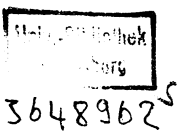
Alfred Hamerle, Gerhard Tutz

Diskrete Modelle zur Analyse von Verweildauer und Lebenszeiten

Campus Verlag
Frankfurt/ New York



CIP-Titelaufnahme der Deutschen Bibliothek

Hamerle, Alfred:
Diskrete Modelle zur Analyse von Verweildauer und Lebenszeiten / Alfred Hamerle ; Gerhard Tutz. –
Frankfurt/Main ; New York : Campus Verlag, 1989
(Campus : Forschung ; Bd. 568)
ISBN 3-593-33946-3
NE: Tutz, Gerhard:; Campus / Forschung

Umschlaggestaltung: Atelier Warminski, Büdingen
Druck und Bindung: KM-Druck, Groß-Umstadt
Printed in Germany

Unser besonderer Dank gilt Frau Beatrix Becker, die sehr professionell und mit großer Akribie die Übertragung des Manuskripts besorgt hat. Den Herrn Prof. Dr. Ludwig Fahrmeir und Dr. Thomas Meindl danken wir für die kritische Durchsicht von Teilen des Manuskripts. Den Herausgebern danken wir für die Aufnahme der Monographie in dieser Reihe. Schließlich ist es den Autoren eine angenehme Pflicht, dem Campus Verlag für eine angenehme Zusammenarbeit zu danken.

Konstanz und Regensburg, im September 1988

Alfred Hamerle
Gerhard Tutz

Vorwort

Ziel dieser Monographie ist die Darstellung von zeitdiskreten Modellen zur Analyse von Verweildauern und Lebenszeiten. Diese Verfahren stellen eine Ergänzung zu den in stetiger Zeit formulierten Hazardratenmodellen dar, die in der Literatur umfassend behandelt werden. Sie können dann eingesetzt werden, wenn die Lebenszeiten bzw. Verweildauern nicht exakt bestimmt werden können, sondern z.B. lediglich Zeitintervalle angebbar sind, in denen die in Frage stehenden Zustandswechsel bzw. Ereignisse eingetreten sind. In vielen Fällen ist nur eine derartige Datenbasis verfügbar. Deshalb sind die hier dargestellten Methoden für einen breiten Leserkreis von Bedeutung. Die Verfahren können nicht nur in der Medizin, sondern auch in den Wirtschafts- und Sozialwissenschaften, der Ökologie, der Psychologie oder in der Zuverlässigkeitstheorie eingesetzt werden. Die behandelten Beispiele entstammen dementsprechend auch unterschiedlichen Wissenschaftsdisziplinen.

Den größten Teil der Monographie nimmt die Darstellung von Modellen ein, bei denen die Zeitdauer von einem Anfangszustand bis zum Erreichen eines bestimmten (absorbierenden) Endzustands untersucht wird. Diese Situation ist typisch für Lebenszeit- bzw. Überlebenszeitstudien. Darüber hinaus werden auch Mehr–Zustands–Modelle (Competing–Risks–Modelle) und Mehr–Episoden–Modelle eingehend behandelt. Letztere liegen dann vor, wenn im Laufe der Zeit mehrfache Übergänge möglich sind oder wenn ein bestimmtes Ereignis wiederholt auftreten kann. Schließlich werden im letzten Kapitel einige Probleme erörtert, die bei der expliziten Modellierung unbeobachteter Populationsheterogenität, die nicht durch die Kovariablen erfaßt wird, auftreten.

Der Text eignet sich als Lehrbuch für die Vermittlung der Verfahren der zeitdiskreten Analyse von Verweildauern und Lebenszeiten für Studenten oder zum Selbststudium sowie als Handbuch und Nachschlagewerk für den Anwender in der Forschung. Neben der Behandlung der statistischen Grundlagen wurde besonderer Wert auf die konkrete Anwendung anhand empirischer Datensätze und den Einsatz geeigneter EDV–Programme gelegt.

INHALTSVERZEICHNIS

1. Grundlegende Begriffe der Analyse von Verweildauern und Lebenszeiten

1.1 Einführung

1.2 Einbeziehung von Kovariablen: Regressionsmodelle

1.3 Zensierte Daten

1.4 Statistische Grundkonzepte

1.4.1 Grundlegende Begriffe bei stetig gemessener Zeit

1.4.2 Grundlegende Begriffe bei diskret erhobenen Zeiten

2. Diskrete Verweildauern ohne explizite Berücksichtigung exogener Variablen: Die Sterbetafel

2.1 Methode der Sterbetafel

2.2 Zugrundeliegendes Modell und Varianzschätzer

2.2.1 Nichtzensierte Daten

2.2.2 Zensierte Daten

3. Modelle für den Ein-Episoden-Fall

3.1 Das gruppierte Cox-Modell

3.1.1 Grundmodell

3.1.2 Erweiterungen des Modells

3.2 Proportionalität der diskreten Hazardfunktionen

3.3 Logistische Modelle

3.4 Sequentielle Modelle auf der Basis latenter Variablen

3.5 Maximum-Likelihood-Schätzung

3.6 Anwendungsbeispiel

4. Die Einbeziehung von zeitabhängigen Kovariablen

4.1 Modelldarstellung

4.2 Beziehung zwischen Survivorfunktion und Hazardrate

4.3 Maximum-Likelihood-Schätzung

4.4 Möglichkeiten zur Konstruktion von speziellen zeitabhängigen Kovariablen

5. Exponentialmodelle mit konstantem Hazard in den Intervallen

5.1 Modelldarstellung

5.2 Maximum-Likelihood-Schätzung

5.3 Anwendungsbeispiel

6. Competing-Risks-Modelle

6.1 Parametrisierung der ursachenspezifischen Hazardrate

6.2 Maximum-Likelihood-Schätzung

7. Modelle für den Mehr-Episoden-Fall

7.1 Episodenspezifische Hazardraten

7.2 Maximum-Likelihood-Schätzung

7.3 Anwendungsbeispiel

8. Modelle mit Einbeziehung unbeobachteter Populationsheterogenität

Anhang

Literaturverzeichnis

1. Grundlegende Begriffe der Analyse von Verweildauern und Lebenszeiten

1.1 Einführung

Die statistische Analyse von Zeitverläufen bzw. Verlaufsdaten untersucht die Länge der Zeitintervalle zwischen aufeinanderfolgenden Zustandswechseln bzw. Ereignissen. Sie informiert für jede Untersuchungseinheit über die Zeitpunkte der Zustandswechsel bzw. des Eintreffens bestimmter Ereignisse und über die Abfolge dieser Ereignisse. Beispiele hierfür sind die Lebens- oder Überlebenszeiten in medizinischen Studien, die Dauer der Arbeitslosigkeit in möglicherweise mehreren aufeinanderfolgenden Perioden, die Lebensdauer von politischen oder gesellschaftlichen Organisationen, die Zeitdauer zwischen der Markteinführung eines Produkts und dem Kauf durch die Konsumenten, die aufeinanderfolgenden Perioden, in denen ein technisches Gerät nach jeweiliger Reparatur störungsfrei arbeitet, die Dauer von Lernprozessen, die Zeitdauer bis zum Umzug in eine andere Region bei Wanderungs- und Mobilitätsanalysen, die Zeitdauer bis zur Rückfälligkeit von Straftätern, etc.

Zusätzlich zu den Verweildauern bzw. Lebenszeiten werden für jede Untersuchungseinheit eine Reihe von weiteren Kovariablen erhoben, von denen einige ebenfalls zeitabhängig sein können, und die einzeln und/oder in Kombination die Verweildauern bzw. Lebenszeiten beeinflussen. Ein wichtiges Ziel der statistischen Analyse besteht in der quantitativen Ermittelung des Ausmaßes des Einflusses dieser exogenen oder endogenen Variablen.

Aufgrund der Entwicklung und Anwendung der Verfahren in verschiedenen Bereichen wie z.B. Medizin, Demographie, Sozialwissenschaften, Psychologie, Wirtschaftswissenschaften oder Technik ist die Terminologie sehr uneinheitlich. So wird — je nach Anwendungsbereich — die in einem Zustand verbrachte Zeit als Verweil- bzw. Aufenthaltsdauer, Lebens- bzw. Überlebenszeit, Ankunftszeit, Wartezeit oder Dauer der Episode bezeichnet. Zur Modellierung derartiger zeitabhängiger Prozesse — ohne Berücksichtigung von Kovariablen und mit stetig gemessener Zeit — wurden lange Zeit homogene Markov-Prozesse, Semi-Markov-Prozesse sowie Erneuerungsprozesse eingesetzt, oder die exakten Verweildauern wurden vernachlässigt und lediglich die

Übergänge mit Markov-Ketten, vorwiegend 1. Ordnung, untersucht. Diese Modelle sind jedoch sehr restriktiv, insbesondere erlauben sie nicht ohne weiteres die Einbeziehung von exogenen Variablen. Gelegentlich wurden für die vorliegende Problemstellung Logit- und Probitmodelle vorgeschlagen, wie z.B. von Egle (1979), der für Arbeitslose die Wahrscheinlichkeit untersuchte, in einem bestimmten Zeitintervall wieder Arbeit zu finden, in Abhängigkeit von personenbezogenen Kovariablen. Solche Analysen sind aber stets abhängig von dem gewählten Zeitintervall. Erst in jüngerer Zeit (z.B. Cox 1972) wurden in der Biostatistik für den Spezialfall von Überlebenszeiten (nur eine Zeitdauer; ein absorbierender Endzustand) Regressionsansätze vorgestellt, für die dann auch geeignete Methoden der Parameterschätzung entwickelt wurden (Kalbfleisch/Prentice 1973, Cox 1975). Mittlerweile existieren eine Reihe von Lehrbüchern und Monographien über "Survival-Analysis", z.B. Kalbfleisch/Prentice (1980), Elandt-Johnson/Johnson (1980), Lee (1980), Miller (1981), Lawless (1982), Cox/Oakes (1984), Schuhmacher (1983).

In den Sozialwissenschaften wurde die Analyse von Verweildauern und Zeitverläufen unter dem Stichwort "Event-History-Analysis" untersucht. Man vergleiche dazu beispielsweise Coleman (1981), Tuma (1982), Tuma/Hannan/Groeneveld (1979), Diekmann/Mitter (1984), Tuma/Hannan (1984), Andress (1985) und Blossfeld/Hamerle/Mayer (1986).

In der Ökonomie werden die Regressionsmodelle zur Analyse von Verweildauern vorwiegend zur Untersuchung der Dauer der Arbeitslosigkeit vorgeschlagen, vor allem von J. Heckman und Mitautoren (siehe z.B. Heckman 1978, Flinn/Heckman 1982, Heckman/Singer 1982, 1984 a, b, Heckman/Borjas 1980, aber auch Lancaster 1979).

Im einfachsten Fall wird die Zeitdauer von einem Anfangszustand bis zu dem Erreichen eines bestimmten (absorbierenden) Zielzustands untersucht. Man spricht dann von *Ein-Episoden-Modellen* mit einem Zielzustand. Existieren mehrere (absorbierende) Zielzustände, handelt es sich um *Mehr-Zustands-Modelle*, die in der Biostatistik meist als Competing-Risks-Modelle bezeichnet werden. *Mehr-Episoden-Modelle* liegen vor, wenn im Laufe der Zeit mehrfache Übergänge möglich sind oder wenn ein bestimmtes Ereignis (z.B. Arbeitslosigkeit oder ein bestimmter Defekt bei einem Gerät) wiederholt auftreten kann.

Im überwiegenden Teil der Literaturbeiträge wird davon ausgegangen, daß die Zeitpunkte, zu denen ein Zustandswechsel stattfinden kann, exakt angegeben werden können. In solchen Fällen handelt es sich um stochastische Prozesse mit stetiger Zeit und endlichem Zustandsraum. Die Zeit ist eine stetige Variable, (Ereignisse bzw. Zustandswechsel können zu jedem beliebigen Zeitpunkt erfolgen), die Zustandsvariable hingegen besitzt nur endlich viele Ausprägungen. Einen Überblick über Mehr–Episoden– und Mehr–Zustands–Modelle, verschiedene Anwendungsmöglichkeiten in Medizin, Marketing, Ökonometrie, Psychologie und Soziologie sowie weitere Literaturhinweise findet man bei Hamerle (1984).

In vielen Fällen ist jedoch die exakte Angabe der Zeitpunkte der Zustandswechsel nicht möglich. In diesen Fällen können lediglich Zeitintervalle angegeben werden, in denen Zustandswechsel aufgetreten oder bestimmte Ereignisse eingetreten sind. Bei anderen Anwendungen ist die Anzahl gleicher Beobachtungswerte (Ties) bei den gemessenen Verweildauern sehr hoch. Dies hat zur Folge, daß die für eine Reihe von Modellen (z.B. Cox-Modell) gewonnenen Parameterschätzungen nicht mehr brauchbar sind. Darüber hinaus ist in den Literaturbeiträgen, die sich mit der Ableitung der asymptotischen Eigenschaften der Schätzungen beschäftigen (z.B. Andersen/Gill (1982), Borgan (1984)), die Annahme enthalten, daß Ties nur mit der Wahrscheinlichkeit 0 auftreten. Deshalb ist es in all diesen Fällen zweckmäßig, diskrete Modelle zu verwenden. Solche Modelle sind Gegenstand des vorliegenden Beitrags.

1.2 Einbeziehung von Kovariablen: Regressionsmodelle

In jeder Episode wird für jedes Individuum bzw. Objekt ein Vektor von Kovariablen bzw. prognostischen Faktoren erhoben, von denen einige auch zeitabhängig sein können. Die Anzahl der Kovariablen kann von Episode zu Episode variieren. Es kann sich dabei um stetige oder um kategoriale Merkmale handeln. Bei kategorialen Merkmalen geht man in Analogie zur Varianzanalyse über zu einer Kodierung der einzelnen Kategorien durch Dummy-Variablen. Dazu gibt es mehrere Möglichkeiten.

Eine Möglichkeit besteht in der sogenannten *(0,1)-Kodierung*. Besitzt ein Merkmal A I Kategorien (Faktorstufen), so lassen sich diese durch $I-1$ Dummy-Variablen erfassen in der Form

$$x_i^A = \begin{cases} 1, & \text{falls Kategorie } i \text{ der Variablen A vorliegt;} \\ 0, & \text{sonst,} \end{cases} \qquad (1\text{–}1)$$

$i = 1, \ldots, I-1$.

Die i-te Dummy-Variable $x_i^A (i = 1, \ldots, I-1)$ kodiert dabei nur das Vorliegen bzw. Nicht-Vorliegen der i-ten Ausprägung. Das Vorliegen der I-ten (Referenz-)Kategorie ist implizit erfaßt durch die Kodierung $x_i^A = 0$ für $i = 1, \ldots, I-1$. Die Wahl der I-ten Kategorie als Referenzkategorie ist nicht zwingend, man kann eine beliebige Kategorie dafür auswählen.

Mit $x_1^A, x_2^A, \ldots, x_{I-1}^A$ lassen sich somit sämtliche Kategorien der Variablen A kodieren. Die zugehörigen Koeffizienten β_i werden (wie in der Varianzanalyse) *Haupteffekte* genannt.

Eine unmittelbar an die Varianzanalyse angelehnte Darstellung ergibt sich durch die Effekt–Kodierung. Die Merkmalsdarstellung erfolgt dann mit den $I-1$ Dummy-Variablen

$$x_i^A = \begin{cases} 1, & \text{falls Kategorie } i \text{ der Variablen A vorliegt;} \\ -1, & \text{falls Kategorie } I \text{ der Variablen A vorliegt;} \\ 0, & \text{sonst,} \end{cases} \qquad (1\text{–}2)$$

$i = 1, \ldots, I-1$.

Die Effektkodierung (1.2) ist eine unmittelbare Konsequenz der in der Varianzanalyse üblichen Restriktionen. Dort wird die Summe der Effekte einer

Variablen A a priori gleich 0 gesetzt. Der Effekt der Referenzkategorie I ist dann die negative Summe der ersten $I-1$ Effekte, und daraus folgt die Kodierung -1 bei Vorliegen der Referenzkategorie. Für weitere Details vergleiche man Hamerle/Kemény/Tutz (1984).

Als Einflußgrößen im Rahmen eines Regressionsansatzes für Verweildauern und Lebenszeiten, insbesondere mit kategorialen prognostischen Faktoren, kommen auch *Interaktionswirkungen* in Frage. Sie messen den gemeinsamen Einfluß einer bestimmten Kombination von Kategorien von zwei oder mehr unabhängigen Merkmalen. Formal werden sie durch Produkte von Dummy-Variablen in den Regressionsansatz einbezogen. Der *Datenvektor* x wird erweitert um die Zwei-Faktoren-Interaktionen, wie z.B. $x_i^A \cdot x_j^B$, bzw. Drei-Faktoren-Interaktionen, z.B. $x_i^A \cdot x_j^B \cdot x_k^C$.

Alle quantitativen Kovariablen einer Person sowie die Kodierungen für sämtliche Haupteffekte und der im Modell enthaltenen Interaktionswirkungen der kategorialen Kovariablen werden zum (geeignet dimensionierten) Daten- oder Designvektor x zusammengefaßt. Die Dimension von x kann von Episode zu Episode variieren.

In der Regel muß man davon ausgehen, daß für die k-te Episode zumindest ein Teil der Vorgeschichte des Prozesses von Bedeutung ist, z.B. die Dauer der vorangegangenen Episoden. Der relevante Teil der Vorgeschichte wird in den aktuellen Kovariablenvektor aufgenommen. Darüber hinaus können neue Einflußgrößen hinzukommen, die bei den vorangegangenen Episoden keine Rolle spielten oder nicht gemessen werden konnten. Man vergleiche dazu das Beispiel in Kapitel 6.

Von besonderer Bedeutung ist die Art des Einwirkens der Kovariablen auf die Verweildauern bzw. Lebenszeiten. Im allgemeinen wird — wie bei herkömmlichen Regressionsansätzen — davon ausgegangen, daß der Einfluß der Kovariablen oder prognostischen Faktoren linear in den Parametern erfolgt, also über eine Linearkombination

$$\gamma = x'\beta$$

mit einem unbekannten Parametervektor β. Es sind aber auch andere Ansätze möglich.

Wie bereits zu Beginn dieses Abschnittes erwähnt, können einige Kovariablen ebenfalls zeitabhängig sein. Dies ist etwa dann der Fall, wenn eine Therapie nur während eines bestimmten Zeitraumes angewendet wird und wenn man überprüfen möchte, ob die Therapie oder das Medikament auch nach der Anwendung eine Wirkung besitzt. Neben der Versuchs- sei auch eine Kontrollgruppe in die Studie aufgenommen. Man definiert dann zwei Dummy-Variablen, etwa $x_1(t)$ und $x_2(t)$ mit

$$x_1(t) = \begin{cases} 1, & \text{während der Behandlung eines Patienten;} \\ 0, & \text{sonst} \end{cases}$$

$$x_2(t) = \begin{cases} 1, & \text{nach Abschluß der Behandlung eines Patienten;} \\ 0, & \text{sonst.} \end{cases}$$

Sind die zugehörigen Regressionskoeffizienten negativ, so ist die Behandlung effektiv und verringert die Hazardrate, d.h. die Wahrscheinlichkeit, daß zum Zeitpunkt t der Zielzustand eintritt, wenn er bis zum Zeitpunkt t noch nicht eingetreten ist (die Regressionsansätze werden für die Hazardraten formuliert; man vergleiche die Ausführungen im nächsten Abschnitt). Ist darüber hinaus der erste Koeffizient absolut signifikant größer als der zweite, sinkt der Effekt nach dem Absetzen der Behandlung.

In den nächsten Abschnitten gehen wir zunächst davon aus, daß die Kovariablen zeitunabhängig sind. Die Einbeziehung von zeitabhängigen Kovariablen wird in Kapitel 4 gesondert behandelt.

1.3 Zensierte Daten

Ein zusätzliches Problem, das bei der Analyse von Verlaufsdaten auftritt, ist die Zensierung. Da das Ende des gesamten Beobachtungszeitraums in der Regel vorgegeben ist, ist die Verweildauer bzw. Lebenszeit eines Individuums unter Umständen nicht abgeschlossen. In einem solchen Fall spricht man von rechts zensierten Daten. Beispielsweise treten die Untersuchungsobjekte zu bestimmten Zeitpunkten in die Untersuchung ein, etwa am Tag der Diagnosestellung oder der Operation, und danach wird ihre Verweildauer oder Lebenszeit über einen Zeitraum hinweg bis zu einem Stichtag verfolgt. In einem solchen Fall kann es sein, daß die Verweildauer oder Lebenszeit am Stichtag noch andauert. Ferner kann ebenfalls keine exakte Lebenszeit oder Verweildauer ermittelt werden, wenn die Personen während der Studie aus anderen Gründen ausscheiden, z.B. wegen eines Umzugs oder Wechsels in eine andere Klinik und daher zur Weiterverfolgung nicht mehr zur Verfügung stehen. Für die verschiedenen Möglichkeiten der Entstehung von zensierten Daten vergleiche man z.B. Nelson (1972).

Bei der Maximum-Likelihood-Schätzung können rechts zensierte Beobachtungen berücksichtigt werden. Zu diesem Zweck ist der Zensierungsmechanismus, der den Daten zugrundeliegt, genau zu analysieren und in ein statistisches Modell zu fassen. Im folgenden werden drei Modelle kurz skizziert, die für Anwendungen von besonderem Interesse sind.

Zensierungsmodell I

In Modell I ist für jedes Individuum i, $i = 1, \ldots, n$, ein fester Beobachtungszeitraum c_i vorgegeben. Die Verweildauer des Individuums i sei wieder repräsentiert durch die Zufallsvariable T_i. Beobachtbar ist in diesem Modell lediglich $min(T_i, c_i)$ und ein Zensierungsindikator δ_i mit $\delta_i = 1$, wenn $T_i \leq c_i$ und $\delta_i = 0$, wenn $T_i > c_i$. Eine in der Anwendung häufig gewählte Variante, die sogenannte Typ I-Zensierung, setzt $c_i = c$ für alle i, und die Konstante c wird vorgegeben.

Zensierungsmodell II (Typ II-Zensierung)

In diesem Modell wird die Untersuchung beendet, wenn eine vorher festgelegte Anzahl von Zustandswechsel bzw. Ereignissen stattgefunden hat. Da-

mit wird das Ende c des Beobachtungszeitraumes eine Zufallsvariable. Dieses Zensierungsmodell eignet sich besonders für die Analyse von Lebenszeiten im technischen Bereich.

Zensierungsmodell III (random censoring)

Hier werden die Zensierungszeiten C_i als Zufallsvariablen vorausgesetzt, die von den Verweildauern T_i unabhängig sind. Beobachtbar ist dabei wieder $min(T_i, C_i)$ und der Zensierungsindikator δ_i mit $\delta_i = 1$ für $T_i \leq C_i$ und $\delta_i = 0$ für $T_i > C_i$. In den folgenden Kapiteln wird bei der Maximum-Likelihood-Schätzung stets von diesem Zensierungsmechanismus ausgegangen.

Für weitere Zensierungsmechanismen vergleiche man z.B. Kalbfleisch/Prentice (1980), Kap. 5, oder Lawless (1982), Kap. 1.4, 3 und 4.

Gelegentlich existiert auch die Möglichkeit der Zensierung von links, d.h. die Zeitspanne, die ein Individum bzw. Objekt bereits im in Frage stehenden Zustand verbracht hat, ist unbekannt. Dieser Fall ist schwieriger zu behandeln als Zensierung von rechts, da es im allgemeinen nicht möglich ist, die Auswirkungen der nicht bekannten Vorgeschichte auf zukünftige Ereignisse einzuschätzen. Im folgenden setzen wir stets voraus, daß entweder der Startzeitpunkt und der Startzustand fest vorgegeben sind (ohne Beschränkung der Allgemeinheit dann $t_0 = 0$) oder daß die Vorgeschichte des Prozesses von dem Beobachtungszeitraum den weiteren Verlauf des Prozesses nicht beeinflußt. Zur Einbeziehung linkszensierter Daten bei der Schätzung von Hazardraten-Modellen in stetig gemessener Zeit vergleiche man Hamerle (1988).

1.4 Statistische Grundkonzepte

1.4.1 Grundlegende Begriffe bei stetig gemessener Zeit

In diesem Abschnitt wird ausschließlich der Ein-Episoden-Fall mit einem absorbierenden Zielzustand behandelt, also die Zeitdauer zwischen einem Anfangszustand und dem Erreichen eines bestimmten Endzustandes. Viele der hierfür entwickelten statistischen Konzepte können auf komplexere Situationen wie mehrere aufeinanderfolgende Episoden oder mehrere Endzustände (competing risks) übertragen werden.

Die Dauer der Episode, d.h. die Verweildauer oder Lebenszeit, wird repräsentiert durch eine nicht negative Zufallsvariable T. Dichte und Verteilungsfunktion von T seien $f(t)$ bzw. $F(t)$. Eine wichtige Rolle spielt bei Lebenszeit-Modellen die zu $F(t)$ komplementäre Wahrscheinlichkeit, nämlich die Wahrscheinlichkeit, den Zeitpunkt t zu "erleben" bzw. zu "überleben". Die Funktion

$$S(t) = P(T \geq t) \tag{1-3}$$

heißt *Survivorfunktion*. Für stetiges T gilt

$$S(t) = 1 - F(t) \quad . \tag{1-4}$$

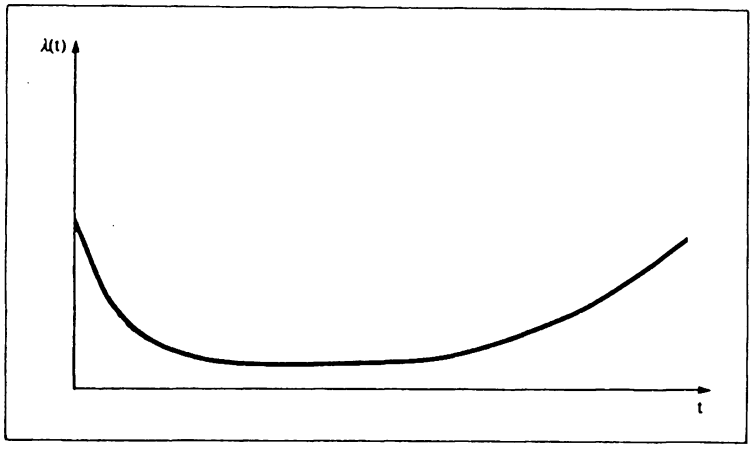
Eine weitere wichtige Funktion zur Beschreibung der Verteilung von T ist die *Hazardrate* (*Intensitäts-* oder *Risikofunktion*). Sie ist bestimmt durch

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t) \tag{1-5}$$

Die Hazardrate kann aufgefaßt werden als der Grenzwert der bedingten Wahrscheinlichkeit, daß die Episode im Intervall $[t, t + \Delta t)$ zu Ende geht unter der Voraussetzung, daß das Individuum den Beginn dieses Intervalls erlebt.

Die Hazardrate stellt ein zentrales Konzept bei der Analyse von Verlaufsdaten dar. Überlebt ein Individuum den Zeitpunkt t, so informiert die Hazardrate über "den weiteren Verlauf". Häufig besitzt man bei praktischen

Anwendungen zumindest qualitative Vorinformationen über die Hazardrate. Betrachtet man beispielsweise das Sterberisiko einer Population, so hat die Hazardrate typischerweise einen "badewannenförmigen" Verlauf.



"Badewannenförmige" Hazardrate des Sterberisikos einer Population

Zu Beginn des Prozesses ist das Sterberisiko aufgrund der Kindersterblichkeit kurz nach der Geburt relativ hoch, fällt dann ab und bleibt über einen bestimmten Zeitraum konstant auf niedrigem Niveau, bis es mit zunehmenden Alter wieder anwächst.

Der Zusammenhang zwischen Hazardrate, Survivorfunktion und Dichtefunktion ist (vgl. z.B. Kalbfleisch/Prentice (1980), S. 6)

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{1-6}$$

$$S(t) = exp(-\int_0^t \lambda(u)\,du) \tag{1-7}$$

$$f(t) = \lambda(t) \cdot exp(-\int_0^t \lambda(u)\,du) \tag{1-8}$$

Sowohl Dichte- bzw. Verteilungsfunktion als auch die Hazardrate und Survivorfunktion beschreiben die Verteilung der Verweildauer bzw. Lebenszeit eindeutig. Kennt man eine der Größen, so lassen sich im Prinzip die anderen daraus ermitteln.

Werden Kovariablen in die Analyse einbezogen — etwa durch einen p-dimensionalen Vektor x —, so werden $f(t \mid x)$, $S(t \mid x)$ und $\lambda(t \mid x)$ jeweils bei gegebenem Kovariablenvektor x definiert, also z.B.

$$\lambda(t \mid x) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t, x) \qquad (1\text{–}9)$$

(1.6) bis (1.8) können ohne Schwierigkeiten auf Modelle mit Kovariablen übertragen werden.

