



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK  
SONDERFORSCHUNGSBEREICH 386



Aydemir, Aydemir, Dirschedl:

## Das lineare Regressionsmodell von Aalen zur Analyse von Überlebenszeiten unter Berücksichtigung zeitveränderlicher Kovariablen

Sonderforschungsbereich 386, Paper 25 (1996)

Online unter: <http://epub.ub.uni-muenchen.de/>

Projektpartner



# Das lineare Regressionsmodell von Aalen zur Analyse von Überlebenszeiten unter Berücksichtigung zeitveränderlicher Kovariablen

Sibel Aydemir<sup>1</sup>, Ülker Aydemir<sup>2</sup>, Peter Dirschedl<sup>1</sup>

## Abstract

Data from clinical studies often contain time-dependent covariates, e.g. events like transplantation or an adverse drug reaction, or the changing measurements of laboratory data. The common approach uses only the covariate information at time  $t=0$  for regression analyses, but this baseline analysis is not very satisfying. This paper applies the linear counting process by Aalen for failure time analysis, modified to deal with time-dependent covariates. In the main part we describe methods to estimate and visualize the cumulated regression function with respect to time-dependent covariates. After introducing a test for significance of the influence of covariates we display different methods to investigate model validity depending on martingale residuals, or by use of the Arjas plot. Coding and interpretation problems are shortly discussed. Results are illustrated with data from the Stanford Heart Transplantation Study and a study on Oropharynx carcinoma.

---

<sup>1</sup> Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie der Ludwig-Maximilians-Universität München

<sup>2</sup> Münchner Forschungsverbund Öffentliche Gesundheit-Public Health, Ludwig-Maximilians-Universität München

# 1 Einführung

Bei der Analyse von Ereigniszeiten hat sich besonders das semiparametrische Cox-proportional-hazards-Regressionsmodell durchgesetzt, nicht zuletzt wegen seiner einfachen Anwendung und der Verfügbarkeit in Softwarepaketen. Allerdings ist die dafür erforderliche Voraussetzung der proportional-hazards in der Praxis nicht immer erfüllt. Als Alternative zum Cox-Regressionsmodell stellt Aalen (1989) für die Analyse von Ereigniszeiten ein nichtparametrisches lineares Modell vor, das im Unterschied zum Cox-Regressionsmodell gestattet, die zeitliche Entwicklung der Kovariableneffekte mitzuschätzen. Der Ansatz der Analyse im linearen Modell besteht in der Betrachtung kumulierter Regressionsfunktionen, die über einen Zählprozeßansatz geschätzt werden. Um den Einfluß einer Kovariablen in einem bestimmten Zeitpunkt zu ermitteln, wird die Steigung des kumulierten Regressionsfunktionsplots in diesem Moment betrachtet. Das lineare Modell bietet die Möglichkeit den Einfluß der Kovariablen im gesamten Beobachtungszeitraum auf Signifikanz zu testen. Die Überprüfung der Validität des Modells kann mit Hilfe von Arjas-Plots oder Martingalresiduen vorgenommen werden. Das Schätzverfahren, die Inferenz und die Diagnostik im linearen Modells demonstriert Aalen (1989, 1993) an Beispieldatensätzen aus der Medizin, die allerdings ausschließlich fixe, d.h. zum Baselinezeitpunkt  $t=0$  ermittelte Kovariablen enthalten.

In Anlehnung an diese Arbeiten wurde im Rahmen des Teilprojekts B3 des Sonderforschungsbereichs 386 das lineare Aalen-Modell auf die Situation zeitveränderlicher Kovariablen angewendet, die speziell in klinischen Studien von großer Bedeutung sind. Unter zeitveränderlichen Kovariablen werden dabei Kovariablen mit ein- oder mehrmaligen Zustandsänderungen im Beobachtungszeitraum verstanden. Das Paradebeispiel einer Kovariablen mit einer einmaligen Zustandsänderung ist die Transplantation. Messungen von Kovariablen im Verlauf ("updated covariates"), wie beispielsweise das Auftreten von klinischen Symptomen oder Veränderungen von Laborwerten, die zu den jeweiligen Untersuchungsterminen eines Patienten erhoben werden, bzw. das Auftreten von Nebenwirkungen während der Beobachtung sind weitere Beispiele für zeitveränderliche Kovariablen.

Im Gegensatz zu zeitveränderlichen Kovariablen spricht man von fixen Kovariablen, wenn Kovariablenwerte nur einmal zu Beginn des Beobachtungszeitraumes bzw. zum sogenannten Baselinezeitpunkt  $t=0$  ermittelt werden. Von den Kovariablen, die fix oder zeitveränderlich sein können, sind die Effekte der Kovariablen zu unterscheiden, die ihrerseits entweder konstant im gesamten Beobachtungszeitraum sind, wie z.B. im Cox-Regressionsmodell, oder dynamisch geschätzt werden können, wie in dem hier betrachteten Aalen-Modell.

Die vorliegende Arbeit beschäftigt sich mit dem linearen Zählprozeßansatz von Aalen bei Vorliegen einer zeitveränderlichen Struktur der Kovariablen. Nach einer Vorstellung des linearen Ansatzes in Abschnitt 2, wird in Abschnitt 3 auf die Schätzung und vor allem auch auf die Interpretation der kumulierten Regressionsfunktionen eingegangen. In Abschnitt 4 wird die Inferenz im linearen Modell behandelt. Der Abschnitt 5 beschäftigt sich mit der Frage nach der goodness-of-fit des betrachteten Modells. Hier werden auf der Grundlage von Martingalresiduen einige Möglichkeiten zur Überprüfung der Validität des Modells präsentiert. Schätzung und Interpretation kumulierter Regressionsfunktionen und das Testen von Kovariableneinflüssen auf Signifikanz werden in Abschnitt 6 am Beispiel der Daten der Stanford Heart Transplantation Study illustriert. Abschließend erfolgt in Abschnitt 7 eine Diskussion des linearen Modells. Die Vorteile, Besonderheiten und Schwächen des Aalen'schen Ansatzes werden dem Cox-proportional-hazards-Regressionsmodell, das sich als Standardverfahren der Survivalanalyse etabliert hat, bzw. dem auf zeitveränderliche Kovariablen erweiterten sogenannten time-dependent Cox-Modell gegenübergestellt.

