Statistische Modellierung von Verlaufsdaten in der Analyse sozialer Probleme, Teil I: Grundlagen[1]

von Wolfgang Ludwig-Mayerhofer

Abstract

Although „dynamic" notions like the concept of career have a long history in the sociology of social problems, still most research is static, referring to the number and the characteristics of people belonging to problem groups at a given point in time. Methods of „event history analysis" (also called „survival analysis" or „analysis of failure time") provide flexible tools for analyses which take into account the changes that take place on the individual level. They model processes where we have information (a) about the time individuals spend in a given social state, and (b) about the state they occupy after a change has occurred, if any. This paper provides an introduction to the basic concepts of event history analysis, that is, survivor, density, and hazard function, and illustrates these by using example data from the Socio-Economic Panel (SOEP). A companion paper, to appear in the next volume, will give an overview on various types of multivariate models and related issues.

Zusammenfassung

Obwohl dynamische Konzepte wie der Karriere-Begriff schon lange in der Soziologie sozialer Probleme etabliert sind, werden vielfach immer noch statische Untersuchungen durchgeführt, die sich auf die Anteile und die Merkmale von Personen in bestimmten sozialen Lagen zu bestimmten Zeitpunkten beziehen. Methoden der Analyse von Verlaufsdaten („Ereignisanalyse", „Survivalanalyse") erlauben es, in flexibler Weise Veränderungen auf der Individualebene zu untersuchen. Sie modellieren Prozesse, zu denen Informationen (a) über die Zeit, die Individuen in einer sozialen Lage verbringen, und (b) über die Lage, die sie im Anschluß einnehmen (falls eine Veränderung geschehen ist), vorliegen. Diese Arbeit führt in die Grundbegriffe dieser Modelle ein (Survivor-, Dichte- und Hazardfunktion) und erläutert sie anhand eines Beispieles aus dem Sozio-ökonomischen Panel (SOEP). Eine weitere Arbeit, die für die nächste Ausgabe der Zeitschrift vorgesehen ist, wird verschiedene Möglichkeiten der multivariaten Auswertung solcher Daten diskutieren.

1. Zur Problemstellung

In der sozialwissenschaftlichen Analyse sozialer Probleme herrscht häufig eine statische Sichtweise vor. Untersucht wird der *Umfang* eines sozialen Problems, also die Anzahl von Armen, Kranken, Straffälligen, Arbeitslosen usw., zu einem gegebenen *Zeitpunkt*, und wenn nach relevanten *Merkmalen* der jeweiligen Personen(gruppen) gefragt wird, bezieht sich dies dementsprechend auf die zu einem be-

stimmten Zeitpunkt Betroffenen. In letzter Zeit ist jedoch zunehmend deutlich geworden, daß derartige auf Bestandsdaten bezogene Analysen nur *einen* Aspekt der Problematik erfassen und die sehr beachtliche „Dynamik" des untersuchten Sachverhalts auf der Individualebene vernachlässigen. Denn während der Anteil der Armen, Arbeitslosen usf. an der Population häufig über Jahre konstant bleibt oder sich nur langsam ändert, ist auf der Ebene der betroffenen Personen eine oftmals sehr ausgeprägte „Mobilität" zu beobachten. Viele soziale Lagen wie Armut, Arbeitslosigkeit usw. sind reversibel; sie werden im Zeitverlauf zwar einerseits von sehr viel mehr Personen eingenommen, als die Bestandsdaten nahelegen, andererseits jedoch auch wesentlich häufiger und schneller wieder verlassen.[2]

Im Grunde sollte dies für Sozialwissenschaftler gar nicht überraschend sein. So gehört zum gängigen Vokabular in der Analyse sozialer Probleme der „Karriere"-Begriff, also die Vorstellung eines sich stufenweise aufschaukelnden, möglicherweise aber auch reversiblen *Prozesses*, in dem ein Individuum zu einem „sozialen Problemfall" werden, aber unter Umständen seine problematische Lebenslage auch wieder verlassen kann.[3] Ganz grundsätzlich läßt sich festhalten, daß viele Fragestellungen in der Analyse sozialer Probleme den *Aspekt der Zeit, also eine dynamische Sichtweise* thematisieren: Wie lange dauert die Arbeitslosigkeit von Individuen? Wann ist mit dem - erstmaligen oder erneuten - Auftreten einer Krankheit, wann ist u.U. mit einer sozialen Marginalisierung aufgrund dieser Krankheit zu rechnen? Wie schnell werden sanktionierte Straftäter erneut straffällig?

In den letzten 20 Jahren wurde die Entwicklung statistischer Modelle vorangetrieben, mit deren Hilfe solche Untersuchungsfragen angemessen analysiert werden können.[4] In vielen Bereichen - etwa der Arbeitsmarkt- und Mobilitätsforschung oder der Demographie - gehören sie zum etablierten Methodenbestand. Doch obwohl diese Verfahren bereits relativ gut entwickelt und erprobt sind, haben sie noch wenig Aufnahme in die Standard-Methodenlehrbücher gefunden. Das erschwert sowohl ihre Anwendung im Forschungsprozeß als auch die Rezeption einschlägiger Ergebnisse. In dieser Arbeit sollen die Grundlagen und Anwendungsmöglichkeiten dieser Verfahren anhand von einfachen Beispielen dargestellt werden. Das Ziel der Arbeit ist didaktisch; es geht darum, die grundlegende Logik der Verfahren zu erläutern und intuitiv verständlich zu machen, während ein Anspruch auf Originalität oder Innovation nicht erhoben wird. Der statistische Hintergrund der Verfahren soll nur ganz rudimentär behandelt werden. Das geschieht nicht, weil dieser Hintergrund irrelevant wäre, sondern in der Absicht, zunächst einen einfachen Zugang zu ermöglichen, der dann entsprechend vertieft werden kann. Dafür versuche ich zumindest gelegentlich, Fragen des praktischen Vorgehens bei der Datenanalyse anzusprechen. Die Arbeit hat zwei Teile. In diesem Teil werden die elementaren Grundlagen und Konzepte dargestellt; in einem zweiten Teil, der im folgenden Heft der *Sozialen Probleme* erscheinen soll, werden die wichtigsten für Sozialwissenschaftler relevanten Techniken der Datenanalyse vorgestellt und diskutiert.

2. Zur Unangemessenheit vertrauter statistischer Modelle

„Karrieren“ oder „Verläufe“ sind zeitliche Phänomene. Sie lassen sich, vereinfacht formuliert, anhand der *Dauer* beschreiben, die Untersuchungseinheiten (Personen, Familien etc.) in bestimmten Zuständen verbleiben, z.B. in Arbeitslosigkeit oder Krankheit, oder umgekehrt in Beschäftigung oder Gesundheit (ausführlicher dazu siehe Abschnitt 3.1). Man könnte daher zu der Annahme verleitet werden - und dies geschieht offensichtlich immer noch häufig, so beispielsweise bei Jacobs/ Ringbeck (1992) -, daß diese Datenkonstellation problemlos mit dem klassischen statistischen Verfahren der linearen Regression auf der Basis des Kleinst-Quadrat-Schätzers („OLS“) analysiert werden kann, da es sich bei der untersuchten Zielvariablen, der Dauer, um ein metrisches, intervallskaliertes Merkmal zu handeln scheint. Tatsächlich sprechen gravierende Gründe gegen die Anwendung der linearen Regression auf Verlaufsdaten.

Der zweifellos wichtigste Grund ist darin zu sehen, daß die interessierende Dauer *im allgemeinen nicht bei allen Untersuchungseinheiten gemessen werden kann*. Denn diese Dauer ist erst vollständig bekannt, wenn ein Individuum den betreffenden Zustand wieder verläßt, wenn also ein Zustandswechsel geschieht. Im Regelfall tritt aber bei einer kleineren oder größeren Anzahl der Fälle kein Zustandswechsel auf (Personen bleiben arbeitslos, werden nicht [erneut] straffällig, usw.). Ferner ist häufig der Beobachtungszeitraum faktisch begrenzt, so daß selbst dann, wenn ein Zustandswechsel untersucht wird, der bei jedem Individuum eintritt - etwa der Tod -, dieser vom Forscher nicht in allen Fällen beobachtet werden kann. Schließlich ist damit zu rechnen, daß bei einer Längsschnittuntersuchung ein Teil der Personen schon vor dem geplanten Untersuchungsende ausscheidet. In all diesen Fällen ist zwar bekannt, daß ein Individuum sich bis zu einem bestimmten Zeitpunkt in dem untersuchten Ausgangszustand befand, es fehlt aber die „positive“ Information über den „Zustandswechsel“ oder „Übergang“ in einen neuen Zustand. Man spricht hier von *(rechts-)zensierten Daten* oder einfach „Zensierungen“.

Liegen zensierte Daten vor, so ist einsichtig, daß das lineare Regressionsmodell inadäquat ist. Weder kann man die zensierten Fälle einfach weglassen, noch kann man so tun, als handele es sich bei den zensierten Beobachtungsdauern um vollständige Beobachtungen. In beiden Fällen werden nämlich die wahren Dauern unterschätzt, und zwar *in systematischer Weise*.[5] Aber noch weitere Gründe sprechen gegen das lineare Regressionsmodell: Die abhängige Variable weicht oft sehr stark von der Normalverteilung ab, und sie ist manchmal nicht exakt (stetig), sondern nur annäherungsweise, nämlich in diskreten Intervallen gemessen worden.

Sofern keine Untersuchungseinheiten vorzeitig ausgefallen sind, mithin die Beobachtungsdauer bei allen Untersuchungseinheiten identisch ist und Zensierungen - nicht eingetretene Zustandswechsel - nur am Ende der Beobachtungsdauer auftreten, wäre als alternative Auswertungsstrategie an ein Regressionsmodell für binäre abhängige Variablen (z.B. Logit- oder Probit-Modell) zu denken, dessen abhängige Variable als „Zustandswechsel bis zum Ende der Beobachtungsdauer eingetreten:

ja/nein“ formuliert werden kann. Trotzdem vergibt man sich auch hier wichtige Analysemöglichkeiten, da nur der Anteil der Zustandswechsel bis zu einem einzigen, möglicherweise recht willkürlich gewählten Zeitpunkt berücksichtigt wird. Sehr oft ist jedoch die „Zeitabhängigkeit“ des untersuchten Prozesses von Interesse: Ist das „Rückfallrisiko“ von Straffälligen zu jedem Zeitpunkt gleich, oder gibt es zunächst einen Abschreckungseffekt, der mit zunehmendem Abstand nachläßt? Hält die Wirkung einer ärztlichen Behandlung eine Zeit lang an, um dann abrupt aufzuhören, ist eine kontinuierliche, oder vielleicht gar keine Abnahme der Wirksamkeit zu beobachten? Wenn solche Fragen von Bedeutung sein können, was sich beim gegenwärtigen Wissensstand nur selten ausschließen läßt, verschenkt eine rein zeitpunkt-orientierte Betrachtung sehr gewichtige Informationen (vgl. hierzu insgesamt Breslow 1991 oder Yamaguchi 1991: 9).

Alle geschilderten Probleme lassen sich jedoch (zumindest im Prinzip) durch statistische Modelle lösen, die in den letzten Jahren unter Begriffen wie „Event History Analysis“ (Tuma/Hannan 1984), „Verlaufsdatenanalyse“ (Andreß 1992), „Ereignisanalyse“ (Blossfeld/Hamerle/Mayer 1986) oder „Survivalanalyse“ (Lee 1980) bekannt geworden sind. Die Vorzüge dieser Verfahren lassen sich nicht nur negativ, in Abgrenzung von ungeeigneten Verfahren beschreiben, es ist auch zu betonen, daß damit ganz neue Untersuchungsmöglichkeiten eröffnet werden. Zum einen kann, wie schon erwähnt, der „Zeitabhängigkeit“ der Prozesse - im Sinne von Änderungen der Wahrscheinlichkeit, den untersuchten Ausgangszustand zu verlassen - häufig durch die Wahl einer entsprechenden Verteilungsannahme für die abhängige Variable Rechnung getragen werden. Zweitens und vor allem ist auf die Möglichkeit zu verweisen, auch *solche erklärenden Variablen zu berücksichtigen, deren Werte sich selbst im Zeitverlauf ändern* (vgl. Andreß 1992: 16 ff.).

Bevor ich fortfahre, möchte ich einige Hinweise auf weiterführende Literatur geben. Als wesentliche *Lehrbuch-Einführungen* sind zu nennen: Allison (1984), Andreß (1992), Blossfeld/Hamerle/Mayer (1986), Diekmann/Mitter (1984a), Kalbfleisch/Prentice (1980), Lawless (1982), Lee (1980), Namboodiri/Suchindran (1987) und Yamaguchi (1991), unter denen die Arbeit von Lee mit Abstand am leichtesten verständlich (und mit vielen Beispielen versehen), allerdings nicht mehr ganz auf dem neuesten Stand ist. Unter den deutschsprachigen Lehrbüchern ist dasjenige von Blossfeld/Hamerle/Mayer (1986) am ausführlichsten und grundsätzlich auch sehr anwendungsorientiert, jedoch sollten unbedingt Hinweise auf neuere EDV-Programme beachtet werden, wie sie sich bei Andreß (1992) und in dieser Arbeit finden. - In den folgenden Ausführungen verweise ich jeweils auf *Fundstellen in den drei deutschen Lehrbüchern* mit den Abkürzungen *A* für Andreß, *BHM* für Blossfeld/Hamerle/Mayer und *DM* für Diekmann/Mitter. - Das Buch von Tuma/Hannan (1984), das sich ebenso wie die Arbeiten von Tuma/Hannan/Groeneveld (1979) und Tuma (1982) als Einführung versteht, ist dagegen teilweise relativ anspruchsvoll und nicht sehr eingängig. Die Monographie von Lancaster (1990) und die Sammelbände von Crouchley (1987) und Diekmann/Mitter (1984b) sind ebenfalls nur mit recht weit fortgeschrittenen Kenntnissen zugänglich. - Unter den

nicht-monographischen Arbeiten seien als Einführungen Carroll (1983), Hutchison (1988a), Kiefer (1988), Teachman (1983) sowie Kap. 6 in Toutenburg (1992) genannt, als Übersichtsartikel (teilweise auch zu neueren und komplexeren Verfahren bzw. Problemen) Diekmann (1988), Diekmann/Mitter (1990, 1993), Hutchison (1988b), Meinken (1992) sowie Petersen (1990, 1991a, 1993); in sehr mathematisch-straffer Form führt auch Arminger (1984, 1988) in alle wesentlichen Begriffe und Modelle ein. Ebenfalls sehr hilfreich ist das Manual des Programms TDA (Rohwer 1993), mit dem auch die in dieser Arbeit vorgestellten Beispiele berechnet wurden.

3. Grundkonzepte der Verlaufsdatenanalyse

In diesem Abschnitt erläutere ich zunächst die Art der Daten, die der Verlaufsdatenanalyse zugrunde liegen. Anschließend stelle ich die wichtigsten Grundkonzepte an Beispielen dar mit dem Ziel, ein intuitives Verständnis dieser Konzepte zu ermöglichen und aufzuzeigen, wie mit dem Problem der rechtszensierten Daten umgegangen werden kann.

3.1 Vorbemerkungen zur Datenstruktur

Eine Beschreibung von Prozessen setzt sich aus zwei Aspekten zusammen. Erstens benötigt man Informationen über die sozialen Lagen, Positionen usw. - allgemein: *Zustände* -, die Individuen in diesem Prozeß einnehmen können, etwa „arbeitslos - beschäftigt“. Die Gesamtheit der für eine bestimmte Untersuchungsfrage relevanten Zustände wird als *Zustandsraum* bezeichnet. Dabei interessiert man sich vor allem für die *Zustandswechsel* - oder „Übergänge“ oder „Ereignisse“[6] - von einem Ausgangs- in einen Zielzustand. Zweitens benötigt man Angaben über die *Verweildauer* im Ausgangszustand, anders gesagt, über die Dauer bis zum Eintreten eines Zustandswechsels (man spricht daher auch von Wartezeiten [im Ausgangszustand] oder Ankunftszeiten [im Zielzustand]). Beide Aspekte, Zustandsraum und Verweildauer, charakterisieren eine *Episode* (englisch „spell“, was auch manchmal in deutschsprachigen Arbeiten gebraucht wird) (vgl. A: 45 ff.; BHM: 27 ff.; DM: 33 ff.). Wenn, wie es häufig der Fall ist, nur je ein einziger Ausgangs- und Endzustand vorliegt (bzw. untersucht wird), genügt auch die Information, ob die Episode mit einem Zustandswechsel endete oder nicht (also rechtszensiert ist).

Wie beim Beispiel einer arbeitslosen Person unmittelbar deutlich wird, sind manche Prozesse, etwa der Erwerbsverlauf, durch mehr als zwei Zustände gekennzeichnet - im genannten Beispiel kann die Arbeitslosigkeit nicht nur in eine Beschäftigung münden, sondern auch in eine berufliche Erstausbildung, eine berufliche Weiterbildung, Umschulung, längere Krankheit und insbesondere bei Frauen auch in Unterbrechungen der Erwerbstätigkeit, schließlich in den vorzeitigen oder altersgemäßen Ruhestand. Allgemein spricht man, wenn aus einem Zustand ein Übergang in mehrere andere Zustände möglich ist, davon, daß die Personen *kon-*

kurrierenden Risiken (englisch: *competing risks)* ausgesetzt sind. Auf die Behandlung dieses Problems gehe ich in Teil II kurz ein und beschränke mich hier auf Zwei-Zustands-Modelle, also Modelle, bei denen ein Wechsel nur aus einem Ausgangszustand in *einen* anderen Zielzustand stattfinden kann.

Auch unter dieser vereinfachenden Voraussetzung bleibt noch als weiteres Problem, daß viele Episoden im Leben eines Individuums *mehrfach auftreten* können. Es ist aber ohne weiteres einsichtig, daß wiederholte Episoden möglicherweise anderen Bedingungen unterliegen als erste. Beispielsweise mag jemand, der zum zweiten Mal oder noch häufiger arbeitslos wurde, bei der Arbeitssuche im Vorteil sein, weil er bereits früher Erfahrungen mit dem Arbeitsamt und bei der Stellensuche machte; auf der anderen Seite könnte allein die Tatsache einer wiederholten Arbeitslosigkeit etwa potentielle Arbeitgeber mißtrauisch werden lassen. Dieses Beispiel verdeutlicht, daß die verschiedenen Episoden eines Individuums untereinander *abhängig* sein können, was in der Datenauswertung berücksichtigt werden muß. Es ist daher im allgemeinen ratsam, zwischen ersten und weiteren Episoden zu unterscheiden und bei Vorliegen mehrerer Episoden entsprechende Mehr-Episoden-Modelle zu schätzen (vgl. Teil II).

Anzumerken ist vielleicht auch noch, daß die Definition eines „Zustandes" grundsätzlich eine *inhaltliche* Frage ist. So könnte man es durchaus für sinnvoll erachten, einen „Zustandswechsel" anzunehmen, wenn auf einer eigentlich kontinuierlichen Skala (etwa dem Einkommen) ein bestimmter Schwellenwert über- oder unterschritten wird, etwa eine wie auch immer definierte „Armutsgrenze". Ob dies sinnvoll ist oder nicht, kann natürlich nicht mit Hilfe der Statistik, sondern nur aus den jeweiligen substantiell-inhaltlichen Annahmen einer Fachdisziplin entschieden werden.[7]

Nun einige Bemerkungen zum Aspekt der *Verweildauern*, genauer ihrer *Messung*. Die meisten der später diskutierten Modelle basieren auf der Annahme stetig gemessener Zeiten. Nun muß man in der Forschungspraxis häufig Abstriche von dieser Annahme machen. So können *erstens* manche Zustandswechsel gar nicht jederzeit, sondern nur zu bestimmten Zeitpunkten stattfinden: Politische Wahlen finden in mehrjährigem Abstand statt, in der Schule wird einmal im Jahr über die Versetzung in die nächst höhere Klasse entschieden, usf. Für solche *diskreten Zeitpunkte* wurden bereits einige Verfahren entwickelt (näheres in Teil II). *Zweitens* liegen auch bei im Prinzip stetiger Zeit oft nicht sehr genaue Messungen vor, so daß man etwa nur weiß, in welcher Woche oder in welchem Monat, allgemein: in welchem Zeitraum oder Intervall ein Zustandswechsel eingetreten ist; man spricht von sog. *gruppierten* oder *aggregierten* Verweildauern. In beiden Fällen werden in aller Regel oft mehrere Zustandswechsel zum gleichen Zeitpunkt bzw. im gleichen Zeitraum auftreten (sog. „Ties"), was für manche Verfahren im Prinzip gleichfalls unerwünscht ist (siehe z.B. Hutchison 1988a).[8]

Nun ist zu konstatieren, daß die Probleme teilweise auf der abstrakten Ebene größer aussehen als in der Praxis. Das beginnt mit der Frage, wann eine Messung exakt genug ist, um als „stetig" bzw. „nicht aggregiert" aufgefaßt werden zu kön-

nen. Soll die Zeit bis zum Auftreten eines Zustandswechsels in Sekunden, Minuten, Stunden, Tagen, Wochen, Monaten oder gar Jahren gemessen werden? Diese Frage muß sicherlich unterschiedlich beantwortet werden, je nachdem um welchen Zustand es sich handelt (die Dauer von Ehen ist im Durchschnitt wesentlich länger als die von Arbeitslosigkeitsepisoden). Es geht also offensichtlich vor allem um die *relative Genauigkeit der Messung*. Nach einigen neueren Arbeiten (z.B. Arminger 1984; Galler 1986; Petersen/Koput 1992) kann man davon ausgehen, daß zumindest bei in der Zeit konstanten Risiken (vgl. Abschnitt 3.2 und Teil II) einigermaßen valide Ergebnisse erzielt werden können, wenn die Zeitskala, auf der die Dauer gemessen wird, maximal die Hälfte des Medians der Verweildauer (vgl. Abschnitt 3.2) ausmacht. Wie sich die Dinge bei zeitlich variablen Risiken verhalten, wie sie in der Praxis häufiger auftreten, ist dagegen weniger eindeutig zu sagen.

Für gruppierte oder aggregierte Verweildauern gibt es ein exploratives Verfahren, den sog. Life-Table-Schätzer, dessen wir uns weiter unten ausführlich bedienen werden. Für eine multivariate Analyse solcher Daten existieren dagegen nur wenige Ansätze, die zudem in den verfügbaren Statistikprogrammen kaum implementiert sind (vgl. als Überblick Heijtan 1989). Sofern man von einer hinreichenden relativen Meßgenauigkeit ausgehen kann, werden daher in der Praxis zumeist Verfahren angewendet, die sich eigentlich auf kontinuierliche Verweildauern beziehen. Für diesen Fall wurde aufgrund theoretischer Überlegungen (Hujer/Schneider 1986, 1989) wie von Simulationsstudien (Petersen 1991b, 1993; Petersen/Koput 1992) vorgeschlagen, von den gemessenen Dauern eine halbe Zeiteinheit abzuziehen. Die Überlegung ist einfach: Wenn man annimmt, daß die Zustandswechsel sich in etwa gleichmäßig auf das Zeitintervall verteilen, so kann der Mittelpunkt des Intervalls als beste Schätzung des Eintritts des Zustandswechsels gelten. Ersichtlich gilt dies nur dann, wenn die Annahme gleicher Verteilung zutrifft, technisch gesprochen, wenn die Hazardfunktion im Zeitverlauf konstant ist (s. unten). Ist sie das nicht, müssen u.U. andere Transformationen der beobachteten Dauern vorgenommen werden (Petersen/Koput 1992) - wobei sich natürlich die Katze ein wenig in den Schwanz beißt, da sich die Zeitabhängigkeit der Hazardfunktion in der Regel erst anhand der empirischen Auswertung zeigt.[9] Offensichtlich können solche Transformationen die Ergebnisse gerade im Hinblick auf die Zeitabhängigkeit der Verweildauern nicht unbeträchtlich beeinflussen (vgl. Ludwig-Mayerhofer 1992).

Nun wird im allgemeinen angenommen, daß zwar - nicht zuletzt aufgrund des Problems gruppierter Messungen - Aussagen über die Zeitabhängigkeit von Prozessen oft nicht eindeutig zu treffen sind, daß aber die Einflüsse von erklärenden Variablen von diesem Problem weitgehend unberührt bleiben (Galler/Pötter 1992). Wie wir in Teil II sehen werden, muß auch diese Annahme, so richtig sie im Kern sein dürfte, im Einzelfall nicht unbedingt zutreffen.

Ganz kurz sei auf einen weiteren Aspekt hingewiesen. Im allgemeinen untersucht man einen konkreten Prozeß - etwa den Prozeß der Arbeitslosigkeit, der Drogenabhängigkeit, der Gesundung - ab seinem Beginn. Das heißt, alle Individuen

beginnen den Prozeß zum Zeitpunkt „Null“, etwa beim Eintritt in die Arbeitslosigkeit, beim ersten Drogenkonsum, bei Beginn der Erkrankung oder der Behandlung. Wir sprechen hier von einer Analyse der *Prozeßzeit*, und dieses Verfahren wird in der weitaus größten Zahl der Untersuchungen zugrundegelegt. Es ließen sich jedoch andere Auffassungen von der relevanten Zeit denken. Man könnte den Prozeß auf einen anderen „Null-Zeitpunkt“ beziehen, etwa auf den Beginn des Erwerbslebens, auf die Geburt, usf. (vgl. DM: 25; Blossfeld/Hamerle 1989). Hierauf kann an dieser Stelle nicht weiter eingegangen werden, und im folgenden gehe ich immer von Analysen der Prozeßzeit aus.

Schließlich ist noch auf das Problem der *Zensierungen* einzugehen. Oben wurde unterstellt, daß die untersuchten Individuen *von Anfang an* beobachtet wurden, also ab dem Zeitpunkt, zu dem sie den Ausgangszustand eingenommen haben; nur das *Ende* wird unter Umständen nicht beobachtet (weswegen man von *rechts*zensierten Daten spricht). Dabei werden häufig verschiedene Zensierungstypen unterschieden.[10] Der wesentliche Gesichtspunkt läßt sich jedoch relativ einfach formulieren: Die Zensierungen sollen zufällig erfolgen, was häufig auch so formuliert wird, daß Zensierungen und „Ereignisse“, also Zustandswechsel, voneinander unabhängig sein sollen. In der Praxis läßt sich diese Annahme allerdings oft nicht überprüfen, sondern nur mehr oder wenig glaubhaft vertreten.[11]

Wenn dagegen eine Episode nicht von Anfang an beobachtet werden konnte, d.h., wenn zu Beginn der Beobachtung nicht bekannt ist, wie lange die Untersuchungseinheiten sich bereits im Ausgangszustand befinden, spricht man von *linkszensierten Daten*. Ein Beispiel wäre eine Untersuchung von Personen im Arbeitslosenbestand, bei denen nicht erhoben wird, wie lange sie bereits arbeitslos sind, und somit nur die Arbeitslosigkeitsdauer ab Beginn der Untersuchung (und nicht ab Beginn der Arbeitslosigkeit) gemessen werden kann.[12] Liegen Linkszensierungen vor, sollte der Datensatz *nicht* oder nur mit größter Vorsicht ausgewertet werden, da hier ja grundsätzlich keine Angabe zur Dauer der jeweiligen Episode gemacht werden kann, ganz unabhängig davon, ob ihr Ende - der Übergang in den Zielzustand - beobachtet werden kann oder nicht (A: 94 f.; BHM: 26). Allenfalls wenn gut begründete Annahmen vorliegen, daß das Ausmaß der Linkszensierung nur sehr geringfügig ist und/oder daß die „Geschichte“ des Prozesses bis zum Beginn der Beobachtung ohne Einfluß auf den weiteren Verlauf ist, könnte an einen Versuch der Auswertung gedacht werden. Gerade letztere Annahme ist aber in der Regel eher zweifelhaft.

