1. Konferenz der SAS-Benutzer in Forschung und Entwicklung (KSFE)

an der

Humboldt-Universität zu Berlin

20./21. Februar 1997

Ein SAS-Macro zur Einbindung kubischer Splinefunktionen ins proportionale Hazardsmodell von Cox

Harald HEINZL und Alexandra KAIDER

Abteilung für Klinische Biometrie

Institut für Medizinische Computerwissenschaften der Universität Wien

Adresse: Spitalgasse 23, A-1090 Wien, Österreich

Fax: + 43 / 1 / 40400 / 6687

E-mail: Harald.Heinzl@akh-wien.ac.at, Alexandra.Kaider@akh-wien.ac.at

Zusammenfassung:

Das proportionale Hazardsmodell von Cox ist das am häufigsten verwendete Regressionsmodell zur Analyse von Lebensdauerdaten im medizinischen Bereich. Die Einbeziehung kubischer Splinefunktionen ermöglicht eine flexible Modellierung von nicht-linearen Effekten stetiger Kovariablen und von Wechselwirkungen der Zeit mit Kovariablen. Die Verwendung weit verbreiteter Softwarepakete wie SAS oder BMDP ist problemlos möglich. Ein SAS-Macro zur Implementierung dieser Methodik wird vorgestellt. Interessant dabei ist, daß dieses Macro selbst keine Berechnungen durchführt, sondern statt dessen ein SAS-Programm erzeugt. Dies bietet den Benutzern den Vorteil, bei Bedarf kleinere Adaptionen im Programm (insbesondere beim graphischen Teil) ohne großen Aufwand durchführen zu können.

Schlüsselwörter:

SAS-Macro, PROC PHREG, PROC IML, PROC GPLOT, automatische Programmgenerierung, kubische Splinefunktionen, Analyse von Lebensdauerdaten, proportionales Hazardsmodell von Cox, nichtlineare Effekte, zeitabhängige Effekte, zeitabhängige Kovariablen

1. Einführung

Das proportionale Hazardsmodell von Cox (1972) wird im medizinischen Bereich sehr häufig zur Analyse von zensierten Lebensdauerdaten verwendet. Die Hazardfunktion des Patienten $h(t;X)$ wird in eine Baseline-Hazardfunktion $h_0(t)$ und einen Term, in den die Patientencharakteristika $X=(X_1,\ldots,X_p)$ eingehen, multiplikativ aufgefächert,

$$h(t;X)=h_0(t)\exp(X\beta),$$

dabei steht $t\geq 0$ für die Zeit. Die Regressionskoeffizienten $\beta=(\beta_1,\ldots,\beta_p)^t$ werden mittels partiellem Maximum-Likelihood (Cox 1975) geschätzt, und man bezeichnet die Hazard-Ratio

$$\frac{h\{t;(\ldots,X_j+1,\ldots)\}}{h\{t;(\ldots,X_j,\ldots)\}}=\exp(\beta_j)$$

auch als „relatives Risiko“, wenn die Variable X_j um eine Einheit erhöht wird.

Ein großer Vorteil des Cox’schen Ansatzes liegt darin, daß zeitabhängige Patientencharaktaristika direkt ins Modell aufgenommen werden können. Wir unterscheiden daher zwischen zeitabhängigen $X_1(t),\ldots,X_q(t)$ und fixen $X_{q+1},\ldots,X_p$ Kovariablen.

Wir wollen uns hier mit Möglichkeiten beschäftigen, das Cox-Modell in verschiedenen Bereichen flexibler einsetzbar zu machen:

- bei der Modellierung von kontinuierlichen fixen Kovariablen (Durrleman & Simon 1989)
- bei der Modellierung von Wechselwirkungen von fixen Kovariablen mit der Zeit, (Hess 1994)
- bei der Modellierung von Wechselwirkungen von binären zeitabhängigen Kovariablen mit der Zeit (Heinzl, Kaider & Zlabinger 1996)

Dazu verwenden wir *natürliche kubische Regressionssplines*. Diese ermöglichen es einerseits, die gewünschte Modellierungs-Flexibilität zu erreichen, andererseits können

statistische Standardsoftwarepakete, wie SAS oder BMDP, problemlos verwendet werden. Der letztgenannte Vorteil ist aber nur hypothetisch, denn bereits die Programmierung eines kubischen Splines mit nur drei Knoten in SAS/PHREG ist sehr unübersichtlich und erfordert ein großes Maß an Konzentration. Ein Umstand, der die Verbreitung und den Routineeinsatz derartiger Techniken naturgemäß hemmt.

