Allg. Statist. Archiv 73, 213—238, ISSN 0002-6018
© Vandenhoeck & Ruprecht 1989

A. Abhandlungen

Hazardraten-Modelle in den Wirtschafts- und Sozialwissenschaften*)

Von HANS-PETER BLOSSFELD, Berlin, ALFRED HAMERLE, Konstanz, und KARL ULRICH MAYER, Berlin

Zusammenfassung: In den letzten Jahren hat in der empirischen Wirtschafts- und Sozialforschung die Verfügbarkeit von Ereignissen ständig zugenommen. Diese Daten informieren über die Zeitdauer bis zu einem Zustandswechsel oder dem Eintreffen bestimmter Ereignisse. Zuerst wird ein kurzer Überblick über geeignete statistische Methoden präsentiert, insbesondere die Modellierung von Regressionsansätzen mit Hilfe der Hazard- bzw. Übergangsrate. Dann werden Schätz- und Testmethoden mit möglicherweise zensierten Daten behandelt. Schließlich wird die praktische Umsetzung der Verfahren anhand eines Beispiels zur Analyse des Berufswechselverhaltens von Männern demonstriert. Die Analyse basiert auf den Daten der Lebensverlaufsstudie, die vom Max-Planck-Institut für Bildungsforschung, Berlin, durchgeführt wird.

Summary: In economics and sociology, the increasing availability of event history data permits the application of new statistical techniques to estimat models for economic and social processes. First we present a short overview of the most important statistical concepts, in particular the hazard rate approach. Of special importance is the inclusion of explanatory variables in parametric or semiparametric regression models for durations. We discuss estimation methods and hypotheses testing in the presence of censored data. In the last section we study the risk of job change for men in the Federal Republic of Germany as an example of the application of event history analysis. The analysis is based on data from the German Life History Study.

I. Einleitung

In den letzten Jahren ist in den Wirtschafts- und Sozialwissenschaften das Interesse an der Erhebung von Längsschnittdaten deutlich gestiegen. Dabei treten neben den traditionellen Panel- oder Zeitreihenstudien zunehmend Untersuchungen in den Vordergrund, die auf Ereignisdaten zurückgreifen. Diese Daten informieren bei jeder Untersuchungseinheit über die genauen Zeitdauern bis zu einem Zustandswechsel beziehungsweise bis zum Eintreffen bestimmter Ereignisse und deren Abfolge. Beispiele hierfür sind: die Arbeitslosigkeitsphasen von Individuen in ökono-

*) Für hilfreiche Hinweise zu einer früheren Fassung dieses Aufsatzes möchten wir Herrn Professor Dr. Horst Rinne danken.

mischen Studien (vgl. z.B. Heckman/Borjas, 1980; Flinn/Heckman, 1983; Kiefer/Lundberg/Neumann, 1985; Hujer/Schneider, 1988; Hamerle, 1989); die Zeitdauer zwischen beruflichen Aufstiegen in Untersuchungen zur sozialen Mobilität (vgl. z.B.

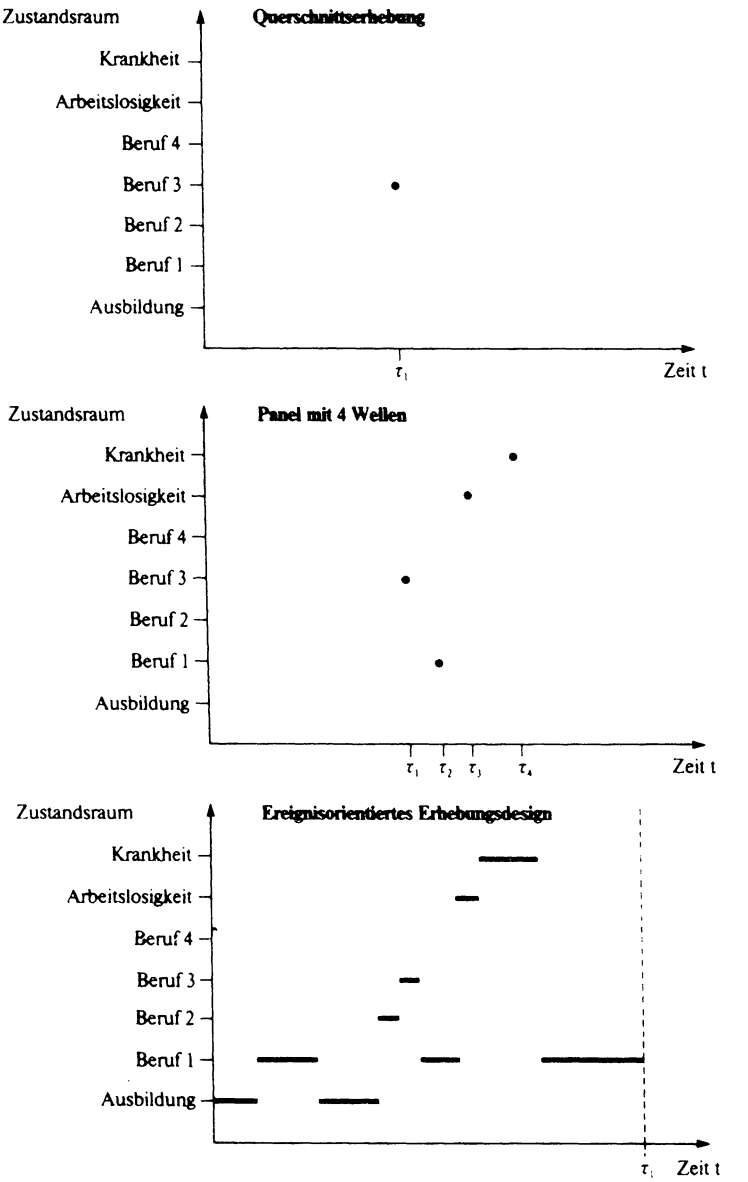

Abbildung 1:
Erfassung des Bildungs- und Berufsverlaufs einer Person mit Hilfe einer Querschnittserhebung, eines Panels und eines ereignisorientierten Erhebungsdesigns

Sørensen/Tuma, 1981; Sørensen, 1984, 1984; Tuma, 1985; Carroll/Mayer, 1986; Mayer/Carroll, 1987; Blossfeld, 1989); die Zeitdauern bis zum Erstkauf eines neu eingeführten Produkts im Marketing (vgl. Hamerle, 1987); die „Lebensdauer" von Firmen oder Arbeitsgruppe in der Organisationsforschung (vgl. z. B. Hannan/Freeman, 1977; Caroll/Dalacroix, 1982; Freeman/Carroll/Hannan, 1983; Carroll, 1984; Carroll/Huo, 1985, 1986); die Zeitspanne zwischen Wohnungswechseln oder Wanderungen in der Raumforschung (vgl. z. B. Sandefuhr/Scott, 1981; Courgeau, 1984, 1985; Wagner, 1989); das Alter bis zur Heirat und der Geburt des ersten Kindes in der Bevölkerungsforschung (vgl. z. B. Michael/Tuma, 1985; Sørensen/Sørensen, 1986; Diekmann, 1987; Huinink, 1987; Wu, 1988); die Verweildauer der Kinder im elterlichen Haushalt bis zum Auszug in der Jugend- und Familiensoziologie (vgl. z. B. Mayer/Wagner, 1986); die Dauer von Lernprozessen in der psychologischen und pädagogischen Forschung (Felmlee/Eder, 1983) usw.

Die zunehmende Beliebtheit der Erhebung von Ereignisdaten in den Wirtschafts- und Sozialwissenschaften beruht auf einer Reihe von Vorteilen, die diese Daten gegenüber Querschnitts- und Paneldaten zur Untersuchung vieler ökonomischer und sozialer Prozesse besitzen. Ein kleines Beispiel, in dem der Bildungs- und Berufsverlauf einer Person mit Hilfe eines Querschnitts, eines Panels und eines ereignisorientierten Designs erhoben wird (Abb. 1), soll dies verdeutlichen. Zur Charakterisierung des Karriereverlaufs werden dabei sieben Zustände (Ausbildung, Beruf 1, Beruf 2, Beruf 3, Beruf 4, Arbeitslosigkeit und Krankheit) unterschieden, in denen sich diese Person befinden kann.

Aus Abbildung 1 ist zunächst zu erkennen, daß in der *Querschnittserhebung* der Bildungs- und Berufsverlauf der Person nur durch einen einzigen Punkt, den Zustand zum Zeitpunkt der Stichprobenziehung, repräsentiert wird. Etwas mehr Informationen liefert das vierwellige *Panel,* in dem die Zustände der Person schon zu vier verschiedenen Zeitpunkten beobachtet werden können. Allerdings ist unklar, wie sich der Prozeß zwischen diesen vier Wellen des Panels entwickelt hat. Erst ein *ereignisorientiertes Erhebungsdesign,* bei dem die Zustandsänderungen und ihre genauen Zeitpunkte erfaßt werden, erlaubt es, den Bildungs- und Berufsverlauf in seinen einzelnen Phasen und für jeden beliebigen Zeitpunkt detailliert zu rekonstruieren.

An diesem Beispiel wird folgendes deutlich:

— Der Einsatz von Querschnittsanalysen impliziert in der Regel eine *Gleichgewichtsannahme.* Das heißt, die zu einem bestimmten Zeitpunkt sich ergebende Verteilung ist nur dann aussagekräftig, wenn der zugrundeliegende Prozeß in der Zeit einigermaßen stabil bleibt. Bei größeren Schwankungen und Wandlungsprozessen ist die Momentaufnahme eines Querschnitts nicht angemessen, weil die Analyseergebnisse dann davon abhängen, wann die Erhebung durchgeführt wurde. Panel- und ereignisorientierte Daten tragen dagegen dem *Wandel und der Dynamik* explizit Rechnung.