3.2 Survivor-, Dichte- und Hazardfunktion

Wie im zweiten Abschnitt verdeutlicht wurde, ist es wegen der rechtszensierten Daten nicht sinnvoll, die Dauer der beobachteten Episoden unmittelbar als abhängige Variable zu verwenden. Stattdessen werden einige andere Funktionen von Verweildauern herangezogen, die sich auch beim Vorliegen von Rechtszensierungen schätzen lassen. Ich will im folgenden *zuerst* ein *fiktives Beispiel* bringen, in welchem keine Zensierungen auftreten, und zeigen, daß die zu erläuternden Funk-

tionen sich in diesem Fall tatsächlich ganz einfach auf die beobachteten Daten zurückführen lassen. Gleichzeitig soll aber gezeigt werden, wie man diese Funktionen in einer Art und Weise schätzen kann, die auch bei rechtszensierten Daten anwendbar ist. Dies wird im Anschluß anhand eines *zweiten* Beispiels mit Daten aus dem „*Sozio-ökonomischen Panel" (SOEP)* (vgl. Projektgruppe „Das sozio-ökonomische Panel" 1990) noch einmal ausführlich dargestellt.

Die nachfolgenden Beispiele wenden *eine spezifische Form* der Schätzung der relevanten Größen an, die sog. „Life Table-Schätzung" (A: 139 ff.; BHM: 42 f., 116 ff.; DM: 60 ff.).[13] Es handelt sich hierbei um die einfachste Form der Verlaufsdatenanalyse, aber gerade deshalb ist sie für eine Einführung am sinnvollsten.[14] Denn dieses Verfahren bezieht sich auf gruppierte Verweildauern (wobei stetige Verweildauern in gruppierte Form gebracht werden), während bei stetigen Dauern auf weniger anschauliche Grenzwertbetrachtungen zurückgegriffen werden muß.

Darstellung 1: *Grunddaten (fiktive Arbeitslosigkeitsdauern, n = 10, Dauern in Wochen)*

Person	Dauer	Person	Dauer
A	1	F	6
B	2	G	7
C	3	H	8
D	4	I	9
E	5	J	10

Wir beginnen mit einem ganz einfachen, konstruierten Beispiel von zehn Untersuchungspersonen (vgl. *Darstellung 1*).[15] Auch wenn es grundsätzlich beliebig ist, welcher Prozeß untersucht wird, sei der besseren Verständlichkeit halber davon ausgegangen, daß es sich um die Dauer von Arbeitslosigkeitsepisoden handeln soll. Da wir in diesem Beispiel zensierte Daten unberücksichtigt lassen wollen, sollen die gemessenen Verweildauern die tatsächliche Dauer bis zum Verlassen der Arbeitslosigkeit repräsentieren. Wie gesagt, betrachten wir Zeitdauern, die gruppiert oder aggregiert sind, und zwar auf der Basis von Wochen. Eine Arbeitslosigkeitsdauer von 1 Woche[16] soll bedeuten, daß die Arbeitslosigkeit *während der ersten Woche* beendet wurde, usw.[17] Ersichtlich verläßt jede Woche genau eine Person die Arbeitslosigkeit, so daß nach 10 Wochen sämtliche Personen nicht mehr arbeitslos sind.

In *Darstellung 2* werden nun einige Funktionen vorgestellt, mit denen sich die Daten aus Darstellung 1 unter dem Aspekt des zeitlichen Verlaufs charakterisieren lassen. Dabei wird nun sozusagen nicht mehr die einzelne Untersuchungsperson, sondern die *Zeit* zum Untersuchungsgegenstand, denn die einzelnen Zeilen beziehen sich jetzt auf die zehn Wochen, die der hier untersuchte Prozeß insgesamt dauerte (vgl. Spalte 1).[18] In der zweiten Spalte ist angegeben, bei wievielen Personen

(Untersuchungseinheiten) *zu Beginn* dieses Intervalls *noch kein Zustandswechsel (Ereignis)* stattgefunden hatte; man kann auch von der „Risikomenge“ sprechen, nämlich der Zahl der Personen, die jeweils noch dem „Risiko“ (statistisch gesprochen!) ausgesetzt sind, ein Ereignis zu „erleiden“, d.h., die Arbeitslosigkeit zu verlassen. Notwendigerweise steht hier in der ersten Zeile die Gesamtzahl der Untersuchungseinheiten. Die dritte Spalte zeigt, bei wievielen Personen *während* des Intervalls *ein Zustandswechsel eingetreten ist*. Der Wert in der zweiten Spalte abzüglich des Wertes der dritten Spalte ergibt demzufolge den Wert der zweiten Spalte in der darauffolgenden Woche.

Darstellung 2: *Life-Table-Schätzung für die fiktiven Beispieldaten aus Darstellung 1*

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0 bis <1	10	1	0.100	0.900	1.000	0.100	0.105
1 bis <2	9	1	0.111	0.889	0.900	0.100	0.118
2 bis <3	8	1	0.125	0.875	0.800	0.100	0.133
3 bis <4	7	1	0.143	0.857	0.700	0.100	0.155
4 bis <5	6	1	0.167	0.833	0.600	0.100	0.182
5 bis <6	5	1	0.200	0.800	0.500	0.100	0.222
6 bis <7	4	1	0.250	0.750	0.400	0.100	0.286
7 bis <8	3	1	0.333	0.667	0.300	0.100	0.400
8 bis <9	2	1	0.500	0.500	0.200	0.100	0.667
9 bis <10	1	1	1.000	0.000	0.100	0.100	2.000
(Ende letztes Intervall)					0.000		

Erläuterungen zu Darstellung 2: (vgl. auch Text)

(1): Intervall (Woche)
(2): n zu Beginn des Intervalls
(3): Ereignisse (Übergänge aus der Arbeitslosigkeit) während des Intervalls
(4): Bedingte „Sterbewahrscheinlichkeit“ für das Intervall
(5): Bedingte „Überlebenswahrscheinlichkeit“ p_i für das Intervall
(6): Survivorfunktion $S(t_i)$
(7): Dichtefunktion $f(t_i)$
(8): Hazardfunktion $r(t_i)$

Die Werte der vierten Spalte geben an, wie groß der *Anteil* der Personen mit einem Ereignis während dieses Intervalls ist; die fünfte Spalte zeigt den Komplementärwert hierzu, den Anteil der „Überlebenden“ (der Personen ohne Zustandswechsel, hier also der immer noch Arbeitslosen) am Ende des Intervalls. Beides gilt bezogen auf die zweite Spalte, also die Personen, die zu Beginn des Intervalls noch arbeitslos waren. So waren nach dem ersten Intervall, also am Ende der ersten Woche, neun von zehn, also 90 Prozent der Personen, die zu Beginn der Woche arbeitslos waren, immer noch arbeitslos, am Ende der zweiten Woche waren es acht von neun, also 88,9 Prozent. Anders formuliert: Die fünfte Spalte stellt nicht die Wahrscheinlichkeit schlechthin, statistisch gesprochen: die unbedingte Wahrscheinlich-

keit dar, das betreffende Intervall zu überleben (die Wahrscheinlichkeit bezogen auf alle Personen überhaupt), sondern die Überlebenswahrscheinlichkeit unter der Bedingung, daß man bis zum Beginn des Intervalls „überlebt" hatte. Man spricht daher von der *„bedingten Überlebenswahrscheinlichkeit"*, die ich mit p_i - Wahrscheinlichkeit p für das i-te Intervall - bezeichnen will.

In den Spalten 6 bis 8 sind nun die entscheidenden Größen eingetragen. Spalte 6 enthält die *Survivor-Funktion S(t)*. Diese gibt die Wahrscheinlichkeit an, daß ein Individuum bis zum Zeitpunkt t „überlebt", d.h., daß bis zum Zeitpunkt t (hier: dem Beginn des betreffenden Intervalls) das Individuum sich noch im Ausgangszustand befindet, anders formuliert: daß noch kein Zustandswechsel oder „Ereignis" stattgefunden hat.[19] Formal schreiben wir:

$$S(t) = P(T \geq t) \text{ .}^{20} \tag{1}$$

Zu Beginn des Prozesses befinden sich definitionsgemäß alle Personen im Ausgangszustand, daher hat S(t) zu Beginn des Prozesses, also $S(t_1)$,[21] immer einen Betrag von 1. In unserem Beispiel nun geht jede Woche genau eine Person aus der Arbeitslosigkeit ab, d.h. jeweils 10 Prozent der Ausgangsstichprobe. Die „Überlebenswahrscheinlichkeit" nimmt also pro Woche um 0,1 ab, anders gesagt, die Survivorfunktion verringert sich in jedem Zeitintervall um den Betrag von 0,1. Die Wahrscheinlichkeit, mindestens eine Woche arbeitslos zu bleiben, beträgt also 0,9 (oder 90 Prozent), die Wahrscheinlichkeit, mindestens zwei Wochen arbeitslos zu bleiben, beträgt 0,8 (oder 80 Prozent), usw.

Daß die Survivorfunktion sich hier unmittelbar aus den Daten ablesen läßt, liegt natürlich daran, daß bewußt ein einfaches Beispiel ohne Zensierungen gewählt wurde. Daher soll sogleich eine allgemeinere Möglichkeit betrachtet werden, S(t) zu berechnen (A: 142; BHM: 43 f.; DM: 65; etwas ausführlicher bei Kiefer 1988: 648 f.; Petersen 1991a: 274 f.). Die bedingte Überlebenswahrscheinlichkeit p_i für das erste Intervall, also p_1, beträgt 0,9, und dies ist offenkundig auch die Wahrscheinlichkeit, den Beginn des zweiten Intervalls zu „erleben", also $S(t_2)$. $S(t_3)$, die Wahrscheinlichkeit, auch das zweite Intervall zu „überleben" und mithin zu Beginn des dritten Intervalls immer noch arbeitslos zu sein, ergibt sich aus der Wahrscheinlichkeit, das erste Intervall *und* das zweite Intervall zu überleben, was nach den Gesetzen der Wahrscheinlichkeitsrechnung durch die Multiplikation ausgedrückt wird, also $p_1 \cdot p_2$ oder $S(t_2) \cdot p_2$. $S(t_4)$, die Wahrscheinlichkeit, daß eine Person mindestens 3 Wochen oder bis zum Beginn der vierten Woche arbeitslos bleibt, ergibt sich aus $p_1 \cdot p_2 \cdot p_3$ - die Person muß die erste Woche *und* die zweite Woche *und* die dritte Woche „überleben" - oder $S(t_3) \cdot p_3$. Allgemein ergibt sich S(t) also jeweils aus dem Wert von $S(t_{i-1})$ multipliziert mit p_{i-1}, formal:

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \cdot \hat{p}_{i-1} \text{ ,} \tag{2}$$

oder, da ja $S(t_{i-1})$ sich wiederum aus Survivorfunktion und bedingter Überlebenswahrscheinlichkeit des vor-vorherigen Intervalls berechnen läßt,

$$\hat{S}(t_i) = \prod_{j=1}^{i-1} \hat{p}_j \ . \tag{3}$$

Die nächste Größe (Spalte 7), die *Dichteverteilung der Verweildauern f(t)*, ist häufig nicht direkt von Interesse, sie wird jedoch benötigt, um den zentralen Begriff der Hazardfunktion verständlich zu machen, der nachfolgend erörtert wird. Die (diskrete) Dichteverteilung ist die (unbedingte) Wahrscheinlichkeit, daß in einem Intervall ein Zustandswechsel (Ereignis) eintritt. Formal läßt sich dies so schreiben:

$$f(t) = P(t \le T < t+1) \ . \tag{4}$$

Auch sie läßt sich in unserem einfachen Beispiel aus der Ausgangstabelle ablesen: In jeder Woche tritt genau ein Ereignis ein, was bezogen auf die Gesamtstichprobe einer Wahrscheinlichkeit von 0,1 entspricht.

Allgemein ergibt sich aber f(t) aus S(t) (bzw. umgekehrt). Dazu muß man sich nur vor Augen halten, daß S(t) für ein beliebiges Intervall nichts anderes ist als die Wahrscheinlichkeit, bis zum vorherigen Intervall überlebt zu haben, abzüglich der Wahrscheinlichkeit, in diesem Intervall zu „sterben“, also in den Zielzustand zu wechseln. Umgekehrt läßt sich also für f(t) - bezogen auf ein konkretes Zeitintervall - schreiben:

$$\hat{f}(t_i) = \hat{S}(t_i) - \hat{S}(t_{i+1}) \ .^{22} \tag{5}$$

Nur der Vollständigkeit halber sei noch kurz erwähnt, daß statt S(t), der Überlebenswahrscheinlichkeit, ebensogut deren Gegenteil betrachtet werden kann, die (kumulierte) „Sterbewahrscheinlichkeit“ F(t), also die Wahrscheinlichkeit, bis zum Intervall t_i *nicht* zu überleben (bis dahin in den Zielzustand zu wechseln). F(t) ist offensichtlich der Komplementärwert zu S(t), formal:

$$F(t) = 1 - S(t) = P(T \le t) \ . \tag{6}$$

F(t) zu Beginn eines Intervalls ist nichts anderes als die Summe der bis dahin aufgetretenen intervallspezifischen Wahrscheinlichkeiten f(t), es gilt also:

$$F(t_i) = \sum_{j=1}^{i-1} f(t_j) \ . \tag{7}$$

Zentral für die Verlaufsdatenanalyse ist die *Hazardfunktion r(t)* (Spalte 8 in Darstellung 2).[23] Folgende Überlegung kann diese verständlich machen: Wenn man betrachtet, wieviele Personen in einem Zeitintervall den Ausgangszustand verlassen, so sollte man auch berücksichtigen, wieviele Personen *bis dahin überhaupt noch im Ausgangszustand verblieben waren*. Alle anderen Personen sind ja gar nicht mehr arbeitslos und nicht mehr dem „Risiko" ausgesetzt, eine Beschäftigung zu finden oder die Arbeitslosigkeit anderweitig zu verlassen. Im Beispiel: In der ersten Woche verläßt ebenso eine Person die Arbeitslosigkeit wie in der achten oder neunten Woche; im ersten Fall handelt es sich jedoch um ein Zehntel der zu diesem Zeitpunkt noch Arbeitslosen, in den anderen beiden Fällen sind es ein Drittel bzw. die Hälfte! Offensichtlich ist also im ersteren Zeitraum die Chance, die Arbeitslosigkeit zu verlassen, wesentlich geringer als im letzteren Zeitraum - *bezogen auf die jeweils noch Arbeitslosen*. Abstrakter: Wir können die (unbedingte) Wahrscheinlichkeit, die Arbeitslosigkeit in einem bestimmten Intervall zu verlassen, auch beziehen auf die Wahrscheinlichkeit, bis dahin überhaupt arbeitslos geblieben zu sein, also auf die Survivorfunktion:

$$r(t) = P(t \le T < t+1 \mid T \ge t) = \frac{f(t)}{S(t)} \quad .^{24} \tag{8}$$

Die Größe r(t) ist die einzige, die sich in unserem Beispiel nicht unmittelbar aus den Daten ablesen läßt. Bei ihrer Schätzung ist zu bedenken, daß wir von wöchentlichen Intervallen ausgehen und S(t) sich immer auf den Anfang des Intervalls bezieht. Es wäre aber sinnvoll, die Wahrscheinlichkeit, die Arbeitslosigkeit zu verlassen, auf die Survivorfunktion „während" des Intervalls zu beziehen, wofür sich hier der Mittelwert von $S(t_i)$ und $S(t_{i+1})$ anbietet. Daher beträgt die Hazardfunktion für das erste Intervall nicht einfach 0,1 sondern 0,1 / ((1 + 0,9) · 0,5) = 0,1 / 0,95 = 0,105, und in der letzten Zeile von Darstellung 2 hat die Hazardfunktion sogar einen Betrag von 2, da $f(t_{10})$ durch den Durchschnitt von 0,1 (=$S(t_{10})$) und 0 (= $S(t_{Ende})$) dividiert wird.[25] Für den hier dargestellten Life-Table-Schätzer ergibt sich also:

$$\hat{r}(t_i) = \frac{\hat{f}(t_i)}{(\hat{S}(t_i) + \hat{S}(t_{i+1}))/2} \quad . \tag{9}$$

Die Hazardfunktion kann man sich vorstellen als die *momentane Neigung oder Tendenz zu einem Zustandswechsel*, und insofern ist sie aus inhaltlichen Gründen gut geeignet, die Dynamik des Prozesses auszudrücken. Insbesondere sei noch einmal darauf hingewiesen, daß die Hazardfunktion sich im Verlauf des untersuchten Prozesses ändern kann. In unserem fiktiven Beispiel nimmt sie kontinuierlich zu, was bedeutet, daß die Wahrscheinlichkeit, die Arbeitslosigkeit zu verlassen, von Woche zu Woche größer wird. Es sind aber auch ganz andere Verläufe

denkbar. Weiter unten wird gezeigt, daß die Hazardfunktion auch aus „formaler" Sicht wichtig ist, da sich alle hier dargestellten Größen auf sie zurückführen lassen. Die Tatsache, daß die Hazardfunktion im Grunde eine unanschauliche (und unbeobachtbare) Größe ist, sollte nicht weiter irritieren. Die gleiche Feststellung gilt auch für die „momentane Geschwindigkeit" eines bewegten Objekts, und doch lassen sich hieraus sehr sinnvolle Aussagen ableiten, z.B. über die Zeit, die benötigt wird, um eine bestimmte Entfernung zurückzulegen. Und auch wer in die Geheimnisse der Infinitesimalrechnung, anhand derer sich ein Konzept wie die „momentane Geschwindigkeit" verstehen ließe, nicht eingeweiht ist, kennt den grundlegenden Sachverhalt, daß eine höhere Geschwindigkeit impliziert, daß man schneller am Ziel ist. Genau das gleiche gilt aber auch für die Hazardrate.[26]

Nachdem nunmehr die zentralen Größen S(t), f(t) und r(t) eingeführt sind, sollen sie im folgenden noch einmal für den Fall rechtszensierter Beobachtungen dargestellt werden.

Wir greifen hier auf Daten aus dem Sozio-ökonomischen Panel zurück *(Darstellung 3)*. Es handelt sich um die jeweils erste (nicht linkszensierte) Arbeitslosigkeitsepisode der *männlichen* Erwerbspersonen aus der *deutschen* Teilstichprobe;[27] verwendet werden die Daten aus den ersten sieben Erhebungswellen. Die Untersuchungspersonen sind nach dem Alter zu Beginn der Arbeitslosigkeit in drei Gruppen eingeteilt: Bis zu 30 Jahre, 31 bis 50 Jahre und über 50 Jahre. Später werden wir darauf eingehen, wie mögliche Unterschiede zwischen diesen Gruppen analysiert werden können. Aus Platzgründen verzichte ich auf eine Wiedergabe der Rohdaten, die sich ja leicht „zurückrechnen" lassen. Außerdem werden für die drei Altersgruppen nur die Daten der ersten 12 Monate dargestellt, tatsächlich sind die beobachteten Arbeitslosigkeitsdauern teilweise länger. Es werden nur die Übergänge in eine Vollzeitbeschäftigung als Ereignis gewertet. Es handelt sich um 630 Episoden, von denen 418 in eine Vollzeitbeschäftigung münden. Auch hier haben wir gruppierte Daten, wobei wegen der großen Stichprobe in den meisten Zeitintervallen mehrere Übergänge in die Beschäftigung stattfinden.

In diesem Beispiel treten nun rechtszensierte Daten auf (vgl. Spalte (3) in Darstellung 3).[28] Erstens haben einige Personen die entsprechenden Daten nicht vollständig angegeben, außerdem sind einige Personen während des Untersuchungszeitraumes ausgeschieden (Teilnahmeverweigerung, Nicht-Anwesenheit zum Befragungszeitpunkt, Umzug ohne Angaben über neuen Wohnort), und ferner waren einige Personen am Ende der siebten Erhebungswelle arbeitslos.[29] Schließlich gibt es Personen, die aus der Arbeitslosigkeit in einen anderen Zielzustand als eine Vollzeitbeschäftigung übergehen. Auch diese werden bei einer Analyse, die sich auf die Vollzeitbeschäftigung konzentriert, wie rechtszensierte Beobachtungen behandelt.

Darstellung 3: *Life-Table-Schätzung zur Arbeitslosigkeitsdauer von Männern*

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Alter bis 30 Jahre									
0 bis < 1	355	22	46	344.0	0.134	0.866	1.000	0.134	0.143
1 bis < 2	287	18	50	278.0	0.180	0.820	0.866	0.156	0.198
2 bis < 3	219	12	49	213.0	0.230	0.770	0.710	0.163	0.260
3 bis < 4	158	6	22	155.0	0.142	0.858	0.547	0.078	0.153
4 bis < 5	130	10	21	125.0	0.168	0.832	0.469	0.079	0.183
5 bis < 6	99	11	10	93.5	0.107	0.893	0.391	0.042	0.113
6 bis < 7	78	4	13	76.0	0.171	0.829	0.349	0.060	0.187
7 bis < 8	61	5	7	58.5	0.120	0.880	0.289	0.035	0.127
8 bis < 9	49	2	4	48.0	0.083	0.917	0.255	0.021	0.087
9 bis <10	43	1	4	42.5	0.094	0.906	0.233	0.022	0.099
10 bis <11	38	3	2	36.5	0.055	0.945	0.211	0.012	0.056
11 bis <12	33	3	3	31.5	0.095	0.905	0.200	0.019	0.100
Alter 31 bis 50 Jahre									
0 bis < 1	191	5	25	188.5	0.133	0.867	1.000	0.133	0.142
1 bis < 2	161	1	18	160.5	0.112	0.888	0.867	0.097	0.119
2 bis < 3	142	4	34	140.0	0.243	0.757	0.770	0.187	0.276
3 bis < 4	104	2	17	103.0	0.165	0.835	0.583	0.096	0.180
4 bis < 5	85	1	15	84.5	0.178	0.822	0.487	0.086	0.195
5 bis < 6	69	3	8	67.5	0.119	0.881	0.400	0.047	0.126
6 bis < 7	58	1	7	57.5	0.122	0.878	0.353	0.043	0.130
7 bis < 8	50	4	5	48.0	0.104	0.896	0.310	0.032	0.110
8 bis < 9	41	0	5	41.0	0.122	0.878	0.278	0.034	0.130
9 bis <10	36	0	1	36.0	0.028	0.972	0.244	0.007	0.028
10 bis <11	35	0	4	35.0	0.114	0.886	0.237	0.027	0.121
11 bis <12	31	1	5	30.5	0.164	0.836	0.210	0.034	0.179
Alter über 50 Jahre									
0 bis < 1	84	2	0	83.0	0.000	1.000	1.000	0.000	0.000
1 bis < 2	82	1	3	81.5	0.037	0.963	1.000	0.037	0.038
2 bis < 3	78	0	4	78.0	0.051	0.949	0.963	0.049	0.053
3 bis < 4	74	2	2	73.0	0.027	0.973	0.914	0.025	0.028
4 bis < 5	70	2	3	69.0	0.043	0.957	0.889	0.039	0.044
5 bis < 6	65	2	1	64.0	0.016	0.984	0.850	0.013	0.016
6 bis < 7	62	1	1	61.5	0.016	0.984	0.837	0.014	0.016
7 bis < 8	60	1	0	59.5	0.000	1.000	0.823	0.000	0.000
8 bis < 9	59	5	0	56.5	0.000	1.000	0.823	0.000	0.000
9 bis <10	54	2	0	53.0	0.000	1.000	0.823	0.000	0.000
10 bis <11	52	2	0	51.0	0.000	1.000	0.823	0.000	0.000
11 bis <12	50	8	0	46.0	0.000	1.000	0.823	0.000	0.000

Erläuterungen zu Darstellung 3 (vgl. auch Text):

(1): Intervall
(2): n zu Beginn des Intervalls
(3): (Rechts-)Zensierungen c_i während des Intervalls
(4): Ereignisse d_i während des Intervalls
(5): Risikomenge n' für das Intervall
(6): Bedingte „Sterbewahrscheinlichkeit“ für das Intervall
(7): Bedingte „Überlebenswahrscheinlichkeit“ p_i für das Intervall
(8): Survivorfunktion $S(t_i)$
(9): Dichtefunktion $f(t_i)$
(10): Hazardfunktion $r(t_i)$

Quelle: Das Sozio-ökonomische Panel, eigene Datenaufbereitung (jeweils erste Arbeitslosigkeitsepisode aus dem „Erwerbskalender“, Welle 1 bis 7).

Beim Vorliegen von Rechtszensierungen können die entscheidenden Größen zur Charakterisierung des Prozesses nicht mehr, wie in unserem ersten Beispiel, direkt aus den beobachteten Verweildauern abgelesen werden. Die „Lösung“ des Problems basiert auf der schon oben dargestellten Überlegung, den Beobachtungszeitraum in - möglichst kleine - Intervalle zu zerlegen, für jedes dieser Intervalle die bedingte Überlebenswahrscheinlichkeit p_i und daraus die relevanten Größen zu berechnen. D.h., wir nehmen - wie sich ja schon bei der „Transformation“ von Darstellung 1 in Darstellung 2 gezeigt hat - nicht mehr direkt auf die Verweildauer einer jeden Untersuchungsperson Bezug, sondern fragen umgekehrt für jede Zeiteinheit (jedes Intervall), welchen Beitrag jede Person zur Definition der relevanten Größen, insbesondere der Risikomenge und damit der bedingten Überlebenswahrscheinlichkeit, leistet.[30]

Hierzu geht man von folgender Überlegung aus. p_i ist der Anteil der „überlebenden“ Personen im i-ten Intervall, und zwar bezogen auf die Zahl der Personen, die während dieses Intervalls überhaupt dem „Risiko“ ausgesetzt waren, die Arbeitslosigkeit zu verlassen oder allgemeiner: in den Zielzustand überzugehen. Zu dieser Risikomenge gehören *auch* die Personen mit rechtszensierten Beobachtungsdauern, solange die Zensierung noch nicht eingetreten ist. Wenn man nun die Annahme, daß die Zensierungen zufällig erfolgen, im konkreten Fall für gültig erachtet, dann sollte der Anteil p_i in den einzelnen Intervallen nicht davon beeinflußt werden, wieviele Dauern gerade in diesem Intervall zensiert wurden; es kommt nur darauf an, eine angemessene Formulierung für die Risikomenge zu finden. Dabei ist zu bedenken, daß die Zensierungen (ebenso wie die „Ereignisse“) vermutlich nicht erst am Ende des Intervalls stattfanden, sondern irgendwann während des jeweiligen Intervalls erfolgt sein können, daß also die zensierten Fälle nicht während des gesamten Intervalls dem „Risiko“ ausgesetzt waren, die Arbeitslosigkeit zu verlassen. Als „Risikomenge“ kann daher nicht einfach die Anzahl der Personen aufgeführt werden, die bis zu Beginn des Intervalls noch „überlebt“ haben (weder die Arbeitslosigkeit verlassen haben noch zensiert waren - vgl. Spalte 2). Üblicherweise wird man davon ausgehen, daß „im Durchschnitt“ die zensierten Fälle während

der Hälfte des Intervalls zur Risikomenge zählten; es wird also von der Zahl der Personen, die bis zu Beginn des Intervalls „überlebten“, die Hälfte der während des Intervalls durch Zensierung ausgeschiedenen Personen abgezogen (vgl. Spalte 2, 3 und 5 in Darst. 3).