Es erschien uns sinnvoll, ein SAS-Macro anzufertigen, welches den Benutzer möglichst optimal bei der Bearbeitung der Formeln unterstützt, ihn aber gleichzeitig nicht zu sehr in seinen Möglichkeiten einschränkt. Wir entschieden uns daher, durch das Macro ein lauffähiges SAS-Programm zu generieren, welches der Benutzer - nach Wunsch und Bedarf - frei weiterbearbeiten kann.

Neben kurzen Abschnitten, die sich mit den theoretischen Grundlagen der Modelle bzw. von kubischen Splines befassen, wollen wir im folgenden die Struktur des SAS-Macros beschreiben und seine Verwendung anhand eines Beispiels demonstrieren.

2. Mehr Flexibilität beim Cox Modell

Nehmen wir an, wir haben eine klinische Studie an insgesamt *n* Patienten durchgeführt, und deren Daten $(y_1, s_1, z_1) \ldots (y_n, s_n, z_n)$ stehen uns zur Verfügung. Dabei bezeichnet y_i die beobachtete Überlebenszeit des Patienten. Die Statusvariable s_i gibt an, ob der Patient einen Ausfall durch z.B. Tod, Rückfall nach erfolgter Heilung, oder Infektion erlitten hat (s_i =1), oder ob wir dies bei ihm nicht beobachten konnten (s_i =0), solche Beobachtungen nennen wir zensiert. Der Einfachheit halber beschränken wir uns auf eine - vorerst unspezifizierte - Kovariable *Z*.

2.1. Modellierung des nichtlinearen Effekts einer stetigen Kovariable

Seien $z_1 \ldots z_n$ die Realisationen des stetigen prognostischen Faktors *Z*. Dieser könnte das Patientenalter (in Jahren), der Serum Bilirubin-Spiegel (in mg/dl), oder der systolische Blutdruck (in mm Hg) sein. Wenn wir *Z* als Kovariable in ein Cox Modell einbringen,

$$h(t; z_i) = h_0(t) \exp(\beta z_i),$$

dann nehmen wir an, daß eine Erhöhung von *Z* um eine Einheit das Ausfallsrisiko um das $\exp(\beta)$-fache verändert. Anders ausgedrückt, die log-Hazard-Ratio-Funktion (LHR) mit Bezug auf *Z* wird als linear in *Z* angenommen:

$$\mathrm{LHR(Z)} = \log\left(h(t; Z)/h_0(t)\right) = \beta Z$$

Diese Art der Modellierung ist dann problematisch, wenn *Z* nichtlinearen Einfluß besitzt. Insbesondere bei neuen prognostischen Faktoren mit unbekannter Wirkungsweise erscheint es sinnvoller zu sein, mehr Flexibilität bei ihrer Modellierung zuzulassen,

$$\mathrm{LHR(Z)} = \log\left(h(t; Z)/h_0(t)\right) = f(Z).$$

Über die Approximation von $f(Z)$ mit Hilfe kubischer Splinefunktionen siehe Kapitel 3.

2.2. Bewertung von Wechselwirkungen der Zeit mit einer fixen Kovariable

Wir wollen jetzt annehmen, daß $z_1 \ldots z_n$ die Realisationen eines fixen prognostischen Faktors *Z* sind, welcher entweder stetiges oder dichotomes Skalenniveau besitzt. Wir wollen uns wieder die log-Hazard-Ratio-Funktion (LHR) ansehen, diesmal aber mit Bezug auf die Zeit *t*. Es liegt in der Definition des *proportionalen Hazardsmodelles*, daß LHR(t) konstant über die Zeit bleibt,

