Lebensdaueranalysen mit zeitveränderlichen Kovariablen

Heinrich Potuschak

Adresse:
Institut für Angewandte Statistik, Universität Linz
A-4040 Linz/Austria
Tel.: +43 732 2468 8586, E-mail: Heinrich.Potuschak@jku.at

Zusammenfassung

Das Ziel von Regressionsanalysen innerhalb einer Lebensdaueranalyse ist - unter der Annahme eines theoretischen Ausfallgesetzes - , a) für jede Kovariable einen Koeffizienten zu schätzen, der die Größe, Richtung und Signifikanz angibt, mit der er die Lebensdauerverteilung beeinflusst, und b) den Verlauf der Überlebenswahrscheinlichkeit eines Individuums mit frei wählbarer Kovariablen-Konstellation zu prognostizieren. Solche Analysen werden von allen einschlägigen Programmpaketen angeboten, jedoch nur bei Vorliegen von konstanten Kovariablen, aber nicht von zeitveränderlichen; eine Ausnahme bildet das Cox-Modell, unter dem zwar eine Koeffizientenschätzung, aber keine weiteren Analysen angeboten werden.
Dieser Aufsatz ist eine Kurzfassung meiner gleichnamigen Dissertation und soll darstellen, wie mit zeitveränderlichen Kovariablen

a) in parametrischen Modellen die Koeffizienten zu schätzen sind,
b) im Cox-Modell die Survivorfunktion samt Standardfehlern zu schätzen ist,
c) in allen Modellen eine Residualanalyse machbar ist.

1. Die Kovariablen

1.1. Typen von Kovariablen:

Kalbfleisch und Prentice (1980, Kap.5.3) unterscheiden zwischen internen und externen Kovariablen. Interne Kovariablen ändern ihre Ausprägungen in wechselhafter kausaler Abhängigkeit voneinander und vom Zeitverlauf. Ihr Vorliegen erfordert spezielle Modellannahmen und wird hier nicht weiter berücksichtigt. Beispiele dafür sind Behandlungsmethoden, Transplantationen oder Medikamentendosierungen in Abhängigkeit vom Zustand eines Patienten – im Gegensatz zu Zufall oder Planung im Voraus.

Externe Kovariablen, mit quantitativen oder qualitativen Ausprägungen, werden als feste Größen betrachtet. Sie lassen sich weiter einteilen in:

1. Konstante Kovariablen. Beispiele dafür sind Geschlecht, Alter bei Behandlungsbeginn, Art des ersten Symptoms, erstmals betroffenes Organ, usw.
2. Globale Kovariablen. Ihre Ausprägungen ändern sich für alle Individuen gleichartig zur selben Kalenderzeit. Durch Längsschnitterhebungen entstehen individuell verschiedene Zeitpunkte. Beispiele sind die Gesetzeslage, Luftverschmutzung, medizinische Standards, usw.
3. Kontrollierte Kovariablen. Deren Ausprägungsänderungen werden vom Untersuchungsleiter im Voraus geplant, wie z.B. verschieden starke Dosierungen oder Stressfaktoren.
4. Definierte Kovariablen. Diese werden zu Analysezwecken funktional abhängig von der Zeit angesetzt, beispielsweise zur Hypothesenprüfung auf Zeitabhängigkeit der Hasardrate.
5. Zufällig veränderliche Kovariablen. Diese verändern ihre Ausprägungen auf nicht kausale Art. Beispiele dafür sind Familienstand, Einkommen, Blutdruck, Therapiewechsel, der Zustand Transplantiert/nicht T., usw. Während die ersten vier Typen von Natur aus keine internen Kovariablen sein können, hat bei diesem die kausale Unabhängigkeit tatsächlich gegeben zu sein.

1.2. Die Notation:

Die Daten seien von n Individuen mit den Indizes i=1,...,n und von je p Kovariablen mit k=1,...,p erhoben. Die Lebensdauern lauten t_i und die Zensierungsindikatoren c_i.
Bei konstanten Kovariablen sind die Elemente der Datenmatrix X mit x_{ik} indiziert. Bei zeitveränderlichen müssen folgende Daten je Individuum und je Kovariable erhoben werden: Die Anzahlen m_{ik} der Wertewechsel, die Zeitpunkte w_{ikj} dieser Wechsel und die fixen Ausprägungen x_{ikj} zwischen den Wechselzeitpunkten. Der letzte Index j der dreidimensionalen $\underline{w}$ läuft, falls $m_{ik}>0$, von 1 bis m_{ik} ; derjenige von $\underline{x}$ von 1 bis $m_{ik}+1$.

> Um die ungewohnte, aber notwendige Notation klarzustellen, folgt ein rein beispielhafter Datensatz des Individuums i: $t_i = 100$, $c_i = 1$, $\underline{m}_i = \{1, 0, 2\}$, $\underline{w}_i = \{ \{50\}, \{\}, \{30, 80\} \}$, $\underline{x}_i = \{ \{0, 1\}, \{-1\}, \{1., -0.5, 0.5\} \}$.
> Es handelt sich hier um drei Kovariablen, wobei sich die Ausprägung $\underline{x}_{i1}$ der ersten $m_{i1}=1$ Mal zum Wechselzeitpunkt $w_{i11}=50$ ändert, für $0 \le t < 50$ $x_{i11}=0$ beträgt und für $50 \le t \le 100$ $x_{i12}=1$.

Für Zeitintervalle, in denen eine Ausprägung nicht konstant, sondern linear verläuft, müssten zusätzlich die Anstiege a_{ikj} bekannt sein. Als Ausprägung im Zeit-intervall $[w_{i,k,j-1}, w_{ikj})$ ist dann $x_{ikj} + a_{ikj} \cdot t$ zu berücksichtigen.

