SAS-MAKROS ZUR ENTWICKLUNG UND VALIDIERUNG VON PROGNOSEMODELLEN AUF BASIS DER LOGISTISCHEN REGRESSION

GETESTET IN SAS VERSION 9.4

Name:	Sandra Müller
Studiengang:	Mathematische Biometrie, Universität Ulm

Name:	Prof. Dr. Rainer Muche
Institution:	Universität Ulm
Abteilung:	Institut für Epidemiologie und Medizinische Biometrie
	Schwabstraße 13
	89075 Ulm
Telefon:	0731 – 50 26903

Zeitraum:	
Beginn:	04.04.2016
Ende:	29.05.2016

1 PROGNOSEMODELLE MIT HILFE VON SAS-MAKROS

Im Institut für Epidemiologie und Medizinische Biometrie der Universität Ulm entstand im Rahmen einer Habilitationsschrift im Jahre 2004 [1] eine Auswertungsstrategie, mit deren Hilfe es möglich ist, komplexe Prognosen, die auf logistische Regressionsmodelle aufbauen, schnell und einfach mit SAS zu erstellen.

Dafür wurden mehrere SAS-Makros entwickelt, die alle zusammen genommen ein umfangreiches Paket darstellen, das von der deskriptiven Analyse der Variablen bis hin zur kompletten Modellierung und Validierung des Prognosemodells alle wichtigen Punkte abdeckt.

Die Makros wurden an einem Beispieldatensatz aus dem Reha-Forschungs-Verbund Ulm [4] entwickelt und getestet, der auch mir für meine Arbeit zu Grunde lag. Dieser Datensatz [5] enthält Informationen über Reha-Aufenthalte von Patienten, wie zum Beispiel Dauer des Aufenthaltes und Art der Reha. Die erfasste Zielgröße ist das Beziehen einer Erwerbsunfähigkeitsrente bis zu 2 Jahre nach dem stationären Reha-Aufenthalt. Mit Hilfe der Makros soll ein Prognosemodell erstellt werden, das zuverlässig Angaben über einen neuen Reha-Patienten machen kann um durch intensivere Reha oder Nachbeobachtungen die Zeit bis zum Erhalt der Rente zu verlängern oder diese gar ganz zu umgehen.

In einem Praktikum, dass Bestandteil des Studienganges "Medizinische Dokumentation und Informatik" der Fachhochschule Ulm ist, wurde im Jahr 2005 die Aufgabe gestellt diese Makros, die in SAS Version 8 erstellt wurden, in der SAS Version 9.1 zu testen. Eventuell auftretende Probleme, die durch den Versionswechsel verursacht werden sollten gefunden und behoben werden, damit die Makros auch in der nächsten SAS-Generation benutzt werden können. Diese Aufgabenstellung liegt nun auch meinem Praktikum für den Studiengang „Mathematische Biometrie“ der Universität Ulm zu Grunde. Die Makros, die für SAS Version 9.1 angepasst wurden, sollen auf ihre Funktionsfähigkeit in SAS Version 9.4 überprüft und gegebenenfalls aufgetretene Fehler behoben werden.

Das komplette Paket besteht aus 102 Makro-Dateien, davon enthalten 14 Dateien die eigentlichen Makroaufrufe (Hauptmakros). In 65 Dateien sind Fehlermeldungen gespeichert, die verbleibenden 23 Makros sind Hilfsmakros, die von anderen Makros aufgerufen werden. Die Hilfsmakros kommen teilweise von externen Stellen, anderen Universitäten etc.

Während meines Praktikums habe ich mich hauptsächlich auf die Hauptmakros konzentriert. Diese werde ich im Folgenden hier vorstellen, kurz die Funktionen erläutern und dann aufgetretene Probleme mit entsprechenden Lösungsvorschlägen darstellen.

Zum Testen der Makros habe ich denselben Datensatz benutzt, an dem auch die Makros entwickelt wurden. Beim ersten Erproben wurde die Situation dargestellt, die auch in der Habilitationsschrift [1] angegeben wird. Das bedeutet, dass das Makro mit allen dort beschriebenen Einstellungen der Parameter gestartet wurde. Verglichen wurde der Output des Makros mit dem Output, der in der Habilitationsschrift abgedruckt ist, bzw. der in Version 8 als Vergleich erzeugt wurde. Außerdem wurden die Meldungen im LOG-Fenster nach Warnungen und Fehler untersucht und unter Umständen erzeugten Dateien auf ihre Korrektheit überprüft.

Aufgetretene Fehler wurden untersucht um Ursachen herauszufinden und zu beheben. Die Vorgehensweise dabei kann sehr unterschiedlich sein, hilfreich ist aber in den meisten Fällen die Isolation von einzelnen Befehlen und Aufrufen (z. B. PROC LOGISTIC) um an einfachen Programmen Änderungen schnell feststellen zu können.

Da die Makros in der Praxis aber nicht nur mit ihrer Standardeinstellung verwendet werden, ist es notwendig auch alle weiteren Parameter, mit denen das Makro aufgerufen werden kann, zu testen. Dazu habe ich jeden einzelnen Parameter verstellt und untersucht, ob das ausgegebene Ergebnis den Erwartungen entsprach. Wenn das nicht der Fall war hab ich auch hier nach den Problemen gesucht und diese (wenn möglich bzw. notwendig) auch behoben.

