Bootstrap-Korrektur von Somer's D im logistischen Regressionsmodell

Tim Friede
Abteilung Medizinische Biometrie
Universität Heidelberg
Im Neuenheimer Feld 305, 69120 Heidelberg
eMail : friede@imbi.uni-heidelberg.de

Zusammenfassung

Somer's D ist eine Korrelationsstatistik, die im logistischen Regressionsmodell ein Maß für den Zusammenhang zwischen den beobachteten Werten und den Modellwerten ist. Bei datenabhängiger Wahl der Einflußgrößen wird dieser Zusammenhang durch Somer's D überschätzt. Es wird ein Verfahren vorgestellt, daß die beschriebene Verzerrung von Somer's D korrigiert. Insbesondere wird auf die Umsetzung des Verfahrens in SAS eingegangen. Ein Beispiel illustriert das Vorgehen.

1. Einleitung

In der Epidemiologie sowie in der Biometrie wird die logistische Regression zur Modellierung dichotomer Outcome-Variablen wie z.B. Heilungserfolg oder Todesfall benutzt. Da sich die geschätzten Regressionsparameter als (adjustierte) Odds ratios interpretieren lassen, erhält man einen anschaulichen Zugang zu diesen Modellen.

Häufig ist unklar, welche Einflußgrößen letzlich für das jeweilige Modell von Interesse sind, so daß eine Variablenselektion anhand des erhobenen Datensatzes durchgeführt wird. Dies führt dazu, daß Statistiken, die über die Güte der Modellanpassung Auskunft geben sollen, verzerrt geschätzt werden. Im logistischen Regressionsmodell ist Somer's D eine Statistik zur Beschreibung der Korrelation zwischen den beobachteten Werten und den Modellwerten der abhängigen Zufallsvariable. Somer's D ist somit eine interessante Statistik im Rahmen der Modelldiagnostik. Führt eine von den vorliegenden Daten beeinflußte Variablenselektion zum endgültigen Modell, so überschätzt Somer's D die wahre Korrelation. Durch Bootstrapping kann eine Schätzung der genannten Statistik so korrigiert werden, daß sie annähernd unverzerrt ist.

Im Abschnitt 2 wird zunächst das logistische Regressionsmodell kurz dargestellt. Anschließend wird im Abschnitt 3 Somer's D definiert. Nach einführenden Bemerkungen zum Bootstrap wird im Abschnitt 4 der Algorithmus zur Bootstrap-Korrektur von Somer's D vorgestellt. Im Abschnitt 5 folgen einige Hinweise zur Umsetzung des Algorithmus' in SAS. Das Verfahren wird dann anhand eines Beispiels im Abschnitt 6 illustriert. Den Schluß bildet eine kurze Diskussion.

2. Die logistische Regression

Im folgenden geht es um die Situation, daß eine dichotome Zielgröße durch eine gewisse Anzahl von Einflußgrößen modelliert werden soll. Als Beispiel sei hier der Fall genannt, daß die dichotome Größe Depressionen mit ihren Ausprägungen „vorhanden" und „nicht vorhanden" durch die Einflußgrößen Geschlecht mit den Ausprägungen „männlich" und „weiblich" und Alter in Jahren modelliert werden soll.

Ohne Einschränkung gehen wir davon aus, daß die dichotome Zielgröße Y die Ausprägungen 0 und 1 annehmen kann und daß wir an einer Modellierung der Wahrscheinlichkeit für $Y = 1$ interessiert sind. Den Vektor der Einflußgrößen für die Beobachtung i bezeichnen wir mit x_i. Wir schreiben dann: $p_i := P(Y = 1 \mid x_i)$.

Das logistische Regressionsmodell sieht mit den eingeführten Bezeichnungen wie folgt aus:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + x_i'\beta. \tag{1}$$

Mit den Schätzern $\hat{\beta}_0$ und $\hat{\beta}$ für die Parameter β_0 und β ergibt sich für die Modellwerte $\hat{p}_i$:

$$\text{logit}(\hat{p}_i) = \hat{\beta}_0 + x_i'\hat{\beta} \tag{2}$$

Schon in der Einleitung ist bemerkt worden, daß sich die Parameter β als adjustierte Odds Rations interpretieren lassen. Dazu müssen die Parameter lediglich mit Hilfe der Eulerschen *e*-Funktion transformiert werden.