- Aber auch wenn in der empirischen Realität beträchtliche Stabilität vorherrschen sollte, besitzen Panel- und ereignisorientierte Daten im Vergleich zu Querschnitten den Vorteil eines *höheren Informationsgehalts.* So können Querschnittsdaten zunächst einmal als Spezialfall von Panel- und ereignisorientierten Daten angesehen werden, weil sich aus diesen Querschnitte ohne weiteres rekonstruieren lassen. Im empirischen Anwendungsfall kann überdies nur die Erhebung von Panel- oder ereignisorientierten Daten Aufschluß darüber geben, ob über die Zeit tatsächlich Stabilität vorliegt. Schließlich dürften die im Vergleich zu Querschnitten bei Panel- oder ereignisorientierten Daten vorliegenden Informationen über die Vorgeschichte dazu beitragen, die Erklärungs- und Prognosekraft statistischer Modelle zu verbessern.
- Bleibt im Panel der Verlauf zwischen den einzelnen Erhebungszeitpunkten offen, so ermöglicht das ereignisorientierte Erhebungsdesign dagegen die *Rekonstruktion des kontinuierlichen Prozesses.* Zwar ist auch das Panel zur Erfassung des zeitlichen Verlaufs geeignet, wenn die Zustandsänderungen zu fest vorgegebenen Zeitpunkten stattfinden, die mit den Erhebungsintervallen übereinstimmen (z. B. die monatliche Erfassung des Monatseinkommens), oder wenn eine kontinuierliche Variable (z. B. das Körpergewicht eines Menschen) sinnvoll nur auf der Basis zeitdiskreter Erhebungen gemessen werden kann; aber alle anderen Veränderungen nicht-metrischer Variablen, die zu beliebigen Zeitpunkten eintreten können, erfordern zur vollständigen Rekonstruktion eine genaue Registrierung von Art und Zeitpunkt der Zustandsänderungen. Das ereignisorientierte Erhebungsdesign erweist sich damit in vielen konkreten Anwendungsgebieten als das eigentlich adäquate Instrumentarium, um Wandlungsprozesse adäquat abbilden zu können.
- Denkt man schließlich an die dynamische Analyse *komplexer Kopplungs- und Rückkopplungsprozesse* im wirtschafts- und sozialwissenschaftlichen Bereich, dann scheint die kontinuierliche Erhebung nicht-metrischer Variablen mit Hilfe ereignisorientierter Designs eine wichtige Möglichkeit zu sein, empirische Wandlungs- und Veränderungstendenzen zu erfassen. Dies gilt insbesondere dann, wenn die Ereignisse dieser parallelen Prozesse nicht nur zu beliebigen Zeitpunkten eintreten, sondern darüber hinaus auch zeitverzögert aufeinander einwirken.

Die adäquate Abbildung der Veränderungen nicht-metrischer Merkmale, die zu beliebigen Zeitpunkten eintreten können, sowie der hohe Informationsgehalt von Ereignisdaten sind also große Vorzüge des ereignisorientierten Datendesigns, das dem steigenden Interesse an der Analyse von Prozessen und Verläufen in den Wirtschafts- und Sozialwissenschaften entgegenkommen. Zur Analyse dieser Datenstrukturen stellt die Statistik heute eine große Zahl von Modellen, Ansätzen und Methoden zur Verfügung, die ursprünglich vor allem aus der Medizin, der Biometrie, der Demographie und der Technik stammen (vgl. z. B. Cox/Oakes, 1984;

Elandt-Johnson/Johnson, 1980; Kalbfleisch/Prentice, 1980; Lawless, 1982) und erst in jüngster Zeit zunehmend auch von den Wirtschafts- und Sozialwissenschaften aufgegriffen und weiterentwickelt wurden (vgl. z. B. Coleman, 1981; Tuma/Hannan, 1984; Diekmann/Mitter, 1984; Allison, 1984; Heckman/Singer, 1985; Blossfeld/Hamerle/Mayer, 1986, 1989; Kiefer, 1988; Mayer/Tuma, 1989). Wir können im vorliegenden Aufsatz nur einen kurzen Überblick über einige wichtige Merkmale dieser Verfahren geben. Dazu stellen wir zunächst einige statistische Grundkonzepte der Ereignisanalyse dar und konzentrieren uns auf die Frage, wie Kovariablen in die Regressionsmodelle zur Analyse von Ereignisdaten einbezogen und die unbekannten Parameter geschätzt werden. Ein konkretes Anwendungsbeispiel soll im Anschluß daran die Umsetzung dieser Verfahren in der Forschung und die Interpretation von Analyseergebnissen demonstrieren.

II. Statistische Grundkonzepte (Ein-Episoden-Fall)

Der einfachste Fall der Ereignisanalyse liegt dann vor, wenn lediglich die Zeitdauer vom Eintritt in einen Anfangszustand bis zum Erreichen eines bestimmten Endzustandes gemessen wird. Anwendungen findet man vor allem bei der Untersuchung von Lebens- beziehungsweise Überlebenszeiten in medizinischen Studien, aber auch zum Beispiel bei der Analyse der Lebensdauer politischer oder gesellschaftlicher Organisationen. Viele der für den Ein-Episoden-Fall entwickelten statistischen Konzepte können auf komplexere Situationen, wie mehrere aufeinanderfolgende Episoden oder mehrere Endzustände (competing risks), übertragen werden (vgl. Blossfeld/Hamerle/Mayer, 1986, 1989).

A. Dichte- und Verteilungsfunktion, Survivorfunktion, Hazardrate

Im folgenden werden wichtige statistische Kenngrößen der Ereignisanalyse im Ein-Episoden-Fall mit einem Anfangs- und einem Endzustand eingeführt. Dabei wird zunächst von einer homogenen Population ausgegangen, das heißt, interindividuelle Heterogenität in bezug auf verschiedene Merkmale bleibt unberücksichtigt.

Die Dauer der Episode wird im statistischen Modell repräsentiert durch eine nicht-negative stetige Zufallsvariable T. Die *Dichte-* und die *Verteilungsfunktion* der Episodendauer T ($T \geq 0$) seien mit $f(t)$ beziehungsweise mit $F(t)$ bezeichnet. Dabei gilt wie üblich der Zusammenhang

$$F(t) = P(T \leq t) = \int_0^t f(u)\,du, \tag{2.1}$$

und an allen Stellen, an denen $f(t)$ stetig ist, gilt

$$f(t) = F'(t). \tag{2.2}$$

Die *Survivorfunktion*

$$S(t) = P(T \geq t) \tag{2.3}$$

gibt die Wahrscheinlichkeit dafür an, daß ein Individuum den Zeitpunkt t „erlebt", das heißt, daß bis zu diesem Zeitpunkt noch kein Ereignis eingetreten ist und die Episode noch andauert. Für kontinuierlich gemessene Zeitdauern gilt

$$S(t) = 1 - F(t). \tag{2.4}$$

Die Survivorfunktion ist in Abhängigkeit von der Zeit monoton fallend (vgl. Abbildung 2).

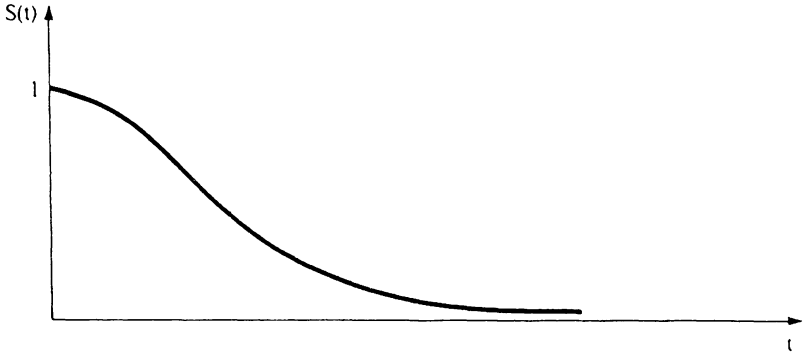

Abbildung 2: Typischer Verlauf einer Survivorfunktion

Die *Hazardrate* ist

$$\lambda(t) = \lim_{\substack{\Delta t \to 0 \\ \Delta t > 0}} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t). \tag{2.5}$$

Man beachte, daß die Werte der Hazardrate keine (bedingten) Wahrscheinlichkeiten sind. Sie sind zwar stets nicht-negativ, können aber größer als Eins sein. Für kleines Δt kann $\lambda(t)\Delta t$ als Approximation der bedingten Wahrscheinlichkeit $P(t \leq T < t + \Delta t | T \geq t)$ aufgefaßt werden. Andere Bezeichnungen für die Hazardrate sind *Intensitäts-* oder *Risikofunktion, Übergangsrate* oder *Mortalitätsrate.*
Das Integral

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.6}$$

wird als *kumulative Hazardrate* bezeichnet.

Die Hazardrate stellt ein zentrales Konzept bei der Analyse von Verlaufsdaten dar. „Überlebt" ein Individuum den Zeitpunkt t, so informiert die Hazardrate über „den weiteren Verlauf". Häufig besitzt man bei praktischen Anwendungen zumindest qualitative Vorinformationen über die Hazardrate. Dies soll an dem Beispiel des Sterberisikos einer Population verdeutlicht werden. Die Hazardrate hat hier typischerweise einen „badewannenförmigen" Verlauf (vgl. Abbildung 3).

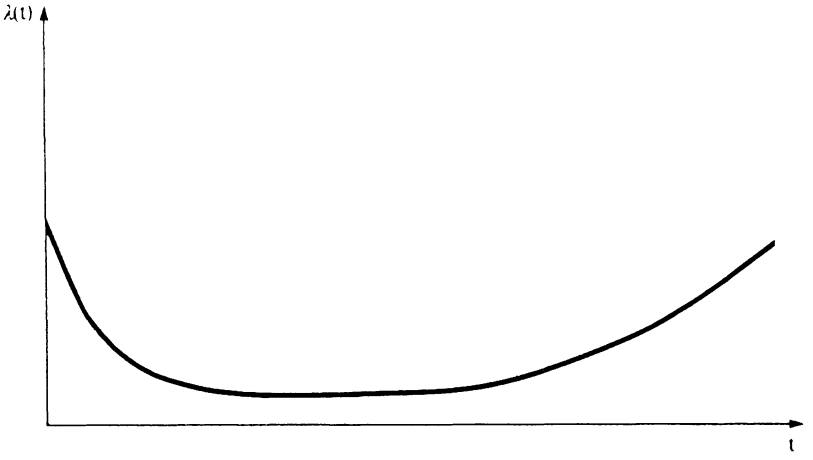

Abbildung 3: Hazardrate mit „badewannenförmigem" Verlauf

Zu Beginn des Prozesses ist das Sterberisiko wegen der Kindersterblichkeit relativ hoch, es fällt dann und bleibt über einen bestimmten Zeitraum konstant auf niedrigem Niveau, bis es mit zunehmendem Alter wieder anwächst. Ähnlich verhält sich die Hazardrate bei vielen technischen Geräten. Aufgrund von „Kinderkrankheiten" und „Defekten beim ersten Einschalten" ist das Ausfallrisiko zunächst relativ hoch, fällt dann ab und wächst wieder, wenn Alterungsprozesse und Materialermüdungserscheinungen auftreten. Daneben sind natürlich auch andere Formen der Hazardrate denkbar, zum Beispiel ständig zunehmende oder abnehmende Hazardraten.