1.1 VORBEREITUNGEN ZUR REGRESSIONSANALYSE

Bevor auf einen Datensatz eine Regressionsanalyse angewandt wird ist es sinnvoll diesen nach verschiedenen Gesichtspunkten zu untersuchen um inhaltliche Fehler zu vermeiden und mögliche Über- oder Unterschätzungen der Regressionsparameter zu verhindern. Die folgenden Makros beinhalten also Auswertungsstrategien, um den vorhanden Datensatz auf die multivariate Regressionsanalyse vorzubereiten.

1.1.1 PM_DESRIPTION.MAC.SAS

Bevor überhaupt mit irgendeiner statistischen Auswertung begonnen wird empfiehlt es sich, dass man sich Klarheit über die vorhandene Datenstruktur verschafft. Dazu dient dieses Makro, das (wie der Name schon sagt) eine deskriptive Darstellung des Datensatzes erzeugt.

Das Makro erstellt Kreuztabellen, die die Häufigkeiten der kategoriellen Merkmale gruppiert nach der Zielgröße enthalten. Für die Darstellung der stetigen Merkmale werden mit Hilfe der PROC UNIVARIATE Stichprobenumfang, Mittelwert und Standardabweichung sowie die Perzentile in 20er-Schritten und der Median ausgegeben. Klassiert nach den Perzentilen und gruppiert nach der Zielgröße werden zusätzlich Kreuztabellen erstellt. Für alle stetigen Variablen kann man sich zudem noch Histogramme generieren lassen, die Aussagen zur Verteilung zulassen.

Die Ausgabe des Makros liefert daneben auch Informationen über fehlende Werte im Datensatz, gruppiert nach stetigen und kategoriellen Merkmalen und der Zielgröße, die für das Makropaket immer dichotom sein muss (in diesem Fall Erwerbsunfähigkeitsrente ja oder nein). Außerdem kann man sich anzeigen lassen welche Beobachtung die meisten fehlenden Werte hat.

Fehlerbeschreibung:

Das Makro gibt Histogramme aller stetigen Variablen aus. Durch Makrovariablen kann das Format und der Speicherort gewählt werden. Diese Angaben werden jedoch ignoriert und alle Graphiken werden als PNG im Arbeitsordner gespeichert.

Fehlerbehebung:

Das Format und der Speicherort der Histogramme werden in GOPTIONS eingestellt. Diese Angaben und andere Formatierungen der Graphik scheinen allerdings ignoriert zu werden. Das Problem wird gelöst, in dem die ODS Graphikausgabe kurzzeitig unterbrochen wird. Dazu wird in Zeile 1760 vor den GOPTIONS das Statement

```
ODS GRAPHICS OFF;
```

eingefügt. Nach der Ausgabe des Histogramms über die PROC UNIVARIATE wird diese Einstellungen durch Einfügen von

```
ODS GRAPHICS;
```

in Zeile 1771 wieder aufgehoben. Die komplette Ausgabe des Histogramms einer Variablen ist dann wie folgt programmiert:

```
ODS GRAPHICS OFF;

GOPTIONS DEVICE=&ext   GSFNAME=grafout GSFMODE=replace FONTRES=PRESENTATION
KEYMAP=NONE FTEXT="Arial" HTITLE=1 HTEXT=1 XMAX=6 in YMAX=6 in XPIXELS=1800
YPIXELS=1800;
FILENAME grafout &path;

PROC UNIVARIATE DATA=&data NOPRINT;
      VAR &var;
      HISTOGRAM &var;
      OUTPUT OUT=&var N=N MEAN=MEAN STD=STD MIN=MIN pctlpre=P_ pctlpts=20 to
      80 by 20 MAX=MAX MEDIAN=MEDIAN;
RUN;
ODS GRAPHICS;
```

1.1.2 PM_MULTICOLLIN.MAC.SAS

Multikollinearität untersucht die Zusammenhänge zwischen einzelnen Variablen um Zusammenhänge im Datensatz zu erkennen und unter Umständen Variablen, die durch andere beschrieben werden, aus dem Modell nehmen zu können. Wichtige Kenngrößen die dazu ermittelt werden sind zum Beispiel die Korrelationskoeffizienten nach Spearman. Das Makro nutzt in der Hauptsache die SAS-Prozedur PROC REG um die Kollinearitätsdiagnostik zu berechnen. Zur Übersichtlichkeit werden nur Variablen angezeigt, die in diesen Werten über einer (vom Benutzer) festgelegten Schranke liegen. Im Reha-Datensatz wurde zum Beispiel eine starke Korrelation zwischen der Anzahl der bewilligten Tage und der Heilverfahrensart festgestellt.