Die „bedingte Überlebenswahrscheinlichkeit“ für das i-te Intervall, p_i, wird also im Falle von Rechtszensierungen folgendermaßen geschätzt:

$$\hat{p}_i = \frac{Anzahl\ der\ Überlebenden}{Umfang\ der\ Risikomenge} .$$

Die Risikomenge im i-ten Intervall (Spalte 5 in Darst. 3) - wir bezeichnen sie mit n'_i im Unterschied zur Gesamtzahl n_i der Personen zu Beginn des Intervalls - läßt sich nach den obigen Erläuterungen berechnen, indem man von n_i die Hälfte der Zensierungen des betreffenden Intervalls - als c_i bezeichnet - abzieht:

$$n'_i = n_i - c_i / 2 . \tag{10}$$

Die Anzahl der Überlebenden, also die Zahl der Personen, die *kein* Ereignis „erlitten“, wird ebenfalls auf die Risikomenge bezogen. Wenn wir mit d_i die Anzahl der „Ereignisse“ während des Intervalls bezeichnen (Spalte 4 in Darst. 3), so wird die Anzahl der Überlebenden durch (n'_i - d_i) geschätzt (und nicht durch n_i - d_i). Damit ergibt sich (vgl. Spalte 7 in Darst. 3):

$$\hat{p}_i = \frac{n'_i - d_i}{n'_i} . \tag{11}$$

Mit dieser Größe lassen sich nun die Survivor-, Dichte- und Hazardfunktion, die in den Spalten 8 bis 10 von Darstellung 3 wiedergegeben sind, nach den Formeln (3), (5) und (9) berechnen. Die *wesentlichen Ergebnisse für die Beispielsdaten* sollen hier kurz zusammengefaßt werden:

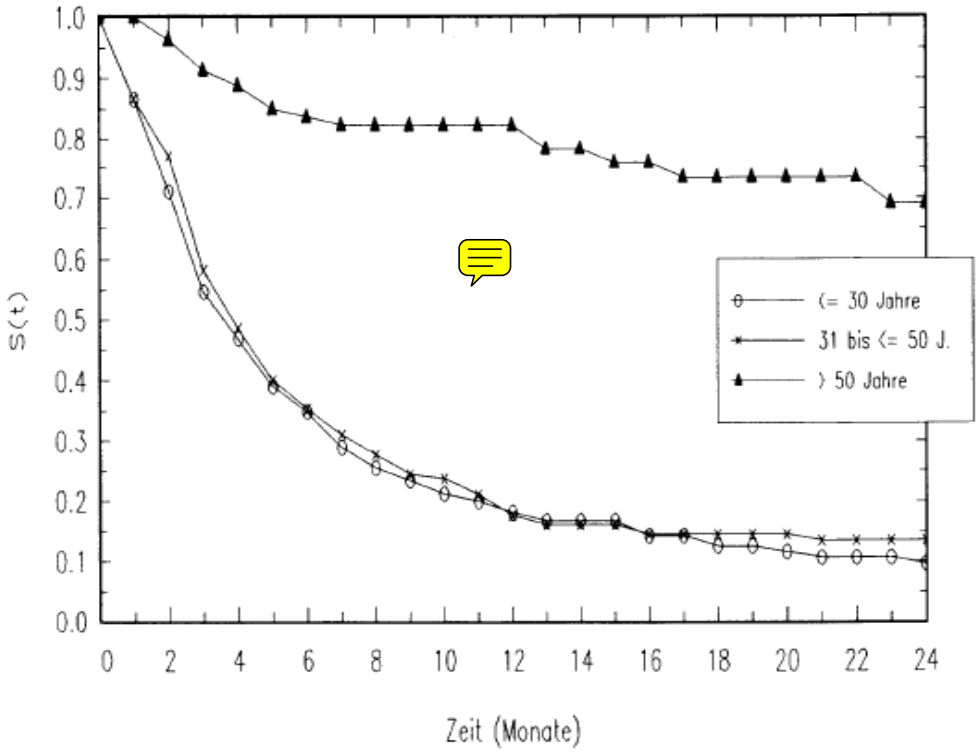
1. Am leichtesten verständlich sind sicherlich die Aussagen, die sich auf die Survivorfunktion beziehen. Wir können z.B. feststellen, daß nach 6 Monaten (also zu Beginn des 7. Intervalls!) *in den beiden jüngeren Altersgruppen* nur mehr ca. 40 Prozent in der Arbeitslosigkeit verblieben sind, es haben bis dahin also bereits 60 Prozent aller Übergänge in die Vollzeitbeschäftigung stattgefunden. Nach 9 Monaten beträgt der Anteil der noch arbeitslosen Personen in diesen beiden Gruppen nur mehr knapp 25 Prozent, über drei Viertel der Abgänge in eine Vollzeitbeschäftigung sind also bis zu diesem Zeitpunkt erfolgt. Die Unterschiede zwischen den beiden jüngeren Altersgruppen sind nur ganz geringfügig; die Survivorfunktion der 31- bis 50jährigen liegt ganz dicht bei oder nur wenig über der der jüngsten Altersgruppe. - Sehr große Unterschiede im Vergleich hierzu sind bei den *über 50jährigen* zu beobachten, bei denen nach 6 wie nach 9 Monaten über 80 Prozent noch keinen Übergang in eine Vollzeitbeschäftigung vollzogen haben.

2. Die Hazardfunktion gibt Auskunft über die „Risiken“ (inhaltlich natürlich: Chancen), in den einzelnen Monaten eine Vollzeitbeschäftigung aufzunehmen. Die wesentlich niedrigeren Verbleibsquoten in der Arbeitslosigkeit in den beiden jüngeren Altersgruppen schlagen sich in dementsprechend höheren Werten in der Hazardfunktion nieder. Von Interesse ist auch der zeitliche Verlauf der Hazardfunktion, die offenbar zunächst ansteigt und dann wieder abnimmt.

Besonders eingängig lassen sich die Ergebnisse aber *graphisch* darstellen, und ich will ihre Kommentierung anhand der folgenden Abbildungen noch einmal aufnehmen.[31]

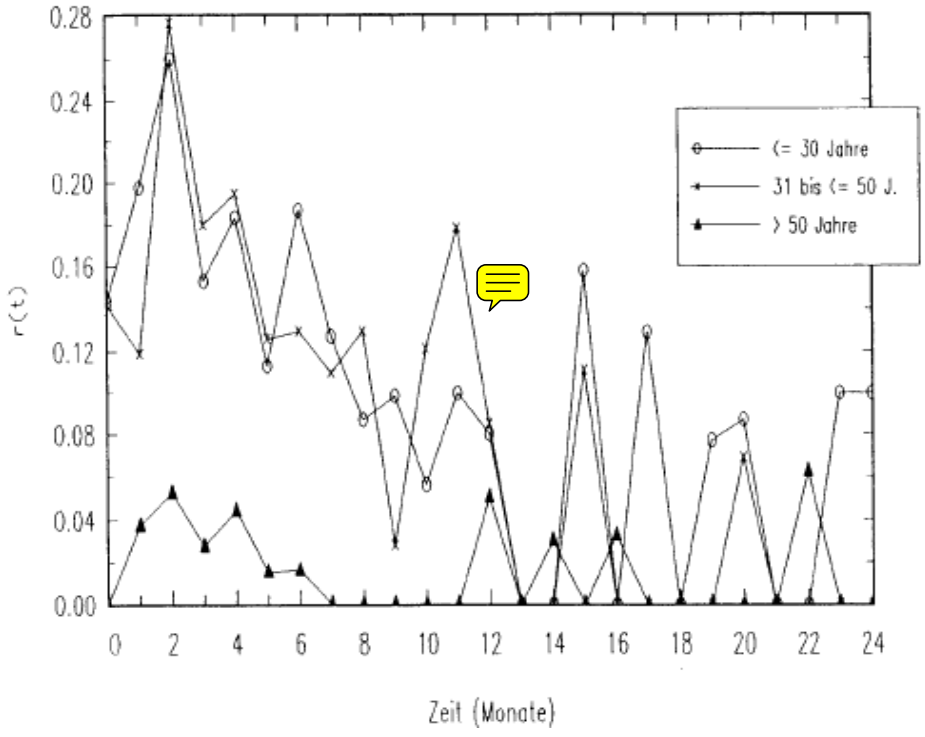
Die Survivorfunktion *(Darstellung 4)* verdeutlicht sehr plastisch, ob bzw. welche Unterschiede zwischen den Gruppen vorhanden sind; im vorliegenden Beispiel läßt sich weitaus schneller als durch Inspektion der Ausgangstabelle erkennen, daß die jüngste und die mittlere Altersgruppe sich in ihren Beschäftigungschancen kaum unterscheiden, während die Gruppe der über 50jährigen schlechtere Beschäftigungschancen aufweist.

Darstellung 4: *Life-Table-Schätzer der Survivor-Funktionen für die Beispielsdaten aus Darstellung 3 (Arbeitslosigkeit, Männer)*



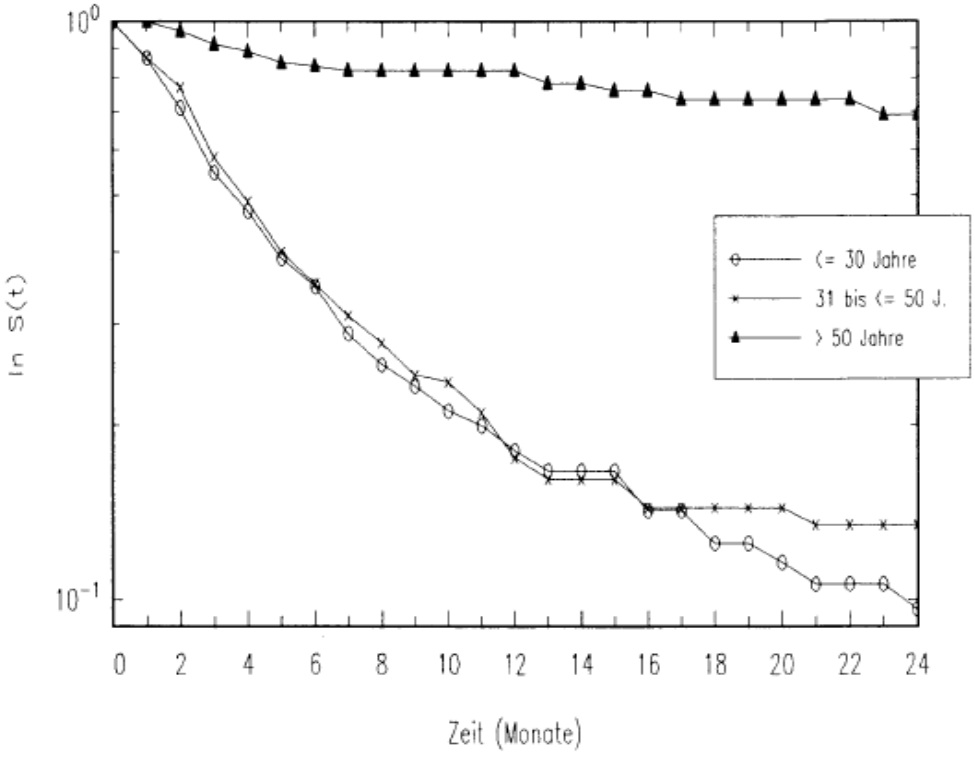
Ganz allgemein kann man eine Survivorfunktion wie eine Kreuztabelle lesen, die zeigt, welcher Anteil der Stichprobe bzw. der einzelnen Gruppen zu einem Zeitpunkt die Arbeitslosigkeit bereits verlassen hat und welcher Anteil noch nicht (wobei ersterer Wert - es handelt sich dabei um die oben erwähnte kumulierte „Sterbewahrscheinlichkeit“! - üblicherweise nicht ausgegeben wird, da er sich aus der Differenz von 1 und S(t) ergibt) - dies aber eben nicht nur für einen einzelnen Zeitpunkt, sondern für viele Zeitpunkte bzw. Zeitintervalle. Darüber hinaus lassen sich auch sehr einfach - wenn auch nicht ganz genau - Angaben zu den Verweildauern ablesen. So wird man sich häufig für die mittlere Verweildauer interessieren. Hierzu verwendet man üblicherweise nicht den Mittelwert, da dieser oft durch einige wenige lange Verweildauern beeinflußt wird, sondern den Median, also den Zeitpunkt, zu dem genau die Hälfte der Untersuchungspersonen ein Ereignis hatte, anders gesagt: zu dem S(t) den Wert von 0,5 hat. Dieser liegt in den beiden jüngeren Altersgruppen etwa bei 4 Monaten.[32] Für die älteste Gruppe läßt sich der Median allerdings gar nicht schätzen, weil auch zum Ende des Beobachtungszeitraumes gerade erst 30 Prozent der Männer über 50 Jahre die Arbeitslosigkeit verlassen haben. Grundsätzlich läßt sich aber nicht nur der Median, sondern der Wert eines jeden beliebigen Perzentils zum Vergleich heranziehen.

Darstellung 5: *Life-Table-Schätzer der Hazard-Funktionen für die Beispielsdaten aus Darstellung 3 (Arbeitslosigkeit, Männer)*



Die Hazardfunktion wird in *Darstellung 5* gezeigt. Ersichtlich ist diese wegen der häufigen starken Sprünge etwas verwirrender als die Survivorfunktion.[33] Die wesentlich besseren Chancen der beiden jüngeren Gruppen im Vergleich zu den über 50jährigen sind aber auch hier sehr gut zu erkennen. Ebenfalls deutlich wird, daß die Werte in den ersten 8 bis 10 Monaten höher liegen als im weiteren Verlauf; zumindest läßt sich dies für die beiden jüngeren Gruppen sagen, während bei der älteren Gruppe die Werte von Anfang an so niedrig liegen, daß ein Abfallen kaum möglich ist.[34] Auffällig ist auch, daß die höchsten Werte in allen drei Gruppen im dritten Intervall liegen, d.h., bis dahin steigen die Werte der Hazardfunktion sogar an, um dann allmählich zurückzugehen. Mit anderen Worten: Im dritten Monat der Arbeitslosigkeit sind die (relativen) Chancen, eine Beschäftigung aufzunehmen, am höchsten. Alles in allem zeigt sich jedenfalls an dieser Darstellung sehr gut, daß die Hazardfunktion - vorbehaltlich der Einbeziehung weiterer erklärender Variablen - *zeitabhängig*, also nicht während der gesamten Dauer des Prozesses gleich ist.[35]

Darstellung 6: *Logarithmierte Survivor-Funktionen (Life-Table-Schätzer) für die Beispielsdaten aus Darstellung 3 (Arbeitslosigkeit, Männer)*



Um die Zeitabhängigkeit des Prozesses zu untersuchen, wird häufig auch vorgeschlagen, die Survivorfunktion auf einer logarithmischen Skala abzutragen (vgl. *Darstellung 6*). Besteht keine Zeitabhängigkeit, muß sich eine Gerade ergeben, ansonsten werden die Kurven steiler oder flacher, je nachdem ob die Hazardfunktion zu- oder abnimmt. Im vorliegenden Beispiel wird vor allem die Abnahme der Hazardfunktion etwa ab dem achten Monat deutlich, bei genauem Hinsehen läßt sich auch erkennen, daß der Plot im dritten Monatsintervall besonders steil ist.

Zum Abschluß dieses Teils möchte ich noch einmal darauf eingehen, daß die bislang verwendeten Formeln auf einer Betrachtung basieren, die die eigentlich stetige oder kontinuierliche Zeit in Intervalle zerlegt. Man kann sich aber leicht vorstellen, daß man sich einer kontinuierlichen Betrachtung annähert, indem man die Zeitintervalle möglichst klein werden, also tendenziell gegen Null gehen läßt. Damit ergeben sich folgende Definitionen für Dichte- und Hazardfunktion bei stetiger Zeit, die ich hier deshalb anführe, weil sie in den Lehrbüchern oder Übersichtsartikeln manchmal im Vordergrund stehen:

$$f(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} \text{ ,}^{36} \tag{12}$$

und entsprechend

$$r(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \text{ .} \tag{13}$$

Diese „abstrakten“ Größen können natürlich nicht mehr unmittelbar aus den Daten abgelesen oder einfach berechnet werden; hierzu müssen komplexere Schätzverfahren angewandt werden (vgl. Teil II). An dieser Stelle sei nur kurz gezeigt, daß sich die Größen f(t) und S(t) auch im stetigen Fall aus der Hazardfunktion bzw. Übergangsrate r(t) berechnen lassen, was ein wichtiger Grund dafür ist, daß in den komplexeren Schätzverfahren r(t) im Mittelpunkt steht.

Zunächst ergibt sich aus der Beziehung r(t) = f(t) / S(t) die umgekehrte Beziehung f(t) = r(t) · S(t). Es muß also nur gezeigt werden, daß S(t) sich aus r(t) berechnen läßt. Der Beweis hierfür übersteigt den Charakter einer Einführung (vgl. etwa DM: 195), es läßt sich aber zeigen, daß gilt:

$$S(t) = \exp\left[-\int_0^t r(u)du\right] . \tag{14}$$

Der Ausdruck ∫r(u)du ist das Integral der Hazardfunktion r(t) von 0 bis zum Zeitpunkt t und wird auch als *kumulierte Hazardfunktion* bezeichnet. Intuitiv ver-

ständlich machen läßt sich diese Größe an dem obigen Beispiel mit gruppierten Verweildauern, wenn man daran denkt, daß ein Integral durch eine Summierung - hier also: vom Beginn des Prozesses bis zum Zeitpunkt t - angenähert werden kann. Tatsächlich läßt sich auch in den Beispieldaten zeigen, daß sich S(t) zu jedem Zeitintervall aus der Summe aller bis dahin „angefallenen“ r(t) nach folgender Gleichung ergibt:[37]

$$S(t_i) = \exp\left[-\sum_{j=0}^{i-1} r(t_j)\right]. \qquad (15)$$

Anhang

Statistik-Software zur Verlaufsdatenanalyse

Die hier vorgestellten einfachen Life-Table-Schätzer können mit allen verbreiteten Statistikprogrammen (SPSS, BMDP, SAS, SYSTAT) berechnet werden. Für komplexere Verfahren reichen deren Möglichkeiten (mit Ausnahme von SAS) im allgemeinen nicht aus. Daher sei schon hier auf das Programm TDA (Transition Data Analysis) hingewiesen, welches von Götz Rohwer entwickelt wurde und das umfassendste Programm zur Verlaufsdatenanalyse darstellt, das derzeit verfügbar ist (erhältlich gegen Einsendung von 4 Disketten - 1,44 MB, MS-DOS-formatiert - und eines frankierten Rückumschlags; Anschrift: Universität Bremen, Fachbereich 8, Postfach 330 440, 28334 Bremen). Weitere Hinweise auf Spezialsoftware enthält Teil II.

Anmerkungen

1 Ich danke Günter Albrecht und Götz Rohwer für die kritische Durchsicht des Manuskripts. Für die noch vorhandenen Fehler bin ich natürlich selbst verantwortlich.

2 Vgl. für Einkommensarmut bzw. allgemein Einkommenslagen die Analysen von Bonß/Plum (1990) und Rohwer (1991, 1992); für Sozialhilfe Leisering/Zwick (1990) und Voges/Rohwer (1991); für Arbeitslosigkeit Ludwig-Mayerhofer (1990, 1992).

3 Vgl. beispielsweise zu „Patientenkarrieren“ Gerhardt (1986) und Keupp (1987), zu „Arbeitslosigkeitskarrieren“ (deren Häufigkeit vielfach überschätzt wird) Andreß (1989) und Wagner (1990), für den Bereich der Prostitution Hess (1978). Siehe auch, auf der Grundlage qualitativer Forschung, allerdings (vielleicht etwas zu) skeptisch gegenüber möglichen „karrierehaften“ Verfestigungsprozessen von Abweichung, Mutz/Kühnlein (1992).

4 Eine Warnung vor zu großer Euphorie hinsichtlich der hier vorgestellten Modelle formuliert Esser (1987). - Ich will im übrigen keinesfalls bestreiten, daß sich „Prozesse“, „Karrieren“ usw. auch mit Methoden der „qualitativen“ Forschung untersuchen lassen. Allerdings stellt sich auch hier die Frage, ob die häufig zu findende Vorstellung einer grundsätzlichen Gegensätzlichkeit von qualitativer und quantitativer Forschung nicht jedenfalls teilweise auf Mißverständnissen beruht (vgl. etwa Diekmann 1987 und Ostner 1987).

5 Vgl. die gute Darstellung bei Andreß (1984, S. 250 ff.) Das muß wohl „Die ersten 10 Berufsjahre“ sein! sowie das eindrucksvolle Beispiel bei Schneider (1991, S. 229 ff.). Dem steht nicht entgegen, daß in der Literatur gelegentlich *modifizierte* Verfahren der linearen Regression vorgeschlagen werden, die allerdings kaum Verbreitung gefunden haben.

6 Daher die Bezeichnungen „Analysis of Transition Data“ (Lancaster) oder „Event History Analysis“ (Tuma) bzw. „Ereignisanalyse“ (BHM).

7 Wenn die Annahme diskreter „Zustände“ nicht gerechtfertigt ist, also ein kontinuierlicher Zustandsraum angenommen wird, kann die Modellierung mit stochastischen Differentialgleichungen erfolgen, vgl. Blossfeld/Hannan/Schömann (1988).

8 Ein gravierendes Problem ist hier (wie anderswo) der uneinheitliche Sprachgebrauch. So verwendet Andreß (1992, S. 40) den Begriff „aggregiert“ noch ein einem anderen Sinn als ich es hier - im Anschluß an Petersen (1991b) bzw. Petersen/Koput (1992) - getan habe; Hamerle/Tutz (1989) gebrauchen den Begriff „diskret“ dagegen im Sinne von „gruppiert“ bzw. „aggregiert“. Dementsprechend wird in der Literatur manchmal zwischen den verschiedenen Fällen nicht hinreichend unterschieden.

9 Außerdem ist zu beachten: Hujer/Schneider (1986) schlagen vor, zensierte Zeiten sogar um eine ganze Zeiteinheit zu kürzen, während die Simulationsstudie von Petersen/Koput (1992) davon ausging, daß Zensierungen exakt gemessen wurden - eine nicht ganz realistische Annahme.

10 Einer der wichtigsten Zensierungsmechanismen ist die faktisch beschränkte Beobachtungsdauer von Untersuchungen: Alle Individuen werden über den gleichen Zeitraum beobachtet, einige haben aber am Ende den Zielzustand noch nicht erreicht. Ebenso können aber aus den unterschiedlichsten Gründen - z.B. wegen Panelmortalität, aber auch wegen des Studiendesigns - verschieden lange Beobachtungsdauern vorliegen. Eine ausführliche Diskussion findet sich etwa bei Lawless (1982, S. 31 ff.).

11 Das von DM: 91 vorgeschlagene Verfahren, Zensierungen als Ereignisse zu betrachten und so die Zensierungsmuster verschiedener Gruppen zu vergleichen, ist nur eine ad-hoc-Lösung. Weder ist die Gleichheit der Zensierungsmuster in den Gruppen eine Garantie der Unabhängigkeit von Zensierungen und Ereignissen, noch müssen Unterschiede zwischen den Gruppen notwendig auf eine Abhängigkeit hindeuten. Man kommt also ohne theoretische Überlegungen oder zusätzliche Untersuchungen nicht aus.

12 Allerdings wäre es auch unabhängig von diesem Problem zumeist nicht sinnvoll, beim Arbeitslosenbestand eines bestimmten Zeitpunktes anzusetzen, da in diesem jeweils die längerfristig Arbeitslosen überrepräsentiert sind. Hier wie auch sonst kann je nach Fragestellung eine Zugangsstichprobe sinnvoller sein, die jedoch auch mit Problemen behaftet ist (vgl. Diekmann/Mitter 1990, S. 432; Diekmann/Mitter 1993, S. 53).

13 Deutschen spricht man dagegen von „Sterbetafel“. Diese Ausdrücke verweisen ebenso wie der später eingeführte Begriff der „Survivor-Funktion“ darauf, daß zentrale Impulse zur Entwicklung der hier vorgestellten Verfahren aus dem Bereich der Lebensversicherung stammen.

14 Ähnlich elementare Einführungen finden sich bei Fleiss/Dunner/Stallone/Fieve (1976) sowie bei Peto et al. (1977, Abschnitt 18 ff.).

15 Mit Tukey (1977) halte ich die getrennte Numerierung von Tabellen und Abbildungen für unnötig (und gelegentlich eine Quelle von Konfusion), fasse beides unter dem Oberbegriff „Darstellung“ (Tukey: „exhibit“) zusammen und numeriere durchgängig.

16 In den statistischen Verfahren geht es natürlich um abstrakte Zeiteinheiten; sie „wissen“ nicht, ob sich die Verweildauern wie hier auf Wochen oder aber auf Sekunden, Tage, Jahre oder beliebige andere Zeiteinheiten beziehen. In manchen Programmen - etwa BMDP - läßt sich die verwendete Zeiteinheit explizit angeben, dies ist jedoch nur für die Beschriftung von Tabellen oder Graphiken von Bedeutung.

17 Für die Auswertung müssen die Daten allerdings anders kodiert werden, da die Statistikprogramme eine Dauer von 1 so interpretieren würden, daß die Dauer eine ganze Zeiteinheit betrug,

konkret also, daß die Arbeitslosigkeit *nach einer Woche, d.h. genau zum Beginn der zweiten Woche* verlassen wurde. Für das hier vorgestellte Verfahren muß bei einer Beendigung während des ersten Zeitintervalls ein Wert aus dem Bereich $0 \leq T < 1$ gewählt werden (analog für die folgenden Intervalle). Man beachte also, daß eine Dauer von genau 0 von den meisten Programmen für einfache (non-parametrische) Auswertungen zugelassen und so interpretiert wird, daß das Ereignis zwischen dem Zeitpunkt 0 und dem nächsten Zeitpunkt liegt. Bei komplexeren parametrischen Verfahren (siehe Teil II), wo Angaben zur Verweildauer als exakte Messungen aufgefaßt werden, wäre eine Kodierung mit 0 allerdings nicht sinnvoll.

18 Faktisch fungieren natürlich nach wie vor die Untersuchungspersonen als Analyseeinheit - schließlich sind die Daten in Darstellung 2 nur eine andere Form des Arrangements der Daten aus Darstellung 1.

19 Genauer gesagt handelt es sich um eine Schätzung (hier: der Überlebenswahrscheinlichkeit), insoweit davon ausgegangen wird, daß es sich bei den Untersuchungseinheiten praktisch immer um eine Stichprobe und nicht um die Grundgesamtheit handelt. Dies gilt auch für die anderen Funktionen, die im folgenden erläutert werden.

20 Der Ausdruck ist zu lesen: S(t) ist die Wahrscheinlichkeit, daß die konkrete Verweildauer einer Untersuchungseinheit, T, größer oder gleich einer bestimmten (beliebigen) Zeit t ist. - Manche Autoren definieren S(t) als $P(T > t)$ (Lee 1980, S. 10; Heckman/Singer 1986, S. 1693). Die substantiellen Unterschiede beider Definitionen sind unerheblich. Es ist nur genau darauf zu achten, wie die einzelnen Programme S(t) ausgeben; beispielsweise wird beim Life-Table-Schätzer von SPSS die Survivorfunktion immer für das *Ende* des jeweiligen Intervalls dargestellt. Grundsätzlich ist darauf hinzuweisen, daß die Literatur hinsichtlich der Definition der relevanten Größen, insbesondere des Zeitbezugs, etwas uneinheitlich ist, wodurch sich teilweise unterschiedliche Formeln ergeben.

21 Die tiefergestellten Indices sollen anzeigen, daß es sich jeweils um die Survivorfunktion - oder andere Größen - für ein bestimmtes Intervall handelt. Streng genommen müßte S(t) zu Beginn des Prozesses als $S(t_0)$ bezeichnet werden, da das erste Intervall eben zum Zeitpunkt 0 beginnt und S(t) sich immer auf den Beginn des Intervalls bezieht. Da wir jedoch vom ersten, zweiten Intervall usw. sprechen, würde es Verwirrung stiften, wenn der Zeitpunkt von S(t) immer von der Ordnungszahl des Intervalls abweichen würde.