Aus Definition (2.5) folgt unmittelbar die Beziehung zwischen Hazardrate und Survivorfunktion

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{2.7}$$

und da T als stetig vorausgesetzt wurde, gilt auch

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \tag{2.8}$$

Umgekehrt ergibt sich die Survivorfunktion in Abhängigkeit von der Hazardrate, wenn man $\lambda(t)$ integriert und die Beziehungen (2.7) und (2.8) verwendet.

$$\int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{1 - F(u)}du = -\ln(1 - F(u))\Big|_0^t$$

$$= -\ln(1 - F(t)) = -\ln S(t). \tag{2.9}$$

Dies führt zur wichtigen Beziehung

$$S(t) = \exp(-\int_0^t \lambda(u)du). \tag{2.10}$$

Die Dichtefunktion f(t) ergibt sich aus (2.7) und (2.10) in Abhängigkeit von der Hazardrate

$$f(t) = \lambda(t) \cdot S(t) = \lambda(t) \cdot \exp(-\int_0^t \lambda(u)du). \tag{2.11}$$

Aus den Beziehungen (2.1) bis (2.11) wird ersichtlich, daß jede der drei Größen f(t), S(t) und λ(t) zur Beschreibung der Verteilung der Episodendauer herangezogen werden kann. Ist eine der Größen festgelegt, so sind die beiden anderen eindeutig daraus ableitbar. Da in der Ereignisanalyse die Hazard- beziehungsweise Übergangsrate das mathematisch einfacher zu handhabende Konzept ist, wird diese in der Regel zur Modellierung herangezogen. Für jede Spezifikation der Hazardrate existiert jedoch eine äquivalente Spezifikation der Wahrscheinlichkeitsverteilung von T. Beide Spezifikationen enthalten dieselben Parameter und liefern insbesondere dieselbe Likelihoodfunktion zur Parameterschätzung. Der Hazardratenansatz ist keine völlig neue Modellierung und gestattet auch nicht, zusätzliche Parameter zu identifizieren. Er ist aber in der Regel einfacher und trägt dem Umstand Rechnung, daß die abhängige Variable eine Zeitdauer ist.

B. Spezielle Wahrscheinlichkeitsverteilungen für die Dauer der Episode

a) Exponentialverteilung — zeitunabhängige Hazard- beziehungsweise Übergangsrate

Eine der am häufigsten verwendeten Verteilungen für Verweildauer und Lebenszeiten ist die Exponentialverteilung. Sie ist charakterisiert durch eine im Zeitablauf konstante Hazardrate

$$\lambda(t) = \lambda, \qquad t \geq 0, \ \lambda > 0.$$

Für Dichte- und Survivorfunktion folgen

$$S(t) = \exp(-\lambda t),$$

$$f(t) = \lambda \exp(-\lambda t).$$

Für die „mittlere Verweildauer" erhält man

$$E(T) = \frac{1}{\lambda}.$$

Je größer das „Risiko" λ des Eintreffens eines Ereignisses ist, desto kürzer ist die erwartete Verweildauer. Für die Varianz ergibt sich

$$Var(T) = \frac{1}{\lambda^2}.$$

b) Weibull-Verteilung

Die Weibull-Verteilung stellt eine Verallgemeinerung der Exponentialverteilung dar und wurde bislang häufig bei der Untersuchung der Lebenszeit technischer Geräte verwendet. Die Hazardrate ist

$$\lambda(t) = \lambda\alpha(\lambda t)^{\alpha-1} \qquad (t > 0)$$

mit den Parametern $\lambda > 0$ und $\alpha > 0$. Für den Spezialfall $\alpha = 1$ erhält man wieder die Exponentialverteilung. Die Harzard- beziehungsweise Übergangsrate der Weibull-Verteilung ist monoton steigend für $\alpha > 1$, abnehmend für $\alpha < 1$ und konstant für $\alpha = 1$. Das Weibull-Modell ist sehr flexibel und daher für eine Vielzahl von Modellen für Verweildauern und Lebenszeiten angemessen.
Die Survivorfunktion ist

$$S(t) = \exp(-(\lambda t)^{\alpha})$$

und die Dichtefunktion

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

Für den Erwartungswert $E(T)$ der Verweildauer ergibt sich

$$E(T) = \Gamma\left(\frac{1+\alpha}{\alpha}\right)/\lambda,$$

wobei $\Gamma(\cdot)$ die Gamma-Funktion ist. Die Varianz ist

$$Var(T) = \left(\Gamma\left(\frac{\alpha+2}{\alpha}\right) - \left(\Gamma\left(\frac{\alpha+1}{\alpha}\right)\right)^2\right)/\lambda^2.$$

Für eine Beschreibung weiterer wichtiger Verteilungen für Verweildauern und Lebenszeiten vergleiche man beispielsweise Kalbfleisch und Prentice (1980) oder Blossfeld, Hamerle und Mayer (1986, 1989).

III. Einbeziehung von Kovariablen: Regressionsmodelle

A. Quantitative und qualitative Kovariablen

Neben der Verweildauer beziehungsweise Lebenszeit werden in der Regel für jedes Individuum oder Objekt eine Reihe von weiteren Kovariablen oder prognostischen Faktoren erhoben, und ein wichtiges Ziel der statistischen Analyse besteht in der quantitativen Ermittlung des Einflusses dieser exogenen oder endogenen Variablen.

Bei den Kovariablen kann es sich um *quantitative* oder um *qualitative Merkmale* handeln. Ein quantitatives Merkmal x_j wird wie in der herkömmlichen multiplen Regression mit einem Parameter β_j gewichtet und mit $x_j \beta_j$ in das Modell aufgenommen. Bei kategorialen Merkmalen geht man in Analogie zur Varianzanalyse über zu einer Kodierung der einzelnen Kategorien durch Dummy-Variablen.

Eine Möglichkeit für die Kodierung der Kategorien qualitativer Merkmale besteht in der *(0,1)-Kodierung* („cornered effects"). Besitzt ein Merkmal A I Kategorien (Ausprägungen, Klassen, Faktorstufen), so lassen sich diese durch I − 1 Dummy-Variablen erfassen in der Form

$$x_i^{\wedge} = \begin{cases} 1 & \text{falls Kategorie i der Variablen A vorliegt} \\ 0 & \text{sonst} \end{cases}, \qquad i = 1, \ldots, I-1.$$

Die i-te Dummy-Variable $x_i^{\wedge}$ $(i = 1, \ldots, I-1)$ kodiert dabei nur das Vorliegen beziehungsweise Nicht-Vorliegen der i-ten Ausprägung. Das Vorliegen der I-ten (Referenz-)Kategorie ist implizit erfaßt durch die Kodierungen $x_i^{\wedge} = 0$ für $i = 1, \ldots, I-1$. Die Wahl der I-ten Kategorie als Referenzkategorie ist prinzipiell beliebig. Im Hinblick auf die Interpretation der Ergebnisse sollte jedoch eine Kategorie gewählt werden, auf die sich alle anderen Ausprägungen leicht beziehen lassen, da die Parameter β_j jeweils die „Abstände" der j-ten Ausprägung zur Referenzkategorie darstellen.

Besonders einfach ist der Spezialfall eines dichotomen unabhängigen Merkmals. Dann ist I = 2, und man erhält nur eine Dummy-Variable

$$x^{\wedge} = \begin{cases} 1 & \text{falls Kategorie 1 vorliegt} \\ 0 & \text{falls Kategorie 2 vorliegt.} \end{cases}$$

Im allgemeinen Fall lassen sich mit $x_1^{\wedge}, x_2^{\wedge}, \ldots, x_{I-1}^{\wedge}$ sämtliche Kategorien der qualitativen Variablen A kodieren. Die zugehörigen Regressionskoeffizienten β_j werden gewöhnlich wie in der Varianzanalyse *Haupteffekte* genannt. Die (0,1)-Kodierung ist insbesondere bei Ansätzen mit gemischt quantitativ/qualitativen Kovariablen zweckmäßig. Bei ausschließlich qualitativen Kovariablen wird häufig auch die *Effekt-Kodierung* („centered effects") verwendet, die unmittelbar an die herkömmli-

che Varianzanalyse angelehnt ist. Man vergleiche dazu beispielsweise Hamerle, Kemény und Tutz (1984, S. 214).

Im Rahmen von Regressionsmodellen für Verweildauern und Lebenszeiten, insbesondere bei kategorialen unabhängigen Merkmalen, kommen auch *Interaktionswirkungen* als Einflußgrößen in Frage. Sie messen den gemeinsamen Einfluß einer bestimmten Kombination von Kategorien von zwei oder mehreren unabhängigen Merkmalen. Formal können sie in einfacher Weise durch die Bildung entsprechender Produkte der Dummy-Variablen in den Regressionsansatz einbezogen werden. Für die Zwei-Faktor-Interaktionswirkungen der Faktoren A und B ergeben sich die Produkte $x_i^A x_j^B$, $i = 1,...,I-1$, $j = 1,...,J-1$, für die Drei-Faktor-Interaktionen die Produkte $x_i^A x_j^B x_k^C$, usw. Die Werte der quantitativen Kovariablen einer Person beziehungsweise eines Objekts i sowie die Kodierungen für sämtliche Haupteffekte und im Modell enthaltene Interaktionswirkungen der qualitativen Kovariablen werden in einem Daten- oder Designvektor x zusammengefaßt.