Fehlerbeschreibung:

Der Spearman-Rangkorrelationskoeffizient soll nur für alle Variablenkombinationen mit einem Koeffizienten größer einem gewählten Wert (hier: 0.6) und kleiner 1 ausgegeben werden. In der Übersicht erscheinen allerdings auch die stetigen Variablen BMI und Alter zu sich selbst mit einem Koeffizienten von jeweils 1:

```
Spearman Rangkorrelationskoeffizienten >0.6 und <1

Variable1    Variable2        spearman

  bmi        bmi               1.00000
  alter      alter             1.00000
  hvart      tagebewilligt     0.86650
```

Fehlerbehebung:

Im Output werden die Korrelationskoeffizienten zwar mit 1.0 angegeben, tatsächlich erreichen sie diesen Wert aber nicht ganz. Die Zeilen 1932 bis 1936 im Makro zeigen, wie die Tabelle für den Output generiert wird:

```
DATA spearman&i;
      SET spearman&i;
      IF &var_string=1 THEN DELETE;
      IF &var_string < &spearman THEN DELETE;
RUN;
```

Es werden zunächst alle Korrelationskoeffizienten berechnet und anschließend alle Kombinationen gelöscht, deren Wert unter 0.6 oder gleich 1 ist. Problematisch ist hierbei die Abfrage auf exakte Gleichheit. Deshalb wird Zeile 1934 ersetzt durch:

```
IF &var_string > 0.999 THEN DELETE;
```

1.1.3 PM_MISSING.MAC.SAS

Dieses Makro ersetzt fehlende Werte im Datensatz. Dazu kann der Benutzer verschiedene Möglichkeiten einstellen. Die Single Imputation ersetzt fehlende Werte stetiger Variablen zum Beispiel durch den Mittelwert oder den Median der anderen vorhandenen Werte innerhalb einer Variablen. Fehlende Werte kategorieller Variablen werden durch eine Missing-Kategorie, die vom Benutzer angegeben werden kann, ersetzt. Bei der Multiple Imputation werden mehrere Datensätze erstellt, die fehlende Werte über eine zu Grunde gelegte Verteilung (zum Beispiel Normalverteilung) ersetzt. So entstehen mehrere verschiedene Datensätze, die getrennt voneinander ausgewertet werden können um dann am Schluss diese Ergebnisse wieder zusammenzuführen. Die SAS-Prozedur PROC MI bietet mit seinen Statements die erforderlichen Ersetzungsmethoden.

Die Ausgabe dieses Makros besteht außer aus den Datensätzen auch aus einer Übersicht über die Verteilung der fehlenden Werte und einer Zusammenfassung über die Art der Ersetzung(en). Der Output stimmte in Version 9.4 mit der Referenz aus der Habilitationsschrift [1] überein.

1.1.4 PM_MI_ANALYZE.MAC.SAS

Die Datensätze, die über die multiple Imputation des Makros PM_MISSING.MAC.SAS erzeugt wurden, können mit diesem Makro ausgewertet werden. Zuvor muss jedoch für jeden dieser Datensätze mit Hilfe der Makros PM_LOGREG.MAC.SAS (siehe Kapitel 2.2.2) sowie PM_ROC.MAC.SAS (siehe Kapitel 2.3.2) ein Regressionsmodell entworfen werden. Diese Modelle lassen sich mit dem PM_MI_ANALYZE.MAC.SAS-Makro zusammenfassend darstellen. In SAS gibt es dazu die Prozedur PROC MIANALYZE.

Das Makro hat in Version 9.4 keine Auffälligkeiten gezeigt.

1.1.5 PM_INFLUENCE.MAC.SAS

Dieses Makro ermittelt die einflussreichste Beobachtung im Datensatz. Diese Beobachtungen können unter Umständen durch ihre extreme Abweichung zu anderen Beobachtungen großen Einfluss auf das zu ermittelnde Modell haben, und dieses verzerren. Daher kann es sinnvoll sein diese Beobachtungen bei der weiteren Modellbetrachtung nicht weiter zu berücksichtigen. Die Ermittlung solcher Beobachtungen erfolgt durch Methoden der Ausreißer- und / oder Residualanalyse. Zusätzlich können die Ergebnisse graphisch ausgegeben werden. Dieses Makro benutzt zur Berechnung seiner Ausgabe die Methoden der PROC LOGISTIC mit dem Parameter 'difchisq'. Das steht für die Veränderung der Pearson-Chi-Quadrat-Statistik bei Entfernung einer Beobachtung.

Die Ausführung dieses Makros mit den verschiedenen Testparametern ergab keine Hinweise auf Versionskonflikte.

1.2 REGRESSIONSANALYSE-MAKROS

Nachdem der Datensatz inhaltlich so weit aufgeschlossen ist, dass fehlenden Werte ersetzt, unbedeutende Variablen und einflussreiche Beobachtungen erkannt bzw. eliminiert wurden, kann mit der Berechnung des eigentlichen Regressionsmodells begonnen werden. Dazu bietet das Paket die folgenden Makros an.

