte Werte
Weitere Statements	
`var ;`	Auflistung der verwendeten Variablen
`monotone ;`	spezifiziert die Ersetzungsmethode

Als Ersetzungsmethode wurde hier die Regressionsmethode für Daten mit monotonem missing pattern gewählt
("`monotone method = regression;`").
Je nach Skalierung und Verteilung der Variablen und nach missing pattern stehen verschiedene Ersetzungsverfahren zu Verfügung ([4], [6]).

Der Output von PROC MI enthält Informationen zum missing pattern sowie zu Parametern und Varianzen der Multiplen Imputation. Vor allem wird jedoch ein Datensatz erzeugt, in dem die fehlenden Werte ersetzt sind. Je nach Anzahl der Imputationen (hier *m*=5) wird eine Datenmatrix erzeugt, die den Ursprungsdatensatz *m* mal untereinander enthält (gekennzeichnet durch eine neue Variable *_imputation_* mit den Werten 1, 2, ..., *m*). In jedem der *m* Blöcke sind die fehlenden Werte (unterschiedlich) ersetzt. Dieser Output-Datensatz wird nun zur statistischen Analyse verwendet (Schritt 2 des MI-Verfahrens), zum Beispiel für eine lineare Regression.

```
proc reg data = mi_out
         outest = reg_out covout;
            model x1 = x2-x5;
            by _imputation_;
run;
```

Um jeden der *m* Blöcke getrennt auszuwerten, wird das Statement "`by _imputation_;`" verwendet.
Der übliche PROC REG-Output enthält Varianzanalyse-Tabellen sowie Parameterschätzer für jeden der *m* Datenblöcke. Durch die Option "`outest = reg_out covout`" wird zusätzlich ein Datensatz erzeugt, der die Parameter- und Kovarianzschätzer der linearen Regression enthält.

Dieser Datensatz wird nun für den dritten MI-Schritt (Kombination der Ergebnisse) genutzt. Die Prozedur PROC MIANALYZE faßt die Ergebnisse der statistischen Analyse zusammen. Als Input-Datensatz verwendet sie den vorherigen Output-Datensatz (Parameter- und Kovarianzschätzer der linearen Regression).

```
proc mianalyze data = reg_out;
            var intercept x2-x5;
run;
```

Im `var`-Statement werden die Variablen angegeben, deren Schätzer zusammengefaßt werden sollen (Achtung: Die Syntax von PROC MIANALYZE ist in SAS V9 unterschiedlich. Statt "`var ...;`" muß in V9 "`modeleffects ...;`" stehen. [5], [6]).
Der Output von PROC MIANALYZE enthält schließlich neben Angaben zu Varianzschätzern des Imputationsverfahrens die endgültigen Parameterschätzer mit Standardfehler und Konfidenzintervall.

4. Fazit

Die beiden neuen SAS-Prozeduren zur Multiplen Imputation sind relativ einfach in ihrer Programmierung. Eine wesentlich komplexere Aufgabe ist die Überlegung, nach welchem Muster fehlende Werte auftreten, für welche der fehlenden Werte welches Imputationsverfahren geeignet ist um sie zu ersetzen, welche Abhängigkeitsstrukturen in den Daten bestehen (d.h. welche Variablen zur Ersetzung der fehlenden Werte verwendet werden können), usw. Das heißt trotz der neuen MI-Prozeduren wäre es wohl nach wie vor am besten, von Beginn an darauf zu achten, fehlende Werte zu vermeiden.

Literatur

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

[2] Rubin, D. (1978). Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse, Imputation and Editing of Faulty or Missing Survey Data, U.S. Department of Commerce, pp 1-23.

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

[4] SAS V8-Dokumentation zu PROC MI, http://support.sas.com/rnd/app/papers/miv802.pdf

[5] SAS V8-Dokumentation zu PROC MIANALYZE, http://support.sas.com/rnd/app/papers/mianalyzev802.pdf

[6] SAS V9 OnlineDoc, http://v9doc.sas.com/sasdoc/