$$\mathrm{LHR(t)} = \log\left(h(t;Z)/h_0(t)\right) = \beta Z \,.$$

Eine Erhöhung von *Z* um eine Einheit hat also während der gesamten Beobachtungszeit des Patienten immer den gleichen Einfluß auf seine Ausfallsrate. Dies muß aber nicht notwendigerweise so sein. Zum Beispiel verlieren manche prognostischen Faktoren in der Onkologie ihre anfänglich sehr hohe Bedeutung im Laufe der Zeit, d.h. LHR(t) strebt gegen null. Des öfteren beobachtet man auch bei einer neuen Chemotherapie, daß sie im Vergleich zur Standardtherapie nur einen vorübergehenden Überlebensvorteil bewirkt, d.h. die beiden Überlebenskurven klaffen zu Beginn der Studie zwar auseinander, nähern sich aber dann wieder an, bzw. äquivalent formuliert: LHR(t) hat anfangs negative Werte (Vorteil für neue Therapie), kreuzt dann die Abszisse (Ausfallsrate bei beiden Therapien gleich) und ist danach positiv (es sterben nun relativ mehr Leute unter der neuen Therapie).

Für das Cox Modell bedeutet dies, daß um die proportionale Hazardseigenschaft zu erhalten, die Wirkung von *Z* zeitabhängig angesetzt werden sollte:

$$\mathrm{LHR(t)} = \log\left(h(t;Z)/h_0(t)\right) = g(t)Z$$

Über die Approximation von $g(t)$ mit Hilfe zeitabhängiger Kovariablen und kubischer Splinefunktionen siehe Kapitel 3.

2.3. Bewertung von Wechselwirkungen der Zeit mit einer binären zeitabhängigen Kovariablen

Die Werte $z_1 \ldots z_n$ enthalten hier beobachtete Zeitdauern, wann ein bestimmtes binäres zeitabhängiges Ereignis eingetreten ist. Das heißt konkret, wenn der *i*-te Patient in die Studie aufgenommen wird, startet er in Gruppe 0, bei Eintritt des besagten Ereignisses zum Zeitpunkt z_i wechselt er in Gruppe 1 und verbleibt dort bis zum Ende der Studie. Es kann natürlich sein, daß das binäre zeitabhängige Ereignis während des Follow-up des Patienten nicht eintritt, dann gilt $z_i > y_i$, und der Patient verbleibt in der Gruppe 0 bis zu seinem Ausscheiden aus der Studie. Es ist von klinischem Interesse, welchen Einfluß das binäre zeitabhängige Ereignis auf den Ausfall des Patienten hat.

Ein berühmtes Beispiel liefert das *Stanford Heart Transplant Programme*. Patienten wurden in die Studie aufgenommen, sobald sie aus medizinischen Gründen für eine Herztransplantation vorgesehen wurden. Wann immer ein Spenderherz zur Verfügung stand, wurde der aus medizinischer Sicht passendste Patient der Warteliste transplantiert. Um den Bezug zu oben herzustellen: Patient *i* startet die Studie auf der Warteliste (Gruppe 0) und nach einer mehr oder weniger langen Wartezeit z_i bekommt er ein Spenderherz transplantiert (er wechselt in Gruppe 1). Patienten können naturgemäß sowohl auf der Warteliste als auch nach der Transplantation versterben. Es stellt sich nun die Frage, ob und wie die Patienten von diesem Eingriff profitieren.

Üblicherweise beantwortet man solche Fragen durch Definition einer binären zeitabhängigen Kovariablen $x_i(t)$, $i = 1 \ldots n$,

$$x_i(t) = \begin{cases} 0 & 0 < t < z_i \\ 1 & t \geq z_i \end{cases}$$

und schätzt ein Cox Modell

$$h_i(t) = h_0(t) \exp\left(\beta x_i(t)\right).$$

Die Baseline-Hazardfunktion $h_0(t)$ ist die Hazardfunktion eines Patienten, der die ganze Zeit in Gruppe 0 verbleibt. Die log-Hazard-Ratio-Funktion (LHR) mit Bezug auf die Zeit *t* für den *i*-ten Patienten hat folgende Form, $i = 1 \ldots n$,

$$\mathrm{LHR}_i(t) = \log\left(h_i(t)/h_0(t)\right) = \begin{cases} 0 & x_i(t) = 0 \\ \beta & x_i(t) = 1 \end{cases}$$