In der Survival-Analyse hat es sich als zweckmäßig erwiesen, die Hazardrate in Abhängigkeit von den Kovariablen zu modellieren. Ein besonders einfaches Modell ist das *Exponential-Regressionsmodell* mit

$$\lambda(t \mid x) = exp(x'\beta) \qquad (1\text{–}10)$$

Die Hazardrate in (1.10) ist zeitunabhängig. Individuen mit verschiedenen Kovariablen besitzen verschiedene Hazardraten, die jedoch jeweils über die Zeit hinweg konstant sind. Eine Erweiterung auf zeitabhängige Hazardraten liefert das *Weibull-Regressionsmodell*

$$\lambda(t \mid x) = \alpha\lambda_0(\lambda_0 t)^{\alpha-1} exp(x'\beta) \qquad (1\text{–}11)$$

Das Weibull-Regressionsmodell gehört zur Klasse der *Proportional–Hazards–Modelle*. Der Quotient der Hazardraten von zwei Individuen mit verschiedenen Kovariablenvektoren ist unabhängig von der Zeit. Eine naheliegende Verallgemeinerung besteht darin, von einer Hazardrate der Form

$$\lambda(t \mid x) = \lambda_0(t) g(x; \beta) \qquad (1\text{–}12)$$

auszugehen, wobei $\lambda_0(t)$ eine nicht spezifizierte "Baseline"-Hazardrate ist. Die Spezifikation

$$g(x; \beta) = exp(x'\beta) \qquad (1\text{–}13)$$

ist das *Cox-Modell*, das von Cox (1972) eingeführt wurde und das mittlerweile breite Anwendung gefunden hat.

1.4.2 Grundlegende Begriffe bei diskret erhobenen Zeiten

Die Zeitachse wird zerlegt in $q+1$ Intervalle

$$[a_0, a_1), [a_1, a_2), \ldots, [a_{q-1}, a_q), [a_q, \infty) \quad ,$$

wobei in der Regel $a_0 = 0$ gesetzt und für a_q das Ende des Beobachtungszeitraums genommen wird. Für das Zeitintervall $[a_{t-1}, a_t)$ schreiben wir auch kurz t.

Die Verweildauer bzw. Lebenszeit wird repräsentiert durch eine positive Zufallsvariable T. T nimmt nur ganzzahlige Werte an, und $T = t$ bedeutet, daß im Intervall $[a_{t-1}, a_t)$ ein Übergang bzw. Zustandswechsel stattgefunden hat.

Neben der Verweildauer bzw. Lebenszeit wird für jedes Individuum bzw. Objekt ein p-dimensionaler Vektor x von Kovariablen bzw. prognostischen Faktoren erhoben. Die Kovariablen werden hier als *zeitunabhängig vorausgesetzt.* Die Einbeziehung von zeitabhängigen Kovariablen wird in Abschnitt 4 erörtert.

In Analogie zu (1.2) und (1.4) können im diskreten Fall Hazardrate und Survivorfunktion definiert werden. Die Hazardrate ist gegeben durch

$$\lambda(t \mid x) = P(T = t \mid T \geq t, x) \quad \text{für } t = 1, \ldots, q \tag{1–14}$$

(1.14) ist die bedingte Wahrscheinlichkeit dafür, daß ein Individuum im Zeitintervall t den Endzustand erreicht, gegeben die Kovariablen und gegeben, daß das Individuum den Beginn des Zeitintervalls erreicht hat.

Die bedingte Wahrscheinlichkeit, das Zeitintervall t zu "überleben", ist dann

$$P(T > t \mid T \geq t, x) = 1 - \lambda(t \mid x). \tag{1–15}$$

Eine Möglichkeit für einen Regressionsansatz besteht darin, die Hazardrate (1.14) in Abhängigkeit von den Kovariablen zu modellieren, etwa in der Form

$$\lambda(t \mid x) = g(\beta_{0t} + x'\beta) \tag{1–16}$$

mit $g(\cdot) \in (0,1)$. β_{0t} bringt dabei den Beitrag einer "Baseline"-Hazardrate für das Intervall t ohne Berücksichtigung der Kovariablen zum Ausdruck.

In der Literatur wurden bereits eine Reihe möglicher Spezifikationen für g in (1.16) vorgeschlagen. Die wichtigsten werden in Kapitel 3 ausführlich behandelt. An dieser Stelle wird lediglich der bei Anwendungen besonders häufig gewählte logistische Ansatz kurz dargestellt. Die Hazardrate ist dann

$$\lambda(t \mid x) = \frac{exp(\beta_{0t} + x'\beta)}{1 + exp(\beta_{0t} + x'\beta)} \qquad t = 1, \ldots, q \quad , \tag{1-17}$$

und die bedingte Wahrscheinlichkeit, daß im Falle des Erreichens von Intervall t in diesem Intervall kein Übergang bzw. Zustandswechsel stattfindet, ist

$$1 - \lambda(t \mid x) = \frac{1}{1 + exp(\beta_{0t} + x'\beta)} \quad . \tag{1-18}$$

Die Survivorfunktion ist

$$S(t \mid x) = P(T \geq t \mid x) \quad , \tag{1-19}$$

die (unbedingte) Wahrscheinlichkeit, das Zeitintervall t zu "erleben". Den Zusammenhang zwischen Survivorfunktion und Hazardrate erhält man durch sukzessive Anwendung von

$$P(T \geq k \mid x) = P(T \geq k \mid T \geq k-1, x) \cdot P(T \geq k-1 \mid x)$$

und mit (1.15) durch

$$S(t \mid x) = \prod_{k=1}^{t-1} (1 - \lambda(k \mid x)) \quad . \tag{1-20}$$

Schließlich erhält man für die (unbedingte) Sterbe- oder Ausfallwahrscheinlichkeit oder allgemein für die Wahrscheinlichkeit, den Endzustand im Zeitintervall t zu erreichen, gegeben die Kovariablen,

$$\begin{aligned} P(T = t \mid x) &= P(T = t \mid T \geq t, x) \cdot P(T \geq t \mid x) \\ &= \lambda(t \mid x) \cdot \prod_{k=1}^{t-1} (1 - \lambda(k \mid x)) \end{aligned} \tag{1-21}$$

Weiter ergibt sich für die diskrete Hazardfunktion

$$\lambda(t \mid x) = \frac{P(T = t \mid x)}{P(T \geq t \mid x)} = \frac{S(t \mid x) - S(t+1 \mid x)}{S(t \mid x)} \quad .$$

2. Diskrete Verweildauern ohne explizite Berücksichtigung exogener Variablen: Die Sterbetafel

Eine der einfachsten und gebräuchlichsten Methoden zur Analyse von Verweildauern stellt die Methode der Sterbetafel dar. Wie der Begriff schon nahelegt, wurde das Verfahren vorwiegend von Demographen und Versicherungsstatistikern in der Form der Populationssterbetafel angewandt.

Neben der Populationssterbetafel sind vor allem die Kohorten-Sterbetafel und die klinische Sterbetafel gebräuchlich. Die Kohorten-Sterbetafel betrachtet die Überlebenszeit einer Kohorte, d.h. einer Gruppe von Individuen, die in einem bestimmten Zeitraum geboren wurden (vgl. Chiang 1968). Die im weiteren betrachtete klinische Sterbetafel geht im Gegensatz zu den beiden anderen Methoden nicht von bevölkerungsstatistischen Zahlen aus, sondern von Daten, wie sie im Rahmen kontrollierter Studien auftreten. Dabei muß es sich nicht, wie in klinischen Studien meist, tatsächlich um Überlebenszeiten, d.h. das Endereignis "Tod" handeln. Ebenso kann sich die Verweildauer auf die Länge eines Krankenhausaufenthaltes oder die Zeit der Arbeitslosigkeit einer Risikogruppe beziehen. Trotzdem werden im weiteren aus Gründen der Konvention meist die Begriffe Überlebenszeit und Sterbetafel (anstatt Verweildauer) gebraucht.

Charakteristisch für das Verfahren der Sterbetafeln ist, daß der Einfluß exogener Merkmale in der Sterbetafel nicht explizit modelliert wird. Im Vordergrund steht vielmehr die möglichst präzise Bestimmung der Überlebenszeiten einer definierten Population. Der Einfluß exogener Merkmale ergibt sich erst indirekt durch den Vergleich der Überlebenszeiten verschiedener Populationen oder Gruppen, die durch das Vorhandensein bzw. Fehlen bestimmter Merkmale charakterisiert sind. Das Verfahren ist im Grunde nonparametrisch, auch wenn manche Aussagen über die Eigenschaften von Schätzverfahren parametrisierte Familien von Verteilungen zugrundelegen.

2.1 Methode der Sterbetafel

Die Zeitachse sei wiederum zerlegt in $q+1$ Intervalle $I_k = [a_{k-1}, a_k)$, $k = 1, \ldots, q+1$, wobei $a_o = 0$ und $a_{q+1} = \infty$. Die Einteilung sei so gewählt, daß die Untergrenze a_q des letzten Intervalls $[a_q, \infty)$ den letztmöglichen Beobachtungszeitpunkt markiert.

Die Hazardrate des k–ten Intervalls

$$\lambda_k = P(T \in [a_{k-1}, a_k) \mid T \geq a_{k-1}) \tag{2-1}$$

bezeichnet die bedingte Wahrscheinlichkeit, das k–te Intervall nicht zu überdauern, gegeben das Zeitintervall wurde erreicht.

Bezeichne

$$p_k = P(T \geq a_k \mid T \geq a_{k-1})$$

die Wahrscheinlichkeit, das k–te Intervall zu überdauern, gegeben es wird erreicht, und

$$P_k = P(T \geq a_k)$$

die absolute Wahrscheinlichkeit, das k–te Intervall zu überdauern.

Man erhält unmittelbar $p_k = 1 - \lambda_k$. Durch sukzessive Anwendung von

$$P(T \geq a_i) = P(T \geq a_i \mid T \geq a_{i-1}) P(T \geq a_{i-1}) \tag{2-2}$$

erhält man mit $P(T \geq a_o) = 1$ unmittelbar

$$P_k = P(T \geq a_k \mid T \geq a_{k-1}) \ldots P(T \geq a_1 \mid T \geq a_o) P(T \geq a_o) = p_k \cdot \ldots \cdot p_1 \tag{2-3}$$

Gleichung (2.3) ist zentral für die Methode der Sterbetafeln. Schätzungen für P_k erhält man aus Schätzungen $\hat{p}_k$, indem Gleichung (2.2) in der Form

$$\hat{P}_k = \hat{p}_k \cdot \ldots \cdot \hat{p}_1 \tag{2-4}$$

auf die Schätzungen angewandt wird.

Die erhobenen Daten sind:

n Gesamtzahl der Beobachtungen zu Beginn der Studie

d_k Anzahl der Fälle, für die das Ereignis "Tod" im k–ten Intervall ($k = 1, \ldots, q$) auftritt,

w_k Anzahl der Zensierungen im k–ten Intervall, d.h. diejenigen Fälle, die zwar das k–te Intervall erreichen, von denen aber weder der Eintritt des Ereignisses "Tod" in diesem Intervall noch das Erreichen des nächsten Intervalls bekannt ist.

Die Anzahl n_k der Fälle, die im k–ten Intervall zur Risikomenge gehören, ergibt sich durch

$$n_1 = n$$

und

$$n_k = n_{k-1} - d_{k-1} - w_{k-1} \qquad \text{für } k = 2, \ldots, q.$$

Liegen im k–ten Intervall keine Zensierungen vor, läßt sich die Hazardrate λ_k des k–ten Intervalls unmittelbar durch die relative Häufigkeit d_k/n_k schätzen. Gilt jedoch $w_k > 0$, wird diese Schätzung die Hazardrate eher unterschätzen. Das übliche Schätzverfahren nach der Sterbetafel–Methode nimmt mit

$$\hat{\lambda}_k = \frac{d_k}{n_k - w_k/2} \tag{2–5}$$

eine Korrektur vor, die die Risikomenge des k–ten Intervalls "verkleinert". Als tatsächlicher Umfang der Risikomenge im k–ten Intervall wird $n_k - w_k/2$ betrachtet. Eine Rechtfertigung für diese willkürliche, wenn auch vernünftig scheinende Korrektur läßt sich nur durch (willkürliche) Annahmen über den zugrundeliegenden Zensierungsprozeß geben.

Mit $\hat{p}_k = 1 - \hat{\lambda}_k$ erhält man aus (2.5) Schätzungen $\hat{P}_k = \hat{p}_k \ldots \hat{p}_1$ der Überlebenswahrscheinlichkeit zum Zeitpunkt a_k. $\hat{P}_k$ als Schätzung der Survivor–Funktion $S(a_k)$ an der Stelle a_k wird als *kumulative Überlebensrate* bezeichnet.

Aus dieser grundlegenden Schätzung lassen sich einige weitere ableiten. Die Schätzungen der Verweildauer zum Zeitpunkt der Intervallmitten $m_k = (a_k -$

$a_{k-1})/2, \quad k = 1, \ldots, q$, erhält man aus

$$\hat{P}(T \geq m_k) = (\hat{P}_k + \hat{P}_{k-1})/2 = \hat{P}_{k-1}(1 + \hat{p}_k)/2. \tag{2-6}$$

Die geschätzte Ereigniswahrscheinlichkeit im k–ten Intervall ergibt sich unmittelbar durch

$$\hat{P}(T \in [a_{k-1}, a_k)) = \hat{P}_{k-1} - \hat{P}_k, \tag{2-7}$$

und für die Sterbewahrscheinlichkeit im k–ten Intervall, bezogen auf eine Zeiteinheit, erhält man die Dichte

$$\hat{f}_k = \frac{\hat{P}_{k-1} - \hat{P}_k}{h_k} = \frac{\hat{P}_{k-1}\hat{\lambda}_k}{h_k} \quad , \tag{2-8}$$

wobei $h_k = a_k - a_{k-1}$ die Länge des k–ten Intervalls bezeichnet.

Gleichung (2.8) läßt sich auf die Schätzung einer zugrundeliegenden stetigen Verweildauer beziehen. Während λ_k die Hazardrate des k–ten Intervalls darstellt, läßt sich in der stetigen Betrachtungsweise eine "mittlere Hazardfunktion" im k–ten Intervall schätzen durch

$$\lambda(m_k) = \frac{\hat{f}_k}{P(T \geq m_k)} = \frac{2\hat{\lambda}_k}{h_k(1 + \hat{p}_k)}$$

Datenaufbereitung und Schätzung werden veranschaulicht anhand einer Studie zum malignen Melanom, die an der M. B. Anderson Tumor Clinic durchgeführt wurde (MacDonald, 1963). Die Darstellung lehnt sich an Clark & Gross (1975) an.

TABELLE 2.1 STERBETAFEL ZUM MALIGNEN MELANOM IN EINER STUDIE DER M.D. ANDERSON TUMOR CLINIC

(nach Cross & Clarke, 1975)

k	$[a_{k-1}, a_k)$ in Jahren	n_k	w_k	$n_k - \frac{w_k}{2}$	d_k	$\hat{p}_k = 1 - \hat{\lambda}_k$	$\hat{P}_k$	$\hat{f}_n$	$\lambda(m_k)$
1	[0,1]	913	96	865.0	312	.639	.639	.361	.441
2	[1,2]	505	74	468.0	96	.795	.508	.131	.228
3	[2,3]	335	62	304.0	45	.852	.433	.075	.160
4	[3,4]	228	30	213.0	29	.864	.374	.059	.146
5	[4,5]	169	40	149.0	7	.953	.356	.018	.048
6	[5,6]	122	37	103.5	9	.913	.325	.031	.091
7	[6,7]	76	17	67.5	3	.956	.311	.014	.045
8	[7,8]	56	12	50.0	1	.980	.305	.006	.020
9	[8,9]	43	8	39.0	3	.923	.281	.024	.080
10	[9,∞]	32	-	32.0	32	.000	-		

2.2 Zugrundeliegendes Modell und Varianzschätzer

2.2.1 Nichtzensierte Daten

Für nichtzensierte Daten entspricht $d = (d_1, \ldots, d_{q+1})$, mit $d_{q+1} = n - d_1 - \ldots - d_q$, einer Multinomialverteilung mit n Beobachtungen und dem Wahrscheinlichkeitsvektor $\pi = (\pi_1, \ldots, \pi_{q+1})$, $\pi_{q+1} = 1 - \pi_2 - \ldots - \pi_q$, d.h. $d \sim M(n, \pi)$, wobei $\pi_k = P_{k-1} - P_k$, $k = 1, \ldots, q$ mit $P_o = 1$. Als Maximum–Likelihood–Schätzung erhält man standardmäßig $\hat{\pi}_k = d_k/n$ und wegen

$$\lambda_k = P(T \in [a_{k-1}, a_k) \mid T \geq a_{k-1}) = \pi_k/(\pi_k + \ldots + \pi_{q+1})$$

erhält man die ML–Schätzung

$$\hat{\lambda}_k = \frac{\hat{\pi}_k}{\hat{\pi}_k + \ldots + \hat{\pi}_{q+1}} = \frac{d_k}{d_k + \ldots + d_{q+1}} = \frac{d_k}{n - (d_1 + \ldots + d_{k-1})} \quad .$$

Bezeichnet wiederum $n_k = n_{k-1} - d_{k-1}$ den Umfang der Risikomenge, erhält man wegen $n_1 = n$ unmittelbar

$$\hat{\lambda}_1 = \frac{d_1}{n}, \quad \hat{\lambda}_2 = \frac{d_2}{n_2}, \ldots, \quad \hat{\lambda}_{q+1} = \frac{d_{q+1}}{n_{q+1}}$$

und damit die Standardschätzung (2.5) der Sterbetafel. Als Momente der entsprechenden Schätzungen $\hat{P}_k = (1 - \hat{\lambda}_k) \ldots (1 - \hat{\lambda}_1)$ erhält man

$$E(\hat{P}_k) = P_k$$
$$var(\hat{P}_k) = P_k(1 - P_k)/n$$

und für $k < r$

$$cov(\hat{P}_k, \hat{P}_r) = P_r(1 - P_k)/n.$$

Entsprechend erhält man

$$E(\hat{\lambda}_k) = \lambda_k \quad ,$$
$$var(\hat{\lambda}_k) = p_k(1 - p_k)E(\frac{1}{n_k})$$

und für $k < r$

$$cov(\hat{\lambda}_k, \hat{\lambda}_r) = 0 \quad ,$$

obwohl $\hat{\lambda}_k, \hat{\lambda}_r$ i.a. nicht unabhängig sind (vgl. Lawless, 1982).

Die Erwartungstreue der Schätzungen und die Möglichkeit, Konfidenzintervalle anzugeben, ergeben sich unmittelbar aus den Momenten, wobei die p_k's durch die Schätzungen $\hat{p}_k$ zu ersetzen sind.

2.2.2 Zensierte Daten

Beim Auftreten zensierter Daten ist neben der stetigen Verweildauer T die Dauer C, die eine Beobachtung bis zu ihrer Zensur in der Studie verbleibt, von Interesse. Zu jeder Beobachtung i gehört das Paar (T_i, C_i) der Zufallsvariablen Verweildauer und Dauer bis zur Zensur. Insbesondere wird angenommen, daß die Tupel $(T_i, C_i), \quad i = 1, \ldots, n$ unabhängige Wiederholungen sind und die Zensurzeit C_i unabhängig von der Verweildauer T_i ist. Beobachtet wird $t_i = \min(T_i, C_i)$ und der Zensierungsindikator

$$\delta_i = \begin{cases} 1, & \text{falls } t_i \text{ nicht zensiert, d.h. } T_i \leq C_i, \\ 0 & \text{sonst.} \end{cases}$$

Bezeichne

$$\pi_k^d = P(T \in [a_{k-1}, a_k), \quad T \leq C)$$

die Wahrscheinlichkeit, daß die Verweildauer im k–ten Intervall endet und auch beobachtet wird,

$$\pi_k^w = P(C \in [a_{k-1}, a_k), \quad C < T)$$

bezeichne die Wahrscheinlichkeit einer Zensur im k–ten Intervall.

Man erhält für den Beobachtungsvektor

$$d = (d_1, w_1, \ldots, d_q, w_q, r)$$

mit $r = n - \sum_{i=1}^{q}(d_i + w_i)$ eine Multinomialverteilung mit Wahrscheinlichkeitsvektor

$$\pi = (\pi_1^d, \pi_1^w, \ldots, \pi_q^d, \pi_q^w, \pi_r),$$

wobei $\pi_r = 1 - \sum_{i=1}^{q}(\pi_i^d + \pi_i^w)$.

Der Standardschätzer $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{q+1})$ der Sterbetafelmethode beruht für zensierte Daten nicht auf der üblichen ML–Schätzung der Multinomialverteilung. Da

$$\hat{\lambda}_k = \frac{d_k/n}{n_k/n - w_k/2n}$$

eine stetige, partiell differenzierbare Funktion der Größen d_k, w_k ist, erhält man für $\hat{\lambda}_k$ asymptotisch wie für die ML–Schätzung eine Normalverteilung. Allerdings konvergiert $\hat{\lambda}_k$ gegen

$$\lambda_k^* = \frac{\pi_k^d}{\pi_k^n - \pi_k^w/2}$$

mit $\pi_k^n = E(n_k/n)$ und ist damit nicht konsistent, da i.a.

$$\lambda_k = P(T \in [a_{k-1}, a_k) \mid T \geq a_{k-1}) \quad \neq \quad \lambda_k^*$$

gilt. Die Stärke der Verzerrung läßt sich exakt bestimmen, wenn spezifische Modelle für Verweildauern und Zensurmechanismus zugrundegelegt werden. Crowley (1970) zeigt für exponentialverteilte Verweildauern mit gleichverteiltem Zensurmechanismus, daß die asymptotische Verzerrung von $\hat{\lambda}_k$ und $\hat{P}_k$ für größeres q relativ klein wird.

Als Schätzungen für die Varianzen verwendet man im Fall zensierter Daten

$$\hat{v}ar(\hat{\lambda}_k) = \frac{\hat{\lambda}_k - \hat{\lambda}_k^2}{n_k - w_k/2}$$

und die von Greenwood (1926) vorgeschlagene Formel

$$\hat{v}ar(\hat{P}_k) = P_k^2 \sum_{i=1}^{k} \frac{\hat{\lambda}_i}{(n_i - w_i/2)\hat{p}_i} \quad ,$$

die sich als Approximation der asymptotischen Varianz ableiten läßt (Details siehe Lawless, 1982, S.64ff). Beide Schätzer tendieren dazu, die Varianz zu überschätzen.

Die Sterbetafelmethode ist insbesondere dann akzeptabel, wenn der Stichprobenumfang nicht zu klein und die Intervalle nicht zu groß sind. Weiter

sollte der Zensierungsmechanismus einigermaßen gleichmäßig über die Intervalle verteilt sein und nicht zu viele zensierte Daten auftreten. Die Methode ermöglicht dann eine nonparametrische Schätzung der Survivorfunktion, die als Grundlage zur Wahl stärkerer Modelle, wie z.B. einer exponentialverteilten Verweildauer dienen kann.

3. Modelle für den Ein–Episoden–Fall

Die im folgenden dargestellten Modelle behandeln den Fall eines absorbierenden Endzustandes. Die Kovariablen werden dabei als zeitunabhängig vorausgesetzt, so daß die in Abschnitt 1. 4. 2 dargestellten Zusammenhänge zwischen diskreter Hazardrate, Survivorfunktion und Ausfallwahrscheinlichkeit gelten.

Das erste betrachtete Modell ergibt sich unmittelbar aus dem Cox–Modell für stetige Verweildauern und wird daher auch als gruppiertes Cox–Modell bezeichnet. Die auch anzutreffende Bezeichnung als gruppiertes Proportional–Hazards–Modell ist insofern irreführend als die diskrete Hazardrate nicht mehr proportional ist. Erweiterungen des Modells durch Aranda–Ordaz (1983) modellieren zusätzlich additive Effekte.

In Abschnitt 3. 2 wird ein Modell betrachtet, das von der Proportionalität der diskreten Hazardrate ausgeht. Die Modellklasse des darauffolgenden Abschnitts geht vom logistischen Ansatz aus, der im Kontext der Verweildauermodelle zu logistischen Modellen in den Hazardraten führt. Einen wesentlich allgemeineren Ansatz stellen die sequentiellen Modelle in Abschnitt 3.4 dar. Die meisten anderen betrachteten Modelle lassen sich als Spezialfälle davon verstehen. Die separate Betrachtung der Spezialfälle ist dadurch begründet, daß sie meist anders abgeleitet werden, so z.B. das gruppierte Cox–Modell aus der stetigen Version.

3.1 Das gruppierte Cox-Modell

3.1.1 Grundmodell

Im Cox–Modell für *stetige Zeit* t wird die Hazardrate modelliert durch

$$\lambda(t \mid x) = \lambda_0(t) exp(x'\beta). \tag{3–1}$$

$\lambda_0(t)$ ist hier die "Baseline"–Hazardrate, die unabhängig vom Kovariablenvektor x ist und deren Form nicht weiter eingeschränkt ist, wie es z.B. beim Weibull–Modell der Fall ist, für das $\lambda_0(t) = \alpha\lambda_0(\lambda_0 t)^{\alpha-1}$ gilt.

Das Cox–Modell wird auch als allgemeines *Modell mit proportionalem Hazard* bezeichnet, da das Hazard–Verhältnis für zwei Kovariablenvektoren x_1, x_2 nicht von der Zeit abhängt. Man erhält

$$\frac{\lambda(t \mid x_1)}{\lambda(t \mid x_2)} = exp((x_1 - x_2)'\beta) \tag{3–2}$$

Das Verhältnis des Hazards zweier durch verschiedene Kovariablen gekennzeichneten Personengruppen bleibt bei Gültigkeit des Modells über die gesamte Zeit hinweg konstant.

Die analoge Konstanzeigenschaft erhält man wegen

$$S(t \mid x) = exp(-\int_0^t \lambda_0(u)\, exp(x'\beta)\, du) = exp(-exp(x'\beta) \int_0^t \lambda_0(u)\, du)$$

auch für die logarithmierte Survivorfunktion

$$\frac{lnS(t \mid x_1)}{lnS(t \mid x_2)} = exp((x_1 - x_2)'\beta). \tag{3–3}$$

Aus der Survivorfunktion ergibt sich mit

$$S_0(t) = exp(-\int_0^t \lambda_0(u)\, du)$$

unmittelbar eine alternative Formulierung des Cox–Modells mit

$$S(t \mid x) = S_0(t)^{exp(x'\beta)} \quad .$$

Sei nun die Zeitachse zerlegt in die folgenden vorgegebenen Intervalle $[a_0, a_1)$, $\ldots, [a_q, \infty)$. Man erhält für die stetige Verweildauer T_s

$$P(T_s \geq a_t \mid x) = S(a_t \mid x) = exp(-\int_0^{a_t} \lambda(u)\, du)$$
$$= exp(-exp(x'\beta) \int_0^{a_t} \lambda_0(u)\, du) = S_0(a_t)^{exp(x'\beta)}$$

und mit

$$\theta_t := ln(\int_0^{a_t} \lambda_0(u)\, du)$$

erhält man unmittelbar

$$-ln\, P(T_s \geq a_t \mid x) = exp(\theta_t + x'\beta) \tag{3-4}$$

Man beachte, daß man die Form der stetigen Hazardrate in (3.1) nicht unmittelbar für die diskrete Hazardrate übernehmen kann. Man muß vielmehr wie in der eben durchgeführten Ableitung zuerst die Wahrscheinlichkeitsverteilung der diskreten Zufallsvariable Verweildauer ermitteln.

Für die diskrete Zeitdauer $T \in \{1, \ldots, q+1\}$ ist Modell (3.4) wegen $P(T > t) = P(T_s \geq a_t)$ äquivalent formulierbar in der üblichen Form des *gruppierten Cox-Modells* (Kalbfleisch/Prentice, 1973)

$$ln(-ln\, P(T > t \mid x)) = \theta_t + x'\beta \quad . \tag{3-5}$$

für $t = 1, \ldots, q$.

Der Parameter θ_t, $t = 1, \ldots, q$, läßt sich hier als ein Parameter auffassen, in den unmittelbar die "Baseline"-Hazardrate eingeht. Die "Baseline"-Hazardrate wird nicht selbst geschätzt, sondern nur in ihrer "verdichteten" Form als Parameter θ_t. Das Modell läßt sich als Cox-Modell bei diskreten Beobachtungen betrachten. Gilt für die zugrundeliegende stetige Verweildauer das Cox-Modell, aber die Dauer wird nur diskret beobachtet, so gilt für die diskreten Beobachtungen Modell (3.5). Modell (3.5) resultiert auch, wenn für die Verweildauer im Cox-Modell eine diskrete Verteilung anstatt einer Gruppierung in Intervallen angenommen wird (vgl. Kalbfleisch/Prentice 1980, S.36). Der Parametervektor β in (3.5) ist identisch mit dem entsprechenden Gewichtsvektor des stetigen Cox-Modells.

