Skripte zur Methodenlehre, V

Einführung in die statistische Analyse von Zustandsverläufen

G. Rohwer
U. Pötter

Version 2 Februar 2000

Vorbemerkung. Bei den „Skripten zur Methodenlehre“ handelt es sich um Texte, die als Leitfäden für Seminare zur sozialwissenschaftlichen Methodenlehre dienen sollen. Das vorliegende Skript beschäftigt sich mit statistischen Methoden zur Analyse von Längsschnittdaten. Dabei orientieren wir uns an Anwendungen dieser Methoden in der empirischen Sozialforschung, bei denen es in erster Linie um eine Untersuchung von Lebensverläufen geht und Daten dementsprechend in Gestalt von Zustandsverläufen gegeben sind. Darauf bezieht sich die im ersten Kapitel eingeführte Terminologie.

Der Text enthält zahlreiche Übungsaufgaben, die während der Bearbeitung des Stoffes gelöst werden sollten. Die meisten Aufgaben können mit Bleistift und Papier gelöst werden; für einige Aufgaben ist die Verwendung eines Taschenrechners hilfreich.

Für Anwendungen der Methoden in der empirischen Sozialforschung, bei denen man es meist mit größeren Datensätzen zu tun hat, muß man allerdings Computer und geeignete Statistikprogramme verwenden. Der Text enthält deshalb einen Anhang, anhand dessen man das Programm TDA kennenlernen kann, mit dem die meisten Fragestellungen der Verlaufsdatenanalyse bearbeitet werden können. Mit den Aufgaben dieses Anhangs kann man sich entweder parallel zur Behandlung des Haupttextes oder in einem sich anschließenden Workshop beschäftigen.

Über die hier behandelten statistischen Methoden gibt es eine sehr umfangreiche Literatur. Wer sein Wissen über die statistischen Aspekte der Methoden vertiefen möchte, sei auf Lawless (1982) und Cox und Oakes (1984) hingewiesen. Für eine weiterführende Diskussion von Anwendungen in der empirischen Sozialstrukturforschung sei auf Blossfeld und Rohwer (1995) verwiesen.

mete.tex Februar 2000
Homepage: `http://www.stat.ruhr-uni-bochum.de [lecture notes]`

Inhaltsverzeichnis

Kapitel 1

Einführung

In diesem Kapitel besprechen wir Grundzüge des begrifflichen Rahmens, der in den nachfolgenden Kapiteln vorausgesetzt wird.

1.1 Objekte und Lebensverläufe

Wir beziehen uns zunächst ganz allgemein auf Objekte. Jedes Objekt existiert in der Form eines Lebensverlaufs: Es wird geboren, dann macht es einen gewissen Entwicklungsprozeß durch, und schließlich stirbt es. Unser Ziel ist es, uns mit einigen statistischen Begriffen und Modellen zu beschäftigen, die vorgeschlagen worden sind, um Lebensverläufe beschreiben und über ihre Entwicklung nachdenken zu können.

Wie wir sehen werden, sind diese Begriffe und Modelle sehr allgemein. Bei ihrer Verwendung in der empirischen Sozialforschung ist darauf zu achten, daß wir es dann meistens mit spezifischen Objekten zu tun haben, nämlich sozialen Akteuren (sowohl individuelle als auch korporative Akteure), die selbst Anteil daran nehmen, wie sich ihre Lebensverläufe entwickeln. Wir werden die Objekte, mit denen wir uns beschäftigen, in allgemeiner Weise als *Individuen* bezeichnen.

1.2 Verhalten und Zustände

Man kann Individuen unter zwei komplementären Betrachtungsweisen vergegenständlichen: als Objekte, die sich *verhalten* können, und als Objekte, die sich *in wechselnden Zuständen befinden* können. Der in diesem Text behandelte Ansatz geht von der zweiten Betrachtungsweise aus: Lebensverläufe von Individuen werden als Folgen von Zuständen konzipiert.

1.3 Der Zustandsraum

Ausgangspunkt ist also die Konzeption eines Zustandsraum. Wir setzen voraus, daß es stets nur eine endliche Menge möglicher Zustände gibt und bezeichnen den Zustandsraum mit dem Symbol $\tilde{Y}$. Der Lebensverlauf eines Individuums besteht dann in einer Folge von Zuständen aus dem vorgegebenen Zustandsraum. Die Aufenthaltsdauer in den Zuständen ist unbestimmt, und es wird auch nicht vorausgesetz, daß alle Zustände durchlaufen werden müssen. Aus diesem Ansatz folgt, daß der hier verwendete Begriff des Lebensverlaufs wesentlich davon abhängt, welcher Zustandsraum vorausgesetzt wird.

Man beachte, daß ein Zustandsraum *eindeutig* sein muß. Damit ist gemeint, daß sich die zu betrachtenden Individuen zu jedem Zeitpunkt in genau einem der möglichen Zustände befinden müssen.

Wir sprechen von einem *vollständigen Zustandsraum*, wenn der Zustandsraum insbesondere die beiden Quasi-Zustände *noch nicht geboren* und *gestorben* umfaßt. Um Lebensverläufe vollständig zu erfassen, ist ein vollständiger Zustandsraum erforderlich.

1.4 Biographieschema

Unter einem *Biographieschema* verstehen wir die Festlegung einer Menge möglicher (ggf. unvollständiger) Lebensverläufe in einem Zustandsraum. Ein Biographieschema kann graphisch durch ein *Zustandsdiagramm* veranschaulicht werden. Es besteht dann aus einem gerichteten Graphen, in dem die möglichen Zustände durch Knoten, die möglichen Übergänge durch gerichtete Kanten repräsentiert werden.

AUFGABE 1.1 Konzipieren Sie einen vollständigen Zustandsraum für die Erfassung von Erwerbsverläufen, der die folgenden Zustände unterscheidet: (1) erwerbstätig, (2) arbeitslos, (3) weder erwerbstätig noch arbeitslos.

AUFGABE 1.2 Konzipieren Sie mit dem Zustandsraum aus Aufgabe 1.1 ein Biographieschema.

1.5 Mehrdimensionale Zustandsräume

Die Konzeption eines Zustandsraum muß durch den Modellkonstrukteur vorgegeben werden. Dies hängt davon ab, welche Aspekte realer Lebensverläufe erfaßt werden sollen, z.B. Erwerbsverläufe oder Ausbildungsverläufe oder Beziehungsverläufe. Man kann mehrere solcher Aspekte durch einen mehrdimensionalen Zustandsraum repräsentieren. Als symbolische Form eines m-dimensionalen Zustandsraum hat man dann

$$\tilde{Y} = \tilde{Y}_1 \times \cdots \times \tilde{Y}_m$$

Andererseits ist es möglich, stattdessen einen einfachen (eindimensionalen) Zustandsraum zu verwenden, bei dem jede mögliche Kombination von Zuständen in $\tilde{Y}_1, \ldots, \tilde{Y}_m$ als ein gesonderter Zustand im kombinierten Zustandsraum $\tilde{Y}$ repräsentiert wird.

AUFGABE 1.3 Konzipieren Sie einen vollständigen Zustandsraum für die Zustände: (1) unverheiratet, (2) verheiratet. Bilden Sie dann aus diesem und dem in Aufgabe 1.1 konzipierten Zustandsraum einen zweidimensionalen Zustandsraum.

AUFGABE 1.4 Konzipieren Sie für den zweidimensionalen Zustandsraum aus Aufgabe 1.3 ein Biographieschema.

1.6 Die Zeitachse

Die Grundvorstellung besteht darin, Lebensverläufe als ein zeitlich geordnetes „Durchwandern" von Zustandsräumen aufzufassen. Es ist also erforderlich, sich explizit auf eine Zeitachse zu beziehen. Hierfür gibt es zwei Möglichkeiten.

- Wir können uns eine Zeitachse als eine Folge von Zeitstellen vorstellen, z.B. Stunden, Tage, Wochen oder Monate. Man spricht dann von einer *diskreten Zeitachse*, und zur numerischen Repräsentation können die natürlichen Zahlen verwendet werden.

- Wir können uns eine Zeitachse als einen kontinuierlichen Zeitfluß vorstellen, d.h. von der Annahme ausgehen, daß Zeitstellen beliebig teilbar sind. Man spricht dann von einer *kontinuierlichen* oder *stetigen Zeitachse* und verwendet zur numerischen Repräsentation die reellen Zahlen.

Wir werden zunächst von einer diskreten Zeitachse ausgehen. Dies hat den Vorteil, daß von einer *Folge* von Zeitstellen gesprochen werden kann. Statistische Modelle verwenden jedoch häufig eine kontinuierliche Zeitachse, so daß wir uns später auch dieser Vorstellung bedienen werden.

1.7 Ereignisse als Zustandswechsel

Wir haben bisher Lebensverläufe als Folgen von Zuständen betrachtet, wobei die Aufenthaltsdauer in jedem der möglichen Zustände von unterschiedlicher Dauer sein kann. Stattdessen kann man das Augenmerk auch auf die Zustandswechsel richten, also auf die Übergänge von einem

Box 1.1 Datensatz 1

ID	Geburt	Beginn des Studiums	Ende des Studiums
1	1970	1990	1995
2	1975	1994	1999
3	1973	1991	1996
4	1970	1989	1995
5	1975	1993	1999
6	1973	1993	1996
7	1970	1988	1995
8	1975	1995	1999
9	1973	1992	1997

gegebenen in einen neuen Zustand. Diese Zustandswechsel werden auch *Ereignisse* genannt.[1]

AUFGABE 1.5 Geben Sie eine Liste aller Ereignisse an, die in dem Biographieschema, das in Aufgabe 1.2 konzipiert wurde, möglich sind.

1.8 Kalenderzeit und Prozeßzeit

Wenn man sich auf reale Individuen und deren Lebensverläufe beziehen will, muß man zunächst immer von einer Kalenderzeitachse ausgehen. Man spricht gelegentlich auch von einer historischen Zeitachse. Für die Modellbildung verwendet man stattdessen meistens eine *Prozeßzeitachse*. Es handelt sich um eine Zeitachse, bei der der Nullpunkt durch das Eintreten eines Ereignisses definiert wird. Zum Beispiel könnte man eine Prozeßzeitachse konzipieren, die mit der Geburt beginnt oder mit der Aufnahme eines Studiums oder dem Beginn einer Eheschließung.

AUFGABE 1.6 Betrachten Sie die Daten in Box 1.1. Konzipieren Sie dazu einen Zustandsraum und ein Biographieschema. Stellen Sie die Daten auf einer Prozeßzeitachse dar, deren Zeiteinheiten Jahre sind und die mit dem Beginn des Studiums beginnt.

[1]Es sei angemerkt, daß das Wort 'Ereignis' dadurch eine spezifische Bedeutung bekommt. Wer sich für eine gründlichere Diskussion interessiert, sei auf Galton (1994) verwiesen.

1.9 Verlaufsdiagramme

Ein *Verlaufsdiagramm* ist ein Diagramm, bei dem die horizontale Achse die Zeitachse und die vertikale Achse den Zustandsraum repräsentiert. Dabei kann die Zeitachse entweder eine Kalenderzeitachse oder eine Prozeßzeitachse sein. Solche Diagramme sind oft nützlich, um exemplarisch einzelne oder auch mehrere Verläufe darzustellen.

AUFGABE 1.7 Stellen Sie die ersten drei Verläufe aus dem Datensatz in Box 1.1 zunächst in einem Verlaufsdiagramm dar, bei dem die Zeitachse eine Kalenderzeitachse ist, dann in einem Verlaufsdiagramm, bei dem die Zeitachse die Prozeßzeitachse ist, die mit dem Beginn des Studiums beginnt.

1.10 Kohorten

In der empirischen Sozialforschung wird oft der Begriff *Kohorte* verwendet, um eine Menge von Individuen zu bezeichnen, die ein Ereignis eines bestimmten Typs in der gleichen Kalenderzeitstelle erfahren haben. Zum Beispiel bilden alle Individuen, die im Jahr 1970 geboren worden sind, eine Geburtskohorte. Dabei muß natürlich angegeben werden, auf welche Grundgesamtheit von Individuen man sich beziehen möchte. Und außerdem muß die Dauer der Zeitstelle fixiert werden, die zur Definition von Kohorten dienen soll.

AUFGABE 1.8 Betrachten sie die Daten in Box 1.1. Wieviel Geburtskohorten gibt es? Erstellen Sie eine Tabelle, in der die Individuen den Geburtskohorten zugeordnet werden. Machen Sie dann das gleiche für die Kohorten von Studienanfängern.

Kapitel 2

Statistische Beschreibungen

In diesem Kapitel beginnen wir mit einer Diskussion der Frage, wie Lebensverläufe beschrieben werden können. Zwei komplementäre Betrachtungsweisen können eingenommen werden. Man kann versuchen, Lebensverläufe spezifischer Individuen ins Auge zu fassen und in ihrer jeweils einmaligen Entwicklung zu beschreiben. Andererseits kann man eine *vergleichende Betrachtungsweise* einnehmen. Dies setzt voraus, daß man sich auf eine Mehrzahl vergleichbarer Lebensverläufe beziehen kann. Vergleichbarkeit ist allerdings kein Merkmal, das Lebensverläufen "an und für sich" zukommt oder nicht zukommt, sondern Gesichtspunkte für einen Vergleich kommen stets durch den Sozialforscher zustande. Er ist es, der Lebensverläufe vergleichen möchte und dafür die ihm wichtig erscheinenden Gesichtspunkte definiert.

Für den hier zu behandelnden statistischen Ansatz kommen die Gesichtspunkte für einen Vergleich von Lebensverläufen durch die Definition eines Biographieschemas zustande. Wir nehmen im folgenden an, daß ein Biographieschema definiert worden ist und daß man sich auf eine vorgegebene Menge von Individuen beziehen kann, deren Lebensverläufe (meistens nur ausschnitthaft) durch das vorgegebene Biographieschema verglichen werden können. Wir bezeichnen diese Menge von Individuen mit dem Symbol Ω.

Diese Voraussetzungen erlauben es, Lebensverläufe mit statistischen Begriffen zu beschreiben. Was damit gemeint ist, wird sogleich deutlicher werden, wenn wir die beiden Grundbegriffe, *statistische Variable* und *statistische Verteilung*, eingeführt haben.

2.1 Statistische Variablen

Eine statistische Variable ist eine Abbildung (auch Funktion genannt), die jedem Individuum aus einer vorgegebenen Menge einen bestimmten Wert in einem Merkmalsraum zuordnet. Zur symbolischen Repräsentation verwenden wir die Schreibweise

$$X : \Omega \longrightarrow \tilde{X}$$

Hier ist X eine statistische Variable, die jedem Individuum $\omega \in \Omega$ einen Merkmalswert $X(\omega)$ aus dem Merkmalsraum $\tilde{X}$ zuordnet. Wir setzen voraus, daß es für den Merkmalsraum eine numerische Repräsentation gibt. In dieser Einführung betrachten wir zwei Arten numerischer Repräsentationen. Wenn $\tilde{X}$ durch eine Teilmenge der natürlichen Zahlen repräsentiert werden kann, nennen wir X eine *diskrete Variable*. Wenn $\tilde{X}$ durch einen zusammenhängenden Teilbereich der reellen Zahlen repräsentiert werden kann, nennen wir X eine *kontinuierliche Variable*. Variablen können außerdem danach unterschieden werden, ob es sich um einen qualitativen, ordinalen oder quantitativen Merkmalsraum handelt. Eine diskrete numerische Repräsentation kann für alle drei Arten von Variablen verwendet werden, eine kontinuierliche numerische Repräsentation ist im allgemeinen nur bei quantitativen Variablen sinnvoll.

2.2 Zustandsvariablen

Der Begriff der statistischen Variablen kann nun verwendet werden, um Lebensverläufe zu repräsentieren. Vorausgesetzt wird ein Biographieschema, also insbesondere ein Zustandsraum $\tilde{Y}$ und eine Zeitachse $\tilde{T}$, die zunächst als eine diskrete Prozeßzeitachse angenommen wird, also

$$\tilde{T} = \{\, 0, 1, 2, 3, \dots \}$$

Weiterhin wird eine endliche Menge von Individuen, Ω, vorausgesetzt. Dann können die Zustände, in denen sich die Individuen befinden, durch *statistische Zustandsvariablen* erfaßt werden. Für jede Zeitstelle $t \in \tilde{T}$ gibt es eine Zustandsvariable

$$Y_t : \Omega \longrightarrow \tilde{Y}$$

$Y_t(\omega)$ ist der Zustand, in dem sich das Individuum $\omega \in \Omega$ in der Zeitstelle t befindet. Der Lebensverlauf jedes Individuums ist dann durch eine Folge von Zuständen:

$$(Y_0(\omega), Y_1(\omega), Y_2(\omega), \dots)$$

gegeben. Da wir angenommen haben, daß Zustandsräume stets nur eine endliche Anzahl unterschiedlicher Zustände enthalten, handelt es sich bei Zustandsvariablen stets um diskrete Variablen.

2.3 Partielle Lebensverläufe

Die Idee, Lebensverläufe durch Folgen von Zuständen zu repräsentieren, bereitet dann keine Schwierigkeiten, wenn es sich um vollständige Lebensverläufe handelt. Jeder Lebensverlauf mündet dann in einem Endzustand, in dem Quasi-Zustand *gestorben*. In der empirischen Sozialforschung werden jedoch meistens nur partielle Lebensverläufe untersucht. Man muß dann festlegen, welchen Teil von Lebensverläufen man betrachten möchte. Dafür gibt es zwei Möglichkeiten. In beiden Fällen beginnt man mit einem Anfangsereignis, dessen Eintritt den Beginn des partiellen Lebensverlaufs markiert; zum Beispiel: Geburt eines Individuums, Beginn eines Studiums, Eintritt in das Erwerbsleben. Dies erlaubt es, eine entsprechende Prozeßzeitachse zu definieren. Um die Entwicklung partieller Lebensverläufe auf dieser Prozeßzeitachse zu erfassen und zu vergleichen, gibt es dann zwei Möglichkeiten.

a) Man kann einen festen Zeitraum fixieren; zum Beispiel die ersten 20 Jahre seit der Geburt, oder 6 Jahre seit dem Beginn eines Studiums. Das heißt, man fixiert auf der vorgegebenen Zeitachse eine maximale Zeitstelle t^* und erhält dann für alle Individuen aus Ω partielle Lebensverläufe gleicher Länge, nämlich

$$(Y_0(\omega), Y_1(\omega), Y_2(\omega), \dots , Y_{t^*}(\omega))$$

Hierbei muß natürlich ein geeigneter Zustandsraum vorausgesetzt werden, der es erlaubt, alle Lebensverläufe für die vorgegebene Zeitspanne zu definieren.

b) Eine andere Möglichkeit besteht darin, daß man die partiellen Lebensverläufe enden läßt, wenn eines aus einer vorgegebenen Menge möglicher Ereignisse eintritt. Statistiker sprechen dann manchmal von einem „absorbierenden Endzustand“, der durch das Eintreten eines solchen Ereignisses erreicht wird. Analog kann man von „absorbierenden Endereignissen“ sprechen, die einen partiellen Lebensverlauf beenden. Um eine Menge absorbierender Endzustände zu fixieren, verwenden wir das Symbol $\tilde{Y}^*$. Es muß gelten, daß $\tilde{Y}^*$ eine Teilmenge des Zustandsraums ist, also $\tilde{Y}^* \subset \tilde{Y}$. Jeder individuelle Lebensverlauf wird dann so lange erfaßt, bis zum ersten Mal ein Zustand in $\tilde{Y}^*$ erreicht wird.