2. Die Ausfallzeit

2.1. Die Lebensdauer:

T sei eine stetige positive Zufallsvariable. Sie ist die Zeit zwischen Nullpunkt und Eintreffen des Ereignisses, d.h. die Dauer eines Zustands. Beispiele dafür sind: Die Zeit zwischen Spitalsaufnahme und Heilung, zwischen Behandlungsbeginn und Tod, zwischen Ansteckung und Ausbruch der Krankheit, zwischen Großjährigkeit und Heirat, zwischen Heirat und Scheidung, zwischen Verlust des Arbeitsplatzes und Beginn an einem neuen, usw. Die Wahrscheinlichkeitsverteilung von T kann durch jede der fünf folgenden Funktionen beschrieben werden, d.h. aus jeder einzelnen sind die übrigen vier eindeutig berechenbar:

1. Die Verteilungsfunktion $F(t) = P(T \le t)$.
2. Die Überlebens- oder Survivorfunktion $S(t) = P(T \ge t)$, also die Wahrscheinlichkeit, den Zeitpunkt t zu er- oder überleben.
3. Die Dichte $f(t) = \frac{\partial F(t)}{\partial t}$. Im diskreten Fall gibt $f(t_j)$ die unbedingte Wahrscheinlichkeit an, im Zeitintervall t_j auszufallen.
4. Die Intensitäts-, Risiko- oder Hasardrate $h(t) = \lim_{\Delta \downarrow 0} \frac{P(t \le T < t + \Delta | T \ge t)}{\Delta}$. Im diskreten Fall gibt $h(t_j)$ die bedingte Wahrscheinlichkeit an, im Zeitintervall t_j auszufallen, gegeben sein Beginn wurde erlebt.
5. Die integrierte oder kumulative Hasardrate $H(t) = \int_0^t h(t)dt$.

Die wichtigsten theoretischen Beziehungen, mit denen obige fünf Funktionen eindeutig zusammenhängen, lauten:

$$S(t) = 1 - F(t) = -\int_0^t f(t)dt = e^{-H(t)}, \quad H(t) = -\ln[S(t)], \quad \text{und} \quad h(t) = \frac{f(t)}{S(t)}.$$

Die folgende Situation wird nicht weiter abgehandelt: T ist als diskrete Zufallsvariable zu definieren, wenn entweder aus erhebungstechnischen Gründen die Zeiten in zu langen Intervallen gemessen sind, oder die Lebensdauer das Ergebnis eines Zählprozesses ist – wie z.B. die Anzahl der Versuche bis zum Eintreffen eines Erfolges. Eine derartige Analyse verliefe

wesentlich gleichartig, nur müssten $H(t_i) = \sum_{j=1}^{i} h(t_j)$ als kumulierte Summe und $S(t_i) = \prod_{j=1}^{i-1} [1 - h(t_j)]$ als Produkt berechnet werden, wonach $S(t_i) \neq e^{-H(ti)}$ gilt. Mit der numerisch kleinen Änderung $h(t) = -\ln[1-h(t)]$ bliebe auch dieser Zusammenhang weiter aufrecht.

2.2. Die Zensierung

Typisch für Lebensdauern sind Beobachtungen zensierter Zeiten. Im Fall von Rechtszensuren sind es solche, von denen nur eine Mindestdauer bekannt ist. Gründe dafür sind a) Stichtage, an denen die Datenerhebung endet, b) geplante Versuche, die nach einer bestimmten Zeit oder Anzahl von Beobachtungen enden, und c) zufällige Ausfälle von Individuen, z.B. durch Unfalltod, Auswanderung oder freiwilligem Behandlungsende.
Um während der Analyse die echten Ausfälle von den zensierten zu unterscheiden, wird jedem Individuum eine Binärzahl c_i zugeordnet. Für eine exakte Modellbeschreibung und für die Simulation von Lebensdauern wird eine Zufallsvariable C mit einer von T unabhängigen Verteilung angenommen: die beobachtete Ausfallzeit t_i ist dann gleich dem Min[t,c], und c_i wird 1 bzw. 0 gesetzt, wenn t kleiner bzw. größer als c, die Realisation von C, ist.

3. Modellbildung

3.1. In parametrischen Modellen

In der Verteilungsannahme wird eine bestimmte Verteilung samt ihren Scharparametern festgelegt. Von jedem Individuum der Stichprobe wird angenommen, nach diesem Verteilungsgesetz auszufallen; die individuellen Ausfallverteilungen unterscheiden sich nur in einem (oder mehreren) Parameter, dessen Größe von den Kovariablen abhängt.
In einer Strukturannahme wird dieser Verteilungsparameter aus den Scharparametern ausgewählt, und formuliert, in welcher algebraischen Form er von den beobachteten Kovariablen beeinflusst wird. Bei Vorliegen mehrerer Scharparameter wird üblicherweise der Lageparameter λ ausgewählt, möglicherweise noch einen Formparameter, meist mit α bezeichnet. Die übrigen festen sowie die Strukturparameter gilt es zu schätzen. Dies sind die Koeffizienten $\underline{\beta}$ und beschreiben das Ausmaß, in dem die einzelnen p

zeitkonstanten Kovariablen auf den gewählten Verteilungsparameter wirken. Neben anderen Möglichkeiten setzt man ihren Einfluss loglinear an:

mit $\lambda(\underline{x}) = \exp[-\underline{\beta}.\underline{x}']$ an.

Dieser Einfluss ist loglinear und kann als $\exp[-\sum_{k=0}^{p}\beta_k \cdot x_k]$ oder $\prod_{k=0}^{p}\exp[-\beta_k \cdot x_k]$ gelesen werden. Mittels $\alpha(\underline{x}) = \exp[-\sum_{k=p+1}^{2p+1}\beta_k \cdot x_k]$ kann zusätzlich noch der Formparameter strukturiert werden. Auch andere, etwa additive Strukturgleichungen mit beliebigen Funktionen, etwa $1/(1+\underline{\beta}.\underline{x}')$ sind plausibel und kommen in Frage; jedoch wird das Schätzverfahren aufwändiger, da Nebenbedingungen einzuhalten sind, um den Definitionsbereich des modellierten Parameters nicht zu verletzen.
Wie bei allen Regressionsanalysen ist ein zusätzlicher Koeffizient β_0 zusammen mit einer Dummyvariablen $X_0 = \underline{1}$ eingeführt worden. Nur bei der Interpretation der Ergebnisse ist darauf zu achten, ob die Kovariablen als Rohwerte, zentriert oder standardisiert angesetzt worden sind. Kovariablen, deren Ausprägungen κ Kategorien aufweisen, sind in $\kappa-1$ dichotome aufzulösen.

3.2. Im Cox-Modell

Dieses Modell wurde 1972 von D.R. Cox in (JRSS B, 34, 187-201) eingeführt. Es wird keine parametrische Verteilungsannahme getroffen. Als Verteilungsparameter wird die Hasardrate gewählt, ohne über ihren konkreten Verlauf eine Annahme zu treffen; sie wird erst im Anschluss an die Koeffizienten geschätzt, und nur zu den Zeitpunkten, an denen Ereignisse stattfinden. Als Strukturannahme wird ein log-linearer Einfluss der Kovariablen gewählt, der multiplikativ auf diese Basis-Hasardrate einwirkt:

$h[t|\underline{x}(t)] = h_o(t)\cdot\exp[\underline{\beta}.\underline{x}(t)']$.

