Ein SAS-Makro Paket für die Entwicklung und Validierung eines logistischen Regressionsmodells

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

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 werden SAS-Makros beschrieben, die für eine Modellierung basierend auf logistischen Regressionsmodellen entwickelt wurden. Die Untersuchung der Prognosemöglichkeiten erfolgt in drei Schritten: Modellentwicklung, Bestimmung der Prognosegüte und Modell-validierung.
Mit den 14 beschriebenen SAS-Makros ist ein Werkzeug vorhanden, mit dem die Durchführung einer vollständigen Modellierung eines logistischen Prognosemodells möglich. Speziell die Möglichkeiten der Modellvalidierung, die in der bisherigen Praxis selten genutzt werden, sollten so in Zukunft zu jeder Prognosemodellierung herangezogen werden.

Keywords: Prognosemodell, Logistische Regression, Modellvalidierung

1 Einleitung

„Prognose ist eine Vorhersage über den zukünftigen Verlauf einer Krankheit nach ihrem Beginn“ [6]. 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 Einschätzung 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. Die Modellierung und Überprüfung der Prognosegüte erfolgt dabei im Wesentlichen in drei Schritten:
(1) Modellentwicklung,
(2) Bestimmung der Prognosegüte und
(3) Modellvalidierung.

2 Logistische Regression

Die logistische Regression ist seit langem das Standardverfahren für die Analyse binärer Zielgrößen [10]. Die Modellgleichung zur Schätzung, ob ein Ereignis eintrifft (Y=1), gegeben einige Einflussgrößen X_1, X_2, ... , X_k , wird modelliert als:

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

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 geeignetste Lösung ist die über die Prozedur PROC LOGISTIC. 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 [7,12] 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.