In der empirischen Sozialforschung wird hauptsächlich die zweite Herangehensweise verwendet. Sie hat zur Folge, daß die individuellen (partiellen) Lebensverläufe im allgemeinen eine unterschiedliche zeitliche Ausdehnung bekommen. Einige Individuen erreichen einen absorbierenden Endzustand schon nach kurzer Zeit, andere brauchen dafür länger.

2.4 Statistische Verteilungen

Grundlegend für statistische Beschreibungen ist der Begriff der statistischen Verteilung. Vorausgesetzt wird, daß man sich auf eine statistische Variable beziehen kann, also auf ein Kollektiv Ω und eine Abbildung X, die jedem Mitglied des Kollektivs einen Wert in einem Merkmalsraum, $\tilde{X}$, zuordnet. Die Idee ist, daß man sich bei einer statistischen Beschreibung nicht für die jeweils individuellen Merkmalswerte der Mitglieder des Kollektivs interessiert, sondern nur dafür, wie sich die Mitglieder auf die möglichen Merkmalswerte verteilen. Diese Betrachtungsweise kommt gut in folgenden Worten der "Declaration on Professional Ethics" zum Ausdruck, die vom *International Statistical Institute* erstellt worden ist:

> "Statistical data are unconcerned with individual identities. They are collected to answer questions such as 'how many?' or 'what proportions?', not 'who?'. The identities and records of cooperating (or non-cooperating) subjects should therefore be kept confidential, whether or not confidentiality has been explicitly pledged."[1]

Eine statistische Verteilung wird deshalb als eine Funktion

$$\mathrm{P} : \mathcal{A}(\tilde{X}) \longrightarrow [0,1]$$

definiert. $\mathcal{A}(\tilde{X})$ ist eine Menge von Teilmengen des Merkmalsraums $\tilde{X}$. Dabei wird üblicherweise vorausgesetzt, daß es sich um eine Mengenalgebra handelt, die bezüglich der mengentheoretischen Basisoperationen (Vereinigung, Durchschnitt und Komplement) abgeschlossen ist. Die Elemente von $\mathcal{A}(\tilde{X})$ werden wir *Merkmalsmengen* nennen. Die Funktion P kann dann folgendermaßen spezifiziert werden: Sie soll für jede Merkmalsmenge $\tilde{x} \in \mathcal{A}(\tilde{X})$ den Anteil der Mitglieder von Ω angeben, deren Merkmalswerte in dieser Merkmalsmenge liegen. Also in einer expliziten Definition:

$$\mathrm{P}(\tilde{x}) := |\{\omega \in \Omega \,|\, X(\omega) \in \tilde{x}\}| \,/\, |\Omega|$$

Es ist erkennbar, wie durch diese Definition eine Bezugnahme auf individuelle Mitglieder von Ω verschwindet und es nur noch darauf ankommt,

[1] International Statistical Institute 1986, S. 238.

wieviele Mitglieder an den jeweiligen Merkmalsmengen teilhaben.

Um uns flexibler auf Merkmalsmengen beziehen zu können, werden wir auch noch einige abkürzende Schreibweisen verwenden; insbesondere die folgenden:

$$\mathrm{P}(X \in \tilde{x}) := \mathrm{P}(\tilde{x})$$

$$\mathrm{P}(X = x) := \mathrm{P}(\{x\})$$

Bei quantitativen Variablen wird auch noch die Schreibweise

$$\mathrm{P}(X \leq x) := \mathrm{P}(\{\omega \in \Omega \,|\, X(\omega) \leq x\}$$

verwendet und im allgemeinen als *(kumulative) Verteilungsfunktion* von X bezeichnet. Die meistens verwendete Symbolik is

$$F(x) := \mathrm{P}(X \leq x)$$

Aufgabe 2.1 Es sei Ω ein Kollektiv mit 10 Mitgliedern und es gebe die folgenden Merkmalswerte einer Variablen X:

$$3, 2, 3, 1, 4, 3, 1, 3, 4, 2$$

(a) Geben sie den Merkmalsraum an. (b) Definieren Sie eine Algebra von Merkmalsmengen durch die Potenzmenge des Merkmalsraums. (c) Berechnen Sie die statistische Verteilung der Variablen X und geben Sie das Resultat für alle möglichen Merkmalsmengen in einer Tabelle an. (d) Nehmen Sie an, daß es sich um eine quantitative Variable handelt. Berechnen Sie dann die Verteilungsfunktion der Variablen und geben Sie das Resultat in einer Tabelle an.

Aufgabe 2.2 Zeigen Sie, daß die Verteilungsfunktion P additiv ist, d.h. daß folgendes gilt: Wenn $\tilde{x}_1$ und $\tilde{x}_2$ zwei disjunkte Merkmalsmengen sind, dann gilt

$$\mathrm{P}(\tilde{x}_1 \cup \tilde{x}_2) = \mathrm{P}(\tilde{x}_1) + \mathrm{P}(\tilde{x}_2)$$

2.5 Zustandsverteilungen

Zu überlegen ist, wie statistische Beschreibungen von Lebensverläufen entwickelt werden können. Das kann man zunächst auf ganz einfache

Box 2.1 Datensatz 2

```
ID   t = 1  2  3  4  5  6
-------------------------
 1       0  0  1  1  0  0
 2       1  0  0  0  1  1
 3       1  1  0  0  0  0
 4       0  0  0  1  1  1
 5       0  1  1  1  0  0
 6       1  1  0  0  1  1
```

Weise dadurch machen, daß man sich auf die Zustandsvariablen Y_t bezieht, die in Abschnitt 2.2 zur Repräsentation von Lebensverläufen eingeführt worden sind. D.h. man kann für jede Zeitstelle $t \in \tilde{T}$ die statistische Verteilung der Zustandsvariablen Y_t berechnen. Bezieht man sich nur auf eine einzige Zeitstelle, spricht man von einer *Querschnittsverteilung*. Eine Querschnittsverteilung ergibt natürlich noch kein Bild der Entwicklung von Lebensverläufen. Eine Möglichkeit, hier weiterzukommen, besteht darin, die Querschnittsverteilungen für alle Zeitstellen der Zeitachse zu berechnen. Wir sprechen dann von *diachronen Zustandsverteilungen*. Man kann das Ergebnis in einer Tabelle oder in einem Schaubild darstellen.

AUFGABE 2.3 Betrachten Sie die Daten in Box 2.1. Es handelt sich um Erwerbsverläufe bei 6 Individuen. Es gibt zwei Zustände: 1 = erwerbstätig, 0 = nicht erwerbstätig. Berechnen sie die diachrone Zustandsverteilung und stellen Sie diese Verteilung (a) in einer Tabelle und (b) in einem Schaubild dar.

Problematik. Diachrone Zustandsverteilungen liefern sinnvolle statistische Beschreibungen, wenn es sich um nicht wiederholbare Zustände handelt. Zum Beispiel: 0 = noch nie verheiratet gewesen, 1 = verheiratet oder mindestens einmal verheiratet gewesen. Wenn es sich jedoch um wiederholbare Zustände handelt, wie z.B. bei Erwerbsverläufen, können diachrone Zustandsverteilungen irreführend werden, weil sie keine Rückschlüsse auf die individuellen Verläufe gestatten.

AUFGABE 2.4 Konstruieren Sie ein Beispiel, um diese Problematik sichtbar zu machen. Es soll zwei Zustände geben: 1 = arbeitslos, 0 = nicht arbeitslos. Konstruieren Sie dann zwei Varianten für 6 individuelle Verläufe, so daß der Anteil der arbeitslosen Personen in jeder Zeitstelle 1/3 beträgt. Bei der ersten Variante sollen 2 Personen immer, 4 Personen nie

arbeitslos sein. Bei der zweiten Variante sollen alle Personen gleichmäßig von Arbeitslosigkeit betroffen sein.

Kapitel 3

Verweildauerverteilungen

In diesem Kapitel werden einige Begriffe diskutiert, die dazu dienen können, die Verweildauern in den durch ein Biographieschema vorgegebenen Zuständen statistisch darzustellen. Soweit wir uns dabei auf Daten beziehen, wird angenommen, daß vollständige Beobachtungen verfügbar sind. Die Problematik unvollständiger (zensierter) Beobachtungen wird im nächsten Teil behandelt.

3.1 Episoden

Gegeben ein Biographieschema, stellen wir uns einen Lebensverlauf als ein sequentielles Durchwandern des zugehörigen Zustandsraums vor. Ein Individuum beginnt in einem gewissen Zustand und hält sich eine mehr oder weniger lange Zeit in diesem Zustand auf, dann wechselt es in einen neuen Zustand und hält sich in diesem neuen Zustand mehr oder weniger lange auf, usw. Wir können uns einen Lebensverlauf also auch als eine Folge von *Episoden* vorstellen, d.h. Aufenthaltsdauern in einem gegebenen Zustand bis ein Wechsel in einen neuen Zustand erfolgt. Eine einzelne Episode läßt sich durch vier Angaben charakterisieren:

- durch einen *Anfangszustand*, mit dessen Auftreten die Episode beginnt;
- durch einen *Endzustand*, oder *Folgezustand*, mit dessen Auftreten die Episode beendet wird;
- durch eine *Anfangszeitstelle*, die angibt, wann der Anfangszustand zum erstenmal eingenommen wird; und
- durch eine *Endzeitstelle*, die angibt, wann der Endzustand zum erstenmal eingenommen wird.

Der Begriff der Episode (verwendet wird auch gelegentlich das englische Wort *Spell*) erlaubt es, ein allgemeines Schema für die Representation von Lebensverlaufsdaten zu definieren. Box 3.1 illustriert dies Schema anhand von vier Verläufen. Der Zustandsraum umfaßt vier Zustände; 1 ist der Anfangszustand, 4 ist der (absorbierende) Endzustand. Jede Zeile

Box 3.1 Schema für Episodendaten (Datensatz 3)

ID	SN	ORG	DES	TS	TF
1	1	1	2	0	10
1	2	2	3	10	15
1	3	3	4	15	20
2	1	1	4	0	15
3	1	1	3	0	16
3	2	3	4	19	18
4	1	1	2	0	6
4	2	2	3	6	11
4	3	3	2	11	17
4	4	2	4	17	23

in dem Schema bezieht sich auf eine Episode, und für jedes Individuum gibt es also so viele Zeilen, wie ihr Lebensverlauf Episoden aufweist. Die die Spalten benennenden Abkürzungen sind folgendermaßen zu verstehen:

- `ID` ist die Identifikationsnummer der Individuen,
- `SN` ist die laufende Nummer der Episode,
- `ORG` ist der Anfangszustand der Episode,
- `DES` ist der Endzustand der Episode,
- `TS` ist die Anfangszeitstelle der Episode,
- `TF` ist die Endzeitstelle der Episode.

Wir werden ein solches Schema ein *Episodendatenschema* nennen.

Aufgabe 3.1 Konstruieren Sie für den Datensatz 1 (Box 1.1) zunächst ein Biographieschema und stellen Sie die Daten dann in einem Episodendatenschema dar.

Aufgabe 3.2 Konstruieren Sie für den Datensatz 2 (Box 2.1) zunächst ein vollständiges Biographieschema und stellen Sie die Daten dann in einem Episodendatenschema dar.

3.2 Statistischer Begriffsrahmen

Wir setzen jetzt die in Abschnitt 2 begonnene Diskussion fort, wie Lebensverläufe statistisch beschrieben werden können. Die Idee, die wir im

weiteren verfolgen, besteht darin, sich zunächst auf einzelne Episoden zu konzentrieren, genauer gesagt, auf die Gesamtheit der Episoden, die in einem bestimmten, der Beschreibung vorausgesetzten Anfangszustand beginnen. Wir setzen außerdem voraus, daß wir diese Episoden auf einer Prozeßzeitachse beschreiben wollen, die mit dem Eintritt des Anfangszustands beginnt. Die Gesamtheit der Episoden, auf die wir uns beziehen wollen, kann dann durch eine zweidimensionale statistische Variable

$$(T, D)$$

repräsentiert werden. T erfaßt die Zeitdauer der Episode, d.h. die Verweildauer im Ausgangszustand, und D erfaßt den Folgezustand, dessen Eintreten die Epiosde abschließt.

Es ist klar, daß sich die Darstellung vereinfacht, wenn eine Episode in nur einem möglichen Folgezustand enden kann. Dann kann D nur einen möglichen Wert annehmen und braucht nicht explizit erfaßt zu werden. Oder anders gesagt, eine Episode wird dann vollständig durch ihre Dauer, den Wert von T, charakterisiert.

3.3 Ein möglicher Folgezustand

Wenn es nur einen möglichen Folgezustand gibt, genügt es, die Verweildauervariable T zu betrachten. Eine statistische Beschreibung zielt dann darauf, die statistische Verteilung dieser Verweildauervariablen zu ermitteln und darzustellen. Die begrifflichen Hilfsmittel hängen davon ab, ob man sich die Zeitachse als diskret oder stetig vorstellen will. In beiden Fällen können wir die Verteilung durch eine (kumulative) Verteilungsfunktion

$$F(t) = \mathrm{P}(T \le t)$$

charakterisieren. Ebenfalls unabhängig von der Art der Zeitachse kann man einen weiteren in der Verweildaueranalyse oft verwendeten Begriff definieren, die *Survivorfunktion*. Sie ergibt sich unmittelbar aus der Verteilungsfunktion durch die Definition

$$G(t) = 1 - F(t)$$

Eine Unterscheidung wird allerdings erforderlich, wenn wir uns auf eine zeitstellenbezogene Ereignisdichte beziehen wollen. Im diskreten Fall kann man dann eine diskrete Dichtefunktion

$$f(t) = \mathrm{P}(T = t)$$

verwenden. Im stetigen Fall wird der Ausdruck $\mathrm{P}(T = t)$ problematisch, und es ist zweckmäßig, zunächst von Zeitintervallen auszugehen, also Ausdrücken der Art

$$\mathrm{P}(t \le T < t + \Delta)$$

wobei Δ die Dauer des Zeitintervals angibt, das an der Stelle t beginnt. Es ist klar, daß der Wert eines solchen Ausdrucks von Δ abhängt, und man definiert deshalb die Ereignisdichte *pro Zeiteinheit* durch

$$f(t) = \lim_{\Delta \to 0} \frac{\mathrm{P}(t \le T < t + \Delta)}{\Delta}$$

Schließlich ist die Unterscheidung auch noch für den Begriff der *Übergangsrate* relevant, der in vielen Ansätzen der Verweildaueranalyse eine zentrale Rolle spielt. Die Idee ist, eine zeitstellenbezogene Ereignisdichte *unter der Bedingung* zu betrachten, daß das Ereignis noch nicht eingetreten ist. Im diskreten Fall lautet die Definition

$$r(t) = \mathrm{P}(T = t \,|\, T \ge t)$$

Im stetigen Fall verwendet man die Definition

$$r(t) = \lim_{\Delta \to 0} \frac{\mathrm{P}(t \le T < t + \Delta \,|\, T \ge t)}{\Delta}$$

Aufgabe 3.3 Zeigen Sie zunächst für den diskreten, dann für den stetigen Fall, daß die Begriffe 'Verteilungsfunktion', 'Survivorfunktion', 'Dichtefunktion' und 'Übergangsrate' äquivalent sind, d.h. daß sie wechselseitig auseinander abgeleitet werden können. Zeigen Sie insbesondere, daß folgende Zusammenhänge gelten. Im diskreten Fall:

$$r(t) = f(t)/G(t-1)$$

und

$$G(t) = \prod_{\tau=1}^{t} (1 - r(\tau))$$

Und im stetigen Fall:

$$r(t) = f(t)/G(t)$$

und

$$G(t) = \exp\left\{ - \int_0^t r(\tau)\, d\tau \right\}$$

Aufgabe 3.4 Berechnen Sie mit dem Datensatz 1 (Box 1.1) die diskrete Übergangsrate für die Beendigung des Studiums.

Aufgabe 3.5 Betrachten Sie im Datensatz 2 (Box 2.1) zwei Gruppen von Episoden: Episoden, die im Zustand 0 beginnen, und Episoden, die im Zustand 1 beginnen. Verwenden Sie nur die nicht-zensierten Episoden, d.h. diejenigen Episoden, für die aus dem Datenbestand erkennbar ist, daß sie durch den Übergang in einen neuen Zustand beendet werden. Berechnen Sie dann die Übergangsraten für den Übergang in den Zustand 1 und für den Übergang in den Zustand 0.

3.4 Mehrere mögliche Folgezustände

Wenn eine Episode in zwei oder mehr möglichen Folgezuständen enden kann, genügt es nicht, nur die Verweildauervariable T zu betrachten, sondern man muß sich direkt auf die zweidimensionale Variable (T, D) beziehen. Die Aufgabe besteht dann darin, eine zweidimensionale Verteilung zu ermitteln und darzustellen. Um einen Zugang zu dieser Aufgabe zu finden, ist es zweckmäßig, mit der Idee einer *zielzustandsspezifischen Übergangsrate* zu beginnen. Im diskreten Fall lautet die Definition

$$r_d(t) = \mathrm{P}(T = t, D = d \,|\, T \ge t)$$

wobei d einen der möglichen Folgezustände bezeichnet. Im stetigen Fall lautet die Definition

$$r_d(t) = \lim_{\Delta \to 0} \frac{\mathrm{P}(t \le T < t + \Delta, D = d \,|\, T \ge t)}{\Delta}$$

Die Menge der möglichen Folgezustände werden wir im folgenden stets mit dem Symbol $\tilde{D}$ bezeichnen und dabei als Konvention annehmen, daß

$$\tilde{D} = \{1, \ldots, m\}$$

ist, wenn es m mögliche Folgezustände gibt.