$\underline{\beta}.\underline{x}(t)'$ ist das Vektorprodukt $\sum_{k=1}^{p}\beta_k \cdot x_k(t)$. Ein Koeffizient β_0 wird nicht angesetzt, da er im von Kovariablen unabhängigen Faktor $h_o(t)$ aufgenommen ist.
$\underline{x}(t)$ bezeichnet den Gesamtverlauf einer Kovariablen für $0 \le T \le t$, aber $h[t|\underline{x}(t)]$ ist die Hasardrate zum Zeitpunkt t, gegeben die Ausprägungen genau zum Zeitpunkt t. Da h(t) als bedingte Grenzwahrscheinlichkeit definiert ist, im Intervall $[t,t+\Delta)$ auszufallen, gegeben der Zeitpunkt t ist erlebt, wird sie von extern - aber nicht intern - zeitveränderlichen Kovariablen auch nur mit deren Ausprägungen zum Zeitpunkt t beeinflusst.

Diese Strukturannahme impliziert folgende Verteilungseigenschaft: Man denke an die Kovariablen $\underline{x}(t)$ und $\underline{y}(t)$ zweier Individuen, die ihre Ausprägungen entweder nie oder gleichzeitig und proportional ändern, und bilde den Quotienten der parametrisierten Hasardraten. Es kürzt sich $h_o(t)$ und übrig bleibt eine abschnittsweise von $\underline{w}$ begrenzte konstante Funktion. Da die Schätzung der Koeffizienten diese PH-Eigenschaft - d.h. theoretisch proportional verlaufende Hasardraten von Individuen mit fixierten Kovariablen - ausnutzt, bedeutet dies, dass sie theoretisch erfüllt sein und geprüft werden muss; entweder mit einem Modelltest oder durch inhaltliche Überlegung. Ein Gegenbeispiel dafür sind die nicht proportional verlaufenden Raten zweier Arbeitsloser mit verschiedener Ausbildung, wenn als Lebensdauer die Zeit der Stellensuche analysiert wird.

3.3. Identifikation

Wählt man die bekanntesten Lebensdauerverteilungen aus, unter deren Modellannahme auch alle Programmpakete Analysen anbieten, und betrachtet man von jeder die fünf theoretischen Funktionen gegen die Zeit aufgetragen, findet man nur bezüglich der Hasardrate charakteristische Unterschiede, die zu einer Typisierung geeignet sind. Bezüglich ihres Verlaufes lassen sich die Verteilungen in solche mit konstanter, steigender, fallender, eingipfeliger und badewannenförmiger Hasardrate einteilen. Nun kann aus der Kenntnis der empirisch geschätzten Hasardrate auf den Typus des theoretischen Verteilungsgesetzes geschlossen werden, was die Menge der in Frage kommenden Verteilungen stark einengt. Da die empirische Hasardrate keineswegs homogen ist und durch die Raten der einzelnen Kovariablen-Konstellationen vermischt wird, ist es notwendig, nach allen Ausprägungen (stetige werden diskretisiert) aller Kovariablen zu gruppieren, und die Hasardraten jeder Gruppe zu betrachten. Stellen sich beispielsweise alle homogenen Hasardraten als - naturgemäß in verschiedenen Ausführungen - steigend heraus, kommt das Modell einer Weibullverteilung in Frage. Da auch homogene empirische Hasardraten $\hat{h}(t)$ stark fluktuieren, sind sie entweder geglättet darzustellen, oder noch einfacher ist es, ihre kumulierten $\hat{H}(t)$ zu plotten und aus deren Krümmungsverhalten auf die Steigungen $\hat{h}(t)$ zu schließen. Kann kein gemeinsamer Hasard-Typus gefunden werden, muss ein verteilungsfreies Modell gewählt werden. Liegen zeitveränderliche Ausprägungen vor, kann nach ihnen nicht oder höchstens nach ihren Mittelwerten gruppiert werden. Dies kann in beiden Fällen zu Fehlschlüssen führen, wie sie bei Mischverteilungen üblich sind.

3.4. Modelltest

Wie bei jedem Regressionsmodell wird mittels einer Residualanalyse geprüft, ob die Verteilung der Residuen – d.h. die Abweichungen der abhängigen Variablen von ihren Schätztwerten – modellkonform ist. Ein Normalverteilungstest der Differenzen $\hat{t}_i - t_i$ wie in LR-Modellen ist wegen eventuell zensierter t_i kaum möglich und wird in der Modellannahme auch nicht verlangt. Stellt man die geschätzten individuellen Überlebenswahrscheinlichkeiten $\hat{S}[t_i | \underline{x}_i(t)]$ formal als $\exp\left[-1 \cdot \hat{H}[t_i | \underline{x}_i(t)]\right]$ dar, erkennt man am Bau dieser Funktion die Survivorfunktion $S(y) = e^{-\lambda \cdot y}$ einer Exponentialverteilung mit $\lambda=1$ und $\hat{H}$ in der Rolle als Zufallsvariable Y. Die empirischen Hasardraten $\hat{H}_i$ sind ebenso zensiert wie die beobachteten t_i und ihre Survivorfunktion wird verteilungsfrei als Kaplan-Meier-Schätzer berechnet. Trägt man ihre negativen Logarithmen gegen die Zeit auf, dürfen sie nur zufällig um eine 45%-Gerade streuen; diese ist die theoretische kumulative Hasardrate $H(y) = y$ der Residuen. Dieses Vorgehen ist kein Test im engeren Sinn und kann nur eine grob verletzte Modellannahme aufdecken, da auch mit direkt simulierten exponentialverteilten Zeiten eine Streuung um diese Gerade entstehen würde.