Eine äquivalente Darstellung des Modells (3.5) erhält man in den Wahrscheinlichkeiten durch

$$P(T = t \mid x) = exp(-exp(\theta_{t-1} + x'\beta)) - exp(-exp(\theta_t + x'\beta)) \qquad (3\text{–}6)$$

für $t = 1, \ldots, q+1$, wobei $\theta_0 = -\infty$, $\theta_{q+1} = \infty$ gesetzt wird. Für die diskrete Hazardrate erhält man

$$\begin{aligned} \lambda(t \mid x) &= 1 - exp(-exp(\theta_t + x'\beta))/exp(-exp(\theta_{t-1} + x'\beta)) \\ &= 1 - \{exp(-exp(x'\beta))\}^{exp(\theta_t) - exp(\theta_{t-1})}. \end{aligned} \qquad (3\text{–}7)$$

Während für das zugrundeliegende Cox-Modell die Proportionalität des (stetigen) Hazards gilt, gilt sie, wie man aus (3.7) unmittelbar erhält, für die diskrete Hazardfunktion des Modells i. a. nicht mehr. Die Proportionalität der logarithmierten Survivorfunktion allerdings überträgt sich auf die diskrete Survivorfunktion. Man erhält

$$\frac{lnS(t \mid x_1)}{lnS(t \mid x_2)} = exp((x_1 - x_2)'\beta) \quad ,$$

wobei der zeitabhängige Parameter θ_t verschwindet.

Eine geringfügig einfachere Darstellung der Hazardrate, die für die Maximum-Likelihoodschätzung in Abschnitt 3. 4 von Bedeutung ist, erhält man durch die Umparametrisierung mit

$$\gamma_t := ln(exp(\theta_t) - exp(\theta_{t-1})) \qquad \text{für } t = 1, \ldots, q \quad ,$$

wobei $\theta_0 = -\infty$.

Dann ist das Modell (3.7) äquivalent zum Modell

$$\lambda(t \mid x) = 1 - exp(-exp(\gamma_t + x'\beta)) \qquad (3\text{–}8)$$

für $t = 1, \ldots, q$.

3.1.2 Erweiterungen des Modells

Eine unmittelbare Verallgemeinerung des gruppierten Proportional-Hazard Modells stellt das von Aranda-Ordaz (1983) vorgeschlagene Modell dar. Unter Hinzunahme eines weiteren Parameters α formuliert Aranda-Ordaz das Modell

$$\begin{aligned} ln(-ln(1-\lambda(t \mid x))) &= \gamma_t + x'\beta \qquad \text{für } \alpha = 0 \\ [\{-ln(1-\lambda(t \mid x))\}^{\alpha} - 1]/\alpha &= \gamma_t + x'\beta \qquad \text{für } \alpha \neq 0 \end{aligned} \tag{3-9}$$

für $t = 1, \ldots, q$.

Für den Spezialfall $\alpha = 0$ erhält man nach kurzer Ableitung

$$\begin{aligned} ln(-ln(P(T > t \mid x))) &= ln(e^{\gamma_1} + \ldots + e^{\gamma_t}) + x'\beta \\ &= \theta_t + x'\beta \end{aligned}$$

und damit das gruppierte Cox-Modell in umparametrisierter Form.

Für den Spezialfall $\alpha = 1$ erhält man das Modell

$$-ln(1-\lambda(t \mid x)) = (1+\gamma_t) + x'\beta \tag{3-10}$$

Modell (3.10) läßt sich als diskrete Version eines *additiven* stetigen Modells verstehen. Anstatt wie im Cox-Modell die multiplikative Form $\lambda(t \mid x) = \lambda_0(t)exp(x'\beta)$ für die stetige Zeit t anzunehmen, läßt sich auch ein additives stetiges Modell

$$\lambda(t \mid x) = \lambda_0(t) + x'\beta \tag{3-11}$$

für stetige Zeit t zugrundelegen. Betrachtet man nun die diskrete Version,wobei für die stetige Zeitdauer gilt $T_s \in [a_{r-1}, a_r)$ genau dann, wenn für die diskrete Dauer gilt $T = r$, erhält man das diskrete Modell

$$-ln(1-\lambda(t \mid x)) = \rho_t - \rho_{t-1} + (a_t - a_{t-1})x'\beta \tag{3-12}$$

wobei $\rho_r = \int_0^{a_r} \lambda_0(u)\, du$.

Setzt man $\varepsilon_t = \rho_t - \rho_{t-1}$ und nimmt eine konstante Intervallbreite $\Delta = a_t - a_{t-1}$ an, entspricht das Modell (3.12) dem Spezialfall (3.10) des Modells von Aranda-Ordaz.

Das Gesamt–Modell (3.9) umfaßt somit in diskreter Version sowohl multiplikative Modelle von der Art des Cox–Modells, als auch additive Hazard–Modelle der Form (3.11).

Für den Spezialfall $\alpha = 1$ erhält man eine ähnliche Eigenschaft wie für das gruppierte Cox–Modell. Während im gruppierten Cox–Modell das Verhältnis der logarithmierten Survivorfunktion zweier verschiedener Einflußvektoren von der Zeit unabhängig ist, ist für das Modell von Aranda–Ordaz für $\alpha = 1$ das Verhältnis

$$ln\{\frac{S(t \mid x_1)}{S(t \mid x_2)}\} = (t-1)(x_2 - x_1)'\beta$$

nur durch den Faktor $t-1$ von der Zeit abhängig. Dies ergibt sich unmittelbar aus der Survivorfunktion

$$ln\, P(T \geq t \mid x) = -(t-1)(1 + x'\beta) - \gamma_{t-1} - \ldots - \gamma_1 \quad .$$

Eine weitere Version des angeführten Modells (3.9) betrachten Tibshirani & Ciampi (1983), die den Einflußterm $\gamma_t + x'\beta$ ersetzen durch den polynomialen Term

$$\gamma_t + \sum_{j=1}^{s} x'\beta_j(a_t^j - a_{t-1}^j).$$

Tibshirani & Ciampi (1983) demonstrieren an einem Datensatz, daß diese Modellvariante eine bessere Anpassung erzielen kann.

3.2 Proportionalität der diskreten Hazardfunktionen

Das Cox–Modell für stetige Zeit

$$\lambda(t \mid x) = \lambda_0(t) exp(x'\beta) \tag{3–13}$$

führt im diskreten Fall zum gruppierten Cox–Modell (3.5), wobei die zeitunabhängige Proportionalität der Hazardraten für Gruppen mit verschiedenen Kovariablenvektoren verlorengeht.

Ein Modell, in dem diese Eigenschaft für die diskreten Hazardraten gilt, erhält man durch eine Formulierung der *diskreten* Hazardraten analog zu (3.13) mit

$$\lambda(t \mid x) = \lambda_{0t} exp(x'\beta) \qquad \text{für } t = 1, \ldots, q \tag{3–14}$$

Für Modell (3.14) gilt

$$\frac{\lambda(t \mid x_1)}{\lambda(t \mid x_2)} = exp((x_1' - x_2')\beta) \qquad \text{für } t = 1, \ldots, q$$

und es wird daher im weiteren als *Modell mit proportionaler diskreter Hazardrate* bezeichnet.

Das Modell ist jedoch kein gruppiertes Cox–Modell, wie sich aus der entsprechenden Survivorfunktion

$$S(t \mid x) = \prod_{k=1}^{t-1} (1 - \lambda(k \mid x)) = \prod_{k=1}^{t-1} (1 - \lambda_{0k} exp(x'\beta))$$

unmittelbar ersehen läßt.

Modell (3.14) ist die rein diskrete Variante des in Kapitel 5 behandelten Exponentialmodells mit konstantem Hazard in den Intervallen. Da $\lambda(t \mid x)$ in Modell (3.14) eine bedingte Wahrscheinlichkeit ist, gilt die Restriktion $0 \leq \lambda(t \mid x) \leq 1$. Dies hat im Gegensatz zu den anderen hier betrachteten Modellen zur Folge, daß der zulässige Bereich des Gewichtsvektors β eingeschränkt ist. Wie bei verallgemeinerten linearen Modellen mit identischer Linkfunktion ist insbesondere bei stetigen Einflußgrößen darauf zu achten, daß die Hazardratenschätzungen $\hat{\lambda}(t \mid x)$ innerhalb des zulässigen Bereichs liegen.

3.3 Logistische Modelle

Ein weitverbreitetes Modell in der Regressionsanalyse mit kategorialer abhängiger Variable ist das logistische Modell.

Für dichotome abhängige Variable $Y \in \{0,1\}$ ist das Modell von der Form

$$P(Y = 1 \mid x) = \frac{exp(\theta + x'\beta)}{1 + exp(\theta + x'\beta)}$$

$$P(Y = 0 \mid x) = 1 - P(Y = 1 \mid x) \quad .$$

Betrachtet man als abhängige Variable die bedingte diskrete Verweildauer $T \mid T \geq t, x$ (für festes t) und unterscheidet nur die beiden Ereignisse $\{T = t \mid T \geq t, x\}$ und $\{T > t \mid T \geq t, x\}$, ist das entsprechende logistische Modell von der Form

$$P(T = t \mid T \geq t, x) = \frac{exp(\theta_t + x'\beta)}{1 + exp(\theta_t + x'\beta)} \qquad (3\text{–}15)$$

Legt man eine sequentielle Betrachtungsweise zugrunde, so daß immer wenn t erreicht ist, ein dem logistischen Modell (3.15) entsprechender Zufallsprozeß abläuft (vgl. Abschnitt 3.4), erhält man das *logistische Modell für die Hazardraten*

$$\lambda(t \mid x) = \frac{exp(\theta_t + x'\beta)}{1 + exp(\theta_t + x'\beta)} \qquad \text{für } t = 1, \ldots, q\,,$$

das von Cox (1972) vorgeschlagen und von Thompson (1977) ausführlich behandelt wird. Ein äquivalente Formulierung des Modells erhält man mit

$$ln \frac{P(T = t \mid x)}{P(T \geq t \mid x)} = \theta_t + x'\beta \qquad \text{für } t = 1, \ldots, q\,.$$

Die "Ausfallwahrscheinlichkeiten" ergeben sich für $t = 1, \ldots, q$ zu

$$P(T = t \mid x) = \frac{exp(\theta_t + x'\beta)}{\prod_{k=1}^{t} (1 + exp(\theta_k + x'\beta))} \quad .$$

Eine spezielle Variante des logistischen Modells wird von Mantel/Hankey (1978) betrachtet. Der zeitabhängige Parameter θ_t wird ersetzt durch eine zeitabhängige Funktion $h(t)$. Im Modell

$$\lambda(t \mid x) = \frac{exp(h(t) + x'\beta)}{1 + exp(h(t) + x'\beta)}$$

ist $h(t)$ eine von t abhängige, festgesetzte Funktion, z.B. mit $h(t) = \sum_{i=0}^{r} \gamma_i t^i$ ein Polynom r–ten Grades. Die Koeffizienten γ_i des Polynoms werden dann als Parameter mitgeschätzt. Insbesondere für eine große Anzahl q von Intervallen und niedrigem Grad r des Polynoms läßt sich damit die Anzahl der neben β zu schätzenden Parameter von q auf $r+1$ verringern.

TAB. 3.1: ÜBERSICHT ÜBER DIE BEHANDELTEN DISKRETEN VERWEILDAUERMODELLE

	Formulierung in $\lambda(t\|x)$	alternative Modelldarstellung
Gruppiertes Cox-Modell	$\lambda(t\|x) = 1 - \exp(-\exp(\gamma_t + x'\beta))$	$\ln(-\ln P(T>t\|x)) = \theta_t + x'\beta$ $t=1,\ldots,q,\ \theta_o = -\infty$
Modell nach Aranda-Ordaz	$\lambda(t\|x) = 1 - \exp(-\exp(\gamma_t + x'\beta))\ \alpha = 0$ $\lambda(t\|x) = 1 - \exp\{(1+\lambda(\gamma_t + x'\beta)^{1/\alpha}\}\ \alpha \neq 0$	$\ln(-\ln(1-\lambda(t\|x)) = \theta_t + x'\beta\ \alpha = 0$ $[\{-\ln(1-\lambda(t\|x))\}^{\alpha} - 1]/\alpha = \theta_t + x'\beta\ \alpha \neq 0$
Proportionalität der diskreten Hazard-raten	$\lambda(t\|x) = \lambda_{ot}\exp(x'\beta)$	$\ln\lambda(t\|x) = \ln\lambda_{ot} + x'\beta$
Logistisches Modell	$\lambda(t\|x) = \exp(\theta_t + x'\beta)/(1+\exp(\theta_t + x'\beta))$	$\ln(P(T=t\|x)/P(T>t\|x) = \theta_t + x'\beta$

3.4 Sequentielle Modelle auf der Basis latenter Variablen

Eine allgemeine Familie diskreter Verweildauermodelle, die das logistische Modell (3.15) als Spezialfall enthält, resultiert aus einem sequentiellen Mechanismus. Dieser in der Konstruktion kategorial-ordinaler Regressionsmodelle verwendete Ansatz (Amemyia 1975, Tutz 1987, 1988) läßt sich problemlos auf die Modellierung diskreter Verweildauern übertragen. Der sequentielle Ansatz basiert auf Vorstellungen über die Wirkungsweise latenter Variablen. Die resultierenden Parametrisierungen lassen sich jedoch auch ohne die Rekursion auf die latenten Variablen interpretieren. Diese sind eher als Motivation zur Modellbildung zu verstehen.

Seien $U_t(x) = u_t(x) + \varepsilon_t$ nicht beobachtbare Zufallsvariablen, wobei $u_t(x) = E(U_t(x))$ den Erwartungswert der vom Kovariablenvektor x abhängigen Zufallsvariable $U_t(x)$ bezeichnet und ε_t eine Störvariable mit Verteilungsfunktion F darstellt. Der Prozeß beginnt mit dem ersten erreichbaren Zustand, der dem Intervall $[a_0, a_1)$ entspricht. Die diskrete Verweildauer endet in diesem Anfangszustand entsprechend dem Mechanismus

$$T = 1 \quad \text{gegeben} \quad x \quad \Longleftrightarrow \quad U_1(x) < \theta_1.$$

Damit endet die Verweildauer, wenn die latente Variable $U_t(x)$ eine fixe Schwelle θ_1 nicht überschreitet. Gilt hingegen $U_1(x) \geq \theta_1$, wird eine neue latente Variable $U_2(x)$ mit unabhängiger Störgröße ε_2 realisiert, und es gilt

$$T = 2 \quad \text{gegeben} \quad T \geq 2, x \quad \Longleftrightarrow \quad U_2(x) < \theta_2.$$

Der Prozeß endet mit $U_2(x) < \theta_2$. Wird jedoch die Schwelle θ_2 überschritten, wird der Prozeß analog fortgesetzt, wobei allgemein gilt

$$T = t \quad \text{gegeben} \quad T \geq t, x \quad \Longleftrightarrow \quad U_t(x) < \theta_t.$$

Diesen sequentiellen Mechanismus zugrundegelegt, erhält man mit der Verteilungsfunktion F der Störgrößen ε_t unmittelbar das Modell

$$\lambda(t \mid x) = P(T = t \mid T \geq t, x) = F(\theta_t - u_t(x)).$$

Die latenten Variablen $U_t(x)$ lassen sich als die Summe der die Lebensdauer verlängernden Kräfte vorstellen. Überschreiten sie eine bestimmte Schwelle,

wird zumindest das nächste Intervall erreicht. Die Zufallsvariablen $U_t(x)$ und insbesondere die Erwartungswerte $u_t(x)$ hängen dabei von dem Kovariablenvektor x ab und werden dem stochastischen Charakter empirischer Prozesse entsprechend von einer Störung (ε_t) überlagert. Der Erwartungswert der latenten Variablen ist dabei spezifisch für ein Intervall, was durch den Index t zum Ausdruck kommt. Alternativ läßt sich $U_t(x)$ auch als die im t-ten Intervall laufende stetige Lebensdauer interpretieren. Dann ergeben sich auch die Schwellen unmittelbar als die Differenz der Intervallgrenzen $a_t - a_{t-1}$. Im weiteren wird jedoch von unbekannten und damit zu schätzenden Schwellen ausgegangen.

Wählt man mit $u_t(x) = x'\beta_t$ einen linearen Ansatz, erhält man das allgemeine Modell

$$\lambda(t \mid x) = F(\theta_t - x'\beta_t), \qquad t = 1, \ldots, q, \tag{3–16}$$

wobei x keine Konstante mehr enthält.

Wählt man für F die logistische Verteilung $F(z) = 1/(1 + exp(-z))$ ergibt sich mit

$$\lambda(t \mid x) = \frac{exp(\theta_t - x'\beta_t)}{1 + exp(\theta_t - x'\beta_t)} \tag{3–17}$$

eine Erweiterung des logistischen Modells (3.15). Während in (3.15) der Einfluß der Kovariablen nicht vom Intervall bzw. der Verweildauerkategorie abhängt, ist in (3.17) der Gewichtsvektor β_t intervallspezifisch.

Prinzipiell lassen sich in (3.16) alle stetigen Verteilungsfunktionen wählen. Aus Gründen der Identifizierbarkeit empfiehlt sich jedoch eine Beschränkung auf stetige Verteilungsfunktionen, die streng monoton zumindest auf dem Bereich $\{z \mid 0 < F(z) < 1\}$ sind. Lineare Transformationen lassen das Modell (3.16) unverändert, d.h. Modelle mit der Verteilungsfunktion $F(z)$ bzw. $G(z) = F(az + b)$, $a > 0$, sind zueinander äquivalent. Es genügt daher, immer nur die Verteilungsform dieser Äquivalenzklassen zu untersuchen. Bei entsprechend reichhaltiger Struktur der Kovariablenvektoren erhält man daraus die einzigen Äquivalenzklassen (vgl. Tutz, 1988).

Eine Probit-Variante des allgemeinen Modells (3.16) ergibt sich aus der An-

nahme einer Standardnormalverteilung für ε_t mit

$$\lambda(t \mid x) = \int_{-\infty}^{\theta_t - x'\beta_t} (2\pi)^{-\frac{1}{2}} exp(-z^2/2)\, dz$$

Ebenso kann man für F auch die Exponentialverteilung $F(z) = 1 - exp(-z)$, wenn $z \geq 0$, wählen. Diese führt zu dem einfachen Modell

$$\lambda(t \mid x) = 1 - exp(\theta_t - x'\beta_t).$$

Die Wahl $F(z) = exp(z)$ für $z < 0$ führt zum Modell (3.14) der diskreten proportionalen Hazardraten aus Abschnitt 3.2.

Man erhält, wählt man für F die doppelte Exponentialverteilung $F(z) = 1 - exp(-exp(z))$, unmittelbar das gruppierte Cox–Modell in der Formulierung (3.8), wobei dort die Schwellenwerte durch γ_t bezeichnet sind und der additive Term ein positives Vorzeichen besitzt. Das negative Vorzeichen des additiven Terms in 3.16 resultiert aus der Ableitung aus dem latenten Mechanismus, entspricht jedoch nur einer einfachen Umparametrisierung. Das gruppierte Cox–Modell läßt sich auch aus einem kumulativen Ansatz ableiten, der Schwellenwerte auf völlig andere Art zugrundelegt (vgl. McCullagh (1980)). Eine enge Verwandschaft zwischen kumulativen und sequentiellen Modellen ergibt sich für wenige Verteilungen wie die doppelte Exponentialverteilung und die Exponentialverteilung. Für letzte allerdings erhält man keine Äquivalenz der Ansätze (vgl. Tutz (1988)).

Durch Spezifizierung der Verteilungsfunktion F wird ein Modell der sequentiellen Modellklasse (3.16) bestimmt. Die daraus resultierenden Modelle sind im allgemeinen verschieden. Die Wahl der Verteilungsfunktion läßt sich zum einen abhängig machen von inhaltlichen Überlegungen zur Adäquatheit eines Ansatzes, zum anderen von statistischen Kriterien wie Anpassungsmaßen und Residuenanalyse. Die Betrachtung einer parametrisierten Familie von Verteilungen und damit parametrisierter Linkfunktionen, wie sie z.B. Pregibon (1980) für verallgemeinerte lineare Modelle vorschlägt, verlangt besondere Schätzverfahren, wenn die verteilungsspezifischen Parameter mitgeschätzt werden sollen. Dasselbe gilt auch für den betrachteten Ansatz von Aranda-Ordaz (1983).

3.5 Maximum–Likelihood–Schätzung

In diesem Abschnitt wird die Maximum–Likelihood–Schätzung der unbekannten Modellparameter für Ein–Episoden–Modelle bei einem Zielzustand in allgemeiner Form behandelt. Dabei ist ein bei Verlaufsdaten zusätzlich auftretendes Problem zu beachten, die Zensierung. Wie bereits in Abschnitt 1.3 dargestellt, können in der Regel nicht alle Verweildauern bzw. Lebenszeiten bis zum Ende beobachtet werden. Man erhält dann für einen Teil der Stichprobe rechts zensierte Daten. Dies ist insbesondere auch bei kurzen Panel-Erhebungen der Fall, bei denen versucht wird, durch Retrospektivbefragung Informationen über wichtige Ereignisse zwischen Panel-Wellen zu erhalten. Stehen nur wenige Panel-Wellen zur Verfügung, kann es durchaus vorkommen, daß die untersuchten Verweildauern (z.B. Dauer der Arbeitslosigkeit) zum Zeitpunkt des aktuellen Interviews noch nicht abgeschlossen sind. Eine ähnliche Situation entsteht jedoch auch, wenn Personen während der Studie aus anderen Gründen ausscheiden, z.B. wegen eines Umzugs oder Klinikwechsels in einer medizinischen Studie, und zur weiteren Befragung bzw. Untersuchung nicht mehr zur Verfügung stehen.

Im statistischen Modell wird die Zensierung durch einen Zensierungsindikator

$$\delta_i = \begin{cases} 1, & \text{falls } t_i \text{ nicht zensiert ist,} \\ 0, & \text{falls } t_i \text{ zensiert ist,} \end{cases} \quad i = 1, \ldots, n, \tag{3–18}$$

zum Ausdruck gebracht. Es gibt verschiedene statistische Modelle für Zensierungsmechanismen. Man vergleiche dazu die Ausführungen in Abschnitt 1. 3. Eine wichtige Variante ist das Modell des "random censoring" (Modell 3 in 1. 3). Dabei werden die Zeiten T_i und die Zensierungszeiten C_i als unabhängige Zufallsvariablen aufgefaßt, wobei C_i die Zeit des i–ten Individuums bis zum Ausscheiden aus der Untersuchung bezeichnet. Beobachtbar in einer Untersuchung ist nur die Zufallsvariable $min(T_i, C_i)$ zusammen mit der Indikatorfunktion, für die gilt $\delta_i = 1$, wenn $T_i \leq C_i$, $\delta_i = 0$, wenn $C_i < T_i$. Die Wahrscheinlichkeit beim "random censoring" für eine unzensierte Beobachtung $(t_i, \delta_i = 1)$, d.h. $T_i = t_i$ und $T_i < C_i$, ist dann (ohne Berücksichtigung der Kovariablen)

$$P(T_i = t_i)P(C_i \geq t_i) \quad , \tag{3–19}$$

während die Wahrscheinlichkeit für eine zensierte Beobachtung $(t_i, \delta_i = 0)$, d.h. $T_i \geq t_i$ und $C_i = t_i$,

$$P(T_i \geq t_i)P(C_i = t_i) \tag{3-20}$$

beträgt. Dabei wird vorausgesetzt, daß die Zensierung jeweils zu Beginn eines Intervalls erfolgt. Es wird somit nur die Minimalinformation ausgenutzt, daß der Beginn des Zeitintervalls erreicht wurde. Diese Annahme kann durch andere ersetzt werden (z.B. Zensierungen erfolgen grundsätzlich am Ende eines Intervalls), wobei dann geringfügige Modifikationen erforderlich sind.

(3.19) und (3.20) lassen sich zusammenfassen und für den Beitrag des Individuums i zur Likelihoodfunktion — ohne Kovariablen — ergibt sich

$$P(T_i = t_i)^{\delta_i} P(T_i \geq t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i} \tag{3-21}$$

Unterstellt man, daß die Verteilung der Zensierungszeiten C_i nicht von den für die Verteilung von T_i relevanten Parametern abhängt, insbesondere bei Hinzunahme von Kovariablen nicht von den Regressionskoeffizienten (Zensierungsmechanismus ist "nicht informativ"), können die beiden letzten Faktoren in (3.21) zu einem in Abhängigkeit von den die Verteilung von T_i determinierenden Parametern konstanten Term c_i zusammengefaßt werden, und die Gesamtlikelihoodfunktion ist von der Form

$$L = c \prod_{i=1}^{n} P(T_i = t_i)^{\delta_i} S_i(t_i)^{1-\delta_i} \tag{3-22}$$

mit $c = \prod c_i$ und der Survivorfunktion $S_i(t_i) = P(T_i \geq t_i)$. Berücksichtigt man schließlich noch die Kovariablen, erhält man für den Beitrag von Individuum i zur Likelihoodfunktion bei gegebenem Kovariablenvektor x_i

$$L_i = P(T_i = t_i \mid x_i)^{\delta_i} P(T_i \geq t_i \mid x_i)^{1-\delta_i} = \lambda(t_i \mid x_i)^{\delta_i} P(T_i \geq t_i \mid x_i) \quad .$$

Wegen (1.19) folgt

$$L_i = \lambda(t_i \mid x_i)^{\delta_i} \prod_{k=1}^{t_i - 1} (1 - \lambda(k \mid x_i)) \tag{3-23}$$

und

$$L = \prod_{i=1}^{n} L_i. \tag{3-24}$$

Der größte Teil der betrachteten Verweildauermodelle läßt sich im Rahmen der verallgemeinerten Modelle schätzen. Dabei wird die formale Ähnlichkeit der Likelihoodfunktion (3.21) zur Likelihood in verallgemeinerten linearen Modellen ausgenutzt.

Verallgemeinerte lineare Modelle sind sehr ausführlich in Fahrmeir/Hamerle (1984), Kap. 7 oder McCullagh/Nelder (1983) beschrieben. Bei einem (univariaten) verallgemeinerten linearen Modell geht man aus von n unabhängigen Zuvallsvariablen y_i mit $E(y_i) = \mu_i$, also

$$y_i = \mu_i + \varepsilon_i \qquad \text{mit } E(\varepsilon_i) = 0, \quad i = 1, \ldots, n\,.$$

Neben y_i werden erklärende Variablen $x_i = (x_{i0} = 1, x_{i1}, \ldots, x_{ip})'$ erhoben, die y_i als Linearkombination $\gamma_i = x_i'\beta$ beeinflussen, und zwar ist der Erwartungswert $\mu_i = E(y_i)$ mit $x_i'\beta$ über eine Linkfunktion g durch

$$g(\mu_i) = x_i'\beta$$

verbunden. Die Umkehrfunktion g^{-1} wird Responsefunktion genannt, und es gilt

$$\mu_i = g^{-1}(x_i'\beta) \quad .$$

Die Responsefunktion g^{-1} wird auch gelegentlich mit h bezeichnet. Darüber hinaus wird noch vorausgesetzt, daß die Dichte von y_i zu einer Exponentialfamilie gehört. Dies ist jedoch für die hier betrachteten diskreten Verweildauer–Modelle stets erfüllt. Gilt insbesondere $y_i \in \{0, 1\}$, so ist

$$\mu_i = P(y_i = 1 \mid x) \quad ,$$

und für die Likelihoodfunktion erhält man

$$L = \prod_{i=1}^{n} P(y_i = 1 \mid x_i)^{y_i}(1 - P(y_i = 1 \mid x_i))^{1-y_i} \qquad (3\text{–}25)$$

Betrachtet man nun die bedingte diskrete Verweildauer einer Person $T_i \mid T_i \geq t, x$ (für festes t) und unterscheidet nur die beiden Ereignisse $\{T_i = t \mid T_i \geq t, x_i\}$ und $\{T_i \geq t \mid T_i \geq t, x_i\}$ bei gegebenem Kovariablenvektor x_i, so sind bis auf das Modell von Aranda–Ordaz für $\alpha \neq 0$ alle behandelten Modelle von der Form

$$P(T_i = t \mid T_i \geq t, x_i) = h(\theta_t + x_i'\beta)\,, \qquad (3\text{–}26)$$

ht von t abhängt. Der Ausdruck
en. Es gilt

$\tilde{\beta}$

$\prime'$)

$_t - (0, \ldots, 1, \ldots, 0)$ den t-ten Einheitsvektor bezeichnet.