## **2 Das lineare Regressionsmodell**

Es werden  $n$  Individuen im Hinblick auf ein Ereignis beobachtet. Sei mit  $\lambda_i(t)$  die Intensität für das Auftreten eines Ereignisses zum Zeitpunkt  $t$  des  $i$ -ten Individuums bezeichnet, d.h.  $\lambda_i(t)$  ist die Wahrscheinlichkeit dafür,

daß ein Ereignis im Intervall  $t+dt$  eintritt, unter der Bedingung, daß zuvor noch kein Ereignis eingetreten ist. Sei  $\lambda(t)$  der  $n$ -dimensionale Intensitätsvektor bestehend aus den individuellen Intensitäten (=hazards)  $\lambda_i(t)$ ,  $i=1,\dots,n$ . Ferner wird angenommen, daß in der Regressionsanalyse  $r$  Kovariable berücksichtigt werden.

Das lineare Modell für die Intensität  $\lambda(t)$  lautet

$$\lambda(t) = Y(t)\alpha(t), \quad (1)$$

wobei  $Y(t)$  die  $n \times (r+1)$ -Datenmatrix, bestehend aus dem Intercept 1 und den  $r$  Kovariablen  $Z_j$ ,  $j = 1, \dots, r$ , gemessen zum Zeitpunkt  $t$ , und  $\alpha(t) = (\alpha_0(t), \alpha_1(t), \dots, \alpha_r(t))'$  der Vektor der Regressionsfunktionen sind. Die Regressionsfunktion  $\alpha_j(t)$ ,  $j = 1, \dots, r$ , gibt dabei den Einfluß der  $j$ -ten Kovariable zur Zeit  $t$  wieder,  $\alpha_0(t)$  bezeichnet die Baselineparameterfunktion.

Durch den linearen Ansatz wird somit unter Berücksichtigung der zeitlichen Struktur der Kovariablen eine dynamische Modellierung der Regressionsfunktionen ermöglicht.

Die Konstruktion von  $Y(t)$  zum Zeitpunkt  $t$  erfolgt folgendermaßen :

- falls zum Zeitpunkt  $t$  das  $i$ -te Individuum unter Risiko ist, d.h. bisher noch kein Ereignis hatte und nicht zensiert ist, entspricht die  $i$ -te Zeile der  $Y(t)$ -Matrix dem Vektor

$$Z^i(t) = \left(1, Z_1^i(t), Z_2^i(t), \dots, Z_r^i(t)\right)',$$

wobei  $Z_j^i(t)$ ,  $j = 1, \dots, r$ , die Kovariablenwerte des  $i$ -ten Individuums zum Zeitpunkt  $t$  sind.

- falls zum Zeitpunkt  $t$  das  $i$ -te Individuum nicht mehr im Risikoset enthalten ist, so ist die  $i$ -te Zeile der  $Y(t)$ -Matrix mit Nullen besetzt.

Im Zusammenhang mit der zeitlichen Struktur von  $Y(t)$  müssen folgende zwei Annahmen zugrunde gelegt werden :

- (A1) Die Stichprobenfunktionen sind linksstetig.
- (A2) Der Wert der Matrix  $Y(t)$  hängt nur von der Vergangenheit, und nicht von der Zukunft ab.

Die Annahme (A1) impliziert, daß bei Vorliegen von zeitveränderlichen Kovariablen diejenige Belegung der Kovariablenmatrix  $Y(t)$  zu wählen ist, die unmittelbar vor einem Ereigniszeitpunkt zutrifft. (In praxi wird der Kovariablenwert beim letzten Ereignis gewählt.)

Liegen ausschließlich Kovariablen mit ihren zum Baselinezeitpunkt  $t=0$  ermittelten Meßwerten vor, bleibt die Belegung der Kovariablenmatrix für die Zeitdauer, die der Patient unter Risiko ist, konstant. Solange der  $i$ -te Patient unter Risiko ist, läßt sich also die  $i$ -te Zeile von  $Y(t)$  schreiben als

$$Z^i = \left(1, Z_1^i, Z_2^i, \dots, Z_r^i\right)' ,$$

wobei  $Z_j^i, j = 1, \dots, r$ , die Kovariablenwerte des  $i$ -ten Individuums gemessen zum Baselinezeitpunkt  $t=0$  sind.

### 3 Schätzung im linearen Regressionsmodell

Anstelle der Regressionsfunktionen  $\alpha(t)$  werden im linearen Regressionsmodell die kumulierten Regressionsfunktionen geschätzt.

Der Vektor der kumulierten Regressionsfunktionen sei mit  $A(t)$  bezeichnet, d.h.

$$A(t) = \left(A_0(t), A_1(t), \dots, A_r(t)\right)' = \left(\int_0^t \alpha_0(s) ds, \int_0^t \alpha_1(s) ds, \dots, \int_0^t \alpha_r(s) ds\right)' .$$

Seien  $T_1 < T_2 < \dots$  geordnete Ereigniszeitpunkte. Es wird angenommen, daß keine gleichzeitigen Ereignisse (Bindungen, ties) auftreten. Im Falle von Bindungen in den Daten wird zu den (ganzzahligen) Ereigniszeiten eine auf  $(0,1)$  gleichverteilte Zufallszahl addiert, um die Ereigniszeitpunkte in eine steigende Ordnung zu bringen. Der Schätzer für die kumulierten Regressionsfunktionen ist gegeben durch

$$A^*(t) = \sum_{T_k \leq t} X(T_k) I_k, \quad (2)$$

wobei  $X(t) = \left[ Y(t)' Y(t) \right]^{-1} Y(t)'$  die generalisierte Inverse von  $Y(t)$  ist, und  $I_k$  ein  $n$ -dimensionaler Vektor, der an  $k$ -ter Stelle mit 1, sonst mit Nullen besetzt ist.

Eine unverzerrte Schätzung  $A^*(t)$  für die kumulierten Regressionsfunktionen ist allerdings nur möglich, solange die Matrix  $Y(t)$  vollen Rang besitzt und somit invertierbar ist, d.h. wenn  $\text{rg}(Y(t))=r+1$ .

$A^*(t)$  ist asymptotisch normalverteilt mit der Kovarianzmatrix

$$\text{cov}\left(A^*(t)\right) = \Omega^*(t) = \sum_{T_k \leq t} X(T_k) I_k^D X(T_k)',$$

wobei  $I_k^D$  eine  $(n \times n)$ -Diagonalmatrix ist, mit  $I_k$  als Hauptdiagonale.