Aufgabe 3.6 Betrachten Sie in Box 3.1 alle Episoden, die im Zustand 1 beginnen. Bestimmen Sie die Menge $\tilde{D}$ der möglichen Folgezustände und berechnen Sie für jeden Zustand $d \in \tilde{D}$ die Übergangsrate $r_d(t)$.

Aufgabe 3.7 Wenn Episoden in mehreren möglichen Folgezuständen enden können, kann man auch von den Unterscheidungen abstrahieren und stattdessen nur einen möglichen Folgezustand betrachten, nämlich das Verlassen des Anfangszustands. Man kann dann die Episoden so betrachten, als ob es nur einen möglichen Folgezustand gibt und die in Abschnitt 3.3 eingeführten Begriffsbildungen verwenden. Zeigen Sie, daß folgender Zusammenhang gilt:

$$r(t) = \sum_{d \in \tilde{D}} r_d(t)$$

wobei $\tilde{D}$ die Menge der möglichen Folgezustände bezeichnet. Verifizieren Sie diesen Zusammenhang an den Rechenergebnissen der Aufgabe 3.6.

AUFGABE 3.8 Bei Episoden mit mehreren möglichen Folgezuständen kann man folgendermaßen sog. *Sub-Survivorfunktionen* definieren:

$$G_d(t) = \exp\left\{ - \int_0^t r_d(\tau)\, d\tau \right\}$$

(a) Überlegen Sie sich, ob bzw. wie man diese Sub-Survivorfunktionen inhaltlich interpretieren kann. (b) Zeigen Sie, daß folgender Zusammenhang zum Begriff der Survivorfunktion gilt:

$$G(t) = \prod_{d \in \tilde{D}} G_d(t)$$

Kapitel 4

Zensierte Beobachtungen

Bisher haben wir angenommen, daß für die Verweildauervariable T, bzw. (T, D) bei mehreren möglichen Folgezuständen, vollständige Beobachtungen verfügbar sind, daß also die Episoden für alle Individuen abgeschlossen sind und wir die Verweildauern und Folgezustände kennen. Das ist bei den in der Praxis ermittelbaren Daten oft nicht der Fall. In diesem Kapitel behandeln wir einen wichtigen Spezialfall unvollständiger Daten, sog. rechts zensierte Beobachtungen.

4.1 Rechts zensierte Beobachtungen

Man sagt, daß die Beobachtung einer Episode bei einem Individuum *rechts zensiert* ist, wenn man zwar weiß, wie lange sich das Individuum schon im Anfangszustand aufhält, aber nicht weiß, wie lange es noch in diesem Zustand bleiben wird und welcher der möglichen Folgezustände dann eintreten wird. Die Situation ist dann folgende: Wir unterstellen eine statistische Variable (T, D) mit einer Menge möglicher Folgezustände $\tilde{D}$. Unsere Beobachtungen für $i = 1, \ldots, n$ Individuen liefern uns jedoch nicht unmittelbar Werte von (T, D), sondern Werte einer Variablen (T^*, D^*). D^* kann Werte in einer Menge

$$\tilde{D}^* = \tilde{D} \cup \{0\}$$

annehmen, wobei 0 der Anfangszustand der Episode ist und infolgedessen kein Element von $\tilde{D}$ sein kann.[1] Die Beobachtungen sind in Form von Werten

$$(t_i^*, d_i^*) \qquad \text{für } i = 1, \ldots, n$$

gegeben, und der Zusammenhang mit den unterstellten Werten (t_i, d_i), also den Werten der als theoretischer Rahmen angenommenen Variablen (T, D), wird folgendermaßen hergestellt:

a) Wenn $d_i^* \in \tilde{D}$, liegt eine nicht zensierte Beobachtung vor, und es gilt: $t_i = t_i^*$ und $d_i = d_i^*$.

[1] Entsprechend unserer Konvention, für $\tilde{D}$ positive natürliche Zahlen zu verwenden, ist also $\tilde{D}^* = \{0, 1, \ldots, m\}$, wenn es m mögliche Folgezustände gibt.

b) Wenn $d_i^* = 0$, liegt eine zensierte Beobachtung vor; über den Folgezustand ist also nichts bekannt, es gilt jedoch $t_i > t_i^*$.

Diese Form der Repräsentation zensierter Beobachtungen erlaubt es, sie auf einfache Weise in einem Episodendatenschema (vgl. Abschnitt 3.1) kenntlich zu machen. Sie werden dadurch kenntlich gemacht, daß man für den Endzustand der Episode ihren Anfangszustand einsetzt, und für die Endzeitstelle diejenige Zeitstelle, bis zu der man weiß, daß sich das Individuum im Anfangszustand der Episode aufgehalten hat.

AUFGABE 4.1 Stellen Sie die Daten des Datensatzes 2 (Box 2.1) in einem Episodendatenschema dar, wobei rechts zensierte Episoden durch die eben genannte Konvention kenntlich gemacht werden.

4.2 Berechnung von Survivorfunktionen

Wir behandeln zunächst eine Situation, in der es nur einen möglichen Folgezustand gibt. Wir können also $\tilde{D}^* = \{0, 1\}$ annehmen, wobei 0 zensierte, 1 unzensierte Beobachtungen kennzeichnet. Wie lassen sich dann Survivorfunktionen berechnen, wenn einige Beobachtungen rechts zensiert sind? Eine genaue Berechnung ist offenbar nicht möglich, denn bei den zensierten Beobachtungen kennt man nur t_i^*, nicht jedoch t_i. Wir können jedoch untere und obere Grenzen für die unbekannte Survivorfunktion $G(t)$ berechnen.

a) Eine untere Grenze, wir bezeichnen sie mit $G^{\vdash}(t)$, erhält man, wenn man für die zensierten Beobachtungen annimmt, daß der Anfangszustand unmittelbar nach dem Zensierungszeitpunkt verlassen wird, also $t_i = t_i^*$ oder, bei einer diskreten Zeitachse, $t_i = t_i^* + 1$.

b) Eine obere Grenze, durch $G^{\dashv}(t)$ bezeichnet, erhält man, wenn man für die zensierten Beobachtungen annimmt, daß der Anfangszustand erst nach einer „beliebig langen" Verweildauer verlassen wird. Es genügt jedoch, die Verweildauern der zensierten Episoden so anzusetzen, daß sie länger sind als die längste unzensierte Verweildauer.

Die unbekannte Survivorfunktion $G(t)$ liegt sicherlich zwischen diesen Grenzen, d.h.

$$G^{\vdash}(t) \leq G(t) \leq G^{\dashv}(t)$$

Die Breite der Intervalle (abhängig von t) hängt natürlich von dem Anteil zensierter Beobachtungen ab und davon, wie sie sich auf der Zeitachse

Box 4.1 Datensatz 4

ID	DUR	CEN
1	17	1
2	5	0
3	22	1
4	13	1
5	2	0
6	9	1
7	12	0
8	15	1

verteilen. Je nachdem liefern die Daten mehr oder weniger viel Information über die Survivorfunktion $G(t)$.

AUFGABE 4.2 Berechnen Sie für die Daten in Box 4.1 untere und obere Grenzen der Survivorfunktion. Stellen Sie dann das Ergebnis in einem Schaubild dar.

4.3 Das Kaplan-Meier-Verfahren

Wenn man etwas nicht genau kennt, wie in diesem Fall die Survivorfunktion $G(t)$, neigen Statistiker dazu, sich Verfahren auszudenken, wie man das, was man nicht kennt, trotzdem möglichst sinnvoll *schätzen* kann. Ein für diesen Zweck ausgedachtes Verfahren stammt von E. L. Kaplan und P. Meier (1958). Um das Verfahren darzustellen, wird zunächst eine diskrete Zeitachse angenommen. Dann gibt es, wie in Abschnitt 3.3 gezeigt worden ist, folgenden Zusammenhang zwischen der Survivorfunktion und der Übergangsrate:

$$G(t) = \prod_{\tau=1}^{t} (1 - r(\tau))$$

Die Idee ist nun, zunächst die Übergangsraten $r(t)$ zu schätzen und dann daraus die Survivorfunktion $G(t)$ zu berechnen. Wenn es keine zensierten Beobachtungen gibt, ist unmittelbar einsichtig, wie man die Übergangsraten berechnen kann, nämlich durch

$$r(t) = \frac{E(t)}{R(t)}$$

Dabei ist $E(t)$ die Anzahl der Individuen, die in der Zeitstelle t den Ausgangszustand der Episode verlassen; und $R(t)$ ist die Anzahl der Individuen, bei denen es in der Zeitstelle t noch möglich ist, daß sie den Ausgangszustand verlassen, also die Anzahl derjenigen Individuen, die den Ausgangszustand nicht schon vorher verlassen haben.

Wenn es zensierte Beobachtungen gibt, kennen wir zwar weder $E(t)$ noch $R(t)$, jedoch zwei vergleichbare Größen. Nämlich $E^*(t)$, die Anzahl der Individuen, deren Verlassen des Ausgangszustands in der Zeitstelle t wir beobachten können; und $R^*(t)$, die Anzahl der Individuen, bei denen ein Verlassen des Ausgangszustands in t noch beobachtet werden könnte, weil sie nicht schon vorher den Ausgangszustand verlassen und/oder rechts zensiert sind. Mithilfe dieser beobachteten Größen kann dann eine beobachtete Übergangsrate

$$r^*(t) = \frac{E^*(t)}{R^*(t)}$$

und daraus schließlich durch Anwendung der Formel (die jetzt eine Definition ist)

$$G^*(t) = \prod_{\tau=1}^{t} (1 - r^*(\tau))$$

eine Survivorfunktion $G^*(t)$ berechnet werden. Offenbar is $G^*(t)$ eine sinnvolle Schätzung für $G(t)$, wenn man voraussetzen kann, daß $r(t)$ sinnvoll durch $r^*(t)$ geschätzt werden kann.

Das gleiche Verfahren kann natürlich angewendet werden, wenn man annimmt, daß die beobachteten (zensierten und nicht zensierten) Verweildauern als exakte Zeitangaben auf einer kontinuierlichen Zeitachse interpretiert werden können. Man erhält dann eine Treppenfunktion, die genau in denjenigen Zeitpunkten Sprungstellen aufweist, in denen mindestens ein Ereignis stattfindet.

Aufgabe 4.3 Berechnen Sie für die Daten in Box 4.1 die Survivorfunktion $G^*(t)$ mit dem Kaplan-Meier-Verfahren. Stellen Sie dann das Ergebnis in einem Schaubild dar, das außerdem die unteren und oberen Schranken, $G^{\vdash}(t)$ und $G^{\dashv}(t)$, zeigt. Beachten Sie, daß $r^*(t)$ nur für diejenigen Zeitstellen berechnet zu werden braucht, in denen mindestens ein Ereignis stattfindet, also $E^*(t) \neq 0$ ist.

4.4 Mehrere Folgezustände

Das Kaplan-Meier-Verfahren läßt sich auch dann verwenden, wenn die Episoden in zwei oder mehr möglichen Folgezuständen enden können. Es werden dann Sub-Survivorfunktionen geschätzt, also

$$G_d^*(t) = \prod_{\tau=1}^{t} (1 - r_d^*(\tau))$$

wobei $d \in \tilde{D}$. Die zielzustandsspezifischen Übergangsraten können durch

$$r_d^*(t) = \frac{E_d^*(t)}{R^*(t)}$$

geschätzt werden, wobei jetzt $E_d^*(t)$ die Anzahl der Individuen ist, bei denen in der Zeitstelle t ein Übergang in den Folgezustand d festgestellt werden kann. Man beachte, daß in diesem Fall der multiplikative Zusammenhang

$$G^*(t) \approx \prod_{d \in \tilde{D}} G_d^*(t)$$

nur näherungsweise gilt.

4.5 Selbst-Konsistenz

Das Kaplan-Meier-Verfahren kann auch mit der Idee einer Selbst-Konsistenz begründet werden, die wir kurz diskutieren wollen. Die Idee ist nicht auf rechts zensierte Daten beschränkt, sondern allgemeiner, und wir besprechen sie deshalb zunächst für eine beliebige diskrete Variable

$$X : \Omega \longrightarrow \tilde{X}$$

Wenn uns für alle Mitglieder von Ω genaue Beobachtungen vorliegen, kann natürlich ohne weiteres die Verteilungsfunktion

$$\mathrm{P}(X = x) \qquad \text{für alle } x \in \tilde{X}$$

berechnet werden (vgl. Abschnitt 2.4). Jetzt nehmen wir jedoch an, daß wir die genauen Werte nicht kennen, sondern für jedes $\omega \in \Omega$ nur eine Teilmenge von $\tilde{X}$, in der der Variablenwert $X(\omega)$ liegt. Um den Gedankengang einfacher darstellen zu können, stellen wir uns vor, daß es für die Mitglieder von Ω Nummern, $i = 1, \dots, n$, gibt. Die beobachteten Werte der Variablen seien durch Merkmalsmengen

$$\tilde{x}_i \subseteq \tilde{X}$$

gegeben. Dann können wir zwar die Verteilungsfunktion P nicht genau berechnen; wir können jedoch zunächst untere und obere Grenzen ermitteln. In einem ersten Schritt definieren wir:

$$\mathrm{pmin}(\tilde{x}_i, \tilde{x}) := \begin{cases} 1 & \text{wenn } \tilde{x}_i \subseteq \tilde{x} \\ 0 & \text{andernfalls} \end{cases}$$

$$\mathrm{pmax}(\tilde{x}_i, \tilde{x}) := \begin{cases} 0 & \text{wenn } \tilde{x}_i \cap \tilde{x} = \emptyset \\ 1 & \text{andernfalls} \end{cases}$$

wobei $\tilde{x}$ eine beliebige Teilmenge von $\tilde{X}$ sein kann. Dann ergeben sich untere und obere Grenzen für P durch die Definitionen:

$$\mathrm{P}^{\vdash}(\tilde{x}) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{pmin}(\tilde{x}_i, \tilde{x})$$

$$\mathrm{P}^{\dashv}(\tilde{x}) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{pmax}(\tilde{x}_i, \tilde{x})$$

Wie man sich leicht überlegen kann, gilt

$$\mathrm{P}^{\vdash}(\tilde{x}) \leq \mathrm{P}(\tilde{x}) \leq \mathrm{P}^{\dashv}(\tilde{x})$$

Die Frage ist nun, wie man sinnvoll eine „mittlere“ Verteilungsfunktion definieren kann, die zwischen den beiden Grenzen liegt; denn die Verteilung P kennt man nicht, und man kann sie (ohne weitere Annahmen) auch nicht aus den Daten schätzen. Eine Überlegung wäre die folgende. Man nimmt an, daß das Individuum i an der Merkmalsmenge $\tilde{x}$ in dem Maße partizipiert, wie sich $\tilde{x}_i$ und $\tilde{x}$ überschneiden. Diese Idee führt zu folgender Definition einer Verteilungsfunktion:

$$\bar{\mathrm{P}}(\tilde{x}) := \frac{1}{n} \sum_{i=1}^{n} \frac{|\,\tilde{x}_i \cap \tilde{x}\,|}{|\,\tilde{x}_i\,|}$$

Wir nennen sie *mittlere Verteilungsfunktion*. Dies ist jedoch ein Spezialfall einer allgemeineren Idee. Wenn wir uns nämlich auf eine beliebige Verteilungsfunktion $\tilde{\mathrm{P}}$ beziehen, erscheint es sinnvoll zu sagen, daß das Individuum i an $\tilde{x}$ entsprechend der Größe von

$$\frac{\tilde{\mathrm{P}}(\tilde{x}_i \cap \tilde{x})}{\tilde{\mathrm{P}}(\tilde{x}_i)}$$

partizipiert. Die mittlere Verteilungsfunktion resultiert dann als ein Spezialfall, wenn man nämlich für $\tilde{\mathrm{P}}$ eine Gleichverteilung annimmt.

Ein zweiter Ansatz ergibt sich durch die Idee der Selbst-Konsistenz. Die Idee ist, nach einer Verteilung $\tilde{\mathrm{P}}$ zu suchen, so daß die folgende Funktionalgleichung erfüllt ist:

$$\tilde{\mathrm{P}}(\tilde{x}) := \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\mathrm{P}}(\tilde{x}_i \cap \tilde{x})}{\tilde{\mathrm{P}}(\tilde{x}_i)} \qquad (4.5.1)$$

Eine Lösung kann meistens mit einem iterativen Verfahren gefunden werden: Man beginnt mit einer beliebigen, z.B. der mittleren Verteilungsfunktion; dann wendet man die Formel an, um daraus eine neue Verteilungsfunktion zu berechnen; dann wiederholt man die Berechnung mit der neuen Verteilungsfunktion, usw., bis sich keine wesentlichen Änderungen mehr ergeben. Wenn man eine Lösung gefunden hat, wird sie als *selbst-konstistente Verteilung* bezeichnet.

AUFGABE 4.4 Berechnen Sie mit den Daten

$$\tilde{x}_1 = \{1\},\ \tilde{x}_2 = \{2\},\ \tilde{x}_3 = \{3\},\ \tilde{x}_4 = \{1,2\},\ \tilde{x}_5 = \{2,3\}$$

die Verteilungen $\mathrm{P}^{\vdash}$, $\mathrm{P}^{\dashv}$, $\bar{\mathrm{P}}$ und $\tilde{\mathrm{P}}$.

AUFGABE 4.5 Man kann zeigen, daß das Kaplan-Meier-Verfahren eine selbst-konsistente Verteilung erzeugt. Überlegen Sie sich, wie mit dem beschriebenen Verfahren eine selbst-konsistente Verteilungs- oder Survivorfunktion für den Datensatz 4.1 berechnet werden kann.

Kapitel 5

Regressionsmodelle für Zustände

In den weiteren Kapiteln beschäftigen wir uns mit statistischen Modellen für die Analyse von Abhängigkeiten zwischen Variablen. Um den Ausgangspunkt zu fixieren, erinnern wir uns an die beiden Formen, die wir zur Repräsentation von Verlaufsdaten eingeführt haben. Einerseits haben wir Folgen von Zustandsvariablen betrachtet: Y_t, wobei der Index t sich auf einer diskreten Zeitachse bewegen kann. Andererseits haben wir uns auf einzelne Episoden bezogen und diese durch eine zweidimensionale Variable (T, D) repräsentiert. Beide Varianten können als Ausgangspunkt für Modellkonstruktionen dienen. In diesem Kapitel gehen wir von der ersten Variante aus.