1.2.1 PM_UNI_LOGREG.MAC.SAS

Bei der univariaten logistischen Regression wird der Einfluss jeweils einer Variablen auf die Zielgröße untersucht. Mit diesem Vorgehen werden zwei Ziele verfolgt: Einerseits können die Voraussetzungen für den Einsatz einer Variablen untersucht werden, andererseits werden so Variablen, die nur geringen Einfluss auf die Zielgröße haben, gefunden und müssen im Modell nicht weiter berücksichtigt werden. Das hat zum Vorteil, dass sich die Variablenanzahl im Modell reduziert, was auch nach Vorauswahl der inhaltlich relevanten Daten nicht immer optimal der Fall ist.

Der Output stimmte in Version 9.4 mit der Referenz aus der Habilitationsschrift [1] überein.

1.2.2 PM_LOGREG.MAC.SAS

Da hier das eigentliche Regressionsmodell entwickelt wird ist dieses Makro das wichtigste im ganzen Paket. Die eigentlichen Regressionsberechnungen (über die SAS-Prozedur PROC LOGISTIC) finden im Untermakro LOGREG2.MAC.SAS statt.

Fehlerbeschreibung Graphiken:

Das Makro erzeugt eine Graphik der ROC-Kurve des finalen Modells im Arbeitsordner, obwohl dies laut der Makro-Beschreibung nicht vorgesehen ist.

Fehlerbehebung:

Bei der Programmierung der Makros war es mit der damaligen SAS Version noch nicht möglich mit Hilfe der PROC LOGISTIC eine ROC-Kurve zu erstellen. Deshalb wurde dies anders gelöst und in ein eigenes Makro, PM_ROC.MAC.SAS, ausgelagert. Durch die Änderungen in SAS Version 9.4 wird dieses separate Makro obsolet und könnte in das Makro PM_LOGREG.MAC.SAS integriert werden. Dies wäre allerdings mit einem nicht geringen Programmieraufwand und anschließenden Validierungstests verbunden. Deshalb soll die bisherige Aufteilung der Makros erhalten bleiben und die Ausgabe der ROC-Kurve in PM_LOGREG.MAC.SAS unterdrückt werden. Dies geschieht in den Zeilen 399 und 892 im Hilfsmakro LOGREG2.MAC.SAS durch die Option PLOTS=NONE im Aufruf von PROC LOGISTIC.

Fehlerbeschreibung numerische Abweichungen:

Im Abschnitt “Association of Predicted Probabilities and Observed Responses” der Ausgabe gibt es numerische Abweichungen. Bei den Prozenten der konkordanten, diskordanten und gebundenen Paare sind diese jeweils geringer als ein Prozentpunkt. Bei den Assoziationsmaßen Somers‘ D und Gamma liegen die Abweichungen im Bereich von 10^{-3}.

Fehlerbehebung:

Zunächst wurde eine genauere Zahlendarstellung durch die Umstellung von SAS 32bit auf 64bit als Grund der numerischen Abweichungen vermutet. Dies konnte durch einen Test mit SAS Version 9.3 32bit und einem anschließendem Vergleich der Ergebnisse jedoch ausgeschlossen werden.

Eine solche numerische Abweichung wurde auch in einem Forenbeitrag der SAS Support Community [8] diskutiert. Als Grund wurde hierbei eine geänderte Standardbelegung der Option BINWIDTH im MODEL-Statement der Prozedur PROC LOGISTIC angegeben. Dieser Parameter gibt die Intervalllänge an, in der die berechneten Wahrscheinlichkeiten bei der Bestimmung der konkordanten, diskordanten und gebundenen Paare zusammengefasst werden. Je größer der angegebene Wert ist, desto mehr gebundene Paare gibt es. Bisher verwendete SAS aus Performanzgründen eine BINWIDTH von 0.002. Ab SAS Version 9.4 wird hingegen bei einem binären Zielereignis und weniger als 5.000.000 Beobachtungen standardmäßig ohne Binning, also mit einem Wert von 0, gerechnet. Dies ist laut einer Usage Note [9] insbesondere der Fall, wenn ein ROC Statement angegeben ist, die Daten der ROC-Kurve gespeichert werden oder eine ROC-Kurve ausgegeben wird. In diesen Fällen werden alle Angaben der Option BINWIDTH ignoriert und es wird kein Binning verwendet. Die genannten Ausnahmefälle treten im Makro PM_LOGREG.MAC.SAS und im Hilfsmakro LOGREG2.MAC.SAS jedoch nicht auf.

Die Berechnung ohne Binning ist genauer als mit und sollte damit auch weiterhin erhalten bleiben. Gleichzeitig soll aber auch ein Vergleich der Ausgabe mit der aus SAS Version 9.1 unter den gleichen Voraussetzungen möglich sein. Dies betrifft auch alle anderen Makros, die das Makro PM_LOGREG.MAC.SAS und dessen Ergebnisse direkt verwenden. Daher wird in den Hauptmakros PM_LOGREG.MAC.SAS, PM_ROC.MAC.SAS, PM_BOOTSTRAP_VALIDATION.MAC.SAS, PM_CROSSVALIDATION.MAC.SAS, PM_EXTERNAL_VALIDATION.MAC.SAS, PM_SHRINKAGE_VALIDATION.MAC.SAS und den Hilfsmakros LOGREG2.MAC.SAS, BOOT_MODEL_VALIDATION.MAC.SAS und BOOT_MEAN_VALIDATION.MAC.SAS eine neue Makrovariable BINWIDTH eingefügt. Zudem wird die neue Fehlermeldung error66.sas erstellt, die bei einer Eingabe der BINWIDTH kleiner 0 oder größer gleich 1 ausgegeben wird.

