Ein SAS-Makro für die Bootstrap-Validierung logistischer Prognosemodelle

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

Zusammenfassung

Prognosen sind in der Medizin ein wichtiges Instrument zur Entscheidungsfindung. Einmal kann der Patient bei Kenntnis seiner Prognose entsprechend mit entscheiden bei der Therapiewahl, aber auch der Arzt braucht bei der Auswahl von Therapien entsprechende Informationen. Die Prognose beruht meist auf der Kenntnis vieler Faktoren, so dass oft multiple Regressionsmodelle als Prognosemodelle benutzt werden. Da die Zielgröße oft dichotom ist (z.B. Heilung ja/nein), werden in diesem Zusammenhang in der Regel multiple logistische Regressionsmodelle eingesetzt. Die Prognosegüte solcher Modelle wird meist durch Reklassifikation bestimmt. Dabei werden die Patientendaten in das Modell eingesetzt, die Wahrscheinlichkeit für das Zielereignis berechnet und mit den beobachteten Ereignissen verglichen. Es werden einige Parameter für diese Prognosegüte in der Literatur angegeben. Bei diesem Vorgehen sind diese Gütemaße in der Regel zu optimistisch, da die Regressionskoeffizienten mit der ML-Methode optimal für den Datensatz geschätzt wurden. Für eine geplante Anwendung dieses Prognosemodells auf neue Daten ist daher die realistische Prognosegüte zu bestimmen. Eine Variante dieser Modellvalidierung lässt sich mit der Bootstrap-Methode durchführen. Dabei werden durch Ziehen mit Zurücklegen aus dem Originaldatensatz „neue“ Datensätze erzeugt, anhand derer durch Refit des Modells eine empirische Verteilung der Gütemaße erzeugt und der Optimismus geschätzt werden kann.

Keywords: Logistische Regression, Bootstrap, Modellvalidierung

1 Einleitung

„Prognose ist eine Vorhersage über den zukünftigen Verlauf einer Krankheit nach ihrem Beginn“ [4]. Genaue Prognosen sind in der Medizin in vielen Situationen wichtig. Einmal kann der Patient, wenn er die Prognose kennt, eigenständige Entscheidungen treffen bzw. kompetent mitreden über zukünftige Therapien. Mediziner verwenden die Kenntnis der Prognose, um zusätzliche diagnostische Tests anzufordern oder bestimmte Therapien einzuleiten. Ausserdem können Prognosen auch zur Evaluation neuer Therapien genutzt werden, indem sie als Zielgröße zum Vergleich in einer Studie hergenommen

wird. Kennt man Prognosen, können diese auch als Ein- und Ausschlusskriterien bzw. als Schichtkriterium in klinischen Studien dienen.

Die Prognosen werden oft auf Grundlage von Prognosemodellen geschätzt. Dabei definieren Wyatt und Altman ein Prognosemodell folgendermaßen: „Prognostic models are more complex tools for helping decision making that combine two or more items of patient data to predict clinical outcomes.“ [8]. Zum Einsatz kommen in der Regel multiple Regressionsmodelle. Hier wird das Augenmerk auf das logistische Regressionsmodell gelegt, welches zur Modellierung dichotomer Zielgrößen standardmäßig eingesetzt wird [6]. Das Ziel der Prognosemodellierung ist eine genaue Prognose zukünftiger Ereignisse, nicht unbedingt die Untersuchung der Kausalbeziehung / Ätiologie der Krankheit. Das Modell sollte die Zielgröße gut beschreiben (Modell-Fit), mehr Wert wird allerdings auf die Optimierung der Prognosegüte gelegt.

2 Prognosegüte

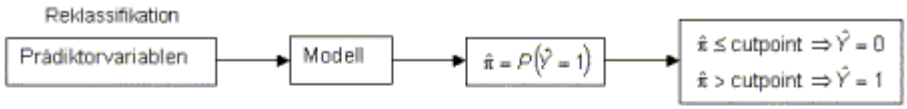
Unter der Prognosegüte wird verstanden, wie gut die Zielgröße eines Patienten durch das Modell vorhergesagt werden kann. Das Ziel dabei ist es, möglichst wenig Fehler zu machen bei dieser Prognose, da dies ernste Konsequenzen für den Patienten z.B. durch eine falsche Therapieentscheidung nach sich ziehen würde.