Diese Art der Modellierung ist wieder unflexibel. Um beim Herztransplantationsbeispiel zu bleiben: Ist das „relative Sterberisiko“ eines Transplantierten im Vergleich zu einem Nichttransplantierten unmittelbar nach der Transplantation wirklich gleich hoch wie nach einem oder nach drei Monaten? Eine Modifikation könnte daher so aussehen,

$$\mathrm{LHR}_i(t) = \log\left(h_i(t)/h_0(t)\right) = \ell\left(t - z_i\right),$$

wobei $\ell(u) = 0$ für $u<0$. Per Definition hat jeder Patient seine eigene log-Hazard-Ratio-Funktion. Da aber diese bis zum Zeitpunkt des Eintretens des binären zeitabhängigen Ereignisses z_i immer gleich null ist, genügt es den gemeinsamen Teil ab z_i graphisch darzustellen.

Über die Approximation von $\ell(u)$ mit Hilfe zeitabhängiger Kovariablen und kubischer Splinefunktionen siehe Kapitel 3.

3. Kubische Splines

Ein kubischer Spline besteht aus kubischen Polynomen, die in Teilbereichen $\left(t_j, t_{j+1}\right]$, $j = 1 \ldots k-1$, definiert sind. Man bezeichtnet $t_1 < \ldots < t_k$ als die Knoten des Splines. Durch Zusatzbedingungen kann man die „Glattheit“ des Splines steuern. Die am häufigsten und deshalb auch hier verwendetete Bedingung ist die Stetigkeit bis einschließlich der 2. Ableitung. Oft wird der Spline auf die Intervalle $\left(a, t_1\right]$ und $\left(t_k, b\right)$ ausgedehnt, wobei häufig $a = -\infty$ (oder $a = 0$) und $b = +\infty$ gilt. Da bei Regressionsproblemen oftmals nur sehr wenige Datenpunkte in den Randintervallen liegen, wird zur Vermeidung von unerwünschter Überanpassung zumeist die Linearität des Splines in $\left(a, t_1\right]$ und $\left(t_k, b\right)$ gefordert. Man nennt einen kubischen Spline mit linearen Rändern *natürlich*.

Der Ausdruck für einen natürlichen kubischen Spline mit *k* Knoten $t_1 < \ldots < t_k$ ist

$$C(u) = \beta_0 + \beta_1 u + \sum_{j=1}^{k-2} \theta_j C_j(u),$$

wobei $C_1(u) \ldots C_{k-2}(u)$ kubische Terme bezeichnen,

$$C_j(u) = \left(u - t_j\right)_+^3 - \frac{\left(u - t_{k-1}\right)_+^3 \left[t_k - t_j\right]}{\left[t_k - t_{k-1}\right]} + \frac{\left(u - t_k\right)_+^3 \left[t_{k-1} - t_j\right]}{\left[t_k - t_{k-1}\right]}, \qquad j = 1 \ldots k-2.$$

Die Notation $(.)_+$ steht für $(a)_+ = \max(0, a)$. Beachte, daß $C(.)$ eine lineare Funktion in den Parametern $\beta_0, \beta_1, \theta_1 \ldots \theta_{k-2}$ ist, und daher mit statistischer Standardsoftware geschätzt werden kann (im Falle der Cox Regression z.B. mit PROC PHREG von SAS oder mit Prozedur 2L von BMDP). Des weiteren erlaubt die Linearität in den Parametern die Verwendung von Standardverfahren der Inferenzstatistik (für Tests und Konfidenzintervalle bzw. -bänder). Detailliertere Informationen dazu und zu den Fragen der Bestimmung der Knotenanzahl *k* und der Knotenpositionen $t_1 < \ldots < t_k$ findet man in Durrleman & Simon (1989), Hess (1984) und Heinzl, Kaider & Zlabinger (1996).

Wir wollen uns nun ansehen, wie mit einem natürlichen kubischen Spline $C(.)$ die Funktionen $f(.)$, $g(.)$ und $\ell(.)$ aus Kapitel 2 approximiert werden können.