Die Modelle vom Typ (3.26) lassen sich somit darstellen durch

$$\lambda(t \mid x_i) = h(\tilde{x}'_{t_i}\tilde{\beta}) .$$

Zum Beispiel erhält man für das gruppierte Cox–Modell in der Form (3.8)

$$\lambda(t \mid x) = 1 - exp(-exp(\theta_t + x'\beta))$$

die Funktion

$$h(y) = 1 - exp(-exp(y)) \quad .$$

Man betrachte t_i unabhängige dichotome Zufallsvariablen $Y_i, \ldots, Y_{it_i}$. Dann gilt für die entsprechende Likelihood

$$L_i = \prod_{r=1}^{t_i} P(Y_{ir} = 1)^{y_{ir}} (1 - P(Y_{ir} = 1))^{1-y_{ir}}$$

und, falls die Y_{ir} bedingte Zufallsvariablen in Abhängigkeit von Einflußvariablen x_i sind, gilt

$$L_i = \prod_{r=1}^{t_i} P(Y_{ir} = 1 \mid x_i)^{y_{ir}} (1 - P(Y_{ir} = 1 \mid x_i)^{1-y_{ir}} . \qquad (3\text{–}27)$$

Dies entspricht der Likelihood von t_i Beobachtungen eines verallgemeinerten linearen Modells mit der Responsefunktion $P(Y_{ir} = 1 \mid x_i) = h(x'_i\beta)$ für eine Responsefunktion h. Setzt man

$$P(Y_{ir} = 1 \mid x_i) = \lambda(r \mid x_i) \quad ,$$

erhält man für den beobachteten Vektor $y_i = (y_{i1}, \ldots, y_{it_i}) = (0, 0, \ldots, 0, 1)$ aus (3.27) die Likelihood (3.23) einer unzensierten Beobachtung der i–ten Person, wenn $\delta_i = 1$ gilt. Betrachtet man nur $t_i - 1$ Zufallsvariablen, erhält man mit

$$L_i = \prod_{r=1}^{t_i-1} P(Y_{ir} = 1 \mid x_i)^{y_{ir}} (1 - P(Y_{ir} = 1 \mid x_i))^{1-y_{ir}}$$

die Likelihoodfunktion (3.23) für $\delta_i = 0$, d.h. einer zensierten Beobachtung, wobei wiederum $P(Y_{ir} = 1 \mid x_i) = \lambda(r \mid x_i)$ gesetzt wird und der Beobachtungsvektor $y_i = (y_{ir}, \ldots, y_{it_i-1}) = (0, \ldots, 0)$ vorliegt.

Man erhält somit, wenn $\lambda(t \mid x_i) = h(\tilde{x}'_{t_i}\tilde{\beta})$ in Abhängigkeit von einem Parametervektor $\tilde{\beta}$ modelliert ist, dieselbe Likelihood, wie für ein verallgemeinertes lineares Modell, wobei für $\delta_i = 1 \quad t_i$ Beobachtungen $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0, 1)$ und analog für $\delta_i = 0 \quad t_i - 1$ Beobachtungen $(y_{i1}, \ldots, y_{it_i-1}) = (0, \ldots, 0)$ zu setzen sind.

Die beobachteten Daten für ein verallgemeinertes Modell sind von der Form $(y_i, x_i), i = 1, \ldots, n$, wobei y_i eine Realisation der abhängigen Variablen bei gegebenem Einflußvektor x_i darstellt. Die Daten sind dann auflistbar in der Form

Abh. Variable	Unabh. Variablenvektor
y_1	x'_1
y_2	x'_2
$\vdots$	$\vdots$
y_n	x'_n

und die Beobachtung für das Individuum bzw. Objekt i generiert gewöhnlich die i–te Zeile der Designmatrix.

Die ML–Schätzung des Verweildauermodells

$$\lambda(t \mid x_i) = h(\tilde{x}'_{t_i}\tilde{\beta})$$

ergeben sich dann als ML–Schätzungen des verallgemeinerten linearen Modells

$$P(Y = 1 \mid x_i) = h(\tilde{x}'_{t_i}\tilde{\beta}),$$

wobei die Anzahl der Beobachtungen und damit die Anzahl der Zeilen der Designmatrix künstlich erhöht werden. Liegt für das i–te Individuum ein tatsächlicher Übergang in den absorbierenden Endzustand (d.h. eine nicht zensierte Beobachtung) vor, erhält man zu dieser Beobachtung die t_i Zeilen

	abhängige dichotome Var $y \in \{0,1\}$	unabhängiger Variablenvektor
1	0	(e_1', x_i')
2	0	(e_2', x_i')
$\vdots$	$\vdots$	$\vdots$
t_i	1	(e_{t_i}', x_i')

für $\delta_i = 1$, $T_i = t_i$.

Der Teil der Datenmatrix für eine zensierte Beobachtung ist

	abhängige dichotome Var $y \in \{0,1\}$	unabhängiger Variablenvektor
1	0	(e_1', x_i')
2	0	(e_2', x_i')
$\vdots$	$\vdots$	$\vdots$
$t_i - 1$	0	(e_{t_i-1}, x_i')

für $\delta_i = 0$, $C_i = t_i$.

Man berechnet somit die Likelihood–Schätzungen des verallgemeinerten linearen Modells mit

$$N = \sum_{i=1}^{n} (\delta_i t_i + (1 - \delta_i)(t_i - 1))$$

Beobachtungen.

Man beachte, daß hier die Äquivalenz der Likelihoodfunktion von diskreten Verweildauermodellen und verallgemeinerten linearen Modellen nur nach dem Gesichtspunkt der Schätzbarkeit gilt. Das beschriebene Verfahren ermöglicht es, die Maximum–Likelihood–Schätzungen der Parameter von dis-

kreten Verweildauermodellen aus Programmpaketen für verallgemeinerte lineare Modelle wie GLIM oder GLAMOUR zu gewinnen. Dagegen sind die in der Theorie der verallgemeinerten linearen Modelle geltenden asymptotischen Aussagen (vgl. z.B. Fahrmeir/Kaufmann (1985)) und damit die Gültigkeit von Verteilungsaussagen für Teststatistiken nicht ohne weiteres übernehmbar. Eine approximative Maximum–Likelihood–Schätzung für das gruppierte Cox–Modell mit kategorialen Einflußgrößen wird von Pierce et al. (1979) abgeleitet.

3.6 Anwendungsbeispiele

(a) Dauer des Krankenhausaufenthaltes

Im folgenden wird das in Abschnitt 3.3 beschriebene logistische Modell zur Analyse der Dauer des Krankenhausaufenthaltes nach Unfällen im Schulsport verwendet. Untersucht wird die Dauer des stationären Aufenthaltes von $n = 554$ im Jahre 1981 bei Ballspielen im Schulsport verletzten Schülern in Abhängigkeit von demographischen (Alter, Geschlecht, Schulart), die Unfallgenese beschreibenden (verletzungsauslösende Spielphase) sowie biometrischen Kovariablen (Verletzungsdiagnosen). Dieser Datensatz, den man in Kemény/Rothmeier/Hamerle (1986) ausführlich beschrieben findet, ist ein Teil einer Studie über Krankenhausverweildauern, die im Rahmen einer von der gesetzlichen Unfallversicherung in den Jahren 1982 und 1984 erstellten medizinischen Dokumentation zur Rehabilitation Unfallverletzter und arbeitsbedingter Erkrankter durchgeführt wird. Insgesamt wurden 64 Kovariablen erhoben. Es wurde das zeitstetige Cox–Modell zugrundegelegt und mit Hilfe einer Variablenselektion konnte die Anzahl der Kovariablen auf vier signifikante Einflußgrößen reduziert werden, nämlich auf Schulart "Hauptschule", verletzter Körperteil "Kniescheibe, Oberschenkel, Hüfte", verletzter Körperteil "Kniegelenk, Schienbein, Unterschenkel, Knöchelbereich" und Verletzungsart "Fraktur". Für Details vergleiche man Kemény et al (1986).

Die Dauer des stationären Krankenhausaufenthaltes wird in Tagen gemessen, so daß der Datensatz viele gleiche Meßwerte (Ties) enthält. Deshalb er-

scheint es sinnvoll, zur Validierung der Resultate auch ein diskretes Modell anzupassen. Es wurden 11 Zeitintervalle gebildet: 14, 15, 16, 17, 18-19, 20-21, 22-24, 25-26, 27-30, 31-37, über 37 (Tage). In einer ersten Teilauswertung wurde der logistische Ansatz (3.15) zugrundegelegt, wobei nur die oben erwähnten vier signifikanten Kovariablen verwendet wurden. Die Maximum-Likelihood-Schätzung der unbekannten Modellparameter wurde nach der im letzten Abschnitt beschriebenen Vorgehensweise durchgeführt. Dabei wurde das Programm GAUSS von L. E. Edlefsen und S. D. Jones eingesetzt. Für die Regressionskoeffizienten und deren (geschätzte) Standardabweichung ergab sich:

Kovariable	$\hat{\beta}_j$	$\hat{\sigma}_j$
Verletzter Körperteil Kniescheibe, Oberschenkel Hüfte	-1.152	0.222
Verletzter Körperteil Kniegelenk Schienbein, Unterschenkel, Knöchelbereich	-0.564	0.123
Verletzungsart Fraktur	-0.622	0.126
Schulart Hauptschule	-0.264	0.121

Die Ergebnisse stimmen in Richtung und Größenordnung mit denjenigen des Modells mit stetig gemessener Zeit überein. Für eine Interpretation der Resultate vergleiche man wieder Kemény et al. (1986).

(b) Dauer der Arbeitslosigkeit

In einer Studie zur Dauer der Arbeitslosigkeit liegt die Verweildauer diskret vor. Die betrachteten Kovariablen und die Diskretisierung der abhängigen Variable Dauer ergeben sich aus Tabelle 3.1.

Tabelle 3.1: Variablen zur Dauer der Arbeitslosigkeit

		Ausprägungen	Kategorie
x_0	Dauer der Arbeitslosigkeit	bis 1 Monat	1
		1–2 Monate	2
		3–5 Monate	3
		6–11 Monate	4
		12-23 Monate	5
		über 23 Monate	6
(A)	Alter	metrisch	
(G)	Geschlecht	männlich	1
		weiblich	2
(S)	Staatsangehörigkeit	Deutscher	1
		Ausländer	2
(GE)	Gesundheitliche	nein	1
	Einschränkungen	ja	2
(E)	Erwerbstätigkeit vor	ja	1
	der Arbeitslosenmeldung	nein	2
(L)	Finanzielle Leistungen	ja	1
	des Arbeitsamtes	nein	2

In einer Teilstichprobe von 3655 Arbeitslosen mit beruflicher Ausbildung wurde das logistische Modell (3.15) zugrundegelegt. Für das Modell mit sämtlichen Zwei-Faktor-Interaktionen resultierte eine Log-Likelihood von -4856.00, für das Modell mit Haupteffekten erhielt man den Wert -4903.63. Die bedingte Log-Likelihood-Teststatistik von 85.26 ergibt eine wesentlich schlechtere Anpassung des Modells ohne Zwei-Faktor-Interaktionen. Nach Auswahl der nicht-signifikanten Interaktionseffekte ergaben sich die ML-Schätzungen in Tabelle 3.2. Die Log-Likelihood weist mit -4856.84 fast keine Veränderung im Verhältnis zum Modell mit sämtlichen Zwei-Faktoren-Interaktionen auf. Zugrundegelegt wurde jeweils die 0-1–Kodierung der Kovariablen.

Tabelle 3.2: Maximum-Likelihood-Schätzungen für das logistische Modell nach Interaktionsreduktion

Variable	ML-Schätzer	Varianzen	Alphaquantil
θ_1	0.582	1.348	0.615
θ_2	1.388	1.349	0.231
θ_3	1.599	1.352	0.169
θ_4	0.900	1.359	0.439
θ_5	0.903	1.369	0.440
θ_6	1.010	1.387	0.391
A	-0.072	0.000	0.000
G	0.873	0.356	0.143
S	-1.005	0.904	0.290
GE	-1.295	0.906	0.173
E	0.369	0.306	0.504
L	-0.237	0.193	0.588
A*G	-0.009	0.000	0.114
A*GE	0.011	0.000	0.099
A*E	0.024	0.000	0.114
A*L	0.029	0.000	0.000
G*S	0.672	0.171	0.104
G*GE	0.586	0.043	0.004
G*E	-0.482	0.110	0.146
G*L	-1.052	0.021	0.000
S*GE	0.650	0.768	0.457
GE*L	0.354	0.054	0.130
E*L	-0.434	0.101	0.172

Die entsprechende Auswertung für das gruppierte Cox-Modell ergab eine Log-Likelihood von -4858.76 für das Modell mit allen Zwei-Faktoren-Interaktionen und -4901.63 für das Modell mit Haupteffekten. Es wurden wiederum alle Zwei-Faktoren-Interaktionen weggelassen, die sich als nicht signifikant erwiesen.

Die Schätzungen für das reduzierte Modell sind in Tabelle 3.3 wiedergegeben. Die Log-Likelihood ist mit -4859.58 wiederum kaum verändert im Verhältnis

zum Modell mit allen Zwei-Faktor-Interaktionen.

Maximum-Likelihood-Schätzungen für das gruppierte Cox-Modell nach Interaktionsreduktion

Variable	ML-Schätzer	Varianzen	Alphaquantil
θ_1	0.198	0.931	0.836
θ_2	0.833	0.932	0.387
θ_3	0.994	0.933	0.303
θ_4	0.457	0.938	0.636
θ_5	0.477	0.945	0.623
θ_6	0.548	0.957	0.575
A	-0.060	0.000	0.000
G	0.645	0.246	0.193
S	-0.890	0.605	0.252
GE	-1.222	0.616	0.119
E	0.307	0.227	0.518
L	-0.126	0.138	0.734
A*G	-0.006	0.000	0.194
A*GE	0.012	0.000	0.044
A*E	0.020	0.000	0.111
A*L	0.021	0.000	0.000
G*S	0.551	0.110	0.097
G*GE	0.478	0.030	0.005
G*E	-0.430	0.084	0.138
G*L	-0.784	0.015	0.000
S*GE	0.591	0.517	0.411
GE*L	0.321	0.037	0.095
E*L	-0.328	0.074	0.227

Ein Vergleich der beiden Modelle zeigt, daß das Logit-Modell nur eine geringfügig bessere Anpassung aufweist. Die Werte der geschätzten Parameter sind durchwegs vergleichbar. Die absoluten Schätzwerte des Logit-Modells sind jedoch fast immer etwas größer als die Schätzungen des gruppierten Cox-Modells. Zu weiteren Anwendungen für andere Berufsschichten vergleiche man

Dübler (1988).

Das umfassendere Modell von Aranda–Ordaz wurde für die Parameter 0.1, 0.2 und 0.3 geschätzt. Für größeres α war keine Konvergenz der Schätzung zu erreichen. Die Log-Likelihoodwerte ergaben sich für das reduzierte Modell mit denselben Interaktionen wie in den Tabellen 3.1 und 3.2 zu -4859.80, -4860.05, -4860.62 und -4860.59. Die kleinste Log–Likelihood ergibt sich somit für das Modell mit $\alpha = 0.0$. Das gruppierte Cox–Modell erweist sich damit als das am besten angepaßte Modell, wobei die Unterschiede der Anpassungsgüte minimal sind.

Zum Vergleich der Schätzwerte werden in Tabelle 3.3 die Schätzungen für das Modell mit dem vergleichsweise großen $\alpha = 0.3$ angegeben. Im Vergleich zu Tabelle 3.2 ergeben sich Veränderungen bei einigen der Werte, so z.B. bei den Haupteffekten S, GE. Für alle Schätzungen bleibt jedoch das Vorzeichen gleich und für die meisten die Größenordnung vergleichbar.

Tabelle 3.3: Maximum-Likelihood-Schätzungen für das Modell von Aranda–Ordaz mit $\alpha = 0.3$

Variable	ML-Schätzer	Varianzen	Alphaquantil
θ_1	-0.466	0.185	0.279
θ_2	0.029	0.186	0.945
θ_3	0.157	0.187	0.715
θ_4	-0.284	0.189	0.512
θ_5	-0.282	0.192	0.520
θ_6	-0.212	0.199	0.633
A	-0.039	0.000	0.000
G	0.527	0.126	0.137
S	-0.189	0.051	0.405
GE	-0.332	0.040	0.100
E	0.226	0.093	0.459
L	-0.196	0.064	0.440
A*G	-0.006	0.000	0.076
A*GE	0.005	0.000	0.176

A*E	0.012	0.000	0.151
A*L	0.017	0.000	0.000
G*S	0.371	0.064	0.144
G*GE	0.347	0.015	0.005
G*E	-0.257	0.034	0.166
G*L	-0.610	0.007	0.000
GE*L	0.206	0.019	0.134
E*L	-0.232	0.032	0.196

4. Die Einbeziehung von zeitabhängigen Kovariablen

4.1 Modelldarstellung

In den vorangegangenen Kapiteln wurde vorausgesetzt, daß die Kovariablen zu Beginn der Episode eines Individuums gemessen werden und nicht von der Zeit abhängen. In einer Reihe von Anwendungen können jedoch auch die Kovariablen mit der Zeit bzw. Verweildauer variieren. Beispiele hierfür sind Alter, Einkommen, Familienstand oder eine Therapie, die nur während eines bestimmten Zeitraumes angewendet wird.

Für die Art der zeitlichen Veränderung der Kovariablen gibt es verschiedene Möglichkeiten. Eine einfache deterministische Form der Zeitabhängigkeit besteht beispielsweise bei der Variablen Alter, wenn das Alter einer Person zu Beginn des Beobachtungszeitraumes bekannt ist. Kalbfleisch/Prentice (1980, Kap. 5) sprechen in einem solchen Fall von *definierten* Kovariablen. Daneben existieren Kovariablen, die ihrerseits Realisierungen eines stochastischen Prozesses sind. Kalbfleisch/Prentice (1980, Kap. 5) unterscheiden hier zwischen *externen* und *internen* zeitabhängigen Kovariablen. Bei externen zeitabhängigen Kovariablen wird der Pfad des Kovariablenvektors nicht von der Verweildauer eines Individuums beeinflußt. Umgekehrt kann er aber sehr wohl auf die Verweildauer einwirken. Umweltfaktoren, soziale oder ökonomische Rahmenbedingungen zählen häufig zu diesem Typ von zeitabhängigen Kovariablen. Im Gegensatz dazu wird bei internen zeitabhängigen Kovariablen der Kovariablenprozeß von den Individuen, deren Verweildauer analysiert wird, selbst generiert, d.h. die Beobachtung einer konkreten Ausprägung der Kovariablen im Zeitintervall t hängt von den Ergebnissen des Verweildauer-Prozesses selbst ab. Bei internen Kovariablen handelt es sich häufig um parallel verlaufende Prozesse, die mit dem Verweildauer-Prozeß wechselseitig zusammenhängen. Eine formale Definition von externen und internen zeitabhängigen Kovariablen erfolgt in Abschnitt 4. 2.

Die explizite Erhebung der Kovariablenwerte kann von Anwendungsfall zu Anwendungsfall variieren. Bei Untersuchungen in regelmäßigen Abständen, etwa bei Quartalsuntersuchungen in der Medizin, werden die Kovariablenwerte immer an den Schnittgrenzen der Intervalle gemessen. Für das Zeitin-

tervall $[a_{t-1}, a_t)$ werden die Kovariablenwerte $x(t)$ zum Zeitpunkt a_{t-1} erhoben, und am Ende des Intervalls, d.h. bei der nächsten Untersuchung, wird registriert, ob das in Frage stehende Ereignis eingetreten ist. Eine andere Vorgehensweise wird z.B. bei Retrospektivfragen im sozio–ökonomischen Panel gewählt. Hier werden Daten in Bezug auf Erwerbstätigkeit und Einkommensentwicklung auf Monatsebene erfragt. Beispielsweise sollen die Personen ankreuzen, ob sie in einem bestimmten Monat voll erwerbstätig, arbeitslos, in einer Umschulung etc. waren. Es wird also eine bestimmte Zeiteinheit (hier: Monat) zugrundegelegt, und dieses Zeitintervall ist die kleinste Einheit, in der die Daten erhoben werden können. Die aufeinanderfolgenden Zeitintervalle $t = 1, 2, \ldots, q$ entsprechen dann den aufeinanderfolgenden Monaten $1, 2, \ldots, q$.

Wie in den vorangegangenen Abschnitten wird davon ausgegangen, daß die Kovariablen durch eine Linearkombination auf die bedingten Übergangswahrscheinlichkeiten einwirken. Für den im Modell enthaltenen zeitunabhängigen Kovariablenvektor x_1 sei diese Linearkombination $x_1'\gamma$. Für die Einwirkung der zeitabhängigen Kovariablen gibt es mehrere Möglichkeiten, die im einzelnen skizziert werden.

1. Sei der Kovariablenpfad bis zum Intervall t gegeben durch

$$\tilde{z}(t) = (x(1)', \ldots, x(t)'),$$

wobei $x(s)$, $s = 1, \ldots, t$, den Kovariablenvektor zum s–ten Intervall bezeichnet. Es werden Parametervektoren θ_{ts} definiert, deren Dimension gleich der Dimension von $x(s)$ ist. Ferner sei

$$\theta_t = \begin{pmatrix} \theta_{t1} \\ \vdots \\ \theta_{tt} \end{pmatrix} \tag{4–1}$$

der Gesamtvektor bis zum Zeitpunkt t, und es wird angenommen, daß die Einwirkung des Kovariablenpfades der zeitabhängigen Kovariablen auf die Hazardrate des Intervalls t durch $\tilde{z}(t)'\theta_t$ erfolgt. In diesem allgemeinen Modell ist auch der Gewichtsvektor θ_t zeitabhängig. Die Hazardrate im t–ten Intervall wird beeinflußt vom gesamten Kovariablenvektor der Vergangenheit, und der Einfluß ist für das t–te Intervall spezifisch.

2. Einen Spezialfall von (4.1) erhält man, wenn in θ_t die Spezifikation

$$\theta_{ts} = \theta_{-(t-s)} \tag{4-2}$$

getroffen wird, mit $\theta_0 := \theta$. Die Einwirkung der zeitabhängigen Kovariablen bis zum Zeitintervall t läßt sich dann durch

$$(x(t-r)', \ldots, x(t)') \begin{pmatrix} \theta_{-r} \\ \vdots \\ \theta_{-1} \\ \theta \end{pmatrix}$$

modellieren, wobei $\theta_{ts} = 0$ gesetzt wird für $t - s > r$. Dies entspricht der Modellierung von Time–Lags, deren Ordnung durch die Festlegung $\theta_{-s} = 0$ für $s > r$ zu fixem r vorgegeben werden kann.

Bei der Modellierung von Time–Lags bis zur ersten Ordnung erhält man den Einfluß der Kovariablen auf die Hazardrate des Zeitintervalls t in der Form

$$x_1'\gamma + x(t-1)'\theta_{-1} + x(t)'\theta \quad .$$

Der Einfluß der Kovariablenvektoren der Vergangenheit hängt jetzt nur noch vom Abstand zum gegenwärtig betrachteten Zeitpunkt ab. Der Gewichtsvektor hängt nicht mehr vom betrachteten Zeitpunkt t selbst ab.

3. Der für die Anwendung wichtige einfachste Fall ergibt sich durch die zusätzliche Voraussetzung

$$\theta_{-(t-s)} = 0 \quad \text{für } s = 1, \ldots, t-1 \tag{4-3}$$

Hier spielen nur noch die Kovariablenwerte des aktuellen Zeitintervalls eine Rolle, und ihr Einfluß erfolgt über $x(t)'\theta$.

Die unter 1. vorgestellte allgemeine Parametrisierung ist nur sinnvoll, wenn die Anzahl q der Zeitintervalle nicht zu groß ist, denn die Zahl der zu schätzenden Parameter wird mit zunehmendem q sprunghaft erhöht. Es sind dann riesige Datenmengen erforderlich, um die Parameter mit ausreichender Genauigkeit schätzen zu können. Bei großer Zahl der Intervalle ist es zweckmäßig, sich auf Modelle der Form (4.2) mit Time–Lags niederer Ordnung oder auf die Spezifikation (4.3) zu beschränken.

Die dargestellten Möglichkeiten der Einwirkung der zeitabhängigen Kovariablen sowie der Einfluß der zeitunabhängigen Kovariablen können in einheitlicher Form erfolgen. Dazu definieren wir den Gesamtparametervektor β durch

$$\beta = \begin{pmatrix} \gamma \\ \theta_1 \\ \vdots \\ \theta_q \end{pmatrix}$$

mit θ_t aus (4.1). Ferner werden allgemeine Designvektoren $z(t)$ definiert mit

$$z(t) = \begin{pmatrix} x_1 \\ 0 \\ \vdots \\ 0 \\ \tilde{z}(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Die Dimensionierung der Teilvektoren und Nullvektoren in $z(t)$ wird derart vorgenommen, daß bei der Skalarproduktbildung

$$z(t)'\beta = x_1'\gamma + \tilde{z}(t)'\theta_t \tag{4–4}$$

gilt, für $t = 1, \ldots, q$. Im folgenden gehen wir stets von der allgemeinen Darstellung (4.4) aus.

Die Definition der Hazardrate wird folgendermaßen erweitert

$$\lambda(t \mid z(t)) = P(T = t \mid T \geq t, z(t)) \tag{4–5}$$

$\lambda(t \mid z(t))$ ist die (bedingte) Wahrscheinlichkeit für einen Zustandswechsel bzw. Übergang im Zeitintervall t, gegeben das Intervall und der Pfad des Kovariablenvektors bis zum Zeitintervall t.

Zur Modellierung der Hazardrate in Abhängigkeit von den Kovariablen können die in Kapitel 3 beschriebenen Modelle entsprechend erweitert werden. Man erhält beispielsweise

1. Logistisches Modell

$$\lambda(t \mid z(t)) = \frac{exp(z(t)'\beta)}{1 + exp(z(t)'\beta)} \tag{4-6}$$

2. Gruppiertes Cox–Modell

$$\lambda(t \mid z(t)) = 1 - exp(-exp(z(t)'\beta)) \tag{4-7}$$

3. Modelle mit "Proportionalität" der diskreten Hazardraten

$$\lambda(t \mid z(t)) = exp(z(t)'\beta) \tag{4-8}$$

Modell (4.8) ist eine direkte Verallgemeinerung des Modells (3.14). Es ist jedoch zu beachten, daß in (4.8) die Proportionalität der diskreten Hazardraten im allgemeinen nicht mehr gilt, da die Kovariablen zeitabhängig sind. Darüber hinaus folgen auch hier aus der Restriktion $0 \leq \lambda(t \mid z(t)) \leq 1$ Einschränkungen für den zulässigen Bereich des Parameterraums.

Bei Einbeziehung von zeitabhängigen Kovariablen kann (4.7) nicht mehr als "gruppierte Version" des Cox–Modells mit stetig gemessener Zeit abgeleitet werden.

4.2 Beziehung zwischen Survivorfunktion und Hazardrate

Werden zeitabhängige Kovariablen im Modell berücksichtigt, können die in Kapitel 1 abgeleiteten Beziehungen zwischen Hazardrate, Survivorfunktion und der Wahrscheinlichkeitsverteilung von T nicht ohne weiteres übernommen werden. Zunächst werden die zu Beginn des Kapitels ausgeführten Definitionen für verschiedene Typen von zeitabhängigen Kovariablen formalisiert.

Sei T die Verweildauer mit $T \in \{1, \ldots, q\}$. Zusätzlich werden die Indikatorvariablen

$$T^{(s)} = \begin{cases} 1, & \text{falls im Zeitintervall } s \text{ ein Übergang stattfindet;} \\ 0 & \text{sonst,} \end{cases}$$

$s = 1, \ldots, q$, definiert. Dabei wird auf den Index i für das Individuum verzichtet.

Externe zeitabhängige Kovariablen werden definiert durch die Eigenschaft

$$\begin{aligned} &P(x(t+1), \ldots, x(q) \mid T^{(1)}, \ldots, T^{(t)}, x(1), \ldots, x(t)) = \\ &P(x(t+1), \ldots, x(q) \mid x(1), \ldots, x(t)), \end{aligned} \tag{4–9}$$

$t = 1, \ldots, q$. (4.9) beinhaltet, daß die zukünftige Entwicklung des Kovariablenpfades nicht von den Variablen $T^{(1)}, \ldots, T^{(t)}$, die die Verweildauer des Individuums bis zum Zeitintervall t festlegen, abhängt.

Bei *internen* zeitabhängigen Kovariablen wird der Kovariablenprozeß vom Individuum, dessen Verweildauer im in Frage stehenden Zustand untersucht wird, selbst generiert. Hier hängt es entscheidend davon ab, welche Verweildauern analysiert werden und wie die inhaltlichen Relationen zwischen Verweildauer und Kovariablen sind. Handelt es sich um Lebenszeitstudien, wie sie z.B. in der Medizin üblich sind, dann gilt

$$P(T \geq t \mid x(1), \ldots, x(t)) = 1\,, \tag{4–10}$$

wenn aus der Bedingung $x(1), \ldots, x(t)$ mit Sicherheit folgt, daß das Individuum das Intervall t erreicht hat. Gilt für interne zeitabhängige Kovariablen (4.10), nennen wir sie "strikt informativ". Handelt es sich jedoch bei Verweildauer und Kovariablen um zwei parallel verlaufende Prozesse wie z.B.