Die Prognose auf Grundlage des Modells wird anhand der Daten folgendermaßen im Rahmen einer Reklassifikation bestimmt. Durch das Einsetzen der Patientendaten berechnet man auf Grundlage der logistischen Regression die Wahrscheinlichkeit für das Outcome. Die Durchführung der Modellberechnung für das logistische Regressionsmodell kann in SAS unter anderem (hauptsächlich) mit der Prozedur PROC LOGISTIC erfolgen. Ein beispielhafter Aufruf dieser Prozedur lautet folgendermaßen:

```
PROC LOGISTIC DATA=reha ;
  CLASS au (PARAM=REF REF='<3 Mo. AU');
  MODEL event(EVENT=LAST)= alter au
        / CLODDS=pl RSQUARE;
RUN;
```

Nach Berechnung dieser (Modell-)Wahrscheinlichkeiten auf Grundlage der Patientendaten können diese mit den beobachteten Ereignissen verglichen werden. Dazu kann z.B. direkt der Rangkorrelationskoeffizient zwischen einer dichotomen und einer stetigen Variablen (Somer’s D) berechnet werden. Meist werden aber nach Wahl eines Grenzwertes (Cut-Point) die Modellwerte klassiert (Wert größer Cut-Point: Prognose Outcome=1 ; Wert kleiner Cut-Point: Prognose Outcome=0). In Kombination mit den beobachteten Werten erhält man so eine Vierfeldertafel, anhand der die bekannten Gütemaße Sensitivität,

Spezifität und Youden-Index (= Sensitivität + Spezifität -1) berechnet werden können.



Classification-Table

	$Y = 0$	$Y = 1$	
$\hat{\pi} \leq$ cutpoint $\Rightarrow \hat{Y} = 0$	a	b	$a + b$
$\hat{\pi} >$ cutpoint $\Rightarrow \hat{Y} = 1$	c	d	$c + d$
	$a + c$	$b + d$	n

Diese Gütemaße sind allerdings direkt abhängig vom gewählten Cut-Point. Durch Variation der Cut-Points erhält man die ROC-Kurve. Ein allgemeines, Cut-Point unabhängiges Prognosegütemaß leitet sich daraus ab: AUC (= Fläche unter der ROC-Kurve). Dieses Maß liegt zwischen 0 und 1, wobei 1 optimal ist. Harrell [5] beschreibt u.a. noch zwei weitere Gütemaße: den Brier Score und Emax, auf die hier nicht eingegangen werden kann.

3 Modellvalidierung

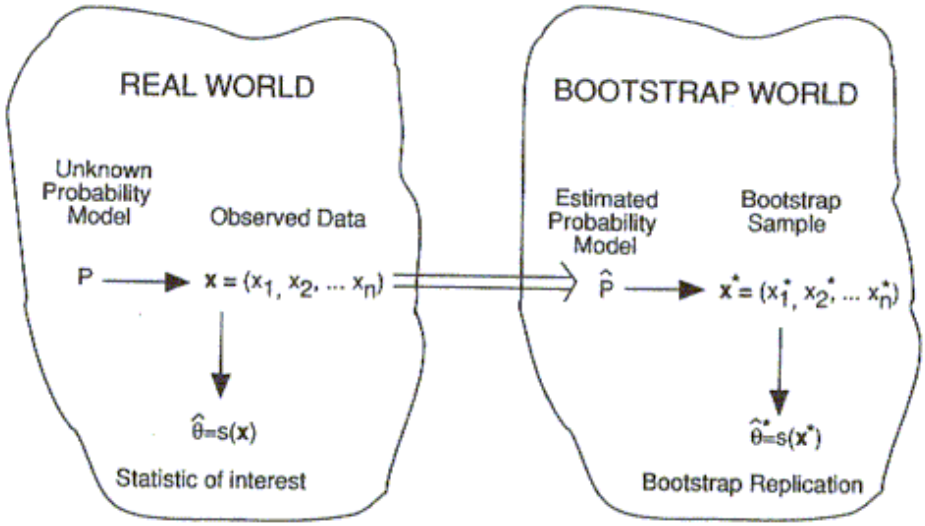
Bei der Nutzung der so berechneten Prognosegüte für ein Modell ergibt sich allerdings ein Problem: diese Werte sind als Gütemaße für die Nutzung des Modells an neuen, unabhängigen Daten zu optimistisch geschätzt. Dieser Überoptimismus kommt dadurch zustande, dass das Modell durch die Schätzung der Regressionskoeffizienten mit der Maximum-Likelihood-Methode optimal an die Daten angepasst ist und demnach die Untersuchung der Prognosegüte anhand derselben Daten „optimal“ wird.