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

$$\eta = x'\beta$$

mit einem unbekannten p-dimensionalen Parametervektor β. Die Parameter $\beta_1,...,\beta_p$ repräsentieren die Einflußgewichte der Kovariablen. Im Gegensatz zur klassischen multiplen Regression geht man aber nicht davon aus, daß die Linearkombination $\eta = x'\beta$ die Verweildauern beziehungsweise Lebenszeiten T direkt beeinflußt, sondern in der Regel eine Funktion von T, etwa $\ln T$.

Ein weiterer wichtiger Unterschied zur herkömmlichen Regression liegt darin, daß in den hier behandelten Verweildauermodellen einige Kovariablen selbst zeitabhängig sein können. Dies ist beispielsweise dann der Fall, wenn eine bestimmte medizinische Therapie nur während eines bestimmten Zeitraums angewendet wird. Das Untersuchungsziel könnte dann darin bestehen, den Einfluß dieser Therapie während der eigentlichen Anwendung oder in ihren Nachwirkungen zu überprüfen. Dafür definiert man zwei Dummy-Variablen, etwa $x_1(t)$ und $x_2(t)$ mit

$$x_1(t) = \begin{cases} 1 & \text{während des Zeitraums der Teilnahme einer Person an Therapie bzw. Programm,} \\ 0 & \text{sonst} \end{cases}$$

$$x_2(t) = \begin{cases} 1 & \text{nach Abschluß der „Behandlung" für eine Person, die an Therapie bzw. Programm teilgenommen hat,} \\ 0 & \text{sonst.} \end{cases}$$

Werden die Regressionsansätze wie gewöhnlich in den Hazardraten formuliert und sind die zugehörigen Regressionskoeffizienten signifikant negativ (positiv), dann ist die Therapie effektiv und verringert (vergrößert) die Wahrscheinlichkeit für einen baldigen Zustandswechsel. Ist darüber hinaus der erste Koeffizient absolut signifikant größer (kleiner) als der zweite, dann sinkt (steigt) der Effekt nach dem Absetzen der Therapie.

Eine Möglichkeit der Analyse des Einflusses von Kovariablen auf die Verweildauern beziehungsweise Lebenszeiten besteht darin, ein Regressionsmodell zu formulieren, bei dem die Verteilung der Verweildauer beziehungsweise Lebenszeit von den Kovariablen abhängt. Bezeichnet x den Vektor der Kovariablen, so ist ein Modell für die Verweildauer beziehungsweise Lebenszeit T bei gegebenem Kovariablenvektor x zu spezifizieren. Da, wie bereits im letzten Abschnitt ausgeführt, bei der Analyse von Verweildauern die Hazardrate das mathematisch einfachere Konzept ist, liegt es nahe, diese in Abhängigkeit von den Kovariablen zu modellieren. Für das Exponentialmodell erhält man

$$\lambda(t|x) = \exp(x'\beta). \tag{3.1}$$

Hängen die Kovariablen nicht von der Zeit ab, ist die Hazardrate in (3.1) zeitinvariant.

Die Hazardrate des Weibull-Regressionsmodells ist gegeben durch

$$\lambda(t|x) = \delta\,\lambda_0(\lambda_0 t)^{\delta-1}\exp(x'\beta). \tag{3.2}$$

Das Weibull-Modell gehört zur Klasse der „Proportional-Hazards"-Modelle, da der Quotient der Hazardraten zweier Individuen nicht von der Zeit abhängt. Eine Erweiterung stellt der semi-parametrische Ansatz dar, der von Cox (1972) vorgeschlagen wurde. Die Hazardrate des Proportional-Hazardsmodells von Cox ist

$$\lambda(t|x) = \lambda_0(t)\exp(x'\beta). \tag{3.3}$$

$\lambda_0(t)$ ist die beliebige, nicht spezifizierte Grundhazardrate. Dadurch wird mehr Flexibilität in der Modellierung erreicht, allerdings sind bei der Parameterschätzung andere Methoden zu verwenden als bei den bisher betrachteten Modellen.

IV. Parameterschätzung

Nach der Konstruktion eines statistischen Modells für die vorliegende Ereignisgeschichte sind die unbekannten Parameter aus den erhobenen Daten zu schätzen. In diesem Abschnitt wird ausschließlich die Maximum-Likelihood-Methode behandelt, die den Erfordernissen der Ereignisanalyse am besten gerecht wird. Bei der Anwendung der Maximum-Likelihood-Schätzung — wie auch bei anderen Schätzverfahren — muß für jedes Stichprobenelement eine Realisation des in Frage ste-

henden zufälligen Merkmals vorliegen. Da in der Ereignisanalyse das Ende des gesamten Beobachtungszeitraums in der Regel vorgegeben ist, ist die Dauer der Episode unter Umständen nicht abgeschlossen. Man spricht in einem solchen Fall von rechtszensierten Daten. Die Stichprobenrealisation t_i eines Individuums besagt dann lediglich, daß die Dauer der Episode *mindestens* t_i Zeiteinheiten beträgt. Die exakte Zeitdauer läßt sich nicht angeben. In der Regel liegt eine Stichprobe vor, bei der einige Werte t_i exakte Zeitdauern sind, während es sich beim Rest um zensierte Daten handelt. Man bringt dies mit Hilfe eines *Zensierungsindikators* δ_i zum Ausdruck mit

$$\delta_i = \begin{cases} 1 & \text{falls } t_i \text{ nicht zensiert ist} \\ 0 & \text{falls } t_i \text{ zensiert ist,} \end{cases} \qquad i = 1,...,n.$$

Die Möglichkeit, die zensierten Daten einfach zu ignorieren und den Stichprobenumfang zu reduzieren, ist nicht zu empfehlen, da dies verzerrte Resultate zur Folge haben kann.

Die Maximum-Likelihood-Methode bietet die Möglichkeit, rechtszensierte Daten explizit im Schätzvorgang zu berücksichtigen. Zu diesem Zweck ist der Zensierungsmechanismus, der den Daten zugrundeliegt, genau zu analysieren und in ein statistisches Modell zu fassen. Für das Zustandekommen von zensierten Daten sind je nach Anwendungsbereich mehrere statistische Konzepte denkbar. Ein häufig verwendetes Modell (random censoring) setzt voraus, daß die Zeitdauern T_i und die Zensierungszeiten C_i stochastisch unabhängige Zufallsvariablen sind und daß darüber hinaus die Verteilungen der Zensierungszeiten nicht von den Parametern abhängen, die die Verteilung der Verweildauern determinieren. Dann ergibt sich für die Likelihoodfunktion

$$L = c \cdot \prod_{i=1}^{n} f_i(t_i|x_i)^{\delta_i} S_i(t_i|x_i)^{1-\delta_i}. \tag{4.1}$$

Berücksichtigt man den Zusammenhang zwischen Hazardrate und Survivorfunktion (vgl. (2.10)), so erhält man

$$L = c \cdot \prod_{i=1}^{n} \lambda_i(t_i|x_i)^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_i(u|x_i)\,du\right). \tag{4.2}$$

Verwendet man ein parametrisches Regressionsmodell, so ist der parametrische Ansatz für $\lambda(t|x)$ in (4.2) einzusetzen, und die logarithmierte Likelihoodfunktion wird in Abhängigkeit von den unbekannten Parametern maximiert. Dazu sind in der Regel iterative Verfahren, etwa Newton- oder modifizierte Newton-Verfahren zu verwenden.

Die Likelihoodfunktion für das *Proportional-Hazards-Modell von Cox* mit der Hazardrate $\lambda(t|x) = \lambda_0(t)\exp(x'\beta)$ ist

$$L(\beta, \lambda_0(t), x_1, ..., x_n) = \prod_{i=1}^{n} [\lambda_0(t_i) \exp(x_i'\beta)]^{\delta_i} \exp[- \int_0^{t_i} \lambda_0(u) \exp(x_i'\beta) du]. \quad (4.3)$$

(4.3) enthält die unbekannte Baseline-Hazardrate $\lambda_0(t)$, das heißt nicht nur die unbekannten Parameter β, sondern darüber hinaus noch die „Nuisance-Funktion" $\lambda_0(t)$. Deshalb kann (4.3) zur Schätzung von β nicht herangezogen werden. Cox (1972, 1975) schlug vor, die Likelihood (4.3) zu faktorisieren. Seien $t_{(1)} < ... < t_{(k)}$ die Zeitdauern der Individuen, die nicht zensiert sind ($k \leq n$), und sei $R(t)$ die *„Risikomenge"*, das heißt die Menge der Individuen, deren Episode unmittelbar vor dem Zeitpunkt t noch nicht beendet ist und die nicht zensiert sind. Aus (4.3) erhält man dann durch Erweiterung

$$L(\beta, \lambda_0(t); x_1, ..., x_n) = \prod_{i=1}^{k} \frac{\exp(x_{(i)}'\beta)}{\sum_{l \in R(t_{(i)})} \exp(x_l'\beta)} \sum_{l \in R(t_{(i)})} \lambda_0(t_{(i)}) \exp(x_l'\beta) \prod_{i=1}^{n} S_0(t_i)^{\exp(x_i'\beta)}$$

mit

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

Den ersten Faktor

$$PL(\beta; x_1, ..., x_n) = \prod_{i=1}^{k} \frac{\exp(x_{(i)}'\beta)}{\sum_{l \in R(t_{(i)})} \exp(x_l'\beta)}, \quad (4.4)$$

der nur von β abhängt, bezeichnete Cox (1972, 1975) als „partial likelihood" und schlug vor, (4.4) wie eine gewöhnliche Likelihood-Funktion zu behandeln und in Abhängigkeit von β zu maximieren.