22 Anzumerken ist, daß die Formulierungen für f(t) und im folgenden für r(t) eine Intervallbreite von 1 unterstellen. Die häufig vorzufindenden komplizierteren Formeln entstehen dadurch, daß man beliebige Intervallbreiten definieren kann und dies entsprechend berücksichtigen muß (besonders mißlich ist dies bei DM: 65 ff., da sie die Intervallbreite h_i zwar in ihren Formeln verwenden, aber nirgends einführen). Im Beispiel: Da eine Woche sieben Tage enthält, läßt sich auch eine auf Tage bezogene Dichtefunktion berechnen, indem man die „wöchentliche" Dichte von 0,1 unseres Beispiels durch 7 dividiert.

23 Ich gebrauche den Begriff Hazardfunktion als Oberbegriff. Im Fall stetiger Zeiten spricht man im allgemein von Hazard*rate*, aber manche Autoren (bzw. Statistikprogramme) gebrauchen diesen Begriff auch bei diskreten oder gruppierten Zeiten. Ferner wird häufig der Begriff „Übergangsrate" gebraucht. Vielfach wird dieser Begriff synonym zum Begriff Hazardrate verwendet (z.B. BHM: 31), andere gebrauchen den Begriff *Übergangs*rate für die Wahrscheinlichkeit eines je spezifischen Übergangs (z.B. Arbeitslosigkeit - Beschäftigung) und den Begriff *Hazard*rate für die Wahrscheinlichkeit, daß irgendein beliebiger Übergang (von möglicherweise mehreren, unterschiedlichen Übergängen) stattfindet (z.B. DM: 51).

24 Der Ausdruck $P(t \leq T < t+1 | T \geq t)$ ist zu lesen als „Wahrscheinlichkeit, daß der Zustandswechsel zwischen t und t+1 liegt (daß also die Verweildauer T eines Individuums in den Zeitraum zwischen t und t+1 fällt), gegeben, daß bis zum Zeitpunkt t noch kein Zustandswechsel stattfand".

25 Daher wird auch häufig betont, daß die Hazardfunktion keine echte Wahrscheinlichkeit ist, denn Wahrscheinlichkeiten können nur Werte zwischen 0 und 1 annehmen. Das unterscheidet sie von der „bedingten Sterbewahrscheinlichkeit“, obwohl natürlich die Ähnlichkeit mit dieser recht groß ist.

26 Vgl. ferner zu diesem Aspekt Petersen (1991a, S. 295). Tatsächlich läßt sich aus den hier dargestellten Größen teilweise die geschätzte durchschnittliche Dauer bis zum Zustandswechsel ableiten, doch ist dies häufig gerade dann nicht möglich, wenn die Hazardrate im Zeitverlauf variiert (vgl. DM: 47).

27 Genauer: Aus allen Haushalten der deutschen Teilstichprobe (sog. „Stichprobe A“) wurden diejenigen Personen ausgewählt, die über alle 7 Wellen eine deutsche Staatsangehörigkeit hatten. - Es wurden auch die Daten zu den weiblichen Arbeitslosen ausgewertet, die Männer sind aber im konkreten Fall „interessanter“ - unter rein *didaktischen* Gesichtspunkten. Eine *inhaltliche* Untersuchung zur Arbeitslosigkeit, bei der es u.a. ganz zentral um den Vergleich von Frauen und Männern geht, befindet sich in Vorbereitung.

28 Im Datensatz befinden sich also 22 Personen im Alter bis 30 Jahre, die nur einen Monat beobachtet werden konnten und am Ende der Beobachtungsdauer noch arbeitslos waren, 46 Personen dieser Altersgruppe mit einer Beobachtungsdauer von 1 Monat, die während dieser Zeit eine Vollzeitbeschäftigung antraten, usw. Die Daten müssen also auf jeden Fall eine dichotome Variable als Zensierungsindikator enthalten, der angibt, ob eine Beobachtungsdauer mit einem Zustandswechsel endete oder nicht. (Auch bei den Daten in Darstellung 1 ist zur Auswertung eine solche Variable notwendig, sie kann jedoch als Konstante jederzeit beliebig erzeugt werden). Bei Mehrzustandsmodellen tritt an die Stelle des dichotomen Zensierungsindikators eine Variable mit mehreren Ausprägungen.

29 Natürlich waren diese nicht sieben Jahre lang arbeitslos. Die meisten der am Ende des siebten Jahres arbeitslos Gebliebenen waren erst während dieses Jahres arbeitslos geworden.

30 Dieser Gedankengang findet sich am deutlichsten bei Lawless (1982, S. 52 ff.), sowie, etwas knapper, bei Petersen (1990, S. 263 ff.; 1991a, S. 274 f.).

31 Die nachfolgend dargestellten Plots werden auch von den einschlägigen Statistikprogrammen ausgegeben. Im Gegensatz zu Darstellung 3 werden die ersten 24 Monate wiedergegeben.

32 Formeln zu exakten Schätzung des Medians finden sich z.B. bei Andreß (1992, S. 143). Im konkreten Fall beträgt die Schätzung 3,61 Monate für die bis 30jährigen und 3,86 für die über 30- bis 50jährigen.

33 Im Programm TDA sind Verfahren zur „Glättung“ der hier gezeigten Funktionen implementiert, wodurch eine übersichtlichere Darstellung möglich wird. Mir geht es hier aber gerade um die Verdeutlichung, wie die ursprünglichen, nicht geglätteten Funktionen aussehen. Die Glättung mag in einem Fall wie dem vorliegenden, wo die Dauern gar nicht exakt gemessen wurden, vielleicht auch die Datenqualität überstrapazieren.

34 Bei genauem Hinsehen läßt sich aber sagen, daß die Werte für die älteste Gruppe in den ersten 6 Monaten - abgesehen vom allerersten Monat! - stets über Null liegen, während danach immer wieder längere oder kürzere Intervalle mit einer Hazardfunktion von Null beobachtet werden, so daß man auch hier davon sprechen kann, daß im Durchschnitt die Hazardfunktion anfangs höhere Werte annimmt.

35 Nur kurz angemerkt sei, daß sich die hier beobachtete Form der Hazardfunktion - Anstieg bis zum dritten Monat und nachfolgender Rückgang - bei *allen* explorativen Analysen der männlichen Arbeitslosen zeigt, welche Variable auch immer man heranziehen mag. Das kann als Indiz dafür gelten, daß es sich hierbei nicht um ein Artefakt handelt (vgl. Ludwig-Mayerhofer 1992 und Klein 1992).

36 Der senkrechte Pfeil im Ausdruck „$\Delta t \downarrow 0$“ soll bedeuten, daß Δt sich dem Grenzwert 0 von oben, also aus dem positiven Wertebereich nähert.

37 Da $S(t_i)$ immer zu Beginn eines Intervalls definiert wurde, dürfen natürlich nur die $r(t_i)$ bis zum Beginn des i-ten Intervalls herangezogen werden. Bei dieser Überlegung ergibt sich ebenfalls $S(t_1) = 1$, da ja die kumulierte Hazardrate zu Beginn des Prozesses 0 ist und exp(-0) = 1. - Wegen Rundungen, und weil Formel (15) nur eine Näherung darstellt, ergeben sich im Beispiel beim Nachrechnen kleine Abweichungen.

Literatur

Allison, P. D., 1984: Event History Analysis. Regression for Longitudinal Event Data. Beverly Hills: Sage.

Andreß, H.-J., 1984: Die ersten zehn Berufsjahre. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (BeitrAB 87).

Andreß, H.-J., 1989: Instabile Erwerbskarrieren und Mehrfacharbeitslosigkeit - ein Vergleich mit der Problemgruppe der Langzeitarbeitslosen. MittAB 22: 17-32.

Andreß, H.-J., 1992: Verlaufsdatenanalyse (Historical Social Research/Historische Sozialforschung, Supplement/Beiheft No. 5). Köln: Zentrum für Historische Sozialforschung.

Arminger, G., 1984: Modelltheoretische und methodische Probleme bei der Analyse von Paneldaten mit qualitativen Variablen. Vierteljahreshefte zur Wirtschaftsforschung: 470-479.

Arminger, G., 1988: Modelle zur Analyse qualitativer Variablen in stetigem Zeitverlauf. S. 77-91 in: Meier, F. (Hrsg.), Prozeßforschung in den Sozialwissenschaften. Anwendungen zeitreihenanalytischer Methoden. Stuttgart, New York: G. Fischer.

Blossfeld, H.-P./Hamerle, A./Mayer, K. U., 1986: Ereignisanalyse. Frankfurt/New York: Campus.

Blossfeld, H.-P./Hamerle, A., 1989: Using Cox Models to Study Multiepisode Processes. Sociological Methods & Research 17: 432-448.

Bonß, W./Plum, W., 1990: Gesellschaftliche Differenzierung und sozialpolitische Normalitätsfiktion. Zeitschrift für Sozialreform 36: 692-715.

Breslow, N., 1991: Use of the Logistic and Related Models in Longitudinal Studies of Chronic Disease Risk. S. 163-197 in: Dwyer, J. H./Feinleib, M./Lippert, P./Hoffmeister, H. (Hrsg.), Statistical Models for Longitudinal Studies of Health. (Monographs in Epidemiology and Biostatistics, Vol. 16). New York, Oxford: Oxford University Press.

Carroll, G. R., 1983: Dynamic Analysis of Discrete Dependent Variables: A Didactic Essay. Quality & Quantity 17: 425-460.

Crouchley, R. (Hrsg.), 1987: Longitudinal Data Analysis. Aldershot: Avebury.

Diekmann, A., 1987: Lebensverläufe und Verlaufsdatenanalyse - Statistische Auswertungsmethoden von Ereignisdaten. S. 171-196 in: Voges, W. (Hrsg.), Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske + Budrich.

Diekmann, A., 1988: Ereignisdatenanalyse - Beispiele, Probleme und Perspektiven. ZUMA- Nachrichten 23: 7-25.

Diekmann, A./Mitter, P., 1984a: Methoden zur Analyse von Zeitverläufen. Stuttgart: Teubner.

Diekmann, A./Mitter, P., 1984b: Stochastic Modelling of Social Processes. Orlando: Academic Press.

Diekmann, A./Mitter, P., 1990: Stand und Probleme der Ereignisanalyse. S. 404-441 in: Mayer, K. U. (Hrsg.), Lebensverläufe und sozialer Wandel. (Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 31). Opladen: Westdeutscher Verlag.

Diekmann, A./Mitter, P., 1993: Methoden der Ereignisanalyse in der Bevölkerungssoziologie: Stand und Probleme. S. 20-65 in: Diekmann, A./Weick, S. (Hrsg.), Der Familienzyklus als sozialer Prozeß. Bevölkerungssoziologische Untersuchungen mit den Methoden der Ereignisanalyse. (Sozialwissenschaftliche Schriften, Heft 26). Berlin: Duncker & Humblot.

Esser, H., 1987: Warum die Routine nicht weiterhilft. Überlegungen zur Kritik an der "Variablen-Soziologie". S. 230-245 in: Müller, N./Stachowiak, H. (Hrsg.), Problemlösungsoperator Sozialwissenschaft, Band 1. Stuttgart: Enke.

Fleiss, J. L./Dunner, D. L./Stallone, F./Fieve, R. R., 1976: The Life Table. A Method for Analyzing Longitudinal Studies. Archives of General Psychiatry 33: 107-112.

Galler, H. P., 1986: Übergangsratenmodelle bei intervalldatierten Ereignissen. Statistische Hefte 27: 1-22.

Galler, H. P./Pötter, U., 1992: Zur Robustheit von Schätzmodellen für Ereignisdaten. S. 379-405 in: Hujer, R./Schneider, H./Zapf, W. (Hrsg.), Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel. Frankfurt/New York: Campus.

Gerhardt, U., 1986: Patientenkarrieren. Frankfurt a.M.: Suhrkamp.

Hamerle, A./Tutz, G., 1989: Diskrete Modelle zur Analyse von Verweildauern und Überlebenszeiten. Frankfurt/New York: Campus.

Heckman, J./Singer, B., 1986: Econometric Analysis of Longitudinal Data. S. 1689-1763 in: Griliches, Z./Intriligator, M. D. (Hrsg.), Handbook of Econometrics, Volume III. Amsterdam: Elsevier.

Heijtan, D. F., 1989: Inference from Grouped Continuous Data: A Review. Statistical Science 4: 163-183.

Hess, H., 1978: Das Karriere-Modell und die Karriere von Modellen. S. 1-30 in: Hess, H./ Störzer, H. U./Streng, F. (Hrsg.), Sexualität und soziale Kontrolle. Heidelberg: Kriminalistik.

Hujer, R./Schneider, H., 1986: Semi-parametrische und parametrische Ratenmodelle. Eine anwendungsbezogene Einführung in die statistischen Grundlagen mit Programmbeispielen. Frankfurt: Sonderforschungsbereich 3, Arbeitspapier Nr. 200.

Hujer, R./Schneider, H., 1989: The Analysis of Labor Market Mobility Using Panel Data. European Economic Review 33: 530-536.

Hutchison, D., 1988a: Event History and Survival Analysis in the Social Sciences I. Background and Introduction. Quality & Quantity 22: 203-219.

Hutchison, D., 1988b: Event History and Survival Analysis in the Social Sciences II. Advanced Applications and Recent Developments. Quality & Quantity 22: 255-278.

Jacobs, H./Ringbeck, A., 1992: Zweiter Zwischenbericht zum Projekt "Hilfen zur Überwindung von Sozialhilfebedürftigkeit" im Auftrag des Bundesministeriums für Familie und Senioren. Köln: ISG (Institut für Sozialforschung und Gesellschaftspolitik).

Kalbfleisch, J. D./Prentice, R. L., 1980: The Statistical Analysis of Failure Time Data. New York: Wiley.

Keupp, H., 1987: Psychisches Leid als gesellschaftlich produzierter Karriereprozeß. S. 341-366 in: Voges, W. (Hrsg.), Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske + Budrich.

Kiefer, N. M., 1988: Economic Duration Data and Hazard Functions. Journal of Economic Literature 26: 646-679.

Klein, T., 1992: Zur Zeitabhängigkeit der Wiederbeschäftigungsrate Arbeitsloser. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 134-138.

Lancaster, T., 1990: The Econometric Analysis of Transition Data. Cambridge: Cambridge University Press.

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

Lee, E. T., 1980: Statistical Methods for Survival Data Analysis. Belmont, CA: Lifetime Learning.

Leisering, L./Zwick, M., 1990: Heterogenisierung der Armut? Alte und neue Perspektiven zum Strukturwandel der Sozialhilfeklientel in der Bundesrepublik Deutschland. Zeitschrift für Sozialreform 36: 715-745.

Ludwig-Mayerhofer, W., 1990: Arbeitslosigkeit im Erwerbsverlauf. Zeitschrift für Soziologie 19: 345-359.

Ludwig-Mayerhofer, W., 1992: Fakt und Artefakt in der Analyse von Arbeitslosigkeitsverläufen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 124-133.

Meinken, H., 1992: Die Modellierung zeitstetiger sozialer Prozesse - Untersuchungsmethoden für Lebensverlaufsereignisse. S. 67-88 in: Andreß, H. J./Huinink, J./Meinken, H./Rumianek, D./Sodeur, W./Sturm, G. (Hrsg.), Theorie, Daten, Methoden. Neue Modelle und Verfahrensweisen in den Sozialwissenschaften. München: R. Oldenbourg.

Mutz, G./Kühnlein, I., 1992: Zur Reformulierung des Karriere-Modells: Theoretische Skizze und empirische Ergebnisse einer Fallstudie. mps-texte 1: 37-50.

Namboodiri, N. K./Suchindran, C. M., 1987: Life Table Techniques and Their Applications. Orlando: Academic Press.

Ostner, I., 1987: Scheu vor der Zahl? Die qualitative Erforschung von Lebenslauf und Biographie als Element einer feministischen Wissenschaft. S. 103-124 in: Voges, W. (Hrsg.), Methoden der Biographie- und Lebenslaufforschung. Opladen: Leske + Budrich.

Petersen, T., 1990: Analyzing Event Histories. S. 259-288 in: von Eye, A. (Hrsg.), Statistical Methods in Longitudinal Research, Volume II. San Diego: Academic Press.

Petersen, T., 1991a: The Statistical Analysis of Event Histories. Sociological Methods and Research 19: 270-323.

Petersen, T., 1991b: Time-Aggregation Bias in Continuous-Time Hazard-Rate Models. S. 263-290 in: Marsden, P. V. (Hrsg.), Sociological Methodology 1991. Cambridge, MA: Basil Blackwell.

Petersen, T., 1993: Recent Advances in Longitudinal Methodology. Annual Review of Sociology 19: 425-454.

Petersen, T./Koput, K. W., 1992: Time-Aggregation Bias in Hazard-Rate Models With Covariates. Sociological Methods and Research 21: 25-51.

Peto, R./Pike, M. C./Armitage, P./Breslow, N. E./Cox, D. R./Howard, S. V./Mantel, N./ McPherson, K./Peto, J./Smith, P. G., 1977: Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient II: Analysis and Examples. British Journal of Cancer 35: 2-39.

Projektgruppe „Das Sozio-ökonomische Panel“, 1990: Das Sozio-ökonomische Panel für die Bundesrepublik Deutschland nach fünf Wellen. Vierteljahreshefte für Wirtschaftsforschung: 141-151.

Rohwer, G., 1991: Einkommensmobilität privater Haushalte 1984-1989. S. 379-408 in: Rendtel, U./Wagner, G. (Hrsg.), Lebenslagen im Wandel: Zur Einkommensdynamik in Deutschland seit 1984. Frankfurt/New York: Campus.

Rohwer, G., 1992: Einkommensmobilität und soziale Mindestsicherung. Einige Überlegungen zum Armutsrisiko. S. 367-379 in: Leibfried, S./Voges, W. (Hrsg.), Armut im modernen Wohlfahrtsstaat. (Sonderheft 32 der Kölner Zeitschrift für Soziologie und Sozialpsychologie). Opladen: Westdeutscher Verlag.

Rohwer, G., 1993: TDA Working Papers, Bremen, Ms.

Schneider, H., 1991: Verweildaueranalyse mit GAUSS. Frankfurt/New York: Campus.

Teachman, J. D., 1983: Analyzing Social Processes: Life Tables and Proportional Hazards Models. Social Science Research 12: 263-301.

Toutenburg, H., 1992: Moderne nichtparametrische Verfahren der Risikoanalyse. Heidelberg: Physica.

Tukey, J. W., 1977: Exploratory Data Analysis. Reading, MA: Addison & Wesley.

Tuma, N. B., 1982: Nonparametric and Partially Parametric Approaches to Event-History Analysis. S. 1-60 in: Leinhardt, S. (Hrsg.), Sociological Methodology 1982. San Francisco: Jossey--Bass.

Tuma, N. B./Hannan, M. T./Groeneveld, L. P., 1979: Dynamic Analysis of Event Histories. American Journal of Sociology 84: 820-854.

Tuma, N. B./Hannan, M. T., 1984: Social Dynamics. Orlando: Academic Press.

Voges, W./Rohwer, G., 1991: Zur Dynamik des Sozialhilfebezugs. S. 510-531 in: Rendtel, U./Wagner, G. (Hrsg.), Lebenslagen im Wandel: Zur Einkommensdynamik in Deutschland seit 1984. Frankfurt/New York: Campus.

Wagner, M., 1990: Arbeitslosenkarrieren. Journal für Sozialforschung 30: 5-23.

Yamaguchi, K., 1991: Event History Analysis. Newbury Park: Sage.

Statistische Modellierung von Verlaufsdaten in der Analyse sozialer Probleme, Teil II: Datenauswertung

von Wolfgang Ludwig-Mayerhofer

Abstract

This paper resumes the discussion from an introductory article on event history analysis (»survival analysis«, »analysis of failure times«) which appeared in the first part of this volume. It describes various approaches to the analysis of hazard functions and the assessment of the influence of covariates. After introductory remarks on non-parametric and semi-parametric estimation, an extensive discussion covers parametric models which may take into account variation of the hazard function over time. In addition, models for discrete time hazard functions, time-dependent covariates, competing risks, and repeated events are treated. All models are illustrated by an example from the German Socio-Economic Panel (SOEP).

Zusammenfassung

Die Arbeit baut auf dem einführenden Artikel zur Verlaufsdatenanalyse auf, der im vorangegangenen Heft dieser Zeitschrift erschienen ist. Sie stellt verschiedene Möglichkeiten vor, Hazardfunktionen und die Einflüsse von Kovariaten auf diese zu analysieren. Nach non-parametrischen und semi-parametrischen Analyseverfahren werden ausführlich Modelle der parametrischen Analyse diskutiert, welche die Veränderlichkeit der Hazardfunktion in der Zeit berücksichtigen können. Ferner werden Modelle für diskrete Verweildauern, zeitveränderliche Kovariaten, mehrere Zielzustände und wiederholte Ereignisse erörtert. Alle Modelle werden anhand eines Beispiels aus dem Sozio-ökonomischen Panel (SOEP) erläutert.

1. Übersicht zu statistischen Verfahren der Verlaufsdatenanalyse

In diesem Teil der Arbeit wird dargestellt, welche verschiedenen Möglichkeiten zur statistischen Analyse von Verlaufsdaten, insbesondere zur Schätzung von Hazardfunktionen oder -raten in Abhängigkeit von Kovariaten, zur Verfügung stehen. Wie schon in Teil I verweise ich häufig auf weiterführende Stellen in den drei deutschsprachigen Lehrbüchern mit den Kürzeln *A* für Andreß (1992a), *BHM* für Blossfeld/Hamerle/Mayer (1986) und *DM* für Diekmann/Mitter (1984).

Die statistische Analyse von Verlaufsdaten steht grundsätzlich vor den gleichen Problemen wie jede andere Datenauswertung. Da in aller Regel keine Vollerhebung, sondern eine Stichprobe vorliegt, möchte man wissen, in welchem Bereich die »wahren«, also die in der untersuchten Population gültigen Parameter mit hinlänglicher Wahrscheinlichkeit liegen werden. Ferner wird man in sozialwissenschaftlichen Untersuchungen zumeist die Frage beantworten wollen, ob Einflüsse von erklärenden Variablen auf den untersuchten Prozeß, wie sie sich in den explo-

rativen Auswertungen in Teil I am Beispiel der Arbeitslosigkeitsdauern aus dem Sozio-ökonomischen Panel hinsichtlich des Alters gezeigt haben, inferenzstatistisch absicherbar sind. Auf dieses Beispiel greife ich auch in diesem Teil zurück.

Im folgenden soll nur kurz auf einfache, nicht-parametrische Verfahren der Analyse eingegangen werden; etwas ausführlicher sollen semi-parametrische und noch umfassender parametrische Verfahren dargestellt werden. In den *nicht-parametrischen Verfahren* geht es – wie bei der in Teil I ausführlich dargestellten Life-Table-Schätzung – um die möglichst »datennahe« Schätzung der relevanten Funktionen und um einfache Prüfungen von Unterschieden zwischen verschiedenen Gruppen. Die Einfachheit ist gleichzeitig Vorzug, aber auch Nachteil dieser Verfahren. Da sie mit sehr wenigen Annahmen verbunden sind, wird vermieden, dem Datenmaterial ein Modell zu oktroyieren, das ihm möglicherweise nicht angemessen ist. Andererseits sind sie für eine multivariate Analyse nicht geeignet.

Semi-parametrische und parametrische Verfahren versuchen, den Verlauf des untersuchten Prozesses durch einen oder mehrere Parameter wiederzugeben. Hier werden die Survivorfunktion oder andere Funktionen nicht mehr explizit für jeden einzelnen Zeitpunkt (bzw. jedes Zeitintervall) angegeben, sondern durch einige wenige Kennzahlen, eben Parameter, charakterisiert. Gleichzeitig ist auf diese Weise – anders als mit den nicht-parametrischen Verfahren – eine simultane Berücksichtigung einer Vielzahl möglicher Einflüsse, also eine multivariate Analyse des untersuchten Prozesses möglich. Im allgemeinen wird der Einfluß der relevanten erklärenden Variablen auf die Hazardfunktion untersucht, da diese, wie wir gesehen haben, als den übrigen Funktionen zugrundeliegend aufgefaßt werden kann und im übrigen besser zwischen verschiedenen Verläufen unterscheidet, weil aus sehr unterschiedlichen Hazardraten zumindest oberflächlich recht ähnliche Survivorfunktionen resultieren können.

Die grundsätzliche Logik der Verfahren ist ganz ähnlich wie in den üblichen multivariaten Analysetechniken, etwa der linearen oder logistischen Regression. Untersucht wird, ob bestimmte Merkmale der Untersuchungseinheiten ceteris paribus, also unter statistischer Kontrolle der übrigen Merkmale, einen Einfluß auf den untersuchten Prozeß, hier: auf die Hazardfunktion, haben. Es werden Koeffizienten geschätzt, die angeben, ob bzw. wie die erklärende(n) Variable(n) - im Rahmen der Verlaufsdatenanalyse spricht man üblicherweise von Kovariaten - eine höhere oder niedrigere Hazardfunktion, also einen schnelleren oder langsameren Übergang in den Zielzustand, bewirken. Darüber hinaus lassen sich mit Hilfe der *parametrischen* Verfahren auch Koeffizienten schätzen, die die spezifische Verlaufsform der Hazardrate charakterisieren. Als *semi-parametrisch* wird dagegen ein von Cox (1972) entwickeltes Schätzverfahren bezeichnet, in welchem nur der Einfluß der Kovariaten geprüft wird, die dem Prozeß zugrundeliegende »Basisrate« jedoch unspezifiziert bleibt (weshalb das Verfahren auch nicht vollständig parametrisch, sondern eben nur semi-parametrisch ist).[1] Der Anspruch des letztgenannten Modells ist also bescheidener als der der parametrischen Verfahren. Je nach Forschungsinteresse kann hierin ein Vor- oder ein Nachteil gesehen werden. Geht man

von der Voraussetzung aus, daß Hypothesen über den Verlauf der Basisrate nur dann getestet werden sollen, wenn man dafür begründete theoretische Annahmen hat, so bietet das Cox-Modell eine ausgezeichnete Möglichkeit der multivariaten Analyse, wenn keine solchen Annahmen vorliegen. Umgekehrt läßt sich die Position vertreten, daß man mit dem Cox-Modell gerade die Möglichkeit verschenkt, die Zeitabhängigkeit des Prozesses zu analysieren, daß damit also ein unnötiger Verzicht auf relevante Informationen verbunden ist (Brüderl/Diekmann 1995).

2. Nicht-parametrische Verfahren

In Zusammenhang mit den nicht-parametrischen Verfahren der Verlaufsdatenanalyse sind drei Aspekte wichtig: 1. Die Schätzung der einschlägigen Funktionen, 2. die Berechnung von Konfidenzintervallen und 3. die Möglichkeit einfacher Gruppenvergleiche.