Das Makro FL_HEINZE.MAC.SAS:

Bei manchen (meist kleinen) Datensätzen kann es vorkommen, dass die Zielgröße allein über eine bestimmte Wertekonstellation der Einflussgrößen vorhergesagt werden kann (ein Regressionsmodell ist dann eigentlich nicht mehr notwendig), man spricht dabei von einer (quasi) complete separation. Dieser Fall bringt Probleme beim Erstellen des Regressionsmodells mit sich. Zum gezielten Abfangen dieses Falles wird das FL-Makro von Georg Heinze [7] aktiviert.

Diese Situation kommt beim vorliegenden Datensatz nicht vor, zum Testen habe ich ein kleines Hilfsprogramm geschrieben, und einen weiteren Datensatz aus der Abteilung verwendet, der das entsprechende Problem erzeugt. Das so manuell aufgerufene Makro zeigte keinen Versionskonflikt.

1.3 MAKROS ZUR MODELLGÜTE

Nachdem das Regressionsmodell erstellt wurde gilt es hier nun zu prüfen wie gut es denn zu den zu den beobachteten Daten passt. Hierzu dienen die nächsten beiden Makros.

1.3.1 PM_GOF.MAC.SAS

Für die Anpassungsgüte (Goodness of Fit) wird der Zusammenhang zwischen den beobachteten Werten y und den durch das Modell geschätzten Werten $\hat{y}$ untersucht. Auch hier kommt wieder die PROC LOGISTIC zum Einsatz.

Dieses Makro greift auf mehrere Untermakros zu, das Makro von Kuss [3] stellt 5 verschiedene statistische Tests zur Prüfung der Goodness of Fit zur Verfügung. Da dieses Makro allerdings in SAS Version 6 erstellt wurde, und es damals noch kein CLASS-Statement in der PROC LOGISTIC gab, ist das Makro von Friendly notwendig, das eine manuelle Dummy-Kodierung der kategoriellen Variablen durchführt. [6]

Fehlerbeschreibung:

Beim Hosmer and Lemeshow Goodness-of-Fit Test gibt Unterschiede in der Klasseneinteilung und damit numerische Abweichungen bei χ^2 und dem dazugehörigen p-Wert.

Fehlerbehebung:

Die 710 Beobachtungen wurden bisher auf zehn Klassen mit jeweils 69 bis 73 Beobachtungen aufgeteilt. In SAS Version 9.4 ist die Aufteilung allerdings gleichmäßig mit 71 Beobachtungen pro Klasse. Laut dem SAS User's Guide werden bei der Klasseneinteilung Gruppen von Beobachtungen mit der gleichen geschätzten Wahrscheinlichkeit nicht getrennt, was zu unterschiedlichen Klassengrößen führen kann. [9] Es besteht daher die Vermutung, dass auch hier eine unterschiedliche BINWIDTH die Ursache der Abweichungen ist. Leider wird jede Angabe eines Wertes für die BINWIDTH ignoriert, wenn die Option LACKFIT zur Berechnung des Hosmer and Lemeshow Goodness-of-Fit Tests angegeben wird. Die Vermutung lässt sich daher nicht bestätigen.

Da die Berechnung mit einer gleichmäßigen Klasseneinteilung vermutlich genauer ist, bleibt das Makro PM_GOF.MAC.SAS unverändert.

1.3.2 PM_ROC.MAC.SAS

Die ROC-Analyse bestimmt die Prognosegüte, unter anderem Sensitivität und Spezifität bei unterschiedlichem Cut-Point (Der Cut-Point bestimmt die Merkmalsausprägungen einer Beobachtung mit der diese in eine der Gruppen der Zielgröße einsortiert wird.). Das Makro liefert Auskunft über die Verteilung von Sensitivität und Spezifität und stellt diese grafisch dar. Hauptsächlich wird hier mit DATA-Steps gearbeitet.

Fehlerbeschreibung:

Für die ROC-Analyse werden insgesamt vier Graphiken erzeugt und gespeichert. Zwei davon, die Graphiken gplot1 und gplot3, enthalten Beschriftungen. Diese werden in der SAS Version 9.4 nicht nur einmal, sondern vier- bzw. fünfmal angezeigt.

Fehlerbehebung:

Die Graphik gplot1 zeigt eine ROC-Kurve mit eingezeichnetem Punkt für den maximalen Youdenindex. Dieser ist mit der zugehörigen Sensitivität und Spezifität beschriftet. Beide Informationen sind aber auch in der Überschrift enthalten. Die Beschriftung ist somit überflüssig und kann einfach entfernt werden.