Die Anwendung der Partial Likelihood (4.4) setzt voraus, daß die Zeitdauern t_i ausreichend genau gemessen werden können, so daß keine gleichen Meßwerte (Verbundwerte; Ties) auftreten. Bei praktischen Anwendungen treten jedoch häufig gleiche Meßwerte auf, da entweder nur ungenau gemessen wird oder nur Zeitintervalle angegeben werden können, in denen Ereignisse stattfinden. In solchen Fällen muß die Partial Likelihood korrigiert werden. Breslow (1974) schlägt vor, (4.4) durch

$$PL(\beta; x_1, ..., x_n) = \prod_{i=1}^{k} \frac{\exp(s_i'\beta)}{[\sum_{l \in R(t_{(i)})} \exp(x_l'\beta)]^{d_i}} \quad (4.5)$$

zu approximieren. Dabei ist d_i die Anzahl der gleichen Verweildauerzeiten zum Zeitpunkt $t_{(i)}$, und s_i ist die Summe der Kovariablenvektoren dieser d_i Individuen.

Parametertests oder die Konstruktion von Konfidenzintervallen können in der üblichen Weise durchgeführt werden. Die Tests beruhen auf den asymptotischen Eigenschaften, insbesondere der asymptotischen Normalverteilung der Maximum-Likelihood-Schätzungen beziehungsweise der Maximum-Partial-Likelihood-

Schätzungen beim Cox-Modell. Man vergleiche dazu etwa Andersen und Gill (1982) und Borgan (1984). Die praktische Durchführung der Tests sowie weiterer Verfahren zur Residuenanalyse und zur Konstruktion von Modelltests (etwa zur Überprüfung der Annahme proportionaler Hazardraten) ist zum Beispiel in Blossfeld, Hamerle und Mayer (1986, 1989) beschrieben.

Die statistische Analyse von Verweildauern ermöglicht — wie bereits erwähnt — auch die Einbeziehung zeitabhängiger Kovariablen. Diese können ihrerseits Realisierungen eines stochastischen Prozesses sein. Die Hazardrate ist dann $\lambda(t|x(t))$. Ein Analogon zur Survivorfunktion kann ebenfalls angegeben werden. Allerdings ist eine entsprechende Interpretation nicht immer gewährleistet. Darüber hinaus erfordert die Ermittlung der Survivorfunktion eine Integration über den gesamten Zeitpfad bis zum Zeitpunkt t der Kovariablen. Dies kann erhebliche numerische Probleme verursachen. Die Berechnungen vereinfachen sich beträchtlich, wenn die zeitabhängigen Kovariablen Treppenfunktionen über die Zeit sind. Ein solches Beispiel wird im nächsten Abschnitt behandelt. Mittlerweile sind zur numerischen Auswertung von ereignisorientierten Datensätzen eine Reihe von Programmpaketen verfügbar, zum Beispiel die Prozedur LIFEREG in SAS, BMDP2L oder RATE (vgl. Blossfeld/Hamerle/Mayer, 1989). Über weitere Softwarepakete zur Analyse von Verweildauern und Lebenszeiten informiert Kemény (1986).

Schließlich ist noch anzumerken, daß eine Reihe von Erweiterungen und Verallgemeinerungen der Verfahren möglich ist, insbesondere im Hinblick auf die Einbeziehung mehrerer Endzustände (competing risks) oder die Analyse mehrerer Zeitdauern oder Episoden für ein Individuum. Für weitere Details vergleiche man beispielweise Kalbfleisch und Prentice (1980), Cox und Oakes (1984) oder Blossfeld, Hamerle und Mayer (1986, 1989). Einen Überblick über diskrete Hazardratenmodelle, bei denen lediglich Zeitintervalle angebbar sind, in denen Ereignisse oder Zustandswechsel stattgefunden haben, findet man in Hamerle und Tutz (1989).

V. Ein Analysebeispiel

Als ein Beispiel für die Anwendung der Ereignisanalyse und die Interpretation ihrer Ergebnisse sollen die Mechanismen des Berufswechselverhaltens von Männern untersucht werden. Die Analyse basiert dabei auf den Daten der Lebensverlaufsstudie, die gegenwärtig vom Max-Planck-Institut für Bildungsforschung in Berlin durchgeführt wird (Mayer/Brückner, 1989). In der Lebensverlaufsstudie wurden 2.171 deutsche Personen aus den Geburtsjahrgängen 1929—31, 1939—41 und 1949—51 repräsentativ in bezug auf deren räumliche Verteilung über die Bundesrepublik Deutschland ausgewählt und befragt. Die Erhebung erstreckte sich von Oktober 1981 bis Mai 1983 (vgl. Mayer/Brückner, 1989; Blossfeld, 1987; Huinink,

1988). Das Ziel der Befragung war es, mit Hilfe von standardisierten Interviews die Lebensläufe dieser Personen retrospektiv mit detaillierten Zeitangaben zu erfassen, um sie so einer dynamischen Längsschnittanalyse zugänglich zu machen.

Da die Daten nicht nur den Bildungs- und Berufsverlauf umfassen, sondern auch die Verläufe in anderen Lebensbereichen (Familie, Wohnung usw.) zugänglich machen, können die Effekte von parallelen Prozessen (z. B. aus der Familiengeschichte das Ereignis Heirat) auf die Erwerbskarriere (z. B. die Stabilität von Berufsverläufen) analysiert werden. Darüber hinaus kann überprüft werden, wie sich Merkmale der Vorgeschichte auf den jeweils späteren Berufsverlauf auswirken. Insgesamt bietet diese Datenbasis gute Voraussetzungen, um an ihrem Beispiel einige methodisch interessante Varianten der Ereignisanalyse darzustellen.

Beginnen wir das Beispiel mit der Schätzung eines Exponential-Modells, bei dem zur Erklärung des Berufswechselverhaltens der Männer die Variablen Bildung (BILDG), Prestige (PRES), Anzahl der vorher ausgeübten Berufstätigkeiten (BANZ), Berufserfahrung zu Beginn jeder Tätigkeit (BERF) und, zur Unterscheidung der drei Geburtskohorten, die Dummy-Variablen KOHO2 und KOHO3 herangezogen werden (zur Definition der Variablen siehe Anhang):

$$\lambda(t|x) = \exp(x'\beta).$$

Das Exponentialmodell geht von der Vorstellung einer konstanten Hazardrate aus und impliziert die Annahme proportionaler Risiken. Es läßt sich einfach interpretieren und wird in der Forschungspraxis häufig als Basis oder Referenz-Modell benutzt, mit dem dann die Schätzungen komplexerer Verteilungs-Modelle verglichen werden. Die Maximum-Likelihood-Schätzungen für dieses Modell wurden mit Hilfe des Programms RATE (vgl. Tuma, 1986) berechnet. Die Ergebnisse der Schätzung sind als Modell (1) in Tabelle 1 zu finden.

Zunächst erhält man für das vorliegende Exponential-Modell, wenn man es auf der Basis eines Likelihood-Quotienten-Tests mit einem Exponential-Modell ohne Kovariablen vergleicht, einen χ^2-Wert von 705,90 mit sechs Freiheitsgraden. Die eingeführten Kovariablen können also zur Erklärung des Berufswechselrisikos bei Männern etwas beitragen, und die Nullhypothese, keiner der zusätzlich aufgenommenen β-Koeffizienten ist von Null verschieden, muß verworfen werden. Eine Signifikanzprüfung der einzelnen Regressionsparameter wird durchgeführt, indem man die Koeffizienten $\hat{\beta}_i$ durch ihre geschätzten asymptotischen Standardabweichungen $s(\hat{\beta}_i)$ dividiert. Unter der Hypothese $H_0: \beta_i = 0$ sind diese Prüfgrößen näherungsweise standardnormalverteilt. Geht man von einer Signifikanzwahrscheinlichkeit von 0,05 und beidseitiger Fragestellung aus, dann haben die Kovariablen einen signifikanten Effekt, wenn der Betrag ihrer standardisierten Koeffizienten größer als der Wert 1,96 ist. Dies ist außer bei der Konstanten β_0(KONST) bei den Variablen PRES, BANZ, BERF, KOHO2 und KOHO3 der Fall. Nur die Variable

Bildung (BILDG) hat keinen signifikanten Einfluß auf die Rate des Berufswechsels bei Männern.

Die Wirkung einer Kovariablen x_i kann man anschaulich interpretieren, indem man bei Konstanthaltung der jeweils anderen Variablen in $x'\beta$ zeigt, um wieviel Prozent sich die Rate bei der Erhöhung der Kovariablen x_i um einen bestimmten Wert Δx_i verändert. So ergibt sich beispielsweise bei einer Erhöhung der Anzahl der vorher ausgeübten Berufe (BANZ) um eine Einheit eine Erhöhung der Rate um etwa 19 Prozent $[(\exp(0{,}171) - 1) \cdot 100\,\% = 18{,}7\,\%]$. Eine Erhöhung des Prestiges (PRES) um 20 Einheiten führt dagegen zu einer Verminderung der Neigung zum Berufswechsel um etwa 10 Prozent $[(\exp(-0{,}005)^{20} - 1) \cdot 100\,\% = -9{,}9\,\%]$. Die gleichzeitige Veränderung von BANZ um eine Einheit und von Prestige (PRES) um 20 Einheiten, was einem beruflichen Aufstieg entspräche, erhöht die Rate allerdings nur um etwa 8 Prozent $[(\exp(0{,}171)^1 \cdot \exp(-0{,}005)^{20} - 1) \cdot 100\,\% = 7{,}4\,\%]$ und nicht um 8,8 Prozent $[18{,}70\,\% - 9{,}9\,\% = 8{,}8\,\%]$.