4. Parameterschätzung

4.1. Das Schätzprinzip

Das Maximum-Likelihood-Prinzip lautet, die Struktur- und eventuellen festen Parameter so zu schätzen, dass die Funktion $ML(\underline{b})$, das ist das Produkt der Wahrscheinlichkeiten aller beobachteten Lebensdauern, maximal wird. Dazu werden die Dichten $f[t_i|\underline{x}_i(t)]$ der vollständigen, und die Überlebenswahrscheinlichkeiten $S[t_i|\underline{x}_i(t)]$ der zensierten Zeiten multipliziert:

$$ML(\underline{b}) = \prod_{i \in D} f_i \cdot \prod_{i \in C} S_i \; .$$

Mit $\underline{x}(t)_i$ ist der Gesamtverlauf der Kovariablenausprägungen $\underline{x}_i$ bezeichnet, mit D und C die Indexmengen aller Individuen mit vollständigen bzw. zensierten Lebensdauern. Dividiert man das erste Produkt durch $\prod_{i \in D} S_i$ und multipliziert man das zweite damit, entsteht daraus $ML(\underline{b}) = \prod_{i \in D} h_i \cdot \prod_{i=1}^{n} S_i$. Da der Logarithmus $LL(\underline{b})$ an der selben Stelle wie $ML(\underline{b})$ ein Maximum besitzt, lautet die Zielfunktion nun

$LL(\underline{b}) = \sum_{i \in D} \ln h_i - \sum_{i=1}^{n} H_i$.

h_i und H_i sind Funktionen von $\underline{b}$ und $\underline{x}_i(t)$ an den Stellen t_i. $LL(\underline{b})$ wird für jene $\underline{b}$ maximal, für welche die ersten Ableitungen $\varphi(k) = \frac{\partial LL(\underline{b})}{\partial b_k}$ Null werden. Dies ist ein quadratisches nichtlineares Gleichungsystem und eindeutig lösbar, wenn die Informationsmatrix mit den Elementen $Inf_{k\ell} = \frac{\partial^2 LL(\underline{b})}{\partial b_k \cdot \partial b_\ell}$ negativ definit ist. Eine von mehreren Optimalitätseigenschaften derart geschätzter Parameter ist die asymptotische Normalverteilung der Strukturparameter $\underline{b} \sim N(\underline{\beta}, -Inf^{-1})$. Eine weitere ist die asymptotische χ^2_{p-q}-Verteilung von $-2\cdot(LL_p - LL_q)$, wobei p die Parameterzahl eines Obermodells ist, und q diejenige eines reduzierten.

4.2. Schätzung in parametrischen Modellen

Um die Likelihoodfunktion aufzustellen und maximieren zu können, muss der Einfluss zeitveränderlicher Kovariabler auf die Verteilung von $T|\underline{x}(t)$, besonders auf ihre Funktionen h, H und S hergeleitet werden. Der Funktionswert $h[t_i|\underline{x}_i(t)]$ ist durch die Verteilungs- und Strukturannahme und die Kovariablenausprägungen $\underline{x}(t_i)$ – wie in Kap. 3.2. erklärt –zum Zeitpunkt t_i bestimmt. Der Verlauf von $h[t|\underline{x}(t)]_i$ ist abschnittsweise einheitlich definiert, wobei seine Intervallgrenzen die Vereinigungsmenge der Wechselzeitpunkte $\underline{w}_i$ sind, da sich die Hasardrate bei jedem Wertewechsel ändert. Definiert man die dreidimensionalen $\underline{w}$ mittels $\bigcup_{k=1}^{p} w_{ikj} = w_{ij}$ in individuell einheitliche mit Umfängen m_i um, lässt sich die kumulative Hasardrate als Funktionswert einer stückweise stetigen Funktion berechnen. Sind $H_j(t)$ die kumulativen Hasardraten von m_i gedachten Individuen mit durchwegs fixen Kovariablen $\underline{x}_j$, dann gilt in der Episode $w_{j-1} \le t < w_j$

$$H[t|\underline{x}_j(t)] = \sum_{\ell=1}^{j-1} H_\ell(w_\ell) + \int_{w(j-1)}^{t} h_j(u)du = H[w_{j-1}|\underline{x}(t)] + H_j(t) - H_j(w_{j-1}).$$

Mit den Definitionen $w_{i,0} = 0$ und $w(i,m_i+1) = t_i$ lautet der für die LL-Funktion benötigte Funktionswert an der Stelle t_i in programmtechnisch bequemer Form

$$H_i = H[t_i|\underline{x}_i(t)] = \sum_{j=1}^{mi+1} \left[H_{ij}(w_{ij}) - H_{ij}(w_{i,j-1}) \right].$$

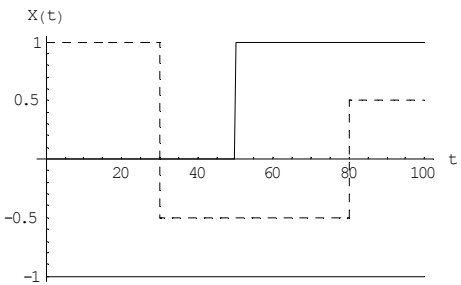
Wenn am Ende des Schätzverfahrens die ML-Schätzer gefunden sind, haben sie, wie in Kap. 4.4. erklärt, die Eigenschaft von Residuen.

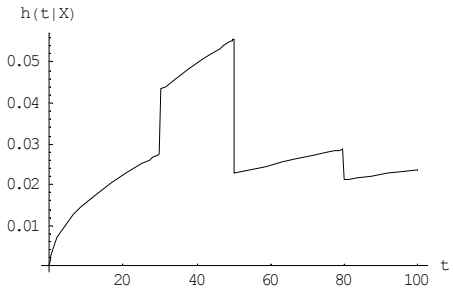
Um den Verlauf von $S[t|\underline{x}_i(t)]$ graphisch darzustellen, ist die Beziehung $S = e^{-H}$ anzuwenden und $w(i,m_i+1)$ nach oben offen zu lassen: $S(t|\underline{x}_1) = \exp[-H(t|\underline{x}_1)]$, $S(t|\underline{x}_2) = S(w_1|\underline{x}) \cdot \frac{S_2(t)}{S_2(w_1)}$, $S(t|\underline{x}_3) = S(w_2|\underline{x}) \cdot \frac{S_3(t)}{S_3(w_2)}$, usw.

Mit diesen Bildungsgesetzen kann unter jeder parametrischer Verteilungsannahme nun die LL-Funktion samt ihren Fehlergleichungen und der Informationsmatrix aufgestellt werden. Für vier spezielle Verteilungen ist dies theoretisch, mit Rechenprogrammen und Simulationsstudien in meiner Dissertation dargestellt. Unter Annahme einer Exponentialverteilung ist das Vorgehen identisch mit der Methode des Episoden-Splittings, beschrieben in Blossfeld/Hamerle/ Mayer (1986). Unter anderen Modellannahmen werden Analysen von den großen Programmpaketen nicht angeboten.

Beispiel 4.2.