### **Interpretation der geschätzten kumulierten Regressionsfunktionen**

Für  $j \geq 1$  kann die geschätzte kumulierte Regressionsfunktion  $A_j^*(t)$  als eine empirische Funktion betrachtet werden, die den Einfluß der  $j$ -ten Kovariable beschreibt. Wird  $A_j^*(t)$  gegen die Zeit geplottet, so gibt die Steigung dieser Funktion den Einfluß der  $j$ -ten Kovariable wieder. Bei konstantem  $\alpha_j(t)$  entspricht der Plot von  $A_j^*(t)$  gegen  $t$  einer Geraden.

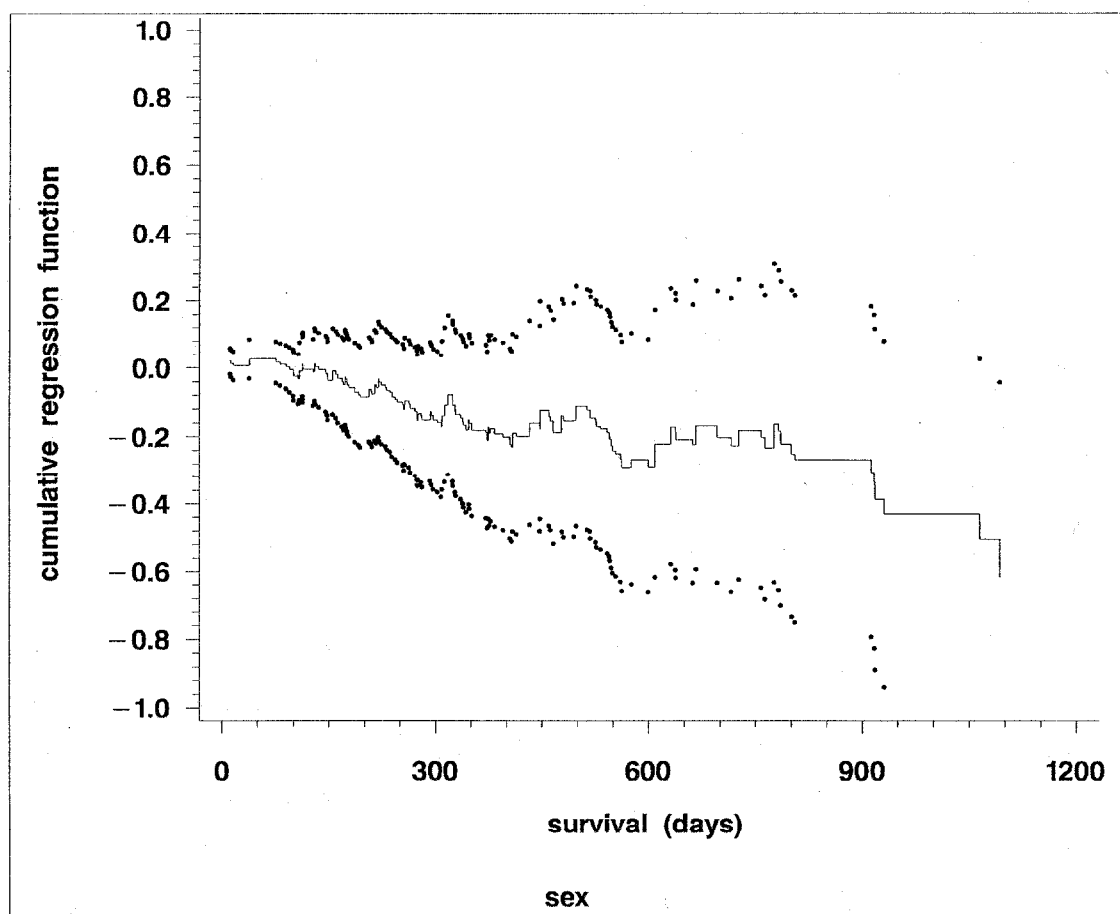
Da die Interpretation kumulierter Regressionsfunktionen bei fixen und zeitveränderlichen Kovariablen gleich ist, werden an dieser Stelle zur Illustration die Daten einer klinischen Studie verwendet, die ausschließlich fixe Kovariablen enthält. Diese in Kalbfleisch&Prentice (1980) publizierte Daten stammen aus einer klinischen Studie zum Vergleich einer Testtherapie mit einer Standardtherapie beim Oropharynxkarzinom und sind bereits mit dem linearen Aalen-Modell analysiert worden (Aalen, 1989). Zu Beginn der Studie wurden bei 195 Patienten verschiedene Kovariablen wie beispielsweise Geschlecht, Alter, Gesundheitszustand des Patienten sowie Grad und Lokalisation des Tumors erhoben. Die Abbildungen 1 und 2 zeigen die kumulierten Regressionsfunktionen mit punktwisen Konfidenzen ( $1.96 \cdot \text{Standardabweichung}$ ) für die Kovariablen Geschlecht (1=männlich, 2=weiblich) und Gesundheitszustand (1=nicht eingeschränkt, 2=beschränkt arbeitsfähig, 3=Hilfe erforderlich, 4=bettlägerig). Wie in Abbildung 1 zu erkennen ist, besitzt Geschlecht über den gesamten Beobachtungszeitraum hinweg einen Einfluß auf die Intensität. Aus der negativen Steigung der kumulierten Regressionsfunktion der Kovariable Geschlecht läßt sich ein geringeres Risiko bzw. eine geringere Mortalität der Frauen gegenüber den Männern ableiten (tatsächlich ist hier nur ein Trend erkennbar, der entsprechende Test ist nicht signifikant). Betrachtet man den Plot der kumulierten Regressionsfunktion des gesundheitlichen Zustands, so ist lediglich im ersten Jahr ein positiver Einfluß der Kovariable auf die Intensität erkennbar, ab dem ersten Jahr scheint dieser Effekt zu verschwinden.