Über die Beziehung

$$E(T|x) = \frac{1}{\lambda(x)} = \frac{1}{\exp(x'\beta)}$$

kann man bei der Exponential-Verteilung auch direkt angeben, wie sich bei Konstanthaltung aller restlichen Kovariablen die durchschnittliche Verweildauer $E(T|x)$ verändert, wenn man den Wert der unabhängigen Variablen x_i um den Betrag Δx_i erhöht:

$$\delta_{\Delta x_i} = \left(\frac{1}{\exp(\beta_i)^{\Delta x_i}} - 1\right) \cdot 100\,\%.$$

Danach folgt bei Erhöhung der Anzahl der vorher bereits ausgeübten Berufe (BANZ) um eine Einheit eine Verminderung der durchschnittlichen Verweildauer im Beruf um etwa 16 Prozent $[(1/\exp(0{,}171) - 1) \cdot 100\,\% = -15{,}8\,\%]$. Eine Erhöhung des Prestiges (PRES) um 20 Einheiten führt dagegen zu einer Erhöhung der Verweildauer im Beruf um 11 Prozent $[(1/\exp(-0{,}005)^{20} - 1) \cdot 100\,\% = 11{,}0\,\%]$. Die gleichzeitige Veränderung von BANZ um eine Einheit und des Prestiges (PRES) um 20 Einheiten, was wieder einem beruflichen Aufstieg entsprechen würde, vermindert die durchschnittliche Verweildauer aber um 6,5 Prozent $[(1/(\exp(0{,}171)^1 \exp(-0{,}005)^{20}) - 1) \cdot 100\,\% = -6{,}5\,\%]$ und nicht nur um etwa 5 Prozent $[10{,}98\,\% - 15{,}75\,\% = -4{,}77\,\%]$.

Für beliebige Subgruppen lassen sich natürlich auch Prognosen über die durchschnittliche Verweildauer, den Median der Verweildauer, die in einer bestimmten Zeitspanne durchschnittlich eintretenden Ereignisse und die Wahrscheinlichkeit geben, bis zu einem bestimmten Zeitpunkt noch im selben Zustand zu sein. Betrachtet man beispielsweise einen Mann der Kohorte von 1939—41 (KOHO2 = 1 und KOHO3 = 0), der in einem mit 50 Prestige-Punkten (PRES = 50) bewerteten Be-

ruf arbeitet und der vorher bereits 10 Berufe (BANZ = 10) ausgeübt sowie dabei eine Berufserfahrung von 100 Monaten (BERF = 100) gesammelt hat, so kommt man zu folgender Prognose-Gleichung für die Rate des Berufswechsels[1]:

$$\hat{\lambda} = \exp(-4{,}338 - 0{,}005 \cdot 50 + 0{,}171 \cdot 10 - 0{,}009 \cdot 100 + 0{,}179 \cdot 1) = 0{,}0274.$$

Daraus folgt, daß man bei dieser Person eine mittlere Verweildauer von 36,5 Monaten [$1/\hat{\lambda} = 1/0{,}0274 = 36{,}496$] erwartet, die weit unter der des Durchschnitts von 98,04 Monaten liegt. Darüber hinaus kann man einen Median der Verweildauer im Beruf von $\hat{M}^* = 0{,}6934 \cdot 36{,}5 = 25{,}3$ Monaten prognostizieren und in einem Jahr durchschnittlich $\hat{\lambda}v = 0{,}0274 \cdot 12 = 0{,}33$ Berufswechsel erwarten. Die Wahrscheinlichkeit schließlich, daß diese Person noch nach acht Jahren in ihrem Beruf arbeitet, beträgt 7,2 Prozent [$\hat{S}(96) = \exp(-0{,}0274 \cdot 96) = 0{,}072$], während sie sich beim Durchschnitt aller Männer auf 37,6 Prozent beläuft. Durch entsprechende Prognosen für weitere Subgruppen kann man insgesamt ein sehr differenziertes Bild von dem Berufswechselverhalten der Männer und der Bedeutung unterschiedlicher Einflußfaktoren geben.

Das Modell (1) in Tabelle 1 hat zeitkonstante Kovariablen einbezogen. Unter zeitkonstanten Kovariablen verstehen wir hier, daß diese zu Beginn der Episode k gemessen (oder aktualisiert) werden und ihre Werte über die Verweildauer

Tabelle 1: Schätzungen der Raten des Berufswechsels bei Männern

	Modell		
Kovariablen	(1)	(2)	(3)
KONST	− 4,338*	− 4,283*	− 3,492*
BILDG	0,013	0,025	0,007
PRES	− 0,005*	− 0,004*	− 0,004*
BANZ	0,171*	0,173*	0,160*
BERF	− 0,009*	− 0,007*	− 0,008*
KOHO2	0,179*	0,159*	0,124*
KOHO3	0,486*	0,415*	0,341*
HEIRAT		− 0,174*	
TDEP			0,8266
χ^2	705,9	969,9	868,4
df	6	7	7

* Statistisch signifikant auf dem 0,05 Niveau.

[1] Die Variable Bildung (BILDG) braucht bei der Prognose nicht berücksichtigt zu werden, da ihr $\hat{\beta}$-Koeffizient nicht signifikant von Null verschieden ist.

$v_k = t - t_{k-1}$ hinweg unverändert bleiben (vgl. Abbildungen 4(c)). Zeitveränderliche Kovariablen können hingegen ihren Wert innerhalb der Episode k verändern. Bei diskreten zeitveränderlichen Kovariablen bleiben die Werte dabei über gewisse Subintervalle

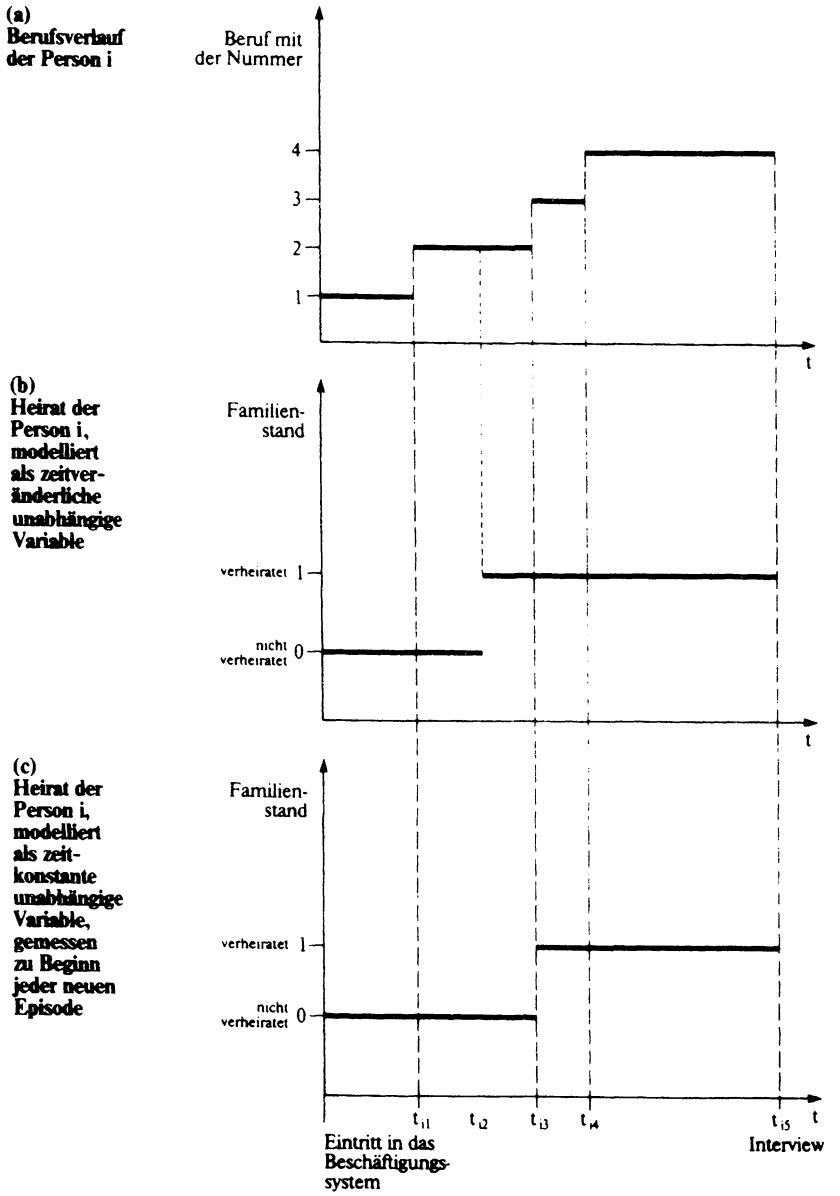

Abbildung 4: Modellierung des Einflusses der diskreten Variablen Heirat (a) auf den Berufsverlauf, (b) als zeitveränderliche unabhängige Variable und (c) als zeitkonstante unabhängige Variable, gemessen zu Beginn jeder neuen Episode

$$v_k, v_k = \sum_{i=1}^{s} v_{k_i}$$

konstant (vgl. Abbildung 4(b)), während sich stetige zeitveränderliche Variablen kontinuierlich verändern.

In den meisten Fällen wird man es in der Wirtschafts- und Sozialwissenschaft mit Variablen zu tun haben, die sich nicht stetig in der Zeit verändern beziehungsweise nicht stetig über die Zeit gemessen werden können. Diese folgen in der Zeit einer Treppenfunktion und wirken direkt auf die Verweildauer ein, indem sie die Rate innerhalb der Episoden verändern. Geht man beispielsweise davon aus, daß sich bei Männern das Ereignis „Heirat" im Familiensystem stabilisierend auf den Erwerbsprozeß im Beschäftigungssystem auswirkt (vgl. Abbildung 4), dann läßt sich dieser Zusammenhang über die Einführung einer zeitveränderlichen Kovariablen testen.