Um die Konstruktion der Funktionen h, H und S zu demonstrieren, wird mit den Kovariablen des Beispiels aus Kap. 2.2. fortgefahren. Ihr zeitveränderlicher Verlauf ist im folgenden linken Bild dargestellt. Die Anzahl der vereinigten Wechselzeitpunkte ist $m_i = 3$, die Zeitabschnitte mit einheitlichem Verlauf sind von $\underline{w}_i = \{30, 50, 80\}$ eingegrenzt, die abschnittsweise konstanten Kovariablen lauten $\underline{x}_i = \{ \{0,-1,1\}, \{0,-1,-0.5\}, \{1,-1,-0.5\}, \{1,-1,0.5\} \}$.





Als Verteilungsannahme ist eine Weibullverteilung mit den Scharparametern λ und α gewählt. Um die Funktionszusammenhänge zu demonstrieren, wird als Ausgangs-Funktion

$S(t) = \exp[-(\lambda \cdot t)^{\alpha}]$ verwendet.

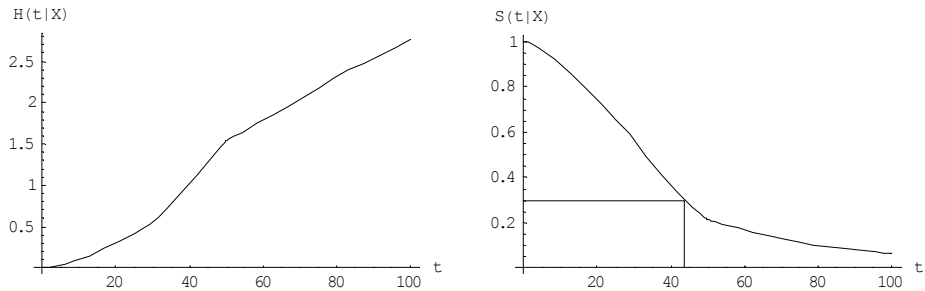
$H(t) = (\lambda \cdot t)^{\alpha}$ folgt aus $-\ln[S(t)]$.

$h(t) = \alpha \cdot \lambda^{\alpha} \cdot t^{\alpha-1}$ folgt aus $\frac{\partial H(t)}{\partial t}$.

$f(t) = \alpha\cdot\lambda^{\alpha}\cdot t^{\alpha-1}\cdot\exp[-(\lambda\cdot t)^{\alpha}]$ folgt aus $\frac{-\partial S(t)}{\partial t}$ oder aus $h(t)\cdot S(t)$.

Die Strukturannahme lautet $\lambda(\underline{x}) = \exp[-\underline{\beta}.\underline{x}']$ und α fest.
Es werden α=1.5 und die Strukturparameter $\underline{\beta}$ ={4., 0.6, 0.4, 0.2} gewählt. Diese Werte können ebenso als Schätzer $\hat{\alpha}$ und $\underline{b}$ angesehen werden, wie sie im Lauf des Iterationsverfahrens zur Maximierung von $LL(\underline{b},\hat{\alpha})$ entstehen könnten.

Die abschnittsweise konstanten Größen $\underline{\beta}.\underline{x}'$ ergeben sich als {3.8, 3.5, 4.1, 4.3}. Aus ihnen folgen die strukturierten Verteilungsparameter $\underline{\lambda}(\underline{x}) \simeq$ {0.022, 0.030, 0.017, 0.014} unter der Annahme durchwegs konstanter $\underline{x}$. Daraus berechnet sich die abschnittsweise einheitlich verlaufende Hasardrate $h[t|\underline{x}_i(t)]$, wie sie oben rechts abgebildet ist. Im letzten Abschnitt z.B. lautet sie $1{,}5\cdot 0{,}014^{1,5}\cdot t^{0,5}$. Ihr Funktionswert h_i, der in der LL-Gleichung gebraucht wird, berechnet sich an der Stelle t_i=100 als 0.024.
Im nächsten linken Bild ist die an den Stellen $\underline{w}_i$ geknickt verlaufende Funktion $H[t|\underline{x}_i(t)]$ abgebildet. Im ersten Abschnitt lautet sie $(0{,}022\cdot t)^{1,5}$ und im zweiten $(0{,}022\cdot 30)^{1,5} + (0{,}030\cdot t)^{1,5} - (0{,}030\cdot 30)^{1,5}$. Für die LL-Gleichung wird der Funktionswert $H_i \simeq 2.76$ an der Stelle Stelle t_i=100 gebraucht.



Am rechten Bild lässt sich veranschaulichen, wie Lebensdauern simuliert werden. Ausgehend von einer Realisation einer in (0,1) gleichverteilten Zufallsvariablen U, z.B. u_i=0.3, auf der Ordinate sucht man horizontal den Schnittpunkt mit der theoretischen Survivorfunktion, dann vertikal den Punkt auf der Zeitachse. In diesem Fall würde $t_i \simeq 43.7$ realisiert werden; das selbe Ergebnis findet man vom Punkt -ln(0.3) aus in der linken Abbildung. Bei numerischer Vorgangsweise hat man t_i aus der Gleichung $S[t_i|\underline{x}_i(t)] = u_i$ zu lösen. Bei fixen Kovariablen ist dies unter vielen Verteilungsannahmen in geschlossener Form möglich; bei zeitveränderlichen erst dann, wenn der Abschnitt $w_{i,j-1} < u_i \le w_{ij}$ festgestellt ist .

4.3. Parameterschätzung im Cox-Modell

Formt man wie in Lawless (1982, S. 357) die ML-Funktion in ein Produkt von drei Faktoren um, und maximiert man nur den ersten, die partial likelihood

$$PL(\underline{b}) = \prod_{i\in D} \frac{\exp[\underline{b}.\underline{x}_i(t_i)']}{\sum_{j\in R(i)} \exp[\underline{b}.\underline{x}_j(t_i)']} = \prod_{i\in D} \frac{E_{ii}}{\sum_{j\in R(i)} E_{ij}},$$

dann weisen, wie von mehreren Autoren gezeigt wurde (siehe Kalbfleisch/ Prentice, 1980, Kap. 4.2.3), derart geschätzte Strukturparameter asymptotische Eigenschaften wie solche der vollen Likelihood auf; diese und die des diskreten Modells werden in meiner Dissertation weiter behandelt.