### **Anmerkungen zur Kodierung kategorialer Kovariablen**

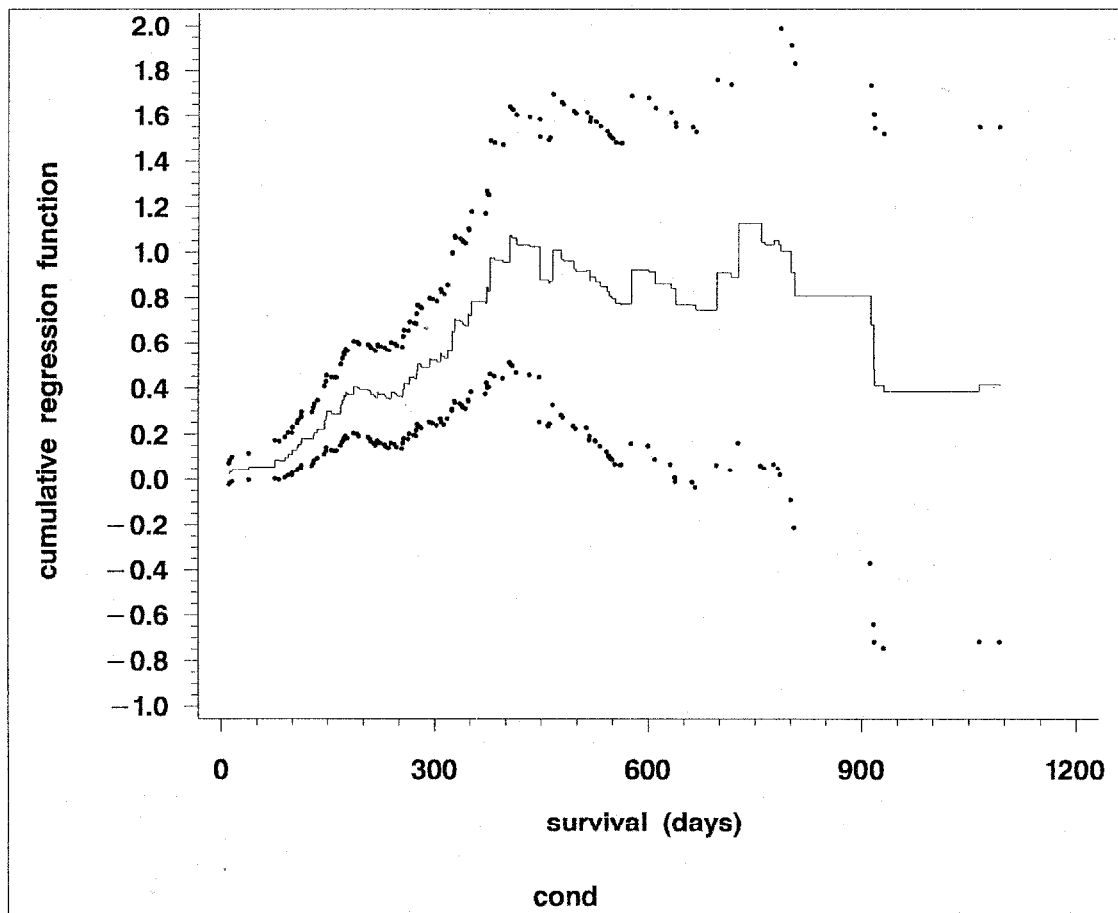
Der lineare Ansatz impliziert bei kategorialen Kovariablen, daß die Kodierung der Kategorien jeweils im Abstand von einer Einheit erfolgt. Eine Abweichung der Codes der Kategorien um mehr als eine Einheit beeinflußt folglich die Schätzung der kumulierten Regressionsfunktion. Während sich beispielsweise (bei obigem Beispiel) eine Kodierung des Geschlechts mit 0/1 anstelle von 1/2 nicht auf die Schätzung auswirkt, führt eine Kodierung mit 1/3 zu einer Halbierung der Schätzwerte der kumulierten Regressionsfunktion. Die Kodierungsproblematik ist vor



allein bei Subgruppenanalysen zu beachten. Hier kann eine Patientenselektion dazu führen, daß bei mehrkategorialen Kovariablen eine Kategorie vollkommen eliminiert wird. In diesem Fall ist eine Neukodierung der verbleibenden Kategorien unerlässlich. Natürlich ist bei mehrkategorialen Kovariablen generell zu beachten, daß die Ordinalität zwischen den Kategorien gegeben sein muß. Andernfalls ist eine Dummykodierung der Kategorien erforderlich.



**Abbildung 1 :** Kovariable "*Geschlecht*"  
geschätzte kumulierte Regressionsfunktion mit  
punktweisen Konfidenzen



**Abbildung 2 :** Kovariable "*Gesundheitszustand*"  
 geschätzte kumulierte Regressionsfunktion mit  
 punktwisen Konfidenzen

## Schätzung der kumulierten Intensitätsfunktion und der Survivorfunktion

Sei mit  $\Lambda_i(t)$  die kumulierte Intensitätsfunktion des  $i$ -ten Individuums bzw. mit  $\Lambda(t)$  der Vektor der kumulierten individuellen Intensitäten bezeichnet. Aus den geschätzten kumulierten Regressionsfunktionen läßt sich direkt die Schätzung für die kumulierte Intensitätsfunktion ableiten als

$$\Lambda_i^*(t) = A^*(t)' Z^i(t)', \quad i = 1, \dots, n$$

bzw.

$$\Lambda^*(t) = A^*(t)' Y(t)' \quad (3)$$

mit der Kovarianzmatrix

$$\begin{aligned} \text{cov}(\Lambda^*(t)) &= \text{cov}(A^*(t)' Y(t)') \\ &= Y(t) \text{cov}(A^*(t)) Y(t)' = Y(t) \Omega^*(t) Y(t)'. \end{aligned}$$

Entsprechend lautet die Schätzung für die Survivorfunktion

$$P^*(t) = \exp(-\Lambda^*(t))$$

und für deren Kovarianzmatrix

$$\begin{aligned} \text{cov}(P^*(t)) &= \text{cov}(\Lambda^*(t)) \cdot [P^*(t)]^2 \\ &= Y(t) \Omega^*(t) Y(t)' \cdot [P^*(t)]^2. \end{aligned}$$

Im Gegensatz zu anderen Regressionsverfahren ist bei diesem nichtparametrischen linearen Ansatz die geschätzte Survivorfunktion nicht notwendigerweise monoton im gesamten Beobachtungszeitraum. Da eine Restriktion der Intensität auf den nicht-negativen Wertebereich nicht vorhanden ist, kann in manchen Zeitpunkten die Survivorfunktion ansteigen und entsprechend die kumulierte Intensitätsfunktion fallen.

## 4 Inferenz im linearen Regressionsmodell

Das lineare Regressionsmodell bietet die Möglichkeit, den Einfluß einer Kovariablen auf das Überleben zu testen.