Dauer des gegenwärtigen Beschäftigungsverhältnisses und Heirat, so können sich diese Prozesse zwar gegenseitig beeinflussen, aber es gilt in der Regel nicht (4.10). Solche internen zeitabhängigen Kovariablen sind "nicht strikt informativ".

Es läßt sich formal eine "Survivorfunktion"

$$S(t \mid x(1)\ldots,x(t)) = P(T \geq t \mid x(1),\ldots,x(t)) \qquad (4\text{–}11)$$

definieren. (4.11) besitzt allerdings keine Survivor–Interpretation im Sinne von Kapitel 1 als unbedingte Wahrscheinlichkeit, das Intervall t "zu erleben", gegeben die Kovariablen. Bei strikt informativen Kovariablen wird dies unmittelbar aus (4.10) deutlich. Aber auch bei nicht informativen Kovariablen ist (4.11) keine Survivorfunktion im üblichen Sinne. Insbesondere müssen sich die Wahrscheinlichkeiten $P(T = t \mid x(1),\ldots,x(t))$ nicht zu 1 aufsummieren.

Es läßt sich allerdings wie in Kapitel 1 eine Beziehung zwischen Survivorfunktion und Hazardrate ableiten. Eine Umformung von (4.11) ergibt mit $\tilde{x}(t) = (x(1),\ldots,x(t))$

$$S(t \mid \tilde{x}(t)) = \prod_{s=1}^{t-1} P(T > s \mid T \geq s, \tilde{x}(t)). \qquad (4\text{–}12)$$

Die Faktoren auf der rechten Seite von (4.12) sind jedoch nicht ohne weiteres gleichzusetzen mit $1 - \lambda(s \mid \tilde{x}(s))$, da in (4.12) der Kovariablenpfad bis zum Zeitintervall t gegeben ist.

Es gilt allgemein für Ereignisse A_s, B_s, C_s

$$P(A_s \mid B_s \cap C_s) = \frac{P(C_s \mid A_s \cap B_s)P(A_s \mid B_s)}{P(C_s \mid B_s)}$$

und man setzt

$$A_s = \{T > s\}, \quad B_s = \{T \geq s, \tilde{x}(s)\}, \quad C_s = \{x(s+1),\ldots,x(t)\} \ .$$

Daraus folgt

$$P(T > s \mid T \geq s, \tilde{x}(t)) = P(T > s \mid T \geq s, \tilde{x}(s))\tilde{Q}_s \qquad (4\text{–}13)$$

mit

$$\tilde{Q}_s = \frac{P(x(s+1), \ldots, x(t) \mid T > s, \tilde{x}(s))}{P(x(s+1), \ldots, x(t) \mid T \geq s, \tilde{x}(s))}.$$

Bei den externen zeitabhängigen Kovariablen hängt die Verteilung $x(s+1)$, $\ldots, x(t)$ nicht von $T_1, \ldots, T_s$ ab, und man erhält $\tilde{Q}_s = 1$. Damit ergibt sich bei externen zeitabhängigen Kovariablen

$$S(t \mid \tilde{x}(t)) = \prod_{s=1}^{t-1} (1 - \lambda(s \mid \tilde{x}(s))) \tag{4-14}$$

in Analogie zu (1.19). Daraus folgt, daß bei externen zeitabhängigen Kovariablen bei der Bildung der Likelihoodfunktion wie bei den Modellen mit zeitunabhängigen Kovariablen verfahren werden kann. Bei internen zeitabhängigen Kovariablen, insbesondere bei strikt informativen Kovariablen, ist dies nicht möglich. Im nächsten Abschnitt wird eine Möglichkeit der Konstruktion der Likelihoodfunktion vorgestellt, die auch für diese Typen von Kovariablen geeignet ist.

4.3 Maximum–Likelihood–Schätzung

In diesem Abschnitt wird eine Vorgehensweise beschrieben, die für zeitabhängige Kovariablen anwendbar ist. Die Daten von Individuum i seien (t, δ_i) und $z_i(t_i)$ mit $z_i(t_i) = (x_i(1), \ldots, x_i(t_i))$. Die individuelle Zensierungszeit sei repräsentiert durch eine ganzzahlige Zufallsvariable $C_i \in \{1, \ldots, q\}$. Zur Konstruktion der Likelihoodfunktion berechnen wir die Wahrscheinlichkeiten für die beiden Ereignisse $\{t_i, \delta_i = 1, z_i(t_i)\}$ und $\{t_i, \delta_i = 0, z_i(t_i - 1)\}$. Hier ist bereits die Annahme enthalten, daß Zensierungen stets zu Beginn eines Zeitintervalls stattfinden. Dies bedeutet, daß die Kovariablen $z_i(t_i)$ nicht mehr erhoben werden, falls t_i ein zensierter Zeitpunkt ist.

Es ergibt sich

$$\begin{aligned} P(t_i, \delta_i = 1, z_i(t_i)) &= P(T_i = t_i, C_i > t_i, z_i(t_i)) \\ &= \prod_{s=1}^{t_i - 1} P(T_i > s, C_i > s, z_i(s) \mid H_{i,s-1}) \\ &\quad \cdot P(T = t_i, C_i > t_i, z_i(t_i) \mid H_{i,t_i-1}) \end{aligned} \tag{4-15}$$

mit $H_{i,s} = \{T_i > s, C_i > s, z_i(s)\}, \quad H_{i0} = \{T_i > 0, C_i > 0\}$.

(4.15) erhält man durch Anwendung der Formel

$$P(A_1 \cap \ldots \cap A_{t_i}) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \\ \cdots P(A_{t_i} \mid A_1 \cap \ldots \cap A_{t_i - 1})$$

mit $A_s = \{T_i > s, C_i > s, z_i(s)\}, \quad s = 1, \ldots, t_i - 1$ und $A_{t_i} = \{T_i = t_i, C_i > t_i, z_i(t_i)\}$.

Aus (4.15) resultiert mit

$$P(A \cap B \mid C) = P(A \mid B \cap C) P(B \mid C)$$

die Form

$$P(t_i, \delta_i = 1, z_i(t_i)) = \prod_{s=1}^{t_i - 1} P(T_i > s \mid T_i > s - 1, C_i > s, z_i(s), H_{i,s-1}) \\ \cdot P(z_i(s), C_i > s \mid H_{i,s-1}) P(T_i = t_i \mid T_i > t_i - 1, C_i > t_i, z_i(t_i), H_{i,t_i-1}) \\ \cdot P(z_i(t_i), C_i > t_i \mid H_{i,t_i-1}). \tag{4–16}$$

Dabei wurden die in $H_{i,s-1}$ enthaltenen Ereignisse $\{T_i > s - 1\}$ explizit aufgeführt, um den Zusammenhang mit der Hazardrate zu verdeutlichen.

Setzt man für den Zusammenhang zwischen Verweildauer und Zensierungsmechanismus voraus, daß gilt

$$P(T_i = s \mid T_i > s - 1, C_i > s, z_i(s)) = P(T_i = s \mid T_i > s - 1, z_i(s)) \\ = \lambda(s \mid z_i(s)),$$

dann folgt

$$P(t_i, \delta_i = 1, z_i(t_i)) = \lambda(t_i \mid z_i(t_i)) \prod_{s=1}^{t_i - 1} (1 - \lambda(s \mid z_i(s))) \\ \cdot \prod_{s=1}^{t_i} P(z_i(s), C_i > s \mid H_{i,s-1}) \tag{4–17}$$

Weiter erhält man unter der Annahme, daß die Zensierung jeweils zu Beginn des Intervalls stattfindet,

$$
\begin{aligned}
&P(t_i, \delta_i = 0, z_i(t_i - 1)) = P(T_i \geq t_i, C_i = t_i, z_i(t_i - 1)) \\
&= \prod_{s=1}^{t_i - 1} P(T_i > s \mid C_i > s, z_i(s), H_{i,s-1}) P(C_i > s, z_i(s) \mid H_{i,s-1}) \\
&\cdot P(C_i = t_i \mid H_{i,t_i-1}) \\
&= \prod_{s=1}^{t_i - 1} (1 - \lambda(s \mid z_i(s))) P(C_i > s, z_i(s) \mid H_{i,s-1}) \cdot P(C_i = t_i \mid H_{i,t_i-1})
\end{aligned}
\tag{4–18}
$$

(4.17) und (4.18) können zusammengefaßt werden zu

$$
P4(t_i, \delta_i, z_i(t_i)^{\delta_i} z_i(t_i - 1)^{1-\delta_i}) = \lambda(t_i \mid z_i(t_i))^{\delta_i} \prod_{s=1}^{t_i - 1} (1 - \lambda(s \mid z_i(s))) \cdot Q_i
\tag{4–19}
$$

mit

$$
\begin{aligned}
Q_i = &P(z_i(t_i), C_i > t_i \mid H_{i,t_i-1})^{\delta_i} P(C_i = t_i \mid H_{i,t_i-1})^{1-\delta_i} \\
&\cdot \prod_{s=1}^{t_i - 1} P(z_i(s), C_i > s \mid H_{i,s-1}) .
\end{aligned}
$$

(4.19) ist der Beitrag des i–ten Individuums zur Likelihoodfunktion. Die Herleitung dieses Beitrages erfordert keine speziellen Annahmen über die Kovariablenprozesse $z_i(t)$. Es kann sich dabei auch um vom Individuum selbst generierte Prozesse handeln, die parallel zur untersuchten Verweildauer beobachtet werden können. Dies schließt auch interne zeitabhängige Kovariablen im Sinne von Kalbfleisch/Prentice (1980), Kap. 5. 2, ein.

Die ersten beiden Faktoren auf der rechten Seite von (4.19) sind in völliger Analogie zu (3.21), allerdings ohne daß bei der Herleitung vom Konzept der Survivorfunktion Gebrauch gemacht wurde. Ist der letzte Faktor Q_i auf der rechten Seite von (4.19) "nicht informativ", d.h. er hängt nicht von den die Hazardrate determinierenden Parameter ab, kann dieser Faktor bei der Bildung der Likelihoodfunktion vernachlässigt werden. Andernfalls kann (4.19) unter Vernachlässigung des letzten Faktors als "Partial Likelihood" im Sinne von Cox (1975) verwendet werden.

Enthält $z_i(t)$ nur externe Kovariablen und sind Zensierungen ausgeschlossen,

dann ist mit (4.9) Q_i von der Form

$$Q_i = \prod_{s=1}^{t_i} P(z_i(s) \mid z_i(s-1))$$

und kann damit in der Likelihood vernachlässigt werden. Bei Vorliegen von Zensierungen allerdings sind dafür Zusatzannahmen über den Zensierungsmechanismus notwendig.

Im folgenden verwenden wir als Likelihoodfunktion unter Vernachlässigung von Q_i

$$L = \prod_{i=1}^{n} \lambda(t_i \mid z_i(t_i))^{\delta_i} \prod_{s=1}^{t_i-1} (1 - \lambda(s \mid z_i(s))) \quad . \tag{4–20}$$

4.4 Möglichkeiten zur Konstruktion von speziellen zeitabhängigen Kovariablen

In diesem Abschnitt werden für zeitabhängige Kovariablenprozesse einige Spezialfälle des allgemeinen Modellansatzes erörtert.

Verschiedene Zeitskalen für die Kovariablen

In der Definition der Hazardrate $\lambda(t \mid x(t))$ gilt der Zeitparameter t für alle Individuen. Dabei muß es sich nicht um die Kalenderzeit handeln. Im allgemeinen wird t die Zeit seit dem Eintreffen eines individuenspezifischen Ereignisses sein, gemessen in bestimmten Zeitintervallen, etwa Monaten oder Quartalen. Für die Kovariablenprozesse hingegen können verschiedenen Zeitskalen herangezogen werden. Dies soll durch einige Beispiele veranschaulicht werden.

Seien $t_{Geb(i)}$ und $t_{Ein(i)}$ die Zeitpunkte der Geburt und des Eintritts in den in Frage stehenden Zustand für das Individuum i. Bei vielen empirischen Studien ist letzterer Zeitpunkt identisch mit dem Eintreten des Individuums

i in die Studie. Es werden folgende zeitabhängige Kovariablen definiert:

$$x_{i1}(t) = 1$$
$$x_{i2}(t) = log(t - t_{Ein(i)} + 1)$$
$$x_{i3}(t) = t - t_{Geb(i)}$$

Dabei bedeutet t die Kalenderzeit (in diskreter Form als Zeitintervall, eventuell gemessen ab einem bestimmten "Nullpunkt"). $x_{i1}(t)$ und $x_{i2}(t)$ können als Teile einer "Grundhazardrate" aufgefaßt werden, ähnlich den Regressionsmodellen für Verweildauern bei stetig gemessener Zeit. $\beta_2 x_{i2}(t)$ stellt approximativ ein diskretes Analogon zur Weibull-Grundhazardrate dar. $\beta_3 x_{i3}(t)$ repräsentiert einen möglichen Alterseffekt. Die Zeitskala kann auch dahingehend abgeändert werden, daß z.B. bei der Kovariablen $x_{i3}(t)$ das Alter nicht in Monaten, sondern in Jahren gemessen wird, falls die Daten in Monaten vorliegen. Dies wird erreicht durch die Festsetzung

$$x_{i4}(t) = x_{i3}(t)/12.$$

Eine spezielle Variante der Verweildauerabhängigkeit wird von Mantel/Hankey (1978) betrachtet. Die Abhängigkeit von t wird durch ein Polynom $h(t) = \sum_{s=0}^{r} \gamma_s t^s$ vom Grade r modelliert. Die Koeffizienten γ_s des Polynoms werden dann als Parameter mitgeschätzt. Soll in $h(t)$ die Zeit seit Eintritt in die Studie eingehen, können zur Bildung des Polynoms die Kovariablen $x_{i0}(t) = 1$, $x_{is}(t) = (t - t_{Ein(i)})^s$ für $s = 1, \ldots, r$ herangezogen werden.

Intervallspezifische Kovariablen

Zur Einbeziehung von intervallspezifischen Kovariablen gibt es mehrere Möglichkeiten. Eine erste Variante besteht darin, für jedes Zeitintervall $t = 1, \ldots, T-1$ eine Dummy-Variable zu generieren, etwa

$$x_{i1}(t) = \begin{cases} 1, & \text{für } t = r; \\ 0, & \text{sonst,} \end{cases} \qquad r = 1, \ldots, T-1.$$

Auf diese Weise werden $T-1$ zusätzliche Kovariablen in den Regressionsansatz aufgenommen. Die dazugehörigen Parameter bilden wieder eine Art "Grundhazardrate", die allen Individuen gemeinsam ist. Die Parameter haben dieselbe Bedeutung wie die Parameter θ_t des logistischen oder des gruppierten Cox-Modells in Abschnitt 3.

Es ist zu beachten, daß die Zahl der zu schätzenden Modellparameter stark anwächst, wenn die Zahl der Zeitintervalle zunimmt, auch bei endlichem Beobachtungszeitraum. Eine Möglichkeit, die Zahl der die Grundhazardrate bildenden Parameter zu begrenzen, besteht darin, jeweils für mehrere aufeinanderfolgende Zeitintervalle den gleichen Parameter zu wählen. Liegen z.B. Monatsdaten vor, so können Dummy-Variablen

$$x_{i1}(t) = \begin{cases} 1, & \text{für } 1 \leq t \leq 12; \\ 0, & \text{sonst,} \end{cases}$$

$$x_{i2}(t) = \begin{cases} 1, & \text{für } 13 \leq t \leq 24; \\ 0, & \text{sonst, etc.} \end{cases}$$

eingerichtet werden, so daß die Parameter der Grundhazardrate nur von Jahr zu Jahr variieren. Selbstverständlich können auch andere Abstände gewählt werden.

Schließlich ist auch die Modellierung von Verschiebungen der Hazardrate ab einem gewissen Zeitpunkt möglich, die z.B. durch einen "Strukturbruch" zustande gekommen sind. Definiert man eine Dummy-Variable der Form

$$x_{i1}(t) = \begin{cases} 1, & \text{für } t \geq \tau; \\ 0, & \text{sonst,} \end{cases}$$

so mißt der dazugehörige Regressionskoeffizient eine mögliche Veränderung der bedingten Übergangswahrscheinlichkeiten ab dem Zeitpunkt τ.

Dummy-Variablen für Einflußfaktoren, die nur über einen gewissen Zeitraum wirksam sind

Eine spezielle Form von zeitabhängigen Kovariablen dient dazu, den Einfluß von Therapien, Programmen oder Strategien (z.B. Marketingstrategien) zu überprüfen, die nur während eines bestimmten Zeitraums durchgeführt werden. Ein weiteres Ziel kann darin bestehen, eventuelle Nachwirkungen der Therapie oder des Programms zu analysieren. Bezeichnet man Beginn und Ende der Therapie bzw. des Programms mit τ_1 und τ_2, so können folgende Dummy-Variablen eingeführt werden:

$$x_{i1}(t) = \begin{cases} 1, & \text{für } \tau_1 \leq t \leq \tau_2, \text{ falls Person } i \text{ an Therapie} \\ & \text{oder Programm teilgenommen hat;} \\ 0, & \text{sonst} \end{cases}$$

$$x_{i2}(t) = \begin{cases} 1, & \text{für } t > \tau_2, \text{ falls Person } i \text{ an Therapie} \\ & \text{oder Programm teilgenommen hat;} \\ 0, & \text{sonst.} \end{cases}$$

Sind die zugehörigen Regressionskoeffizienten signifikant negativ (positiv), so ist die Therapie bzw. das Programm effektiv und verringert (vergrößert) die bedingte Wahrscheinlichkeit für einen bevorstehenden Zustandswechsel. Ist darüber hinaus der erste Koeffizient absolut signifikant größer als der zweite, sinkt (steigt) der Effekt nach dem Absetzen der Therapie bzw. des Programms.

Eine Erweiterung des Konzeptes besteht darin, daß die Durchführung der Therapie oder des Programms von bestimmten Merkmalen aus der Vorgeschichte des Prozesses abhängig gemacht wird. Es liegen dann individuenspezifische Anfangs- und Endzeitpunkte τ_{i1} und τ_{i2} vor, die vom Eintreten bestimmter Ereignisse oder Ausprägungen von zeitlich variierenden Kovariablen abhängen.

Ein weiteres Anwendungsbeispiel für Kovariablen der eben beschriebenen Form ist die Untersuchung der Wirkung zeitlich befristeter betrieblicher Maßnahmen (z.B. Arbeitsbeschaffungsmaßnahmen) bei der Analyse der Dauer der Arbeitslosigkeit.

5. Exponentialmodelle mit konstantem Hazard in den Intervallen

Die in Kapitel 3 behandelten Modelle betrachten die abhängige Variable Verweildauer als rein kategorial, wobei $T = t$ das Eintreten des Zielereignisses im Intervall $[a_{t-1}, a_t)$ bezeichnet. Das hier behandelte Modell unterscheidet sich von dieser Modellklasse insofern, als die Verweildauer T hier als stetig betrachtet wird. Allerdings wird wieder die Intervalleinteilung $[a_0, a_1), \ldots, [a_q, a_{q+1})$ zugrundegelegt, wobei die Hazardrate $\lambda(t \mid x)$ in stetiger Zeit als konstant in diesen Intervallen vorausgesetzt wird. Obwohl dem Grunde nach stetig, wird das Modell hier betrachtet, einmal wegen der "Diskretisierung" in Intervalle und zum anderen wegen der Beziehung zu den in der Analyse kategorialer Daten weitverbreiteten loglinearen Modellen.

Der Einflußgrößenvektor x wird im folgenden immer als kategorial in einer entsprechenden Kodierung (siehe Abschnitt 1.2) vorausgesetzt. $\lambda(t \mid x)$ bezeichnet die Hazardrate zur stetigen Verweildauer T.

5.1 Modelldarstellung

Für stetig beobachtete Zeit t wird häufig das in Kapitel 1 kurz besprochene Cox–Modell

$$\lambda(t \mid x) = \lambda_0(t) exp(x'\beta)$$

zugrundegelegt. $\lambda_0(t)$ ist in diesem Modell die "Baseline"–Hazardrate, die unabhängig vom Kovariablenvektor x ist, und $\lambda(t \mid x)$ bezeichnet die stetige Hazardrate. Eine spezielle Version dieses Modells ergibt sich aus der Annahme, daß die "Baseline"–Hazardrate innerhalb jeder der vorgegebenen Intervalle $[a_0, a_1), \ldots, [a_q, \infty)$ konstant ist, d.h. es gilt

$$\lambda_0(t) = \lambda_i \qquad \text{für } t \in [a_{i-1}, a_i) \quad , \tag{5-1}$$

wobei $a_{q+1} = \infty$. Man erhält das *Exponentialmodell mit konstantem Hazard in den Intervallen* durch

$$ln\lambda(t \mid x) = ln\lambda_i + x'\beta \tag{5-2}$$

für $t \in [a_{i-1}, a_i), \quad i \leq q$. Sind die Einflußgrößen $x_1, \ldots, x_p$ kategorial, so ist (5.2) in dieser Form nur sinnvoll, wenn die Komponenten von x dichotom

sind. Besitzen die Komponenten x_i mehr als zwei Ausprägungen, benutzt man, wie in Abschnitt 1.2 dargestellt, Dummy–Variablen. Da loglineare Modelle für kategoriale Variablen meist anders formalisiert werden, wird der Zusammenhang der verschiedenen Darstellungen im folgenden kurz skizziert. Die Darstellung im folgenden orientiert sich zuerst am Fall kategorialer Einflußgrößen, da dafür Programme der loglinearen Kontingenztafelanalyse anwendbar sind (Laird& Olivier, 1981). Der allgemeinere Fall mit möglichen metrischen Einflußgrößen wird danach behandelt.

Sei $x' = (x_1, \ldots x_p)$ der Vektor der interessierenden ursprünglichen Variablen mit den möglichen Ausprägungen $\{1, \ldots, m_i\}$ für x_i. Zur Komponente x_i erhält man $m_i - 1$ Dummy–Variablen in Effektkodierung durch

$$x_k^{(i)} = \begin{cases} 1, & x_i = k; \\ -1, & x_i = m_i; \\ 0, & \text{sonst.} \end{cases}$$

Dem Vektor $x' = (x_1, \ldots, x_p)$ entspricht dann ein Vektor von Dummy-Variablen

$$\tilde{x} = (x_1^{(1)}, \ldots, x_{m_1-1}^{(1)}, x_1^{(2)}, \ldots, x_{m_2-1}^{(2)}, \ldots, x_1^{(1)} \cdot x_1^{(2)}, \ldots),$$

wobei als Komponenten in $\tilde{x}$ auch Produkte von Dummy–Variablen zur Kodierung von Interaktionswirkungen auftreten können.

Modell (5.2) ist dann für $t \in [a_{i-1}, a_i)$ von der Form

$$ln\lambda(t \mid \tilde{x}) = ln\lambda_i + \tilde{x}'\beta \quad , \tag{5–3}$$

wobei der Vektor der Dummy–Variablen als Kovariablenvektor eingeht. Der Einflußvektor $\tilde{x}$ enthält kodierte Dummy–Variablen und Produkte davon. Unter Verwendung der Effektkodierung läßt sich der Term $\tilde{x}'\beta$ in der bei loglinearen Modellen üblichen Form darstellen durch

$$u_0 + u_{1(x_1)} + u_{1(x_2)} + \ldots + u_{12(x_1x_2)} + \ldots + u_{1\ldots p(x_1,\ldots,x_p)},$$

wobei x_i die möglichen Ausprägungen $x_i \in \{1, \ldots, m_i\}$ besitze. Analog zu den loglinearen Modellen bezeichnet man $u_1, \ldots, u_p$ als Haupteffekte, $u_{12}, u_{13}, \ldots, u_{p-1,p}$ als 2–Faktoren–Interaktionen und entsprechend Terme, die k Variablen verknüpfen wie $u_{12\ldots k}$, als k–Faktor–Interaktion.

Das Modell (5.3) ist damit für $t \in [a_{i-1}, a_i)$ darstellbar durch

$$ln\lambda(t \mid x) = u + u_{0(i)} + u_{1(x_1)} + \ldots + u_{1\ldots p(x_1\ldots x_p)} \qquad (5\text{–}4)$$

bzw. durch

$$ln\lambda(t \mid x) = u_{0(i)} + \sum_{j_s \in \{1,\ldots,m_{i_s}\}} x_{j_1}^{(i_1)} \ldots x_{j_r}^{(i_r)} u_{i_1\ldots i_r(j_1,\ldots,j_r)} = u_{0(i)} + \tilde{x}'\beta \quad ,$$

wobei der ursprüngliche Variablenvektor x als Einflußvektor in $\lambda(t \mid x)$ eingeht.

Es gilt $u_{0(i)} = ln\lambda_i$ und die für loglinearen Modelle üblichen Nebenbedingungen

$$\sum_{j_s=1}^{m_{i_s}} u_{i_1\ldots i_r(j_1,\ldots,j_r)} = 0$$

für $i_1, \ldots, i_r \in \{1, \ldots, p\}$ müssen erfüllt sein. Der Parametervektor β und die u–Parameter entsprechen einander. Beispielsweise entspricht $u_{1(x_1)}$ der Komponente in β, die mit $\tilde{x}_{x_1}^{(1)}$ multiplikativ verknüpft ist. Zu loglinearen Modellen und deren Darstellung vergleiche man Hamerle/Tutz (1984).

Modell (5.4) ist ein saturiertes Modell, was den Einflußvektor x betrifft, da sämtliche Interaktionsterme bis zur p–Faktor–Interaktion erhalten sind. Vereinfachte Modelle ergeben sich, wenn nur Interaktionsterme niedrigerer Ordnung oder nur Haupteffekte zugelassen werden.

Als Cox–Modell mit speziell gewählter "Baseline"–Hazardrate besitzt Modell (5.3) wiederum die Eigenschaft proportionaler Hazardraten. Es gilt für zwei Einflußvektoren x, y

$$ln\frac{\lambda(t \mid x)}{\lambda(t \mid y)} = u_{1(x_1)} - u_{1(y_1)} + \ldots + u_{1\ldots p(x_1\ldots x_p)} - u_{1\ldots p(y_1\ldots y_p)}. \qquad (5\text{–}5)$$

Das Verhältnis der Hazardraten zweier Einflußvektoren hängt somit nicht von der Zeit t ab.

Modell mit Interaktionen zwischen Kovariablen und Intervall

Eine naheliegende Verallgemeinerung des Modells (5.4) erhält man, wenn die Zeit als eigener Faktor mit möglichen Interaktionen mit den Kovariablen in

das Modell aufgenommen wird. Eine derartige Verallgemeinerung, die das formale Konzept unverändert läßt und damit nach demselben Schätzprinzip abläuft, erhält man mit

$$ln\lambda(t \mid x) = u + u_{0(i)} + \ldots + u_{p(x_p)} + u_{01(ix_1)} + \ldots + u_{01\ldots p(ix_1\ldots x_p)} \quad (5\text{–}6)$$

für $t \in [a_{i-1}, a_i)$, wobei die üblichen Nebenbedingungen unterstellt werden. Damit sind Interaktionen zwischen der Zeit t und den Kovariablenvektoren zugelassen. Die Proportionalität in (5.5) gilt nicht mehr. Modell (5.6) ist saturiert, d.h. jede in den Intervallen konstante Hazardrate $\lambda(t \mid x)$ läßt sich damit modellieren. Ausgehend von diesem saturierten Modell sind einfachere nicht–saturierte Modelle von Interesse. Ein erstes nicht–saturiertes Modell ist Modell (5.4), das dadurch charakterisiert ist, daß die Interaktionen $u_{0i_1\ldots i_r}$ der Zeit mit Kovariablen verschwinden. Das saturierte Modell (5.6) läßt sich so, abhängend von der Empirie, vereinfachen. Zum Beispiel läßt sich die Proportionalität (5.5) durch die Relevanz der Interaktionen zwischen Zeit und Kovariablen testen, da (5.5) nur gilt, wenn diese Interaktionen verschwinden.

Die sich aus dem saturierten Modell (5.6) ergebende Modellhierarchie ergibt sich analog zu den loglinearen Modellen. Hierarchische loglineare Modelle lassen sich eindeutig charakterisieren durch sämtliche Effekte, die keine Marginaleffekte eines anderen Effekts sind. Diese maximalen Effekte werden zur Abkürzung des Modells verwendet, indem die maximalen Effekte durch die Variablenindices angegeben werden. Querstriche trennen die maximalen Effekte voneinander. Entsprechend wird für die Verweildauer x_0 und die Variablen x_1, x_2, x_3 das saturierte Modell (5.6) durch 0123 abgekürzt, das Modell mit nur Ein–Faktor–Effekten u_i, $i = 0,1,2,3$ durch 0/1/2/3 (vgl. Hamerle/Tutz (1984), S. 505). Das Modell (5.5) mit proportionalem Hazard wird in dieser Schreibweise für drei Variablen durch 0/123 abgekürzt.