1. In Teil I, Abschnitt 3.2, wurde bereits ausführlich *eine* Möglichkeit einer einfachen Schätzung der grundlegenden Funktionen, also der Survivorfunktion S(t), der Dichtefunktion f(t) und der Hazardfunktion r(t), erläutert: die Life-Table- oder Sterbetafel-Methode. Ein zweites Verfahren, der sog. *Kaplan-Meier-* oder *Product-Limit-Schätzer* für S(t), basiert auf ganz ähnlichen Überlegungen, allerdings wird hier davon ausgegangen, daß die Verweildauern exakt, also nicht gruppiert gemessen wurden. Wegen der Annahme exakter Messungen wird S(t) für jeden Zeitpunkt berechnet, zu dem ein oder mehrere Ereignisse eingetreten sind. Angesichts dieser »punktuellen« Betrachtungsweise lassen sich f(t) und r(t) nicht unmittelbar schätzen, jedoch »Hazardkomponenten« für die Sprungstellen von S(t), aus denen sich auch eine kumulierte Hazardrate schätzen läßt.[2] Der wesentliche Unterschied zum Life-Table-Schätzer ist darin zu sehen, daß wegen der Annahme exakter Verweildauern eine Korrektur der »Risikomenge« bei Ereignissen nicht erfolgt. Es wird vielmehr angenommen, daß die Zensierungen und dementsprechend die Verringerung der Risikomenge jeweils zwischen Ereignisse fallen. Werden doch Zensierungen und Ereignisse zum gleichen Zeitpunkt beobachtet, werden die Zensierungen so behandelt, als wären sie nach den Ereignissen aufgetreten. Der Kaplan-Meier-Schätzer führt daher im vorliegenden Beispiel im Detail zu etwas anderen Ergebnissen als der Life-Table-Schätzer, die wesentlichen Schlußfolgerungen hinsichtlich des Arbeitslosigkeitsverlaufs und der Einflüsse des Alters sind aber in unserem Beispiel bei beiden Verfahren - wie auch sonst in aller Regel - identisch. Ausführlichere Darstellungen finden sich in den Lehrbüchern (A: 147 ff.; BHM: 44 ff., 124 ff.; DM: 76 ff.).

2. Im Rahmen der beiden genannten Schätzverfahren lassen sich jeweils Standardfehler und hieraus Vertrauensintervalle für S(t) bzw. - nur im Life-Table-Schätzer - für f(t) und r(t) berechnen. Hierzu sei wiederum auf die Lehrbuchliteratur verwiesen (A: 156 ff.; BHM: 45; DM: 66, 78).

3. Wichtiger als die Berechnung von Konfidenzintervallen für einzelne Funktionen sind in sozialwissenschaftlichen Anwendungen die Vergleiche zwischen Grup-

pen, wie etwa in unserem Beispiel zwischen den Altersgruppen.[3] Hierfür sind verschiedene nicht-parametrische Teststatistiken entwickelt worden. Die wichtigsten unter diesen sind zwei »klassische« Testverfahren, der *Log-Rank-Test*, auch als Mantel-Cox-Test oder Verallgemeinerter Savage-Test bezeichnet, sowie eine Teststatistik nach Gehan und Breslow, auch als *Verallgemeinerter Wilcoxon-Test* bezeichnet (vgl. A: 159; BHM: 48, Anwendung S. 128 ff.; DM: 86 ff.).[4] Der erstgenannte Test kann tendenziell eher Unterschiede am rechten Ende der Survivorfunktion entdecken, der zweite reagiert eher auf Unterschiede zu Beginn der Survivorfunktion, so daß in der Praxis unterschiedliche Entscheidungen über die Signifikanz von Unterschieden zustandekommen können. Zwei neuere Statistiken, die ebenfalls in einigen Programmen implementiert sind, wurden von Tarone/Ware (1977) und Prentice (1978; siehe auch Prentice/Marek 1979) entwickelt und liefern im allgemeinen Werte zwischen denjenigen des Log-Rank-Tests und der Gehan/Breslow-Statistik.

Abschließend ist festzuhalten, daß die Anwendungsmöglichkeiten nicht-parametrischer Verfahren für viele sozialwissenschaftliche Fragestellungen sicher begrenzt sind. Die hier angesprochenen Teststatistiken sind eher im Rahmen von kontrollierten Experimenten von Bedeutung, wo durch die randomisierte Zuteilung zu Kontroll- und Experimentalgruppen weitere Einflüsse ausgeschaltet werden können. Bei Untersuchungen, die nicht am Experimental-Paradigma ausgerichtet sind, wird dagegen häufig eine Vielzahl von potentiellen Einflüssen erfaßt, die simultan nur mit multivariaten Verfahren analysiert werden können. Es ist in einer solchen Situation sicherlich auch nicht sinnvoll, zunächst eine Vielzahl nichtparametrischer Signifikanztests zu berechnen und im Anschluß nur die in diesem Schritt signifikanten Einflüsse mit multivariaten Modellen zu prüfen, da wegen Suppressor-Effekten möglicherweise bedeutsame Einflüsse so nicht erkannt werden.

Dagegen sind einfache Berechnungen bzw. graphische Darstellungen von S(t), r(t) oder log S(t), wie sie in Teil I dargestellt wurden, wichtig für eine *explorative Datenanalyse*. Sie erlauben eine optische Inspektion der Daten, durch die eventuell vorhandene Datenfehler erkannt werden können. Ferner führen sie zu ersten Aufschlüssen über die Form der Survivor- bzw. Hazardfunktion, also über eine mögliche Zeitabhängigkeit des untersuchten Prozesses. Metrische Variablen können in mehrere Gruppen zerlegt werden, so daß - vorbehaltlich einer genaueren Prüfung mit multivariaten Verfahren - u.U. nicht-lineare Einflüsse erkannt werden können, wie sie in unserem Beispiel wohl hinsichtlich des Alters vorliegen.

Ob man ein solches exploratives Vorgehen für sinnvoll hält, hängt davon ab, welche Auffassung man von der Datenanalyse hat. Wer diese ausschließlich für den Test von Hypothesen als zulässig erachtet, muß ein exploratives Vorgehen als fragwürdig beurteilen. Nun besteht gewiß die Gefahr, daß man nach einer ausführlichen Datenexploration nur mehr die Zusammenhänge »bestätigt« findet, die man beim Screenen der Daten erst entdeckt hat. Andererseits sollte man gegenüber einer Praxis empirischer Sozialforschung, die abstrakte Modelle auf Daten anwendet,

ohne sich zu vergewissern, ob jene diesen in irgendeiner Weise angemessen sind, ebenfalls skeptisch sein. Daher sind explorative Analysen auch nicht nur eine Vorstufe der Datenauswertung, sondern können sehr wichtig sein, um etwa auffälligen oder unerwarteten Ergebnissen (und Nicht-Ergebnissen) auf die Spur zu kommen. »Hypothesentestende« und »explorative« Datenanalyse, das hat soeben Schnell (1994, vor allem Kap. 11) einmal mehr verdeutlicht, lassen sich ohnehin kaum sinnvoll voneinander abgrenzen; auch wer »explorativ« seine Daten durchforstet, hat dabei im allgemeinen bestimmte Ideen im Kopf, verfolgt also Hypothesen. Sogar der in Lehrbüchern als warnendes Beispiel dargestellte *Idealtypus* des Forschers, der »Alles mit Allem« korreliert, wird als *Realtypus* nur selten völlig willkürlich handeln, da ja schon die Auswahl der erhobenen Variablen nicht zufällig, sondern nach wenigstens impliziten Hypothesen erfolgte.

3. Ein semi-parametrischer Ansatz: Das »Partial-Likelihood«-Verfahren

In dem semi-parametrischen Modell von Cox wird folgende Gleichung zur Charakterisierung der Hazardfunktion geschätzt:[5]

$$r(t;\mathbf{X}) = r_0(t)\exp(\mathbf{X}\boldsymbol{\beta}) \qquad (1)$$

r(t;**X**) ist die Hazardrate für ein Individuum mit gegebenem Kovariaten-Vektor **X**. Diese Hazardrate ergibt sich also aus einer Basisrate $r_0(t)$ und den mit einem (Spalten-)Vektor **ß** gewichteten individuellen Ausprägungen der Kovariaten. Die Größe **Xß** wird mit der Basisrate exponentiell verknüpft. Wie oben erwähnt, wird die Basisrate nicht explizit geschätzt. Die exponentielle Verknüpfung mit **Xß** ist rein technisch motiviert; sie stellt sicher, daß keine negativen Werte für die Hazardfunktion geschätzt werden können. Die Parameter des Modells können aber leicht interpretiert werden, indem ihr Antilogarithmus $\alpha = \exp(ß)$ gebildet wird. Diese α-Parameter geben an, um welchen Faktor das Risiko, in den Zielzustand zu wechseln, erhöht ($\alpha > 1$) oder verringert ($\alpha < 1$) wird (BHM: 147; DM: 98 f. und 128 f.). Das Modell geht übrigens von der Annahme kontinuierlich gemessener Verweildauern aus, es gibt aber auch Vorschläge zu seiner Modifikation für diskrete bzw. gruppierte Dauern (siehe Abschnitt 5).

Aus Gleichung (1) und der soeben erläuterten Interpretation der Parameter ergibt sich, daß das Cox-Modell davon ausgeht, daß die Hazardfunktionen für verschiedene Werte von Kovariaten jeweils *proportional* zueinander sind. Angenommen, wir würden für eine Kovariate mit drei Ausprägungen 0, 1, und 2 einen Koeffi
zienten ß von 0,405 schätzen. Das Modell nimmt an, daß die Hazardfunktion für die Individuen mit der Ausprägung 1 in der betreffenden Variablen *stets* das a = exp(0,405) = 1,5-fache der Hazardfunktion der Individuen mit der Ausprägung 0 beträgt, für Individuen mit der Ausprägung 2 *stets* das $\alpha = \exp(2 \times 0{,}405) = 2{,}25$-

fache. Man spricht daher häufig im Zusammenhang mit dem Cox-Modell vom Proportional-Hazards-Modell. Tatsächlich gehen aber auch einige der später geschilderten parametrischen Modelle von proportionalen Hazardfunktionen aus. Die Annahme der proportionalen Hazardfunktionen kann explorativ mit einfachen Verfahren (siehe z.B. Teachman 1983), aber auch auf komplexere Weise im Rahmen der Modell-Schätzung geprüft werden (A: 254 ff.; BHM: 143 ff.). Allerdings dürfte das Cox-Modell relativ robust gegenüber nicht allzu starken Abweichungen von der Proportional-Hazards-Annahme sein. Liegen stärkere Abweichungen vor, ist eine Analyse im Rahmen des Cox-Modells unter Umständen trotzdem möglich, indem eine sog. stratifizierte Schätzung vorgenommen wird. Konkret: Ist für eine Variable die Annahme der proportionalen Hazards verletzt, so kann die gesamte Stichprobe in Gruppen oder »Schichten« (engl. strata) zerlegt werden, die den einzelnen Ausprägungen dieser Variablen entsprechen. Ist nun innerhalb dieser Gruppen die Annahme proportionaler Hazards gültig, kann ein Modell geschätzt werden, in dem für die verschiedenen Gruppen (Schichten) eine unterschiedliche Basisrate angenommen wird. Ausführliche Beispiele finden sich bei A: 254 ff. und BHM: 139 ff.

Die Modellschätzung erfolgt über ein Verfahren, welches als Partial-Likelihood-(PL-)Schätzung bezeichnet wird (grob gesagt deshalb, weil wegen der nicht explizit geschätzten Basisrate nicht die gesamte Information der Daten für die Schätzung benutzt wird). Trotzdem ist die grundsätzliche »Logik« des Schätzverfahrens nicht anders als in der Maximum-Likelihood-(ML-)Schätzung, wie sie den nachfolgend geschilderten parametrischen Modellen zugrundeliegt. Daher sei hier ganz kurz auf die wesentlichen inferenzstatistischen Gesichtspunkte eingegangen, also auf die Frage der statistischen Signifikanz der geschätzten Parameter.

Die Modellschätzung mit PL oder ML beruht auf iterativen Verfahren, in welchen der Wert einer Likelihood-Funktion maximiert wird, die angibt, wie wahrscheinlich man die gegebenen Daten in der Stichprobe erhalten würde, wenn die geschätzten Parameter den »wahren« Parametern entsprächen. Die Parameter mit der größten Wahrscheinlichkeit werden als die besten Schätzungen der wahren Parameter aufgefaßt. Die mathematische Logik des Verfahrens braucht uns hier nicht zu interessieren,[6] wohl aber die Frage, wie man aus den Resultaten Aufschluß über die inferenzstatistische Absicherung des gesamten Modells bzw. einzelner Effekte erhält.

Zunächst werden in allen Verfahren Standardfehler (engl. standard error, meist abgekürzt als S.E.) für die einzelnen Koeffizienten berechnet, welche bei hinreichend großen Stichproben als normalverteilt gelten können.[7] Daher kann - grob gesagt - ein Koeffizient als auf dem 5-Prozent-Niveau signifikant von Null verschieden angesehen werden, wenn das Verhältnis des Koeffizienten zu seinem Standardfehler (in den Programmen häufig auch als T-Statistik bezeichnet) mindestens 1,96 beträgt. Die entsprechenden Signifikanzniveaus werden aber von den meisten Programmen explizit berechnet.[8]

Noch zuverlässiger (und gleichzeitig universeller einsetzbar) sind jedoch Tests, die auf der Likelihood-Funktion aufbauen bzw. auf deren Logarithmus, im folgenden als Log-Likelihood bezeichnet. Grundsätzlich geht es immer darum, ein gegebenes Modell - nennen wir es M_1 - bzw. dessen Log-Likelihood LL_1 mit einem anderen - Modell M_0 mit Log-Likelihood LL_0 - zu vergleichen, aus welchem im Vergleich zu Modell M_1 einer oder mehrere Parameter weggelassen wurden. Die Größe

$$2\left(LL_1 - LL_0\right) \quad (2)$$

ist χ^2-verteilt mit s Freiheitsgraden, wobei s der Zahl der weggelassenen Parameter entspricht. Dieser Likelihood-Verhältnis- oder Likelihood-Ratio-Test[9] (abgekürzt: LR-Test) kann also eingesetzt werden, um

- zu testen, ob ein einzelner Parameter signifikant von Null verschieden ist, d.h. Modell M_0 wäre in diesem Fall ein Modell, aus dem im Vergleich zu M_1 eine einzelne Variable weggelassen wurde;
- zu testen, ob die Gesamtheit aller Variablen zusammengenommen eine signifikante Erklärungskraft besitzt, d.h. Modell M_0 wäre hier ein sog. Null-Modell, welches nur eine Modellkonstante enthält,[10] oder
- um den Einfluß einer beliebigen Zahl von Variablen zu testen, was z.B. geboten sein kann, wenn eine Variable in mehrere Dummy-Variablen zerlegt wurde und der Einfluß all dieser Dummy-Variablen zusammen geprüft werden soll (A: 205; BHM: 89; DM: 106 u. 141).

Weitere häufig verwendete Teststatistiken, die asymptotisch - d.h. mit zunehmendem Stichprobenumfang - mit dem LR-Test identisch sind, sind die Wald- und die Score-Teststatistik (vgl. A: 203 ff.; BHM: 89; DM: 106). Besonders hinzuweisen ist auf die Möglichkeit im Programm TDA (Rohwer 1994; vgl. Anhang), den Modell-Parametern beliebige Restriktionen aufzuerlegen. So kann z.B. leicht getestet werden, ob zwei oder mehrere Parameter sich signifikant voneinander unterscheiden.[11]

Nun aber zu den Ergebnissen des Cox-Modells für unseren konkreten Fall.[12] In *Darstellung 1* werden, wie im folgenden für weitere Beispiele, die Log-Likelihood für das geprüfte Modell mit den zwei (Dummy-)Variablen für die Altersgruppen »31 bis 50« und »über 50« sowie die ß-Koeffizienten für diese beiden Variablen, deren Standardfehler, die sich hieraus ergebende T-Statistik und das (zweiseitige) Signifikanzniveau angegeben.[13] Zunächst zur inhaltlichen Interpretation: Da im Cox-Modell die Basisrate nicht geschätzt wird, wird keine Regressionskonstante berechnet; das Modell sagt also nichts über die Hazardrate der jüngsten Altersgruppe, sondern nur darüber, wie die Raten der beiden aufgeführten Altersgruppen sich von der der jüngsten Gruppe *unterscheiden*. Der Koeffizient von -0,0494 für die 31- bis 50jährigen ergibt in den Exponenten erhoben einen Wert von etwa 0,95, d.h., die Rate dieser Gruppe beträgt im Schnitt das 0,95fache der Rate der Vergleichsgruppe, liegt also nur geringfügig unter dieser. Dagegen beträgt die Rate der

ältesten Gruppe das exp(-1,8977) = 0,15fache der jüngsten Gruppe, also nicht einmal ein Sechstel. Wie auch nach den explorativen Ergebnissen zu erwarten, ist der Unterschied zwischen der jüngsten Gruppe als Bezugsgruppe und den 31- bis 50jährigen nicht signifikant, dagegen derjenige zwischen der höchsten und der jüngsten Altersgruppe höchst signifikant.

Darstellung 1: *Ergebnisse eines Cox-Modells mit zwei Kovariaten (Erweiterte Beispieldaten aus Teil I, Darstellung 3)*

Variable	Coeff	Error	T-Stat	Signif
Alter 31 bis 50	-0.0494	0.1033	-0.4778	0.3672
Alter über 50	-1.8977	0.2404	-7.8928	1.0000

Log-likelihood: -2361.52

Die Log-Likelihood für das Null-Modell ohne Kovariaten beträgt -2417,26, so daß der LR-Test nach Formel (2) für das Gesamtmodell einen Wert von 2(-2361,52 - (-2417,26)) = 111,49 hat, der mit 2 Freiheitsgraden hoch signifikant ist.[14] Im Rahmen des Cox-Modells wird ferner häufig eine »Globale Chi-Quadrat-Statistik« berechnet, die einem Score-Test entspricht (vgl. dazu BHM: 89 und zur Anwendung S. 145 ff.). Deren Wert beträgt 82,70 und zeigt ebenfalls, daß das Gesamtmodell sich signifikant von einem Modell ohne Kovariaten unterscheidet. Will man z.B. den LR-Test auf die Variable »Alter 31 bis 50« anwenden und prüfen, ob deren Effekt signifikant von Null verschieden ist, so ergibt sich für das Modell, aus dem der Effekt dieser Variablen weggelassen (und damit auf Null gesetzt wurde), eine Log-Likelihood von 2361,63. Der LR-Test, ob dieses Modell signifikant von dem Ausgangsmodell verschieden ist, ergibt 2(-2361,52 - (-2361,63)) = 0,22, d.h., das Modell, welches keinen Unterschied zwischen den beiden jüngeren Altersgruppen postuliert, ist nicht signifikant schlechter und könnte somit als einfacheres Modell dem Ausgangsmodell vorgezogen werden.[15]

4. Parametrische Modelle (kontinuierliche Zeit)

Die parametrischen Modelle für kontinuierliche Verweildauern, die im folgenden vorgestellt werden, haben den Vorzug, nicht nur den Einfluß von Kovariaten - relativ zu einer unspezifizierten »Basisrate« - zu prüfen, sondern auch die zugrundeliegende Hazardfunktion selbst zu modellieren. Im folgenden werden exempla risch die wichtigsten Modelle dargestellt, wobei weiterhin die schon verwendeten einfachen Beispieldaten herangezogen werden. Im Vordergrund stehen die Möglichkeiten, unterschiedliche Verläufe der Hazardrate zu modellieren.

Zunächst ist hier eine Warnung auszusprechen. Definitive Aussagen über den zeitlichen Verlauf sind immer mit Fragezeichen zu versehen, da sie auch auf nicht bzw. nicht ausreichend berücksichtigte Einflüsse (sog. »unbeobachtete Heterogenität«) zurückgehen können. Konkret: Wenn sich in einem Datensatz zwei (oder

mehr) Gruppen befinden, von denen die eine eine hohe und die andere eine niedrige Hazardrate aufweist, so ereignen sich bei den Personen aus der ersten Gruppe die Übergänge in den Zielzustand natürlich wesentlich schneller, so daß am Schluß vor allem die Personen aus der anderen Gruppe mit sehr langen Verweildauern verbleiben. Würde man nicht zwischen diesen Gruppen unterscheiden (also z.B. in der Arbeitslosenstichprobe nicht nach Arbeitslosen bis 50 und solchen über 50 Jahren differenzieren), hätte man den Eindruck, daß die Hazardrate insgesamt immer mehr abnehmen würde, auch wenn sie in Wirklichkeit in beiden Gruppen möglicherweise konstant ist. Auf diesen - schon lange bekannten - Sachverhalt hat kürzlich wieder Klein (1992) hingewiesen (vgl. außerdem Andreß 1988; Arminger 1984). Auf der anderen Seite sollte man sich hierdurch nicht entmutigen lassen, nach Zeitabhängigkeiten zu suchen - denn schließlich gilt die Feststellung, daß unbeobachtete Einflüsse die Resultate statistischer Modellbildung in Frage stellen, ganz grundsätzlich (Lieberson 1985). Daß man keine definitive Aussage machen kann, welches Modell den Daten angemessen ist, sollte sich eigentlich von selbst verstehen; da aber, wie wir sehen werden, die verschiedenen Modelle den Daten eine bestimmte Struktur »aufzwingen«, sollte man in jedem Fall versuchen, auch alternative Modelle zu testen. Gegebenenfalls muß es weiteren Untersuchungen überlassen bleiben, zwischen möglicherweise divergierenden Schlußfolgerungen zu entscheiden.

Im übrigen wird häufig angenommen, daß abgesehen von der Frage der Zeitabhängigkeit die Wahl des konkreten Modells für die Analyse von geringer Bedeutung ist, da es hinsichtlich der zumeist wichtigeren Informationen über die Einflüsse der Kovariaten keine großen Divergenzen zwischen verschiedenen Modellen gibt. Das zeigen empirische Beispiele (Diekmann/Klein 1991; Ziegler/ Brüderl/Diekmann 1988) ebenso wie theoretische Überlegungen (Galler/ Pötter 1992). Allerdings werden wir gerade in unserem Beispiel sehen, daß das nicht in allen Fällen zutrifft und ein - allerdings wahrscheinlich unangemessenes - Modell auch zu inhaltlich anderen Schlußfolgerungen führen kann!

Zur Überprüfung, ob unbeobachtete Heterogenität vorliegt, wurde ein Verfahren entwickelt, welches unter der Annahme einer bestimmten Verteilung (der Gamma-Verteilung) für die Varianz der Fehler des Modells - also der nicht erklärten Anteile - prüfen kann, wie groß die unbeobachtete Heterogenität ist (A: 266 ff.; BHM: 251 ff.). Allerdings wird man realistischerweise davon ausgehen müssen, daß mit

diesem Verfahren fast immer ein relevanter Anteil an unbeobachteter Heterogenität entdeckt wird (vgl. auch Galler/Pötter 1987; Petersen 1993).[16]

Wir beginnen nunmehr mit dem einfachsten parametrischen Modell, dem sog. *Exponentialmodell.* Hier besteht zwischen den Kovariaten **X** und der Rate r(t;**X**) folgende einfache Beziehung:

$$r(t;\mathbf{X}) = \exp(\boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta}) \qquad (4)$$

Darstellung 2: *Ergebnisse verschiedener parametrischer Exponentialmodelle (Erweiterte Beispieldaten aus Teil I, Darstellung 3)*

Variable	Coeff	Error	T-Stat	Signif
Einfaches Exponentialmodell (Log-likelihood: -1303.72)				
Konstante	-1.9351	0.0639	-30.2885	1.0000
Alter 31 bis 50	-0.2175	0.1028	-2.1153	0.9656
Alter über 50	-2.2821	0.2381	-9.5829	1.0000
Piecewise Constant (Log-likelihood: -1244.53)				
Konstante für Monat 1-2	-1.8447	0.0914	-20.1773	1.0000
Konstante für Monat 3	-1.2781	0.1143	-11.1787	1.0000
Konstante für Monat 4-6	-1.7875	0.1087	-16.4466	1.0000
Konstante für Monat 7-12	-2.1761	0.1372	-15.8641	1.0000
Konstante für Monat 13-24	-3.0400	0.2305	-13.1912	1.0000
Konstante für Monat 25 u. höher	-3.6244	0.3395	-10.6756	1.0000
Alter 31 bis 50	-0.0812	0.1034	-0.7853	0.5677
Alter über 50	-1.9874	0.2402	-8.2755	1.0000
Polynom 1. Grades (Log-likelihood: -1254.95)				
Konstante	-1.5832	0.0725	-21.8417	1.0000
Alter 31 bis 50	-0.0686	0.1033	-0.6643	0.4935
Alter über 50	-1.9894	0.2399	-8.2925	1.0000
$ß_{t1}$	-0.0671	0.0087	-7.7399	1.0000
Polynom 4. Grades (Log-likelihood: -1247.23)				
Konstante	-1.8783106	0.0912149	-20.5921450	1.0000
Alter 31 bis 50	-0.0712948	0.1032454	-0.6905368	0.5101
Alter über 50	-1.9932917	0.2402498	-8.2967483	1.0000
$ß_{t1}$	0.1560741	0.0305186	5.1140619	1.0000
$ß_{t2}$	-0.0303144	0.0043719	-6.9338968	1.0000
$ß_{t3}$	0.0012134	0.0002318	5.2344372	1.0000
$ß_{t4}$	-0.0000145	0.0000036	-3.9948453	0.9999

Log-likelihood des *Exponentialmodelles nur mit Konstante*: -1393.38

Wie im Cox-Modell wird die Rate hier (wie auch in den folgenden Modellen) exponentiell mit dem durch **ß** gewichteten Kovariatenvektor verknüpft, um eine negative Schätzung der Rate zu vermeiden.[17] Im Gegensatz zu jenem Modell enthält der Koeffizientenvektor jetzt aber (wie ebenfalls in allen folgenden Modellen) eine Modellkonstante - hier als $ß_0$ bezeichnet -, die die »Basisrate« angibt, also die Rate für diejenigen Fälle, die in allen Kovariaten den Wert 0 aufweisen (in unserem einfachen Beispiel die Altersgruppe bis 30 Jahre). Damit lassen sich nicht nur Aussagen über die *relativen Chancen* der verschiedenen Gruppen treffen, die Arbeitslosigkeit zu verlassen, sondern diese Chancen lassen sich als Hazardrate für die verschiedenen Gruppen explizit »beziffern«. Wie aus der Modellformulierung ersichtlich wird, wird $r(t;\mathbf{X})$ als zeitlich konstant aufgefaßt.

Eine Schätzung dieses einfachen Exponentialmodells kommt zu dem etwas überraschenden Ergebnis, daß der Unterschied in den Hazardfunktionen (und damit in den Arbeitslosigkeitsdauern) zwischen der jüngsten und der mittleren Altersgruppe signifikant ist (vgl. *Darstellung 2*)![18] Im einzelnen lassen sich die Ergebnisse so interpretieren: Für die jüngste Altersgruppe wird eine Hazardrate von exp(-1,9351) = 0,144 geschätzt, für die mittlere Gruppe eine Rate von exp(-1,9351 + (-0,2175)) = 0,116 und für die älteste Gruppe eine Rate von exp(-1,9351 + (-0,2812)) = 0,015.

Wie schneidet dieses Modell im Vergleich mit einem Null-Modell, also einem Exponentialmodell mit einer Konstanten, aber ohne Kovariaten ab? (Da das einfache Exponentialmodell ohne Kovariaten den allereinfachsten Fall eines parametrischen Modells darstellt, wird es im allgemeinen als Null-Modell für alle parametrischen Modellklassen eingesetzt.) Die Log-Likelihood des letzteren Modells beträgt -1393,39, das Modell mit Kovariaten bringt also durchaus eine entscheidende Verbesserung mit 2(-1303,72 - (-1393,38)) = 179,32.

Im vorliegenden Fall müssen wir allerdings aufgrund der explorativen Ergebnisse damit rechnen, daß das einfache Exponentialmodell, auch wenn es gegenüber dem Null-Modell eine erhöhte Erklärungskraft besitzt, den Arbeitslosigkeitsverlauf nicht adäquat modelliert, da die Hazardrate für die einzelnen Gruppen als konstant angenommen wird. Nach den explorativen Analysen in Teil I dürfte die Annahme der Konstanz jedoch kaum gerechtfertigt sein. Man beachte auch, daß nach dem geschätzten Modell die Rate der Altersgruppe 31 bis 50 Jahre nur etwa das 0,8fache der Basisrate (also der Rate für die Vergleichsgruppe bis 30 Jahre) beträgt; auch dies scheint nach den explorativen Darstellungen des Life-Table-Schätzers kaum gerechtfertigt.