5.1 Der Modellansatz

Wir beziehen uns auf eine Prozeßzeitachse $t = 0, 1, 2, \ldots$ und nehmen an, daß eine Folge von Zustandsvariablen

$$Y_t : \Omega \longrightarrow \tilde{Y}$$

gegeben ist. $\tilde{Y}$ ist der Zustandsraum, der zwei oder mehr unterschiedliche Zustände enthalten kann. Wenn sich der Prozeß bis zu einem Zeitpunkt t entwickelt hat, wird er statistisch durch die Verteilung

$$\mathrm{P}(Y_t = y_t, Y_{t-1} = y_{t-1}, \ldots, Y_0 = y_0)$$

erfaßt. Dabei sind $y_0, \ldots, y_t$ mögliche Zustände im Zustandsraum $\tilde{Y}$.

Ein Modell soll es erlauben, über Abhängigkeiten zwischen den Zustandsvariablen nachzudenken zu können. Um unseren Modellansatz einfach zu schreiben, verwenden wir folgende Abkürzung:

$$\bar{Y}_t := (Y_t, Y_{t-1}, \ldots, Y_0)$$

und nennen dies eine *Prozeßvariable*. Auf mögliche Werte wird durch entsprechende Kleinbuchstaben, also

$$\bar{y}_t := (y_t, y_{t-1}, \ldots, y_0)$$

verwiesen; es handelt sich um mögliche Zustandsfolgen. Der Ausgangspunkt für die Modellbildung kann dann durch

$$\mathrm{P}(\bar{Y}_t = \bar{y}_t)$$

fixiert werden. Um Abhängigkeiten zwischen den Zustandsvariablen zu erfassen, werden bedingte Verteilungen verwendet. Dabei nehmen wir an, daß die Ausgangsverteilung, also die Verteilung von Y_0, vorgegeben ist und sich der Prozeß dann sequentiell entwickelt; in symbolischer Notation:

$$\begin{array}{l} Y_0 \\ Y_1 \mid Y_0 \\ Y_2 \mid Y_0, Y_1 \\ \quad\vdots \\ Y_t \mid Y_0, Y_1, \ldots, Y_{t-1} \end{array}$$

Durch sukzessive Anwendung der Regel zur Bildung von bedingten Verteilungen erhält man dann:

$$\mathrm{P}(\bar{Y}_t = \bar{y}_t) = \prod_{\tau=1}^{t} \mathrm{P}(Y_\tau = y_\tau \mid \bar{Y}_{\tau-1} = \bar{y}_{\tau-1})\, \mathrm{P}(Y_0 = y_0) \tag{5.1.1}$$

5.2 Spekulation und Empirie

Der allgemeine Modellansatz (5.1.1) kann in zwei unterschiedlichen Weisen als Ausgangspunkt für weitere Überlegungen dienen. Er kann einerseits als Ausgangspunkt für spekulative Überlegungen zur Prozeßentwicklung, andererseits als ein formaler Rahmen für die Repräsentation von Daten über die Prozeßentwicklung verwendet werden. Um mit dem Modellansatz etwas vertrauter zu werden, beginnen wir mit spekulativen Überlegungen.

Da wir annehmen, daß die Verteilung von Y_0 vorgegeben ist, konzentriert sich die Spekulation auf Annahmen über

$$\mathrm{P}(Y_\tau = y_\tau \mid \bar{Y}_{\tau-1} = \bar{y}_{\tau-1})$$

Eine einfache Annahme könnte zum Beispiel darin bestehen, daß es nur ein einstufiges Gedächtnis gibt:

$$\mathrm{P}(Y_\tau = y_\tau \mid \bar{Y}_{\tau-1} = \bar{y}_{\tau-1}) \;=\; \mathrm{P}(Y_\tau = y_\tau \mid Y_{\tau-1} = y_{\tau-1})$$

Hier wird also angenommen, daß der Zustand, der in einer Zeitstelle eingenommen wird, nur davon abhängt, welcher Zustand in der vorangegangenen Zeitstelle angenommen worden ist. Etwas komplizierter wäre ein zweistufiges Gedächtnis, das man durch die Annahme

$$\begin{aligned} &\mathrm{P}(Y_\tau = y_\tau \mid \bar{Y}_{\tau-1} = \bar{y}_{\tau-1}) \;= \\ &\qquad \mathrm{P}(Y_\tau = y_\tau \mid Y_{\tau-1} = y_{\tau-1}, Y_{\tau-2} = y_{\tau-2}) \end{aligned}$$

ausdrücken kann.

Aufgabe 5.1 Betrachten Sie einen Prozeß mit einem einstufigen Gedächtnis. Der Zustandsraum sei $\tilde{Y} = \{0, 1\}$, zum Prozeßbeginn befinden sich alle Individuen im Zustand 0, und die Übergangswahrscheinlichkeiten werden durch

$$\begin{aligned} \mathrm{P}(Y_\tau = 1 \mid Y_{\tau-1} = 0) &= 1/2 \\ \mathrm{P}(Y_\tau = 1 \mid Y_{\tau-1} = 1) &= 1/3 \end{aligned}$$

angenommen. Konstruieren Sie mithilfe eines Würfels 10 Realisationen dieses Prozesses, für $t = 0, \ldots, 8$, und stellen Sie in einer Tabelle dar, wie sich die Zustandsverteilung im Zeitablauf entwickelt.

Aufgabe 5.2 Nehmen Sie an, daß es 5 unterschiedliche Zustände gibt. Wieviele Übergangswahrscheinlichkeiten sind erforderlich, um einen Prozeß mit einem zweistufigen Gedächtnis vollständig darzustellen?

5.3 Modelle für zwei Zustände

Wir betrachten Prozesse, bei denen es nur zwei unterschiedliche Zustände gibt, also $\tilde{Y} = \{0, 1\}$. Außerdem nehmen wir an, daß es nur ein einstufiges Gedächtnis gibt. Wenn man keine weiteren Einschränkungen vornimmt, müssen für jede Zeitstelle zwei Parameter ermittelt werden:

$$\begin{aligned} \mathrm{P}(Y_t = 1 \mid Y_{t-1} = 0) &= \theta_{10,t} \\ \mathrm{P}(Y_t = 1 \mid Y_{t-1} = 1) &= \theta_{11,t} \end{aligned}$$

Die Größen $\theta_{10,t}$ und $\theta_{11,t}$ nennen wir *Parameter des Prozesses*. Im allgemeinen muß angenommen werden, daß sich diese Prozeßparameter während der Entwicklung des Prozesses verändern können. Eine radikal vereinfachende, aber auch problematische Annahme ist, daß die Prozeßparameter zeitkonstant sind; man spricht dann von einem *stationären*

Prozeß. Der Modellansatz ist dann

$$\mathrm{P}(Y_t = 1 \,|\, Y_{t-1} = 0) \;=\; \theta_{10}$$
$$\mathrm{P}(Y_t = 1 \,|\, Y_{t-1} = 1) \;=\; \theta_{11}$$

AUFGABE 5.3 Verwenden Sie die Daten aus Box 2.1. Berechnen Sie zunächst die zeitstellenspezifischen Prozeßparameter

$$\theta_{ij,t} \qquad \text{für} \quad i,j \in \{0,1\},\; t = 2,\ldots,6$$

Nehmen Sie dann an, daß die Daten aus einem stationären Prozeß stammen und berechnen Sie die zeitkonstanten Prozeßparameter

$$\theta_{ij} \qquad \text{für} \quad i,j \in \{0,1\}$$

5.4 Modelle mit Kovariablen

Bisher haben wir nur eine Prozeßvariable, Y_t, betrachtet, und die Modellbildung bezog sich darauf, herauszufinden, wie der jeweils gegenwärtige Zustand von vorangegangenen Zuständen abhängt. Bei praktischen Anwendungen ist man oft daran interessiert, noch weitere Variablen, sog. *Kovariablen*, zu berücksichtigen. Zwei Arten von Kovariablen können dabei unterschieden werden:

- *Zeitunabhängige Kovariablen*, deren Werte zu Beginn des Prozesses feststehen und sich während des Prozesses nicht ändern können; und
- *Zeitabhängige Kovariablen*, deren Werte sich während des Prozesses verändern können.

Offenbar können zeitunabhängige Kovariablen als ein Spezialfall zeitabhängiger Kovariablen betrachtet werden. Wir betrachten deshalb im folgenden nur zeitabhängige Kovariablen. Einen sinnvollen Begriffsrahmen liefert dann die Vorstellung *paralleler Prozesse*. Den primär interessierenden Prozeß repräsentieren wir wie bisher durch die Zustandsvariablen Y_t, den parallelen Kovariablenprozeß durch eine Folge von Kovariablen X_t; in expliziter Schreibweise:

$$(X_t, Y_t) : \Omega \longrightarrow \tilde{X} \times \tilde{Y}$$

Wie bisher nehmen wir an, daß Y_t eine diskrete eindimensionale Zustandsvariable ist. Bei X_t kann es sich um eine mehrdimensionale Variable handeln, z.B. um eine m-dimensionale Variable

$$X_t = (X_{t1}, \ldots, X_{tm})$$

Box 5.1 Datensatz 5

```
ID   t =  0  1  2  3  4  5  6
-----------------------------
 1   Y    0  0  0  1  1  0  0
     X1   0  0  0  0  0  0  0
     X2  20 21 22 23 24 25 26
 2   Y    1  1  0  0  0  1  1
     X1   0  0  0  0  0  0  0
     X2  22 23 24 25 26 27 28
 3   Y    1  1  1  0  0  0  0
     X1   0  0  0  0  0  0  0
     X2  21 22 23 24 25 26 27
 4   Y    0  0  0  0  1  1  1
     X1   1  1  1  1  1  1  1
     X2  20 21 22 23 24 25 26
 5   Y    0  0  1  1  1  0  0
     X1   1  1  1  1  1  1  1
     X2  22 23 24 25 26 27 28
 6   Y    1  1  1  0  0  1  1
     X1   1  1  1  1  1  1  1
     X2  21 22 23 24 25 26 27
```

Zur Vereinfachung werden wir jedoch annehmen, daß auch X_t eine diskrete Variable ist.

Die Idee ist nun, daß der Zustand in einer Zeitstelle t nicht nur von Zuständen abhängen kann, die in vorangehenden Zeitstellen eingenommen worden sind, sondern auch von den bisher realisierten Werten der Kovariablen. Für die Modellbildung nehmen wir an, daß die Werte der Kovariablen ihrerseits nicht von den bisher realisierten Werten der Prozeßvariablen Y_t abhängig sind.[1] Der allgemeine Modellansatz (5.1.1) kann dann folgendermaßen erweitert werden:

$$\mathrm{P}(\bar{Y}_t = \bar{y}_t) \;=\; \prod_{\tau=1}^{t} \mathrm{P}(Y_\tau = y_\tau \,|\, \bar{Y}_{\tau-1} = \bar{y}_{\tau-1}, \bar{X}_{\tau-1} = \bar{x}_{\tau-1})\; \mathrm{P}(Y_0 = y_0, X_0 = x_0) \tag{5.4.1}$$

Wiederum kann dieser Modellansatz auf vielfältige Weisen vereinfacht werden. Denkt man an die Idee eines Prozesses mit einem einstufigen

[1] Die Kovariablen werden dann *exogen* genannt. Wenn diese Annahme nicht erfüllt ist, spricht man gelegentlich von *interdependenten* Prozessen. Damit werden wir uns in dieser Einführung jedoch nicht näher beschäftigen.

Gedächtnis, kann man das zum Beispiel auch für die Kovariablen annehmen und erhält dann den Modellansatz

$$\begin{aligned}&\mathrm{P}(Y_\tau = y_\tau \,|\, \bar{Y}_{\tau-1} = \bar{y}_{\tau-1}, \bar{X}_{\tau-1} = \bar{x}_{\tau-1}) \;=\\ &\qquad \mathrm{P}(Y_\tau = y_\tau \,|\, Y_{\tau-1} = y_{\tau-1}, X_{\tau-1} = x_{\tau-1})\end{aligned}$$

Als Beispiel betrachten wir den Datensatz 5 in Box 5.1. Er enthält Angaben über die Entwicklung von Zuständen bei 6 Personen. Es gibt zwei mögliche Zustände, $\tilde{Y} = \{0,1\}$, und zwei Kovariablen. Die Kovariable X_1 ist zeitunabhängig, z.B. das Geschlecht der Personen (0 = Männer, 1 = Frauen); die Kovariable X_2 ist zeitabhängig, z.B. das Alter der Personen.

5.5 Binäre Logitmodelle

Zur Modellierung von Prozessen mit Kovariablen werden in der Praxis oft Logitmodelle verwendet, die rechentechnisch verhältnismäßig einfach handhabbar sind. Wir besprechen hier diese Modelle für Prozesse, bei denen es nur zwei mögliche Zustände gibt, also $\tilde{Y} = \{0,1\}$. Verzichtet man zunächst auf die Annahme, daß es sich um einen stationären Prozeß handelt, ist der Modellansatz:

$$\mathrm{P}(Y_t = 1 \,|\, Y_{t-1} = 0, X_{t-1} = x_{t-1}) \;=\; \frac{\exp(\alpha_{10,t} + x_{t-1}\beta_{10,t})}{1 + \exp(\alpha_{10,t} + x_{t-1}\beta_{10,t})}$$

$$\mathrm{P}(Y_t = 1 \,|\, Y_{t-1} = 1, X_{t-1} = x_{t-1}) \;=\; \frac{\exp(\alpha_{11,t} + x_{t-1}\beta_{11,t})}{1 + \exp(\alpha_{11,t} + x_{t-1}\beta_{11,t})}$$

Nimmt man an, daß es sich um einen stationären Prozeß handelt, entfällt auf den rechten Seiten der Zeitindex bei den Modellparametern. Das in Abschnitt 5.3 behandelte Modell ohne Kovariablen ist offenbar ein Spezialfall dieses Logitmodells.

Leider können die Parameter eines Logitmodels im allgemeinen nicht mit einfachen Rechenverfahren aus Daten berechnet werden. Wir werden uns im nächsten Abschnitt mit einem Schätzverfahren beschäftigen. Zuvor behandeln wir einige Aufgaben, um mit dem Modellansatz etwas vertrauter zu werden.

AUFGABE 5.4 Zeichnen Sie den Verlauf der Logitfunktion

$$z = \frac{\exp(x)}{1 + \exp(x)}$$

im Bereich $-3 \le x \le 3$.

AUFGABE 5.5 Unter Verwendung der Notation aus Aufgabe 5.4, entwickeln Sie die Umkehrfunktion, die zeigt, wie x von z abhängt. Berechnen Sie die x-Werte, die den z-Werten 0.5, 0.6 und 0.7 entsprechen.

AUFGABE 5.6 Formulieren Sie ein Logitmodell für die Daten aus Box 5.1 unter der Annahme, daß es sich um Daten aus einem stationären Prozeß handelt. Überlegen Sie sich insbesondere, welche Modellparameter berechnet werden müßten.

5.6 Maximum-Likelihood-Schätzung

In diesem Abschnitt besprechen wir, wie die Parameter eines binären Logitmodells mit der ML-Methode berechnet werden können. Wenn angenommen wird, daß es sich um einen stationären Prozeß handelt, müssen vier Parameter berechnet werden:

$$\alpha_{10}, \beta_{10}, \alpha_{11}, \beta_{11}$$

Die Daten seien in der Form

$$(x_{it}, y_{it}) \qquad \text{für} \quad i = 1, \ldots, N, \; t = 0, \ldots, T$$

gegeben. Dann folgt aus dem Modellansatz:

$$\mathrm{P}(Y_t = y_{i,t} \,|\, Y_{t-1} = y_{i,t-1}, X_{t-1} = x_{i,t-1}) = \begin{cases} \frac{\exp(\alpha_{10} + x_{i,t-1}\beta_{10})}{1+\exp(\alpha_{10} + x_{i,t-1}\beta_{10})} & \text{wenn } \; y_{i,t} = 1, y_{i,t-1} = 0 \\ \frac{1}{1+\exp(\alpha_{10} + x_{i,t-1}\beta_{10})} & \text{wenn } \; y_{i,t} = 0, y_{i,t-1} = 0 \\ \frac{\exp(\alpha_{11} + x_{i,t-1}\beta_{11})}{1+\exp(\alpha_{11} + x_{i,t-1}\beta_{11})} & \text{wenn } \; y_{i,t} = 1, y_{i,t-1} = 1 \\ \frac{1}{1+\exp(\alpha_{11} + x_{i,t-1}\beta_{11})} & \text{wenn } \; y_{i,t} = 0, y_{i,t-1} = 1 \end{cases}$$

Das kann man in einer Formel folgendermaßen ausdrücken:

$$\begin{aligned}&\mathrm{P}(Y_t = y_{i,t} \,|\, Y_{t-1} = y_{i,t-1}, X_{t-1} = x_{i,t-1}) =\\ &\left(\frac{\exp(\alpha_{10} + x_{i,t-1}\beta_{10})^{y_{it}}}{1 + \exp(\alpha_{10} + x_{i,t-1}\beta_{10})}\right)^{1-y_{i,t-1}} \left(\frac{\exp(\alpha_{11} + x_{i,t-1}\beta_{11})^{y_{it}}}{1 + \exp(\alpha_{11} + x_{i,t-1}\beta_{11})}\right)^{y_{i,t-1}}\end{aligned}$$

Betrachtet man diesen Ausdruck als eine Funktion der Modellparameter, wird er als *Likelihood* der Beobachtung i in der Zeitstelle t bezeichnet. Die ML-Methode besteht nun darin, diese Likelihood für alle Beobachtungen und Zeitstellen zusammenzufassen und dann als eine Funktion der Modellparameter zu maximieren. Die *Likelihoodfunktion* sieht also folgendermaßen aus:

$$\mathcal{L}(\alpha_{10}, \beta_{10}, \alpha_{11}, \beta_{11}) = \prod_{i=1}^{N} \prod_{t=1}^{T} \left(\frac{\exp(\alpha_{10} + x_{i,t-1}\beta_{10})^{y_{it}}}{1 + \exp(\alpha_{10} + x_{i,t-1}\beta_{10})} \right)^{1-y_{i,t-1}} \left(\frac{\exp(\alpha_{11} + x_{i,t-1}\beta_{11})^{y_{it}}}{1 + \exp(\alpha_{11} + x_{i,t-1}\beta_{11})} \right)^{y_{i,t-1}}$$

Aus der Maximierung dieser Likelihoodfunktion ergeben sich die ML-Schätzwerte der Modellparameter, die mit

$$\hat{\alpha}_{10}, \hat{\beta}_{10}, \hat{\alpha}_{11}, \hat{\beta}_{11}$$

bezeichnet werden.