Ein weiterer Zugang zur logistischen Regression drängt sich dem Leser auf, wenn er sich die Regressionsgleichung für lineare Modelle vergegenwärtigt und mit (1) vergleicht.

3. Somer's D

Bei Somer's D handelt es sich um eine Rangstatistik, die die Korrelation zwischen den beobachteten Werten y_i und den Modellwerten $\hat{p}_i$ quantifiziert. Um Somer's D definieren zu können, müssen zunächst die Begriffe *konkordante* und *diskordante* Paare eingeführt werden.

Es sei $\hat{p}_F$ ein zu einer Beobachtung mit $Y = 1$ gehörender Modellwert und $\hat{p}_K$ ein zu einer Beobachtung mit $Y = 0$ gehörender Modellwert. Das Paar $(\hat{p}_K, \hat{p}_F)$ heißt *konkordant*, wenn $\hat{p}_F > \hat{p}_K$. Im Fall $\hat{p}_F < \hat{p}_K$, nennen wir es *diskordant*.

Nun können wir Somer's D wie folgt definieren:

$$D = \frac{c - d}{n_1 \cdot n_2}, \tag{3}$$

wobei c die Anzahl der konkordanten Paare, d die Anzahl der diskordanten Paare, n_1 die Anzahl der Beobachtungen mit $Y = 1$ und n_2 die Anzahl der Beobachtungen mit $Y = 0$ ist. Das Produkt $n_1 \cdot n_2$ ist dann die Anzahl aller paarweisen Vergleiche von Modellwerten, die zu Beobachtungen mit $Y = 1$ gehören, und solchen, die zu Beobachtungen mit $Y = 0$ gehören.

4. Korrektur von Somer's D durch Bootstrapping

4.1 Das Problem: Datenabhängige Variablen-Selektion

Bei der Modellbildung besteht der Wunsch, nur die Einflußgrößen ins Modell aufzunehmen, die „wesentlich" sind. Um diesem Wunsch nahe zu kommen, läßt man sich bei der Modellwahl, insbesondere bei geringen a priori Informationen, durch die vorliegenden Daten bei der Modellwahl leiten. Man spricht dabei von datenabhängiger Selektion (*data driven selection*).

Dabei ergibt sich nun folgendes Problem. Durch die datenabhängige Selektion der Einflußgrößen kommt es zu einer verzerrten Schätzung mit positivem Bias, d.h. daß die tatsächliche Korrelation zwischen den beobachteten Werten und den Modellwerten überschätzt wird.

4.2 Die allgemeine Idee des Bootstraps

Bootstrapping ist eine Resampling-Methode. Resampling-Methoden zeichnen sich dadurch aus, daß aus einer Stichprobe (*sample*) neue Stichproben gezogen werden. Beim Bootstrapping zieht man mit Zurücklegen aus der Ausgangsstichprobe eine Stichprobe, deren Umfang genauso groß ist wie der der Ausgangsstichprobe. Wenn man B Bootstrap-Stichproben zieht und für jede Bootstrap-Stichprobe die interessierende Statistik berechnet, erhält man so einen Schätzer für die Verteilungsfunktion der Statistik.

4.3 Algorithmus für die konkrete Situation

Algorithmus (nach Harrell et al. (1996)):

1. Führe die Variablenselektion am Orginaldatensatz durch und berechne für das gewählte logistische Regressionsmodell Somer's D (D_{app}).
2. Ziehe eine Bootstrap-Stichprobe (getrennt für Fälle und Kontrollen).
3. Wähle nach den gleichen Regeln wie oben die Variablen aus.
4. Berechne Somer's D (D_{boot}).
5. Berechne Somer's D für den Orginaldatensatz, aber mit den soeben gewählten Variablen (D_{orig}).
6. Berechne die „Überschätzung“ $D_{boot} - D_{orig}$.
7. Wiederhole die Schritte 2-6 B -mal ($B \geq 100$).
8. Berechne die mittlere „Überschätzung“ $O = \frac{1}{B} \sum D_{boot} - D_{orig}$.
9. Berechne das korrigierte Somer's D: $D_{corr} = D_{app} - O$.