- **Approximation von $f(.)$:**

  Der Baseline-Hazard $h_0(t)$ entspricht $h(t; Z = 0)$, deshalb muß $f(0) = 0$ gelten. Wenn wir $f(Z)$ mittels kubischer Splinefunktionen approximieren wollen, müssen wir diesen Umstand berücksichtigen.

$$\mathrm{LHR(Z)} = \log\left(h(t; Z)/h_0(t)\right) = f(Z)$$
$$\approx C(Z) - C(0) = \beta_1 Z + \sum_{j=1}^{k-2} \theta_j \left(C_j(Z) - C_j(0)\right).$$

  Anmerkung: Wenn der kleinste Knoten nicht-negativ ist, $t_1 \geq 0$, dann gilt $C_j(0) = 0$, $j = 1..k-2$, und damit auch $C(0) = 0$.

  Sei *Z* beispielsweise das Patientenalter. Es mag nun als wenig interessant bzw. als völlig unsinnig erscheinen, die „relativen Risken“ mit Bezug auf ein Alter null anzugeben, statt dessen kann es aber von großem Interesse sein, diese mit Bezug auf ein Alter von, sagen wir, 35 Jahren zu spezifizieren. Die log-Hazard-Ratio-Funktion für einen beliebigen Referenzwert *m* ist daher

$$\mathrm{LHR_m(Z)} = \log\left(h(t; Z)/h(t; m)\right) = f(Z) - f(m)$$
$$\approx C(Z) - C(m) = \beta_1 (Z - m) + \sum_{j=1}^{k-2} \theta_j \left(C_j(Z) - C_j(m)\right).$$

  Beachte, sobald wir $\beta_1, \theta_1 \ldots \theta_{k-2}$ geschätzt haben, können wir den Referenzwert *m* sehr einfach ändern, es ist nur eine Translation in der Ordinate notwendig.

- **Approximation von $g(.)$:**

  Eine Änderung um eine Einheit in *Z* bewirkt

$$\mathrm{LHR}_{+1}(t) = \log\left(h\left(t; Z = z_0 + 1\right) / h\left(t; Z = z_0\right)\right) = g(t)$$

$$\approx C(t) = \beta_0 + \beta_1 t + \sum_{j=1}^{k-2} \theta_j C_j\left(t\right).$$

  Beachte, die Knoten müssen hier nicht-negative Werte annehmen, $0 \le t_1 < \ldots < t_k$.

- **Approximation von $\ell(.)$:**

$$\mathrm{LHR}_i(t) = \log\left(h_i(t)/h_0(t)\right) = \ell\left(t - z_i\right)$$

$$\approx x_i(t)\, C\left(t - z_i\right) = \beta_0 x_i(t) + \beta_1 \left(t - z_i\right)_+ + \sum_{j=1}^{k-2} \theta_j C_j\left(t - z_i\right)$$

  Beachte, auch hier müssen die Knoten nicht-negative Werte annehmen, $0 \le t_1 < \ldots < t_k$.

Abschließend sei bemerkt, daß der Verständlichkeit wegen die Ergebnisse einer Spline-Approximation *immer* graphisch dargestellt werden sollten.

4. Warum SAS?

Obwohl wir SAS bereits viele Jahre verwenden, stellte sich für uns ernsthaft die Frage, ob die rechentechnische Umsetzung der beschriebenen statistischen Methoden mit SAS erfolgen sollte, oder ob wir auf z.B. S-Plus umsteigen sollten. Dabei wurden die Vor- und Nachteile von SAS aus unserer Sicht eingehend erörtert.

Vorteile von SAS:

I. exzellente Umgebung um Daten einzugeben, zu speichern, zu verwalten und zu transformieren

II. weite Verbreitung (SAS ist in vielen Bereichen, u.a. der medizinischen Statistik, *das* Standardprodukt)

III. gute Anbindung an andere im medizin-statistischen Bereich verbreitete Statistikprogramme, z.B. BMDP, StatXact, LogXact, usw.