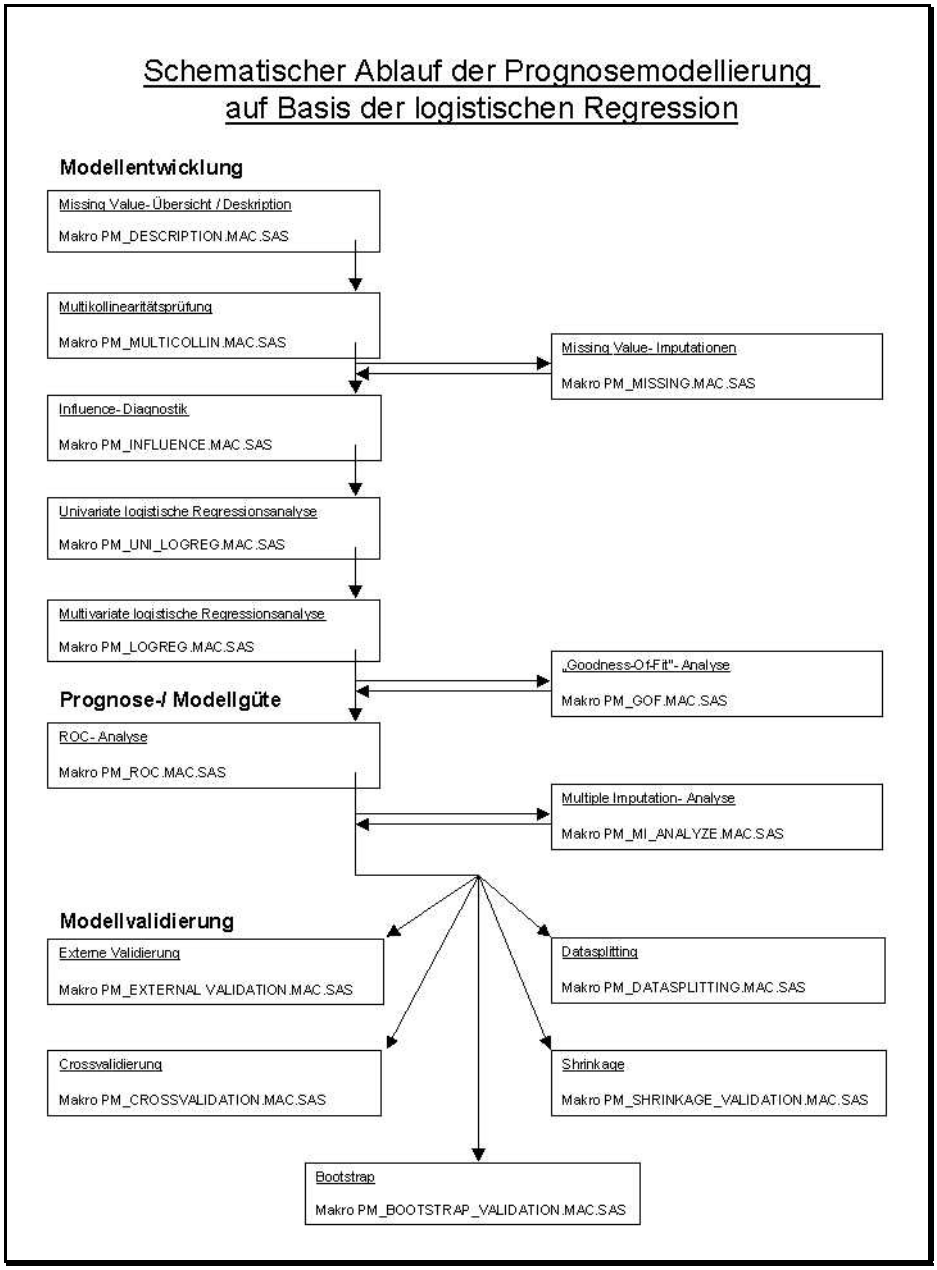

Abb. 2: Schematischer Ablauf der Prognosemodellierung auf Basis der SAS-Makros [12]

4 Allgemeiner 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 Kurzbeschreibung der SAS-Makros

In diesem Beitrag kann jedes Makro nur kurz beschrieben werden. Eine genaue und detaillierte Beschreibung findet sich in [12] bzw. [18]. Die folgende Kurzbeschreibung ist in die drei Oberbereiche der Prognosemodellierung: Modellentwicklung, Prognosegüte und Modellvalidierung aufgeteilt.

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, bevor das logistische Regressionsmodell angepasst wird.

5.1.1 PM_DESCRIPTION.MAC.SAS

Mit diesem Makro werden alle angegebenen Einfluss- sowie die Zielgröße univariat deskriptiv ausgewertet. Je nachdem, ob als stetig oder diskret angegeben werden PROC UNIVARIATE und PROC FREQ zur Analyse herangezogen.
Dabei kann über den Parameter miss= entschieden werden, ob alle Beobachtungen in jeder Variablen oder ein Complete-Case-Datensatz ausgewertet wird.
Zur Untersuchung der Missing-Value Situation im Datensatz lässt sich neben der Auszählung der fehlenden Werte pro Variable auch die Anzahl fehlender Werte pro Beobachtung ausgeben.

5.1.2 PM_MULTICOLLIN.MAC.SAS

Die Untersuchung der Multikollinearität geschieht durch:
- paarweise Korrelationen (Spearman, PROC CORR)
- Varianzinflationsfaktoren (VIF, PROC REG)
- Eigenwertanalyse ([3], PROC REG / COLLINOINT)

Die Auswertungen in PROC REG werden gewichtet mit geschätzten Wahrscheinlichkeiten aus PROC LOGISTIC durchgeführt [1].

5.1.3 PM_MISSING.MAC.SAS / PM_MI_ANALYZE.MAC.SAS

Mit Missing Values kann bei der Auswertung folgendermaßen umgegangen werden:
- Complete-Case-Analyse (miss=0)
- Single Imputation
  (stetig: PROC STDIZE, diskret: zus. Kategorie MISSING)
- Multiple Imputation (Untersuchung des Missing pattern, PROC MI)

Zur Zusammenfassung der Prognosegüten (nach logistischer Regression und ROC-Analyse (Receiver Operating Characteristics)) wird das Makro PM_MI_ANALYZE.MAC.SAS eingesetzt.

5.1.4 PM_INFLUENCE.MAC.SAS

Dieses Makro identifiziert die einflussreichsten Beobachtungen für die Modellierung. Dabei wird hauptsächlich die Veränderung der Pearson-Statistik nach Entfernung einer Beobachtung untersucht. Eine große Veränderung weist auf einen großen Einfluss auf die Parameterschätzung hin. Es werden schrittweise die einflussreichsten Beobachtungen bis zu einer vorgegebenen Schranke identifiziert, aber nicht automatisch aus dem Datensatz eliminiert.

