Ein SAS-Makro Paket für die Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression

Rainer Muche, Christina Ring
Abteilung Biometrie und Med. Dokumentation,
Universität Ulm
Schwabstraße 13
89070 Ulm
rainer.muche@medizin.uni-ulm.de

Christoph Ziegler
Boehringer Ingelheim Pharma GmbH & Co KG
Dep. Med. Data Services

88397 Biberach / Riss

Zusammenfassung

Prognosen zum Krankheitsverlauf oder zum Schweregrad multipler Schädigungen bestimmen die medizinischen Therapie- und Diagnostikentscheidungen direkt oder indirekt. Neben der subjektiven Einschätzung des Arztes können mathematische Modelle für Prognosezwecke entwickelt und validiert werden. Prognosemodelle werden vielfach als verallgemeinerte lineare Regressionsmodelle formuliert. In der Praxis ist die betrachtete Zielgröße häufig dichotom, so dass multiple logistische Regressionsmodelle zum Einsatz kommen. Im Folgenden wird ein Vorgehen für eine Modellierung basierend auf logistischen Regressionsmodellen beschrieben. Die Untersuchung der Prognosemöglichkeiten erfolgt in drei Schritten: Modellentwicklung, Bestimmung der Prognosegüte und Modellvalidierung.
Die für diese Modellierung zugehörenden speziellen 14 SAS-Makros sowie einige zusätzliche Untermakros sind zu einem Paket zusammengefasst, welches im Internet zu bekommen ist. Eine Kurzbeschreibung der Makros ist im Proceedingsband der letzten KSFE-Tagung zu finden, detailliertere Angaben finden sich in einem zugehörigen Buch sowie im Kopf des jeweiligen Makros.

Schlüsselworte: Prognosemodell, Logistische Regression, Modellvalidierung, SAS-Makro

1 Einleitung

„Prognose ist eine Vorhersage über den zukünftigen Verlauf einer Krankheit nach ihrem Beginn“ [2]. Nach dieser Definition können Prognosen in der Medizin die Therapieentscheidungen direkt oder indirekt mitbestimmen und sollten daher so zuverlässig wie möglich erstellt werden. Neben der subjektiven ärztlichen Einschät-

zung zum zukünftigen Krankheitsverlauf können Prognosen auch auf Grundlage entwickelter mathematischer Modelle gegeben werden. Dabei handelt es sich oft um verallgemeinerte lineare Regressionsmodelle, wie z.B. das multiple logistische Regressionsmodell, das im Fall dichotomer Zielgrößen, wie sie häufig im klinischen Alltag beobachtet werden, zur Anwendung kommt.
Im Folgenden wird *eine* Vorgehensweise zur Prognosemodellierung auf Basis der logistischen Regression vorgestellt, deren Umsetzung in der Praxis durch neu entwickelte SAS-Makros bzw. den sinnvollen Einbau bereits vorhandener SAS-Makros unterstützt wird.

2 Logistische Regression

Die logistische Regression ist seit langem das Standardverfahren für die Analyse binärer Zielgrößen [5]. Die Modellgleichung zur Schätzung, ob ein Ereignis eintrifft (Y=1), gegeben einige Einflussgrößen $X_1, X_2, \ldots, X_k$, wird modelliert als:

$$P\left(Y=1 \middle| X_j = x_j\right) = \frac{1}{1+\exp\left(-\left(\alpha + \sum_j \beta_j x_j\right)\right)} \qquad j = 1,\ldots,k$$

Dabei werden die Regressionskoeffizienten $ß_i$ mit der Maximum-Likelihood Methode geschätzt. In SAS kann die logistische Regression mit mehreren Prozeduren umgesetzt werden: PROC LOGISTIC, PROC CATMOD, PROC GENMOD, PROC PROBIT. Die für die Umsetzung in den SAS-Makros am besten geeignete Lösung ist die über die Prozedur PROC LOGISTIC [1], wobei in der Situation der Datenseparation alternativ das FL-Makro von Heinze zur Schätzung genutzt werden kann [4]. Abbildung 1 zeigt den allgemeinen Aufruf der Prozedur mit den für die Programmierung notwendigen Optionen.

```
PROC LOGISTIC DATA= OUTEST= INEST= ;
     CLASS var1 (PARAM=  REF= );
     MODEL ziel (EVENT= ) = var1 var2
           / CLODDS= RSQUARE LACKFIT
             SELECTION= OUTROC=  ;
     OUTPUT OUT=  PRED=  RESCHI=  DIFCHISQ= ;
RUN;
```

Abb. 1: Aufruf der logistischen Regression mit PROC LOGISTIC

3 Umsetzung der Prognosemodellierung

Den Vorschlag für eine Vorgehensweise [3, 6] und sukzessive Abarbeitung der Prognosemodellierung in den drei Schritten (1) Modellentwicklung, (2) Prognosegüte und (3) Modellvalidierung zeigt Abbildung 2. Im Abschnitt 5 werden nach einer kurzen Beschreibung des prinzipiellen Aufrufs der Makros und der technischen Voraussetzungen jeweils einige kurze Hinweise zu den einzelnen Auswertungsschritten gegeben.