Die Teststatistik für die Hypothese

$$H_j : \alpha_j(t) = 0 \quad \forall t$$

wird formuliert durch die j-te Komponente  $U_j$  des Vektors  $U$

$$U = \sum_{T_k} K(T_k) X(T_k) I_k. \quad (4)$$

Die Teststatistik  $U$  berechnet sich als eine gewichtete Summe der Summanden von  $A^*(t)$  mit der Gewichtematrix  $K(t)$ , wobei die Summierung über alle Ereigniszeitpunkte erfolgt.

Analog zur Schätzung ist auch eine Inferenz im nonparametrischen Ansatz nur solange möglich, solange  $Y(t)$  vollen Rang besitzt und somit invertierbar ist.

Die Matrix  $K(t)$  ist eine  $(r+1)$ -dimensionale Diagonalmatrix, in der die Diagonalelemente den Gewichtsfunktionen entsprechen. Die Wahl der Gewichtematrix  $K(t)$  hängt einerseits von den Alternativhypothesen andererseits von den relativen Gewichten der einzelnen Zeitintervalle ab. Eine mögliche Belegung der Gewichtematrix ist, die Gewichtsfunktionen entsprechend der Anzahl  $M(t)$  der Individuen zu wählen, die zu gegebenen Zeitpunkten im Risikoset enthalten sind. In diesem Fall wird  $K(T_k)$  substituiert durch das Skalar  $M(T_k)$ . Als eine weitere Möglichkeit für  $K(t)$ , mit der zudem eine größere Power erreicht wird, schlägt Aalen (1989) folgenden Ausdruck vor

$$K(t) = \left\{ \text{diag} \left[ \left( Y(t)' Y(t) \right)^{-1} \right] \right\}^{-1}.$$

Die Teststatistik  $U$  ist asymptotisch multivariat normalverteilt. Die Schätzung für die Kovarianzmatrix  $V$  der Teststatistik  $U$  lautet

$$\text{cov}(U) = V = \sum_{T_k} K(T_k) X(T_k) I_k^D X(T_k)' K(T_k).$$

Wird nur eine Komponente getestet, z.B.  $H_j$ , so ist die Teststatistik

$$z = U_j V_{jj}^{-1} \quad (5)$$

zu verwenden, die unter der Nullhypothese  $N(0,1)$ -verteilt ist.

### **Simultanes Testen mehrerer Hypothesen (Kovariableneffekte)**

$A$  sei eine  $s$ -dimensionale Subgruppe aus  $\{1, \dots, r\}$ ,  $U_A$  der entsprechende Subvektor von  $U$  und analog  $V_A$  die entsprechende Submatrix der Kovarianzmatrix  $V$  ( $V_A$  ist die geschätzte Kovarianzmatrix von  $U_A$ ).

Die normalisierte Teststatistik zum simultanen Testen aller  $s$  Hypothesen  $H_j$ ,  $j \in A$  lautet

$$U_A' V_A^{-1} U_A$$

und ist  $\chi^2$ -verteilt mit  $s$  Freiheitsgraden.

## **5 Modelldiagnostik**

Die Anwendung eines Modells und die Interpretation der daraus abgeleiteten Ergebnisse machen jedoch nur dann Sinn, wenn die Validität des Modells gezeigt werden kann. Ein besonders nützliches graphisches Verfahren zur Überprüfung der goodness-of-fit des linearen Aalen-Modells ist der Arjas-Plot (Arjas, 1983). Der Arjas-Plot vergleicht für verschiedene Subgruppen von Kovariablenkonstellationen die Anzahl der beobachteten und der erwarteten Ereignisse als eine Funktion über die Zeit.

Eine weitere Möglichkeit zur Beurteilung des Modellfits stellen die von Aalen (1993) in Anlehnung an das Cox-Modell definierten Martingalresiduen dar. Martingalresiduen sind nur für den Zeitraum definiert, in dem eine Schätzung möglich ist, d.h. solange  $Y(t)$  den Rang  $r+1$  besitzt.

Sei  $R$  der Zeitpunkt, zu dem  $Y(t)$  den vollen Rang verliert, so daß alle Individuen, die bis dahin noch kein Ereignis erfahren haben, zensiert werden. Sei mit  $N_i(t)$  der Zählprozeß des  $i$ -ten Individuums bezeichnet, d.h. vor dem Ereigniszeitpunkt des  $i$ -ten Individuums ist  $N_i(t) = 0$  und ab dem Ereigniszeitpunkt ist  $N_i(t) = 1$ . Im Falle von zensierten Beobachtungen bleibt folglich  $N_i(t) = 0$ . Der Martingalresiduenprozeß des  $i$ -ten Individuums ist definiert als

$$M_{res,i}(t) = N_i(t) - \Lambda_i^*(t). \quad (6)$$

D.h.

$$M_{res,i}(t) = \begin{cases} -\Lambda_i^*(t) & \text{für } t < \text{Ereigniszeitpunkt des } i\text{-ten Individuums} \\ 1 - \Lambda_i^*(t) & \text{für } t \geq \text{Ereigniszeitpunkt des } i\text{-ten Individuums.} \end{cases}$$

$M_{res,i}(t)$  sind exakte Martingale.  $M_{res,i}(R)$  ist der letzte Wert des Residualprozesses und wird Martingalresiduum genannt.

Falls das lineare Modell das wahre Modell ist, d.h. ein guter Fit an die Daten gewährleistet ist, gilt für alle Individuen  $i, i=1, \dots, n$  :

$$E(M_{res,i}) = 0.$$

Mit (3) kann der Vektor der Martingalresiduenprozesse geschrieben werden als

$$M_{res}(t) = \sum_{T_K \leq t} (J - Y(T_K)X(T_K))I_K, \quad (7)$$

wobei  $J$  eine  $n \times n$ -Einheitsmatrix ist.

Für eine Beurteilung der goodness-of-fit können zwei komplementäre Ansätze herangezogen werden, zum einen eine graphische Überprüfung der Martingalresiduen mittels eines Martingalresiduenplots, zum anderen ein analytisches Verfahren zur Identifizierung nicht-linearer Effekte der Kovariablen.