5.1.5 PM_UNI_LOGREG.MAC.SAS

Es wird für jede Einflussgröße ein eigenes logistisches Regressionsmodell berechnet und der entsprechende p-Wert ausgegeben. Dabei kann auf den Complete-Case-Datensatz zurückgegriffen werden (miss=0). Kategorielle Variablen werden immer als Dummy-Variablen ins Modell aufgenommen und über das CLASS-Statement der gemeinsame Einfluss aller Dummies dieser Variable untersucht.
Stetige Variablen gehen linear ins Modell ein. Zusätzlich wird für jede stetige Variable eine Überprüfung der besseren Modellierung über „Fractional Polynomials“ bis zum Grad 2 [14] durchgeführt.

5.1.6 PM_LOGREG.MAC.SAS

Die eigentliche multiple logistische Regressionsanalyse wird mit diesem Hauptmakro durchgeführt. Mit diesem Makro wird ein multiples logistisches Regressionsmodell, eventuell mit Stepwise-Variablenselektion, angepasst.
Dies Makro liefert spezielle Ausgabedateien für die weitere Analyse (ROC, Modellvalidierung).
Zur Prüfung des gemeinsamen Einflusses von Variablen kann das TEST-Statement eingesetzt werden, das in PROC LOGISTIC nicht zusammen mit dem CLASS-Statement funktioniert.
Bei Problemen der Parameterschätzung durch eine quasi-complete-Separation wird eine korrigierte Schätzmethode über das FL-Makro [8] automatisch durchgeführt.
In der Abbildung 4 ist der schematische Ablauf des Makros dargestellt.

5.1.7 PM_GOF.MAC.SAS

Dieses Makro dient zur Überprüfung der Modellanpassung. Neben Parametern aus PROC LOGISTIC sind hier spezielle Tests für Sparseness (wenige Beobachtungen pro Merkmalskombination) aus der Literatur integriert (Makros aus [11], [13]), da in dieser Situation u.a. der Hosmer-Lemeshow-Test nicht mehr geeignet ist.

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.