$\underline{x}_i(t_i)$ sind die Kovariablen-Ausprägungen des Individuums i, aktualisiert zu seinem Ereignis-Zeitpunkt t_i . $\underline{x}_j(t_i)$ sind die Ausprägungen von Individuen der Risikomenge R_i (Ausfallszeiten $t_j \geq t_i$), aktualisiert zum Zeitpunkt t_i . Denkt man sich den Zähler und jeden Summanden des Nenners der PL-Funktion mit $h_o(t_i)$ multipliziert, kann man nach Cox(1972, S. 191) jeden ihrer Faktoren als bedingte Wahrscheinlichkeit deuten, Individuum i fällt aus, gegeben ein Individuum aus R_i fällt aus. Da deren Ausprägungen für jedes $i \in D$ aktualisiert werden müssen, wird dies beim Iterationsverfahren zur Parameterschätzung bei jeder Iteration und bei jeder Ereigniszeit durchgeführt. Eine wie in parametrischen Modellen einmalige Vorbereitung abschnittsweise konstanter Kovariabler wäre zu aufwändig, da dies eine Dreiecksmatrix (mit p-dimensionalen Elementen) mit der Dimension n(E), der Anzahl aller Ereigniszeitpunkte, verlangen würde.

Die Schätzmethoden bei Beobachtung verbundener Werte, von Vielfachheiten, sowie bei notwendig gewordener Schichtung sind mit zeitveränderlichen Kovariablen die selben wie mit konstanten; ebenso alle Signifikanztests. Eine Schichtung nach den Ausprägungen einer zeitveränderlichen Kovariablen ist ebenso unmöglich wie der graphische Logminuslog-Plot zur Überprüfung der PH-Eigenschaft einer solchen; die Zeitunabhängigkeit der Hasardrate kann wie bei einer konstanten Kovariablen $\underline{x}_k$ mittels zusätzlicher und definiert zeitabhängiger Kovariabler $\underline{x}_{p+1} = \underline{x}_k \cdot t$ getestet werden.

Im Anschluss an die Koeffizientenschätzung können mit jedem bekannten Programmpaket die Berechnungen der Residuen, Survivorfunktion (als Basisfunktion, als geschätzte Funktion von Mitgliedern der Lernstichprobe, oder als Prognose für neue Individuen) und deren Standardfehler (zu Zeitpunkten, an denen Ereignisse beobachtet sind) durchgeführt werden. Da dies nur bei Vorliegen konstanter Kovariabler angeboten wird, folgen Vorschläge für die Vorgangsweise mit veränderlichen. Sie läuft ohne zusätzliche Annahmen und nach dem gleichen Prinzip ab: Die Linkfunktionen $\exp[\underline{b}.\underline{x}_i] = E_i$, die bei Zeitkonstanz aus den Summen oder Produktzeichen herausgehoben sind, werden

in diese hineingezogen und die Ausprägungen $\underline{x}_i(t)$ je Ereignis-Zeitpunkt t_j als E_{ij} aktualisiert.

5. Residuen und Survivorfunktion im Cox-Modell

In parametrischen Modellen sind $\hat{H}[t_i|\underline{x}_i(t)]$ und $\hat{S}[t|\underline{x}_i(t)]$ nach der Schätztung der Strukturparameter eindeutig bestimmt; die Residuen stehen als Summanden H_i der maximierten LL-Funktion sogar numerisch fest. Im Cox-Modell muss noch die nichtparametrische Basishasardrate $h_o(t)$ geschätzt werden; aus ihr folgen die Funktionen H_o und S_o, und aus ihnen folgen H und S.

5.1. Die Residuen

Die Herleitung richtet sich nach Klein und Moeschberger (1997, Kap. 8 und 11.2), wo auf die Möglichkeit hingewiesen wird, die Residuen bei Vorliegen von zeitveränderlichen Kovariablen und/oder schichtweise zu berechnen.
Formuliert man die Likelihood als Funktion von $\underline{b}$ und $\underline{h}_o$, und hält man die aus der PL geschätzten $\underline{b}$ fest, und schätzt man $\underline{h}_o$ unter dem Gesichtspunkt, $LL(\underline{h}_o)$ zu maximieren, ergeben sich nach einigen Umformungen die n(E) Fehlergleichungen $\frac{\partial LL(\underline{h}_0)}{\partial[h_0(t_v)]} = 0$. Sie sind sukzessive und geschlossen lösbar:

$$h_o(t_v) = d_v \,/\, \sum_{i \in R(v)} E_{iv} \,.$$

Hierbei sind t_v die n(E) verschiedenen Ereigniszeitpunkte, d_v die Anzahlen der Ereignisse zu diesen Zeitpunkten, und R_v die Indexmengen aller Ausfallzeiten $t_i \geq t_v$. Die Abkürzung $E_{iv} = \exp[\underline{b}.\underline{x}_i(t_v)']$ gilt weiterhin.

Da $h_o(t)$ als Funktion mit Sprüngen an den Stellen t_v geschätzt wird, ergibt sich die kumulierte Basishasardrate als Treppenfunktion $H_o(t) = \sum_{t(v) \leq t} h_0(t_v)$. Die Berechnung der Basis-Survivorfunktion $S_o(t)$ mittels der Beziehung $S = e^{-H}$ ist nicht üblich, da der bequeme Zusammenhang $S(t|\underline{x}) = S_o(t)^{\exp[\underline{b}.\underline{x}']}$ verloren ginge; allerdings besteht er nur bei zeitkonstanten Kovariablen und die numerischen Unterschiede verschwänden mit zunehmendem Stichprobenumfang.
Zieht man den zweiten Faktor der Modellgleichung $h[t|\underline{x}(t)] = h_o(t) \cdot \exp[\underline{b}.\underline{x}(t)']$ in die Kumulation von $H[t|\underline{x}(t)]$ hinein, berechnen sich die Residuen als

$$\hat{H}[t_i|\,\underline{x}_i(t)] = \sum_{t(v) \leq t(i)} \frac{d_v \cdot E_{iv}}{\sum_{j \in R(v)} E_{jv}} \,.$$

t_i ist eine individuelle Ausfallzeit; ist sie zensiert, ist auch das Residuum derart zu behandeln (siehe Kap. 4.4). Mit $t_\nu \le t_i$ werden alle Ereigniszeitpunkte kleiner-gleich als t_i durchlaufen, mit $j \in R_\nu$ alle Ausfallzeiten größer-gleich als t_ν.

5.2. Die Survivorfunktion

Die Herleitungen folgen Kalbfleisch und Prentice (1980, Kap. 4.3). Liegen ausschließlich zeitkonstante Kovariable vor, lässt sich die Survivorfunktion als potenzierte Basisfunktion berechnen:

Aus $h(t|\underline{x}) = h_o(t)\cdot\exp[\underline{b}.\underline{x}']$ und $H(t|\underline{x}) = H_o(t)\cdot\exp[\underline{b}.\underline{x}']$ folgt

$S(t|\underline{x}) = \exp[-H(t|\underline{x})] = S_o(t)^{\exp[\underline{b}.\underline{x}']}$.