Eine variable Hazardrate kann man durch verschiedene andere Verteilungsannahmen modellieren, auf die wir noch zurückkommen. Allerdings läßt sich auch das einfache Exponentialmodell durch verschiedene Zusatzannahmen zu folgenden zwei Modellen erweitern:[19]

Piecewise Constant Exponentialmodell:

$$r(t;\mathbf{X}) = \exp(\beta_{0t} + \mathbf{X}\boldsymbol{\beta}) \tag{5}$$

Exponentialmodell mit Polynom-Term für die Zeit:

$$r(t;\mathbf{X}) = \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta} + \beta_{t1}t + \beta_{t2}t^2 \ldots + \beta_{tn}t^n) \tag{6}$$

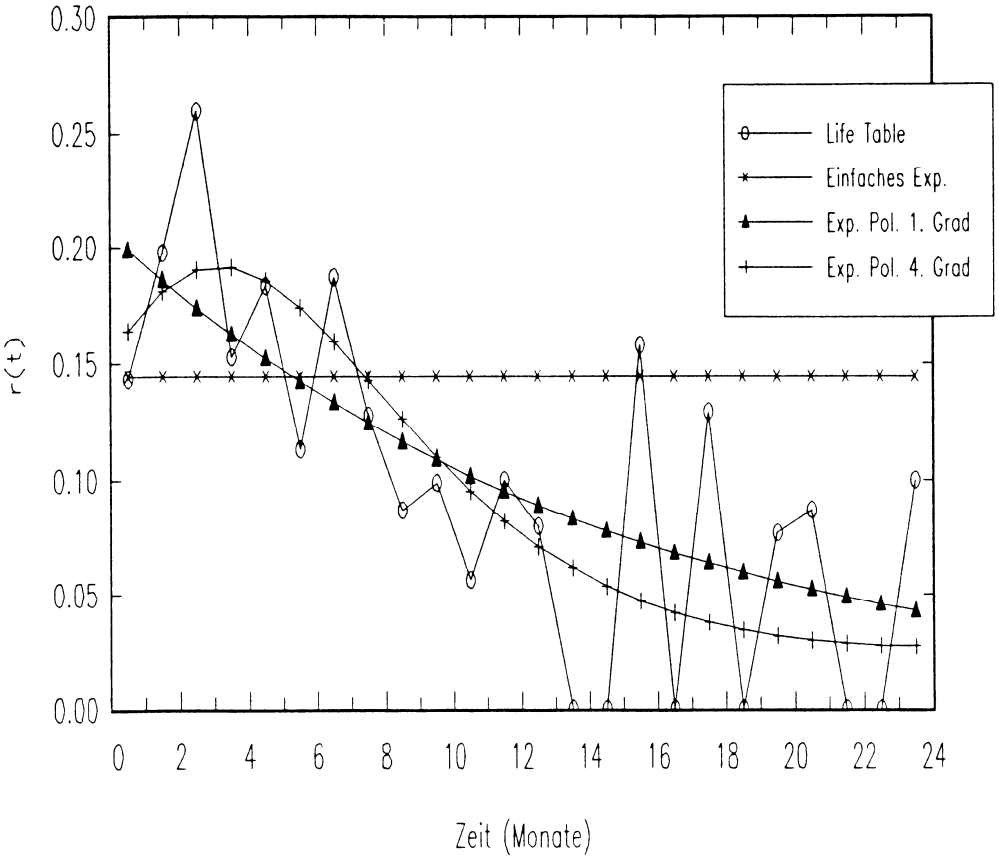
Im *Piecewise Constant Exponentialmodell* (oder periodenspezifischen Exponentialmodell) gibt es nicht nur eine Modellkonstante ß$_0$, sondern mehrere *periodenspezifische* Konstanten ß$_{0t}$ (vgl. A: 231 f. und 235 ff.). Die Perioden beziehen sich auf die Prozeßzeit und können vom Benutzer selbst beliebig spezifiziert werden. Das legt natürlich die Gefahr des »curve fitting« nahe, erlaubt andererseits eine recht flexible Modellstruktur. Für die Beispieldaten wurde ein Modell getestet mit der Annahme von sechs unterschiedlichen Basisraten in den Monaten 1 und 2, 3, 4

bis 6, 7 bis 12, 13 bis 24 sowie über 24 Monate. Offensichtlich entspricht dieses Modell dem beobachtbaren Verlauf der Hazardrate deutlich besser, was sich in einer erheblichen Erhöhung der Log-Likelihood niederschlägt (vgl. Darstellung 2).

Im *Exponentialmodell mit Polynom-Term für die Zeit* werden dagegen zusätzliche ß-Koeffizienten geschätzt - hier als $ß_{tn}$ bezeichnet -, die mit der Zeit in der Form von Polynomen beliebigen Grades verknüpft werden können. Wieder kann bzw. muß vom Benutzer spezifiziert oder ausprobiert werden, welcher Grad des Polynoms zu sinnvollen Ergebnissen führt. Bei den Beispieldaten ergibt sich in einem Modell mit dem Zeitglied 1. Grades eine fallende Rate (vgl. Darstellung 2). Modelle mit einem Glied 2. oder 3. Grades, die grundsätzlich in der Lage sein müßten, einen zuerst steigenden und dann fallenden Verlauf zu modellieren - wie er nach den explorativen Analysen gegeben ist -, haben im vorliegenden Fall weder einen signifikanten Erklärungszuwachs erbracht noch auch den entsprechenden Verlauf modellieren können. Erst ein Polynom 4. Grades führt zur Modellierung einer zunächst steigenden ($ß_{t1}$ ist positiv!) und dann fallenden Rate; der Erklärungszuwachs dieses Modells gegenüber dem Polynom 1. Grades in Höhe von 2(-1247,23 - (-1254,95)) = 15,44 ist auch bei drei Freiheitsgraden (für die drei zusätzlichen Parameter) nach der χ^2-Verteilung signifikant. Trotzdem sollte man nicht annehmen, hiermit »bewiesen« zu haben, daß das Polynom 4. Grades ein besseres Modell impliziert; eine gewisse Plausibilität hierfür wird allerdings durch den höheren Modellfit nahegelegt.

Da dieses Modell in der Lehrbuchliteratur (mit Ausnahme einer kurzen Erwähnung bei A: 233) nicht dargestellt wird, seien noch einmal exemplarisch für einige Zeitpunkte die geschätzten Hazardraten für die Referenzgruppe bis 30 Jahre berechnet. Für den Zeitpunkt 1 ergibt sich - annäherungsweise berechnet - eine Rate von $\exp(-1{,}8783 + 0{,}1561 \times 1 - 0{,}0303 \times 1^2 + 0{,}0012 \times 1^3 - 0{,}0000145 \times 1^4) = 0{,}1735$, für den Zeitpunkt 3 erhält man eine etwas höhere Rate von 0,1918, nach 12 Monaten beträgt sie aber nur mehr 0,0746 und nach 24 Monaten schließlich $\exp(-1{,}8783 + 0{,}1561 \times 24 - 0{,}0303 \times 24^2 + 0{,}0012 \times 24^3 - 0{,}0000145 \times 24^4) = 0{,}0222$. Bei den anderen beiden Altersgruppen sind die jeweiligen gruppenspezifischen Koeffizienten noch mit in den Ausdruck in der Klammer aufzunehmen.

Darstellung 3: *Hazardfunktionen für die Altersgruppe bis 30 (Schätzungen nach Life Table, Exponentialmodell, Exponentialmodell mit Polynom 1. und Polynom 4. Grades)*



Schließlich ist noch darauf hinzuweisen, daß die Modelle mit Polynom in der Praxis teilweise schwer zu handhaben sind, da im allgemeinen spätestens mit dem Polynom-Glied vierten Grades die iterative Modellschätzung numerisch sehr schwierig wird. Auch im vorliegenden Fall ist nicht ganz sicher, ob der Schätzalgorithmus tatsächlich das globale Maximum erreicht hat. Im folgenden werden einige Modelle vorgestellt, die jedenfalls teilweise leichter zu handhaben sind. Doch zuvor möchte ich einige der von den bislang erörterten Modellen geschätzten Hazardraten graphisch zeigen, und zwar exemplarisch für die jüngste Altersgruppe *(Darstellung 3)*. So kann besser verdeutlicht werden, welche unterschiedlichen Implikationen die verschiedenen Modellschätzungen für die Hazardrate haben: die konstante Rate des einfachen Exponentialmodells, die fallende Rate des Exponentialmodells mit einem Polynom-Glied und die erst steigende und dann fallende des Exponentialmodells mit einem vierfachen Polynom.[20] Auch anhand dieser Darstellung wird man zu der Schlußfolgerung kommen, daß das letztgenannte Modell

noch am ehesten geeignet ist, den »datennahen« Life-Table-Schätzer nachzuvollziehen. Angemerkt sei abschließend auch, daß alle bislang besprochenen komplexeren Modelle den Alterseffekt für die mittlere Gruppe im Vergleich zur jüngsten Gruppe als nicht signifikant ausweisen. Das wird auch für die folgenden Modelle gelten, so daß dort nicht mehr eigens darauf hingewiesen wird.

Nunmehr möchte ich noch einige weitere sehr häufig gebrauchte Verteilungen vorstellen, die geeignet sind, nicht konstante Verläufe der Hazardrate zu modellieren (vgl. zu den Ergebnissen *Darstellung 4*).

Das *Weibull-Modell* läßt sich in den äquivalenten Formulierungen

$$r(t;\mathbf{X}) = \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta})\, p\left(\exp(\boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta})\, t\right)^{p-1} \tag{7}$$

bzw.

$$r(t;\mathbf{X}) = \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta})^{p}\, p\, t^{\,p-1} \tag{8}$$

darstellen. Der Parameter p[21] ist als sog. »Shape Parameter« u.a. mit der Prozeßzeit t verknüpft, und somit ist leicht zu sehen, daß r(t) mit zunehmender Zeit monoton ansteigt, wenn p > 1, und monoton absinkt, wenn p < 1. Im Falle von p = 1 erhält man wieder ein einfaches Exponentialmodell. Zu beachten ist, daß in den meisten Fällen - und so auch hier - von den Statistikprogrammen nicht p ausgegeben wird, sondern ln(p), und das heißt, daß ln(p) < 0 eine fallende und ln(p) > 0 eine steigende Hazardrate impliziert. Im vorliegenden Fall wird aufgrund der längerfristig fallenden Rate ein signifikant negativer Wert für ln(p) geschätzt.

Ebenfalls eine monoton steigende oder fallende Rate modellieren kann die sog. *Gompertz-Verteilung*, die sich in unserem Kontext folgendermaßen formulieren läßt:

$$r(t;\mathbf{X}) = \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta}) + \exp\left((\gamma_0 + \mathbf{X}\boldsymbol{\gamma})\, t\right) \tag{9}$$

Hier wird die Zeitabhängigkeit modelliert, indem die Prozeßzeit mit einem weiteren Parameter γ_0 und gegebenenfalls zusätzlich mit einem Parametervektor $\boldsymbol{\gamma}$ und Kovariaten verknüpft wird. Wird nur eine Konstante γ_0 geschätzt, wird der zeitlich variable Hazardratenverlauf für alle Untersuchungseinheiten (bzw. alle Kovariatenkonstellationen) als gleich angenommen, ansonsten können durch die verschiedenen mit dem Parametervektor $\boldsymbol{\gamma}$ verknüpften Kovariaten auch spezifische Verläufe modelliert werden (wobei sich dann u.U. für bestimmte Konstellationen steigende und für andere Konstellationen fallende Verläufe ergeben können) (vgl. A: 228 ff.; BHM: 211 ff.). Bei unseren Beispieldaten ergibt sich (vgl. Darstellung 4), daß nur die Konstante γ_0 signifikant von Null verschieden ist, nicht aber die $\boldsymbol{\gamma}$-Koeffizienten für die beiden Kovariaten, so daß man davon ausgehen kann, daß –

sofern man eine konstant sinkende Rate überhaupt als sinnvolle Modellierung erachtet - der Ratenverlauf für alle Subgruppen gleich ist.

Darstellung 4: *Ergebnisse weiterer parametrischer Modelle (erweiterte Beispieldaten aus Teil I, Darstellung 3)*

Variable	Coeff	Error	T-Stat	Signif
Weibull (Log-likelihood: -1291.22)				
Konstante	-1.9107	0.0758	-25.2143	1.0000
Alter 31 bis 50	-0.1755	0.1223	-1.4354	0.8488
Alter über 50	-2.5326	0.2878	-8.8005	1.0000
ln(p)	-0.1702	0.0359	-4.7380	1.0000
Gompertz, Kovariaten in ß- und g-term (Log-likelihood: -1254.74)				
ß-Konstante	-1.6039	0.0813	-19.7368	1.0000
ß-Alter 31 bis 50	-0.0317	0.1313	-0.2417	0.1910
ß-Alter über 50	-1.8592	0.3417	-5.4406	1.0000
g-Konstante	-0.0619	0.0123	-5.0351	1.0000
g-Alter 31 bis 50	-0.0084	0.0179	-0.4686	0.3606
g-Alter über 50	-0.0209	0.0383	-0.5460	0.4150
Log-logistisch (Typ I) (Log-likelihood: -1248.61)				
Konstante	-1.3210	0.0731	-18.0728	1.0000
Alter 31 bis 50	-0.0876	0.1199	-0.7306	0.5350
Alter über 50	-2.3385	0.2207	-10.5946	1.0000
ln(p)	0.2925	0.0404	7.2349	1.0000
Log-logistisch (Typ II) (Log-likelihood: -1236.48)				
Konstante	-1.3287	0.0707	-18.7821	1.0000
Alter 31 bis 50	-0.0805	0.1032	-0.7800	0.5646
Alter über 50	-1.9901	0.2399	-8.2950	1.0000
ln(l)	-0.7707	0.1541	-5.0010	1.0000
ln(p)	0.5353	0.0761	7.0368	1.0000
Sichel (Log-likelihood: -1278.72)				
Konstante	-1.7516	0.0865	-20.2569	1.0000
Alter 31 bis 50	-0.0914	0.1031	-0.8865	0.6246
Alter über 50	-2.0712	0.2396	-8.6430	1.0000
ln(g)	1.3489	0.0529	25.4761	1.0000

Das Gompertz-Modell impliziert bei negativen γ-Koffizienten, daß die Hazardrate mit der Zeit gegen Null tendiert. Das *Gompertz-Makeham-Modell* enthält eine zusätzliche Konstante α, gegen die die Rate - sofern sie fällt - tendieren würde. Auch diese Konstante kann (muß aber nicht) mit Kovariaten verknüpft werden, so daß sich als vollständige Formulierung ergibt:

$$r(t;\mathbf{X}) = \exp(\alpha_0 + \mathbf{X}\boldsymbol{\alpha}) + \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta}) + \exp((\gamma_0 + \mathbf{X}\boldsymbol{\gamma})\, t) \qquad (10)$$

Dieses Modell ist bereit relativ komplex; es kann zwar sehr verschiedenartige (wenngleich nur monoton steigende oder fallende) Verläufe modellieren, dies wird jedoch mit dem Nachteil einer sehr aufwendigen Modellsuche sowie im übrigen auch mit Problemen bei der Durchführung der Modellschätzung erkauft. Solche Schwierigkeiten deuten dann allerdings im allgemeinen darauf hin, daß das Modell den Daten nicht sehr gut angemessen ist. Auch bei unseren - eigentlich sehr einfachen - Beispieldaten traten solche Probleme auf (der Schätzalgorithmus konvergierte erst nach 75 Iterationen und erbrachte Schätzungen mit extrem hohen Standardfehlern), was als Indikator dafür gelten könnte, daß der unterstellte monotone Verlauf nur eine grobe Annäherung an den tatsächlichen Verlauf ist. (Daher verzichte ich auf eine Wiedergabe der Ergebnisse in Darstellung 4).

Im Gegensatz zu Weibull- und Gompertz-(Makeham-)Modellen sind die beiden folgenden Modelle in der Lage, Hazardfunktionen zu modellieren, die zunächst steigen und erst dann fallen. Relativ häufig angewendet wird die *log-logistische Verteilung* mit folgender Formulierung (A: 292 f.; BHM: 39 f. und 240 f.; DM: 153):[22]

$$r(t;\mathbf{X}) = \frac{\exp(\beta_0 + \mathbf{X}\boldsymbol{\beta})^p \, p\, t^{p-1}}{1 + \left(\exp(\beta_0 + \mathbf{X}\boldsymbol{\beta})\, t\right)^p} \tag{11}$$

Hier ergibt sich bei einem Koeffizienten p von > 1 (bzw. von ln(p) > 0) eine zunächst steigende und dann fallende Hazardrate, während bei p < 1 bzw. ln(p) < 0 die Hazardrate von Anfang an fällt. Im vorliegenden Falle ergibt die Modellschätzung tatsächlich den nach den explorativen Analysen in Teil I, Abschnitt 3 zu erwartenden Verlauf einer zunächst steigenden und anschließend fallenden Hazardrate.

Ein noch flexibleres log-logistisches Modell mit einem weiteren Parameter wurde von Schneider (1991) und Brüderl (vgl. Brüderl/Diekmann 1995) entwickelt. In der Formulierung von Brüderl - diejenige von Schneider unterscheidet sich nur geringfügig - wird folgende Rate geschätzt:

$$r(t;\mathbf{X}) = \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta}) \frac{p(\lambda\, t)^{p-1}.}{1 + (\lambda\, t)^p} \tag{12}$$

Mit dieser Formulierung - zur Unterscheidung von der vorhergehenden Formulierung spreche ich vom log-logistischen Modell Typ II[23] - können noch steiler ansteigende Hazardfunktionen modelliert werden als mit dem zuerst genannten Modell vom Typ I. Dementsprechend weist das Modell vom Typ II eine noch größere Log-Likelihood auf als das Modell vom Typ I, d.h., es ist den Daten vermutlich noch besser angemessen. Nach der Höhe der Log-Likelihood handelt es sich hierbei um das erklärungskräftigste Modell überhaupt; es übertrifft sogar die weiter oben

dargestellten komplexen Exponentialmodelle, obwohl diese noch mehr Parameter zur Modellierung der Hazardfunktion verwenden.

Ebenfalls eine zunächst steigende und dann fallende Rate kann mit Hilfe der *Sichel-Verteilung* modelliert werden (vgl. A: 231; DM: 152 f.):

$$r(t;\mathbf{X}) = \exp(\beta_0 + \mathbf{X}\boldsymbol{\beta})\, t\, \exp(-t / \gamma) \tag{13}$$

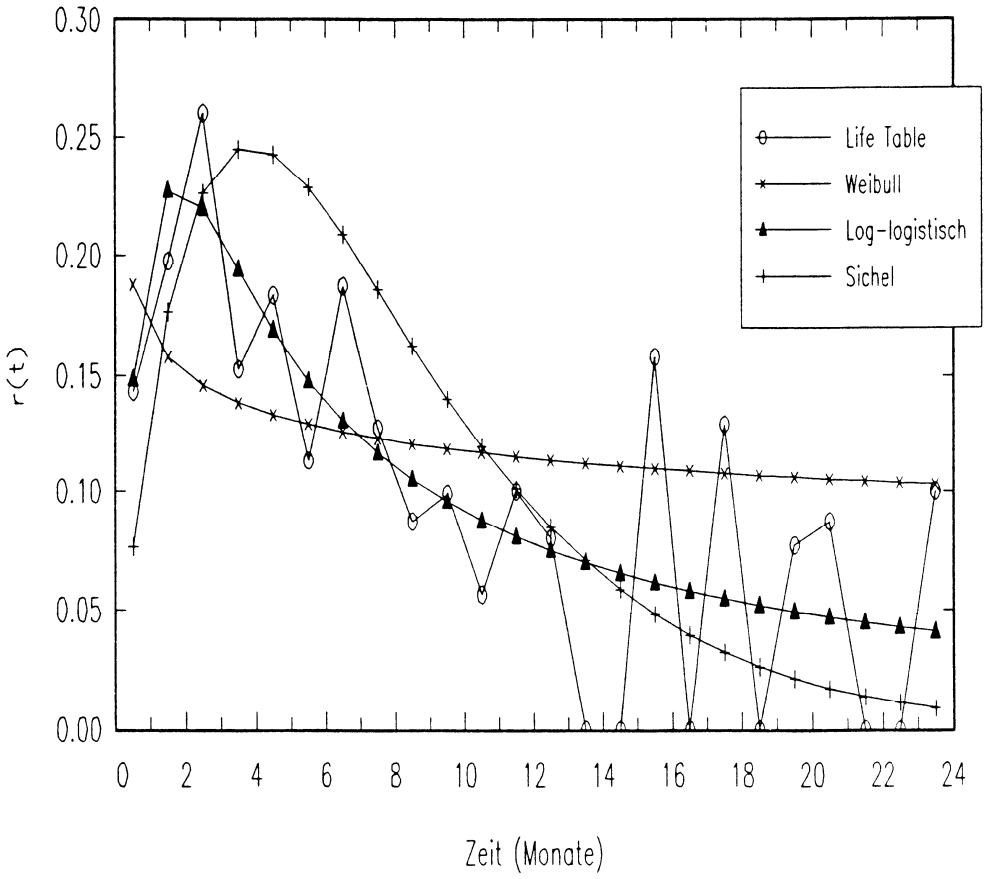
Dieses Modell läßt sich relativ gut inhaltlich interpretieren: Angenommen wird ein Maximum der Rate bei $t = \gamma$ (in unserem Beispiel: exp(1.3489) = 3.85) und ein Wendepunkt bei $t = 2\gamma$.[24] Eine wesentliche Eigenschaft unterscheidet es von dem vorgenannten Modell: Es geht davon aus, daß ein Teil der Untersuchungspopulation niemals den Ausgangszustand verläßt, was in dem Kontext, in dem dieses Modell entwickelt wurde, der Analyse von Heiratsdauern, durchaus realistisch ist, da - trotz ansteigender Scheidungsziffern - die Mehrzahl der Ehen nicht mit einer Scheidung endet. Im vorliegenden Fall ist diese Annahme allerdings möglicherweise unangemessen, und so zeigt dieses Modell eine deutlich geringere Log-Likelihood als die meisten anderen Modelle.

Wenn wir jetzt wiederum einige der erörterten Modelle graphisch vergleichen[25], so wird man auch aufgrund der Verläufe der Hazardraten *(Darstellung 5)* geneigt sein, dem log-logistischen Modell (Typ II) den Vorzug zu geben, welches auch den höchsten Wert der Log-Likelihood aufweist. Zusätzlich möchte ich die von den Modellen vorhergesagten Survivorfunktionen abbilden *(Darstellung 6)*. Zunächst bestätigt sich die weiter oben getroffene Feststellung, daß die sehr unterschiedlichen Hazardraten, die von den verschiedenen Modellen geschätzt werden, zu recht ähnlichen Survivorfunktionen führen. Bei genauem Hinsehen zeigt sich aber doch, daß jedenfalls im mittleren Bereich das Weibull-Modell den Verlauf der Survivorfunktion tendenziell unter- und das Sichelmodell diesen Verlauf tendenziell überschätzt, während das log-logistische Modell insgesamt die größte Nähe zu dem »datennahen« Life-Table-Schätzer aufweist.

Trotzdem möchte ich noch einmal betonen, daß die »Ergebnisse« dieser Modelle teilweise durch die Modellwahl *vorausgesetzt* sind: Auch wenn die Hazardrate tatsächlich ansteigt, führt z.B. ein log-logistisches oder log-normales Modell zu dem »Ergebnis«, daß die Rate nach einem initialen Anstieg absinkt! Umgekehrt führen Weibull- und Gompertz-Modell immer zu dem »Ergebnis« monoton steigender oder fallender Raten, auch wenn kein monotoner Verlauf vorliegt.

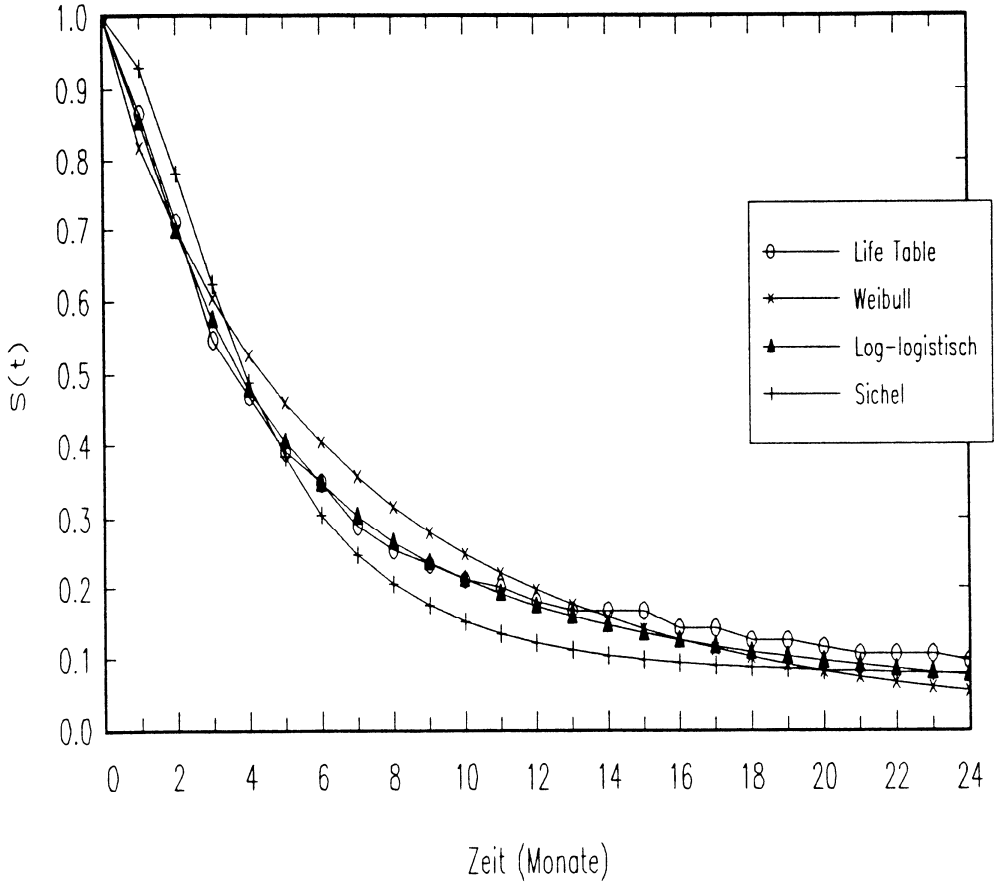
Für weitere Überprüfungen, welches aus der Vielzahl möglicher Modelle den Daten am besten angemessen ist, werden in der Literatur insbesondere Residuenanalysen vorgeschlagen. Deren Darstellung würde hier zu weit führen, und es sei wieder auf die Lehrbuchliteratur verwiesen (A: 272 f.; BHM v.a. S. 189, 217, 237, 246).

Darstellung 5: *Hazardfunktionen für die Altersgruppe bis 30 (Schätzungen nach Life-Table, Weibull-, log-logistischem (Typ II) und Sichelmodell)*



Zum Schluß ist darauf hinzuweisen, daß hier nur die gebräuchlichsten Verfahren parametrischer Analyse dargestellt werden konnten. So befindet sich unter den hier diskutierten Verfahren keines, welches eine zunächst fallende und dann steigende Hazardrate modellieren kann. Daher sind Versuche zu erwähnen, sehr allgemeine Modelle zu formulieren, welche sowohl die hier vorgestellten Modelle als Spezialfall enthalten als auch weitere Verläufe zu modellieren geeignet sind. Zu nennen sind hier Modelle mit einer verallgemeinerten Gamma-Verteilung oder solche mit einer Box-Cox-Transformation, zu denen in der Lehrbuchliteratur jedoch leider nur kurze Hinweise vorliegen (A: 234 f.).[26] Auch hier gilt die oben geäußerte Warnung: Einerseits erlaubt die hohe Flexibilität der Modelle die Prüfung einer Vielzahl von Hypothesen; andererseits legt gerade dies nahe, mit der Interpretation von Ergebnissen, die nicht aufgrund gezielter Hypothesen, sondern durch explorativ es »Herumprobieren« zustandegekommen sind, vorsichtig zu sein.