Deshalb sollte vor Einsatz eines solchen Prognosemodells auf neue Daten dieser Optimismus geschätzt und eventuell die Prognosegüte korrigiert werden. Dies kann anhand einer Modellvalidierung erfolgen. Es werden verschiedene Validierungsstrategien in der Literatur vorgeschlagen [5,7]. Als Golden Standard gilt die externe Validierung anhand eines zweiten, unabhängigen Datensatzes. Dieser steht aber häufig bei der Modellentwicklung nicht zur Verfügung. So kommen sogenannte interne Validierungsverfahren zum Einsatz, die auf Grundlage des vorhandenen Datensatzes diesen Überoptimismus korrigieren sollen.

In einer vergleichenden Arbeit von Steyerberg et al. aus dem Jahre 2001 [7] werden 8 Varianten für die interne Prognosemodellierung anhand der logistischen Regression verglichen. Die Autoren kommen zum Schluss, dass die interne Validierung anhand der Bootstrap-Validierung im Allgemeinen die besten Ergebnisse liefert.

4 Das Bootstrap-Verfahren zur Modellvalidierung

Der Begriff des Bootstrap in der Statistik leitet sich ab von der Redewendung „to pull oneself up by one's bootstrap" und geht auf die „Abenteuer des Baron von Münchhausen" von Rudolf Erich Raspe zurück. Das Hauptprinzip dabei ist es, durch Ziehen mit Zurücklegen (mit gleicher Fallzahl) aus dem Originaldatensatz viele ähnliche Datensätze zu erzeugen, in denen die Auswertung wiederholt und so eine Verteilung der Ergebnisse erzeugt wird [3]. Auf Grundlage dieser Verteilung können so z.B. inferenzstatistische Aussagen getroffen werden.



Das allgemeine Vorgehen bei der Modellvalidierung anhand des Bootstrap-Verfahrens kann in folgende Schritte eingeteilt werden [3,5]:

- Berechnung der Prognosegüte am Originaldatensatz (OD)
- B Bootstrap-Samples mit Ziehen mit Zurücklegen vom Umfang n aus dem Originaldatensatz erzeugen
- in jedem Bootstrap-Sample die Modellparameter neu schätzen (Refit des Modells)
- Anwendung der Bootstrap-Modelle jeweils auf den Originaldatensatz und Berechnung der Prognosegütemaße, Zusammenfassung durch Mittelwertbildung über die B Ergebnisse (**simple Bootstrap**)
- Anwendung der Bootstrap-Modelle jeweils auf das Bootstrap-Sample (Reklassifikation) und Berechnung der Prognosegütemaße
- Ein Schätzer für den Überoptimismus kann bestimmt werden aus der Differenz der jeweiligen Prognosegütemaße der Bootstrap-Modelle bestimmt am Bootstrap-Sample und dem Originaldatensatz
- Das „**enhanced Bootstrap**“ – Ergebnis wird durch Abzug des geschätzten Überoptimismus von der Ausgangsprognosegüte bestimmt.

5 Umsetzung in SAS

Das beschriebene Vorgehen wurde in einem SAS-Makro programmiert. Die Erzeugung der Bootstrap-Stichproben erfolgt dabei mit folgendem Teilprogramm [2]:

```
%macro boot;
%do i=1 %to 100;
data boot&i (drop=j);
      do j=1 to 841;
         select=ceil(ranuni(0)*total);
                  set rehadaten point=select nobs=total;
                  output;
          end;
          stop;
run;
%end;
%mend;

%boot;
```

Zur Durchführung der Modellvalidierung ist im Programm eine Schleife über PROC LOGISTIC und ROC-Analysen vorzunehmen. Der Ablauf des Makros wird in der folgenden Abbildung skizziert [1]:

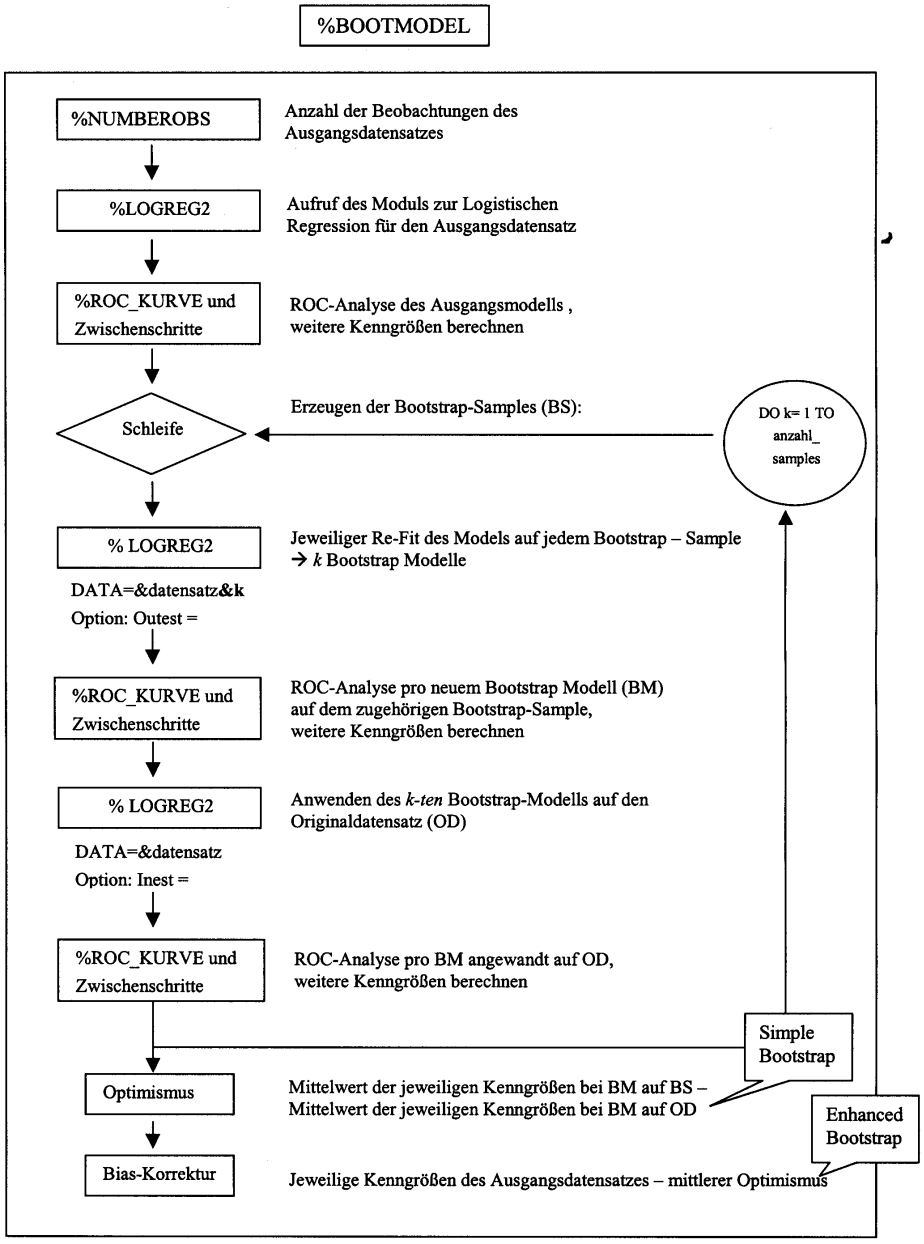


Abbildung 8: Schematischer Ablauf des Makros %BOOTMODEL

Die notwendigen Parameter des Aufrufes des Makros %BOOT lauten wie folgt:

```
%Boot ( variante      = model,      enhanced Bootstrap
        datensatz     = ,           Originaldatensatz
        cvar          = ,           klassierte Variablen
        ref           = ,           Referenzkategorien der klass. Variablen
        xvar          = ,           stetige Variable
  Anzahl_samples      = ,           Anzahl Bootstrap-Samples
        resp_var      = ,           Response-Variable
        event         = 1,          Event-Kategorie für Response-Variable
        miss          = );          0/1 - Missings raus
```

Insgesamt kann die Auswertung mit dem Makro durch 49 Parameter gesteuert werden. Die meisten Parameter können zum Fine-Tuning der zwei eingegliederten Untermakros %LOGREG2 (Logistische Regression (z.B. Variablenselektion, Odds ratios)) und %ROC-KURVE (ROC-Analyse (z.B. Grafiken ROC-Kurve, Konfidenzband Makro Hilgers, AUC Makro Koenig)) genutzt werden. Ausserdem sind zwei weitere, hier nicht dargestellte Bootstrap-Varianten integriert [1].