Aufgabe 5.7 Für die praktische Berechnung verwendet man meistens die *Log-Likelihoodfunktion*, also

$$\ell(\alpha_{10}, \beta_{10}, \alpha_{11}, \beta_{11}) = \log\left(\mathcal{L}(\alpha_{10}, \beta_{10}, \alpha_{11}, \beta_{11})\right)$$

Berechnen Sie die Log-Likelihoodfuntion für die oben angegebene Likelihoodfunktion.

Aufgabe 5.8 Berechnen Sie den Wert der in Aufgabe 5.7 gefundenen Log-Likelihoodfunktion für den Datensatz in Box 5.1 und für die Parameterwerte

$$\alpha_{10} = \beta_{10} = \alpha_{11} = \beta_{11} = 0$$

Aufgabe 5.9 Die ML-Schätzwerte der Modellparameter ergeben sich aus der Maximierung der Log-Likelihoodfunktion. Eine Lösung kann mithilfe eines Statistik-Programms berechnet werden, das in der Lage ist, Logitmodelle zu schätzen.[2] Zunächst erkennt man, daß sich die Likelihoodfunktion in zwei Faktoren separieren läßt, die unabhängig voneinander maximiert werden können. Überlegen Sie sich, wie dementsprechend der Datensatz aus Box 5.1 reorganisiert und aufgeteilt werden kann (einen Hinweis gibt Box 5.2). Überlegen Sie sich dann, wie die Modellparameter mit einem Programm berechnet werden können, das einfache Logitmodelle für eine binäre abhängige Variable schätzen kann.

[2] Wie man das praktisch machen kann, wird im Rahmen der praktischen Übungen im Anhang besprochen.

Box 5.2 Datensatz 5a

ID	t	Y(t)	Y(t-1)	X1(t-1)	X2(t-1)
1	1	0	0	0	20
1	2	0	0	0	21
1	3	1	0	0	22
1	4	1	1	0	23
1	5	0	1	0	24
1	6	0	0	0	25
2	1	1	1	0	22
2	2	0	1	0	23
2	3	0	0	0	24
2	4	0	0	0	25
2	5	1	0	0	26
2	6	1	1	0	27
3	1	1	1	1	21
3	2	1	1	1	22
3	3	1	0	1	23
3	4	0	0	1	24
3	5	0	0	1	25
3	6	0	0	1	26
4	1	0	0	1	20
4	2	0	0	1	21
4	3	0	0	1	22
4	4	1	0	1	23
4	5	1	1	1	24
4	6	1	1	1	25
5	1	0	0	1	22
5	2	1	0	1	23
5	3	1	1	1	24
5	4	1	1	1	25
5	5	0	1	1	26
5	6	0	0	1	27
6	1	1	1	1	21
6	2	1	1	1	22
6	3	0	1	1	23
6	4	0	0	1	24
6	5	1	0	1	25
6	6	1	1	1	26

Aufgabe 5.10 Verwendet man für die Variable X das Alter (X2 in Box 5.1), erhält man folgende ML-Schätzwerte der Modellparameter:

$$\hat{\alpha}_{10} = -3.14,\ \hat{\beta}_{10} = 0.10,\ \hat{\alpha}_{11} = 4.91,\ \hat{\beta}_{11} = -0.16$$

Das für diese Berechnungen verwendete Computerprogramm (TDA) gibt als Wert der maximierten Log-Likelihoodfunktion die Werte -12.14 (für die erste Modellhälfte, $Y_{t-1} = 0$) und -8.88 (für die zweite Modellhälfte, $Y_{t-1} = 1$) an. Berechnen Sie den Wert der Log-Likelihoodfunktion für das Gesamtmodell.

AUFGABE 5.11 Unter Verwendung der in Aufgabe 5.10 angegebenen Schätzwerte für die Modellparameter: Erstellen Sie eine Tabelle, die für die Alterswerte $X = 20, \ldots, 26$ und die Werte $Y_{t-1} = 0, 1$ die geschätzten Wahrscheinlichkeiten

$$\begin{aligned}&\mathrm{P}(Y_t = 0 \,|\, Y_{t-1} = \cdots, X = \cdots) \quad \text{und}\\&\mathrm{P}(Y_t = 1 \,|\, Y_{t-1} = \cdots, X = \cdots)\end{aligned}$$

angibt.

Kapitel 6

Modelle für Verweildauern

Entsprechend der zu Beginn des vorangegangenen Kapitels getroffenen Unterscheidung beziehen wir uns für die weitere Diskussion statistischer Modelle auf einzelne Episoden. Ausgangspunkt ist also die Repräsentation einer einzelnen Episode durch eine zweidimensionale statistische Variable

$$(T, D) \qquad \text{wobei } T \in \tilde{T},\; D \in \tilde{D}$$

$\tilde{D}$ ist die Menge der möglichen Folgezustände der Episode, $\tilde{T}$ ist die Zeitachse. Für die Zeitachse kann man entweder eine diskrete oder eine kontinuierliche numerische Repräsentation wählen. Wir verwenden im folgenden eine kontinuierliche Repräsentation, da die meisten in der Literatur diskutierten und praktisch verwendeten Modelle von dieser Konvention ausgehen. Wir identifizieren $\tilde{T}$ also mit den nichtnegativen reellen Zahlen.

Statistische Modelle können sowohl spekulativen als auch repräsentativen Zwecken dienen. Im ersten Fall geht es darum, Vorstellungen darüber zu bilden, wie Episoden ablaufen *könnten*. Damit beschäftigen wir uns in diesem Kapitel. Wie schon in früheren Kapiteln können wir zwei Situationen unterscheiden: Erstens eine Betrachtungsweise, bei der Episoden nur in einem möglichen Folgezustand enden können; es genügt dann, die Episode durch eine Verweildauervariable (T) zu erfassen. Zweitens eine Betrachtungsweise, bei der es zwei oder mehr mögliche Folgezustände gibt; dann muß von der zweidimensionalen Variablen (T, D) ausgegangen werden.

6.1 Zeitkonstante Raten

Beginnen wir mit einer Episode, für die es nur einen möglichen Folgezustand gibt. Es genügt dann, die Verweildauervariable T zu betrachten, und ein statistisches Modell besteht darin, Annahmen über die Verteilung dieser Variablen zu machen. Wie wir uns schon überlegt haben, gibt es zur Charakterisierung dieser Verteilung vier äquivalente Konzepte: die Verteilungsfunktion $F(t)$, die Survivorfunktion $G(t)$, die Dichtefunktion

$f(t)$ und die Rate $r(t)$. Um Annahmen über die Verteilung von T zu formulieren, kann man also wahlweise einen dieser vier Begriffe verwenden und die jeweils übrigen daraus berechnen.

Um inhaltliche Vorstellungen über den Ablauf einer Episode zu bilden, ist es oft hilfreich, vom Begriff der Rate auszugehen. Die allereinfachste Vorstellung besteht darin, von einer zeitkonstanten Rate auszugehen. Diese Annahme kann so formuliert werden:

$$r(t) = \theta$$

wobei θ ein Modellparameter ist, der in einem vorgegebenen Parameterraum variieren kann. Wir bezeichnen den Parameterraum mit Θ; und da Raten keine negativen Werte annehmen können, können wir Θ mit den nichtnegativen reellen Zahlen identifizieren.

Ein Verteilungsmodell, das auf der Annahme einer zeitkonstanten Rate beruht, wird *Exponentialmodell* genannt. Man spricht von einer *Exponentialverteilung* mit dem Parameter θ.

Die Survivorfunktion für eine Exponentialverteilung ergibt sich aus unserer Basisformel für den Zusammenhang zwischen Rate und Survivorfunktion in folgender Weise:

$$G(t) = \exp\left\{-\int_0^t r(\tau)\,d\tau\right\} = \exp(-\theta t)$$

AUFGABE 6.1 Berechnen Sie die Verteilungs- und Dichtefunktionen für eine Exponentialverteilung mit dem Parameter θ.

AUFGABE 6.2 Zeichnen Sie den Verlauf der Survivor- und Dichtefunktionen für eine Standard-Exponentialverteilung, d.h. für eine Exponentialverteilung mit dem Parameter $\theta = 1$.

6.2 Weibull-Verteilung

Interessanter als die Exponentialverteilung sind Verteilungen, bei denen sich die Rate im Zeitablauf verändern kann. Eine in Anwendungen oft verwendete Verteilung ist die *Weibull-Verteilung*, bei der die Rate monoton steigen oder fallen kann. Die Weibull-Verteilung hat zwei Parameter, die Survivorfunktion sieht folgendermaßen aus:

$$G(t) = \exp\{-(\alpha t)^\beta\}$$

α und β sind die Verteilungsparameter, und es wird vorausgesetzt, daß beide Parameter nur positive Werte annehmen können.

Aus der Survivorfunktion erhält man durch Differenzieren die Dichtefunktion

$$f(t) = \beta\alpha^\beta t^{\beta-1}\exp\{-(\alpha t)^\beta\}$$

und daraus die Rate

$$r(t) = \beta\alpha^\beta t^{\beta-1}$$

AUFGABE 6.3 Zeigen Sie schrittweise, wie sich die Dichtefunktion und die Rate aus der Survivorfunktion der Weibull-Verteilung berechnen lassen.

AUFGABE 6.4 Für welche Parameterwerte der Weibull-Verteilung erhält man die Exponentialverteilung als einen Spezialfall?

AUFGABE 6.5 Zeichnen Sie den Verlauf der Rate der Weilbull-Verteilung auf einer Zeitachse von 0 bis 3, und zwar für die Parameterwerte $\alpha = 1$ und $\beta = 0.5, 1.0, 1.5$.

6.3 Loglogistische Verteilung

Die Exponentialverteilung kann nur zeitkonstante, die Weibull-Verteilung nur monotone Ratenverläufe ausdrücken. Für nichtmonotone Ratenverläufe eignet sich manchmal die sog. *loglogistische Verteilung*. Sie hat, wie die Weibull-Verteilung, zwei Parameter, α und β, die nur positive Werte annehmen können. Die Survivorfunktion sieht folgendermaßen aus:

$$G(t) = \frac{1}{1 + (\alpha t)^\beta}$$

Daraus erhält man durch Differenzieren die Dichtefunktion

$$f(t) = \frac{\beta\alpha^\beta t^{\beta-1}}{(1 + (\alpha t)^\beta)^2}$$

und schließlich die Rate

$$r(t) = \frac{\beta\alpha^\beta t^{\beta-1}}{1 + (\alpha t)^\beta}$$

AUFGABE 6.6 Zeigen Sie schrittweise, wie sich die Dichtefunktion und die Rate aus der Survivorfunktion der loglogistischen Verteilung berechnen lassen.

AUFGABE 6.7 Zeigen Sie, daß die Exponentialverteilung *kein* Spezialfall der loglogistischen Verteilung ist.

AUFGABE 6.8 Zeichnen Sie den Verlauf der Rate der loglogistischen Verteilung auf einer Zeitachse von 0 bis 3, und zwar für die Parameterwerte $\alpha = 1$ und $\beta = 1$, $\beta = 2$.

AUFGABE 6.9 Zeigen Sie, daß die Rate der loglogistischen Verteilung (wenn sie konkav ist) ihr Maximum in der Zeitstelle

$$t_{\max} = \frac{1}{\alpha}(\beta - 1)^{\frac{1}{\beta}}$$

annimmt. Für welche Parameterwerte ist der Verlauf der Rate konkav? Welchen Wert nimmt die Rate in ihrem Maximum an?

6.4 Lognormal-Verteilung

Bei Regressionsmodellen wird oft eine Normalverteilung unterstellt. Da unsere Verweildauervariable T nur positive Werte annehmen kann, liegt es nahe, ihren Logarithmus zu betrachten und dafür eine Normalverteilung zu unterstellen. Allgemein sagt man: eine Variable ist *lognormal* verteilt, wenn ihr Logarithmus normalverteilt ist.[1]

Um diesen Gedanken zu verfolgen, betrachten wir zunächst ein allgemeineres Problem. Es sei X eine kontinuierliche Variable mit der Verteilungsfunktion $F_X(x)$ und der Dichtefunktion $f_X(x)$. Außerdem sei g eine beliebige monoton steigende Funktion. Wir können dann eine neue Variable

$$Y = g(X)$$

bilden. Die Frage ist, wie man die Verteilungs- und Dichtefunktionen von Y, also $F_Y(y)$ und $f_Y(y)$, aus den entsprechenden Funktionen für X berechnen kann; oder umgekehrt, wie man die Verteilung von X aus der Verteilung von Y finden kann. Da die Transformationsfunktion g monoton ist, kann man natürlich auch

$$X = g^{-1}(Y)$$

[1] Wir verwenden hier stets den natürlichen Logarithmus, also die Umkehrfunktion zur Exponentialfunktion.

betrachten, insofern ist das Problem symmetrisch. Zur Beantwortung der Fragen gibt es zunächst folgende Beziehung zwischen den Verteilungsfunktionen:

$$F_X(x) = \mathrm{P}(X \leq x) = \mathrm{P}(Y \leq g(x)) = F_Y(g(x))$$

Beziehungen zwischen den Dichtefunktionen ergeben sich aus Ableitungen der Verteilungsfunktionen. Man findet

$$f_X(x) = \frac{dF_X(u)}{du}\bigg|_{u=x} = \frac{dF_Y(g(u))}{du}\bigg|_{u=x} = f_Y(g(x))g'(x)$$

wobei $g'(x)$ die Ableitung der Transformationsfunktion g bezeichnet.

Wenden wir jetzt diese Überlegung auf eine Situation an, in der die Variable Y normalverteilt und der Zusammenhang zwischen T, unserer Verweildauervariablen, und Y durch die Transformation

$$Y = \log(T)$$

definiert ist. Für die Dichtefunktion von Y haben wir also

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$$

Wenden wir jetzt die oben abgeleitete Transformationsformel an, erhalten wir für die Dichtefunktion von T, die wir jetzt wieder mit $f(t)$ bezeichnen, den Ausdruck

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}$$

Dies ist die Dichtefunktion der Lognormal-Verteilung mit den Paramtern μ und σ. Dabei kann μ beliebige, σ nur positive Werte annehmen.

Ein Nachteil dieser Verteilung ist, daß es für die Verteilungsfunktion und infolgedessen für die Rate keinen geschlossenen, einfach berechenbaren Ausdruck gibt. Die Lognormal-Verteilung ist trotzdem wichtig, weil sie es erlaubt, einen Zusammenhang zwischen Ratenmodellen und üblichen Regressionsmodellen herzustellen.

AUFGABE 6.10 Zeichnen Sie den Verlauf der Dichtefunktion der Lognormal-Verteilung auf einer Zeitachse von 0 bis 3, und zwar für die Parameterwerte $\mu = 0$ und $\sigma = 1$.

6.5 Mehrere Zielzustände

Bisher haben wir Episoden betrachtet, bei denen es nur einen möglichen Folgezustand gibt. Wenn es zwei oder mehr mögliche Folgezustände gibt, muß die zweidimensionale Variable (T, D) betrachtet werden. Wie wir uns schon überlegt haben, kann ihre Verteilung durch zielzustandsspezifische Raten charakterisiert werden:

$$r_d(t) \qquad \text{für } \; d \in \tilde{D}$$

Um zu Verteilungsmodellen zu gelangen, kann man also die in den vorangegangenen Abschnitten diskutierten Annahmen über Raten einfach auf die zielzustandsspezifischen Raten übertragen. Dabei hat man die Möglichkeit, für alle Raten das gleiche Verteilungsmodell, aber mit jeweils unterschiedlichen Verteilungsparametern, zu verwenden; man kann aber auch für jede zielzustandsspezifische Rate ein eigenes Verteilungsmodell annehmen.

Im einfachsten Fall könnte man zum Beispiel für alle zielzustandsspezifischen Raten zeitkonstante Werte annehmen, also

$$r_d(t) = \theta_t$$

6.6 Mischungen

Mischungsmodelle ergeben sich, wenn man annimmt, daß die Gesamtheit Ω aus zwei oder mehr Teilgesamtheiten besteht, bei denen es jeweils unterschiedliche Ratenverläufe gibt. Nehmen wir an, daß es m Teilgesamtheiten gibt. Zum Zeitpunkt $t = 0$ sei der Anteil der j.ten Teilgesamtheit durch π_j gegeben, also

$$\sum_{j=1}^{m} \pi_j \;=\; 1$$

$r_j(t)$ sei die Rate, $f_j(t)$ die Dichtefunktion und $G_j(t)$ die Survivorfunktion für die j.te Teilgesamtheit. Der Anteil der j.ten Teilgesamtheit entwickelt sich dann im Zeitablauf entsprechend

$$\pi_j G_j(t) \Big/ \sum_{k=1}^{m} \pi_k G_k(t)$$

wobei

$$G(t) = \sum_{j=1}^{m} \pi_j G_j(t)$$

die Survivorfunktion in der Gesamtheit ist. Daraus findet man die Dichtefunktion

$$f(t) = -\frac{dG(t)}{dt} = \sum_{j=1}^{m} \pi_j \left(-\frac{dG_j(t)}{dt} \right) = \sum_{j=1}^{m} \pi_j f_j(t)$$

Schließlich findet man für die durchschnittliche Rate in der Gesamtheit den Ausdruck

$$r(t) = \frac{\sum_{j=1}^{m} \; \pi_j f_j(t)}{\sum_{j=1}^{m} \; \pi_j G_j(t)}$$

AUFGABE 6.11 Nehmen Sie an, daß es zwei Teilgesamtheiten gibt und $\pi_1 = \pi_2 = 0.5$ ist. Nehmen Sie außerdem an, daß es in der ersten Teilgesamtheit eine konstante Rate $r_1(t) = 1$, in der zweiten Teilgesamtheit eine konstante Rate $r_2(t) = 2$ gibt. Berechnen Sie die Entwicklung der durchschnittlichen Rate im Zeitablauf und zeigen Sie, daß diese Rate immer kleiner wird.