5. Umsetzung in SAS

Das beschriebene Verfahren zur Bootstrap-Korrekur von Somer’s D im logistischen Regressionsmodell ist vom Autor in das SAS-Makro DBOOT umgesetzt worden. Dem Leser steht dieses Makro auf dem Server des Universitätsrechenzentrums Heidelberg zur Verfügung (web.urz.uni-heidelberg.de/veranstaltungen/ksfe/makros).

In diesem Abschnitt sollen exemplarisch verschiedene Bestandteile des Makros erläutert werden. Für die logistische Regression wird die SAS-Prozedur LOGISTIC verwendet. Diese Prozedur bietet verschiedene Selektionsregeln an, die genutzt werden. Die Selektionsregel wird als Option im MODEL-Statement angegeben. Es stehen dabei vier Regeln zur Wahl, nämlich FORWARD für Vorwärtsselektion, BACKWARD für Rückwärtsselektion, STEPWISE und SCORE. Nähere Informationen zur Variablenselektion findet der Leser in SAS/STAT (6.12).

Ein Nachteil entsteht bei der Verwendung von PROC LOGISTIC. Somer’s D wird zwar berechnet und erscheint in der Ausgabe, läßt sich aber nicht in einen Data Set schreiben (Stand: Release 6.12). Deshalb wurde die Berechnung von Somer’s D in SAS/IML programmiert.

Interessant und deshalb erwähnenswert ist die Implementierung des Ziehens einer Bootstrap-Stichprobe. Nicht zuletzt wegen der Anwendbarkeit in ganz verschiedenen Bootstrap-Verfahren soll der Programmcode an dieser Stelle demonstriert werden. Das Ziehen einer Bootstrap-Stichprobe wurde als Makro realisiert, dessen wesentlicher Bestandteil der Aufruf von SAS/IML darstellt.

```
/*** Makro zum Ziehen einer Bootstrap-Stichprobe ***/

%macro bsample(file);
proc iml;
  use &file;
  read all var _all_ into daten [colname=varname];
  close &file;
  n=nrow(daten);
  m=ncol(daten);
  bsample=j(n,m,.);
  random=j(n,1,.);
  random[1:n,]=ceil(n*ranuni(j(n,1,-1)));
  bsample=daten[random,];
  create bsample from bsample [colname=varname];
  append from bsample;
quit;
%mend bsample;
```

An das Makro BSAMPLE wird als Parameter der Name des Data Sets übergeben, der die Ausgangsstichprobe enthält, aus der die Bootstrap-Stichprobe gezogen werden soll. Das Makro erzeugt dann eine mit Zurücklegen gezogene Stichprobe, die in den Data Set WORK.BSAMPLE geschrieben wird.

6. Beispiel: Depressions-Studie

6.1 Angaben zur Studie

Als Beispiel wird eine Studie verwendet, deren Ziel die Schätzung der Prävalenz von Depressionen in der erwachsenen Bevölkerung war. Die Studie wurde im Sommer 1979 in Los Angeles County durchgeführt. Insgesamt wurden 1003 Personen, die repräsentativ für die gesamte erwachsene, seßhafte und nicht-inhaftierte Bevölkerung waren, von professionellen Interviewern befragt. Ob Depressionen bei der jeweiligen Person vorlagen wurde anhand des CES-D-Indexes, der eine Selbstbeurteilungsskala ist, festgestellt. Bei einem CES-D von 16 und mehr Punkten wurde angenommen, daß Depressionen vorlagen.

Durch die Studie wurde die Prävalenz von Depressionen in der erwachsenen Bevölkerung auf 19% geschätzt. Es stellte sich heraus, daß das Einkommen einen wichtigen Einflußfaktor darstellt. Für weitere Details sei der Leser auf Frerichs et al. (1981) verwiesen.

Das Buch von Afifi & Clark (1996) enthält als Beispieldatensatz eine zufällige Auswahl von 294 Personen aus der oben beschriebenen Studie. Dieser Datensatz wurde zur Illustration der Bootstrap-Korrektur von Somer's D im logistschen Regressionsmodell verwendet. Dabei wurden die in Tabelle 1 genannten Einflußgrößen, die zum Teil durch Vereinfachungen erhobener Merkmale seitens des Autors entstanden sind, modelliert.