Die Modelldarstellung (5.6) geht aus von der üblichen Darstellung der loglinearen Modelle. Die alternative Darstellung mit Dummy–Variablen empfiehlt sich insbesondere bei der Ableitung der Likelihood. Als zusätzliche "Zeit"–Variable sei x_0 eingeführt mit $x_0 = i$, wenn $T \in [a_{i-1}, a_i)$. Modell (5.6) läßt sich dann mit Hilfe der Effekt–Kodierung $\tilde{x}_i^{(0)}$ zur kategorialen Variablen x_0 darstellen durch

$$\begin{aligned} ln\lambda(t \mid x) &= u + x_1^{(0)} u_{0(1)} + \ldots + x_1^{(1)} u_{1(1)} + \ldots \\ &= \sum x_{i_0}^{(0)} x_{j_1}^{(i_1)} \cdots x_{j_r}^{(i_r)} u_{01\ldots r(j_1\ldots j_r)} \end{aligned} \quad (5\text{–}7)$$

für $t \in [a_{i-1}, a_i)$ und damit $x_0 = i$.

Modell (5.7) läßt sich partitionieren durch

$$ln\lambda(t \mid x) = \gamma_i(x) + \tilde{x}'\beta_1 \quad ,$$

wobei in $\gamma_i(x)$ alle Terme zusammengefaßt werden, die von i und damit von der Zeit abhängen. Darin sind auch Interaktionen zwischen Zeit und Kovariablenvektor eingeschlossen. Man beachte, daß $\gamma_i(x)$ wiederum ein linearer Term ist mit

$$\gamma_i(x) = (x_1^{(0)}, \ldots, x_q^{(0)}, x_1^{(0)}x_1^{(1)}, \ldots)'\beta_0 = \tilde{x}_0'\beta_0 \quad ,$$

wobei $x_j^{(0)} = 0$ ist für $j \neq i$ und $x_i^{(0)} = 1$. Liegen beispielsweise keine Interaktionen vor, ist $\tilde{x}_0 = (0, \ldots, 1, \ldots, 0)'$ der i–te Einheitsvektor, wenn $t \in [a_{i-1}, a_i)$. Man erhält damit die allgemeine Modellform

$$ln\lambda(t \mid x) = \tilde{x}_0'\beta_0 + \tilde{x}\beta_1 \quad , \tag{5–8}$$

wobei $\tilde{x}_0$ die Kodierung der Intervalle und die Interaktion zwischen Intervallen und dem Kovariablenvektor x enthält, $\tilde{x}$ eine reine Kodierung des Kovariablenvektors darstellt. Die Interaktion zwischen Intervallen und Kovariablen entspricht der Zeitabhängigkeit der Einflußgrößen.

Im allgemeinen Modell (5.8) ist die Beschränkung auf kategoriale Variablen aufgehoben. Im Vektor $\tilde{x}$ können ebenso metrische Komponenten enthalten sein. Liegt z.B. nur der metrische Kovariablenvektor x vor und keine Zeitabhängigkeit der Einflußgrößenwirkung, erhält man für $t \in [a_{i-1}, a_i)$ das einfache Modell

$$ln\lambda(t \mid x) = \gamma_i + x'\beta_1$$

bzw.

$$ln\lambda(t \mid x) = (x_1^{(0)}, \ldots, x_q^{(0)})'\beta_0 + x'\beta_1.$$

Die zeitabhängige Wirkung wird deutlich an einem einfachen Modell mit skalarem Einfluß x und ausschließlich Interaktionswirkungen. Gelte für $t \in [a_{i-1}, a_i)$

$$\begin{aligned} ln\lambda(t \mid x) &= (x_1^{(0)}, \ldots, x_q^{(0)}, x_1^{(0)}xx_q^{(0)}x, \ldots x_q^{(0)}x)\beta_0 \\ &= \gamma_i + x\beta_{0i} \end{aligned}$$

bzw.

$$ln\lambda(t \mid x) = (x_1^{(0)}, \ldots, x_q^{(0)}, x_1^{(0)}x, \ldots, x_q^{(0)}x)\beta_0,$$

wobei $\beta_0 = (\gamma_1, \ldots, \gamma_q, \beta_{01}, \ldots, \beta_{0q})$.

Hier variieren die Gewichte auf der Einflußgröße x von Intervall zu Intervall. Zu beachten ist, daß zwar die Wirkung der Kovariablen zeitabhängig ist, nicht aber die Kovariablen selbst. Zeitabhängige Kovariablen, die über die Intervalle hinweg als konstant anzunehmen sind, betrachtet Clayton (1987) in einer Übersicht über Verweildauermodelle.

5.2 Maximum–Likelihood–Schätzung

Ein Vorteil des Modells mit konstantem Hazard in den Intervallen besteht darin, daß Programme für loglineare Modelle zur Bestimmung der Maximum-Likelihood–Schätzungen benützt werden können. Die Nutzbarmachung dieses Vorteils geht vor allem auf Holford (1980) zurück. Weitere Ausführungen finden sich bei Holford (1976), Laird & Olivier (1981) und Friedman (1982).

Bezeichne im weiteren $\Delta_i = a_i - a_{i-1}, \quad i = 1, \ldots, q$, die Länge des i–ten Intervalls. Zur Verweildauer t_j des Individuums j bezeichne

$$t_{ij} = min(0, t_j - a_{i-1}, a_i - a_{i-1})$$

die stetige Zeit, die für das Individuum j im Intervall $[a_{i-1}, a_i)$ beobachtet wurde. δ_{ij} sei eine Indikatorvariable zum i–ten Intervall und j–ten Individuum mit $\delta_{ij} = 1$, wenn für das Individuum der Zielzustand im i–ten Intervall eintritt, und $\delta_{ij} = 0$, wenn das Individuum nicht weiter beobachtet wird bzw. der Endzustand in diesem Intervall nicht eintritt. Man erhält mit $k = q+1$ zu jedem Individuum j die Vektoren $t_{j\cdot} = (t_{1j}, \ldots, t_{kj})$ und $\delta_{j\cdot} = (\delta_{1j}, \ldots, \delta_{kj})$, wobei in δ_j höchstens eine Komponente 1, alle anderen gleich 0 sind und in $t_{j\cdot}$ immer dann das gesamte Intervall Δ_i als Komponente erscheint, wenn von dem Individuum bekannt ist, daß es das Intervall $[a_i, a_{i+1})$ erreicht hat.

Für die stetige Survivorfunktion und die Dichte des in den Intervallen konstanten Hazardmodells ohne Berücksichtigung von Kovariablen erhält man für $t \in [a_{i-1}, a_i)$

$$S(t) = exp(-\int_0^t \lambda(u)\, du) = exp(-\sum_{r=1}^{i-1} \lambda_r \Delta_r - \lambda_i(t - a_{i-1}))$$

und

$$f(t) = \lambda_i S(t) \quad .$$

Die Likelihood für Individuum j mit $t_j \in [a_{i-1}, a_i)$ erhält man mit

$$\begin{aligned} L_j = f(t_j)^{\delta_{ij}} S(t_j)^{1-\delta_{ij}} &= \lambda_i^{\delta_{ij}} S(t_j) \\ &= \lambda_i^{\delta_{ij}} exp(-\sum_{r=1}^{i-1} \lambda_r \Delta_r) exp(-\lambda_i (t_j - a_{i-1})) \end{aligned}$$

Für beliebiges t_j erhält man mit den Komponenten der Vektoren $t_{j\cdot}, \delta_{j\cdot}$

$$L_j = \prod_{i=1}^{q} \lambda_i^{\delta_{ij}} exp(-\lambda_i t_{ij}) \quad .$$

Für die Gesamtlikelihood ergibt sich

$$L = \prod_j L_j = \prod_{i=1}^{q} \lambda_i^{d_i} exp(-\lambda_i T_i) \quad ,$$

wobei

$$d_i = \sum_j^n \delta_{ij}$$

die Anzahl der im Intervall $[a_{i-1}, a_i)$ zu Ende beobachteten Individuen darstellt und

$$T_i = \sum_j^n t_{ij}$$

die Zeitdauer der gesamten Stichprobe im Intervall $[a_{i-1}, a_i)$ bezeichnet.

Unter Einbeziehung von Kovariablen ist Modell (5.7) von der Form

$$ln(\lambda(t \mid x_j)) = \gamma_i(x_j) + \tilde{x}_j' \beta_1 \qquad \text{für } t \in [a_{i-1}, a_i)$$

bzw. mit $\lambda_i(x_j) = exp(\gamma_i(x_j))$ von der Form

$$\lambda(t \mid x_j) = \lambda_i(x_j) exp(\tilde{x}_j' \beta_1) \qquad \text{für } t \in [a_{i-1}, a_j) \,.$$

Diejenigen Terme in der Modelldarstellung (5.6), die vom Intervall und damit von i abhängen, wurden in $\gamma_i(x_j)$ zusammengefaßt.

Für nichtinformativen Zensierungsmechanismus erhält man ganz entsprechend die Likelihood für Individuum j durch

$$L_j = \lambda_i(x_j)^{\delta_{ij}} exp(\tilde{x}_j'\beta_1\delta_{ij}) exp(exp(\tilde{x}_j'\beta_1) \cdot (-\sum_{r=1}^{i-1} \lambda_r(x_j)\Delta_r - \lambda_i(x_j)(t_j - a_{i-1})))$$

für $t_j \in [a_{i-1}, a_i)$ bzw. allgemeiner durch

$$L_j = \prod_i \lambda_i(x_j)^{\delta_{ij}} exp(\tilde{x}_j'\beta_1\delta_{ij}) exp(exp(\tilde{x}_j'\beta_1)(-\lambda_i(x_j)t_{ij}))\,.$$

Daraus erhält man die Likelihood

$$L = \prod_j L_j = \prod_i \prod_j \lambda_i(x_j)^{\delta_{ij}} exp(\tilde{x}_j'\beta_1\delta_{ij}) exp(exp(\tilde{x}_j'\beta_1)(-\lambda_i(x_j)t_{ij})). \tag{5-9}$$

Mit den Abkürzungen

$$d_{i,x} = \sum_{\substack{j \\ x_j = x}} \delta_{ij}$$

für die Anzahl aller im Intervall $[a_{i-1}, a_i)$ zu Ende beobachteten Individuen mit Merkmalsvektor x, und

$$T_{i,x} = \sum_{\substack{j \\ x_j = x}} t_{ij}$$

für die Zeitdauer aller Individuen mit Merkmalsvektor x im Intervall $[a_{i-1}, a_i)$ läßt sich die Gesamt–Likelihood darstellen durch

$$\begin{aligned} L &= \prod_{i=1}^{I} \lambda_i(x)^{\sum_x d_{i,x}} exp(\sum_x \tilde{x}'\beta_1 d_{i,x}) exp(-\sum_x \lambda_i(x) T_{i,x} exp(\tilde{x}'\beta_1)) \\ &= \prod_{i=1}^{I} \prod_x \lambda_i(x)^{d_{i,x}} exp(\tilde{x}'\beta_1 d_{i,x}) exp(-\lambda_i(x) T_{i,x} exp(\tilde{x}'\beta_1)), \end{aligned}$$

wobei $\sum_x$ bzw. $\prod_x$ über alle möglichen Ausprägungen des diskreten Vektors x laufen.

Mit $m_{i,x} := T_{i,x}\lambda_i(x)exp(\tilde{x}'\beta_1)$ erhält man die Vereinfachung

$$L = \prod_{i=1}^{I}\prod_{x}\frac{1}{T_{i,x}^{d_{i,x}}} \quad m_{i,x}^{d_{i,x}} \; e^{-m_{i,x}} \quad . \tag{5-10}$$

Da der erste Term in (5.10) keine Parameter, sondern nur Beobachtungen enthält, genügt es, die eingeschränkte Likelihood

$$L_r = \prod_{i=1}^{I}\prod_{x} m_{i,x}^{d_{i,x}} \; e^{-m_{i,x}} \tag{5-11}$$

zu betrachten.

Äquivalenz der Likelihood zu einem Poissonmodell

Die Likelihood läßt sich im Rahmen der loglinearen Modelle maximieren, wie sich aus der Betrachtung des folgenden Modells ergibt: Sei $D_{i,x}$ unabhängig poissonverteilt mit Parameter $\mu_{i,x}T_{i,x}$ für bekanntes $T_{i,x}$, es gilt also $E(D_{i,x} \mid T_{i,x}) = \mu_{i,x}T_{i,x}$. Dann erhält man für die entsprechende Likelihood mit den Realisationen $d_{i,x}$

$$\begin{aligned}\tilde{L} &= \prod_{i}^{I}\prod_{x} e^{-\mu_{i,x}T_{i,x}}\frac{(\mu_{i,x}T_{i,x})^{d_{i,x}}}{d_{i,x}!} \\ &= \prod_{i}^{I}\prod_{x}\frac{1}{d_{i,x}!}(\mu_{i,x}T_{i,x})^{d_{i,x}} \; e^{-\mu_{i,x}T_{i,x}}\end{aligned}$$

Es genügt wiederum, die reduzierte Likelihoodfunktion

$$\tilde{L}_r = \prod_{i}\prod_{x}(\mu_{i,x}T_{i,x})^{d_{i,x}} \; e^{-\mu_{i,x}T_{i,x}} \tag{5-12}$$

zu betrachten.

Aus der formalen Ähnlichkeit zwischen (5.11) und (5.12) ergibt sich, daß L_r äquivalent ist zur reduzierten Likelihood, wie sie in der Analyse von Poisson-Regressionsmodellen auftritt. Damit sind Programme für die Poisson-Regression, wie z.B. GLAMOUR, anwendbar. Im Spezialfall rein kategorialer Einflußgrößen lassen sich auch Programme zur Kontingenztafelanalyse mit poissonverteilten Größen in den einzelnen Zellen anwenden.

Beim Vergleich von L_r und $\tilde{L}_r$ entsprechen sich $m_{i,x}$ und $\mu_{i,x}T_{i,x}$. $m_{i,x}$ ist spezifiziert durch das Modell

$$m_{i,x} = T_{i,x}\lambda_i(x)exp(\tilde{x}'\beta_1) = T_{i,x}exp(\gamma_i(x) + \tilde{x}'\beta_1).$$

Da $\gamma_i(x)$ wiederum ein linearer Term ist, läßt sich im Modell (5.8) setzen

$$\gamma_i(x) + \tilde{x}'\beta_1 = \tilde{x}_0'\beta_0 + \tilde{x}'\beta_1 = (\tilde{x}_0', \tilde{x}')\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} \quad ,$$

wobei $\tilde{x}_0$ die entsprechende Kodierung der Zeitterme bzw. der Interaktion zwischen Zeit und Kovariablen enthält. Man erhält mit $\bar{\beta}' = (\beta_0', \beta_1')$ und $\bar{x}' = (\tilde{x}_0', \tilde{x}')$

$$m_{i,x} = T_{i,x}\ exp(\bar{x}'\bar{\beta}) \quad . \tag{5-13}$$

Wählt man das loglineare Poisson–Modell

$$E(D_{i,x} \mid T_{i,x}) = exp(lnT_{i,x} + \bar{x}'\bar{\beta}) = T_{i,x}exp(\bar{x}'\bar{\beta}) \tag{5-14}$$

für gegebene $T_{i,x}$, so erhält man, da sich die Likelihoodfunktionen der beiden Modelle entsprechen, die ML–Schätzung für β durch Maximieren des Modells (5.14).

Bedingungen für die Existenz und Eindeutigkeit der Maximum–Likelihood–Schätzungen findet man bei Friedman (1982). Darüber hinaus betrachtet Friedman die Eigenschaften der Schätzungen, wenn die tatsächliche Hazardrate nicht konstant in den Intervallen ist und die Intervallänge gegen Null konvergiert, wobei die Anzahl der Intervalle wächst. Diese Vorstellung entspricht Breslows (1974) Ansatz, der unter bestimmten Bedingungen die Äquivalenz zur Parameterschätzung im stetigen Cox–Modell zeigt. Die Verwendung von Likelihood–Ratio–Tests im nächsten Abschnitt basiert auf der Proportionalität zum Poisson–Erhebungsschema (vergleiche Laird/Olivier, 1981).

5.3 Anwendungsbeispiel

In einer Studie zur Arbeitslosigkeit wurden zu 7168 Fällen folgende Variablen erhoben:

		Ausprägungen	Kategorie
x_0	Dauer der Arbeits-	0–0.5 Jahre	1
	losigkeit	0.5–1 Jahre	2
		1–1.5 Jahre	3
		1.5–2 Jahre	4
		2 Jahre	5
x_1	Alter	0–30 Jahre	1
		30–50 Jahre	2
		50 Jahre	3
x_2	Geschlecht	männlich	1
		weiblich	2
x_3	Gesundheitliche	nein	1
	Einschränkungen	ja	2
x_4	Beendigung	selbst	1
	des letzten	durch Arbeitgeber	2
	Arbeitsver-	oder im gegens.	
	hältnisses	Einvernehmen	

Als erstes wurde die Gültigkeit des Modells 0/1234 der Proportionalität der Hazardraten überprüft. Mit einem Loglikelihood-Verhältnis vom Wert 162.67 bei 92 Freiheitsgraden wird das Modell deutlich abgelehnt (vergleiche Meindl (1985)).

Zur Modellauswahl wurden im ersten Schritt die Modelle der Ordnung k betrachtet. Ein Modell M_k heißt dabei von der Ordnung k, wenn sämtliche k-Faktor-Interaktionen enthalten sind. Man erhält damit für die Log-Likelihood-Statistik folgende Werte:

	$lq(M_k)$	FG	kritische W'keit
M_1	256.20	110	0.0
M_2	80.3	81	0.225

Da M_2 den Daten relativ gut angepaßt ist, wurden von M_2 ausgehend die Relevanz der implizierten Effekte untersucht. Bezeichne $M_2 \setminus \{i,r\}$ das Modell 2. Ordnung ohne die Interaktion zwischen X_i und X_r. Angegeben wird jeweils die Loglikelihood $lq(M_2 \setminus \{i,r\})$ und das Verhältnis $lq(M_2 \setminus \{i,r\} \mid M_2)$. Ausgehend von Modell M_2 erweisen sich die Interaktionen u_{02}, u_{04}, u_{34} als unverzichtbar. Das reduzierte Modell 02/04/34/1 mit den maximalen Parametern $u_1, u_{02}, u_{04}, u_{34}$ erweist sich mit $lq = 112.69$ bei 101 Freiheitsgraden mit resultierender Restwahrscheinlichkeit von 0.201 als gut angepaßt.

Modell	$lq(M_2 \setminus \{\})$	FG	krit. Wahrscheinlichkeit	$lq(M_2 \setminus \{\} \mid M_2)$	FG	krit. Wahrscheinlichkeit
$M_2 \setminus \{0,1\}$	102,71	89	0,152	12,41	8	0,134
$M_2 \setminus \{0,2\}$	133,34	85	0,001	43,04	4	0,0
$M_2 \setminus \{0,3\}$	93,31	85	0,252	3,01	4	0,556
$M_2 \setminus \{0,4\}$	131,94	85	0,001	41,64	4	0,0
$M_2 \setminus \{1,2\}$	90,47	83	0,269	0,17	2	0,919
$M_2 \setminus \{1,3\}$	90,66	83	0,265	0,36	2	0,835
$M_2 \setminus \{1,4\}$	90,41	83	0,271	0,11	2	0,946
$M_2 \setminus \{2,3\}$	91,19	82	0,228	0,89	1	0,345
$M_2 \setminus \{2,4\}$	92,93	82	0,192	2,63	1	0,105
$M_2 \setminus \{3,4\}$	97,71	82	0,114	7,41	1	0,006

6. Competing-Risks-Modelle

Competing-Risks-Ansätze wurden bisher nahezu ausschließlich für den Fall stetig gemessener Zeit untersucht. Man vergleiche beispielsweise David/ Moeschberger (1978), Seal (1977), Holt (1978), Prentice/Breslow (1978) oder Prentice et al. (1978). Für diskrete Hazardraten-Modelle für konkurrierende Risiken siehe Hamerle (1985).

6.1 Parametrisierung der ursachenspezifischen Hazardrate

Neben der Verweildauer T wird eine Zustandsvariable Y beobachtet, die Werte aus der Menge der möglichen Zielzustände $\{1, \ldots, m\}$ annehmen kann. In der Analyse von Lebenszeiten in der Medizin werden die Zielzustände als "competing risks" interpretiert und gewöhnlich durch verschiedene Todesursachen repräsentiert; in der Technik können es verschiedenartige Defekte sein, die den Ausfall eines Gerätes bewirken. Die Anzahl der konkurrierenden Risiken (Zielzustände, Ereignisarten) sei m.

Die Kovariablen x werden der einfacheren Handhabung wegen als zeitunabhängig vorausgesetzt. Die in Kapitel 4 beschriebene Vorgehensweise, insbesondere zur Konstruktion der Likelihoodfunktion, kann jedoch auch auf Competing-Risks-Modelle übertragen werden.

Ein geeignetes Instrument zur Analyse der Relationen zwischen den Kovariablen und den verschiedenen Übergängen in die Zielzustände ist die Hazardrate. Eine übergangsspezifische (ursachenspezifische) Hazardrate läßt sich wie folgt definieren (gegeben die Kovariablen):

$$\lambda_j(t \mid x) = P(T = t, Y = j \mid T \geq t, x) \quad . \tag{6-1}$$

(6.1) ist die bedingte Wahrscheinlichkeit dafür, daß ein Individuum im Zeitintervall t in den Zustand j wechselt, gegeben die Kovariablen, und daß das Individuum den Beginn des Zeitintervalls t erreicht hat.

Die Gesamthazardrate

$$\lambda(t \mid x) = \sum_{j=1}^{m} \lambda_j(t \mid x) \tag{6-2}$$

ist die (bedingte) Wahrscheinlichkeit dafür, daß im Zeitintervall t ein Übergang bzw. Zustandswechsel stattfindet, gegeben die Kovariablen, und daß bis zum Beginn dieses Zeitintervalls noch kein Übergang stattgefunden hat.

Die bedingte Wahrscheinlichkeit, das Zeitintervall t "zu überleben", ist dann

$$1 - \lambda(t \mid x) = 1 - \sum_{j=1}^{m} \lambda_j(t \mid x) \quad . \tag{6-3}$$

Für jedes Zeitintervall gibt es demnach $m + 1$ Kategorien, nämlich Wechsel in den Zustand j, $j = 1, \ldots, m$, bzw. kein Übergang im Zeitintervall t. Ist ein Individuum bis zum Beginn des Zeitintervalls t gelangt, so sind die dazugehörigen Wahrscheinlichkeiten gerade $\lambda_1(t \mid x), \ldots, \lambda_m(t \mid x)$ bzw. $1 - \lambda(t \mid x)$.

Eine Möglichkeit für einen Regressionsansatz besteht darin, die übergangsspezifischen Hazardraten in Abhängigkeit von den Kovariablen zu parametrisieren, etwa in der Form

$$\lambda_j(t \mid x) = g_j(\beta_{0t} + x'\beta_1, \ldots, \beta_{0t} + x'\beta_m), \qquad \begin{matrix} t = 1, \ldots, q; \\ j = 1, \ldots, m \end{matrix} \tag{6-4}$$

mit $g_j(\cdot) \in (0, 1)$. β_{0t} bringt dabei den Beitrag einer "Baseline"-Hazardrate ohne Berücksichtigung der Kovariablen zum Ausdruck. Ein Beispiel ist der multivariate logistische Ansatz mit den Hazardraten

$$\lambda_j(t \mid x) = \frac{exp(\beta_{0t} + x'\beta_j)}{1 + \sum_{i=1}^{m} exp(\beta_{0t} + x'\beta_i)} \qquad \begin{matrix} t = 1, \ldots, q; \\ j = 1, \ldots, m. \end{matrix} \tag{6-5}$$

Die Wahrscheinlichkeit, daß im Zeitintervall t kein Übergang bzw. Zustandswechsel stattfindet, ist dann

$$1 - \lambda(t \mid x) = \frac{1}{1 + \sum_{i=1}^{m} exp(\beta_{0t} + x'\beta_i)} \quad t = 1, \ldots, q. \tag{6-6}$$

Das logistische Modell (6.5) bzw. (6.6) wurde zur empirischen Auswertung einer Mammakarzinomstudie eingesetzt, wobei als konkurrierende Risiken die Zustände "Auftreten von Metastasen" und "Tod" festgelegt wurden und für diejenigen Patientinnen, bei denen im Laufe der Zeit Metastasen aufgetreten sind, eine zweite Verweildauer, die "Lebenszeit bis zum Tode" untersucht

wird. Da es sich um ein Mehr-Episoden-Modell handelt, wird die Auswertung erst im nächsten Kapitel vorgestellt.

Eine Verallgemeinerung von (6.5) erhält man mit

$$\lambda_j(t \mid x) = \frac{exp(\beta_{0tj} + x'\beta_{jt})}{1 + \sum_{i=1}^{m} exp(\beta_{0ti} + x'\beta_{it})} \quad \begin{matrix} t = 1, \ldots, q; \\ j = 1, \ldots, m, \end{matrix} \qquad (6\text{–}7)$$

wobei einerseits die Konstante β_{0tj} zusätzlich über die Zielzustände variiert und die Gewichtung der Einflußgrößen durch β_{jt} zeitpunktspezifisch ist. Modell (6.7) entspricht der direkten Parametrisierung der Hazardraten zum Zeitpunkt t als Logit–Modell. Modell (6.7) enthält jedoch eine sehr große Zahl von Parameter, so daß die Existenz z.B. von Maximum–Likelihood–Schätzungen nur für sehr großen Stichprobenumfang zu erwarten ist.

Das parameterökonomischere Modell (6.4) ist dadurch gekennzeichnet, daß mit $\beta_{0t1} = \ldots = \beta_{0tm} = \beta_{0t}$ die Baseline–Hazardrate (ohne Einflußgrößen) nicht spezifisch für den Zielzustand ist. Eine andere Möglichkeit der Parameterreduktion besteht in der Annahme $\beta_{01j} = \ldots = \beta_{0qj} = \beta_{0\cdot j}$. Diese Annahme entspricht der Zeitunabhängigkeit der Baseline–Hazardrate. Ohne Berücksichtigung von Einflußgrößen erhält man zustandsspezifische Hazardraten $\lambda_j(t \mid x)$, die aber über die Zeit hinweg konstant bleiben. Entsprechende Überlegungen gelten für die Gewichtsvektoren β_{jt}, die wie in (6.5) zeitunabhängig angenommen werden können. Zustandsunabhängige Gewichtsvektoren scheinen allerdings weniger angebracht.

Eine Erweiterung des Ansatzes der Exponentialmodelle mit konstantem Hazard in den Intervallen (Kapitel 5) auf den Fall konkurrierender Risiken wird von Larson (1984) betrachtet. Auf eine ausführliche Darstellung wird hier verzichtet.

Es besteht auch die Möglichkeit — zumindest für einen Spezialfall —, das gruppierte Cox–Modell auf den Mehrzustandsfall zu erweitern (vgl. auch Meindl (1988)). Den Ausgangspunkt bilden die zustandsspezifischen Hazardraten in kontinuierlicher Zeit

$$h_j(t) = h_{0j}(t)\, exp(\tilde{x}'\tilde{\beta}_j), \quad j = 1, \ldots, m,$$

insbesondere der Spezialfall

$$h_j(t) = h_0(t)\, exp(x'\beta_j)$$

mit $x' = (1, \tilde{x}')$ und $\beta_j' = (\beta_{0j}, \tilde{\beta}_j')$, $j = 1, \ldots, m$.

Für die Survivorfunktion erhält man

$$S(a_t \mid x) = P(T_s \geq a_t \mid x) = exp(-\sum_{j=1}^{m}(x'\beta_j)\int_0^{a_t} h_0(s)ds),$$

und für die zustandsspezifischen Hazardraten in diskreter Zeit ergibt sich

$$\begin{aligned}
\lambda_j(t \mid x) &= \frac{P(T = t, Y = j \mid x)}{P(T \geq t \mid x)} \\
&= \frac{P(T \in [a_{t-1}, a_t), Y = j \mid x)}{P(T \geq t \mid x)} = \frac{\int_{a_{t-1}}^{a_t} h_j(t \mid x)S(t \mid x)\, dt}{P(T \geq t \mid x)} \\
&= \int_{a_{t-1}}^{a_t} h_0(u)\, exp(x'\beta_j)\, exp(\int_0^u h_0(s)\, ds) \\
&\quad \cdot (-\sum_k exp(x'\beta_k))\, du \mid P(T \geq t \mid x)
\end{aligned}$$

Mit $F(u) = \int_0^u h_0(s)\, ds$ und $c = -\sum_k exp(x'\beta_k)$ resultiert

$$\begin{aligned}
\lambda_j(t \mid x) &= exp(x'\beta_j)\int_{a_{t-1}}^{a_t} F'(u)\, exp(F(u)c)\, du / P(T \geq t \mid x) \\
&= exp(x'\beta_j)\,\frac{1}{c}[exp(cF(a_t) - exp(cF(a_{t-1}))] \mid exp(cF(a_{t-1}))\,. \\
&= \frac{exp(x'\beta_j)}{\sum_k exp(x'\beta_k)}[1 - exp(c\int_{a_{t-1}}^{a_t} h_0(s)\, ds)]
\end{aligned}$$

Setzt man $\gamma_t = ln\,(F(a_t) - F(a_{t-1}))$, erhält man

$$\lambda_j(t \mid x) = \frac{exp(\gamma_t + x'\beta_j)}{\sum_k exp(\gamma_t + x'\beta_k)}[1 - exp(-\sum_k exp(\gamma_t + x'\beta_k))].$$

Im Falle einer beliebigen Grundhazardrate $h_{0j}(t)$ in kontinuierlicher Zeit ist die Übertragung auf den diskreten Fall nicht mehr ohne weiteres möglich.