Die Graphiken werden in beiden Fällen im Hilfsmakro PRINTIT.MAC.SAS erzeugt. Für die Graphik gplot1 ist dabei die Zeile 275 relevant:

```
symbol5 C=RED L=1 W=1 V=CIRCLE pointlabel = ("#sensi:#spezi $/"  j=l c=red);
```

Sie wird durch folgenden Code ersetzt:

```
symbol5 C=RED L=1 W=1 V=CIRCLE; /* Schrift entfernt, da das Label mehrmals in
der Graphik war */
```

In Graphik gplot3 wird der Youdenindex für jeden Cut-Point aufgetragen. Das Maximum des Graphen ist markiert und beschriftet. Auch in diesem Fall wird diese Beschriftung entfernt. Geändert wird hierbei Zeile 412 im Hilfsmakro PRINTIT.MAC.SAS:

```
symbol11 C=RED I=STEPCJ L=1 W=1 V=CIRCLE pointlabel = ("#maxyouden:#cutpoint
$/"  c=red);
```

Die Zeile wird ersetzt durch:

```
symbol11 C=RED I=STEPCJ L=1 W=1 V=CIRCLE;
```

Da die Informationen des Labels nicht bereits in der Überschrift vorhanden sind, wird eine solche zweite Überschrift in Zeile 433 ergänzt:

```
TITLE2 H=1 "Maximaler Youdenindex: &maxyouden, Cutpoint: &cutpoint" ; /* Werte
in Überschrift eingefügt */
```

1.4 MAKROS ZUR MODELLVALIDIERUNG

Bei einem an einem Beispieldatensatz gewonnenem Regressionsmodell muss geprüft werden mit welcher Güte (Prognosegüte) dieses Modell auf neue Daten angepasst werden kann. Es gibt mehrere Arten der Validierung, man kann diese einteilen in interne, temporale und externe Validierungsmethoden.

Bei den **internen** Validierungsmethoden wird der Datensatz, der auch zur Erstellung des Modells zur Verfügung stand, benutzt um das Modell zu testen. Beispielsweise werden mehrere zufällige Datensätze aus dem Orignaldatensatz gewonnen, die dann auf das Modell angewendet werden um seine Güte zu ermitteln. Diese Vorgehensweise muss in der Regel mehrere hundert Mal angewandt werden um zu einem richtigen Ergebnis zu kommen.

Die **temporale** Validierung nutzt den Orignaldatensatz, der mit einem Zufallsverfahren in mehrere, meistens zwei, Datensätze gespalten wird. An einem Teildatensatz wird das Modell entwickelt und an dem anderen validiert. Allerdings stammen beide Datensätze immer noch aus einer gemeinsamen Quelle.

Wenn eine Validierung auf einem zweiten, unabhängigen Datensatz beruht, spricht man von einer **externen** Validierung. Da hier mit komplett neuen Daten gearbeitet wird ist dies die weitestgehende Form der Validierung.

1.4.1 PM_EXTERNAL_VALIDATION.MAC.SAS

Bei der externen Validierung ist, wie oben bereits erwähnt, die Anwendung eines zweiten unabhängigen Datensatzes gefordert. Dieser muss für das Makro strukturgleich sein, das heißt, dass dieselben Variablennamen und Formatierungen in beiden Datensätzen gegeben sein müssen. Das Makro untersucht mit Hilfe der Makros PM_LOGREG.MAC.SAS und PM_ROC.MAC.SAS die Prognosegüte eines festen Modells auf Basis neuer Daten. Dabei unterscheidet das Makro nicht zwischen rein externer und temporaler Validierung, da dies Fragen inhaltlicher Natur sind.

Beim Testen des Makro sind keine weiteren Probleme aufgetreten.

1.4.2 PM_DATASPLITTING.MAC.SAS

Um für die temporale Validierung einen zweiten Datensatz mit den geforderten Eigenschaften (gleiche Variablen und Formatierungen) zu erhalten erzeugt dieses Makro aus einem Quelldatensatz auf Basis eines angegebenen Prozentsatzes zwei zufällige Teil-Datensätze. Dabei greift das Makro auf die SAS-Funktion RANUNI zu.

Dieses Makro ist eins der "einfachen" Makros und wie erwartet sind hier keine Probleme aufgetreten.

1.4.3 PM_CROSSVALIDATION.MAC.SAS

Die Kreuzvalidierung ist eine Weiterführung des Data-Splittings. Dabei wird der Datensatz mehrfach in Gruppen unterteilt und die Auswertungen jeweils auf diese Untergruppen durchgeführt. Es gibt 4 verschiedene Kreuzvalidierungsarten [1], die von dem Makro durchgeführt werden können. Für die einzelnen Modellerstellungen und –validierungen werden wie bei dem Makro für die externe Validierung die Makros PM_LOGREG.MAC.SAS und PM_ROC.MAC.SAS verwendet.

Das Makro hat in Version 9.4 keine Auffälligkeiten gezeigt.

1.4.4 PM_BOOTSTRAP_VALIDATION.MAC.SAS

Das Prinzip des Ziehens mit Zurücklegen kommt bei der Bootstrap-Validierung zum Einsatz. Es werden neue Datensätze gebildet, die dieselbe Größe wie der ursprüngliche Datensatz haben, und ähnliche statistische Eigenschaften. Dabei kann es vorkommen, dass einzelne Beobachtungen aus dem Quelldatensatz einfach, mehrfach oder gar nicht vorkommen. Durch wiederholte Erzeugung solcher Bootstrap-Samples und der Ermittlung der einzelnen Bootstrap-Schätzer kann nun ein validierter Schätzer der Prognosegüte gewonnen werden.