IV. gute Programmierumgebung durch Macro-Möglichkeit und PROC IML

Nachteile von SAS:

V. die SAS-Manuals sind durch ihren Umfang und ihren Aufbau sehr unübersichtlich

VI. Graphiken waren bisher in SAS nur sehr umständlich produzierbar (hier wiegt der Nachteil der schlecht strukturierten, lexikalisch an Details ausgerichteten SAS-Manuals besonders schwer)

Ein ganz wesentlicher Punkt war aber für uns, ob das SAS-Institut die Entwicklung und Verbesserung von anspruchsvollen statistischen Prozeduren in Zukunft wieder ernstnehmen würde, denn einige Zeit lang schien uns dies nicht mehr der Fall zu sein. Als typisches Beispiel soll das bereits öfters erwähnte Cox'sche Regressionmodell dienen, von dem erst sehr spät (Version 6.07) eine eigene Prozedur (PHREG) freigegeben wurde, obwohl dieses Modell seit längerem als Standard in der medizinischen Statistik gilt. Auch andere Punkte liesen sich hier anführen.

Seitdem wir aber die SAS-Versionen 6.11, 6.12 (beide unter Win95) und 6.09E (unter CMS) kennengelernt haben, stellen wir erleichtert fest, daß es offensichtlich richtig war, bei SAS zu bleiben. Unsere Erwartungen an SAS für die Zukunft betreffen daher einerseits

weiterhin vermehrte Anstrengungen im statistischen Bereich und andererseits benutzerfreundlichere Lösungen in Bezug auf Manuals und Graphik.

5. SAS-Macro RCS

Die Anwendung von natürlichen kubischen Splinefunktionen im Cox'schen Regressionsmodell ist im Prinzip mit SAS sehr einfach.

1. Im Rahmen der PROC PHREG

    a. programmiere die Spline-Formeln

    b. schätze die Regressionsparameter

    c. teste die interessierenden Hypothesen

2. Verwende die geschätzten Regressionsparameter und deren Kovarianzmatrix, um die graphische Darstellung der geschätzten Splinefunktion inklusive dazugehöriger Konfidenzbänder mittels der PROC IML vorzubereiten.

3. Stelle die Ergebnisse graphisch dar (PROC GPLOT).

„Im Prinzip sehr einfach" heißt aber leider nicht „tatsächlich einfach". Wenn man sich einerseits die eher komplizierten Spline-Ausdrücke ansieht, und andererseits den Programmieraufwand bis zur fertigen Graphik im SAS vergegenwärtigt, dann erscheint es kaum verwunderlich, daß derartige flexible Modellierungsmethoden nur sehr beschränkt Verwendung finden.

Eine Lösung kann daher nur in einer adäquaten Unterstützung der potentiellen Nutzer dieser Methoden liegen. Dies war auch der Grund, warum wir uns für die Anfertigung des im folgenden beschriebenen SAS-Macros RCS („**r**estricted **c**ubic **s**plines", englisch für „natürliche kubische Splines") entschlossen.

Unser erster Ansatz war folgender: Der Benutzer spezifiziert sein Problem über Macro-Statements, und das Macro erzeugt den PHREG-Output und die dazugehörige Graphik. Wir erkannten allerdings relativ rasch ein Dilemma. Wollen wir ein möglichst einfach zu bedienendes Macro, dann werden die Möglichkeiten des Benutzers enorm eingeengt. Wollen wir hingegen dem Benutzer eine Menge von Optionen offenlassen, dann wird einerseits die Bedienung des resultierenden Macros extrem unübersichtlich, und andererseits stellt sich die Frage, ob der Benutzer in bestimmten, unvorhersehbaren Situationen nicht doch eingeschränkt wird.

Unsere Lösung des Dilemmas war schließlich ein einfach bedienbares Macro, welches anstatt SAS-Output ein lauffähiges SAS-Programm erzeugt. Dieses Programm kann nun bei Bedarf vom Benutzer beliebig modifiziert werden.

6. Beispiel

Hess (1994) verwendet einen relativ bekannten Datensatz, der aus den Aufzeichnungen von 42 Patienten mit akuter Leukämie besteht. Eine bestimmte Therapie (6-MP) wird mit Placebo bezüglich der Remissionsdauer verglichen. REMTIME (in Wochen), REMSTAT (1=Ende der Remission, 0=zensiert) und GROUP (0=Placebo, 1=6-MP) sind die einzigen Variablen dieses Datensatzess. Die klinisch interessierenden Fragen sind: Besteht ein Effekt der 6-MP Therapie, und wenn ja, hängt dieser von der Zeit ab? Letztere Frage kann statistisch reformuliert werden in: Gilt die proportionale Hazards-Annahme oder besteht eine Wechselwirkung zwischen der Zeit und der Therapiegruppe?