Tabelle 1 Modellierte Einflußgrößen auf Depressionen

Variable	Bedeutung	Ausprägungen
SEX	Geschlecht	Männlich / Weiblich
AGE	Alter	In Jahren
INCOME	Einkommen	In 1000 $
DRINK	Alkoholiker	Ja / Nein
HEALTHY	Allg.Gesundheitszustand	Gut / Schlecht
REGDOC	Hausarzt	Ja / Nein
TREAT	Verordnete Therapie	Ja / Nein
BEDDAYS	Tag im Bett verbracht	Ja / Nein
ACUTEILL	Akute Krankheit	Ja / Nein
CHRONILL	Chronische Krankheit	Ja / Nein

6.2 Ergebnisse der Bootstrap-Korrektur von Somer's D

Um das beschriebene Verfahren zu veranschaulichen, werden für zwei Selektionsregeln im logistischen Regressionsmodell die Ergebnisse der Bootstrap-Korrektur von Somer's D dargestellt. Es sei an dieser Stelle ausdrücklich daraufhingewiesen, daß es im allgemeinen nicht sinnvoll ist, verschiedene Selektionsregeln ohne inhaltliche Begründung anzuwenden. Zu Zwecken der Demonstration wollen wir uns aber darüber hinwegsetzen. Für das Modell, das alle in Tabelle 1 genannten Einflußgrößen enthält, ergibt sich ein Somer's D von 0.508.

Durch die Selektionsregel FORWARD wird das Modell mit den Einflußgrößen SEX, INCOME und BEDDAYS ausgewählt. BACKWARD selektiert die Variablen SEX, AGE, INCOME, HEALTHY und BEDDAYS.

Die Tabelle 2 enthält die Ergebnisse für die Bootstrap-Korrektur von Somer's D für die beiden Selektionsregeln FORWARD und BACKWARD mit $B = 100$ Wiederholungen. Es zeigt sich, daß in dem Beispiel der Bias, der durch die datenabhängige Variablenselektion verursacht wird, nicht unerheblich ist. Für die Vorwärtsselektion beträgt die Verzerrung 0.076 und für die Rückwärtsselektion sogar 0.094.

Tabelle 2 Ergebnisse der Bootstrap-Korrektur von Somer's D

Selektionsregel	D_{app}	O	D_{corr}
Forward	0.440	0.076	0.364
Backward0	0.496	0.094	0.402

7. Diskussion

Es wurde ein Ansatz zur Korrektur von Somer's D bei datenabhängiger Variablenselektion vorgestellt. Diese Korrektur beruht auf einem Bootstrap und wurde bereits von Harrell et al. (1996) dargelegt. In dem vorliegenden Beitrag wurde insbesondere auf die Umsetzung in SAS eingegangen und das Verfahren an einem Beispiel illustriert.

An dieser Stelle sei darauf hingewiesen, daß sich die beschriebene Vorgehensweise zur Korrektur von Somer's D im logistischen Regressionsmodell leicht auf andere Statistiken und andere Regressionsmodelle (wie z.B. Cox-Modelle für Überlebenszeiten) übertragen läßt.

Für das beschriebene Vorgehen ist es notwendig, die vorgenommene Variablenselektion zu formalisieren und evtl. zu programmieren. Solange die in SAS implementierten Selektionsregeln benutzt werden, stellt dies kein Problem dar. Wird aber nach anderen Selektionsregeln ausgewählt (z.B. Change-in-Estimate-Methoden), so kann sich die beschriebene Korrektur schwieriger gestalten.

Literatur

Afifi AA, Clark V. Computer-aided multivariate analysis. Chapman & Hall, 1996, third edition.

Frerichs RR, Aneshensel CS, Clark VA. Prevalence of depression in Los Angeles County. American Journal of Epidemiology 113 (1981); 691-699.

Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics, Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors; Statistics in Medicine 15 (1996); 361-387.

SAS Institute Inc. SAS/STAT Software: Changes and Enhancements through Release 6.12, Cary, NC