Darstellung 6: *Survivorfunktionen für die Altersgruppe bis 30 (Schätzungen nach Life-Table, Weibull-, log-logistischem (Typ II) und Sichelmodell)*



5. Modelle für diskrete Verweildauern

In diesem Abschnitt möchte ich kurz auf Modelle eingehen, die angewendet werden können, wenn *diskrete Verweildauern vorliegen.* Grundsätzlich - vgl. die Ausführungen in Teil I, Abschnitt 2 - ist damit gemeint, daß der interessierende Zustandswechsel nicht jederzeit, sondern nur zu fixen Zeitpunkten stattfinden kann (Wahlen, Versetzungen in der Schule). In der Praxis wird jedoch vielfach auch vorgeschlagen, solche Modelle anzuwenden, wenn es sich um stark gruppierte oder aggregierte Dauern handelt (Hamerle/Tutz 1989). So haben z.B. Licht/Steiner (1991) die hier herangezogenen Arbeitslosigkeitsdaten aus dem SOEP mit einem Modell für diskrete Dauern untersucht, weil es sich bei den monatlichen Messungen um gruppierte Daten handelt.[27]

Zwei relativ einfache und doch ziemlich flexible Modelle basieren auf der *logistischen Verteilung* sowie auf der *komplementären Log-Log-Verteilung* (auch um-

gekehrt als doppelte Exponentialverteilung bezeichnet). Die erste Verteilung ist aus der Analyse binärer abhängiger Variablen mit dem Verfahren der logistischen Regression inzwischen hinlänglich vertraut (Urban 1993). Als einfaches Modell erhalten wir auch hier ein Modell ohne Zeitabhängigkeit:

Logistisches Modell für diskrete Zeit ohne Zeitabhängigkeit:

$$r(t;\mathbf{X}) = \frac{\exp(\beta_0 + \mathbf{X}\boldsymbol{\beta})}{1+\exp(\boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta})} \tag{14}$$

Zeitabhängigkeit des Prozesses kann ganz analog zum Exponentialmodell durch Hinzufügen eines Polynom-Terms für die Zeit modelliert werden:

Logistisches Modell für diskrete Zeit mit Zeitabhängigkeit:

$$r(t;\mathbf{X}) = \frac{\exp(\beta_0 + \mathbf{X}\boldsymbol{\beta} + \beta_{t1}t + \beta_{t2}t^2 + \ldots\beta_{tn}t^n)}{1+\exp(\beta_0 + \mathbf{X}\boldsymbol{\beta} + \beta_{t1}t + \beta_{t2}t^2 + \ldots\beta_{tn}t^n)} \tag{15}$$

Dieses Modell hat den Vorzug, daß es grundsätzlich mit jedem Programm realisiert werden kann, welches ein logistisches Regressionsmodell schätzen kann (vgl. Allison 1982; Yamaguchi 1991, Kap. 2). Alternativ dazu wird auch ein Modell vorgeschlagen, welches auf der doppelten Exponentialverteilung bzw. deren Komplementärwert beruht; daher wird es vielfach als komplementäres Log-Log-Modell bezeichnet. Auch hier lassen sich ein einfaches Modell ohne Zeitabhängigkeit und ein Modell mit Polynom-Term für die Zeitabhängigkeit formulieren:

Komplementäres Log-log-Modell für diskrete Zeit ohne Zeitabhängigkeit

$$r(t;\mathbf{X}) = 1-\exp\{-\exp(\boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta})\} \tag{16}$$

Komplementäres Log-log-Modell für diskrete Zeit mit Zeitabhängigkeit

$$r(t;\mathbf{X}) = 1-\exp\{-\exp(\boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}_{t1}t + \boldsymbol{\beta}_{t2}t^2 + \ldots\boldsymbol{\beta}_{tn}t^n)\} \tag{17}$$

In unserem Beispiel führen beide Modelle zu praktisch identischen Ergebnissen. Daher werden nur die Ergebnisse des logistischen Modells vorgestellt *(Darstellung 7)*. Die Ergebnisse ähneln sehr stark denjenigen, die mit dem Exponentialmodell für stetige Zeiten erhalten wurden. Im einfachen, zeitkonstanten Modell ist auch hier der Parameter für die Altersgruppe der 31- bis 50jährigen beinahe auf dem 5-Prozent-Niveau signifikant.

Darstellung 7: *Ergebnisse logistischer Hazardratenmodelle für diskrete Zeit (erweiterte Beispieldaten aus Teil I, Darstellung 3)*

Variable	Coeff	Error	T-Stat	Signif
Zeitkonstantes Modell (Log-likelihood: -1313.86)				
Konstante	-1.8945	0.0685	-27.6468	1.0000
Alter 31 bis 50	-0.2130	0.1094	-1.9461	0.9484
Alter über 50	-2.3404	0.2410	-9.7105	1.0000
Polynom 1. Grades (Log-likelihood: -1264.08)				
Konstante	-1.4773	0.0812	-18.1974	1.0000
Alter 31 bis 50	-0.0521	0.1113	-0.4679	0.3601
Alter über 50	-2.0359	0.2433	-8.3670	1.0000
$ß_{t1}$	-0.0716	0.0092	-7.8140	1.0000
Polynom 4. Grades (Log-likelihood: -1258.07)				
Konstante	-1.8674189	0.1576898	-11.8423561	1.0000
Alter 31 bis 50	-0.0536346	0.1114339	-0.4813137	0.3697
Alter über 50	-2.0370082	0.2438730	-8.3527419	1.0000
$ß_{t1}$	0.1665113	0.0768530	2.1665921	0.9697
$ß_{t2}$	-0.0302462	0.0098949	-3.0567520	0.9978
$ß_{t3}$	0.0011732	0.0004203	2.7915498	0.9948
$ß_{t4}$	-0.0000136	0.0000055	-2.4630662	0.9862

Log-likelihood des *Grundmodells nur mit Konstante*: -1403.38

Allerdings zeigt dieses Modell wiederum eine deutlich schlechtere Modellanpassung als die zeitabhängigen Modelle, unter denen hier die beiden mit Polynom 1. und 4. Grades dargestellt werden, da diejenigen mit Polynom 2. und 3. Grades keine signifikanten Koeffizienten für die Polynomglieder und auch keinen Erklärungszuwachs erbrachten.

Einige weitere Modelle wie etwa ein Cox-Modell für gruppierte Zeiten sowie ein von Aranda-Ordaz vorgeschlagenes, relativ flexibles Modell werden in der einschlägigen Spezialliteratur diskutiert (Hamerle/Tutz 1989: 31 ff.).

Insgesamt läßt sich festhalten, daß vermutlich bei den hier untersuchten Daten die Meßeinheit für die Verweildauern in Form von Monaten klein genug ist, um sie auch mit Modellen für stetige Zeiten analysieren zu können. Die Tatsache, daß Modelle für diskrete Verweildauern, soweit sie vergleichbar sind, zu den gleichen Schlußfolgerungen führen - zunächst kurzfristig steigende, dann sinkende Hazardrate, kann aber als zusätzliche Abstützung der Ergebnisse gewertet werden. Liegen andere Datenstrukturen vor, insbesondere also noch größere Meßintervalle oder tatsächlich diskrete Ereignisse, sind die hier genannten Modelle eine wichtige alternative Auswertungsmöglichkeit.

6. Zeitabhängige (zeitveränderliche) Kovariaten

In diesem Abschnitt möchte ich noch kurz auf die Möglichkeit eingehen, mit den hier besprochenen Verfahren die Einflüsse von Kovariaten zu untersuchen, die sich selbst während des untersuchten Prozesses ändern. Grundsätzlich ist hierin eine der wichtigsten Anwendungsmöglichkeiten zu sehen, die mit anderen statistischen Verfahren nicht in der gleichen Weise erreicht werden kann. Eine ausführliche Erörterung würde aber den Rahmen dieser Arbeit sprengen, ich will daher nur einige Beispiele ansprechen und kurz auf die praktische Durchführung eingehen.

Zunächst ist auf die wichtige Unterscheidung zwischen *extern* und *intern* zeitabhängigen Kovariaten einzugehen. *Interne* Zeitabhängigkeit meint den Sachverhalt, daß sich Kovariaten in Abhängigkeit von der Dauer des untersuchten Prozesses selbst verändern. So könnten Arbeitslose mit zunehmender Arbeitslosigkeitsdauer immer größere Entmutigungseffekte zeigen.[28] Solche Effekte sind für die Datenauswertung problematisch, weil sie schwer von der Zeitabhängigkeit des Prozesses selbst zu trennen sind, so daß man häufig genötigt ist, die Zeitabhängigkeit einfach festzustellen und entweder ihr aufgrund theoretischer Überlegungen eine substantielle Interpretation zu geben oder zu versuchen, diese substantielle Interpretation durch andere Daten oder Analysemethoden zu erhärten (vgl. zum Beispiel der Arbeitslosigkeit Winter-Ebmer 1992).[29]

Externe Zeitabhängigkeit von Kovariaten liegt vor, wenn sich Werte von erklärenden Variablen zwar *während*, aber (mutmaßlich) *nicht infolge* des untersuchten Prozesses ändern. Ein Beispiel hierfür wäre die *Arbeitsmarktlage*, gemessen in regionalen Arbeitslosigkeitsquoten (vgl. Hujer/Schneider 1987, 1992). Diese ändert sich (praktisch) unabhängig vom Arbeitslosigkeitsschicksal des einzelnen Arbeitslosen. Im konkreten Beispiel ist insbesondere an Saisonalitätseffekte zu denken, d.h., an die Verschlechterung der Arbeitsmarktsituation im Winter und ihre Verbesserung im Frühjahr. Aber auch *Individualdaten* lassen sich als extern zeitveränderlich auffassen, z.B. Daten zur Familiensituation. Es ist bekannt, daß Heirat und Geburt von Kindern sich auf das Erwerbsverhalten insbesondere von Frauen auswirken, und es kann nach bisherigen Analysen davon ausgegangen werden, daß dies auch für die Arbeitslosigkeit zutrifft (Ludwig-Mayerhofer 1990). Man kann sich solche Kovariaten so vorstellen: Es gibt, ähnlich wie in unseren Beispieldaten, verschiedene Gruppen von Arbeitslosen, z.B. verheiratete und ledige Personen, mit je unterschiedlichen Hazardraten. Wenn eine Person während der Arbeitslosigkeit heiratet (oder umgekehrt sich trennt oder scheiden läßt), so heißt das praktisch, daß sie während der Arbeitslosigkeit aus der einen in die andere Gruppe wechselt, d.h., die Risikomenge der einen Gruppe verläßt und - während des Prozesses, also nicht zum Zeitpunkt Null, sondern zum Zeitpunkt des Wechsels! - derjenigen der anderen Gruppen zugeschlagen wird.[30]

Allgemein läßt sich also sagen: Bei der Einbeziehung zeitveränderlicher (oder zeitabhängiger) Kovariaten muß - ganz ähnlich wie bei dem im Vordergrund der

Analyse stehenden Prozeß selbst - bekannt sein, zu welchem Zeitpunkt ein Wechsel in der erklärenden Variablen eintritt, und natürlich, welche Werte diese Variable zu den verschiedenen Zeitpunkten hat. D.h. die zeitabhängigen Kovariaten können ohne weiteres mehrmals oder häufig ihre Werte wechseln, wie es z.B. bei einer Untersuchung der Arbeitslosigkeit der Fall wäre, die sich über mehrere Jahre erstreckt und monatliche Arbeitslosenquoten als zeitveränderliche Kovariate einbezieht. In anderen Fällen geht es möglicherweise nur um einmalige Veränderungen, etwa den Zeitpunkt, zu dem eine Person eine Ausbildung abgeschlossen hat. In solchen Fällen genügt es im allgemeinen, diesen Zeitpunkt als Variable in den Datensatz aufzunehmen und Fälle, bei denen die betreffende Änderung nicht eintritt, mit einem geeigneten Wert zu kodieren.

Wie lassen sich solche Variablen mit den hier untersuchten Modellen analysieren? Im Rahmen des *semiparametrischen Cox-Modells* ist das Verfahren grundsätzlich relativ einfach. Es genügt hier, wie soeben geschildert, die entsprechenden Variablen in einer Form im Datensatz verfügbar zu haben, die eine Beziehung auf die Dauer des untersuchten Prozesses ermöglicht. Wie das konkret geschieht, muß in den einschlägigen Programmhandbüchern und der Lehrbuchliteratur nachgesehen werden (BHM: 155 ff.). Allerdings wird, sofern es sich nicht nur um wenige Kovariaten handelt, die Rechenzeit u.U. überproportional lang, so daß sich bei vielen zeitveränderlichen Kovariaten auch im Cox-Modell ein Vorgehen empfiehlt, das bei den *parametrischen Modellen unumgänglich ist*. Dieses Vorgehen besteht darin, die untersuchten Episoden jeweils an der Stelle, an der die zeitveränderlichen Kovariaten ihren Wert ändern, in Teilepisoden zu zerlegen (»Episodensplitting«). Konkret: Wenn ein Individuum z.B. eine Arbeitslosigkeitsdauer von 10 Monaten aufweist und nach 5 Monaten heiratet, so müssen aus der entsprechenden Episode zwei Episoden gemacht werden: Die erste mit einer Dauer von 0 bis 5, die zweite mit einer Dauer von 5 bis 10. Die betreffende zeitabhängige Kovariate müßte in den beiden Teilepisoden Werte aufweisen, die den Zustandswechsel wiedergeben (also z.B. 0 für ledig und 1 für verheiratet), die zeitkonstanten Kovariaten müßten für beide Teilepisoden identisch sein. Die erste Teilepisode wäre als zensiert zu betrachten, da die Person ja nach 5 Monaten noch arbeitslos ist. Ein solches Episodensplitting ist entgegen der Darstellung von BHM (193 ff., v.a. 196) auch problemlos mit SPSS oder vergleichbaren Programmen durchführbar.[31] Zu beachten ist, daß nicht alle Programme in der Lage sind, entsprechende Datensätze auszuwerten; Voraussetzung ist, daß sie andere Anfangszeiten als »0« zulassen.[32] - Die Analyse mit parametrischen Modellen hat natürlich den Vorteil, daß hier zusätzlich die Zeitabhängigkeit der Hazardrate selbst modelliert werden kann.

Abschließend ist darauf hinzuweisen, daß man auf diese Weise auch die *wechselseitige Beeinflussung* von Prozessen untersuchen kann. Wenn wir hier das Beispiel angesprochen haben, daß die Geburt von Kindern die Arbeitslosigkeitsverläufe verändern kann, so könnte man auch der Frage nachgehen, ob umgekehrt die Arbeitslosigkeit von Frauen die Neigung beeinflußt, Kinder zu bekommen - wobei

die Wirkungsrichtung auch vom Familienkontext abhängen könnte, d.h., bei einer günstigen ökonomischen Situation könnte die Arbeitslosigkeit möglicherweise als Gelegenheit gesehen werden, einen Kinderwunsch zu realisieren, während umgekehrt eine ohnehin ungünstige Situation durch die eigene Arbeitslosigkeit noch verschlechtert würde und dadurch eher einem Kinderwunsch entgegenwirken könnte. Entsprechende Analysen für den Zusammenhang von Kindern und Unterbrechungen der Erwerbstätigkeit hat z.B. Huinink (1991, 1992) vorgestellt. Grundsätzlich sind solche Analysen wiederum, wie Huinink in den genannten Arbeiten verdeutlicht, vor allem dann unproblematisch, wenn *keine Dauerabhängigkeit* vorliegt, d.h., wenn z.B. zwar die Arbeitslosigkeit an sich, aber nicht ihre Dauer, sich auf die Geburt von Kindern auswirkt (und umgekehrt). Auch hier steht aber die Entwicklung von statistischen Modellen erst in den Anfängen.

7. Mehrere Zielzustände

Bislang wurden nur Übergänge in einen einzigen Zielzustand untersucht, in unserem Beispiel von der Arbeitslosigkeit in die Vollzeitbeschäftigung. Hier soll kurz darauf eingegangen werden, wie zu verfahren ist, wenn Übergänge in mehrere Zielzustände möglich sind (vgl. A: 77 ff.; BHM: 59 ff., 78 ff., 133 ff., 164 ff.; DM: 51 f., 174 ff.).

Entscheidend ist hier die Annahme, daß die verschiedenen Risiken voneinander *unabhängig* sein sollten. Diese Annahme zeigt eine deutliche Analogie zur Forderung nach der Unabhängigkeit von Zensierungen und Ereignissen, und tatsächlich besteht ein unmittelbarer Bezug dazu. Denn, wie schon in Teil I kurz angedeutet, verfährt man bei der Analyse mehrerer Zielzustände so, daß die Übergänge in die einzelnen Zielzustände separat analysiert werden,[33] wobei jeweils sämtliche anderen Zielzustände als rechtszensierte Daten betrachtet werden (so wie wir auch in unserer Beispielsanalyse immer mit den Übergängen in andere Zustände verfahren sind). Den Analysen für die einzelnen Zielzustände können dabei ohne weiteres verschiedene Modelle des Verlaufs der Hazardfunktion zugrundegelegt werden, also z.B. ein zeitkonstantes Exponentialmodell für den Übergang in Zustand A, ein Weibull-Modell für den Übergang in Zustand B, usw.

Die Annahme der Unabhängigkeit der verschiedenen Zielzustände bzw. Übergänge ist nicht immer plausibel. Auch in unserem Beispiel müssen wir davon ausgehen, daß z.B. Personen mit geringen Chancen einer (Wieder-)Beschäftigung eine höhere Wahrscheinlichkeit aufweisen, aus dem Arbeitslosenbestand auszuscheiden und den Arbeitsmarkt ganz zu verlassen. Das kann dazu führen, daß die Beschäftigungschancen u.U. sogar überschätzt werden. Insofern müssen unsere Ergebnisse mit einer gewissen Vorsicht betrachtet werden.

Eine ausführliche Darstellung der Probleme bei voneinander abhängigen Risiken findet sich bei Klein (1988). Die Diskussion über Möglichkeiten der Schätzung

von Modellen bei abhängigen Risiken hat bislang noch keine leicht anwendbaren Verfahren erbracht (Hinweise hierzu etwa bei Schneider/Hujer 1992, Fn. 8).

8. Karrieren und wiederholte Episoden

Bislang wurden sehr einfache Modelle geschätzt, und das Versprechen einer Analyse von »Karrieren«, wie es in der Einleitung zu Teil I angedeutet wurde, läßt sich damit sicher nur begrenzt einlösen. In diesem Abschnitt sollen kurz einige Möglichkeiten angesprochen werden, komplexere Abfolgen von Episoden zu untersuchen, die sich gegebenenfalls im Sinne von Karrieren interpretieren lassen. D.h., es geht darum, *mehrere Episoden* im Lebenslauf von Individuen in geeigneter Weise aufeinander zu beziehen.

Eine erste Analysemöglichkeit könnte darin bestehen, *verschiedenartige Zustände bzw. Übergänge* zu untersuchen, aber jeweils zu fragen, ob Merkmale des vorangegangenen Zustandes bzw. Prozesses (gegebenenfalls auch mehrerer Zustände oder Prozesse) den aktuellen Prozeß beeinflussen. So könnte man fragen, ob z.B. die Arbeitslosigkeitsdauer durch Merkmale des vorherigen Zustandes beeinflußt wird, ob sich also Merkmale aus der Beschäftigung (oder einem anderen Zustand) *vor* der Arbeitslosigkeit *in* dieser auswirken. Umgekehrt ließe sich fragen, ob die Art oder Dauer von Beschäftigungsverhältnissen davon abhängt, ob oder wie lange Individuen vorher arbeitslos waren. So zeigte sich in einer Untersuchung an Berufsanfängern (Ludwig-Mayerhofer 1992b), daß eine ganz lange Arbeitslosigkeit (über 12 Monate) vor der ersten Beschäftigung zu einem geringerem Anfangseinkommen führt. Weiterhin wurde gefragt, ob Arbeitslosigkeit *vor* der ersten Beschäftigung auch das Risiko *erneuter* Arbeitslosigkeit erhöht. Tatsächlich ist ein solcher Effekt, jedenfalls als direkter, nicht oder kaum zu beobachten; der entscheidende Effekt hinsichtlich des Arbeitslosigkeitsrisikos ist vielmehr das Einkommen aus dem aktuellen Beschäftigungsverhältnis, wodurch immerhin ein indirekter Effekt einer früheren Arbeitslosigkeit denkbar ist.

Bei solchen Analysen ist grundsätzlich nicht anders vorzugehen als bisher geschildert, da jeweils nur ein Zustandswechsel untersucht wird und die Merkmale aus vorangegangen Episoden oder Zuständen als Kovariaten eingesetzt werden. Anders ist dies, wenn man *wiederholte* Episoden der *gleichen Art* untersucht, wenn also z.B. in einer Stichprobe - wie auch in den Daten aus dem SOEP - Individuen mit mehreren (hier: Arbeitslosigkeits-)Episoden enthalten sind. Würde man diese mit einem der gängigen Modelle untersuchen, ohne das mehrfache Auftreten ein und derselben Person zu berücksichtigen, wäre die Annahme der Unabhängigkeit der einzelnen Beobachtungen voneinander verletzt. Es ist also grundsätzlich geboten, zwischen verschiedenen Episoden ein und desselben Individuums zu unterscheiden, zumal sich die Dauer und/oder die Einflüsse von Kovariaten zwischen verschiedenen Episoden unterscheiden können. Zudem wird es auch in diesem Fall

ratsam sein, Aspekte des Verlaufs vor der jeweiligen Episode (der »Vorgeschichte«) einzubeziehen, insbesondere bei den wiederholten Arbeitslosigkeitsepisoden Merkmale der früheren Arbeitslosigkeitsepisoden (vgl. grundsätzlich Hamerle 1989 sowie BHM: 62 ff., Diekmann/Mitter 1993: 56 f.).

Im Rahmen dieser Arbeit ist natürlich keine umfassende Diskussion möglich. Ich will aber die jedenfalls grundsätzlich vorhandene Fruchtbarkeit einer solchen Analyse anhand der Ergebnisse eines einfachen Modells zeigen. Dieses geht in zwei Hinsichten über die bisherigen Auswertungen hinaus: Erstens werden bis zu *vier* Arbeitslosigkeitsepisoden jeder Person analysiert, und zweitens werden die Einflüsse dreier zusätzlicher Variablen aus der »Vorgeschichte« untersucht: der Dauer der jeweils vorangegangenen Arbeitslosigkeitsepisode, der Zeit zwischen dem Ende der vorangegangenen und dem Beginn der gegenwärtigen Arbeitslosigkeitsepisode, und des Zustands, in den die vorangegangene Arbeitslosigkeitsepisode mündete, hier nur dichotomisiert nach Voll- und Teilzeitbeschäftigung vs. alle anderen Zustände (Haushalt, Ausbildung, Bundeswehr usw.). Zugrundegelegt wird ein log-logistisches Modell vom Typ I (*Darstellung 8*).[34] Selbstverständlich ließen sich noch komplexere Einflüsse heranziehen (z.B. bei der dritten Episode die Dauer der ersten *und* der zweiten Episode, usw.).

Vorbehaltlich der Tatsache, daß man die Ergebnisse wegen des Fehlens weiterer relevanter Variablen und angesichts der vor allem für die dritte und vierte Episode schon recht kleinen Fallzahlen nicht überbewerten sollte, zeigt sich, daß es zum einen durchaus Unterschiede zwischen den verschiedenen Episoden und zum anderen auch Einflüsse der Vorgeschichte geben könnte. So ist nach den Ergebnissen bei der zweiten bis vierten Episode auch eine Verschlechterung der Wiederbeschäftigungschancen für die 31- bis 50jährigen festzustellen (über 50jährige haben so wenige dritte und vierte Arbeitslosigkeitsepisoden, daß die Schätzungen sehr unzuverlässig werden). Ferner zeigt sich - teilweise nur als Tendenz -, daß mit längerer Dauer der vorangegangenen Arbeitslosigkeitsepisode auch die Hazardrate für die gegenwärtige Episode zurückgeht und daß die Personen, die schon früher nach der Arbeitslosigkeit direkt in ein Beschäftigungsverhältnis übergingen, bei einer späteren Arbeitslosigkeit wiederum bessere Wiederbeschäftigungschancen haben. Nur der Einfluß der Dauer seit der letzten Arbeitslosigkeit ist inkonsistent.[35]

Die Ergebnisse sind grundsätzlich die gleichen, die man erhalten würde, wenn man jeweils separate Modelle für die erste, zweite, dritte und vierte Arbeitslosigkeitsepisode schätzen würde (so sind die Ergebnisse für die erste Episode identisch mit jenen aus Darstellung 4). Der Vorteil der simultanen Schätzung in einem Modell ist darin zu sehen, daß geprüft werden kann, ob sich die Parameter für die einzelnen Episoden signifikant voneinander unterscheiden. So erbringt eine Modellschätzung unter der Annahme, daß die vier Modellkonstanten für die Basisrate miteinander identisch sind, ein Modell mit einer Log-Likelihood von 1919,61.

Darstellung 8: *Ergebnisse logistischer Hazardratenmodelle (Typ I) für kontinuierliche Zeiten, wiederholte Arbeitslosigkeitsepisoden*

Variable	Coeff	Error	T-Stat	Signif
1. Episode				
Konstante	-1.3210	0.0731	-18.0728	1.0000
Alter 31 bis 50	-0.0876	0.1199	-0.7306	0.5350
Alter über 50	-2.3385	0.2207	-10.5946	1.0000
Konstante	0.2925	0.0404	7.2349	1.0000
2. Episode (214 Episoden mit 158 Übergängen)				
Konstante	-1.1948	0.2720	-4.3929	1.0000
Alter 31 bis 50	-0.6784	0.1874	-3.6204	0.9997
Alter über 50	-1.1617	0.3588	-3.2377	0.9988
Dauer d. vorh. Arbeitslosigk.	-0.0266	0.0205	-1.2996	0.8063
Zeit seit vorher. Arbeitslosigk.	-0.0076	0.0082	-0.9249	0.6450
Zielzustand d. vorher. Arblk.	0.5627	0.2330	2.4151	0.9843
ln(p)	0.3861	0.0651	5.9331	1.0000
3. Episode (96 Episoden mit 68 Übergängen)				
Konstante	-1.1297	0.6889	-1.6399	0.8990
Alter 31 bis 50	-0.3633	0.2118	-1.7157	0.9138
Alter über 50	-0.3835	0.3521	-1.0891	0.7239
Dauer d. vorh. Arbeitslosigk.	-0.0872	0.0243	-3.5934	0.9997
Zeit seit vorher. Arbeitslosigk.	-0.0229	0.0121	-1.8993	0.9425
Zielzustand d. vorher. Arblk.	0.6626	0.6815	0.9723	0.6691
ln(p)	0.6880	0.1000	6.8781	1.0000
4. Episode (43 Episoden mit 28 Übergängen)				
Konstante	-2.2347	0.5389	-4.1465	1.0000
Alter 31 bis 50	-0.6374	0.3228	-1.9749	0.9517
Alter über 50	-0.5415	0.4276	-1.2665	0.7947
Dauer d. vorh. Arbeitslosigk.	-0.1081	0.0391	-2.7646	0.9943
Zeit seit vorher. Arbeitslosigk.	0.0837	0.0231	3.6278	0.9997
Zielzustand d. vorher. Arblk.	1.0436	0.4963	2.1028	0.9645
ln(p)	0.7917	0.1567	5.0514	1.0000

Log likelihood: -1917.96
Log-likelihood des Null-Modells (Exponentialmodell mit 4 Konstanten): -2124.65

Der Likelihood-Ratio-Test ergibt also im Vergleich zum oben dargestellten Modell einen χ^2-Wert von 3,30, der bei 3 Freiheitsgraden (es wird im Vergleich zu vorher nur mehr eine einzige Regressionskonstante geschätzt) nicht signifikant auf dem 5-Prozent-Niveau ist. D.h., wir können - auf der Basis unseres unvollständigen Modells - nicht annehmen, daß wiederholte Arbeitslosigkeitsepisoden ceteris paribus länger sind als frühere.