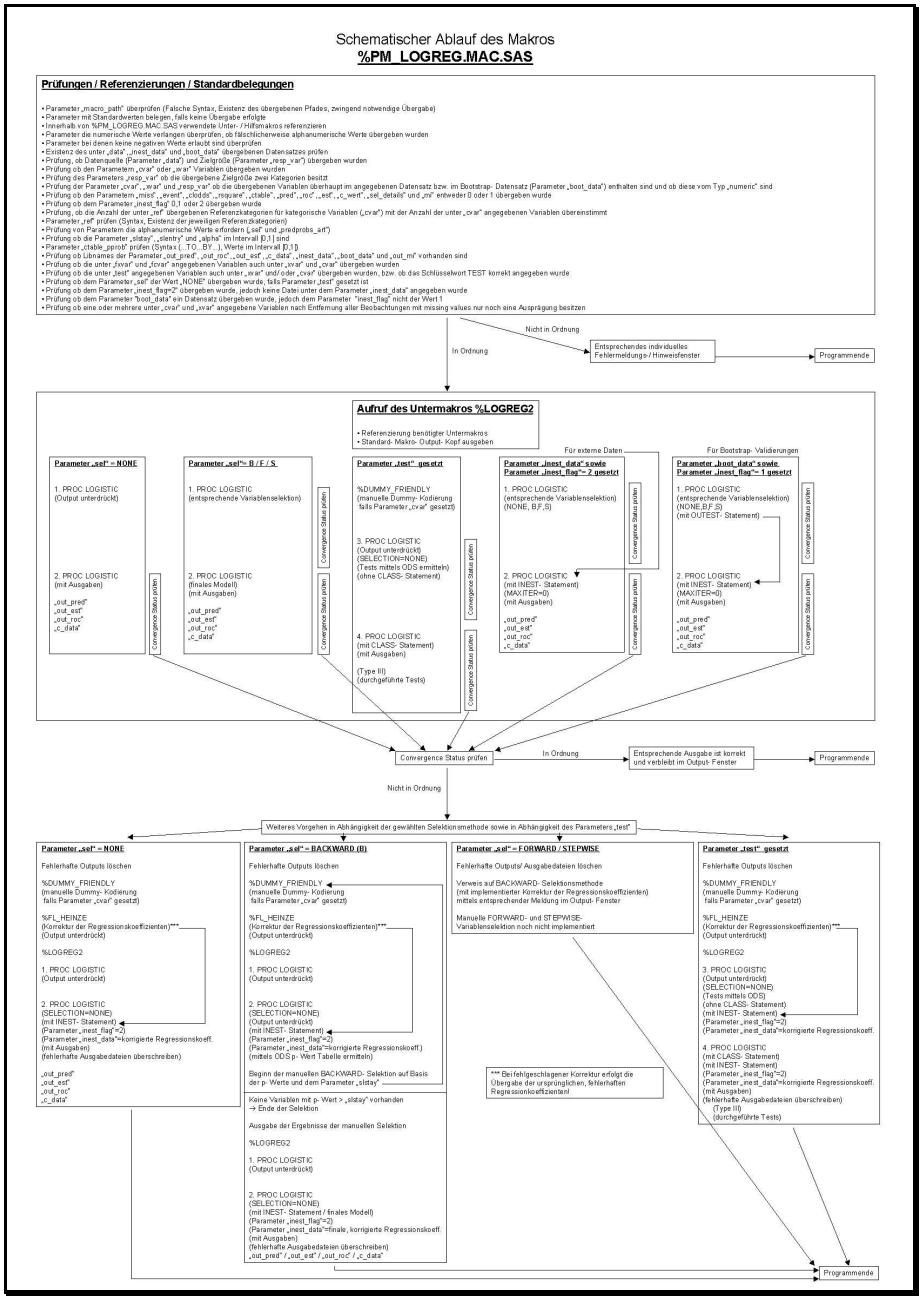

Abb. 4: Schematischer Ablauf von PM_LOGREG.MAC.SAS [18]

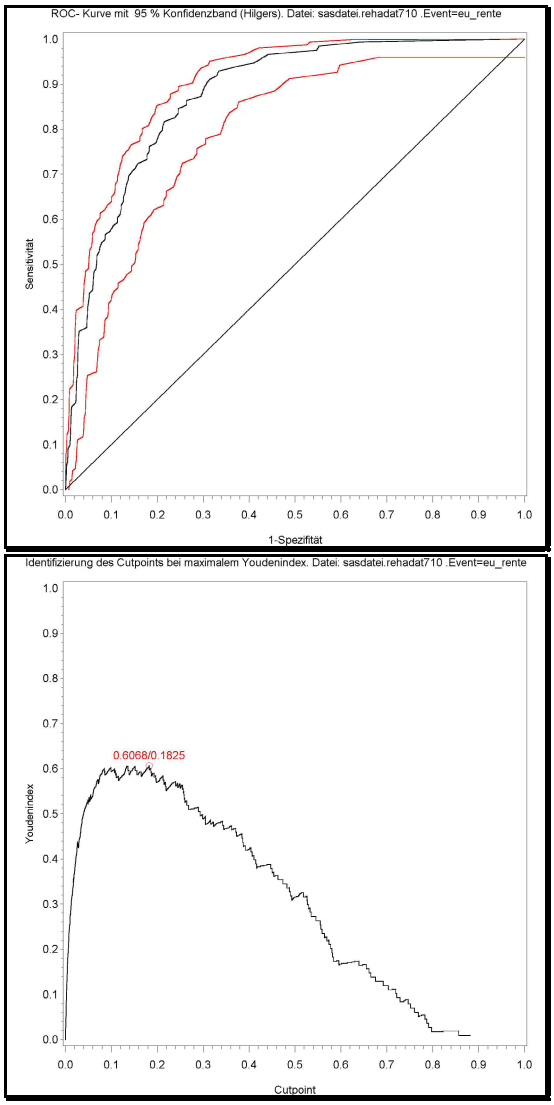

Abb.: 5: Beispielhafte Ausgabe von Grafiken aus PM_ROC.MAC.SAS

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.

5.2.1 PM_ROC.MAC.SAS

Die Prognosegüte wird anhand einer ROC-Analyse mit dem Makro PM_ROC.MAC.SAS durchgeführt. Die oben angegeben Maßzahlen (zusätzlich Konfidenzintervalle für AUC) sowie einige wichtige Grafiken (u.a. ROC-Kurven (inkl. Konfidenzbänder nach Hilgers [9], Zusammenhang Youden-Index und Cutpoint) werden ausgegeben.

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 [16].
Als Output werden jeweils die Cutpoint-abhängigen und -unabhängigen Prognosegütemaße der ROC-Analyse vor und nach der Validierung sowie die absolute und relative Veränderung ausgegeben.

