Ersetzung fehlender Werte in SAS: zwei weiterentwickelte SAS®-Makros

Kathrin Hohl[1], Kati Brodrecht[1], Christoph Ziegler[2], Rainer Muche[1]

[1]Universität Ulm, Abteilung Biometrie und Medizinische Dokumentation, Ulm
kathrin.hohl@medizin.uni-ulm.de, kati.brodrecht@uni-ulm.de, rainer.muche@medizin.uni-ulm.de

[2]Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach/ Riß
christoph.ziegler@bc.boehringer-ingelheim.com

Inhalt

Das Auftreten von fehlenden Werten führt besonders bei multivariaten Analysen, die auf einer vollständigen Datenmatrix beruhen, zu Problemen wie Fallzahlreduktion und Verringerung der Power. Zudem findet ein Informationsverlust statt und die Repräsentativität der ausgewerteten Stichprobe ist fraglich.
Auf der 9. KSFE haben die Autoren zwei in SAS®-Version 8.2 entwickelte Makros für den Umgang mit fehlenden Werten in Datensätzen vorgestellt. Diese Makros stellen dem Benutzer zwei variations- und umfangreiche Module für die Deskription und die Ersetzung fehlender Werte zur Verfügung. Mehrere Ersetzungsmethoden für Single und Multiple Imputation sind möglich. Die Multiple Imputation basiert im Wesentlichen auf der Prozedur PROC MI, welche in SAS®-Version 8.2 lediglich als experimentelle Prozedur vorliegt. Ab der SAS®-Version 9 steht diese Prozedur offiziell zur Verfügung und ist im Vergleich zu der Version 8.2 um einige Ersetzungsmethoden erweitert worden. Aus diesem Anlass wurden die vorhandenen Makros weiterentwickelt und optimiert, so dass sie nun auf Version 9 lauffähig sind und ebenfalls weitere Ersetzungsmethoden dem Benutzer zur Verfügung stellen.
Abhängig vom vorliegenden Missing Pattern des Datensatzes konnten in SAS®-Version 8.2 die MCMC-Methode, der EM-Algorithmus oder die Regressionsmethode zur Ersetzung von fehlenden Werten angewendet werden. Diese Methoden setzten alle eine multivariate Normalverteilung voraus und sind daher nur unter Vorbehalt für die Ersetzung fehlender Werte von kategorialen Variablen einsetzbar. In Version 9 werden speziell für kategoriale Variablen und bei Vorliegen eines monotonen Missing Patterns die Ersetzungsmethoden: Discriminant Function und logistische Regression zur Verfügung gestellt. Ferner besteht die Möglichkeit fehlende Werte stetiger Variablen mit der Predictive Mean Matching Methode zu ersetzten.
Trotz der neu zur Verfügung gestellten Ersetzungsmethoden muss beachtet werden, dass die beste Lösung des Problems mit fehlenden Werten *keine* fehlenden Werte sind!

Literatur

[Allison, 2001] **Allison, Paul D.** (2001): Missing Data. Thousand Oaks, CA: Sage Publications

[Allison, 2005] **Allison, Paul D.** (2005): Imputation of Categorical Variables with PROC MI. SAS User Group International URL: http://www2.sas.com/proceedings/sugi30/113-30.pdf (aufgerufen am 13.07.2005)

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

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

[SAS®, 2002/03] **SAS® Institute Inc.** (2002/03): SAS® OnlineDoc® 9.1. Cary, NC: SAS® Institute Inc. URL: http://support.sas.com/91doc/docMainpage.jsp (aufgerufen am 08.07.2005)

[Yuan, 2000] **Yuan, Y.C.** (2000): Multiple imputation for missing data: concepts and new development. SAS Institute Inc., Cary NC. URL:

http://support.sas.com/rnd/app/papers/multipleimputation.pdf (aufgerufen am 13.07.2005)

Postanschrift: Universität Ulm
Biometrie und Medizinische Dokumentation
Schwabstraße 13
89070 Ulm

Form des Beitrags: Poster

Zielpublikum: Kenntnisse in SAS® und Interesse an dem Umgang mit fehlenden Werten in Datensätzen