Kapitel 7

Ratenmodelle mit Kovariablen

In diesem Kapitel werden Ratenmodelle mit zeitunabhängigen Kovariablen behandelt. Wie im vorangegangenen Kapitel beziehen wir uns auf Episoden, die wir durch eine zweidimensionale Variable (T, D) repräsentieren. Allerdings müssen wir jetzt berücksichtigen, daß es noch Kovariablen gibt. Als Ausgangspunkt haben wir also eine Variable (T, D, X), wobei X die Kovariable(n) bezeichnet. Im allgemeinen kann es sich bei X um eine ein- oder mehrdimensionale Variable handeln. Zunächst behandeln wir Modelle für Episoden, bei denen es nur einen möglichen Folgezustand gibt; Modelle für Episoden mit mehreren Folgezuständen werden in Abschnitt 7.4 besprochen.

7.1 Das Exponentialmodell

Beim Exponentialmodell wird für die Verweildauerverteilung eine Exponentialverteilung unterstellt, also eine zeitkonstante Rate

$$r(t) = \theta$$

Die Idee ist, daß diese Rate von den Werten von Kovariablen abhängig sein kann. Da die Rate nur positive Werte annehmen kann, sollte eine Link-Funktion verwendet werden, die dies garantiert. Beim Standardmodellansatz wird als Link-Funktion eine Exponentialfunktion verwendet. Wenn X eine eindimensionale Kovariable ist, sieht das Modell dann so aus:

$$r(t \,|\, X = x) \;=\; \exp(\beta_0 + x\beta_1)$$

Wenn es m Kovariablen $(X_1, \ldots, X_m)$ gibt, kann man folgende allgemeine Formulierung verwenden:

$$r(t \,|\, X_1 = x_1, \ldots, X_m = x_m) \;=\; \exp(\beta_0 + x_1\beta_1 + \ldots + x_m\beta_m)$$

AUFGABE 7.1 Beweisen Sie folgende Formel für den Mittelwert einer

Box 7.1 Datensatz 6

ID	T	D	X1	X2
1	17	1	8	1
2	5	0	5	0
3	22	1	9	1
4	13	1	7	0
5	2	0	5	0
6	9	1	6	1
7	12	0	5	1
8	15	1	7	1

Exponentialverteilung mit dem Parameter θ:[1]

$$E(T) = \int_0^\infty t\, f(t)\, dt = \int_0^\infty \theta t \exp(-\theta t)\, dt = \frac{1}{\theta}$$

AUFGABE 7.2 Finden Sie eine Formel für den Median einer Exponentialverteilung mit dem Parameter θ.

7.2 Parameterschätzungen

Wir nehmen an, daß uns n Beobachungen

$$(t_i, d_i, x_{i1}, \dots, x_{im}) \qquad \text{für } i = 1, \dots, n$$

gegeben sind. t_i ist der Wert der Verweildauer T, d_i gibt an, ob es sich um eine zensierte ($d_i = 0$) oder unzensierte ($d_i = 1$) Beobachtung handelt, und $x_{i1}, \dots, x_{im}$ sind die Werte der Kovariablen. Zur Modellschätzung wird die ML-Methode verwendet. Für die Bildung der Likelihood-Funktion wird bei nicht-zensierten Beobachtungen die Dichtefunktion

$$f(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im})$$

[1] Verwenden Sie folgende Regel für partielle Integration:

$$\int F(t) g(t)\, dt = F(t)G(t) - \int f(t) G(t)\, dt$$

wobei $f(t) = dF(t)/dt$ und $g(t) = dG(t)/dt$. Setzen Sie $F(t) = \theta t$, $g(t) = \exp(-\theta t)$.

und bei zensierten Beobachtungen die Survivorfunktion

$$G(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im})$$

verwendet; sie sieht also folgendermaßen aus:

$$\mathcal{L}(\beta_0, \dots, \beta_m) = \prod_{i=1}^n f(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im})^{d_i} \\ G(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im})^{1-d_i}$$

Wegen der Beziehung $r(t) = f(t)/G(t)$ kann man diese Likelihood-Funktion auch folgendermaßen schreiben:

$$\mathcal{L}(\beta_0, \dots, \beta_m) = \prod_{i=1}^n r(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im})^{d_i} \\ G(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im})$$

Wenn wir uns jetzt auf das in Abschnitt 7.1 besprochene Exponentialmodell beziehen, finden wir zunächst:

$$r(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im}) = \exp(\beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m)$$

$$G(t_i \,|\, X_1 = x_{i1}, \dots, X_m = x_{im}) = \\ \exp\{-\exp(\beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m)\, t_i\}$$

Daraus findet man die Likelihood-Funktion

$$\mathcal{L}(\beta_0, \dots, \beta_m) = \prod_{i=1}^n \exp(\beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m)^{d_i} \\ \exp\{-\exp(\beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m)\, t_i\}$$

Einen einfacheren Ausdruck erhält man, wenn man zur Log-Likelihood-Funktion übergeht:

$$\ell(\beta_0, \dots, \beta_m) = \sum_{i=1}^n d_i(\beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m) - \\ \exp(\beta_0 + x_{i1}\beta_1 + \ldots + x_{im}\beta_m)\, t_i$$

Im allgemeinen, wenn das Modell Kovariablen enthält, kann man das Maximum dieser Funktion nicht auf analytischem Wege finden, sondern benötigt ein iteratives Verfahren.[2] Einfach geht es nur, wenn das Modell keine Kovariablen enthält. Die Log-Likelihood-Funktion sieht dann

[2] Das wird in einigen Übungen im Anhang besprochen.

folgendermaßen aus:

$$\ell(\beta_0) = \sum_{i=1}^{n} d_i \beta_0 - t_i \exp(\beta_0)$$

Daraus gewinnt man den Gradienten, also die erste Ableitung

$$\frac{\partial \ell(\beta_0)}{\partial \beta_0} = \sum_{i=1}^{n} d_i - t_i \exp(\beta_0)$$

Den ML-Schätzwert für β_0 findet man schließlich an der Nullstelle des Gradienten, also durch

$$\hat{\beta}_0 = \log\left(\frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} t_i}\right)$$

Aufgabe 7.3 Bilden Sie die zweite Ableitung der Log-Likelihoodfunktion für das einfache Exponentialmodell ohne Kovariablen und zeigen Sie anhand dieser Ableitung, daß diese Funktion genau ein globales Maximum hat.

Aufgabe 7.4 Nehmen Sie für die Daten in Box 7.1 ein Exponentialmodell ohne Kovariablen an. Schätzen Sie dafür die zeitkonstante Rate und gewinnen Sie daraus einen Schätzwert für die mittlere Verweildauer.

7.3 Ein allgemeiner Modellansatz

Wie im vorangegangenen Kapitel besprochen worden ist, kann man zahlreiche unterschiedliche Vorstellungen über den Ratenverlauf einer Episode bilden. Bei parametrischen Modellansätzen geht man von einer Verweildauerverteilung aus, die im allgemeinen von einem ein- oder mehrdimensionalen Parameter θ abhängt. Rate, Dichtefunktion und Survivorfunktion können also allgemein folgendermaßen geschrieben werden:

$$r(t \,|\, \theta),\ f(t \,|\, \theta),\ G(t \,|\, \theta)$$

Kovariablen können in einen solchen Modellansatz eingebaut werden, indem man mittels einer Link-Funktion die Verteilungsparameter von den Kovariablen abhängig macht. Wenn also die Verteilung zum Beispiel zwei Parameter hat, also $\theta = (\alpha, \beta)$, kann man sie von den Kovariablen auf folgende Weise abhängig machen:

$$\alpha = g_\alpha(\alpha_0 + x_1\alpha_1 + \ldots + x_m\alpha_m)$$
$$\beta = g_\beta(\beta_0 + x_1\beta_1 + \ldots + x_m\beta_m)$$

Hierbei ist g_α die Link-Funktion für den Verteilungsparameter α, g_β die Link-Funktion für den Verteilungsparameter β. Zur Unterscheidung von den Verteilungsparametern werden $\alpha_0, \ldots, \alpha_m$ und $\beta_0, \ldots, \beta_m$ als *Modellparameter* bezeichnet. Natürlich ist es möglich, die Verteilungsparameter von unterschiedlichen Kovariablen abhängig zu machen oder auch vollständig auf Kovariablen zu verzichten.

Es sollte evident sein, wie sich dieser Ansatz für Verteilungen, die nur einen oder mehr als zwei Verteilungsparameter enthalten, modifizieren läßt. Zum Beispiel ist das in Abschnitt 7.1 behandelte Exponentialmodell ein einfacher Spezialfall dieses allgemeinen Modellansatzes.

Die Likelihood-Funktion kann dann in jedem Fall so formuliert werden, wie wir es für das Exponentialmodell gezeigt haben.

Aufgabe 7.5 Entwickeln Sie die Log-Likelihood-Funktion für ein Weibull-Modell ohne Kovariablen.

Aufgabe 7.6 Entwickeln Sie die Log-Likelihood-Funktion für ein loglogistisches Modell ohne Kovariablen.

7.4 Mehrere Folgezustände

Der im vorangegangenen Abschnitt behandelte allgemeine Modellansatz läßt sich leicht für Episoden verallgemeinern, bei denen es mehrere mögliche Folgezustände gibt. Ausgangspunkt ist eine Verteilungsannahme über die zielzustandsspezifischen Raten, die wir allgemein folgendermaßen schreiben können:

$$r_d(t \,|\, \theta_d) \qquad \text{für } d \in \tilde{D}$$

Daraus gewinnt man zunächst die Gesamtrate

$$r(t \,|\, \theta) \;=\; \sum_{d \in \tilde{D}} r_d(t \,|\, \theta_d)$$

wobei hier θ für die Gesamtheit der zustandsspezifischen Verteilungsparameter θ_d steht. Die Gesamtrate liefert die Survivorfunktion

$$G(t \,|\, \theta) \;=\; \exp\left\{-\int_0^t r(\tau \,|\, \theta)\, d\tau\right\}$$

oder

$$G(t\,|\,\theta) \;=\; \prod_{d \in \tilde{D}} G_d(t\,|\,\theta_d)$$

wobei

$$G_d(t\,|\,\theta_d) \;=\; \exp\left\{ -\int_0^t r_d(\tau\,|\,\theta_d)\,d\tau \right\}$$

ist. Um schließlich die Likelihood-Funktion zu finden, nehmen wir an, daß die Daten so gegeben sind, wie in Abschnitt 7.2 angegeben worden ist, nur daß jetzt d_i mehr als zwei mögliche Werte annehmen kann: Wenn $d_i = 0$ ist, handelt es sich um eine zensierte Beobachtung, wenn $d_i > 0$ ist, handelt es sich um einen Übergang in den Folgezustand d_i. Also kann man die Likelihood-Funktion folgendermaßen schreiben:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} G(t_i\,|\,\theta) \prod_{d \in \tilde{D}} r(t_i\,|\,\theta_d)^{I(d=d_i)}$$

wobei die Indikatorfunktion $I(d = d_i)$ den Wert 1 annimmt, wenn $d = d_i$ ist, andernfalls den Wert 0.

Aufgabe 7.7 Entwickeln Sie zunächst eine Likelihood-Funktion und dann eine Log-Likelihood-Funktion für ein Exponentialmodell, bei dem es drei verschiedene Folgezustände gibt.

7.5 Pseudo-Residuen

Im Unterschied zu gewöhnlichen Regressionsmodellen mit einer quantitativen abhängigen Variablen gibt es bei der ML-Schätzung von Ratenmodellen keine einfache Methode, um zu beurteilen, wie gut ein Modell zu den Daten paßt. Ein gewisses Hilfsmittel für die Modelldiagnostik liefern jedoch sog. Pseudo-Residuen. Um die Idee zu verdeutlichen, beziehen wir uns auf eine Episode mit einem möglichen Folgezustand. Daten seien in der Form

$$(t_i, d_i, x_i) \qquad \text{für } \; i = 1, \ldots, n$$

gegeben. Außerdem wird angenommen, daß bereits ein Modell geschätzt worden ist und die Schätzergebnisse durch

$$r(t\,|\,x;\hat{\theta}),\;\; f(t\,|\,x;\hat{\theta}),\;\; G(t\,|\,x;\hat{\theta})$$

gegeben sind, wobei $\hat{\theta}$ die geschätzten Modellparameter bezeichnet. Dieses Modell kann nun als Beschreibung eines zweistufigen Zufallsgenerators betrachtet werden, der folgendermaßen aussieht:

$$\boxed{X} \xrightarrow{\;\;x\;\;} \boxed{T_x} \longrightarrow (t, x)$$

Der erste Zufallsgenerator liefert einen Wert x für den Kovariablenvektor entsprechend der in den Daten gegebenen Verteilung von X. Der zweite Zufallsgenerator liefert eine Verweildauer t, wobei vom Modell $f(t\,|\,x;\hat{\theta})$ ausgegangen wird, also konditional auf den im ersten Schritt realisierten Wert des Kovariablenvektors. Auf diese Weise kann eine beliebige Menge von Pseudo-Beobachtungen (t_j^*, x_j^*) erzeugt werden $(j = 1, 2, 3, \ldots)$, so daß die Verteilung der x_j^* mit der durch die Daten gegebenen Verteilung von X und die konditionale Verteilung von t_j^* mit dem Modell übereinstimmt. Für jeden möglichen Wert x aus dem Merkmalsraum der Kovariablen X kann man die auf diese Weise erzeugten Werte als Realisierungen einer Zufallsvariablen T_x ansehen, deren Verteilung durch das Modell, d.h. durch $f(t\,|\,x;\hat{\theta})$ definiert ist.

Im nächsten Schritt wird eine Transformation der Zufallsvariablen T_x betrachtet, so daß die Abhängigkeit vom jeweils realisierten Wert des Kovariablenvektors verschwindet. Als Transformation wird

$$T_x \longrightarrow J(T_x)\,, \;\text{ definiert durch }\; t \longrightarrow J(t) = \int_0^t r(\tau \mid x;\hat{\theta})\,\mathrm{d}\tau$$

verwendet, denn die Verteilung von $J(T_x)$ ist dann eine Standard-Exponentialverteilung, d.h. eine Exponentialverteilung mit der konstanten Rate 1. Man sieht das folgendermaßen, wobei ausgenutzt wird, daß es sich um eine monotone Transformation handelt:

$$\begin{aligned}
\mathrm{P}\,(J(T_x) > t) &= \mathrm{P}\,(T_x > J^{-1}(t)) \\
&\equiv G(J^{-1}(t) \mid x;\hat{\theta}) \\
&= \exp\Big\{ -\int_0^{J^{-1}(t)} r(\tau \mid x;\hat{\theta})\,\mathrm{d}\tau \Big\} \\
&= \exp\big\{ -J(J^{-1}(t)) \big\} \\
&= \exp(-t)
\end{aligned}$$

Die Survivorfunktion für die transformierte Zufallsvariable $J(T_x)$ ist also die Survivorfunktion einer Standard-Exponentialverteilung und mithin unabhängig von x.[3]

Diese Tatsache kann ausgenutzt werden, um zu prüfen, ob es plausibel erscheint, daß die als Stichprobe vorliegenden Daten aus dem durch das Modell beschriebenen Zufallsgenerator stammen könnten. Wenn dies der Fall ist, müßte die Anwendung der oben beschriebenen Transformation auf die vorliegenden Daten zu einer Menge transformierter Größen

$$e_i = J(t_i)$$

führen, die näherungsweise einer Standard-Exponentialverteilung folgen. Die Größen e_i werden als *Pseudo-Residuen*, gelegentlich auch als *verallgemeinerte Residuen*, bezeichnet. Um zu prüfen, ob sie näherungsweise einer Standard-Exponentialverteilung folgen, kann ihre Survivorfunktion berechnet werden. Dafür sollte das Kaplan-Meier-Verfahren verwendet werden, um berücksichtigen zu können, daß einige Beobachtungen und mithin die ihnen korrespondierenden Residuen rechts zensiert sein können. Wenn $G_r(t)$ die auf diese Weise berechnete Survivorfunktion bezeichnet, kann man schließlich die Abbildung

$$t \longrightarrow -\log\{G_r(t)\}$$

betrachten. Wenn die Residuen näherungsweise einer Standard-Exponentialverteilung folgen, sollte diese Abbildung näherungsweise einer 45°-Linie entsprechen.

Aufgabe 7.8 Entwickeln Sie eine Formel zur Berechnung von Pseudo-Residuen für ein Exponentialmodell mit Kovariablen.

Aufgabe 7.9 Berechnen Sie die Pseudo-Residuen für das in Aufgabe 7.4 geschätzte Exponentialmodell ohne Kovariablen.

[3] Es sei angemerkt, daß die Verteilung von T_x durch $G(\cdot \,|\, x, \theta)$, nicht durch durch $G(\cdot \,|\, x, \hat{\theta})$, definiert ist. Das Verfahren beruht jedoch darauf, daß θ durch die mit den Daten geschätzten Parameterwerte ersetzt wird.

Kapitel 8

Zeitveränderliche Kovariablen

Im vorangegangenen Kapitel wurde diskutiert, wie Ratenmodelle mit zeitunabhängigen Kovariablen formuliert und geschätzt werden können. Für viele Anwendungen ist es jedoch erforderlich, auch Kovariablen zu berücksichtigen, die ihre Werte während einer laufenden Episode verändern können. Man möchte dann herausfinden, wie die Rate für den Übergang in einen neuen Folgezustand von den jeweils aktuellen, sich während des Episodenverlaufs verändernden Kovariablenwerten bedingt wird. In diesem Kapitel wird besprochen, wie dafür geeignete Ratenmodelle formuliert werden können.

8.1 Konditionale Survivorfunktionen

Als Ausgangspunkt erinnern wir an unsere Basisformel für den Zusammenhang zwischen Rate und Survivorfunktion, nämlich

$$G(t) = \exp\left\{-\int_0^t r(\tau)\, d\tau\right\}$$

Eine *konditionale Survivorfunktion* wird dementsprechend durch

$$G(t \,|\, s) = \exp\left\{-\int_s^t r(\tau)\, d\tau\right\}$$

definiert. Offenbar gilt:

$$G(t) = G(t \,|\, s)\, G(s)$$

Dies läßt sich beliebig wiederholen. Wir können zum Beispiel die Zeitspanne von 0 bis t in k beliebige Subintervalle aufteilen:

$$0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = t$$

Dann finden wir

$$G(t) = \prod_{j=1}^{k} G(t_j \,|\, t_{j-1})$$

8.2 Reformulierte Likelihoodfunktion

Erinnern wir uns jetzt an die Likelihoodfunktionen zur Schätzung von Ratenmodellen. Bei einer Episode mit einem möglichen Folgezustand ist die allgemeine Formulierung:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} r(t_i \,|\, \theta)^{d_i}\, G(t_i \,|\, \theta)$$

Für jedes Individuum i kann man jetzt die Verweildauer t_i in beliebig viele Subintervalle aufteilen, etwa

$$0 = t_{i0} < t_{i1} < \cdots < t_{i,k_i-1} < t_{ik_i} = t_i$$

Dann kann man die Likelihoodfunktion folgendermaßen reformulieren:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} r(t_i \,|\, \theta)^{d_i} \prod_{j=1}^{k_i} G(t_{ij} \,|\, t_{i,j-1}, \theta)$$

Wichtig ist, daß sich durch diese Reformulierung tatsächlich nur die äußere Form der Likelihoodfunktion verändert, nicht jedoch der funktionale Zusammenhang zwischen Modellparametern und Werten der Likelihoodfunktion. Zur Modellschätzung kann man deshalb ebensogut die reformulierte Likelihoodfunktion verwenden.