## 5.1 Martingalresiduenplots

Da die Verteilung der Martingalresiduen gewöhnlich schief, bei hoher Zensierungsrate sogar bimodal ist, ist eine Interpretation der Residuenplots in der Regel schwer. Aalen (1993) schlägt vor, den Martingalresiduenprozeß (MRP) summiert über eine Subgruppe gegen die Zeit zu plotten, um den Fit des linearen Modells für diese betrachtete Subgruppe über die Zeit beurteilen zu können. Bei guter Anpassung des linearen Modells streut der Plot um die Nullachse. Aufgrund der Tatsache, daß der Plot des Martingalresiduenprozesses explizit die Zeitachse enthält, können Zeitintervalle mit guter und ebenso mit schlechter Anpassung leicht diagnostiziert werden. Dies ist ein großer Vorteil des MRP-Plots gegenüber dem Arjas-Plot, der keine Zeitachse enthält, sondern die Anzahl der beobachteten Ereignisse in einer Subgruppe gegen die geschätzte kumulative Intensität dieser Subgruppe vergleicht.

Das Vorgehen beim MRP-Plot ist dabei folgendermaßen :

Die  $n$  Individuen werden bezüglich einer fixen  $g$ -klassigen Kovariable in Gruppen eingeteilt, wobei  $g \ll n$  gilt. Sei  $G$  eine  $g \times n$ -Matrix bestehend aus 0 und 1, die für jedes Individuum die Gruppenzugehörigkeit angibt. Die Summe der MRP in den durch  $G$  definierten Subgruppen ist

$$M_{res}^G(t) = G \cdot M_{res}(t). \quad (8)$$

Die Residualprozesse der einzelnen Subgruppen können dann jeweils gegen die Zeit geplottet werden. Die Kovarianzmatrix berechnet sich mit (7) und (8) als

$$V_{res}^G(t) = \sum_{T_k \leq t} G \left( J - Y(T_K) X(T_K) \right) I_k^D \left( J - Y(T_K) X(T_K) \right)' G'.$$

Für den Test auf die Validität des linearen Modells im Intervall  $(0,t)$  wird angenommen, daß die Kovarianzmatrix  $V_{res}^G(t)$  invertierbar und die Anzahl der Events in den einzelnen Gruppen nicht sehr klein ist. Falls das lineare Modell vorliegt, ist die Teststatistik

$$\left(M_{res}^G(t)\right)' \left(V_{res}^G(t)\right)^{-1} M_{res}^G(t)$$

asymptotisch Chi-Quadrat-verteilt mit  $g$  Freiheitsgraden. Ein Nachteil dieses Vorgehens ist zweifellos die Abhängigkeit (der Ergebnisse) von der subjektiven Gruppierung der Individuen.

## 5.2 Analyse der Martingalresiduen mit Hilfe eines linearen Modells

Zur analytischen Überprüfung der goodness-of-fit wird auf die Martingalresiduen eine multiple lineare Regression angewendet. Als unabhängige Variablen (Regressoren) werden lediglich die auffälligen Kovariablen aus dem linearen Aalen-Modell, ihre quadratischen Terme und die paarweisen Kreuzprodukte dieser Variablen berücksichtigt. Resultieren aus dieser Regressionsanalyse nicht-lineare Effekte für die Kovariablen, so ist dies ein Indiz für eine schlechte Anpassung des linearen Aalen-Modells. Falls der Einfluß einer Kovariablen nicht tatsächlich linear ist, führt die Anwendung des linearen Aalen-Modells zu einer "gekrümmten" Abhängigkeit zwischen dieser Kovariablen und den Residuen.

Sei  $Y_{res}$  die Matrix der unabhängigen Kovariablen, die in die multiple Regressionsanalyse eingehen. Die Schätzung der Kovariableneffekte führt damit zu

$$\beta' = [Y_{res}' Y_{res}]^{-1} Y_{res}' M_{res}$$

Für den Test auf nicht-lineare Effekte wird ein  $q$ -dimensionaler Subvektor  $\beta'_Q$  von  $\beta'$  betrachtet, der lediglich die nicht-linearen Effekte, d.h. die Effekte der quadratischen Terme und der Kreuzproduktterme enthält.



Sei  $C_Q$  die Kovarianzmatrix von  $\beta_Q$ . Dann ist die Testgröße für den Test auf nicht-lineare Effekte

$$\beta_Q' C_Q^{-1} \beta_Q$$

asymptotisch Chi-Quadrat verteilt mit  $q$  Freiheitsgraden. Ist der Test auf nicht-lineare Kovariableneffekte signifikant, wird eine Analyse mit dem linearen Aalen-Modell abgelehnt.

## 6 Beispiel

Im folgenden wird der lineare Zählprozeßansatz angewendet auf die in der Literatur mehrfach analysierten Daten der Stanford Heart Transplantation Study (Kalbfleisch&Prentice, 1980). In die Studie wurden in den Jahren 1967 bis 1974 103 Patienten rekrutiert, von denen 69 innerhalb des Beobachtungszeitraumes ein Herz transplantiert bekamen. Während der Studiendauer verstarben insgesamt 75 Patienten. Folgende Kovariablen wurden erfaßt :

Z1: Alter

Z2: vorangegangene Operation (0=nein, 1=ja)

Z3: Transplantation (0=nein, 1=ja)