Ein Beispiel-Output [1] wird in der nächsten Abbildung gezeigt. Es werden per Default die Cut-point abhängigen Gütemaße Sensitivität, Spezifität und Youden-Index zum Cut-point beim maximalen Youden-Index im Originaldatensatz angegeben. Als Cut-point unabhängige Maße werden AUC, Somer's D, Brier Score und Emax berichtet. Dabei wird jeweils (in der Reihenfolge) das Gütemaß nach Reklassifikation im Originaldatensatz (_ori), das Maß des Bootstrap-Modells angewandt auf den Bootstrap-Sample (_boot) und auf den Originaldatensatz (_new model), der geschätzte Optimismus (dif_) und der korrigierte Wert nach dem enhanced Bootstrap (d_) dargestellt.

Enhanced und Simple Bootstrap 11:33 Tuesday, January 7, 2003
Optimismus und Bias-Korrektur der ROC-Kenngrößen

Obs	sens_ori	sens_ boot	sens_ NewModel	dif_sens	d_sens	spez_ori	spez_ boot	spez_ NewModel
1	0.90654	0.89114	0.84192	0.049229	0.85731	0.71808	0.73279	0.72621

Obs	dif_spez	d_spez	maxyouden_ ori	youden_ boot	youden_ NewModel	dif_ youden	d_youden
1	.006583522	0.71149	0.62462	0.62394	0.56813	0.055812	0.56881

Enhanced und Simple Bootstrap 11:33 Tuesday, January 7, 2003
Optimismus und Bias-Korrektur der Accuracy-Indizes nach Harrell

Obs	auc_ori	auc_boot	auc_new Model	dif_auc	d_auc	som_ori	som_boot
1	0.87786	0.89129	0.86154	0.029747	0.84811	0.75572	0.78258

Obs	som_new Model	dif_som	d_som	brier_ ori	brier_ boot	brier_ newModel	dif_brier
1	0.72308	0.059495	0.69623	0.091954	0.087153	0.096713	.009560418

Obs	d_brier	emax_ori	emax_boot	emax_ newModel	dif_emax	d_emax
1	0.082394	.000016026	.000096479	0.11501	0.11492	0.11490

Abbildung 7: Beispiel-Output des Makros %BOOTMODEL

6 Zusammenfassung

Zusammenfassend kann festgestellt werden, dass eine Validierung von Prognosemodellen dringend notwendig ist, um nicht mit überschätzter Güte die Modelle in einen klinischen Alltag einzubringen. Die Abschätzung realistischer Fehlerraten ist bei der Anwendung der Modelle zwingend notwendig.

Wie sich in mehreren Untersuchungen gezeigt hat ist der Ansatz der Modellvalidierung über Bootstrap-Methoden alternativen Verfahren in vielen Bereichen überlegen. Da die Computer immer schneller werden und auch der Speicher kaum Probleme mehr bereitet, kann dieses Verfahren ohne Probleme um- und eingesetzt werden.

Wie von uns durchgeführte Literatur-Recherchen zeigten, sind bisher keine SAS-Lösungen für die Bootstrap-Modellvalidierung angeboten worden. Das hier beschriebene Makro eignet sich zur Validierung von Prognosemodellen auf der Basis des logistischen Regressionsmodells. Einige Untersuchungen zur Softwarevalidierung unseres Makros sind schon durchgeführt worden, die meisten Untersuchungen stehen allerdings noch aus, so dass das Makro in einiger Zeit nach Abschluss der Untersuchungen und nach einer genauen Dokumentation vom Autor zu erhalten (rainer.muche@medizin.uni-ulm.de) ist.

Literatur

1. Assfalg I. (2003) *Die Bootstrap-Methode zur internen Validierung von Prognosemodellen*. Diplomarbeit FH Ulm
2. Boudreaux D., Cranford K. (1995) Simple random sampling and subsetting strategies using SAS Software. *Observations*, 34-40
3. Efron B., Tibshirani R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York
4. Fletcher R.M., Fletcher S.W., Wagner E.H. (1999). *Klinische Epidemiologie*. Ullstein Medical Verlag, Wiesbaden
5. Harrell F.E. Jr. (2001) *Regression Modeling Strategies*. Springer Verlag, New York
6. Hosmer D.W., Lemeshow S. (2000) *Applied Logistic Regression (2nd Edition)*. John Wiley & Sons, New York
7. 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
8. Wyatt J.C., Altman D.G. (1995) Commentary: Prognostic models: clinical useful or quickly forgotten? *BMJ*, 311, 1539-1541