Für die Darstellung eines weiteren diskreten Hazardraten-Modells für einen Mehrzustandsfall, das aus dem Exponentialmodell in stetig gemessener Zeit abgeleitet wird, vergleiche man Meindl (1988). Da es sich dabei um einen Spezialfall des eben vorgestellten Modells handelt, wird auf eine detaillierte Darstellung verzichtet.

6.2 Maximum-Likelihood-Schätzung

Nimmt man wie in Kapitel 3 an, daß die Zensierung zu Beginn eines Intervalls erfolgt und der Zensierungsmechanismus dem "random censoring" entspricht, erhält man in Analogie zu den Ableitungen in Abschnitt 3.4 als relevanten Beitrag von Individuum i zur Likelihoodfunktion bei gegebenem Kovariablenvektor x_i

$$\begin{aligned} L_i &= P(T_i = t_i, Y_i = y_i \mid x_i)^{\delta_i} P(T_i \geq t_i \mid x_i)^{1-\delta_i} \\ &= \lambda_{y_i}(t_i \mid x_i)^{\delta_i} P(T_i \geq t_i \mid x_i). \end{aligned} \tag{6-8}$$

Die Survivorfunktion läßt sich ebenfalls in Abhängigkeit von der Hazardrate ausdrücken. Es resultiert

$$P(T_i \geq t_i \mid x_i) = \prod_{s=1}^{t_i-1} (1 - \lambda(s \mid x_i))$$

und damit

$$L_i = \lambda_{y_i}(t_i \mid x_i)^{\delta_i} \prod_{s=1}^{t_i-1} (1 - \sum_{j=1}^{m} \lambda_j(s \mid x_i)) \quad . \tag{6-9}$$

Rein formal erhält man (6.9) auch, indem man das Produkt von t_i (für $\delta_i = 1$) bzw. $t_i - 1$ (für $\delta_i = 0$) multinomialverteilten Zufallsvariablen, jeweils mit $m+1$ Kategorien, Stichprobenumfang 1 und unterschiedlichen Vektoren von Zellwahrscheinlichkeiten, bildet. Für $\delta_i = 0$ fallen sämtliche $t_i - 1$ Beobachtungen in die letzte Kategorie, die zugehörigen Wahrscheinlichkeiten sind jeweils $1 - \sum_j \lambda_j(s \mid x_i)$. Für $\delta_i = 1$ ist zusätzlich die t_i-te Beobachtung

aus der Kategorie y_i mit der zugehörigen Wahrscheinlichkeit $\lambda_{y_i}(t_i \mid x_i)$. Dies bedeutet, daß zur numerischen Auswertung Programme für multinomiale Modelle verwendet werden können, insbesondere bei (6.5) bzw. (6.6) für multivariate Logitmodelle. Es ist lediglich die Designmatrix entsprechend zu erweitern.

Dazu betrachtet man zum s-ten Intervall und der Versuchsperson i den multinomialverteilten Vektor

$$z_s^{(i)} = (z_{s1}^{(i)}, \ldots, z_{s,m}^{(i)})' \sim M((\lambda_1(s \mid x_i), \ldots, \lambda_m(s \mid x_i))';\ 1), \qquad (6\text{–}10)$$

wobei $z_{sj} \in \{0, 1\}$.

Wegen (6.3) stellt $(\lambda_1(s \mid x_i), \ldots, \lambda_m(s \mid x_i), 1 - \lambda(s \mid x_i))$ tatsächlich einen Wahrscheinlichkeitsvektor dar, dessen Komponenten sich zu 1 aufsummieren. Weiter sei

$$z_{s,m+1}^{(i)} = 1 - \sum_{r=1}^{m} z_{sr}^{(i)}$$

die im Vektor $z_s^{(i)}$ nicht mehr enthaltene $(m+1)$te Komponente.

Für unabhängige Vektoren $z_1^{(i)}, \ldots z_{t_i-1}^{(i)}$ erhält man als gemeinsame Wahrscheinlichkeit

$$P(z_1^{(i)}, \ldots, z_{t_i-1}^{(i)}) = \prod_{s=1}^{t_i-1} \lambda_1(s \mid x_i)^{z_{s1}^{(i)}} \cdots \lambda_m(s \mid x_i)^{z_{sm}^{(i)}} (1 - \lambda(s \mid x_i))^{z_{s,m+1}^{(i)}} .$$

Für die speziellen Beobachtungsvektoren

$$z_s^{(i)} = (0, \ldots, 0)', \qquad s = 1, \ldots, t_i - 1,$$

erhält man daraus unmittelbar $z_{s,m+1}^{(i)} = 1$ und damit

$$P(z_1^{(i)}, \ldots, z_{t_i-1}^{(i)}) = \prod_{s=1}^{t_i-1} (1 - \lambda(s \mid x_i)) = \prod_{s=1}^{t_i-1} (1 - \sum_{r=1}^{m} \lambda_r(s \mid x_i))$$

und damit (6.9) für $\delta_i = 0$.

Völlig analog erhält man für die unabhängigen Zufallsvektoren $z_1^{(i)}, \ldots, z_{t_i}^{(i)}$ mit den speziellen Beobachtungen

$$z_s^{(i)} = (0, \ldots, 0)' \quad \text{für } s = 1, \ldots, t_i - 1$$

und

$$z_{t_i}^{(i)} = e_{y_i}$$

mit dem y_i-ten Einheitsvektor e_{y_i} die Wahrscheinlichkeit

$$P(z_1^{(i)}, \ldots, z_{t_i}^{(i)}) = \lambda_{y_i}(t_i \mid x_i) \prod_{s=1}^{t_i-1} (1 - \sum_{r=1}^{m} \lambda_r(s \mid x_i))$$

und damit (6.9) für $\delta_i = 1$.

Aus den vorangegangenen Ausführungen wird deutlich, daß wie in Abschnitt 3.4 zur numerischen Berechnung der Maximum-Likelihood-Schätzungen des Modells (6.5) Programme für multinomiale Logit-Modelle bzw. verallgemeinerte lineare Modelle herangezogen werden können, wobei das Design erweitert wird mit $t_i - 1$ gegebenen Responsevektoren $z_1^{(i)}, \ldots, z_{t_i-1}^{(i)}$ zur Person i, wenn $\delta_i = 0$, und t_i Responsevektoren, wenn $\delta_i = 1$.

Im folgenden wird das Design für das Logit–Modell (6.5) angegeben. Für allgemeinere Modelle wie (6.7) wird die Design–Matrix entsprechend umfangreicher. Die Darstellung der Design–Matrix ist insofern von Bedeutung als damit unmittelbar Programmpakete für das multivariate Logit–Modell anwendbar sind. Der Teil der Designmatrix des Logit-Modells (6.5) für die i-te Person, mit $\delta_i = 0$ ergibt sich wie folgt:

Abhängiger Variablenvektor		Designvektoren		
$(z_{11}^{(i)}, \ldots, z_{1m}^{(i)})$	=	$x_{11}^{(i)}$	=	$(e_1', x_i', 0, \ldots, 0)'$
		$\vdots$		
$(0, 0, \ldots\ldots, 0)$		$x_{1m}^{(i)}$	=	$(e_1', 0, \ldots, 0, x_i')'$
$\vdots$		$\vdots$		
$(z_{t_i-1,1}^{(i)}, \ldots, z_{t_i-1,m}^{(i)})$	=	$x_{t_i-1,1}^{(i)}$	=	$(e_{t_i-1}', x_i', 0, \ldots, 0)'$
$(0, 0, \ldots\ldots, 0)$		$\vdots$		
		$x_{t_i-1,m}^{(i)}$	=	$(e_{t_i-1}', 0, \ldots, 0, x_i')'$

Für den Fall $\delta_i = 1$ erhält man nach analogen Überlegungen:

Abhängiger Variablenvektor		Designvektoren		
$(z_{11}^{(i)}, \ldots, z_{1m}^{(i)})$	$=$	$x_{11}^{(i)}$	$=$	$(e_1', x_i', 0, \ldots, 0)'$
$(0, 0, \ldots\ldots, 0)$		$\vdots$		
$\vdots$		$x_{1m}^{(i)}$	$=$	$(e_1', 0, \ldots, 0, x_i')'$
		$\vdots$		
$(z_{t_i-1,1}^{(i)}, \ldots, z_{t_i-1,m}^{(i)})$	$=$	$x_{t_i-1,1}^{(i)}$	$=$	$(e_{t_i-1}', x_i', 0, \ldots, 0)'$
$(0, 0, \ldots\ldots, 0)$		$\vdots$		
		$x_{t_i-1,m}^{(i)}$	$=$	$(e_{t_i-1}', 0, \ldots, 0, x_i')'$
$(z_{t_i,1}^{(i)}, \ldots, z_{t_i,m}^{(i)})$	$=$	$x_{t_i,1}^{(i)}$	$=$	$(e_{t_i}', x_i', 0, \ldots, 0)'$
$(0, \ldots, 1, \ldots, 0)$		$\vdots$		,
		$x_{t_i,m}^{(i)}$	$=$	$(e_{t_i}', 0, \ldots, 0, x_i')'$

wobei $x_{st}^{(i)}$ jeweils eine Zeile der Design–Matrix darstellen.

Die abhängigen Variablenvektoren $z_s^{(i)}$ sind hier jeweils um die letzte Stelle verkürzt dargestellt. Der Grund liegt darin, daß für das in der Schätzung zugrundegelegte multinomiale Logit–Modell nur die ersten m Zustände relevant sind. Für den multinomialverteilten Vektor $z_s^{(i)}$ in (6.5) ergibt sich die letzte 'Auftretenswahrscheinlichkeit' $1 - \lambda(t \mid x_i)$ durch

$$1 - \sum_{j=1}^{m} \lambda_j(t \mid x_i).$$

Das heißt, die letzte Auftretenswahrscheinlichkeit ist durch die ersten m fest bestimmt. Entsprechend ist für den multinomialen Vektor in (6.5) die letzte Ausprägung durch die ersten m eindeutig bestimmt. Gilt $z_{sj}^{(i)} = 0$ für $j = 1, \ldots, m$, muß $z_{s,m+1}^{(i)} = 1$ gelten. Die Notwendigkeit der reduzierten Darstellung im Design ergibt sich daraus, daß nur m Zeilen der Designmatrizen gemäß Modell (6.5) bildbar sind, da nur die Parameter (β_{0t}, β_j'), $j = 1, \ldots, m$ frei variieren können.

7. Modelle für den Mehr-Episoden-Fall

Sind nicht alle Zielzustände absorbierend, so können für jedes Individuum bzw. Objekt mehrere Episoden aufeinanderfolgen. Beispiele hierfür sind Verweildauern in verschiedenen Berufen bei der Untersuchung von Berufskarrieren (vgl. z.B. Blossfeld/Hamerle/Mayer (1986)), die Dauer der Arbeitslosigkeit in möglicherweise mehreren aufeinanderfolgenden Perioden, die Dauer bis zum Umzug in eine andere Region bei Wanderungs- und Mobilitätsanalysen, die Nutzungsdauer von langlebigen Konsumgütern etc.

7.1 Episodenspezifische Hazardraten

Die zufälligen Übergangszeiten werden nun repräsentiert durch nicht negative Zufallsvariablen $T_0 = 0 \leq T_1 \leq T_2 \leq \ldots$ mit den zugehörigen Zustandsvariablen $Y_0, Y_1, Y_2, \ldots$.

Durchläuft ein Individuum den Prozeß bis zur n_i-ten Episode, so wird der Pfad des Individuums folgendermaßen generiert: Das Individuum befindet sich zum Zeitpunkt $T_0 = 0$ im Zustand y_0. Im Zeitintervall t_1 erfolgt der erste Übergang in den Zustand y_1, und es beginnt die zweite Episode. Im Zustand y_1 verweilt das Individuum bis zum Zeitintervall t_2, $t_2 \geq t_1$, und wechselt in diesem Intervall in den Zustand y_2, usw.

Die im letzten Abschnitt dargestellten Competing-Risk-Ansätze stellen Spezialfälle der hier behandelten Modelle dar, wenn nur eine Episode vorliegt. Wegen ihrer besonderen Bedeutung in verschiedenen Bereichen, z.B. zur Analyse von Lebenszeiten, wurde ihnen ein eigener Abschnitt gewidmet.

In jeder Episode wird für ein Individuum bzw. Objekt ein Vektor von Kovariablen x_k erhoben, deren Komponentenzahl von Episode zu Episode variieren kann. Die übergangsspezifische Hazardrate der k-ten Episode ($k = 1, 2, \ldots$) beim Übergang $y_k = j$ wird folgendermaßen definiert:

$$\lambda_j^k(t \mid H_{k-1}, x_k) = P(T_k = t, Y_k = j \mid T_k \geq t, H_{k-1}, x_k). \qquad (7\text{-}1)$$

Dabei wird in H_{k-1} die Vorgeschichte des Prozesses, die sich darstellen läßt durch $\{y_0, t_1, y_1, x_1, \ldots, t_{k-1}, y_{k-1}, x_{k-1}\}$, $H_0 = \{y_0\}$, zusammengefaßt. Man beachte, daß $\lambda_j^k(t \mid H_{k-1}, x_k)$ identisch gleich 0 ist für $t \leq t_{k-1}$.

Die Gesamthazardrate $\lambda^k(t \mid x_k, H_{k-1})$, in der k-ten Episode den Zustand y_{k-1} zu verlassen, ist

$$\lambda^k(t \mid x_k, H_{k-1}) = \sum_j \lambda_j^k(t \mid x_k, H_{k-1}).$$

Die Survivorfunktion kann ebenfalls auf den Mehr–Episoden–Fall erweitert werden. Man definiert

$$S^k(t \mid x_k, H_{k-1}) = P(T_k > t \mid x_k, H_{k-1})$$

für $t \geq t_{k-1}$.

Für die Zeitintervalle, in denen Zustandswechsel stattfinden, ist eine Zusatzannahme notwendig. Eine mögliche Annahme ist, daß in einem Zeitintervall höchstens ein Ereignis stattfinden kann, so daß die k-te Episode erst im Intervall $t_{k-1}+1$ beginnt, wenn die $(k-1)$-te Episode im Intervall t_{k-1} endete. Für kleinere Zeitintervalle dürfte dies unproblematisch sein. Bei größeren Intervallbreiten hingegen kann es vorkommen, daß die $(k-1)$-te Episode im Intervall t_{k-1} endet, die k-te Episode beginnt und ebenfalls bereits in diesem Intervall endet. Dies ist z.B. bei der Mammakarzinomstudie, die am Ende dieses Abschnitts dargestellt wird, gelegentlich der Fall. In solchen Fällen modifizieren wir die Annahme und setzen voraus, daß die k-te Episode in dem gleichen Intervall beginnt, in dem die $(k-1)$-te Episode endete. Unter der Bedingung, daß $T_{k-1} = t_{k-1}$ ist, ist dann $\lambda^k(t_{k-1} \mid x_i) \geq 0$, während im oben beschriebenen Fall $\lambda^k(t_{k-1} \mid x_i) = 0$ gilt. Wir gehen im folgenden von der zuletzt genannten Annahme aus, bei der zuerst beschriebenen Annahme sind lediglich geringfügige Modifikationen notwendig.

Die episoden- und zielzustandsspezifischen Hazardraten (7.1) können wieder in Abhängigkeit von den Kovariablen x_k und der Vorgeschichte H_{k-1} parametrisiert werden. Meist wird dabei der relevante Teil von H_{k-1} in den Kovariablenvektor x_k aufgenommen, und man benützt eine Parametrisierung der Form

$$\lambda_j^k(t \mid x_k) = g(\beta_{0jt}^k + x_k' \beta_j^k) \tag{7–2}$$

$t = 1, \ldots, q;\ j = 1, \ldots, m;\ k = 1, 2, \ldots$. Die Regressionskoeffizienten β_j^k können von der Episode und dem Zielzustand abhängen. Für die empirische

Auswertung am Ende dieses Abschnitts wird ein logistischer Ansatz gewählt mit den Hazardraten

$$\lambda_j^k(t \mid x_k) = \frac{exp(\beta_{0jt}^k + x_k'\beta_j^k)}{1 + \sum_i exp(\beta_{0it}^k + z_k'\beta_i^k)}.$$

Da die Länge des Beobachtungszeitraumes im allgemeinen vorgegeben ist, kann die letzte Episode eines Individuums bzw. Objektes zensiert sein.

7.2 Maximum-Likelihood-Schätzung

Die Schätzprozedur wird nun auf den Mehr-Episoden-Fall verallgemeinert. Zunächst wird vorausgesetzt, daß keine Zensierungen vorliegen. Ferner wird aus Gründen der einfacheren Schreibweise der Index i bei T_{ik}, Y_{ik} etc. weggelassen.

Der Beitrag von Individuum i zur Likelihoodfunktion ist

$$\begin{aligned} L_i = {} & P(T_{n_i} = t_{n_i}, Y_{n_i} = y_{n_i}, x_{n_i}, \ldots, T_1 = t_1, Y_1 = y_1, x_1 \mid y_0) \\ & \prod_{k=1}^{n_i} P(T_k = t_k, Y_k = y_k, x_k \mid H_{k-1}) \end{aligned} \tag{7-3}$$

mit $H_{k-1} = \{t_{k-1}, y_{k-1}, x_{k-1}, \ldots, t_1, y_1, x_1, y_0\}, \quad H_0 = \{y_0\}.$

Man erhält

$$\begin{aligned} L_i = {} & \prod_{k=1}^{n_i} P(T_k = t_k, Y_k = y_k \mid T_k \geq t_k, H_{k-1}, x_k) \\ & \cdot P(T_k \geq t_k \mid H_{k-1}, x_k) P(x_k \mid H_{k-1}). \end{aligned} \tag{7-4}$$

Dies ergibt

$$L_i = \prod_{k=1}^{n_i} \lambda_{y_k}^k(t_k \mid H_{k-1}, x_k) \prod_{s=t_{k-1}}^{t_k - 1} (1 - \lambda^k(s \mid H_{k-1}, x_k)) P(x_k \mid H_{k-1}) \tag{7-5}$$

Sind zensierte Beobachtungen der letzten Episode registriert worden, ist (7.4) analog zu (6.8) zu modifizieren, und es resultiert

$$L_i = \prod_{k=1}^{n_i} [\lambda_{y_k}^k(t_k \mid H_{k-1}, x_k)]^{\delta_k} \prod_{s=t_{k-1}}^{t_k-1} (1 - \lambda^k(s \mid H_{k-1}, x_k)) P(x_k \mid H_{k-1}) \tag{7-6}$$

mit $\delta_k = 1$ für $k = 1, \ldots, n_i - 1$ und $\delta_{n_i} = 0$, falls die n_i-te Episode zensiert ist, und $\delta_{n_i} = 1$ andernfalls. Hängt $P(x_k \mid H_{k-1})$ nicht von den in Frage stehenden Parametern, die die Hazardrate $\lambda_j^k(\cdot)$ determinieren, ab, kann dieser Faktor bei der Maximierung der Likelihoodfunktion vernachlässigt werden. Andernfalls kann (7.6) ohne den Faktor $P(x_k \mid H_{k-1})$ als "Partial Likelihood" (vgl. Cox (1972, 1975) oder Kalbfleisch/Prentice (1980), Kap. 5) betrachtet werden.

Die Gesamtlikelihoodfunktion

$$L = \prod_{i=1}^{n} L_i$$

mit L_i aus (7.6) ohne den Faktor $P(x_k \mid H_{k-1})$ kann auch in einer anderen Form dargestellt werden, die gelegentlich zweckmäßig ist. Mit

$$\delta_{ikj} = \begin{cases} 1, & \text{falls die } k\text{-te Episode von Ind. } i \text{ im} \\ & \text{Zeitintervall } t_{ik} \text{ im Zustand } j \text{ endet;} \\ 0, & \text{sonst} \end{cases}$$

und

$$\varepsilon_{ik} = \begin{cases} 1, & \text{falls Ind.} i \text{ die } k\text{-te Episode erlebt;} \\ 0, & \text{sonst} \end{cases}$$

ergibt sich

$$L = \prod_k \prod_{i=1}^{n} \prod_{j=1}^{m} [\lambda_j^k(t_{ik} \mid H_{i,k-1}, x_{ik})]^{\delta_{ikj}} [\prod_{s=t_{i,k-1}}^{t_{ik}-1} (1 - \lambda^k(s \mid H_{i,k-1}, x_{ik}))]^{\varepsilon_{ik}}. \tag{7-7}$$

Aus (7.6) wird ersichtlich, daß die Likelihoodfunktion bzw. die Log-Likelihoodfunktion getrennt für jedes k maximiert werden kann, wenn die Hazardraten für jede Episode von verschiedenen Parametervektoren abhängen. Für die k-te Episode sind nur diejenigen Mitglieder der Stichprobe heranzuziehen, die mindestens k Episoden erlebt haben.

7.3 Anwendungsbeispiel

In diesem Abschnitt wird anhand eines Datensatzes aus einer Mammakarzinomstudie (vgl. Schedel, 1986) ein Competing-Risks-Modell mit zwei aufeinanderfolgenden Episoden angesetzt.

Bei 748 Patientinnen, die an Brustkrebs operiert wurden, wurde die Lebenszeit nach der Operation registriert, eventuell mit Zensierung. Als konkurrierende Risiken werden die Zustände "Auftreten von Metastasen" und "Tod" festgelegt. Die erste Verweildauer ist als "Krankheitsfreier Verlauf" definiert, während für diejenigen Patientinnen, bei denen im Laufe der Zeit Metastasen aufgetreten sind, die zweite Verweildauer als "Überlebenszeit bis zum Tod" gegeben ist.

Die zeitlichen Verläufe werden in der folgenden Abbildung veranschaulicht.

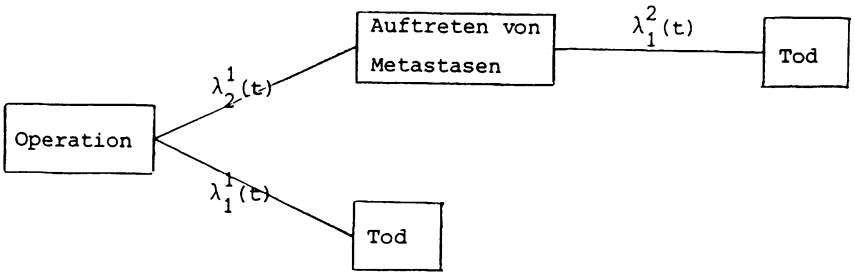


Abbildung 7.1 .

Für jede Patientin wurden folgende prognostische Faktoren erhoben:

- Alter

- TNM-Klassifikationen (der malignen Mammatumoren)

  T Ausdehnung des Primärtumors; 5 Kategorien (zusammengefaßt zu 3 Kat.)

    T1: Größer als 5 cm

    T2: Zwischen 2 und 5 cm

    T3: Kleiner als 2 cm

  N Zustand der regionären Lymphknoten; 4 Kategorien (zusammengefaßt zu 3 Kat.)

    N1: befallen, fixiert oder diffus

    N2: befallen, beweglich

    N3: klinisch negativ

  M Nachweis bzw. Ausschluß von Fernmetastasen zum Zeitpunkt der Operation

- Anteil der befallenen Lymphknoten

- Histologischer Befund (5 Kategorien)

- Metastasenort (falls Metastasen aufgetreten sind)

  (Leber–Lunge–Gehirn–Knochen–Haut–Lymphknoten–Mamma)

Für eine erste Teilauswertung mit dem logistischen Modell wurden die Kovariablen

(1) Alter

(2) Ausdehnung des Primärtumors

(3) Zustand der regionären Lymphknoten

(4) Anteil der befallenden Lymphknoten

ausgewählt und für die zweite Periode der Patientinnen, bei denen Metastasen auftraten, wurden die Kovariablen

(5) Metastasenort (M1–M7)

(6) Anzahl der von Metastasen befallenen Organe

(ein Organ–mehr als ein Organ)

(7) Zeit bis zum Auftreten von Metastasen hinzugenommen.

Es wurden 15 Zeitintervalle gebildet (in Monaten): 0–1, 1–2, 2–4, 5–10, 11–14, 15–19, 20–24, 25–30, 31–35, 36–44, 45–54, 55–74, 75–99, 100–149, über 149.

Für jede Episode wurde wieder der logistische Ansatz für die Hazardrate zugrundegelegt. Während der ersten Episode ist die Hazardrate

$$\lambda_j^1(t \mid x_1) = \frac{exp(\beta_{0jt}^1 + x_1'\beta_j^1)}{1 + \sum_{i=1}^{2} exp(\beta_{0it}^1 + x_1'\beta_i^1)} \quad , \quad j = 1,2,$$

wobei x_1 die Kovariablen (1) bis (4) enthält. Für die anschließende zweite Episode ist die Hazardrate

$$\lambda^2(t \mid x_2) = \frac{exp(\beta_{0t}^2 + x_2'\beta^2)}{1 + exp(\beta_{0t}^2 + x_2'\beta^2)} \quad , \quad t \geq t_1,$$

wobei x_2 außer den Kovariablen (1) bis (4) noch die Merkmale (5) bis (7) enthält.

Die Berechnungen wurden an einem IBM PC-AT unter Verwendung des GAUSS-Programms von Edlefsen und Jones durchgeführt. Die ML-Schätzungen der Regressionskoeffizienten sind in den folgenden Tabellen wiedergegeben, wobei die Werte in Klammern jeweils $\hat{\beta}_i/\hat{\sigma}_i$ enthalten.

1. Episode

	Risiko "Auftreten von Metastasen"	Risiko"Tod"
Alter	-0.0056	0.0158
	(-0.69367)	(2.55964)
T1	0.6438	0.2616
	(2.18089)	(1.03566)
T2	0.0721	0.0427
	(0.31189)	(0.26923)
N1	0.6131	-0.1354
	(2.02985)	(-0.50563)
N2	0.6666	-0.0563
	(2.91154)	(-0.30562)
Anteil der befal-	-0.0038	0.1916
lenen Lymphknoten	(-0.013232)	(0.76307)

2. Episode

Alter	-0.0012
	(-0.1069)
T1	0.0054
	(0.01360)
T2	0.0561
	(0.18496)
N1	-1.0483
	(-2.5517)
N2	-0.7715
	(-2.6021)

Anteil der befallenen Lymphknoten	-0.1196 (-0.3326)
Anzahl der befallenen Organe	-0.7872 (1.40864)
Zeit bis zum Auftreten von Metastasen	-0.0028 (-0.9508)
M1	0.0874 (0.19000)
M2	-0.8413 (-1.4483)
M3	1.0120 (1.20392)
M4	-1.2807 (-2.9715)
M5	0.2324 (0.51037)
M6	0.1319 (0.28713)
M7	-1.2598 (-2.3798)

Im ersten Schritt wurde die Hypothese getestet, daß die entsprechenden Regressionskoeffizienten der Kovariablen (1) bis (4) in allen drei Übergangsarten gleich sind. Es ergab sich ein Wert der Likelihood–Quotienten–Teststatistik von 124.4 bei 12 Freiheitsgraden ($\chi^2(0.95; 12) = 21.03$). Die Hypothese, daß die 6 Koeffizienten der Kovariablen (1) bis (4) für die Zeit bis zum Auftreten von Metastasen (1. Periode) und danach für die Zeit bis zum Tod (2. Periode) gleich sind, mußte bei einem Wert der Likelihood-Quotienten-Teststatistik von 49.1 bei 6 Freiheitsgraden ($\chi^2(0.95; 6) = 12.59$) ebenfalls abgelehnt werden.

In der ersten Periode mit Übergang in den Endzustand "Tod" besitzt lediglich das Alter einen signifikanten Einfluß. Dies ist plausibel und entspricht dem natürlichen Anwachsen des Sterberisikos mit zunehmenden Alter.