Wie an diesem kleinen Beispiel schon deutlich geworden sein könnte, lassen sich durch die Verknüpfung wiederholter Episoden mit Daten aus anderen Prozes

sen oder Zuständen unter Umständen sehr komplexe Prozesse modellieren. Die Forschungspraxis beschränkt sich bislang allerdings ganz weitgehend auf Ein-Episoden-Modelle, und die Analyse komplexer Verlaufsmuster wurde noch kaum in Angriff genommen. Hier stehen für künftige Untersuchungen noch erhebliche Potentiale offen, bei denen allerdings auch ein beträchtlicher Datenerhebungsaufwand erforderlich ist.

9. Abschließende Bemerkungen

Ich hoffe, in dieser Arbeit aufgezeigt zu haben, daß die Verweildaueranalyse für Längsschnittuntersuchungen ganz erhebliche und wichtige Analysemöglichkeiten bietet, die weit über die früher verfügbaren Verfahren hinausgehen. Gleichzeitig hoffe ich, die Grundlagen dieser Modelle soweit erläutert zu haben, daß damit ein Verständnis einschlägiger Forschungsresultate ebenso möglich ist wie - in Verbindung mit der Lehrbuchliteratur - ein Einstieg in eigene Datenauswertungen.

Ich will abschließend noch einmal daran erinnern, daß der Einsatz der hier besprochenen Verfahren stets in Abhängigkeit von der konkreten Fragestellung und den verfügbaren Daten erfolgen muß. Verschiedene inhaltlich durchaus wichtige Probleme sind noch nicht oder jedenfalls nicht definitiv gelöst, etwa in der Analyse voneinander abhängiger konkurrierender Risiken oder in der simultanen Analyse von sich wechselseitig beeinflussenden Prozessen. Allerdings kann das kein Einwand gegen die Anwendung der hier diskutierten Verfahren sein, im Gegenteil. Es muß vielmehr hervorgehoben werden, daß erst die Beschäftigung mit diesen Verfahren die zugrundeliegenden Probleme ins Bewußtsein gehoben hat, die andernfalls vielleicht mit gänzlich unzulänglichen statistischen Mitteln angegangen worden wären, mit der Folge von gravierenden Methodenartefakten. Notwendig ist also, die Verfahren in dem Bewußtsein einzusetzen, daß zwar manche Probleme noch offen sind, daß dem aber nicht durch Abstinenz, sondern durch möglichst sachgerechten Einsatz der Verfahren und die daraus resultierende Akkumulation von Erfahrungen abgeholfen werden kann.

Selbstverständlich soll auch nicht der Eindruck erweckt werden, daß keine anderen Verfahren für Längsschnittuntersuchungen sinnvoll seien. Methoden zur Panelanalyse für metrische Variablen sind längst etabliert (Arminger/Müller 1990). Auch in der Panelanalyse von diskreten (bislang allerdings nur binären oder ordinalen) Variablen wurden in den letzten Jahren erhebliche Fortschritte erzielt (Andreß 1992b; Maddala 1987; Petersen 1993; Schneider/Hujer 1992).

* * * * *

Anhang

Programme zur Verlaufsdatenanalyse.

Wie schon in Teil I erwähnt, stellt TDA das umfassendste Programm zur Analyse von Verlaufsdaten dar. Unter anderem ist die simultane Analyse wiederholter Episoden nur in diesem Programm als Standardprozedur verfügbar, sieht man von einzelnen schwer zugänglichen und sehr speziellen Programmen ab. Auch Panelmodelle für diskrete Daten sind in TDA implementiert. Einzig SAS bietet unter den verbreiteten Programmen eine annähernd große Vielfalt an Verfahren. In BMDP sind immerhin die gängigsten Modelle für kontinuierliche Verweildauern implementiert, während SPSS zur Zeit nur Prozeduren für nicht- und semiparametrische Verfahren enthält, und auch dies nur in der Windows-Version.

Unter den weniger verbreiteten Programmen sind noch zu nennen: LIMDEP (Greene 1992) und PARAT (Schneider 1991), die verschiedene Modelle für kontinuierliche Verweildauern schätzen können, GLAMOUR mit Verfahren für diskrete Verweildauern (Tutz/Georg 1991) sowie EGRET (Statistics and Epidemiology Research Corporation 1991), das inbesondere für Mediziner und Epidemiologen von Interesse ist, allerdings neben nicht- und semiparametrischen Verfahren nur wenige Modelle für kontinuierliche Dauern enthält. Mit dem Programm GLIM können sehr viele Modelle durch benutzerdefinierte Routinen geschätzt werden.

Nur am Großrechner verfügbar ist das Programm RATE (Tuma 1980), welches einmal eines der wichtigsten Programme war (vgl. die Beispiele in BHM), inzwischen aber durch einige andere Programme überholt ist.

Bei allen Programmen ist vor allem bei den multivariaten Verfahren sehr genau darauf zu achten, in welcher Art und Weise die verschiedenen Modelle formuliert (»parametrisiert«) werden. Beispielsweise verwenden BMDP und SAS vielfach Parametrisierungen, die stark von den hier (in Anlehnung vor allem an die deutschsprachige Lehrbuchliteratur) vorgestellten Formulierungen abweichen. Die angegebenen Modellparameter haben teilweise genau das umgekehrte Vorzeichen, so daß ein Effekt mit positivem Vorzeichen tatsächlich einer Verlängerung der Verweildauer, also einer Verringerung der Hazardrate entspricht. Auch die Parameter für die Zeitbezogenenheit der Übergangsrate müssen zum Teil umgerechnet werden (die Formulierungen entsprechen offenbar überwiegend denjenigen bei Kalbfleisch/Prentice 1980, vor allem 24 ff.).

Anmerkungen

1 Es gibt allerdings – soviel im Vorgriff – Techniken, die »Basisrate« dennoch zu schätzen; dies stellt jedoch einen zweiten Schritt zusätzlich zur Schätzung der Modellparameter dar.

2 Ein Beispiel, wo die »Hazardkomponenten« explizit vorgestellt werden, findet sich bei Kiefer (1988).

3 Zur Klarstellung sei darauf hingewiesen, daß sich das Alter in den später vorgestellten multivariaten Modellen auch als metrische Variable behandeln ließe. Für die hier diskutierten nicht-parametrischen Verfahren ist es aber immer erforderlich, metrische Variable zu gruppieren. Allerdings ist dies im Sinne einer explorativen Datenanalyse durchaus nicht nur ein notwendiges Übel, sondern sogar sehr empfehlenswert, um etwaigen nicht-linearen Zusammenhängen auf die Spur zu kommen.

4 Die von SPSS/PC$^+$ (Version 4) ausgegebene Lee-Desu-Statistik führt praktisch zu identischen Ergebnissen wie die Gehan/Breslow-Statistik.

5 Fettgedruckte Buchstaben verweisen darauf, daß sich bei den Größen um Matrizen bzw. Vektoren handelt.

6 Anschauliche Beispiele aus dem Bereich der ML-Schätzung für kategoriale Daten finden sich bei Maier/Weiss (1990: 80 ff.) und Urban (1993: 53 ff.). Konkrete Beispiele für die Verweildaueranalyse finden sich vor allem bei A: 191 ff. zu ML- und 241 ff. zu PL-Schätzungen.

7 Auf die naheliegende Frage, welcher Stichprobenumfang als »hinreichend« gelten kann, findet sich in der einschlägigen Literatur keine klare Antwort. Aus verschiedenen Simulationsstudien, allerdings mit meist einfachen Datensätzen mit nur einer oder zwei Kovariaten, läßt sich ersehen, daß bereits Stichprobenumfänge von 50 bis 100 Fällen zu einigermaßen zuverlässigen Ergebnissen führen (vgl. A: 203 sowie Tuma 1982). Ganz grundsätzlich läßt sich sagen (was aber für alle multivariaten Verfahren gilt), daß natürlich bei kleinen Stichprobenumfängen die Zahl der geprüften Kovariaten ebenfalls nur sehr klein sein sollte. Studien, von denen gelegentlich berichtet wird, in denen die Zahl der geprüften Variablen kaum kleiner oder gar größer ist als der Stichprobenumfang, können nur als unsinnig bezeichnet werden. – Ebenso relevant sind Fragen der Teststärke, also der Vermeidung von ß-Fehlern. Hierzu gibt es leider noch weniger Aufschluß. (Zu nicht-parametrischen Verfahren vgl. die Arbeit von Freedman 1982.)

8 Allerdings legt das die Gefahr nahe, diese Signifikanzniveaus einfach zu übernehmen. Tatsächlich beziehen sich diese immer auf zweiseitige Signifikanztests. Liegen gerichtete Hypothesen vor, müßte man sich an den Werten für einseitige Signifikanztests orientieren (also z.B. 1,645 für ein Signifikanzniveau von 0,05).

9 Der Ausdruck rührt daher, daß Formel 16 identisch ist mit der Formel

$$-2 \ln \frac{L_1}{L_0} \qquad (3),$$

wobei L_0 und L_1 der (nicht logarithmierten) Likelihood entspricht. Grundsätzlich geben aber alle Programme die Log-Likelihood aus.

10 Dies entspricht also etwa dem globalen F-Test eines linearen Regressionsmodells.

11 Grundsätzlich ist dies auch mit der Wald-Statistik möglich, allerdings ist diese in den meisten Programmen nur als Test implementiert, ob einzelne Koeffizienten von Null verschieden sind.

12 Zu beachten ist, daß für die folgenden Beispiele jeweils der gesamte Datensatz herangezogen wurde, also auch die Verweildauern, die mehr als 12 Monate betrugen. Daher führt ein Nachrechnen anhand von Darstellung 3 aus Teil I zu anderen Ergebnissen!

13 Für die Schätzungen wurde das Programm TDA verwendet. Dieses gibt den (auf 4 Stellen gerundeten) Wert 1 - p des Signifikanzniveaus aus, d.h., der Wert von 1,0000 in der Spalte „Signif“ bedeutet, daß das Signifikanzniveau unter 0,00005 liegt.

14 Beim Nachrechnen ergeben sich hier wie anderswo wegen Rundungen leichte Abweichungen.

15 Diese Schlußfolgerung gilt natürlich nur im Rahmen unsers einfachen Beispiels. Es könnte sein, daß sich durch Einbeziehen weiterer wichtiger Variablen die Ergebnisse auch hinsichtlich des Alters ändern!

16 Inzwischen können - mit dem Programm TDA - Modelle mit unbeobachteter Heterogenität nicht für das Exponentialmodell (A: 266 ff.; BHM: 251 ff.), sondern auch für zahlreiche andere Modelle geschätzt werden.

17 Das Exponentialmodell hat allerdings nicht deshalb seinen Namen, sondern weil es unterstellt, daß S(t) exponetiell mit der Zeit abnimmt.

18 Es ist noch einmal daran zu erinnern, daß die hier vorgestellten Ergebnisse mit dem Datensatz über die gesamte Beobachtungsdauer berechnet wurden, also nicht nur mit den Daten für die ersten 12 Monate, die in Darstellung 3 enthalten sind. Zieht man nur diese Daten heran, so erhält man teilweise konträre Ergebnisse. Insbesondere würden diejenigen Verfahren, die nur monoton steigende oder fallende Verläufe modellieren (vgl. die folgende Darstellung von Weibull- und Gompertz-Verteilung), nicht zu den hier berichteten Ergebnissen einer fallenden, sondern zu einer steigenden Hazardrate führen. Die Ergebnisse werden also nicht unerheblich von der Untersuchungsdauer beeinflußt! Dieser Sachverhalt könnte auch die unterschiedlichen Ergebnisse von Klein (1990) und Ludwig-Mayerhofer (1992a) erklären, da in der erstgenannten Arbeit nur vier, in der zweiten jedoch sechs Wellen des Sozio-ökonomischen Panels herangezogen wurden (vgl. hierzu auch Hujer/Schneider 1992: 328).

19 Zu diesen beiden Modellen gibt es in den deutschen Lehrbüchern leider nur eine kurze Fundstelle (A: 231 ff.). Das in BHM (205 ff.) vorgestellte Modell mit periodisierter Verweildauer ist eine noch umfassendere Ausweitung; es werden nicht nur periodenspezifische Konstanten, sondern auch periodenspezifische Koeffizienten für die Kovariateneinflüsse geschätzt.

20 Aus Gründen der Übersichtlichkeit ist die Hazardrate für das Piecewise Constant Exponentialmodell nicht aufgeführt.

21 In den Lehrbüchern werden die einzelnen Bestandteile - man ist geneigt zu sagen: wie könnte es auch anders sein - unterschiedlich bezeichnet: Der hier nach DM mit „p“ bezeichnete Parameter heißt bei BHM „α„ und bei A „γ„. Zu beachten ist, daß die bei A: 233 angegebene Formel 5.58 mißverständlich ist, da es dort den Anschein hat, als wäre γ ein Index zu λ!

22 Dieser Verteilung sehr ähnlich ist die log-normale Verteilung, auf die ich nicht weiter eingehen will (ebenfalls nur kurze Darstellungen bei A: 74, 292 f.; BHM: 54 f.).

23 Diese Bezeichnungen lehnen sich an das Programm TDA an, in dem auch diese beiden Modelle implementiert sind.

24 Es ist auch denkbar, Kovariaten-Einflüsse hinsichtlich des γ-Terms zu schätzen, was zu einem relativ komplexen Modell führt. Daher verzichte ich auf eine Darstellung, zumal die Einflüsse im vorliegenden Fall wiederum nicht signifikant waren.

25 Die Schätzungen des Gompertz-Modells sind in den folgenden Abbildungen nicht enthalten, weil sie sich kaum von denen des Weibull-Modells unterscheiden. Das log-logistische Modell Typ I verläuft ähnlich dem Modell Typ II, aber in den ersten Monaten deutlich flacher.

26 Allerdings können diese Modelle mit dem Programm TDA geschätzt werden; das Manual des Programms enthält auch Erläuterungen zu den Modellen.

27 Man könnte sogar argumentieren, daß es sich um diskrete Daten im Wortsinn handelt, weil die meisten Entlassungen und Einstellungen jeweils zu Monatsende bzw. Monatsanfang erfolgen.

28 Ein weiteres,.oft in Zusammenhang mit Arbeitslosigkeit diskutiertes Beispiel ist der Bezug von Arbeitslosenunterstützung: Der Bezug von Arbeitslosen*hilfe* tritt sehr häufig erst nach Ablaufen des Arbeitslosen*geldes* ein, d.h., Arbeitslosenhilfe ist fast per definition mit längeren Arbeitslosigkeitsdauern verbunden. Ob und wie diese Problematik zu lösen ist, ist in der Literatur umstritten (vgl. unter anderem Hunt 1992).

29 Für eine ausführliche, aber wiederum nicht einfache Diskussion sei auf die Arbeiten von Petersen (1986a, b) verwiesen.

30 Man kann sich dies sogar an den Beispieldaten vorstellen: Da die Personen während des Arbeitslosigkeitsprozesses selbst älter werden, können sie auch von der einen Altersgruppe in die andere wechseln. Bei dieser Konzeption ist allerdings schon wieder fraglich, ob es sich tatsächlich um eine externe zeitveränderliche Kovariate handelt.

31 Eine ausführliche Beschreibung des Vorgehens mit SPSS/PC$^+$ bzw. SPSS for Windows, das aber auch auf die meisten anderen Programme übertragbar sein dürfte, findet sich in Brüderl/Ludwig-Mayerhofer 1994). Die Lehrbuchliteratur enhält leider keine geeigneten Darstellungen; das Beispiel von BHM (195 f.) ist - vorsichtig formuliert - mißverständlich.

32 Das trifft z.B. für BMDP oder SPSS nicht zu. Besonders hinzuweisen ist wieder auf das Programm TDA, das eine eingebaute Funktion zum Episodensplitting enthält, die allerdings nicht ganz einfach zu handhaben und nicht sehr gut dokumentiert ist.

33 Mit dem Programm TDA ist es möglich, mehrere Übergänge in einem einzigen Modell simultan zu schätzen. Der Vorteil ist, daß es auf diese Weise auch möglich ist, Unterschiede zwischen den Parametern für die verschiedenen Modelle zu testen. Der Nachteil ist, daß in diesem Fall allen Übergängen die gleiche funktionale Form der Hazardrate zugrundegelegt wird.

34 Ein Modell vom Typ II konnte nicht geschätzt werden, da der Schätzalgorithmus - trotz Vergabe von Startwerden - nicht konvergierte. Dies sei wiederum als Hinweis darauf gegeben, daß mit zunehmender Komplexität der Modelle auch die Schwierigkeiten mit ihrer Handhabung zunehmen.

35 Da die auf die Dauer in der Vorgeschichte bezogenen Variablen eine ziemlich schiefe Verteilung aufweisen, wäre hier auch daran zu denken, Transformationen dieser Variablen (z.B. den Logarithmus) zu verwenden. An der je nach Episode unterschiedlichen Wirkung der letztgenannten Variablen würde dies im konkreten Fall jedoch nichts Grundsätzliches ändern.

Literatur

Allison, P. D., 1982: Discrete-Time Methods for the Analysis of Event Histories. S. 61-98 in: Leinhardt, S. (Hrsg.), Sociological Methodology 1982. San Francisco: Jossey-Bass.

Andreß, H.-J., 1988: Spezifikationsfehler und unbeobachtete Heterogenität in Regressionsmodellen für Übergangsraten. Kölner Zeitschrift für Soziologie und Sozialpsychologie 40: 93-116.

Andreß, H.-J., 1992a: Verlaufsdatenanalyse (Historical Social Research/Historische Sozialforschung, Supplement/Beiheft No. 5). Köln: Zentrum für Historische Sozialforschung.

Andreß, H.-J., 1992b: Logistische Regressionsmodelle für Paneldaten. Analyse dichotomer Variablen im Zeitverlauf unter besonderer Berücksichtigung unbeobachteter Heterogenität. S. 35-66 in: Andreß, H.-J./Huinink, J./Meinken, H./Rumianek, D./Sodeur, W./Sturm, G. (Hrsg.), Theorie, Daten, Methoden. Neue Modelle und Verfahrensweisen in den Sozialwissenschaften. München: R. Oldenbourg.

Arminger, G., 1984: Modelltheoretische und methodische Probleme bei der Analyse von Paneldaten mit qualitativen Variablen. Vierteljahreshefte zur Wirtschaftsforschung: 470-479.

Arminger, G./Müller, F., 1990: Lineare Modelle zur Analyse von Paneldaten. Opladen: Westdeutscher Verlag.

Blossfeld, H. P./Hamerle, A./Mayer, K. U., 1986: Ereignisanalyse. Frankfurt/New York: Campus.

Brüderl, J./Diekmann, A., 1995: The Log-Logistic Rate Model. Two Generalizations with an Application to Demographic Data. Sociological Methods & Research 1995 (im Erscheinen).

Brüderl, J./Ludwig-Mayerhofer, W., 1994: Aufbereitung von Verlaufsdaten mit zeitveränderlichen Kovariaten mit SPSS. ZA-Information 34: 79-105.

Cox, D. R., 1972: Regression Models and Life-tables (with discussion). Journal of the Royal Statistical Society, Series B, 34: 187-220.

Diekmann, A./Klein, T., 1991: Bestimmungsgründe des Ehescheidungsrisikos. Kölner Zeitschrift für Soziologie und Sozialpsychologie 43: 271-290.

Diekmann, A./Mitter, P., 1984: Methoden zur Analyse von Zeitverläufen. Stuttgart: Teubner.

Diekmann, A./Mitter, P., 1993: Methoden der Ereignisanalyse in der Bevölkerungssoziologie: Stand und Probleme. S. 20-65 in: Diekmann, A./Weick, S. (Hrsg.), Der Familienzyklus als sozialer Prozeß. Bevölkerungssoziologische Untersuchungen mit den Methoden der Ereignisanalyse. (Sozialwissenschaftliche Schriften, Heft 26). Berlin: Duncker & Humblot.

Freedman, L. S., 1982: Tables of the Number of Patients Required in Clinical Trials Using the Logrank Test. Statistics in Medicine 1: 121-129.

Galler, H. P./Pötter, U., 1987: Unobserved Heterogeneity in Models of Unemployment Duration. S. 628-650 in: Mayer, K. U./Tuma, N. B. (Hrsg.), Applications of Event History Analysis in Life Course Research. (Materialien aus der Bildungsforschung, Vol. 30). Berlin: Max-Planck-Institut für Bildungsforschung.

Galler, H. P./Pötter, U., 1992: Zur Robustheit von Schätzmodellen für Ereignisdaten. S. 379-405 in: Hujer, R./Schneider, H./Zapf, W. (Hrsg.), Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel. Frankfurt/New York: Campus.

Greene, W. E., 1992: LIMDEP, Version 6.0. New York: Econometric Software.

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 Überlebenszeiten. Frankfurt/New York: Campus.

Huinink, J., 1991: The Analysis of Interdependent Social Processes - The Example of Life-Course Analysis. S. 601-615 in: Albrecht, G./Otto, H. U. (Hrsg.), Social Prevention and the Social Sciences. Theoretical Controversies, Research Problems, and Evaluation Strategies. Berlin/ New York: de Gruyter.

Huinink, J., 1992: Die Analyse interdependenter Lebensverlaufsprozesse. Zum Zusammenhang von Familienbildung und Erwerbstätigkeit bei Frauen. S. 343-366 in: Andreß, H.-J./Huinink, J./ Meinken, H./Rumianek, D./Sodeur, W./Sturm, G. (Hrsg.), Theorie, Daten, Methoden. Neue Modelle und Verfahrensweisen in den Sozialwissenschaften. München: R. Oldenbourg.

Hujer, R./Schneider, H., 1987: Ökonometrische Ansätze zur Analyse von Paneldaten: Schätzung und Vergleich von Übergangsratenmodellen. S. 219-242 in: Krupp, H.-J./Hanefeld, U. (Hrsg.), Lebenslagen im Wandel: Analysen 1987. Frankfurt/New York: Campus.

Hujer, R./Schneider, H., 1992: Strukturelle und institutionelle Determinanten der Arbeitslosigkeit aus mikroanalytischer Sicht. S. 315-341 in: Hujer, R./Schneider, H./Zapf, W. (Hrsg.), Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel. Frankfurt/New York: Campus.

Hunt, J. 1992: The Effect of Unemployment Compensation on Unemployment Duration in Germany. Diskussionspapier Nr. 50, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin.

Kalbfleisch, J. D./Prentice, R. L., 1980: The Statistical Analysis of Failure Time Data. New York: Wiley.

Kiefer, N. M., 1988: Economic Duration Data and Hazard Functions. Journal of Economic Literature 26: 646-679.

Klein, T., 1988: Zur Abhängigkeit zwischen konkurrierenden Mortalitätsrisiken. Allgemeines Statistisches Archiv 72: 248-258.

Klein, T., 1990: Arbeitslosigkeit und Wiederbeschäftigung im Erwerbsverlauf. Kölner Zeitschrift für Soziologie und Sozialpsychologie 42: 688-705.

Klein, T., 1992: Zur Zeitabhängigkeit der Wiederbeschäftigungsrate Arbeitsloser. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 134-138.

Licht, G./Steiner, V., 1991: Abgang aus der Arbeitslosigkeit, Individualeffekte und Hysteresis. Eine Panelanalyse für die Bundesrepublik Deutschland. S. 182-205 in: Helberger, C./Bellmann, L./Blaschke, D. (Hrsg.), Erwerbstätigkeit und Arbeitslosigkeit. Analysen auf der Grundlage des Sozioökonomischen Panels. (BeitrAB 144). Nürnberg: IAB.

Lieberson, S., 1985: Making It Count. Berkeley: University of California Press.

Ludwig-Mayerhofer, W., 1990: Arbeitslosigkeit im Erwerbsverlauf. Zeitschrift für Soziologie 19: 345-359.

Ludwig-Mayerhofer, W., 1992a: Fakt und Artefakt in der Analyse von Arbeitslosigkeitsverläufen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 124-133.

Ludwig-Mayerhofer, W., 1992b: Arbeitslosigkeit an der "zweiten Schwelle" - was sind die Folgen? Eine Analyse anhand des Sozioökonomischen Panels (1984-1988). S. 172-189 in: Kaiser, M./Görlitz, O. (Hrsg.), Berufliche Bildung im Umbruch - Zur Diskussion der Übergänge in Bildung und Beruf im vereinten Deutschland (BeitrAB 153.2). Nürnberg: IAB.

Maddala, G. S., 1987: Limited Dependent Variable Models Using Panel Data. Journal of Human Resources 22: 307-338.

Maier, G./Weiss, P., 1990: Modelle diskreter Entscheidungen. Wien: Springer.

Petersen, T., 1986a: Estimating Fully Parametric Hazard Rate Models with Time-dependent Covariates. Sociological Methods & Research 14: 219-246.

Petersen, T., 1986b: Fitting Parametric Survival Models With Time-Dependent Covariates. Applied Statistics 35: 281-288.

Petersen, T., 1993: Recent Advances in Longitudinal Methodology. Annual Review of Sociology 19: 425-454.

Prentice, R. L., 1978: Linear Rank Tests with Right Censored Data. Biometrika 65: 167-179.

Prentice, R. L./Marek, P., 1979: A Qualitative Discrepancy between Censored Data Rank Tests. Biometrics 35: 861-867.

Rohwer, G., 1994: TDA Working Papers. Bremen, Ms.

Schneider, H., 1991: Verweildaueranalyse mit GAUSS. Frankfurt/New York: Campus.

Schneider, H./Hujer, R., 1992: Die Analyse struktureller Wandlungsprozesse mit Hilfe von Paneldaten. S. 39-69 in: Hujer, R./Schneider, H./Zapf, W. (Hrsg.), Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel. Frankfurt/New York: Campus.

Schnell, R., 1994: Graphisch gestützte Datenanalyse. München, Wien: Oldenbourg.

Statistics and Epidemiology Research Corporation, 1991: EGRET Reference Manual and Manual Addendum. Seattle: Statistics and Epidemiology Research Corporation.

Tarone, R. E./Ware, J., 1977: On Distribution-free Tests for Equality of Survival Distributions. Biometrika 64: 156-160.

Teachman, J. D., 1983: Analyzing Social Processes: Life Tables and Proportional Hazards Models. Social Science Research 12: 263-301.

Tuma, N. B., 1980: Invoking RATE. Menlo Park: SRI International.

Tuma, N. B., 1982: Nonparametric and Partially Parametric Approaches to Event-History Analysis. S. 1-60 in: Leinhardt, S. (Hrsg.), Sociological Methodology 1982. San Francisco: Jossey-Bass.

Tutz, G./Georg, W., 1991: Diskrete Hazardraten-Modelle in der Shell-Jugendstudie 1985: Eine Anwendung des Programms GLAMOUR. ZA-Information 29: 81-93.

Urban, D., 1993: Logit-Analyse. Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen. Stuttgart: Gustav Fischer.

Winter-Ebmer, R., 1992: Persistenz von Arbeitslosigkeit. Frankfurt/New York: Campus.

Yamaguchi, K., 1991: Event History Analysis. Newbury Park: Sage.

Ziegler, R./Brüderl, J./Diekmann, A., 1988: Stellensuchdauer und Anfangseinkommen bei Hochschulabsolventen. Zeitschrift für Wirtschafts- und Sozialwissenschaften 108: 247-270.