Um nun mit Hilfe des RCS-Macros ein SAS-Programm zur Schätzung eines kubischen Splines mit 3 Knoten zu erzeugen, und damit eine mögliche Wechselwirkung zwischen der Zeit und der Variable GROUP zu testen, müssen wir die folgenden Macro-Statements spezifizieren. Dabei ist zu beachten, daß keiner der verwendeten Datei- und Variablennamen mit einem doppelten Underscore beginnen darf, da es sonst zu unerwünschten Überschneidungen mit erzeugten Hilfsdateien bzw. -variablen kommen kann.

```
%RCS(
  TITLE=%STR(ACUTE LEUKAEMIA DATASET AS PRINTED IN HESS,
1994),
  DATA=GEHAN,  DIRDATA=%STR(B:\DATA\),
  PROGRAM=%STR(E:\STUDY\RESULTS\HESS.SAS),
  TIME=REMTIME, STATUS=REMSTAT,
  COV1=GROUP, WHAT1=1, KNOTS1=6 10 19,
  TIMEUNIT=WEEKS
);
```

Wir haben folgendes spezifiziert:

- den Titel der Analyse (`ACUTE LEUKAEMIA DATASET AS PRINTED IN HESS, 1994`)
- daß sich im Verzeichnis `B:\DATA\` das SAS-Datenfile `GEHAN` befindet, und daß das erzeugte SAS-Programm im Verzeichnis `E:\STUDY\RESULTS\` ins File `HESS.SAS` geschrieben werden soll

- die Zeit- (`REMTIME`) und die Statusvariable (`REMSTAT`), bei letzterer wird übrigens die 0/1-Codierung defaultmäßig verlangt

- eine Kovariable (`GROUP`), die in Form einer Wechselwirkung mit der Zeit (`WHAT1=1`) mit einem kubischen Spline mit 3 Knoten (`KNOTS1=6 10 19`) modelliert werden soll

- die Zeiteinheit (`WEEKS`)

Der Inhalt des erzeugten SAS-Programmfiles „`E:\STUDY\RESULTS\HESS.SAS`“ sieht wie folgt aus (der PROC IML-Code wurde stark gekürzt):

```
LIBNAME __DATA  'B:\DATA\';

TITLE ' ACUTE LEUKAEMIA DATASET AS PRINTED IN HESS, 1994 ';

PROC PHREG DATA=__DATA.GEHAN COVOUT OUTEST=__RCS;
 MODEL REMTIME*REMSTAT(0) =  GROUP  __1_LIN  __1_1  /RL;

 ********** spline modelling of the time-dependent effect;
 ********** of fixed covariate GROUP;
 ********** with 3 knots located at;
 ********** 6 10 19;
 __1_1=((REMTIME-6)**3)*(REMTIME>6)
     -((REMTIME-10)**3)*(REMTIME>10)
     *(19-6)/(19-10)
     +((REMTIME-19)**3)*(REMTIME>19)
     *(10-6)/(19-10);
 __1_LIN=GROUP*REMTIME;
 __1_1=GROUP*__1_1;

 *--------- Testing variable: GROUP ---------;
 EFFECT1: TEST  GROUP, __1_LIN, __1_1;
 NONCON1: TEST  __1_LIN, __1_1;
 NONLIN1: TEST  __1_1;
 RUN;
*====================      End      of      PROC      PHREG
====================;
```

```
*-------------------- Graph for GROUP --------------------;
PROC IML;
 NPOINTS=101;   * Number of points to build the graphic;
 LOWEREND=0;     *Smallest value for X-axis;
 UPPEREND=19;     *Largest value for X-axis;
 [... code deleted ...]
 CREATE __RCS1 VAR { F FE Z X };  APPEND;  CLOSE __RCS1;
 QUIT;

 SYMBOL1 C=RED  L=1 I=JOIN WIDTH=5;
 SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
 SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

PROC GPLOT DATA=__RCS1;
 PLOT F*X=Z / VREF=0 LV=3 NOLEGEND;
 TITLE2 ' GROUP ';
 LABEL X=WEEKS;
 LABEL F=LOG HAZARD RATIO;
RUN;
```