Bezeichnet man für das Individuum i mit $t_{i,k-1}$ den Beginn der Berufsepisode k und mit t_{ik} deren Endzeitpunkt sowie mit v die Verweildauer der Episode k und ist t_i^H der Heiratszeitpunkt des Individuums i, dann ergibt sich der Wert der zeitveränderlichen Dummy-Variablen Heirat $x_{ik}^H(v)$ wie folgt:

$$x_{ik}^H(v) = \begin{cases} 0 & \text{für } t_i^H - t_{i,k-1} \geq v \\ 1 & \text{für } t_i^H - t_{i,k-1} < v \end{cases}$$

Bei der Maximum-Likelihood-Schätzung läßt sich die Einbeziehung diskreter Kovariablen relativ einfach bewerkstellingen. Bezeichnen $t_0 < t_1 < \ldots < t_s$ die Änderungszeitpunkte des Kovariablen-Vektors im Verweildauer-Intervall $[0,t)$ und sei $t_{s+1} = t$, dann kann die kumulative Hazardrate in eine Summe von Integralen zerlegt werden, und die Wahrscheinlichkeit, daß bis zum Zeitpunkt t kein Ereignis auftritt, ergibt sich aus dem Produkt der Survivorfunktionen der Subepisode, in denen der Kovariablen-Vektor unverändert bleibt:

$$S(t|x(t)) = \prod_{r=1}^{s+1} S(t_r|t_{r-1}, x(t_{r-1})).$$

Die konkrete Realisierung der Maximum-Likelihood-Schätzung kann dann in der Weise erfolgen, daß man die beobachteten Verweildauern t_i anhand der s_i Änderungszeitpunkte in $s_i + 1$ eigenständige Subepisoden aufsplittert und die Hazardrate wie im Falle zeitkonstanter Kovariablen schätzt. Die dabei neu zu konstruierende ereignisorientierte Datei enthält dann für jede dieser Subepisoden, in der der Kovariablen-Vektor unverändert bleibt, einen eigenen Satz mit folgenden Informationen (vgl. Blossfeld/Hamerle/Mayer, 1986, 1989):

(1) die Ausprägungen der Kovariablen zu Beginn der Subepisode;
(2) die Verweildauer zu Beginn und am Ende der Subepisode (die Verweildauer als solche ist nur beim Exponential-Modell ausreichend);

(3) eine Zensierungsinformation, ob die Subepisode mit einem Ereignis (ZEN = 1) endete oder nicht (ZEN = 0).

Will man nun, wie im vorliegenden Beispiel, den Familienstand (verheiratet — nicht verheiratet) als zeitveränderliche unabhängige Variable bei der Schätzung des Berufswechselrisikos von Männern in einem Exponential-Modell berücksichtigen, dann bricht man die Berufsepisoden nach dem Zeitpunkt der Heirat auf. Die neue ereignisorientierte Datei wird dabei so aufbereitet, daß für jedes Zeitintervall innerhalb einer gegebenen Berufsepisode, in der die Kovariable Familienstand unverändert bleibt, ein eigener Datensatz erzeugt wird. Der neue ereignisorientierte Datensatz mit den nach dem Heiratszeitpunkt aufgebrochenen Episoden, kann nun wie im Falle zeitkonstanter Kovariablen behandelt und wie im vorliegenden Beispiel in das Programm RATE zur Schätzung eines Exponential-Modells mit zeitveränderlicher Heiratsvariable

$$\lambda(v|x(v)) = \exp(x_k'(v)\beta)$$

verwendet werden.

Das Ergebnis dieser Schätzung ist in Modell (2) von Tabelle 1 zu finden. Der β-Koeffizient der zeitveränderlichen Kovariable HEIRAT ist signifikant und hat erwartungsgemäß ein negatives Vorzeichen. Es besagt, daß sich die Neigung zum Berufswechsel nach einer Heirat deutlich vermindert. Im Vergleich zu den unverheirateten Männern vermindert sich die Mobilitätsrate bei den Ehemännern um 51,03 Prozent [$(0{,}4897 - 1) \cdot 100\% = -51{,}03\%$].

Während sich diskrete zeitveränderliche unabhängige Variablen einfach in die parametrischen Raten-Modelle aufnehmen lassen, indem die ursprünglichen Verweildauern in Subepisoden aufgesplittert werden, innerhalb deren diese dann konstant sind, besteht bei stetigen zeitveränderlichen unabhängigen Variablen diese einfache Möglichkeit nicht (vgl. dazu aber die Approximationsmöglichkeiten in Blossfeld/Hamerle/Mayer, 1986, 1989). Eine unmittelbare Lösung ist nur dann gegeben, wenn die stetigen zeitveränderlichen unabhängigen Variablen eine bestimmte vorgegebene Funktion der Verweildauer sind und direkt ein Weibull-, ein Gompertz-(Makeham-), ein log-logistisches, ein log-normales oder ein Gamma-Modell zur Schätzung herangezogen werden kann.

Bei der Untersuchung der beruflichen Mobilität kann man allerdings die Verweildauer auf einem bestimmten Arbeitsplatz als Proxy-Variable für den Erwerb von berufsspezifischen Kenntnissen und die Akkumulation von Humankapital betrachten (Modell (3) von Tabelle 1). Bei Richtigkeit der Hypothese, daß die Neigung zum Berufswechsel mit zunehmender Akkumulation der berufsspezifischen Kenntnisse abnimmt, erwarten wir bei der Weibull-Verteilung ein signifikantes $\hat{\alpha}$, das zwischen 0 und 1 liegt. Die Aufnahme von Kovariablen in das Weibull-Modell soll in der Weise geschehen, daß der Parameter λ log-linear mit dem Kovariablen-Vektor x verbunden wird: $\lambda(x) = \exp(x'\beta^*)$. Das Weibull-Modell lautet dann wie folgt:

$$\lambda(v|x) = \exp(x'\beta^*)^{\alpha}\alpha v^{\alpha-1}$$

$$= \exp(x'\beta)\alpha v^{\alpha-1}$$

mit

$$\beta = \alpha\beta^*.$$

Es wurde mit dem Programm GLIM (vgl. Roger/Peacock, 1983) geschätzt. Die geschätzten β-Koeffizienten stimmen in Einflußrichtung und Signifikanz weitgehend mit dem Exponential-Modell überein. Die Prüfung der Nullhypothese $H_0: \alpha \geq 1$ gegen die Alternativhypothese $H_1: \alpha < 1$ zeigt aber, daß bei Kontrolle der Kovariablen eine monoton fallende Neigung zum Berufswechsel vorliegt:

$$z = \frac{\hat{\alpha} - 1}{s(\hat{\alpha})} = \frac{0{,}8266 - 1}{0{,}01293} = -13{,}41.$$

Das bedeutet, daß die Neigung zum Berufswechsel mit zunehmendem Erwerb von arbeitsplatzspezifischen Kenntnissen abnimmt.

VI. Abschließende Bemerkungen

Die adäquate Abbildung der Veränderungen nicht-metrischer Merkmale, die zu beliebigen Zeitpunkten eintreten können, sowie der hohe Informationsgehalt von Ereignisdaten sind große Vorzüge des ereignisorientierten Datendesigns, das dem steigenden Interesse an der Analyse von Prozessen und Verläufen in den Wirtschafts- und Sozialwissenschaften entgegenkommt. So stellt sich die Frage, warum ereignisorientierte Datenstrukturen bis heute in den Wirtschafts- und Sozialwissenschaften nur selten erhoben und analysiert worden sind.

Ein Grund dafür ist sicherlich in dem außerordentlich *aufwendigen und kostenintensiven Beobachtungsverfahren* zu suchen, das zur vollständigen Erfassung einer Ereignisgeschichte notwendig ist. Diese kann zunächst *prozeßbegleitend* geschehen, indem die Entwicklung der Merkmale der Untersuchungseinheiten über einen längeren Zeitraum mit dem Erhebungsinstrument verfolgt wird. Allerdings dauert es dabei oft sehr lange, bis die Daten schließlich für die Beantwortung einer Forschungsfrage verfügbar sind, und nicht selten haben sich die Forschungsinteressen dann bereits in eine andere Richtung entwickelt. Ereignisdaten werden deswegen häufig *retrospektiv* erhoben. Der zeitliche Verlauf der Merkmale wird dabei über einen längeren Zeitraum rekonstruiert, wie das auch bei der Lebensverlaufsstudie der Fall war. Diese Art der Datengewinnung stellt manchmal überhaupt die einzige Möglichkeit dar, ereignisorientierte Informationen zu gewinnen; so sind ja beispielsweise die bereits vergangenen Teile der Lebensverläufe der zwischen 1929—31, 1939—41 und 1949—51 Geborenen nur noch retrospektiv zugänglich. Im allgemeinen werden solche Daten aber mit dem Einwand unzureichender Zuverlässigkeit

konfrontiert; insbesondere dann, wenn die zu erinnernden Ereignisse weit in der Vergangenheit zurückliegen. Die retrospektive Erhebung von Ereignisdaten erfordert deswegen ein vergleichsweise sehr hohes Maß an Sorgfalt und Kontrolle, wie es in der Regel nur durch aufwendige Datenrecherchen und zeitraubende Dateneditionen zu erreichen ist. Werden die Daten darüber hinaus nur ein einziges Mal retrospektiv erfragt, so ist die Gefahr groß, daß die Datenbasis relativ schnell veraltet.

Deswegen werden beispielsweise beim Sozio-ökonomischen Panel (Krupp, 1985; Hanefeld, 1987) die Vorteile des *traditionellen Panels mit der retrospektiven Erhebung von Ereignisdaten verbunden.* Mit jeder neuen Panel-Welle stehen dann nicht nur jeweils aktuelle Informationen bereit, sondern durch die retrospektiven Fragen werden auch die wichtigsten Veränderungen und ihre genauen Zeitpunkte zwischen den Wellen erfaßt (zum Vergleich von Panel- und Retrospektivstudien vgl. auch Featherman, 1979—80).

Welches der beschriebenen Verfahren zur Erhebung von Ereignisdaten auch immer herangezogen wird, es handelt sich stets um außerordentlich *aufwendige und kostenintensive Prozeduren.* Indessen besteht inzwischen eine starke, meist inhaltlich motivierte Nachfrage nach dynamischen Analysen von Prozessen und Verläufen im Bereich der Wirtschafts- und Sozialwissenschaften. Sie dürfte zunehmend dazu führen, daß ereignisorientierte Datenstrukturen auch dort bereitgestellt und adäquat analysiert werden.