$S_o(t)$ wird als Treppenfunktion mit Stufen zu den Ereigniszeitpunkten $t_1 < t_2 < ... < t_{n(E)}$ geschätzt. Sie gibt die unbedingte Wahrscheinlichkeit an, mit der ein Individuum mit Kovariablenausprägung $\underline{x} = \underline{0}$ den Zeitpunkt t überlebt.

$S_o(t_\nu) = \prod_{t(j)\le t(\nu)} \alpha_j$ ist das Produkt der bedingten Wahrscheinlichkeiten, jeden Ereigniszeitpunkt bis t_ν zu er- oder überleben. An Stelle von $1-h(t|\underline{x})$, die im stetigen Fall keine Wahrscheinlichkeiten sind, werden solche als $\alpha_j = P(T\ge t_j|T\ge t_{j-1})$ eingeführt. Mit ihnen lassen sich $f(t_i|\underline{x}_i)$ und $S(t_i|\underline{x}_i)$ als Produkte formulieren und in die Zielfunktion $ML(\underline{b},\underline{\alpha})$ einsetzen.

Hält man nun die aus der PL geschätzten $\underline{b}$ fest, und schätzt man $\underline{\alpha}$ unter dem Gesichtspunkt, $ML(\underline{\alpha})$ zu maximieren, ergeben sich nach einigen Umformungen die n(E) sukzessiven Fehlergleichungen

$$\frac{\partial LL(\underline{\alpha})}{\partial \alpha_\nu} = \sum_{i\in D(\nu)} \frac{E_i}{1-\alpha_\nu^{E(i)}} - \sum_{i\in R(\nu)} E_i = 0.$$

Diese sind zu den Ereigniszeitpunkten t_ν, an denen mehrere Ereignisse beobachtet werden, nur auf iterativem Weg lösbar – ansonsten in geschlossener Form. D_ν ist die Indexmenge aller Ereignisse zum Zeitpunkt t_ν, R_ν ist die Risikomenge. Als Abkürzung gelten $E_i = \exp[\underline{b}.\underline{x}_i']$ und $E_{iv} = \exp[\underline{b}.\underline{x}_i(t_\nu)']$.

Bei zeitveränderlichen Kovariablen, die zu allen Ereigniszeitpunkten kleiner-gleich ihrer Ausfallzeit zu aktualisieren sind, lauten die Fehlergleichungen

$$\sum_{i\in D(\nu)} \frac{E_{ii}}{1-\alpha_\nu^{E(i,\nu)}} = \sum_{i\in R(\nu)} E_{iv}\ .$$

Zur Berechnung von $\hat{S}(t_\nu|\underline{z}_\ell) = \left(\prod_{t(j)\le t(\nu)} \alpha_j\right)^{E(\ell)}$ wird bei zeitveränderlichen Kovariablen die Potenz in jeden Faktor hineingezogen: $\hat{S}[t_\nu|\underline{z}_\ell(t)] = \prod_{t(j)\le t(\nu)} \alpha_j^{E(\ell,j)}$.

$\underline{z}_\ell(t)$ sind die Kovariablen eines Stichprobenmitglieds oder eines neuen Individuums, dessen Überlebenswahrscheinlichkeiten prognostiziert werden sollen. t_v ist einer der beobachteten Ereigniszeitpunkte $t_1 < ... < t_{n(E)}$. An diesen Stellen stuft sich die Treppenfunktion $\hat{S}[t|\underline{z}_\ell(t)]$, falls sie graphisch dargestellt werden soll.
Ebenso möglich wäre die Berechnung der Residuen als negative Logarithmen von $\hat{S}[t_i|\underline{x}_i(t)]$, wobei bei zensierten t_i bis zum letzten Ereigniszeitpunkt vor t_i aufmultipliziert wird. Da sie nur für einen graphischen Modelltest gebraucht werden, ist der numerische Unterschied zu $\hat{H}[t_i|\underline{x}_i(t)]$ vernachlässigbar.
Im geschichteten Cox-Modell verlaufen die Berechnungen prinzipiell gleichartig; nur die Schichtungs- oder Gruppierungsvariable selbst kann keine zeitveränderliche sein. Ob eine Schichtung notwendig ist, entscheidet ein PH-Test, ob sie unzulässig ist und gruppiert werden muss, ein Gruppierungstest. Während die Koeffizienten $\underline{b}$ in allen Schichten die gleichen sind (bei Gruppierung sind sie getrennt zu berechnen), werden die Residuen, die Survivorfunktionen und deren Varianzen getrennt geschätzt; die Behandlung zeitveränderlicher Kovariablen ist die selbe wie im Modell ohne Schichtung.

5.3. Die Varianzen der Survivorfunktion

Die asymptotischen Varianzen von $\hat{S}[t|\underline{z}(t)]$ werden an den Stellen der beobachteten Ereigniszeitpunkte t_1 bis $t_{n(E)}$ berechnet und für die Konstruktion von Konfidenzintervallen gebraucht. $\underline{z}$ sind die vorerst zeitkonstanten Ausprägungen eines Individuums aus der Stichprobe oder eines neuen. Die Herleitung richtet sich nach Kalbfleisch und Prentice (1980, Kap. 4.8.2) und Klein und Moeschber-ger (1997, Kap. 8.6).
Zur Herleitung benötigt man die asymptotischen Varianzen von $\hat{H}_0(t_v|\underline{b})$ und von $\underline{b}$, sowie Kenntnisse über Varianzen von Funktionen von Zufallsvariablen.

Es ergibt sich $\mathrm{Var}[\hat{S}(t_v|\underline{z})] = [\hat{S}(t_v|\underline{z})\cdot e^{\underline{b}.\underline{z}'}]^2\cdot[G(t_v)+\underline{a}.\mathrm{Var}(\underline{b}).\underline{a}']$

mit $G(t_v) = \sum_{j=1}^{v}\frac{d_j}{SJ^2}$, $a_k = \sum_{j=1}^{v}\frac{d_j\cdot\sum_{i\in R(j)}(x_{ik}-z_k)\cdot E_i}{SJ^2}$ und $SJ = \sum_{i\in R(j)}E_i$.