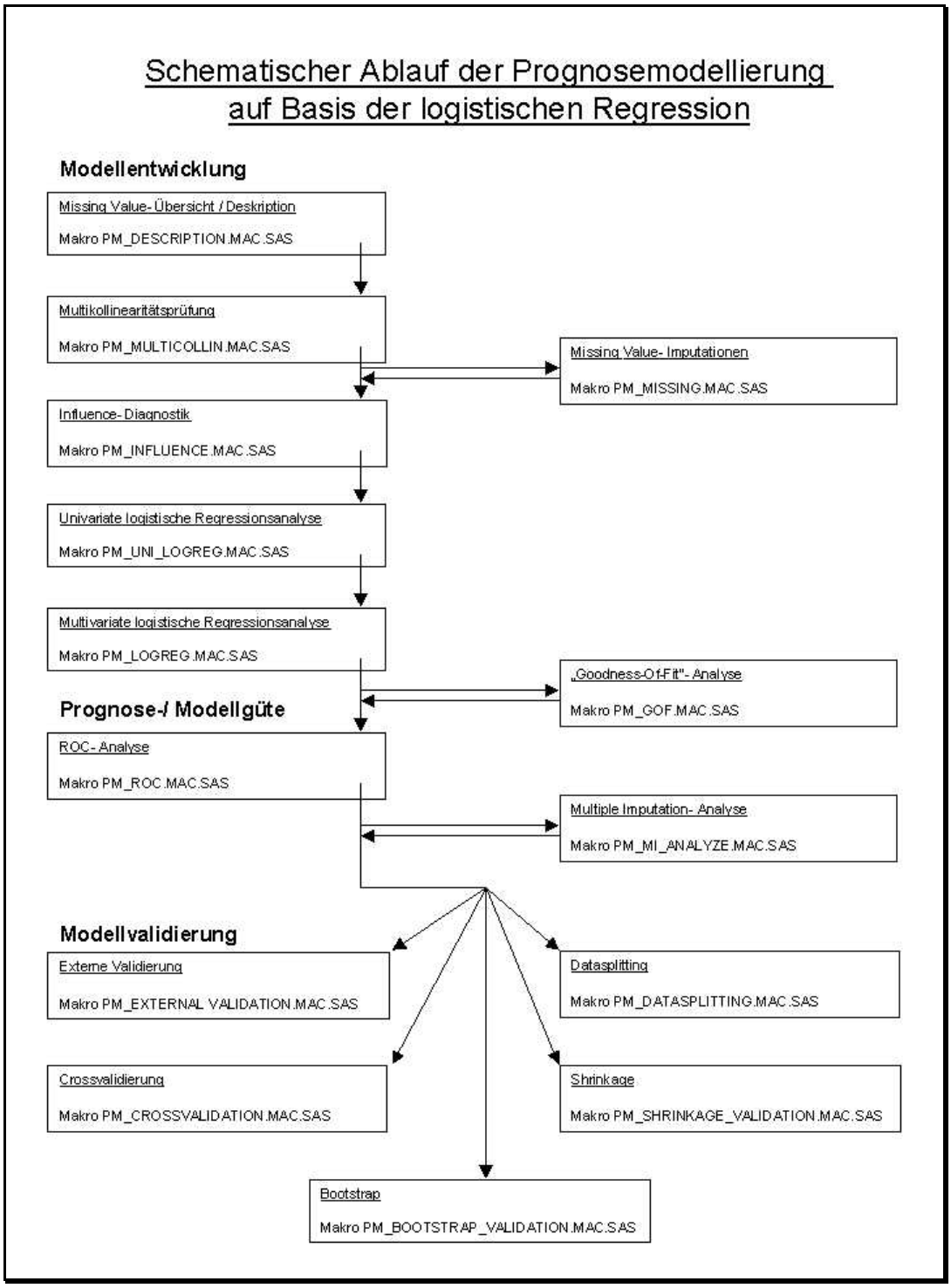

Abb. 2: Ablauf der Prognosemodellierung auf Basis der SAS-Makros [6]