Die erzeugte PHREG-Prozedur hat drei Regressoren, GROUP, __1_LIN und __1_1. Die beiden letzteren werden erst im Programmteil von PHREG definiert. Beachte, daß dabei beide auch mit GROUP multipliziert werden. Dies ist notwendig, um den zeitabhängigen Effekt einer Änderung um *eine* Einheit in der interessierenden Kovariablen (hier: GROUP) zu modellieren.

Das TEST-Statement EFFECT1 ermöglicht einen Wald Test für die Null-Hypothese, daß alle Regressionskoeffizienten mit Bezug zu GROUP gleich null sind, bzw. äquivalent formuliert: „Kein Effekt von GROUP auf die Remissionsdauer.“ Die TEST-Statements NONCON1 und NONLIN1 testen die Null-Hypothesen: „Kein zeitabhängiger Effekt von GROUP.“ und „Der zeitabhängige Effekt von GROUP ist linear.“

Nach der Aufbereitung der Regressionsergebnisse mit einem IML-Programm wird die geschätzte log-Hazard-Ratio-Funktion einer Änderung um eine Einheit in GROUP gegen die Zeit geplottet. Da GROUP dichotom ist (0=Placebo, 1=6-MP), steht eine Änderung um eine Einheit für den Unterschied zwischen den beiden Therapien.

7. Abschließende Bemerkungen

Das beschriebene Macro RCS ist über World-Wide-Web unter

http://www.akh-wien.ac.at/imc/biometrie/rcs.zip

erhältlich. Durch *Entzippen* der übermittelten Datei erhält der Benutzer das Macro selbst (`rcs.mac`), ein Manual in Postscript-Format (`rcsman.ps`) und drei Beispieldatensätze (`miller.dat`, `gehan.dat`, `oakes.dat`). Die Auswertung letzterer ist im Manual beschrieben und bezieht sich auf Beispiele aus Durrleman & Simon (1989), Hess (1994) und Heinzl, Kaider & Zlabinger (1996).

Einige interessante Details des Macros RCS:

- Die gleichzeitige Spline-Modellierung von bis zu 20 Kovariablen ist möglich, die Anzahl der spezifizierbaren Knoten pro Kovariable ist theoretisch unbegrenzt.

- Die Spline-Modellierung ist immer nur optional, d.h. auch die einfache Spezifikation von Kovariablen ist möglich.

- Um mit der beschriebenen statistischen Technik sinnvolle Ergebnisse zu erhalten, sollte man bedenken, daß jeder spezifizierte Knoten dem *Verbrauch* eines zusätzlichen Freiheitsgrads entspricht. Die verwendeten Datensätze sollten daher eine diesem Umstand adäquate Stichprobengröße aufweisen.

- Um numerische Probleme zu vermeiden, sollte die Zeitachse auf relative große Einheiten transformiert werden, z.B. Umwandlung von Angaben in Tagen in eine Monats- oder Jahresskala.

Wir würden es begrüßen, wenn das präsentierte Macro RCS häufige Verwendung finden würde. Wir laden alle Benutzer ein, uns ihre konstruktive Kritik und interessanten Anmerkungen zu übermitteln.

Literaturliste

Cox D.R. (1972): *Regression models and life-tables.* Journal of the Royal Statistical Society Series B 34, 187-220.

Cox D.R. (1975): *Partial likelihood.* Biometrika 62, 269-276.

Durrleman S. & Simon R. (1989): *Flexible regression models with cubic splines.* Statistics in Medicine 8, 551-561.

Hess K.R. (1994): *Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions.* Statistics in Medicine 13, 1045-1062.

Heinzl H., Kaider A. & Zlabinger G. (1996): *Assessing interactions of binary time-dependent covariates with time in Cox proportional hazards regression models using cubic spline functions.* Statistics in Medicine 15, 2589-2601.