5.3.1 PM_EXTERNAL_VALIDATION.MAC.SAS

Bei der externen Validierung wird die Prognosegüte des Modells anhand eines zweiten, unabhängigen Datensatzes bestimmt. Das ist die Methode der Wahl, wenn ein zweiter Datensatz zur Verfügung steht. Leider ist dies selten der Fall, so dass auf Methoden zurückgegriffen werden muss, die auf dem vorhandenen Datensatz basieren (interne Validierungsmethoden).

5.3.2 PM_DATASPLITTING.MAC.SAS

Durch das Data-Splitting wird der Datensatz geteilt. Ein Teil wird zur Modellentwicklung, der Zweite zur Validierung (s. externe Validierung) genutzt. Dabei beruht die Modellentwicklung und Modellvalidierung allerdings auf einer wesentlich geringeren Fallzahl, so dass dieses Verfahren nur selten sinnvoll eingesetzt werden kann.
Das Makro teilt den Datensatz nach vorgegebener Prozentangabe zufällig auf. Für die Validierung ist anschließend PM_EXTERNAL_VALIDATION .MAC.SAS aufzurufen.

5.3.3 PM_CROSSVALIDATION.MAC.SAS

Die Kreuzvalidierung war lange Zeit das Standardverfahren für die Modellvalidierung. Prinzipiell liegt dem Verfahren ein Stichprobenziehen ohne Zurücklegen zugrunde. Das Vorgehen kann folgendermaßen skizziert werden: Datensatz in K Teile teilen; anschließend an K-1 Teilen das Modell entwickeln und am K-ten Teil validieren. Das ganze wird für alle K Teile wiederholt.
Im Makro sind folgende Methoden programmiert: K-fold Crossvalidation, adjusted Crossvalidation [4], Jackknife-Crossvalidation.

5.3.4 PM_BOOTSTRAP_VALIDATION.MAC.SAS

Die Methode der Bootstrap-Validierung [2] ist ebenfalls ein Resampling-Verfahren, basiert aber auf einem Ziehen mit Zurücklegen: Es werden Datensätze gleicher Größe aus dem vorhandenen Datensatz erzeugt. Anhand dieser so erzeugten Datensätze kann die Modellierung und/oder Validierung der Modelle erfolgen. Durch geeignetes Zusammenführen der Einzelergebnisse kann der Bias der Prognosegüte abgeschätzt werden.
Im Makro implementiert sind die Vorschläge von Efron [5]: simple- / enhanced Bootstrap sowie ein Ansatz über Mittelung der Regressionskoeffizienten (Mean Model).

5.3.5 PM_SHRINKAGE.MAC.SAS

Die Shrinkage-Methode korrigiert die geschätzten Regressionskoeffizienten [17], so dass die Prognosegüte anhand des korrigierten Modells bestimmt wird.

Drei Methoden sind im Makro implementiert: heuristischer Shrinkage, globaler Shrinkage [17] sowie ein parameterbezogener Shrinkage-Faktor, der auf Sauerbrei zurückgeht [15].

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 –validierung 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] Assfalg I. (2003) Die Bootstrap-Methode zur internen Validierung von Prognosemodellen. Diplomarbeit FH Ulm

[3] Belsley D.A. (1991) Conditioning diagnostics – Collinearity and weak data in regression. John Wiley & Sons, New York

[4] Davison A.C., Hinkley D.V. (1997) Bootstrap methods and their application. Cambridge University Pressl, Cambridge

[5] Efron B., Tibshirani R.J. (1993) An Introduction to the Bootstrap. Chapman & Hall, New York

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

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

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

[9] Hilgers R. (1991) Distribution-free confidence bounds for ROC curves. Meth. Inform. Med. 30, 96-101

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

[11] Kuss O. (2002) Global goodness-of-fit-tests in logistic regression with sparse data. Stat. Med. 21, 3789-3801

[12] Muche R. (2004) Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression Habilitationsschrift Medizinische Fakultät, Universität Ulm

[13] Pulkstenis E., Robinson T.J. (2002) Two goodness-of-fit tests for logistic regression with continuous covariates. Stat. Med. 21, 79-93

[14] Royston P., Altman D.G. (1994) Regression using fractional polynomials of continuous covariates. Appl. Statist. 43, 429-467

[15] Sauerbrei W. (1999) The use of resampling methods to simplify regression models in medical statistics. Appl. Statist. 48, 313-329

[16] 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

[17] van Houwelingen H., LeCessie S. (1990) Predictive value of statistical models. Stat. Med. 9, 1303-1325

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