Z4: Wartezeit bis zur Transplantation

Z5: Aufnahmejahr in die Studie

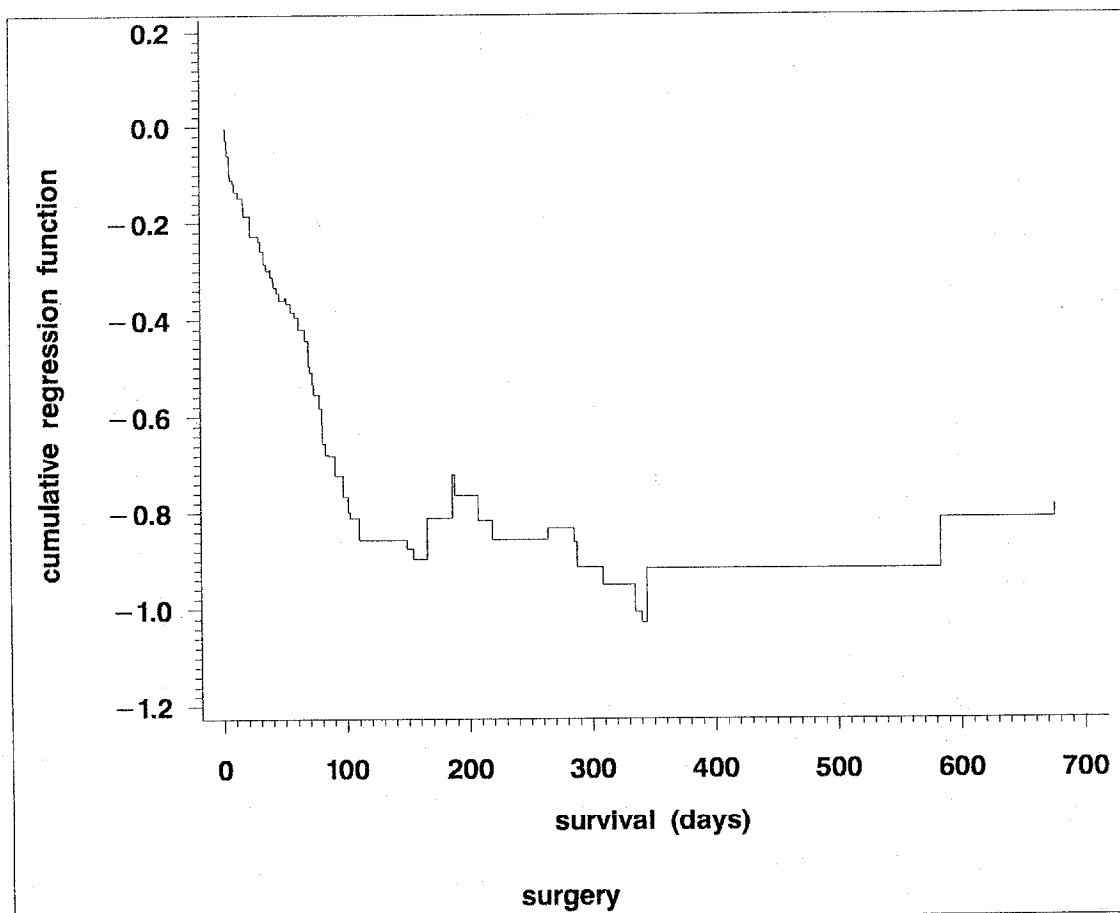
Zusätzlich wurden bei transplantierten Patienten (gewebespezifische) Mismatch-Variablen und ein Mismatch-Score erhoben, die allerdings in nachfolgender Analyse nicht berücksichtigt worden sind. Alle Kovariablen mit Ausnahme der Transplantation wurden zum Baselinezeitpunkt  $t=0$ , d.h. zu Beginn der Studie, ermittelt. Die Transplantation, die erst im Laufe der Beobachtungszeit bei manchen Patienten durchgeführt wurde und deren Zeitpunkt von Patient zu Patient verschieden ist, ist somit die einzige zeitveränderliche Kovariable in diesem Datensatz. In den linearen Zählpro-

zeßansatz gehen also die drei fixen Kovariablen Z1, Z2, Z5 und die zeitveränderliche Kovariable Transplantation, als Kombination aus Z3 und Z4 ein.

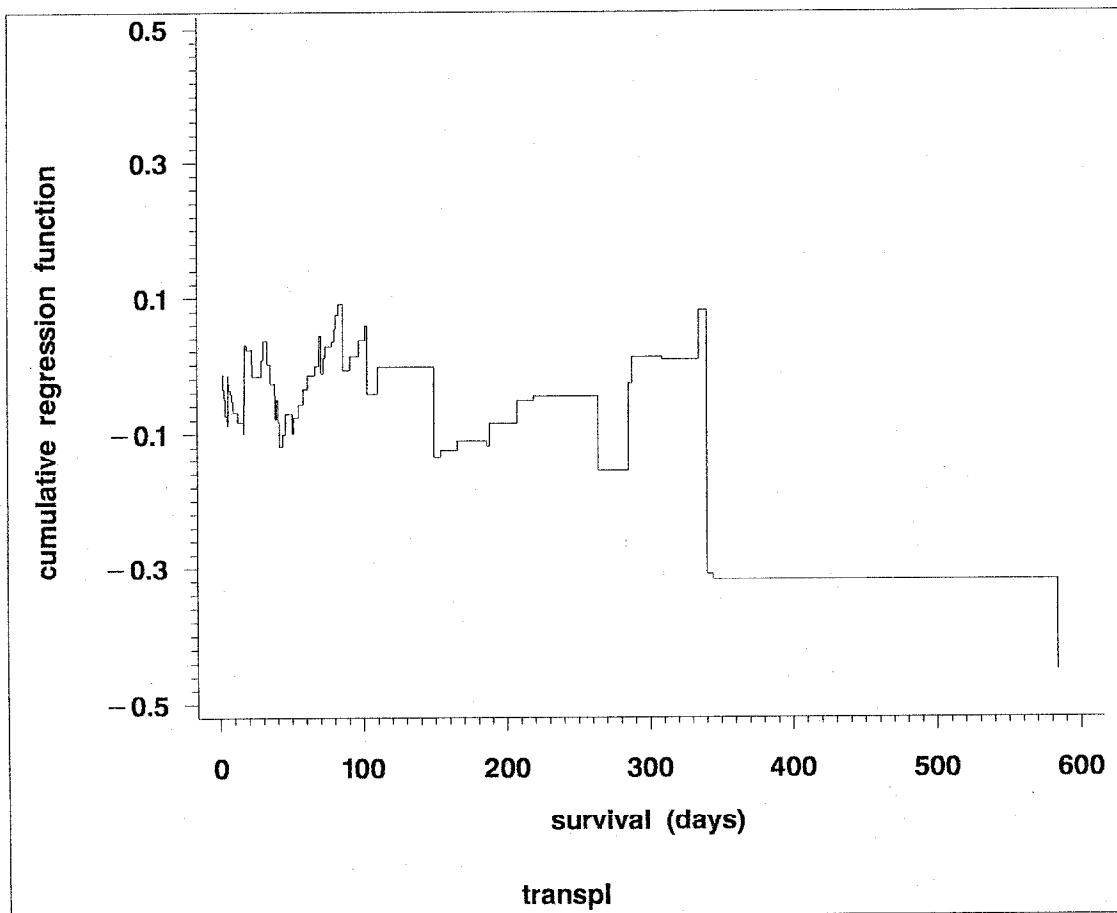
Die geschätzten kumulierten Regressionsfunktionen für die fixe Kovariable "vorangegangene Operation" und die zeitveränderliche Kovariable "Transplantation" sind in den Abbildungen 3 und 4 zu sehen.