Anhang: Übersicht über die im Beispiel verwendeten Variablen

Variablenname	Bedeutung
BANZ	Anzahl der vorher ausgeübten Berufe
BERF	Berufserfahrung in Anzahl von Monaten
BILDG	Ausbildungsniveau in Anzahl von durchschnittlichen Schuljahren zu Beginn der Berufsperiode: 9 Jahre ≙ Volksschul- oder Hauptschulabschluß ohne Berufsausbildung 10 Jahre ≙ Mittlere Reife ohne Berufsausbildung 11 Jahre ≙ Volksschul- oder Hauptschulabschluß mit Berufsausbildung 12 Jahre ≙ Mittlere Reife mit Berufsausbildung 13 Jahre ≙ Abitur 17 Jahre ≙ Fachhochschulabschluß 19 Jahre ≙ Hochschulabschluß

HEIRAT	Familienstand: 0 ≙ unverheiratet 1 ≙ verheiratet
KOHO2	Dummy-Variable für die Kohorte 1939—41: 1 ≙ Kohorte 1939—1941 0 ≙ sonst.
KOHO3	Dummy-Variable für die Kohorte 1949—51: 1 ≙ Kohorte 1949—51 0 ≙ sonst.
KONST	Bezeichnung der Regressionskonstanten
PRES	Prestige, gemessen nach der Prestigeskala von Wegener (1985)
TDEP	Variable, die die Zeitdauerabhängigkeit im parametrischen Modell bezeichnet

Literaturverzeichnis

Allison, P. D. (1984): *Event history analysis. Regression for longitudinal event data.* Beverly Hills, CA: Sage.

Andersen, P. K., & Gill, R. D. (1982): Cox's regression model for counting processes. A large sample study. *Annals of Statistics, 10,* 1100—1120.

Blossfeld, H.-P. (1987): Zur Repräsentativität der Sfb-3-Lebensverlaufsstudie: Ein Vergleich mit Daten aus der amtlichen Statistik. *Allgemeines Statistisches Archiv, 71,* 126—144.

Blossfeld, H.-P. (1989): *Kohorendifferenzierung und Karriereprozeß — Eine Längsschnittstudie über die Veränderung der Bildungs- und Berufschancen im Lebenslauf.* Frankfurt a. M./New York: Campus.

Blossfeld, H.-P., Hamerle, A., & Mayer, K. U. (1986): *Ereignisanalyse: Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften.* Frankfurt a. M./New York: Campus.

Blossfeld, H.-P., Hamerle, A., & Mayer, K. U. (1989): *Event history analysis.* Hillsdale, NJ: Erlbaum.

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

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

Carroll, G. R. (1984): Organizational ecology. *Annual Review of Sociology, 10,* 71—93.

Carroll, G. R., & Delacroix, J. (1982): Organizational mortality in the newspaper industries of Argentina and Ireland. An ecological approach. *Administrative Science Quarterly, 27,* 169—198.

Carroll, G. R., & Huo, Y. P. (1985): Organizational task and institutional environments in ecological perspective: Findings from the local newspaper industry. American Sociological Association Meeting. Washington, D. C.

Carroll, G. R., & Huo, Y. P. (1986): Losing by winning: The paradox of electoral success by organized labor parties in the Knights of Labor era. Technical Report No. OBIR-6. Center for Research in Management, University of California, CA.

Carroll, G. R., & Mayer, K. U. (1986): Job-shift patterns in the Federal Republic of Germany: The effects of social class, industrial sector and organizational size. *American Sociological Review, 51,* 323—341.

Coleman, J. S. (1981): *Longitudinal data analysis.* New York: Basic Books.

Courgeau, D. (1984): Relations entre cycle de vie et migrations. *Population, 3,* 483—514.

Courgeau, D. (1985): Interrelation between spatial mobility, family, and career life-cycle: A French survey. *European Sociological Review, 1,* 139—163.

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

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

Cox, D. R., & Oakes, D. (1984): *Analysis of survival data.* London: Chapman & Hall.

Diekmann, A. (1987): Determinanten des Heiratsalters und Scheidungsrisikos. Eine Analyse sozio-demographischer Umfragedaten mit Modellen und statistischen Schätzmethoden. Habilitationsschrift, Ludwig-Maximilians-Universität, München.

Diekmann, A., & Mitter, P. (1984): *Methoden zur Analyse von Zeitverläufen. Anwendungen stochastischer Prozesse bei der Untersuchung von Ereignisdaten.* Stuttgart: Teubner.

Elandt-Johnson, R. C., & Johnson, N. (1980): *Survival models and data analysis.* New York: Wiley.

Featherman, D. J. (1979—1980): Retrospective longitudinal research. Methodological considerations. *Journal of Economics and Business, 32,* 152—169.

Felmlee, D., & Eder, D. (1983): Contextual effects in the classroom: The impact of ability groups on student attention. *Sociology of Education, 56,* 77—87.

Flinn, Ch. J., & Heckman, J. J. (1983): Are unemployment and out of the labor force behaviorally distinct labor force states? *Journal of Labor Economics, 1,* 28—42.

Freeman, J., Carroll, G. R., & Hannan, M. T. (1983): The liability of newness: Age dependence in organizational death rates. *American Sociological Review, 48,* 692—710.

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

Hamerle, A. (1987): Der Event-History-Ansatz zur Modellierung von Diffusions- und allgemeinen Kaufentscheidungsprozessen. *Marketing, Zeitschrift für Forschung und Praxis, 10,* 248—257.

Hamerle, A. (1989): Multiple-spell regression models for duration data. *Applied Statistics, 38,* 127—138.

Hamerle, A., & Tutz, G. (1989): *Diskrete Modelle zur Analyse von Verweildauern und Lebenszeiten.* Frankfurt: Campus.

Hamerle, A., Kemény, P., & Tutz, G. (1984): Kategoriale Regression. In: L. Fahrmeier & A. Hamerle (Hrsg.), *Multivariate statistische Verfahren* (Kap. 6). Berlin: Springer.

Hanefeldt, U. (1987): *Das Sozio-ökonomische Panel. Grundlagen und Konzeption.* Frankfurt a. M./New York: Campus.

Hannan, M. T., & Freeman, J. (1977): The population ecology of organization. *American Journal of Sociology, 82,* 929—964.

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

Heckman, J. J., & Singer, B. (1985): *Longitudinal analysis of labor market data.* Cambridge: Cambridge University Press.

Huinink, J. (1987): Soziale Herkunft, Bildung und das Alter bei der Geburt des ersten Kindes. *Zeitschrift für Soziologie, 16* (5), 367—384.

Huinink, J. (1988): Die demographische Analyse der Geburtenentwicklung mit Lebensverlaufsdaten. *Allgemeines Statistisches Archiv, 72,* 359—377.

Hujer, R., & Schneider, H. (1988): Unemployment duration as a function of individual characteristics and economic trends. In: K. U. Mayer & N. B. Tuma (Hrsg.), *Event history analysis in life course research.* Madison, WI: University of Wisconsin Press.

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

Kemény, P. (1986): Regressionsmodelle zur Analyse von Verweildauern — ein Software-Vergleich. In: W. Lehmacher, A. Hörmann (Hrsg.), *Statistik-Software. 3. Konferenz über wissenschaftliche Anwendung von Statistik-Software 1985.* Stuttgart/New York: Springer.

Kiefer, N. M. (1988): Economic duration data and hazard functions. *Journal of Economic Literature, XXVI,* 646–679.

Kiefer, N. M., Lundberg, S., & Neumann, G. R. (1985): How long is a spell of unemployment? *Journal of Econ. Statist. Apr. 3,2,* 118–128.

Krupp, H.-J. (1985): Das Sozio-ökonomische Panel. Bericht über die Forschungstätigkeit 1983–1985. Antrag auf Förderung der Forschungsphase 1986–1988. Frankfurt a. M./Berlin.

Lawless, J. F. (1982): *Statistical models and methods for life-time data.* New York: Wiley.

Mayer, K. U., & Wagner, M. (1986): Wann verlassen die Kinder ihr Elternhaus? Untersuchungen zu den Geburtsjahrgängen 1929–31, 1939–41, 1949–51. IBS-Materialien, Institut für Bevölkerungsforschung und Sozialpolitik, Universität Bielefeld.

Mayer, K. U., & Carroll, G. R. (1987): Jobs and classes: Structural constraints on career mobility. *European Sociological Review, 3,* 14–38.

Mayer, K. U., & Brückner, E. (1989): Lebensverläufe und Wohlfahrtsentwicklung. Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929–1931, 1939–1941, 1949–1951. Materialien aus der Bildungsforschung Nr. 35, Max-Planck-Institut für Bildungsforschung, Berlin.

Mayer, K. U., & Tuma, N. B. (Hrsg.) (1989): *Event history analysis in life course research.* Madison, WI: University of Wisconsin Press.

Michael, R. T., & Tuma, N. B. (1985): Entry into marriage and parenthood by young men and women: The influence of familiy background. *Demography, 22,* 515–544.

Roger, J. H., & Peacock, S. D. (1983): Fitting the scale as a GLIM parameter for Weibull, extreme value, logistic and log-logistic regression models with censored data. *GLIM-Newsletter, 6,* 30–37.

Sandefuhr, G. D., & Scott, W. J. (1981): A dynamic analysis of migration: An assessment of the effects of age, family, and career variables. *Demography, 18,* 355–368.

Sørensen, A. B. (1984): Interpreting time dependency in career processes. In: A. Diekmann & P. Mitter (Hrsg.), *Stochastic modelling of social processes* (pp. 89–122). New York: Academic Press.

Sørensen, A. B., & Sørensen, A. (1986): An event history analysis of the process of entry into first marriage. *Current Perspectives on Aging and Life Cycle, 2,* 53–71.

Sørensen, A. B., & Tuma, N. B. (1981): Labor market structures and job mobility. *Research in Social Stratification and Mobility, 1,* 67–94.

Tuma, N. B. (1985): Effects of labor market structure on job-shift patterns. In: J. J. Heckman & B. Singer (Hrsg.), *Longitudinal analysis of labor market data.* Cambridge, MA: Cambridge University Press.

Tuma, N. B. (1986): Invoking rate. Program Manual. Stanford University.

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

Wagner, M. (1989): *Räumliche Mobilität im Lebensverlauf. Eine empirische Untersuchung sozialer Bedingungen der Migration.* Stuttgart: Enke.

Wu, L. (1988): Simple graphical goodness-of-fit tests for hazard rate models. In: K. U. Mayer & N. B. Tuma (Hrsg.), *Event history analysis in life course research.* Madison, WI: University of Wisconsin Press.