Aufgabe 8.1 Entwickeln Sie eine reformulierte Likelihoodfunktion zur Schätzung eines einfachen Exponentialmodells ohne Kovariablen. Zeigen Sie, daß man mit der reformulierten Likelihoodfunktion die gleichen Schätzwerte erhält wie mit der usprünglichen Likelihoodfunktion.

8.3 Zeitveränderliche Indikatorvariablen

Jetzt nehmen wir an, daß die Daten zeitveränderliche Kovariablen enthalten und in folgender Form gegeben sind:

$$(t_i, d_i, x_i, z_{i1}(t), \ldots, z_{im}(t)) \qquad \text{für } i = 1, \ldots, n$$

x_i ist wie bisher ein Vektor mit zeitunabhängigen Kovariablen, deren Werte also spätestens zu Beginn der Episode feststehen. Die Variablen $z_{ij}(t)$ sind dagegen zeitveränderlich. Und zwar nehmen wir an, daß es sich um zeitveränderliche 0-1-Variablen handelt, das heißt: bis zu einem

Box 8.1 Datensatz 7: Episodensplitting

```
 ID  ORG  DES  TS  TF        ID  SPN  ORG  DES  TS  TF
--------------------        ------------------------
  1    0    1   0  10         1    1    0    0   0   3
  2    0    0   0  12         1    2    0    0   3   6
                              1    3    0    1   6  10
                              2    1    0    0   0   5
                              2    2    0    0   5   9
                              2    3    0    0   9  11
                              2    4    0    0  11  12
```

gewissen Zeitpunkt haben sie den Wert 0, dann den Wert 1. Es sei t_{ij} der Zeitpunkt, bei dem die Variable $z_{ij}(t)$ ihren Wert von 0 auf 1 wechselt. Wir können dann die Gesamtheit der Zeitpunkte innerhalb der Episoden $[0, t_i]$ betrachten, zu denen mindestens eine der zeitveränderlichen Kovariablen ihren Wert verändert.[1] Wir stellen uns vor, daß diese Zeitpunkte folgendermaßen der Größe nach geordnet sind:

$$0 = \tau_{i0} < \tau_{i1} < \cdots < \tau_{i,k_i-1} < \tau_{ik_i} = t_i$$

Dann kann man die Likelihood unter Berücksichtigung der zeitunabhängigen und der zeitveränderlichen Kovariablen folgendermaßen schreiben:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} r(t_i \,|\, z_i, x_{i1}(t_i), \ldots, x_{im}(t_i), \theta)^{d_i}$$
$$\prod_{j=1}^{k_i} G(\tau_{ij} \,|\, \tau_{i,j-1}, z_i, x_{i1}(\tau_{i,j-1}), \ldots, x_{im}(\tau_{i,j-1}), \theta)$$

Offenbar berücksichtigt man dadurch für jedes Teilstück des Episodenverlaufs die dafür gegebenen aktuellen Werte der Kovariablen.

8.4 Episodensplitting

Die praktische Umsetzung der eben beschriebenen Idee zur Berücksichtigung zeitveränderlicher Kovariablen geschieht mit der Methode des Episodensplitting. Wir erklären die Methode zunächst an einem Datensatz

[1] Es ist klar, daß zeitveränderliche Kovariablen, die ihre Werte während des Episodenverlaufs nicht verändern, wie zeitunabhängige Kovariablen behandelt werden können.

Box 8.2 Datensatz 8: Episodensplitting

ID	ORG	DES	TS	TF	X
1	0	1	0	10	7
2	0	1	0	8	-1
3	0	1	0	5	8
4	0	0	0	12	9

ID	SPN	ORG	DES	TS	TF	D
1	1	0	0	0	7	0
1	2	0	1	7	10	1
2	1	0	1	0	0	1
3	1	0	1	0	5	0
4	1	0	0	0	9	0
4	2	0	0	9	12	1

ohne Berücksichtigung von Kovariablen. Box 8.1 zeigt den Datensatz 7, zunächst in ungesplitteter Form auf der linken Seite. Es gibt zwei Episoden. Die Episode für das erste Individuum endet mit einem Ereignis, die für das zweite Individuum ist rechts zensiert. Die rechte Seite zeigt, wie die Episoden gesplittet worden sind; in diesem Beispiel ganz willkürlich, die erste Episode in drei, die zweite in vier Splits.

Box 8.2 zeigt ein zweites Beispiel mit einer zeitveränderlichen Variablen (`X`); diese Variable enthält zunächst den Zeitpunkt, zu dem die korrespondierende 0-1-Variable (`D`) ihren Wert von 0 auf 1 verändert. Die rechte Seite der Box zeigt dann den gesplitteten Datensatz.

Appendix A

Exercises with TDA

The following exercises are intended to supplement the more theoretical discussion in the previous chapters with some computer exercises. We assume that participants have access to the computer program TDA (Transition Data Analysis). This program is in the public domain and can be obtained from our home page, `www.stat.ruhr-uni-bochum.de`.

EXERCISE A.1 Assuming that you have access to a computer, begin with creating an environment for work on the exercises.

a) Create your private working directory. (Always work only in this private directory!)

b) Try to execute TDA. Simply type `tda`. The program should show up with a short message.

c) Invoke TDA in *interactive mode*. Simply type

```
tda i
```

The program should show up with a command line, beginning with a colon, that allows you to enter commands. Try simple commands like ‘`time`’ or ‘`mpr(3+4)`’. *Don't forget that each command must be finished by a semicolon.*

d) Try the ‘`help;`’ command.

e) Leave the program with ‘`quit;`’ or ‘`exit;`’.

EXERCISE A.2 Most often we shall use TDA in *batch mode*. This means that one first creates a *command file* containing the commands to be executed by the program and then calls the program to execute the commands in the command file.

a) Become familiar with one of the editors that you can find on your computer.

b) Create a command file, say `my.cf`, containing some commands to be executed by TDA.

Box A.1 Data file `eha1.dat`

```
ID   DUR   CEN
--------------
 1    17     1
 2     5     0
 3    22     1
 4    13     1
 5     2     0
 6     9     1
 7    12     0
 8    15     1
```

c) Invoke TDA to execute the commands in the following way:

```
tda cf=my.cf
```

The program should then show the results on the screen.

d) You can save the results into an output file by invoking the program in the following way:

```
tda cf=my.cf > out
```

Try this and investigate the contents of `out`.

EXERCISE A.3 Create a data file that contains the data shown in Box A.1. Then create a TDA command file that performs the following tasks.

a) Create an internal data matrix, using the `nvar` command.

b) Create a frequency distribution of the censoring variable, `CEN`, using the `freq` command.[1]

c) Calculate the mean value of uncensored durations, using first the `tsel` command to select uncensored cases and then the `dstat` command for descriptive statistics.

Solution: `eha1.cf`

EXERCISE A.4 Create a data file that contains the data shown in Box A.2. Then create a TDA command file that performs the following tasks.

[1] Remember the convention: `CEN`=0 if the observation is censored, `CEN`=1 if the observation is not censored.

Box A.2 Data file `eha2.dat`

```
ID     X
---------
 1    17
 2    -5
 3    22
 4    13
 5    -2
 6     9
 7   -12
 8    15
```

a) Create an internal data matrix, using the `nvar` command.

b) Create new variables, `DUR` and `CEN`, where `DUR` is the absolute value of `X` and `CEN`=1 if `X` is positive and `CEN=0` if `X` is negative.[2]

c) Create a new data file that contains TDA's internal data matrix, using the `pdata` command.

d) Create another new data file that contains only the variables `ID`, `DUR`, and `CEN`, using the `pdata` command and, in addition, the `keep` parameter. This output file should be identical with `eha1.dat` as shown in Box A.1.

Solution: `eha2.cf`

EXERCISE A.5 Use TDA's `edef` command to create an episode data structure based on data file `eha1.dat` (Box A.1). Try two different ways to do this.

a) Origin state is 0, destination state is 1.

a) Origin state is 3, destination state is 9.

Solution: `eha3.cf`

EXERCISE A.6 Having defined an episode data structure with the `edef` command, one can use the `epdat` command to write the episode data into an output file. In addition, one can request a TDA command file

[2] This can be done by defining new variables inside the first `nvar` command, or by using a new `nvar` command.

that describes the data in the output file and can be used to create a new internal data matrix.

a) Try the `epdat` command with the episode data structure created in the previous exercise.

b) Use the command file created by the `epdat` command to read the output file into a new internal data matrix.

Solution: `eha27.cf`

EXERCISE A.7 Use the data shown in Box A.1 and calculate the Kaplan-Meier survivor function for the variable `DUR`.

a) Do this with paper and pencil.

b) Do this with TDA's `ple` command.

   Solution: `eha4.cf`

Check whether you get the same result.

EXERCISE A.8 Consider the output file that you got from the `ple` command in the previous exercise.

a) Calculate an estimate of the median of the distribution by using linear interpolation of the survivor function values. You should get the same result as written at the end of the file (14.2, in this example).

b) Will it always be possible to estimate the median of the distribution?

c) Use the `qo` and `qt` parameters that are offered by the `ple` command to create a table with quantiles.[3]

   Solution: `eha5.cf`

[3]Both, `qo` and `qt`, are optional parameters for the `ple` command, but only one of these parameters can be used in each `ple` command. `qt` must be given with a sequence of time points,

`qt` $= t_1, t_2, t_3, \ldots$

and then provides the corresponding values of the estimated survivor function. `qo` must be given with a *descending* sequence of values between 1 and 0 and then provides the corresponding quantiles.

EXERCISE A.9 Use the output file from the `ple` command in exercise A.7 to create a plot of the survivor function. The steps are:

a) Use the `nvar` command to create an internal data matrix that contains variables for t and $\hat{G}(t)$, as found in the output file from the `ple` command.

b) Use the `xplot` command to create a PostScript file. For example, if the variables are called `T` and `G`, use

```
xplot = T,G;
```

   to create a scatter plot, or

```
xplot(opt=2) = T,G;
```

   to create a line plot.

c) Use the `xshow` command to see the plot on the screen.

See `eha6.cf` for an example. Also try to use this in interactive mode.

EXERCISE A.10 The `xplot` command is mainly intended for interactive use. In order to use all commands that TDA offers to create PostScript plots one should work in batch mode. While we do not intend here to discuss the creation of PostScript plots systematically, you may find an example in the command file `eha7.cf`.

EXERCISE A.11 Create a macro (see `help macro`) that can be used to plot a survivor function that has been estimated with the `ple` command. See the file `macro1.cf` for an example. Assume that you have used the `ple` command to create an output file, say `ple1.out` (see exercise A.7). You may then use the macro in interactive mode, or simply by calling TDA in the following way:

```
tda cf=macro1.cf Plotple=ple1.out
```

Notice that a macro must first be loaded before it can be used.

EXERCISE A.12 The Kaplan-Meier procedure does not directly provide estimates of the rate. An estimate of the rate can be recovered, however, by differentiating a smoothed version of the estimated survivor function. Use paper and pencil to become familiar with this idea.

Box A.3 Data file `eha3.dat`

```
 ID   DUR   CEN
 -------------
  1    17     1
  2     5     0
  3    22     2
  4    13     1
  5     2     0
  6     9     2
  7    12     0
  8    15     1
  9    13     2
 10     8     2
 11    11     1
 12     8     1
```

a) Use the results from exercise A.7 and plot a smoothed version of the survivor function, say $\hat{G}_s(t)$.

b) Graphically differentiate $-\hat{G}_s(t)$ to get an estimate of the density function, say $\hat{f}_s(t)$.

c) Plot $\hat{f}_s(t)/\hat{G}_s(t)$ to get an idea about the rate function.

Exercise A.13 Calculate lower and upper bounds for the Kaplan-Meier estimate of the survivor function (exercise A.7).

a) Calculate a lower bound by assuming that censored observations end with the observed censored duration.

b) Calculate an upper bound by assuming that all censored observations end at the longest observed duration.

c) Create a plot that shows the Kaplan-Meier estimate of the survivor function and its bounds.

Solution: `eha10.cf`

Exercise A.14 Box A.3 shows a data file where episodes may end in one of two different destination states, 1 and 2.

a) Create a data file, `eha3.dat`, that contains these data.

b) Create a command file to set up a corresponding episode data structure. Use the `nvar` command to create an internal data matrix, then use the `edef` command to create an episode data structure with two destination states.

c) Create another episode data structure that recognizes only a single destination state (1 or 2).

Solution: `eha8.cf`

Exercise A.15 Use the episode data structures created in the previous exercise.

a) Based on the first episode data structure that distinguishes two different destination states, use the `ple` command to estimate corresponding sub-survivor functions.

b) Based on the second episode data structure that combines both destination states into a single one, use the `ple` command to estimate a standard survivor functions.

c) Check that the relationship is not additive, but multiplicative:

$$\hat{G}(t) \approx \hat{G}_1(t)\,\hat{G}_2(t)$$

Solution: `eha9.cf`

Exercise A.16 Consider the data in Box A.1. Assume that you can only observe events if they occur at time point 10 or later, resulting in so-called *left truncated* data.

a) Set up a command file that uses only those cases from `eha1.dat` where `DUR` is at least 10.

b) Set up an episode data structure for left truncated data by explicitly providing a positive value (10, in this example) for the starting time.

c) Use the `ple` command to get a Kaplan-Meier estimate of the survivor function for the left truncated data.

d) Compare the result with the survivor function that was estimated from the complete data set. Verify that you have estimated

$$\mathrm{P}(T > t \mid T \geq 10) \;=\; \mathrm{P}(T > t)/\mathrm{P}(T \geq 10)$$

Solution: `eha11.cf`

EXERCISE A.17 Create a table that contains three columns:

a) Values of a time variable, $t = 0\,(0.1)\,5$.

b) Corresponding values of the survivor function of an exponential distribution with $\theta = 2$.

c) Corresponding values of the density function.

Solution: `eha12.cf`

EXERCISE A.18 Create a plot for the survivor function of an exponential distribution with $\theta = 2$, in the range $0 \le t \le 4$.

Solution: `eha13.cf`. To see the plot, use TDA in interactive mode. First, use

```
xopen = plot4.ps;
```

to make `plot4.ps` (or whatever the name of your PostScript file) the currently active plot file. Then use `xshow` to see the plot. Alternatively, you can call TDA as

```
tda xopen=plot4.ps xshow
```

EXERCISE A.19 Consider fitting an exponential distribution to the data shown in Box A.1. The maximum likelihood estimate of the parameter, say θ, is given by

$$\hat{\theta}_{\mathrm{ML}} = \frac{N_u}{T_w} \tag{A.0.1}$$

where N_u is the number of uncensored observations and T_w is the summed duration of all observations. Calculate $\hat{\theta}_{\mathrm{ML}}$ for the data in Box A.1.

EXERCISE A.20 Use TDA's `rate` command to fit an exponential distribution to the data in Box A.1. The command is `rate=2` to estimate a model without covariates. The model is then parameterized as

$$\theta = \exp(\alpha)$$

where θ is the parameter of an exponential distribution. The command provides an ML estimate for the model parameter, α. Check whether you get the same result as you have found in the previous exercise.

Solution: `eha14.cf`

EXERCISE A.21 Derive the likelihood for fitting an exponential distribution to censored data. The parameterization should be

$$\theta = \exp(\alpha)$$

where θ refers to the parameter of the exponential distribution and α is the parameter to be estimated. If `DUR` denotes the duration and `CEN` the censoring indicator, the contribution of the ith observation to the log-likelihood should then be derivable as

$$\texttt{CEN}(i)\,\alpha - \mathrm{DUR}(i)\,\exp(\alpha)$$

Use TDA's `fml` command to maximize the log-likelihood function (i.e., the sum over all individual contributions to the log-likelihood) and find an estimate of α. Compare the result with the estimate found in exercise A.20.

Solution: `eha36.cf`

EXERCISE A.22 When fitting transition rate models to single episode data, TDA's `rate` command uses the likelihood

$$\mathcal{L}(\theta) = \prod_{i \in \mathcal{E}} f(t_i;\theta) \prod_{i \in \mathcal{Z}} G(t_i \,|\, s_i;\theta) \tag{A.0.2}$$

where $\mathcal{E}$ and $\mathcal{Z}$ denote, respectively, the index sets for the uncensored and censored observations; t_i is the ending time and s_i is the starting time in the ith observation. $G(t\,|\,s;\theta)$ denotes the conditional survivor function, defined by

$$G(t \,|\, s;\theta) = \frac{G(t;\theta)}{G(s;\theta)}$$

a) Consider the corresponding likelihood function for the exponential model and derive that parameter estimates will not change when one adds a constant value to all starting and ending times.

Box A.4 Data file `eha8.dat`

ID	TS	TF	CEN
1	0	10	0
1	10	17	1
2	0	3	0
2	3	5	0
3	0	11	0
3	11	22	1
4	0	12	0
4	12	13	1
5	0	1	0
5	1	2	0
6	0	6	0
6	6	9	1
7	0	8	0
7	8	12	0
8	0	10	0
8	10	15	1

b) Check whether TDA does this correctly by adding a constant value, say 10, to the starting and ending times of the durations in Box A.1. This can be done by modifying the command file `eha14.cf` discussed in exercise A.20.

Solution: `eha30.cf`

Exercise A.23 The fact that TDA uses the likelihood function (A.0.2) allows to apply the so-called method of *episode splitting*. Assume that an observation has starting time s_i and ending time t_i. Its contribution to the likelihood should therefore be $G(t_i \mid s_i;\theta)$.[4] Now, the same contribution can also be given by

$$G(t_i \mid s_i;\theta) = G(t_i \mid \tau_i;\theta)\, G(\tau_i \mid s_i;\theta)$$

where τ_i is some time point that splits the period from s_i to t_i into two parts ($s_i < \tau_i < t_i$). For example, consider the data in Box A.4. These data have been derived from the data in Box A.1 by arbitrarily splitting each duration into two parts. Of course, the first part does not end in an event and should always be treated as a censored (sub-) episode.

a) Set up an episode data structure for the data shown in Box A.4 and estimate an exponential model.