Die negative Steigung der kumulierten Regressionsfunktion der Kovariable "vorangegangene Operation" in den ersten 100 Tagen zeigt, daß operierte Patienten im Vergleich zu nicht-operierten Patienten in diesem Zeitraum ein geringeres Risiko besitzen, dieser protektive Effekt der Operation jedoch nach den ersten 100 Tagen verschwindet. Der Test auf die Signifikanz (5) des Kovariableneffekts im gesamten Beobachtungszeitraum ergibt mit einer Teststatistik von  $z=-4.26$  einen auffällig negativen Einfluß einer vorangegangenen Operation auf die Intensität.

Bei Betrachtung der kumulierten Regressionsfunktion der Transplantation ist jedoch keine Steigung erkennbar, woraus gefolgert werden kann, daß eine Transplantation keinen Einfluß auf die hazard bzw. auf das Überleben besitzt. Mit der Teststatistik von  $z=-0.3$  ist auch im gesamten Beobachtungszeitraum kein signifikanter Effekt der Transplantation nachweisbar. Während das Alter mit einer Teststatistik von  $z=2.24$  einen auffällig positiven Einfluß auf die Intensität besitzt, d.h. höheres Alter wirkt sich auffällig negativ auf das Überleben aus, ist bezüglich des Aufnahmejahres in die Studie kein signifikanter Effekt feststellbar (Teststatistik  $z=-1.78$ ).



**Abbildung 3** : kumulierte Regressionsfunktion der Kovariable  
*"vorangegangene Operation "*



**Abbildung 4 :** kumulierte Regressionsfunktion der zeitveränderlichen Kovariable "*Transplantation* "

## 7 Diskussion

Bei Vorliegen von zeitveränderlichen Kovariablen kann die Analyse der Überlebenszeiten mit dem erweiterten time-dependent Cox-Regressionsmodell,

$$h_i(t) = h_0(t) \exp\left(\sum_{j=1}^r \beta_j x_{ij}(t)\right), \quad i = 1, \dots, n,$$

vorgenommen werden. Dabei bezeichnet  $h_i(t)$  die Hazardrate des  $i$ -ten Individuums,  $h_0(t)$  die unspezifizierte Baselinehazardrate,  $x_{ij}(t), j = 1, \dots, r$ , sind die Kovariablenwerte des  $i$ -ten Individuums zum Zeitpunkt  $t$  und  $\beta_j, j = 1, \dots, r$ , die Effekte der entsprechenden Kovariablen. Infolge der zeitlichen Struktur der Kovariablen entfällt im time-dependent Cox-Modell die Annahme der proportional hazards. Die Effekte der Kovariablen werden hier allerdings, wie auch im fixen Cox-proportional-hazards-Modell, über die Zeit konstant geschätzt, so daß eine tatsächliche oder eventuell vorhandene zeitliche Variation der Effekte nicht erkennbar ist. Das lineare Modell von Aalen hingegen gestattet eine Survivalanalyse unter Einbezug zeitveränderlicher Kovariablen, in der zudem von der zeitlichen Struktur der Kovariableneffekte profitiert werden kann.

Da das nichtparametrische lineare Modell anders strukturiert ist als das Cox-Regressionsmodell, ist jedoch ein direkter Vergleich beider Modelle nicht möglich. Während das Cox-Regressionsmodell multiplikativ ist, handelt es sich beim linearen Modell um einen rein additiven Ansatz. Es besteht demnach kein offensichtlicher Zusammenhang zwischen der Steigung der kumulierten Regressionsfunktion und dem relativen Risiko aus der Cox-Regression. So kann es vorkommen, daß das relative Risiko konstant bleibt, während die kumulierte Intensität aus dem linearen Modell steigt bzw. sinkt. Die Tatsache, daß die Intensität auch negative Werte annehmen kann, ist eine Schwäche des linearen Modells, das die Interpretation zusätzlich erschwert. Der ausschlaggebende Grund jedoch, weshalb - trotz dringendem Bedarf an solchen Modellen - ein derartig gut ausgearbeitetes Modell, wie das von Aalen, in der Praxis eher wenig Anwendung findet, war sicherlich das Fehlen von geeigneter Software.

## 8 Software

Das lineare Aalen-Modell wurde sowohl für den Fall von fixen als auch von zeitveränderlichen Kovariablen als Makro in SAS-IML programmiert. Bei Datensätzen mit mehreren zeitveränderlichen Kovariablen, die ihrerseits mehrmals switchen können, kann die derzeitige Version des Programmes (teilweise) längere Rechenzeiten erfordern. Eine Optimierung des Makros zur Verkürzung der Laufzeit ist in Bearbeitung.

Auf Anforderung wird die aktuelle Version (V10/95) der Makros zur Verfügung gestellt. Dazu wende man sich via Email an die erste Autorin ([ayd@ibe.med.uni-muenchen.de](mailto:ayd@ibe.med.uni-muenchen.de)).

## 9 Literatur

**Aalen, O.O.** (1989). *A linear regression model for the analysis of life times*. *Statistics in Medicine*, 8, 907-925.

**Aalen, O.O.** (1993). *Further results on the non-parametric linear regression model in survival analysis*. *Statistics in Medicine*, 12, 1569-1588.

**Altman, D.G., De Stavola, B.L.** (1994). *Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates*. *Statistics in Medicine*, 13, 301-341.

**Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.** (1993). *Statistical models based on counting processes*. Springer-Verlag, New York.

**Arjas, E.** (1988). *A graphical method for assessing goodness of fit in Cox's proportional hazards model*. *JASA*, 83, 204-212.

**Henderson, R., Milner, A.** (1991). *Aalen plots under proportional hazards*. *Applied Statistics*, 40, 401-409.

**Huffer, F.W., McKeague, I.W.** (1991). *Weighted least squares estimation for Aalen's additive risk model*. *JASA*, 86, 114-129.

**Kalbfleisch, J.D., Prentice, R.L.** (1980). *The statistical analysis of failure time data*. Wiley, New York.

**Mau, J.** (1986). *On a graphical method for the detection of time-dependent effects of covariates in survival data*. *Appl. Statist.*, 35, No.3, 245-255.

**Mau, J.** (1988). *A comparison of counting process models for complicated life histories*. *Applied Stochastic Models and Data Analysis*, vol. 4, 283-298.

**Therneau, T.M., Grambsch, P.M., Fleming, T.R.** (1990). *Martingale-based residuals for survival models*. *Biometrika*, 77, 147-160.