Fehlerbeschreibung:

Um bei mehrmaliger Durchführung der Bootstrap-Validierung die gleichen Ergebnisse zu erhalten, kann beim Aufruf des Makros eine positive Zahl als Seed übergeben werden. Dadurch erhält man allerdings auch immer die gleichen Bootstrap-Samples und dadurch nur einen Satz Bootstrap-Schätzer. Werden für die Validierung beispielsweise 200 Samples verlangt, so erhält man 200 identische Samples. Eine sinnvolle Schätzung der Parameter ist damit nicht möglich.

Fehlerbehebung:

Bisher wird in einer Schleife jeweils ein neues Sample mit dem angegebenen Seed erzeugt und ausgewertet. Mit Hilfe der PROC SURVEYSELECT außerhalb dieser Schleife werden nun stattdessen alle Samples auf einmal erzeugt und dann in Teildatensätze aufgespalten.

Der zugehörige Code in den Hilfsmakros BOOT_MEAN_VALIDATION und BOOT_MODEL_VALIDATION jeweils ab Zeile 214 sieht wie folgt aus:

```
/* Gesamten Bootstrap-Datensatz erzeugen */
PROC SURVEYSELECT DATA=&boot OUT=boot_all
```

```
      seed=&random
      method=urs
      samprate=1
      outhits
      rep=&anzahl_samples
      NOPRINT;
RUN;

%DO km=1 %TO &anzahl_samples;
      DATA boot&km (DROP=Replicate);
      SET boot_all (WHERE=(Replicate=&km));
      RUN;
 /* Auswertung des aktuellen Bootstrap-Datensatzes */
%END;
```

1.4.5 PM_SHRINKAGE_VALIDATION.MAC.SAS

Dieses Verfahren ist ein internes Validierungsverfahren, es handelt sich um eine so genannte Kalibrierung. Es wird versucht einen Überoptimismus bei Bestimmung der Regressionskoeffizienten zu verhindern. Dazu werden die beobachteten Werte mit den im Modell vorhergesagten Werten aufgetragen und die Steigung der Geraden beobachtet. Bei gleichen Datensätzen beträgt diese 1, sobald allerdings ein anderer Datensatz angewandt wird, ist die Geradensteigung üblicherweise kleiner 1. Diese Steigung (Shrinkage) kann zur Korrektur der Regressionskoeffizienten, und demnach auch der Prognosegüte, herangenommen werden.

Fehlerbeschreibung:

Die berechneten Größen AUC und SomersD werden aus dem Makro PM_LOGREG.MAC.SAS übernommen und enthalten somit die gleichen numerischen Abweichungen, die durch die Korrektur der Binwidth behoben werden. Außerdem gibt es jedoch weitere Abweichungen bei S_Heur und S_Global, die trotz geänderter Binwidth weiterbestehen. Außerdem werden bei exakter Berechnung fast keine Werte mehr ausgegeben.

Fehlerbehebung:

Die Funktion des Makros ist durch die numerischen Abweichungen und das Fehlen der exakten Berechnung stark eingeschränkt. Um diese Probleme zu lösen ist vermutlich eine generelle Überarbeitung nötig, die aufgrund der Komplexität der Programmierung recht umfangreich und zeitaufwendig ist. Das Makro PM_SHRINKAGE_VALIDATION.MAC.SAS wird deshalb zumindest vorerst nicht weiter unterstützt und aus dem Makropaket entfernt.

1.5 ZUSAMMENFASSUNG UND AUSBLICK

Das Makropaket wurde in SAS Version 8 geschrieben und getestet. Im Jahr 2005 wurde sie zudem in SAS Version 9.1 getestet und gegebenenfalls korrigiert. Meine Aufgabe war dieses auf SAS Version 9.4 zu testen und eventuelle Fehler zu korrigieren. Einige Fehler waren sofort gefunden, andere wiederum waren etwas kniffeliger zu entdecken und zu beheben. Das Testen der Makros bezog sich während meines Praktikums auch nur auf die Hauptmakros, die vom Benutzer selbst aufgerufen werden. Im Folgenden findet sich eine Übersicht über die einzelnen Makros mit einer kurzen Bemerkung über eventuell aufgetretene Fehler.