Liegen zeitveränderliche Ausprägungen $\underline{x}$ und/oder $\underline{z}$ vor, ist der erste Faktor als $[\hat{S}(t_v|\underline{z}(t))]^2$ zu berechnen und $e^{2.\underline{b}.\underline{z}'}$ in die Summierungen von G und a_k hineinzuziehen:

$$G(t_v) = \sum_{j=1}^{v}\frac{d_j\cdot e^{2\cdot\underline{b}.\underline{z}(tj)'}}{SJ^2},\quad a_k = \sum_{j=1}^{v}\frac{d_j\cdot e^{\underline{b}.\underline{z}(tj)'}\cdot\sum_{i\in R(j)}(x_{ikj}-z_{kj})\cdot E_{ij}}{SJ^2}\quad\text{und}\quad SJ = \sum_{i\in R(j)}E_{ij}.$$

Eine ausführliche Herleitung samt Rechenprogramm ist in meiner Dissertation angegeben. Die dort verwendeten Indizierungen sind komplizierter, aber direkt zur Programmierung geeignet; diese muss der Benutzer selbst durchführen, da sie von keinem Programmpaket angeboten wird.

5.4. Simulation

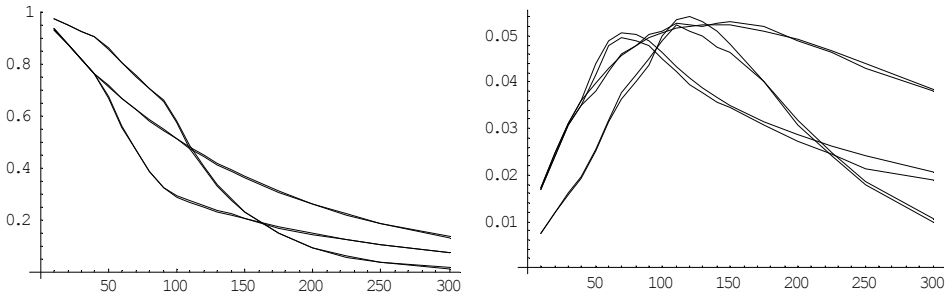
Um die Richtigkeit der Vorgangsweise wenigstens an einem Beispiel zu demonstrieren, wurde folgende Simulation durchgeführt.

Annahmen: ungeschichtetes Cox-Modell,
konstante Basishasardrate $h_o(t) = e^{-5}$, Strukturparameter $\beta_1 = 1$,
eine (p=1) Kovariable $\underline{x}$ mit gleichverteilten binären Ausprägungen,
Anzahlen $\underline{m}$ der Wertewechsel: gleichverteilt 0, 1, oder 2,
ca. L-förmig verteilte Wechselzeitpunkte $\underline{w}$,
drei Kovariablen-Konstellationen neuer Individuen:
$z_1 =$ 0 und zeitkonstant,
$z_{21} =$ 0 für $0 \le t < 47$, $z_{22} = 1$ für $47 \le t < 97$, $z_{23} = 0$ für $t > 97$,
$z_{31} = -1$ für $0 \le t < 47$, $z_{32} = 0$ für $47 \le t < 97$, $z_{33} = 1$ für $t > 97$.
Das untere linke Bild zeigt ihre theoretischen Survivorfunktionen; $S_{th}(t|\underline{z}_3)$ ist die anfangs zuoberst verlaufende.

Nach Simulation einer Stichprobe ($\underline{t}$, $\underline{x}$, $\underline{m}$, $\underline{w}$) des Umfangs n=200 wurde b_1 geschätzt und hierauf mit den Methoden aus den Kap. 5.2 und 5.3 $\hat{S}(t_v|\underline{z}_\ell)$ und die Wurzeln $se[\hat{S}(t_v|\underline{z}_\ell)]$ ihrer Varianzen. $\underline{t}_v$ sind die Ereigniszeitpunkte der Stichprobe, und $\underline{z}_\ell$ die drei Konstellationen. Die jeweils drei Vektoren $\hat{S}_\ell$ und se_ℓ wurden an 60 fixierten äquidistanten Stützstellen gespeichert.

Der obige Vorgang wurde 400 Mal wiederholt, was zu 3 mal 60 Stichprobenverteilungen von $\hat{S}_\ell$ und von se_ℓ mit den Umfängen 400 führt.

$\overline{S}_\ell$ seien die Mittelwerte der Verteilungen von $\hat{S}_\ell$; sie sind im folgenden linken Bild zusammen mit den theoretischen $S_{th}(t|\underline{z}_\ell)$ dargestellt und nahezu deckungsgleich.



$se(\hat{S}_\ell)$ seien die Standardabweichungen der Verteilungen von $\hat{S}_\ell$. Sie sind in der obigen rechten Abbildung gemeinsam mit $\overline{se}_\ell$ dargestellt, den Mittelwerten der Verteilungen von se_ℓ. Am rechten Bildrand, von oben nach unten betrachtet, enden die drei paarweise verschieden geschätzten Verläufe der Standardfehler von $\hat{S}_1$ bis $\hat{S}_3$. Die Übereinstimmungen sind zufriedenstellend.

Die numerischen Werte der Standardfehler sind hoch, wie es in Lebensdaueranalysen typisch ist; an der Stelle t=100 würden die Längen der drei Konfidenzintervalle bei einem Niveau α=5% ca. 0.2 betragen. Im parametrischen Exponentialmodell verringert sich die Intervalllänge auf ca. 0.14. In Stichproben des Umfangs 50 bzw. 800 verdoppeln bzw. halbieren sich diese Längen.

Literaturverzeichnis

Blossfeld H., Hamerle A. und Mayer K., *Ereignisanalyse*, Campus Verlag : 1986

Cox, D.R., *Regression Models and Life-Tables*,
Journal of the Royal Statistical Society B, V34 (1972), 187-220

Cox, D.R., *Partial Likelihood,* Biometrika, V62 (1975), 269-276

Cox, D.R. und Snell, E.J., *A General Discussion of Residuals*,
Journal of the Royal Statistical Society B, V30 (1968), 248-275

Kalbfleisch, J.D. und Prentice, R.L., *The Statistical Analysis of Failure Time Data*, Wiley : 1980

Klein, J.P. und Moeschberger M.L., *Survival Analysis*, Springer : 1997

Lawless, J.F., *Statistical Models and Methods for Lifetime Data*, Wiley :1982

Petersen, T., *Fitting parametric survival models with time-dependent covariates*, Applied Statistics, V35 (1986), 281-288

Meine Dissertation (Heinrich Potuschak, „*Lebensdaueranalysen mit zeitveränderlichen Kovariablen*“) wurde vom Universitätsverlag Rudolf Trauner, Linz 2000, in der Reihe B, Nr. 45 herausgegeben.