Anders ist der Sachverhalt beim konkurrierenden Risiko "Auftreten von Metastasen". Hier verringert fortgeschrittenes Alter eher die Hazardrate, allerdings nicht in signifikantem Ausmaß. Gleichfalls ungünstig ist eine große Ausdehnung des Primärtumors (T1), und auch die zunehmende Involvierung der regionären Lymphknoten (N1, N2) erhöht die Hazardrate beträchtlich.

Wichtige Einflußgrößen für die Lebenszeit nach dem Auftreten von Metastasen sind bestimmte Lokalisationen der Metastasen und der Zustand der Lymphknoten. Bemerkenswert ist, daß die Dauer der 1. Periode, d.h. die Zeit bis zum Auftreten der Metastasen, keinen signifikanten Einfluß auf die restliche Lebenszeit hat.

8. Modelle mit Einbeziehung unbeobachteter Populationsheterogenität

In der Regel werden neben den in das Modell aufgenommenen Kovariablen weitere personen- oder umweltspezifische Merkmale, die nicht erhoben worden oder unbekannt sind, die Übergangs- bzw. Hazardraten beeinflussen. Diese unbeobachteten Variablen wurden bisher in den Modellen nicht berücksichtigt. Wie sich dies auf die Hazardrate auswirken kann, soll an einem Beispiel demonstriert werden.

Angenommen, die Population sei in zwei Teilgesamtheiten unterteilt, die mit x_1 und x_2 gekennzeichnet werden. In jeder der Teilgesamtheiten sei die Übergangsrate konstant, etwa

$$\lambda(t \mid x_1) = \lambda_1,$$

$$\lambda(t \mid x_2) = \lambda_2,$$

und die Wahrscheinlichkeiten, daß ein Individuum der 1. bzw. 2. Teilgesamtheit entstammt, seien $p(x_1)$ und $p(x_2)$. Wird die Unterteilung in die beiden Teilgesamtheiten nicht berücksichtigt, erhält man für die Übergangsrate in der Gesamtpopulation

$$\lambda(t) = \frac{\sum \lambda(t \mid x_i) S(t \mid x_i) p(x_i)}{\sum S(t \mid x_i) p(x_i)} \quad ,$$

wobei $S(t \mid x_i)$ die Survivor-Funktion bezeichnet. Für $\lambda_1 \neq \lambda_2$ ist $\lambda(t)$ nicht mehr konstant über die Zeitintervalle $t = 1, 2, \ldots$. Setzt man speziell

$$\lambda_1 = 0.1, \qquad \lambda_2 = 0.4 \qquad \text{und} \qquad p(x_1) = p(x_2) = 0.5,$$

wird der Verlauf der Übergangsraten in der folgenden Abbildung dargestellt.

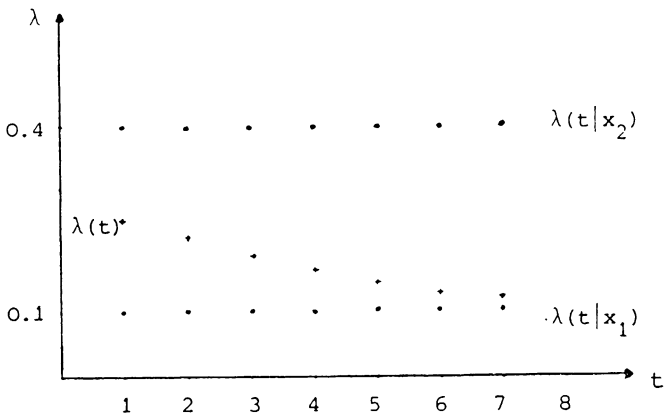


Abb.8.1

Die Nichtberücksichtigung der Populationsheterogenität führt zu einer abnehmenden Übergangsrate.

Modelle mit Einbeziehung unbeobachteter Populationsheterogenität wurden bisher nur für die Ansätze mit stetig gemessener Zeit und hier fast ausschließlich für den Ein-Episoden-Fall behandelt (Ausnahmen bilden Flinn/Heckman (1982) und Newman/McCulloch (1984)). Insbesondere vom theoretischen Standpunkt aus sind die Modelle noch nicht hinreichend untersucht. Im folgenden werden Möglichkeiten dargestellt, unbeobachtete Populationsheterogenität auch bei diskreten Modellen explizit in den Modellansatz mitaufzunehmen. Dabei beschränken wir uns hier auf eine Kurzdarstellung der wichtigsten Konzepte, für eine ausführliche Beschreibung siehe Hamerle (1986).

Wir betrachten zunächst den Fall einer Episode, eines einzigen Endzustands und zeitunabhängiger Kovariablen. Die unbeobachtete Populationsheterogenität kann durch eine (reellwertige) Zufallsvariable α in der Übergangsrate repräsentiert werden. Sie hat dann die Form

$$\lambda(t \mid x, \alpha) = P(T = t \mid T \geq t, x, \alpha) \tag{8-1}$$

und speziell für das logistische Modell lautet der Ansatz

$$\lambda(t \mid x, \alpha) = \frac{exp(\alpha + \beta_{0t} + x'\beta)}{1 + exp(\alpha + \beta_{0t} + x'\beta)} \quad . \tag{8-2}$$

Dabei variiert α von Individuum zu Individuum. Betrachtet man α als Realisierung einer Zufallsvariablen, so spricht man von einem Random-Effect-Modell. Die Verteilung von α in der Population wird mit $G(\alpha)$ bezeichnet. Wählt man statt (8.2) eine Probit-Spezifikation, ergibt sich

$$\lambda(t \mid x, \alpha) = \Phi(\alpha + \beta_{0t} + x'\beta).$$

Dabei ist Φ die Verteilungsfunktion der Standardnormalverteilung.

Für die Verteilung von T bei gegebenem beobachteten Kovariablenvektor ergibt sich

$$P(T = t \mid x) = \int [\lambda(t \mid x, \alpha)]^{\delta} \prod_{j=1}^{t-1} (1 - \lambda(j \mid x, \alpha))\, dG(\alpha) \quad , \tag{8-3}$$

wobei δ wieder ein Indikator für die Zensierung ist ($\delta = 0$, falls T zensiert ist).

Die marginalen Wahrscheinlichkeiten (8.3) können zur Bildung der marginalen Likelihoodfunktion herangezogen werden. Mit

$$\varepsilon_{it} = \begin{cases} 1, & \text{falls die Verweildauer des Individuums } i \\ & \text{im Zeitintervall } t \text{ endet;} \\ 0, & \text{sonst} \end{cases}$$

$$\phi_{it} = \begin{cases} 1, & \text{falls Individuum } i \text{ das Ende} \\ & \text{des Zeitintervalls } t \text{ erreicht,} \\ 0, & \text{sonst} \end{cases}$$

erhält man für die logarithmierte marginale Likelihoodfunktion

$$lnL_M = \sum_{i=1}^{n} ln \int \prod_{t=1}^{q} \lambda(t \mid x_{it}, \alpha)^{\varepsilon_{it}} (1 - \lambda(t \mid x_{it}, \alpha)^{\phi_{it}}\, dG(\alpha). \tag{8-4}$$

Setzt man

$$\lambda(t \mid x_t, \alpha) = F(\alpha + \beta_{0t} + x_t\beta)$$

mit einer Verteilungsfunktion F, etwa der Normalverteilung oder der logistischen Verteilung entsprechend (8.2), und trifft die Annahme, daß $G(\alpha)$ zu einer parametrischen Verteilungsklasse gehört, etwa $N(0;\sigma_\alpha^2)$, so lassen sich die Modellparameter β_{0t} und β zusammen mit den Parametern von $G(\alpha)$ mit Hilfe der Maximum-Likelihood-Methode aus (8.4) schätzen. Dabei kann das Integral in (8.4) durch die Gauß–Hermite–Quadraturformel approximiert werden.

Aus (8.4) wird ersichtlich, daß die Wahl von $G(\alpha)$ die Form der Likelihoodfunktion und damit auch die Schätzung der strukturellen Modellparameter β_{0t} und β beeinflussen kann. Heckmann/Singer (1982) gelangen bei der Analyse eines empirischen Datensatzes für eine angenommene Weibull-Hazardrate mit stetig gemessener Zeit bei verschiedenen $G(\alpha)$ zu recht unterschiedlichen Resultaten für die Schätzungen der $\beta's$. Sie schlagen deshalb eine alternative Strategie vor, auf die wir im nächsten Abschnitt eingehen werden. Zu einem anderen Ergebnis kommen Newman/McCulloch (1984) bei einem Datensatz über die Zeitabschnitte zwischen aufeinanderfolgenden Geburten. Sie wählen als "mischende" Verteilung $G(\alpha)$ verschiedene diskrete Approximationen der Gammaverteilung sowie der Lognormalverteilung und kamen zu dem Ergebnis, daß sich die Schätzungen für die $\beta's$ nur wenig unterscheiden.

Trussel/Richards (1985) zeigen anhand von Simulationen, daß die marginale Likelihood-Schätzung auch sehr sensitiv ist gegenüber einer Fehlspezifikation der Hazardrate. Daher könnten die Resultate von Heckmann/Singer auch darauf zurückzuführen sein, daß das gewählte Weibull-Modell nicht adäquat ist.

Für mehrere aufeinanderfolgende Zeitperioden werden wir uns auf den "Repeated-event"-Fall beschränken. Unterstellt man für jede Episode eine andere Heterogenitätskomponente α_k, so ist eine Annahme über die gemeinsame Verteilung der α_k notwendig, da die einzelnen Komponenten im allgemeinen nicht unabhängig sind. Die Rechtfertigung einer bestimmten Wahl dieser gemeinsamen Verteilung ist — wie schon die Wahl von $G(\alpha)$ — aus dem Sachzusammenhang heraus in der Regel schwierig. Wir treffen hier die vereinfachende Annahme

$$\alpha_k = \gamma_k \alpha \quad ,$$

wobei γ_k ein episodenspezifischer Parameter ist.

Für die Hazardrate wird

$$\lambda^k(t \mid x_{kt}, \alpha_k) = F(\gamma_k \alpha + \beta_{0t}^k + x'_{kt}\beta^k) \tag{8–5}$$

gesetzt mit einer speziellen Verteilungsfunktion F, z.B. der logistischen Verteilung.

Mit

$$\varepsilon_{ikt} = \begin{cases} 1, & \text{falls für Individuum } i \text{ die } k\text{-te Episode im Intervall } t \\ & \text{zu Ende geht;} \\ 0, & \text{sonst} \end{cases}$$

$$\Phi_{ikt} = \begin{cases} 1, & \text{falls für Individuum } i \text{ die } k\text{-te Episode am Ende des} \\ & \text{Intervalls } t \text{ andauert;} \\ 0, & \text{sonst} \end{cases}$$

erhält man für die logarithmierte marginale Likelihoodfunktion

$$\begin{aligned} lnL_M = \sum_{i=1}^{n} ln \int \prod_{k} \prod_{t=1}^{q} & F(\gamma_k \alpha + \beta_{0t}^k + x'_{ikt}\beta^k)^{\varepsilon_{ikt}} \\ & (1 - F(\gamma_k \alpha + \beta_{0t}^k + x'_{ikt}\beta^k))^{\Phi_{ikt}} \, dG(\alpha) \end{aligned} \tag{8–6}$$

aus der sich wieder Schätzungen für die unbekannten Parameter ermitteln lassen, wenn $G(\alpha)$ spezifiziert wird.

Gewöhnlich wird die Annahme getroffen, daß die Heterogenitätskomponente unabhängig ist von den beobachteten Kovariablen. Dies ist insbesondere in sämtlichen Beiträgen, die im folgenden Abschnitt zitiert werden, der Fall. Auf der anderen Seite wird die Einbeziehung unbeobachteter Populationsheterogenität in der Regel durch den Einwand motiviert, man könne in empirischen Anwendungen niemals alle relevanten Einflußgrößen erheben, und man habe bei Nichtberücksichtigung unbeobachteter Merkmale mit einer Verzerrung der Resultate zu rechnen (omitted variables bias). Mit Sicherheit werden jedoch die unbeobachteten Merkmale bei einem Individuum nicht unabhängig sein von den erhobenen Merkmalen. Unterstellt man Unabhängigkeit, so wird das Problem nicht beobachteter Einflußgrößen hinausdefiniert, und der omitted variables bias ist nicht beseitigt. Man vergleiche dazu auch Chamberlain (1980).

Auch bei den in diesem Abschnitt vorgestellten diskreten Modellen ist dieses Problem zu berücksichtigen. Man sollte nicht eine Verteilung $G(\alpha)$, sondern eine Verteilung $G(\alpha \mid x)$ spezifizieren. Dies ist im praktischen Anwendungsfall schwierig. Eine Möglichkeit besteht darin, einen Regressionsansatz

$$\alpha = x'\pi + \varepsilon$$

zu formulieren und in das Modell aufzunehmen. Der Parametervektor π ist mitzuscätzen. Allerdings sind dabei in der Regel Identifikationsprobleme zu berücksichtigen.

Heckman/Singer (1982, 1984a,b) schlagen aufgrund der möglichen Sensitivität der Schätzungen der strukturellen Modellparameter gegenüber der Wahl der Verteilung der unbeobachteten Heterogenität eine simultane Schätzung der Modellparameter und der Verteilung der Heterogenitätskomponente vor, ähnlich der empirischen Bayes-Schätzung (vgl. z.B. Maritz (1971)).

Im folgenden betreffen alle Aussagen den Ein-Episoden-Fall, und der Kovariablenvektor wird als zeitunabhängig vorausgesetzt. Sei die bedingte Verteilung der Verweildauer, gegeben die Kovariablen und die Heterogenitätskomponente, mit $f(t \mid x, \alpha; \beta)$ bezeichnet, die mischende Verteilung sei $G(\alpha)$, und die Mischverteilung ist

$$h(t \mid x; \beta) = \int f(t \mid x, \alpha; \beta)\, dG(\alpha). \tag{8–7}$$

Das Ziel besteht darin, neben den Parameterschätzungen $\hat{\beta}$ auch eine Schätzung $\hat{G}(\alpha)$ zu finden, die mit zunehmendem Stichprobenumfang zumindest nach Wahrscheinlichkeit gegen $G(\alpha)$ konvergiert.

Bevor dieses Ziel näher untersucht werden kann, sind zuerst verschiedene Identifikationsprobleme zu lösen. Läßt man das Schätzproblem gänzlich außer acht, so stellt sich die Frage, ob die Kenntnis der Mischverteilung $h(t \mid x; \beta)$ ausreicht, damit die Integralgleichung (8.7) mit eindeutig bestimmten Funktionen $f(t \mid x, \alpha; \beta)$ und $G(\alpha)$ erfüllt ist. Ohne weitere Zusatzannahmen ist dies sicher nicht gewährleistet. Auch bei Spezifikation der bedingten Verteilung $f(t \mid x, \alpha; \beta)$ ist nicht in jedem Fall gesichert, daß nicht zwei verschiedene mischende Verteilungen $G_1(\alpha)$ und $G_2(\alpha)$ dieselbe Mischungsverteilung ergeben. Für Beispiele vergleiche man Heckman/Singer (1984b).

Für stetige Verweildauermodelle (eine Verweildauer, ein Endzustand) haben Elbers/Ridder (1982) für die Klasse der Proportional-Hazards-Modelle mit

$$\lambda(t \mid x) = \lambda_0(t) exp(x'\beta)$$

die Identifizierbarkeit gezeigt. Eine wichtige Bedingung ist dabei, daß der Kovariablenvektor mindestens eine stetige Komponente enthält. Man vergleiche dazu auch Hougaard (1984) und Heckman/Singer (1984a,b). Eine weitere zentrale Forderung, die in allen Beiträgen enthalten ist, betrifft die Unabhängigkeit der Heterogenitätskomponente und der Kovariablen. Wie im letzten Abschnitt ausgeführt wurde, ist damit das "omitted-variables"-Problem im allgemeinen nicht gelöst.

Kiefer/Wolfowitz (1956) geben allgemeine Bedingungen für die Existenz eines konsistenten Schätzers der mischenden Verteilung und der strukturellen Modellparameter an. Ihr Aufsatz enthält aber keinen Hinweis auf eine konstruktive Vorgehensweise bei der numerischen Ermittelung der ML-Schätzungen. Heckman/Singer (1984b) verifizieren die Kiefer/Wolfowitz-Bedingungen für Proportional-Hazards-Modelle mit zeitabhängigen Kovariablen und möglicherweise zensierten Daten. Darüber hinaus schlagen sie einen nicht parametrischen Maximum-Likelihood-Schätzalgorithmus vor, der auf der theoretischen Charakterisierung der ML-Schätzung bei Mischverteilungen von Lindsay (1983a,b) beruht. Sie propagieren dazu die Verwendung des EM-Algorithmus (vgl. Dempster et al. (1977)).

Eine Annahme über Abhängigkeit bzw. Unabhängigkeit zwischen Kovariablen und Heterogenitätskomponente kann vermieden werden, wenn ein fixed-effect-Ansatz gewählt wird. Hier wird α als individuenspezifischer Parameter aufgefaßt, der mitzuschätzen ist. Allerdings tritt dabei das Problem inzidentieller Parameter auf (vgl. Neymann/Scott (1948)), da mit zunehmender Zahl der Individuen auch die Zahl der individuenspezifischen Parameter α wächst. Die $\beta's$ sind strukturelle Parameter, die für alle Individuen gelten. Bei der Untersuchung der asymptotischen Eigenschaften der ML-Schätzer geht die Zahl der individuenspezifischen Parameter mit der Zahl der Individuen gegen Unendlich. Das hat zur Folge, daß bei einer simultanen Schätzung von β und α auch die strukturellen Parameter β bei endlichem T nicht konsistent geschätzt werden können. Andersen (1973) zeigt dies für den Spezialfall $T = 2$ und nur einem strukturellen Parameter β. Läßt man auch T gegen

Unendlich gehen, so deuten die Ergebnisse von Haberman (1977) darauf hin, daß in diesem Fall β konsistent geschätzt werden kann. Ein exakter Nachweis für die hier vorliegende Situation steht allerdings noch aus. Eine gewöhnliche ML-Schätzung ist demnach nur dann sinnvoll, wenn T relativ groß ist. In diesem Fall kann sie im Prinzip je nach Spezifikation mit Programmen für Logit- oder Probitmodelle durchgeführt werden.

Im Mehr-Episoden-Fall kann man bei Verwendung des logistischen Ansatzes zu einer bedingten Maximum-Likelihood-Schätzung übergehen. Durch Konditionierung mit einer geeigneten suffizienten Statistik kann eine von individuenspezifischen Parametern unabhängige bedingte Likelihood gefunden und zur Schätzung der strukturellen Modellparameter β verwendet werden. Eine ausführliche Darstellung dieser Methode findet man bei Hamerle (1986). Darüber hinaus bietet der fixed-effect-Ansatz stets den Vorteil, daß keine Verteilung für die Heterogenitätskomponente spezifiziert werden muß.

Anhang

Programmpakete:

BMDP
Statistical Software
University of California Press
2223 Fulton Street
Berkeley, CA 94720
USA

GAUSS
L.E. Edlefsen, S.D. Jones
Applied Technical Systems
Kent, WA

GLAMOUR
Programmpaket zur Analyse verallgemeinerter linearer Modelle
Lehrstuhl für Statisik (Prof. Dr. Fahrmeir)
Universität Regensburg
Universitätsstr. 31
8400 Regensburg

GLIM
Generalized Linear Interactive Modelling
Numerical Algorithmus Group
7 Banbury Road
Oxford OX2 6NN
England

SAS
Statistical Analysis System
SAS Institute Inc.
Box 8000
Cary, NC 27511
USA

Andersen, E. B. (1973): Conditional inference and models for measuring. Kopenhagen.

Andersen, P. K., Gill, R. D. (1982): Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100–1120.

Andress, H. J. (1985): Multivariate Analyse von Verlaufsdaten. ZUMA–Methodentexte, Bd. 1, Mannheim.

Aranda-Ordaz, F. J. (1983): An extension of proportional-hazards-model for grouped data. Biometrics 39, 110–118.

Blossfeld, H. P., Hamerle, A., Mayer, K. U. (1986): Ereignisanalyse: Statistische Theorie und Anwendungen in den Wirtschafts- und Sozialwissenschaften. Frankfurt/Main.

Borgan, Ø. (1984): Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. Scand. J. Statistics 11, 1–16.

Breslow, N. E. (1974): Covariance analysis of censored survival data. Biometrics 30, 89–100.

Chamberlain, G. (1980): Analysis of covariance with qualitative data. Review of Economic Studies 47, 225–238.

Chiang, C. L. (1968): Introduction to stochastic processes in Biostatistics. New York.

Coleman, J. (1981): Longitudinal data analysis. New York.

Cox, D. R. (1972): Regression models and life tables (with discussion). J. R. Statist. Soc., B, 34, 187–220.

Cox, D. R. (1975): Partial likelihood. Biometrika 62, 269–275.

Cox, D. R., Oakes, D. (1984): Analysis of survival data. London.

David, H. A., Moeschberger, M. (1978): Theory of competing risks. London.

Dempster, A. P., Laird, N. Rubin, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39, 1–38.

Diekmann, A., Mitter, P. (1984): Methoden zur Analyse von Zeitverläufen. Stuttgart.

Dübler, H. (1988): Empirische Untersuchungen zur Dauer der Arbeitslosigkeit. Manuskript, Universität Regensburg.

Egle, F. (1979): Ansätze für eine systematische Beobachtung und Analyse der Arbeitslosigkeit. Beitr. AB 36, Institut für Arbeitsmarkt und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg.

Elandt-Johnson, R. C., Johnson, N. L. (1980): Survival methods and data analysis. New York.

Elbers, C., Ridder, G. (1982): True and spurious duration dependence: The identifiability of the proportional hazards model. Review of Economic Studies 49, 403–410.

Fahrmeir, L., Hamerle, A. (1984): Multivariate statistische Verfahren. Berlin.

Fahrmeir, L., Kaufmann, H. (1985): Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13, 342–368.

Flinn, C. J., Heckman, J. J. (1982): Models for the analysis of labor force dynamics. In: Basman, R., Rhodes, G. (eds): Advances in Econometrics. Bd. 1, Greenwich, Conn., 35–95.

Friedman, M. (1982): Piecewise exponential models for survival data with covariates. Ann. Statist. 10, 101–113.

Gross, A. J., Clark, V. A. (1975): Survival distributions: Reliability aspects in the Biomedical Sciences. New York.

Haberman, S. (1977): Maximum likelihood estimates in exponential response models. Ann. Statist. 5, 815–841.

Hamerle, A. (1984): Zur statistischen Analyse von Zeitverläufen. Regensburger Diskussionsbeitrag Nr.180, Universität Regensburg.

Hamerle, A. (1985): Regressionsmodelle für diskrete Verweildauern und Lebenszeiten. Zeitschrift für Operations Research, B, 243–260.

Hamerle, A. (1986): Regression analysis for discrete event history or failure time data. Statistical Papers 27, 207–225.

Hamerle, A. (1988): On the incorporation of left censored observations in analysis of survival or duration data. Preprint, Universität Konstanz.

Hamerle, A., Tutz, G. (1984): Zusammenhangsanalysen in Mehrdimensionalen Kontingenztabellen — das loglineare Modell. In: Fahrmeir, L., Hamerle, A. (Hrg.): Multivariate statistische Verfahren, Berlin, 473–574.

Hamerle, A., Kemény, P., Tutz, G. (1984): Kategoriale Regression. In: Fahrmeir, L., Hamerle, A. (Hrg.): Multivariate statistische Verfahren, Berlin, 211–256.

Heckman, J. J., Borjas, G. (1980): Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. Econometrica 47, 247–283.

Heckman, J. J., Singer, B. (1982): Population heterogeneity in demographic models. In: Land, K., Rogers, A. (eds): Multidimensional Mathematical Demography. New York.

Heckman, J. J., Singer, B. (1984a): Econometric duration analysis. Journal of Econometrics 24, 63–132.

Heckman, J. J., Singer, B. (1984b): A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.

Holford, T. R. (1976): Life tables with concomitant information. Biometrics 32, 587–598.

Holford, T. R. (1980): The analysis of rates and of survivorship using log-linear models. Biometrics 36, 299–305.

Holt, J. D. (1978): Competing risks analysis with special reference to matched pair experiments. Biometrika 61, 159–166.

Kalbfleisch, J. D., Prentice, R. L. (1973): Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267–278.

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

Kemény, P., Rothmeir, F., Hamerle, A. (1986): Explorative Variablenselektion und Anpassungstests bei Regressionsmodellen zur Analyse der stationären Aufenthaltsdauer nach Unfallverletzungen im Schulsport. EDV in Medizin und Biologie 16 (im Druck).

Kiefer, J., Wolfowitz, J.(1956): Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 27, 887–906.

Lagakos, S. W. (1979): General right censoring and its impact on the analysis of survival data. Biometrics 35, 139–156.

Laird, N., Olivier, D. (1981): Covariance analysis of censored survival data using log-linear analysis techniques. J. Amer. Statist. Assoc., 231–240

Lancaster, T. (1979): Econometric methods for the duration of unemployment. Econometrica 47, 939–956.

Lawless, J. F. (1982): Statistical models and methods for lifetime data. New York.

Lee, E. T. (1980): Statistical methods for survival data analysis. Belmont.

Lindsay, B. G. (1983a): The geometry of mixture likelihoods: A general theory. Annals of Statistics 11, 86–94.

Lindsay, B. G. (1983b): The geometry of mixture likelihoods, part II: The exponential family. Annals of Statistics 11, 783–792.

Mantel, N., Hankey, B. F. (1978): A logistic regression analysis of response-time data where the hazard function is time dependent. Comm. Statist. A7, 333-347.

Maritz, J. S. (1971): Empirical Bayes Methods. London.

McCullagh, P., Nelder, J.A. (1983): Generalized linear models. London.

Meindl, T. (1985): Loglineare Modelle und allgemeine Mehr-Episoden- und Mehr-Zustands-Modelle zur Analyse der Dauer der Arbeitslosigkeit. Diplomarbeit, Regensburg.

Meindl, T. (1988): Neuere Regressionsansätze zur Marktanalyse: Modelle für geordnete Kategorien und Verweildauern. Dissertation, Universität Regensburg.

Miller, R. G. (1981): Survival analysis. New York.

Nelson, W. (1982): Applied life data analysis. New York.

Neyman, J., Scott, E. L. (1948): Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

Newman, J. L., McCulloch, C. E. (1984): A hazard rate approach to the timing of births. Econometrica 52, 939–961.

Pierce, D., Steward, W., Kopecky, K. (1979): Distribution-free regression analysis of grouped survival data. Biometrics 35, 785–793.

Pregibon, D. (1980): Goodness of link tests for generalized linear models. Applied Statistics 29, 15–24.

Prentice, R. L., Breslow, N. E. (1978): Retrospective studies and failure time models. Biometrika 65, 153–158.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V.,Flourney, N.,Farewell, V. T., Breslow, N. E. (1978): The analysis of failure times in the presence of competing risks. Biometrics 34, 541–554.

Prentice, R. L., Gloeckler, L. A. (1978): Regression analysis of grouped survival data with application to bearst cancer data. Biometrics 34, 57–67.

Schedel, H. (1986): Einflußverschiedener Tumormerkmale des Mammakarzinoms auf das Metastasierungsverhalten und auf die Überlebensraten. Dissertation, Universität München.

Seal, H. L. (1977): Studies in the history of probability and statistics XXXV. Multiple decrements or competing risks. Biometrika 64, 429–439.

Schuhmacher, M. (1983): Analyse von Überlebenszeiten bei nichtproportionalen Hazardfunktionen. Habilitationsschrift, Heidelberg.

Thompson, W. A., Jr. (1977): On the treatment of grouped observations in life studies. Biometrics 33, 463–470.

Tibshirani, R., Ciampi, A. (1983): A family of proportional- and additive-hazards models for survival data. Biometrics 39, 141–147.

Trussel, J., Richards, T. (1985): Correcting for unmeasured heterogeneity in hazard models using the Heckman–Singer procedure. In: Tuma; N. B. (ed.): Sociological Methodology, San Francisco, 242–276.

Tuma, N. B. (1982): Nonparametric and partially parametric approaches to event-history analysis. In: Leinhardt, S. (ed): Sociological Methodology. San Francisco.

Tuma, N. B., Hannan, M. T., Groeneveld, L. P. (1979): Dynamic analysis of event histories. American Journal of Sociology 84, 820–854.

Tuma, N. B., Hannan, M. T. (1984): Social dynamics: Models and methods. New York.

Tutz, G. (1985): Regressionsmodelle mit ordinalen Reaktionsvariablen. Regensburger Diskussionsbeitrag Nr. 184, Universität Regensburg.

Tutz, G. (1987): Regression models with an ordered categorical response. Paper presented at the Second Workshop of Statistical Modelling, Perugia 6–10th of July.

Tutz, G. (1988): Modelle für kategoriale Daten mit ordinalem Skalenniveau—parametrische und nonparametrische Ansätze. Habilitationsschrift Universität Regensburg.