[4]And, if the observation is not censored, also $f(t_i;\theta)$.

Box A.5 Data file `eha9.dat`

ID	DUR	CEN	S
1	17	1	10
2	5	0	3
3	22	1	11
4	13	1	12
5	2	0	1
6	9	1	6
7	12	0	8
8	15	1	10

b) Check whether estimation results are identical with those from exercise A.20.

Solution: `eha31.cf`

Exercise A.24 Episode splitting can be performed with the `edef` command. One only needs to supply variables containing the time points for splitting. To illustrate this option, consider the data in Box A.5. The data are identical to those in Box A.1, we only added a further column (`S`) containing time points for splitting the episodes.

a) Set up a command file that reads data file `edat9.dat` (Box A.5).

b) Use the `edef` command to create an episode data structure and the `split=S` parameter to request episode splitting at the time points given by variable `S`.

c) Use the `epdat` command to create a new output file containing the splitted episodes. Check that the resulting output file contains the same information as the data in Box A.4.

Solution: `eha32.cf`

Exercise A.25 Also TDA's Kaplan-Meier procedure uses conditional survivor functions. (See the description of the `ple` command in the manual.) One can therefore apply the `ple` command to episode data that have been splitted and should get the same result as if the episodes were not split. Check this with the data file created in the previous exercise. Assume that the command file `eha32.cf` contains the command

```
epdat(dtda=t) = eha9a.dat
```

You can then make the file `t` to become the starting point for a new command file, say `eha33.cf`, that reads the data file `eha9a.dat`, creates an episode data structure with the `edef` command, and then requests a Kaplan-Meier estimate of the survivor function with the `ple` command. The resulting survivor function should be identical with the estimate produced in exercise A.7.

Solution: `eha33.cf`

EXERCISE A.26 The technique of episode splitting is mainly used to provide a simple way of incorporating time-varying covariates. It therefore suffices to split episodes at time points where a covariate changes its value. Since episode splitting does not change the information contained in a set of episode data it is possible, however, to split episodes at each possible time point. This is sometimes done when the data are defined on a discrete time axis. It would then be possible to apply, for example, standard procedures for estimating logit and probit models.

a) Set up a command file that splits the episodes in Box A.1 at all integral time points and write the data into a new output file, say `eha1a.dat`.

   Solution: `eha34.cf`

b) Set up a command file that uses `eha1a.dat` to estimate a simple logit model for the event that occurs when an episode ends. (The command is `qreg`.) If the state space is $\{0, 1\}$, where $Y = 1$ denotes the destination state, the model would be

$$\mathrm{P}(Y = 1) = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

   Solution: `eha35.cf`

c) Calculate the estimated probability for the occurrence of an event and derive a comparable estimate from fitting an exponential transition rate model. Compare both estimates.

EXERCISE A.27 Derive the log-likelihood function for the simple logit model without covariates that was used in the previous exercise. Then use the `fml` command to estimate the parameter, α. Compare with the parameter estimate that you got in the previous exercise.

Solution: `eha37.cf`

EXERCISE A.28 Let $F(t)$ denote a distribution function. Then, if r is a random variable that is equally distributed in $[0, 1]$, $F^{-1}(r)$ is a random variable with a distribution described by F.

a) Use this idea to derive a formula that can be used to create exponentially distributed random numbers.

b) Use TDA's operator for equally distributed random numbers (`rd`) and the formula derived under (a) to create 100 exponentially distributed random numbers ($\theta = 2$).[5]

c) Fit an exponential distribution and check the estimated value of θ.

Solution: `eha15.cf`

EXERCISE A.29 Continue with the previous exercise and introduce some censored observations. One possibility is as follows: Let t_i denote the original uncensored duration for case i. Then, for each case i, draw another random number, say r_i, equally distributed in $[0, 1]$, and assume that case i is censored at duration 0.5 if $r_i \leq 0.5$ and $t_i \geq 0.5$.

Solution: `eha16.cf`

EXERCISE A.30 Create data for $n = 100$ cases. Define a dummy variable, say `GRP`, that takes the value 1 for the first 50 cases and value 0 for the remaining 50 cases. For each case create an exponentially distributed duration, $\theta = 2$ if `GRP` $= 1$ and $\theta = 3$ if `GRP` $= 0$.

a) Estimate an exponential model that contains `GRP` as a covariate. Check whether you can recover estimates of the parameters that have been used for data generation from the results of the model estimation.

   Solution: `eha17.cf`

b) Estimate an exponential model for each group separately and compare the parameter estimates with the estimates you got in (a).

   Solution: `eha18.cf`

---

[5] Use the `nvar` command. The number of cases can then be fixed with the parameter `noc=100`.

Box A.6 Data file `eha4.dat`

```
ID  T1  T2  T3  CEN
-------------------
1   50  66  73   1
2   56  71  81   0
3   45  63  88   1
4   70  87  97   1
5   72  90  -1   0
6   58  75  80   1
7   60  77  82   1
8   65  82  -1   0
```

EXERCISE A.31 Having fitted an exponential distribution to a set of durations, one can use a simple graphical method to check goodness-of-fit. The method uses the survivor function of the exponential distribution,

$$G(t) = \exp(-\theta t)$$

The graphical check uses the transformation

$$-\log(G(t)) = \theta t$$

One first estimates the survivor function non-parametrically, e.g., with the Kaplan-Meier procedure, resulting in an estimate $\hat{G}(t)$, and then plots $-\log(\hat{G}(t)$ against t. If the exponential distribution fits the data one should get, approximately, a straight line through the origin.

a) Apply this check to the data created in exercise A.28.

   Solution: `eha28.cf`

b) Apply this check to the data in Box A.1.

   Solution: `eha29.cf`

In both examples, add a straight line representing the fitted exponential distribution. For the first example, use $\theta = 2$; for the second example use $\theta = 0.0526$ as resulting from exercise A.20.

EXERCISE A.32 Consider the data shown in Box A.6. Each case is described by two, or three, dates, given in calendar time. (You may assume that `T1` records birth date, `T2` records end of schooling, and `T3` records first marriage.) `T2` is censored if `T3` is missing, `T3` is censored if `CEN` = 0. `T1` is always observed.

Box A.7 Data file and required output file

```
Data file: eha5.dat

 I   T   X
-----------
 1   3   1
 2   4   2
 3   2   7

Required output file: eha6.dat

 I   T   X  Cnt
---------------
 1   3   1   1
 1   3   1   2
 1   3   1   3
 2   4   2   1
 2   4   2   2
 2   4   2   3
 2   4   2   4
 3   2   7   1
 3   2   7   2
```

a) Create a data file that contains, for each case, information about its first episode, recorded in process time.

b) Create a data file that contains, for each case that has a second episode, information about its second episode, recorded in process time.

Solution: `eha19.cf`

EXERCISE A.33 Modify the command file that you have created in the previous exercise in order to set up an episode data structure, both for first and second episodes.

Solution: `eha20.cf`

EXERCISE A.34 The next step is to create multi-episode data. This can be done with TDA's matrix and loop commands. To learn some of these options, consider the data file, `eha5.dat`, shown in Box A.7.[6] There is an

[6]This example is taken from the paper *Using TDA Matrix Commands and Loops for Data Generation and Selection*. The paper is available in the `contrib` directory of

Box A.8 Command file `eha21.cf`

```
nvar(
    dfile = eha5.dat,
    I = c1,
    T = c2,
    X = c3,
);
mfmt = 2.0;
repeat(n = noc,Case);
    repeat(n = T(Case,1),TCnt);
        mcath(I(Case,1),T(Case,1),X(Case,1),TCnt,Tmp);
        mpra(Tmp) = eha6.dat;
    endrepeat;
endrepeat;
```

ID variable (`I`), a variable that counts time periods (`T`), and some further covariate (`X`). The file contains a single record for each ID number. Now assume that you want a new data file that contains, for each ID number i, `T`(i) records, as shown in the lower part of Box A.7. This can be done with the command file `eha21.cf` shown in Box A.8.

a) The `nvar` commands reads the input data file, `eha5.dat`, and creates the three variables, `I`, `T`, and `X`.

b) The `mfmt` command specifies a print format for the `mpra` command which is used later in the command file.

c) Then follows a `repeat` command that repeats the following commands, until the matching `endrepeat`, a number of times as defined by the `n` parameter. In this case, `n = noc`, that is, the number of cases in the data matrix. In addition, the command creates a $(1, 1)$ matrix `Case` that gets the value `Case` $= 1, \dots, n$ while being in the repeat loop.

d) Then follows a second `repeat` command where the repeat variable, `TCnt`, now runs in the range $1, \dots,$ `T(Case,1)`. The latter expression refers to the value of variable `T` in the current data matrix row as given by `Case`.

e) The inner repeat loop contains two commands. The first one, `mcath` (= horizontal concatenation), creates a row vector, `Tmp`, that consists of the current values of the three variables and, in addition, the current value of `TCnt`.

f) The second command in the inner loop, `mpra`, appends the row vector `Tmp` to the output file `eha6.dat`.

the TDA homepage. We recommend that you also study the other examples discussed in that paper.

Box A.9 Data file `eha7.dat`

```
ID NS SN  TS TF CEN
-------------------
 1  2  1  50 66  1
 1  2  2  66 73  1
 2  2  1  56 71  1
 2  2  2  71 81  0
 3  2  1  45 63  1
 3  2  2  63 88  1
 4  2  1  70 87  1
 4  2  2  87 97  1
 5  1  1  72 90  0
 6  2  1  58 75  1
 6  2  2  75 80  1
 7  2  1  60 77  1
 7  2  2  77 82  1
 8  1  1  65 82  0
```

The final result is the output file `eha6.dat` as shown in the lower part of Box A.7. Note that when running the command file, the matrix and loop commands will not, by default, give any echo in the standard output. Such an echo might be helpful when debugging a command file and can be requested with the `silent=-1` command.

EXERCISE A.35 Now try to transform the data file `eha4.dat` (Box A.6) into a multi-episode data file that should look similar to the file `eha7.dat` shown in Box A.9.

Solution: `eha22.cf`

EXERCISE A.36 Use the data file `eha7.dat`, created in the previous exercise, and the `edef` command, to set up a multi-episode data structure. This should be done on a process time axis where the first episode for each individual begins at time 0.

Solution: `eha23.cf`

EXERCISE A.37 Continue with the previous exercise and consider, for each time point on the process time axis, the cross-sectional distribution of cases in the state space, $\{0, 1, 2\}$, in this example. This will be called a *state distribution*. Use the `epsdat` command to calculate a state distribution for the time points $t = 0, 1, 2, \ldots, 50$.

Solution: `eha24.cf`

EXERCISE A.38 Continue with the multi-episode data created in exercise A.36.

a) Estimate an exponential model without covariates simultaneously for first and second episodes.

b) Estimate separate exponential models for first and second episodes.

c) Derive from the likelihoods of the models that one should get identical parameter estimates.

Solution: `eha25.cf`

EXERCISE A.39 Transform the multi-episode data created in exercise A.36 into sequence data, on a process time axis that runs from 0 to 50. Use the `seqpe` command.

Solution: `eha26.cf`

EXERCISE A.40 Consider the Weibull distribution. Create a command file that plots the survivor function of the Weibull distribution,

$$G_{a,b}(t) = \exp(-(at)^b) \tag{A.0.3}$$

for parameter values $a = 1$ and $b = 2$, in the range $0 \leq t \leq 2$. Use the `plotf` command.

Solution: `eha38.cf`

EXERCISE A.41 Continue with the Weibull distribution.

a) Derive a formula for the inverse survivor function,

$$t = \exp\left\{\frac{\log\left(-\log(G(t))/a^b\right)}{b}\right\} \tag{A.0.4}$$

b) Use this formula to create 100 random durations which are distributed according to a Weibull distribution with $a = 1$, $b = 2$.

Solution: `eha39.cf`

EXERCISE A.42 Use the random data created in the previous exercise.

a) Use the `ple` command to find a Kaplan-Meier estimate of the survivor function.

   Solution: `eha40.cf`

b) Use the macro created in exercise A.11 to see a plot of the estimated survivor function. Assuming that you have written the estimated survivor function into an output file, `wei.ple`, you may use

```
tda cf=macro1.cf Plotple=wei.ple
```

c) Create a plot that shows, simultaneously, the theoretical and the estimated survivor function.

   Solution: `eha41.cf`

EXERCISE A.43 Use TDA's `rate` command to estimate a Weibull model for the data created in exercise A.41. The model number is `rate=7`. Notice that TDA's Weibull model uses the parameterization

$$a = \exp(\alpha) \quad b = \exp(\beta)$$

Calculate the estimated values for a and b and compare with the values that were used for data generation.

Solution: `eha42.cf`

EXERCISE A.44 Consider the Weibull model parameterized with $a = \exp(\alpha)$ and $b = \exp(\beta)$.

a) Derive the log-likelihood for ML estimation of α and β.

b) Use the `fml` command to estimate α and β with the data created in exercise A.41.

   Solution: `eha43.cf`

EXERCISE A.45 Remember the graphical method to check goodness-of-fit of an exponential distribution that was discussed in exercise A.31. Think of a similar method for the Weibull distribution.

a) Derive the formula

$$\log(-\log(G(t))) = b\log(a) + b\log(t)$$

from the survivor function of the Weibull distribution.

b) Use the Kaplan-Meier estimate of the survivor function that was created in exercise A.42 to plot

$$\log(-\log(G(t))) \text{ vs. } \log(t)$$

If the Weibull model fits the data (what should be the case in this example), the plot should exhibit a straight line.

Solution: `eha44.cf`

c) Use the plot to graphically determine estimates of a and b and compare with the values that were used to create the data.

EXERCISE A.46 Continue with the Weibull distribution.

a) Derive a general formula for the median of the Weibull distribution in terms of the parameters, a and b.

b) Calculate the median of a Weibull distribution with $a = 1$ and $b = 2$.

c) Compare with the Kaplan-Meier estimate of the median that was calculated in exercise A.42.

EXERCISE A.47 We now discuss some difficulties that occur when one tries to fit a Weibull model to the data in Box A.1.

a) Try to estimate a Weibull model with TDA's `rate` command. (Use a suitably modified version of command file `eha14.cf` that was used in exercise A.20.) You will find that TDA is not able to estimate a Weibull model with these data when beginning with default starting values.

b) Fix the value for the b parameter to estimate an exponential model as a special case of the Weibull model (see exercise A.40). Since in the TDA parameterization we have $b = \exp(\beta)$, use the constraint

```
con = b2 = 0,
```

You should then get the same estimate for α as was found in exercise A.20.

Solution: `eha45.cf`

c) Now try to fix β at some other value, say $\beta = 1.5$, and check whether you get a better fit. Use the value of the log-likelihood as a criterion. Also write the estimated parameter values into an output file, say `sv`, using the `ppar` parameter for the `rate` command.

Solution: `eha46.cf`

d) Now use these parameter values as starting values to fit an unrestricted Weibull model.

Solution: `eha47.cf`

e) Use a graphical method to check whether the finally estimated Weibull model fits the data in Box A.1.

EXERCISE A.48 We now discuss some options provided by TDA's `rate` command.

a) Add the parameter

prate (tab=0(1)20) = rate.dat,

to the command file `eha47.cf` that was used in the previous exercise. You will get an output file, `rate.dat`, containing the estimated rate, survivor and density functions for the time points $t = 0, 1, \ldots, 20$.

b) Add also the parameter

pres = res.dat,

You will get an output file, `res.dat`, containing the so-called *generalized residuals*. For information about the contents of this file, see Section 6.17.1.6 of the TDA manual.

Solution: `eha48.cf`

c) Use the generalized residuals to check whether the model fits the data.

EXERCISE A.49 We now have finished with a selection of most basic exercises. It remains to apply what we have learnt to more complex data sets that also provide an opportunity to include covariates. For this task we continue with an example data set, `rrdat.1`, that provides observations of job histories for 201 individuals.[7] The variables contained in this data set are shown in Box A.10.

a) Begin with investigating the first records of the data file shown in Box A.11.

b) Set up a command file that reads the data into an internal data matrix.

c) Use the `edef` command to create different versions of single and multi-episode data structures.

d) Find survivor functions with the `ple` command.

e) Estimate transition rate models with the `rate` command.

[7]This data set has been used by Blossfeld and Rohwer (1995) and is also used in the TDA manual, Section 3.3.3.

Box A.10 Variables in data file `rrdat.1`

```
Variable  Column  Description
-----------------------------------------------------
  ID       C1     ID of individual
  NOJ      C2     Serial number of the job
  TS       C3     Starting time of the job
  TF       C4     Ending time of the job
  SEX      C5     Sex (1 men, 2 women)
  TI       C6     Date of interview
  TB       C7     Date of birth
  TE       C8     Date of entry into the labor market
  TM       C9     Date of marriage (0 if no marriage)
  PRES     C10    Prestige score of job i
  PRES1    C11    Prestige score of job i + 1
  EDU      C12    Highest educational attainment
```

Box A.11 First records of data file `rrdat.1`

```
  ID NOJ   TS   TF  SEX   TI   TB   TE   TM PRES PRES1 EDU
  --------------------------------------------------------
   1   1  555  982    1  982  351  555  679   34    -1   17
   2   1  593  638    2  982  357  593  762   22    46   10
   2   2  639  672    2  982  357  593  762   46    46   10
   2   3  673  892    2  982  357  593  762   46    -1   10
   3   1  688  699    2  982  473  688  870   41    41   11
   3   2  700  729    2  982  473  688  870   41    44   11
   3   3  730  741    2  982  473  688  870   44    44   11
   3   4  742  816    2  982  473  688  870   44    44   11
   3   5  817  828    2  982  473  688  870   44    -1   11
```

Literatur

Blossfeld, H.-P., Rohwer, G. 1995. Techniques of Event History Modeling. New Approaches to Causal Analysis, Mahwah. NJ: Lawrence Erlbaum.

Cox, D. R., Oakes, D. 1984. Analysis of Survival Data. London: Chapman and Hall.

Galton, A. 1984. The Logic of Aspect. Oxford: Clarendon.

International Statistical Institute 1986. Declaration on Professional Ethics. International Statistical Review 54, 227–242.

Kaplan, E. L., Meier, P. 1958. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 53, 457–481.

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