Makroname	**SAS Version 9.4**
DESCRIPTION	Graphiken nicht im angegebenen Format und Ordner
MULTICOLLIN	Spearman Rangkorrelationskoeffizient wird für stetige Variablen jeweils mit sich selbst berechnet
MISSING	Keine Fehler aufgetreten
MI_ANALYZE	Keine Fehler aufgetreten
INFLUENCE	Keine Fehler aufgetreten
UNI_LOGREG	Keine Fehler aufgetreten
LOGREG	Numerische Abweichungen in “Association of Predicted Probabilities and Observed Responses” Speichert Graphik einer ROC-Kurve, obwohl nicht in der Dokumentation vermerkt
GOF	Numerische Abweichungen bei “Partition for the Hosmer and Lemeshow Test”
ROC	Zwei der Graphiken enthalten Beschriftungen mehrmals
EXTERNAL_VALIDATION	Keine Fehler aufgetreten
DATASPLITTING	Keine Fehler aufgetreten
CROSSVALIDATION	Keine Fehler aufgetreten
BOOTSTRAP_VALIDATION	Bei angegebenem Seed wird mehrmals der gleiche Bootstrap-Datensatz erzeugt
SHRINKAGE_VALIDATION	Numerische Abweichungen bei AUC, SomersD, S_Heur und S_Global Exakte Berechnung ist nicht mehr funktionsfähig

Die Hilfsmakros, auf die diese Makros zugreifen, kommen teilweise von externen Stellen, anderen Universitäten etc. Sie sind zum Teil in SAS Version 6 geschrieben und stellen daher die Frage, ob die Kompatibilität zu SAS Version 9.4 gewährleistet werden kann. In meinen Tests sind keine Versionskonflikte aufgefallen, allerdings muss das nicht heißen, dass es diese nicht gibt. Vor dem unbedachten Gebrauch der Makros sollte also geklärt werden, ob die Untermakros zu SAS Version 9.4 kompatibel sind. Hier können in den meisten Fällen die Autoren der entsprechenden Makros am besten Auskunft geben.

Bei der Gegenüberstellung der Outputs in Version 9.4 und der Referenz in der Habilitationsschrift [1] fiel mir mehrmals auf, dass ein direkter Vergleich nicht möglich ist. Bei den Makros PM_MISSING und PM_BOOTSTRAP_VALIDATION wurde beispielsweise in der Makrovariablen „random“ kein Seed übergeben, so dass sich die Ergebnisse zufallsbedingt immer unterscheiden. Außerdem wird ab Version 9.4 der Hosmer und Lemeshow Test im Makro PM_GOF mit anderer Klasseneinteilung berechnet, so dass sich auch der Output unterscheidet. Sollte meine Arbeit einmal mit einer neuen SAS Version wiederholt werden, so wäre eine neue Referenz mit angegebenen Seeds und neuer Klasseneinteilung sicher hilfreich. Ich habe deshalb die Word-Datei „SAS-Makros Referenz V9.4.docx“ erstellt, die dies beinhaltet. Anfangs ist der allgemeine SAS Code zum Einbinden der Makros und Ausgabeoptionen enthalten. Anschließend folgen die Makros in der gleichen Reihenfolge wie in der Habilitationsschrift [1], jeweils zuerst der Makroaufruf mit allen Parameterangaben und direkt darauf folgend der Output in Version 9.4. Die Parameterangaben wurden dabei aus der Habilitationsschrift übernommen und sinnvoll ergänzt, um eine bessere Vergleichbarkeit zu gewährleisten.

1.6 LITERATURVERZEICHNIS

[1] MUCHE, R.: Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression, Habilitationsschrift (2004)

[2] MUCHE, R., RING, C., ZIEGLER, C.: Entwicklung und Validierung von Prognosemodellen auf Basis der logistischen Regression – Eine Auswertungsstrategie und deren Umsetzung mit SAS-Makros, Shaker-Verlag, Aachen (2005)

[3] KUSS, O.: Global goodness-of-fit tests in logistic regression with sparse data. Statist. Med. 21: 3789 – 3801 (2002)

[4] JACOBI, E., RÖSCH, M. ALT, B.: Rehabilitationswissenschaftlicher Forschungsverbund Ulm - "Bausteine der Reha". Die Rehabilitation 37 Suppl. 2: 111 – 116 (1998)

[5] KALUSCHA, R., JACOBI, E.: Eine Datenbank zur Effektivitätsbeurteilung: Das Datenbankkonzept des rehabilitationswissenschaftlichen Forschungsverbundes Ulm. *DRV-Schriften* **20:** 218 - 219 (2000)

[6] FRIENDLY, H.: SAS-Makro dummy.sas http://www.psych.yorku.ca/friendly/lab/file/macros/dummy.sas (aufgerufen am 16.1.2004) (2001)

[7] HEINZE, G., SCHEMPER, M.: A solution to the problem of separation in logistic regression. *Statist. Med.* **21:** 2409 - 2419 (2002)

[8] SAS Support Community: Estimation of AUC with proc logistic https://communities.sas.com/t5/SAS-Statistical-Procedures/Estimation-of-AUC-with-proc-logistic/td-p/13009 (aufgerufen am 22.4.2016) (2012)

[9] SAS Knowledge Base: Usage Note 45767: Computing the statistics in "Association of Predicted Probabilities and Observed Responses" table
http://support.sas.com/kb/45/767.html (aufgerufen am 22.4.2016) (2016)

[10] SAS/STAT(R) 14.1 User's Guide: The LOGISTIC Procedure - The Hosmer-Lemeshow Goodness-of-Fit Test
http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_logistic_details32.htm (aufgerufen am 27.5.2016) (2015)