4 Makro-Aufruf und technische Voraussetzungen

Der Aufruf aller Makros ist ähnlich gestaltet. In Abbildung 3 wird der prinzipielle Aufruf der wichtigsten Parameter dargestellt. Mit den Parametern wird der auszuwertende Datensatz (data=), die Zielgröße mit interessierendem Event (resp_var=, event=) sowie die Einflussgrößen (diskret: cvar=, stetig: xvar=) angegeben. Mit dem Parameter miss= können für Complete-Case-Analysen Beobachtungen mit fehlenden Werten aus der Analyse ausgeschlossen werden.

```
%MACRO macroname (  data        =,
                    resp_var    =,
                    event       =1,
                    xvar        =,
                    cvar        =,
                    ref         =,
                    miss        =0,
                    ...
                    weitere spezifische Parameter,
                    ...
                    macro_path  =,
                    );
%MEND macroname;
```

Abb. 3: Prinzipieller Aufruf der SAS-Makros

Zur Nutzung sind einige Hard- und Softwarevoraussetzungen einzuhalten. Folgende Mindestanforderungen werden an das Computersystem gestellt:

- SAS-Installation ab SAS 8.2
- SAS-Module BASE, STAT, GRAPH, IML
- Hardwarevoraussetzungen zur Nutzung von SAS 8.2 (Empfehlung: RAM 512 Mb, Prozessor > 1 Ghz)

Die SAS-Makros nutzen viele externe Programme, u.a. umfangreiche Prüfprogramme. Das gesamte Makropaket besteht aus etwa 100 Programmen und Dateien. Deshalb sind zur Nutzung der Makros einige Voraussetzungen vorgegeben:

- das gesamte Makropaket steht in einem Ordner (Aufruf über macro_path=),
- die auszuwertenden Variablen müssen numerisch sein,
- die Variablen sollten möglichst numerisch formatiert sein,
- es wird eine Variable verlangt, die die Beobachtungen eindeutig identifiziert.

5 Übersicht über die SAS-Makros

In diesem Beitrag werden die Makros nur aufgezählt. Eine Kurzbeschreibung der Makros kann [7] entnommen werden. Eine genaue und detaillierte Beschreibung findet sich in [6] bzw. [9] und im Kopf des jeweiligen Makros. Die folgende Kurzbeschreibung ist in die drei Oberbereiche der Prognosemodellierung: Modellentwicklung, Prognosegüte und Modellvalidierung aufgeteilt.

Die Makros können neben weiteren Informationen von der Internetseite:
http://www.uni-ulm.de/uni/fak/medizin/biodok/v2004/prognosemakros.htm
heruntergeladen werden.

5.1 Makros zur Modellentwicklung

Bei der Modellentwicklung sind verschiedene Untersuchungen des Datensatzes vor der eigentlichen Modellierung notwendig. Dazu gehört die Untersuchung der Variablen (Deskription) und deren Beziehung untereinander (Multikollinearität) genauso wie die Analyse des Einflusses der einzelnen Beobachtungen. Ein spezielles Problem bei der Regressionsanalyse sind fehlende Werte, die einen enormen Einfluss auf das Ergebnis haben können. Die folgenden Makros helfen, diese Untersuchungen durchzuführen und das logistische Regressionsmodell anzupassen:

- PM_DESCRIPTION.MAC.SAS
- PM_MULTICOLLIN.MAC.SAS
- PM_MISSING.MAC.SAS / PM_MI_ANALYZE.MAC.SAS
- PM_INFLUENCE.MAC.SAS
- PM_UNI_LOGREG.MAC.SAS
- PM_LOGREG.MAC.SAS
- PM_GOF.MAC.SAS

5.2 Makro zur Überprüfung der Prognosegüte

Bei der Überprüfung der Prognosegüte stellt sich die Frage**: „Wie gut kann der Outcome des Patienten vorhergesagt werden?“** Die Überprüfung der Prognosegüte geschieht anhand einer Reklassifikation. Dabei werden die Daten der Patienten in die Modellgleichung eingesetzt und so für jeden Patienten die Wahrscheinlichkeit für das Eintreten des Outcome geschätzt. Durch einen Vergleich mit den beobachteten Werten lässt sich die Übereinstimmung untersuchen. Dabei können nach Wahl eines Grenzwertes (Cutpoint) zur Einteilung der Wahrscheinlichkeiten in „groß“ bzw. „klein“ die Kenngrößen wie Sensitivität, Spezifität, prädiktive Werte, Youden-Index etc. bestimmt werden. Daneben lassen sich als Maße einer globalen Prognosegüte (unabhängig von einem Cutpoint) angeben: AUC, Somer´s D, Emax, Brier Score etc.. Diese Kenngrößen werden im Rahmen einer ROC-Analyse erzeugt:

- PM_ROC.MAC.SAS

5.3 Makros zur Modellvalidierung

Nach der Untersuchung der Prognosegüte könnte entschieden werden, ob das Prognosemodell eine ausreichende Güte mit geringen Fehlerraten besitzt, um in der Praxis eingesetzt zu werden. Allerdings ist die Frage: **„Wie gut ist die Prognosegüte für spätere unabhängige Beobachtungen?“** bis hierhin noch nicht beantwortet.
Das Problem besteht darin, dass die Prognosegüte nach der Reklassifikation anhand derselben Patientendaten ermittelt wird, die auch zur optimalen Schätzung der Regressionskoeffizienten zur Verfügung standen. Somit ist von einem Bias in Richtung zu optimistischer Prognosegüten nach der ROC-Analyse auszugehen. Zur Untersuchung dieses Bias sollte eine Modellvalidierung erfolgen. Dafür stehen verschiedenen Verfahren zur Verfügung, die in den folgenden fünf Makros umgesetzt wurden. In der Literatur wird neben der externen Validierung die Bootstrap-Methode favorisiert [8]. Als Output werden in den Makros jeweils die Cutpoint-abhängigen und –unabhängigen Prognosegütemaße der ROC-Analyse vor und nach der Validierung sowie die absolute und relative Veränderung ausgegeben.

- PM_EXTERNAL_VALIDATION.MAC.SAS
- PM_DATASPLITTING.MAC.SAS
- PM_CROSSVALIDATION.MAC.SAS
- PM_BOOTSTRAP_VALIDATION.MAC.SAS
- PM_SHRINKAGE.MAC.SAS

6 Zusammenfassung

Die wichtigsten Probleme der Modellbildung werden in der Literatur folgendermaßen zusammengefasst: nicht spezifizierte Definition der Variablen, Multikollinearität, Nichtberücksichtigung einflussreicher Beobachtungen, nicht erfüllte Modellvoraussetzungen, Nichtlinearität des Zusammenhanges, Überanpassung, unspezifizierte Variablenselektion, keine Wechselwirkungsprüfung sowie fehlende Modellvalidierung.

Die vorgestellte Strategie zur Modellentwicklung und Modellvalidierung anhand eines SAS-Makro-Paketes berücksichtigt all diese Auswertungsprobleme und schafft damit Voraussetzungen, in Zukunft geeignete Prognosemodelle auf Basis der logistischen Regression erstellen und deren praktischen Nutzen genauer ermitteln zu können. Damit tragen die vorgestellten Makros zur Verbesserung der biometrischen Praxis zur Bestimmung zuverlässigerer Prognosen bei.

Literatur

[1] Allison P.D. (1999) Logistic Regression using the SAS System. SAS Institute Books By Users, Cary NC

[2] Fletcher R.M., Fletcher S.W., Wagner E.H. (1999). Klinische Epidemiologie. Ullstein Medical Verlag, Wiesbaden

[3] Harrell F.E. Jr. (2001) Regression Modeling Strategies. Springer Verlag, New York

[4] Heinze G, Schemper M. (2002) A solution to the problem of separation in logistic regression. Stat. Med. 21, 2409-2419

[5] Hosmer D.W., Lemeshow S. (2000) Applied Logistic Regression (2nd Edition). John Wiley & Sons, New York

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

[7] Muche R., Ring C, Ziegler C. (2004) Ein SAS-Makro Paket für die Entwicklung und Validierung eines logistischen Regressionsmodells. In: Beyer D., Ortseifen, C. (Hrsg.): SAS in Hochschule und Wirtschaft. Proceedings der 8. Konferenz für SAS-Anwender in Forschung und Entwicklung (KSFE), Shaker Verlag, Aachen, 173-186

[8] Steyerberg E.W., Harrell F.E.Jr., Borsboom G.J.J.M. et al. (2001) Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol., 54, 774-781

[9] Ziegler Ch. (2003) Ein SAS-Makro-Paket zur Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression. Diplomarbeit FH Ulm